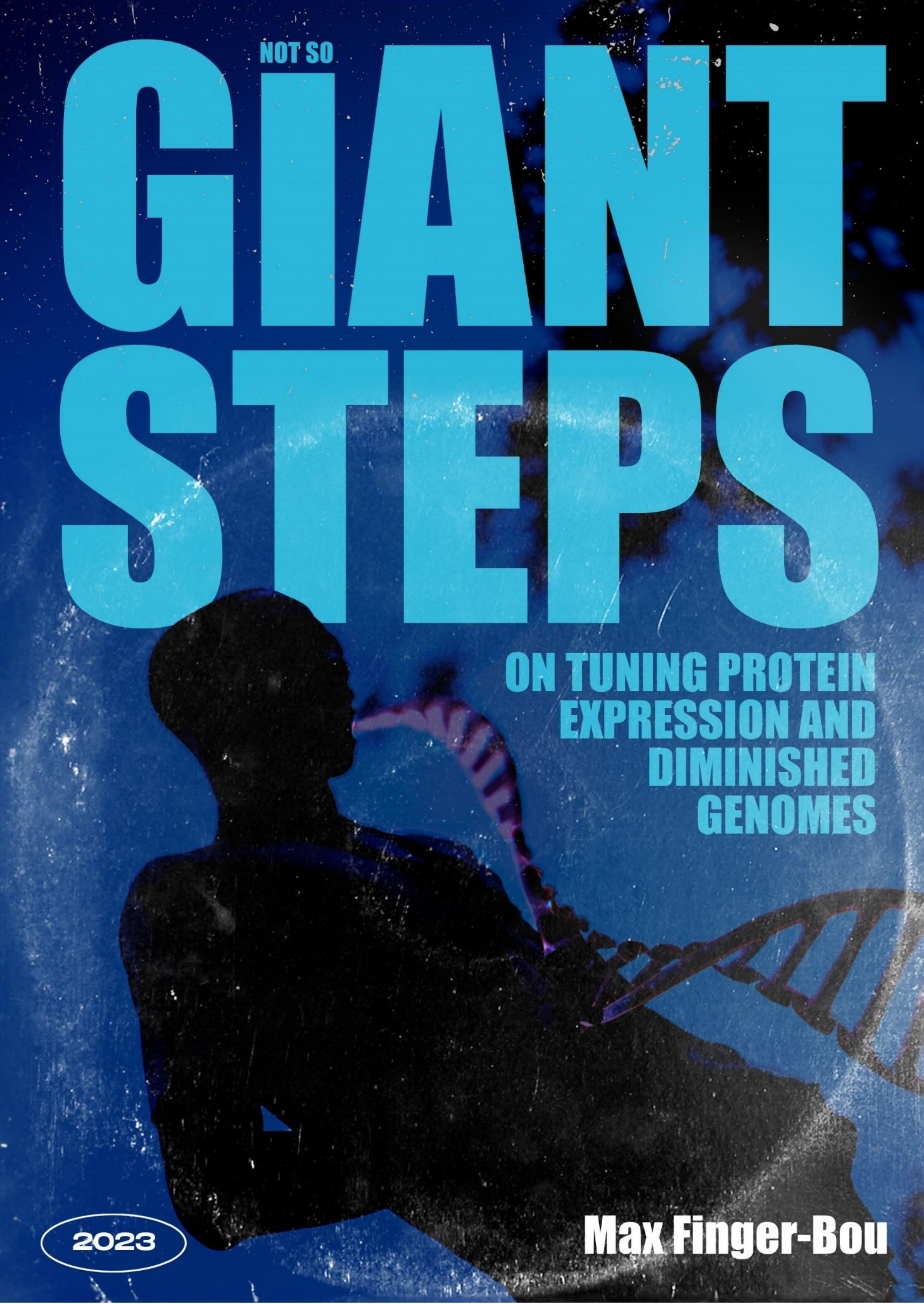


NOT SO

GIANT STEPS

ON TUNING PROTEIN EXPRESSION AND DIMINISHED GENOMES

2023

Max Finger-Bou

Propositions

1. Translation initiation is the major bottleneck in protein expression.
(this thesis)

2. The currently accepted models of prokaryotic DNA repair pathways and their outcomes are incomplete.
(this thesis)

3. Discussing the definition of molecular life is a waste of time.

4. The current academic publishing model is a greatly sophisticated form of slavery.

5. A researcher's training is incomplete without meticulous formation in logical reasoning.

6. Becoming an independent researcher is a paradoxical goal that undermines academic progress.

7. Metaphors are a double-edged sword for bridging science and society.

8. The chaos caused by artificial intelligence in education highlights the need for a paradigm shift in assessment and evaluation practices.

Propositions belonging to the thesis entitled:

Not So Giant Steps On Tuning Protein Expression And Diminished Genomes

Max Finger-Bou

Wageningen, September 8th, 2023

Not So Giant Steps

On Tuning Protein Expression and Diminished Genomes

Max Finger-Bou

Thesis committee

Promotor
Prof. Dr John van der Oost
Personal chair at the Laboratory of Microbiology
Wageningen University & Research

Co-promotors
Dr Raymond H.J. Staals
Associate Professor, Laboratory of Microbiology
Wageningen University & Research

Dr Nico J.P.H. Claassens
Assistant Professor, Laboratory of Microbiology
Wageningen University & Research

Other members
Prof. Dr María Suárez Diez, Wageningen University & Research
Dr Richard A. Notebaart, Wageningen University & Research
Dr Joe Bondy-Denomy, University of California, San Francisco, USA
Prof. Dr Christophe Danelon, Toulouse Biotechnology Institute, France

This research was conducted under the auspices of VLAG Graduate School (Biobased, Biomolecular, Chemical, Food, and Nutrition sciences).

Not So Giant Steps

On Tuning Protein Expression and Diminished Genomes

Max Finger-Bou

Thesis
submitted in fulfillment 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 presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday 8 September 2023
at 4 p.m. in the Omnia Auditorium.

Max Finger-Bou

Not So Giant Steps On Tuning Protein Production And Diminished Genomes

236 pages

PhD thesis, Wageningen University, Wageningen, The Netherlands (2023)

With references, with summary in English

ISBN: 978-94-6447-743-6

DOI: https://doi.org/10.18174/632756

Table of contents

Chapter 1. General introduction and thesis outline

1.1 The early days of biotechnology

Biotechnology, as generally described today, is born from the integration of biology and engineering to generate products or services by using whole organisms, their cells, or parts thereof. Over the last millennia, humanity has made remarkable progress, biotechnology playing a vital role in shaping our relationship with the environment. Despite having a limited understanding of natural resources and their systematic values and mechanisms, humans have used them extensively to their advantage, enabling our species not only to survive, but to thrive.

Perhaps one of the best examples of this is the domestication and breeding of plants and animals, which dramatically influenced the course of human history[1]. Similarly, the technology of fermentation, dated to have been already in use more than 13000 years ago[2], has been applied for thousands of years without a complete understanding of its inner workings while enabling societies to process and preserve different foods[3]. Another example of fermentation and its significance can be found in the emergence of dairying; unknowingly and yet through fermentation, the processing of milk into cheese dramatically increased the durability of milk products, facilitating their transport and turning the raw product into a more digestible asset[4].

1.2 The birth of molecular biology: a brief but intense history

Molecular biology, emerging in the mid-20th century, marked a new era in biotechnology. Innumerable findings and scientific advances related to the mechanisms that enable molecular life have been instrumental in the blooming of biotechnology that we experience today. While deoxyribonucleic acid (DNA) was discovered in 1869[5], it was not until 1952 that Alfred Hershey and Martha Chase helped confirm that it was unequivocally the material responsible for the inheritance of information in biological entities[6]. Soon after, a series of findings culminated in the postulation of the double helix structure of DNA in 1953[7–9].

Studies on DNA replication throughout the same decade led to the discovery of the DNA polymerase[10], which would later become an instrumental part of the revolution in molecular biology. During the 1960s, DNA was proposed and confirmed to be converted into amino acids through a code computing three nucleotides at a time[11,12], and it was found that certain combinations of these nucleotides stop the translation of DNA into protein[13]. It was then discovered that, in fact, an intermediate step between decoding DNA and protein exists, which led to the discoveries of transcription, the different types of RNA and ribosomes[14]. The now often quoted central dogma of molecular biology, first coined by Francis Crick, put forward the idea that the process by which the information is transferred from DNA to protein is not bidirectional, that is, once a protein is generated, the information used to generate such protein cannot be transferred from protein to protein or from protein to nucleic acid[15,16].

The discovery of restriction enzymes[17] prompted the combination of DNA sequences through *in vitro* technology and enabled the technological rise of recombinant DNA[18]. *Thermus aquaticus*, able to withstand temperatures of up to 80ºC, was isolated from a hot spring in the Yellowstone National Park, in the US[19]. The bacterium became a model thermophilic organism, becoming the source of many thermostable enzymes for biochemical and biotechnological experimentation. The isolation of its DNA polymerase[20], able to amplify DNA even at temperatures higher than 75ºC, would later be key in the development of yet another technique that would change the field forever: the polymerase chain reaction (PCR). After several attempts, thermal cycling was included in the experiments that would become the precursor to the one of the most versatile *in vitro* biology techniques, and promising results regarding the amplification of DNA were presented at Cetus Corporation, albeit they were not convincing enough for many[21]. Not much later, the method was improved substantially and enabled the amplification of the beta-globin gene for the diagnosis of sickle-cell anemia[22]. With the appropriate intellectual property claims submitted, a modified version of the original method was published[23], enabling the technique to quickly gain traction in the field. Yet a more final variation of the protocol, implementing the DNA polymerase from *T. aquaticus*, was developed and published[24], quickly becoming a standard and essential technique in the field of molecular biology that is still routinely used today.

1.3 Empowering the central dogma: a journey towards deeper understanding of protein expression

An iconic example of how molecular biotechnology can be used to redesign and improve processes relevant, among others, to human health is that of its effect on the production of human insulin. After the initial discovery of the human hormone in the 1920s, researchers began to harvest insulin from pig pancreatic glands, a process requiring more than two tons of porcine material to yield a mere 200 grams of insulin hormone[25]. Soon after the development of recombinant DNA technology, scientists applied it to introduce the gene coding for insulin into the genome of *E. coli*, successfully enabling the production of a human hormone in bacteria[26].

Today, the field is ever evolving, with developments in gene editing, synthetic biology, and personalized medicine taking place every day. As our understanding of the gene expression process and its underlying mechanisms has advanced, so have our capabilities to manipulate the genome and metabolic pathways of microorganisms to optimize their performance.

DNA is decoded in codons, that is, triplets of nucleotides from a pool of four different nucleotides. Although there are 64 different unique codons, there are only 20 different amino acids being coded by them, and thus most amino acids can be coded by more than one codon. This phenomenon is commonly referred to as the redundancy or degeneracy of the genetic code, and its significance lies in the fact that for given protein of interest, made up of a specific number of amino acids, there is a colossal number of unique DNA sequences that encode the same amino acid sequence. For instance, the medium-sized green fluorescent protein (GFP), made up of 238 amino acids, can be encoded by approximately $3\text{x}10^{110}$ unique open reading frames (ORFs)[27]. Advances in the field have elucidated that different sequences encoding the very same protein can lead to dramatically disparate levels of protein production and functionality thereof, illustrating the complexity of protein expression[28,27].

To begin with, for a gene to be translated into protein, RNA polymerases must first transcribe the DNA into RNA. This necessitates the promoter region to be accessible, as otherwise the polymerase will not be able to realize the process known as transcription initiation. Several factors regulate whether a promoter in a DNA molecule is accessible, but the most notable in prokaryotes are secondary structures arising from sequence complementarity with another sequence and proteins that interact specifically with a certain promoter sequence[29]. How compact the DNA is in the cell can also play a role, as it limits the accessibility to promoters, although the phenomenon is likely more relevant in eukaryotes[30]. Several other steps must be attained correctly for protein synthesis to ensue, many of which involve untranslated regions (UTRs), that is, sequences other than the ORF.

Once a messenger RNA (mRNA), carrying the sequence information to synthesize a protein, is generated, ribosomes must be able to recognize the ribosome binding site (RBS) at the 5'UTR region to begin translation initiation, process regulated by several factors like RBS accessibility and mRNA folding dynamics[28]. Provided that ribosomes can recognize and load onto the mRNA's RBS, several factors modulate protein expression from then onwards. The codon adaptation index (CAI) reflects the deviation in codon usage of a gene sequence compared to a reference set of genes, being one of the many techniques to analyze codon usage bias[31]. Codon bias can also be observed in different tissues of the same organism, and even between its genes[32]. A plethora of types of codon bias has been described so far, and several mechanistic explanations have been hypothesized and experimentally confirmed[27,28].

The codon bias implies that frequent codons in a specific organism will likely be met with a higher abundancy of their cognate transfer RNAs (tRNAs)[33]. While the genetic code can be assumed to be universal, a certain gene sequence from a specific organism with a concrete codon bias might meet completely different abundances of tRNAs when attempting to express it in another organism, hampering the translation of the gene therein. With this in mind, codon optimization algorithms have been developed to alter a DNA sequence from a specific organism to maximize protein expression in another by redesigning it with the most frequent codons in the host species. Nowadays, however, these are considered to be built on oversimplified factors such as codon indices[34] that cannot fully ensure the high or optimal expression of proteins[35] and overlook many important factors such as the effect of codon usage on mRNA secondary structures[36]. Although the discussion is far from concluded, other algorithms aimed at codon harmonization, rather than optimization, are considered to be better in facilitating heterologous protein expression[37], as they attempt to copy the so-called codon landscape, respecting low frequency codons used in the original host context, which are often used to maintain the translation kinetic properties associated with the use of a tRNA at a certain concentration during protein folding[38]. In addition to codon bias, another major feature of mRNA determining protein expression is its half-life[27]. Influenced by secondary structures and UTR sequences, structural sequence elements are one of the factors modulating the stability of mRNA[39], but many others are also at play, such as small antisense RNAs, which are complementary to mRNAs and downregulate their expression by either blocking access to the RBS or by promoting mRNA degradation[40].

Specifically, the heterologous expression of membrane proteins poses tremendous challenges. In addition to the general problems associated with the expression of proteins mentioned above, membrane proteins often contain several highly hydrophobic regions, requiring a specific lipid environment for their proper folding. Moreover, their inclusion in the membrane generally requires the assistance of chaperone proteins and other co-factors that are a limited cellular resource, and hence the saturation of the membrane protein synthesis machinery is often a bottleneck[41,42]. To overcome these obstacles, several strategies have been used, such as codon optimization or harmonization, fusion with solubility-enhancing tags, co-expression with chaperones, or expression in strains specifically engineered to accommodate membrane protein production[43], although a combination of these is often required to successfully produce membrane proteins in settings ranging from industry to academic synthetic biology laboratories.

1.4 Genome engineering tools and DNA repair

Following the revolution in molecular biology of the 20th century, several genome editing tools were developed that facilitate the engineering of microorganisms. In particular, reprogrammable DNA endonucleases such as zinc finger nucleases[44] (ZFNs) and transcription activation-like effector nucleases[45] (TALENs) were a notable addition to the toolbox, as they enabled the induction of DNA double-stranded breaks (DSBs) at specific loci. The discovery of clustered regularly interspaced short palindromic repeats (CRISPR) and the CRISPR associated (Cas) proteins as a prokaryotic adaptive immune system[46–49] completely revolutionized the field of genetic engineering, as they are significantly easier to program than ZFNs or TALENs, tremendously streamlining experiments that depend on the genetic engineering of any organism. As such, CRISPR-Cas tools have gained popularity to the point that they are considered the preferred programmable endonuclease for the genetic engineering of most genetically accessible species[50,51].

ZFNs, TALENs and CRISPR-Cas endonucleases are normally used to generate DSBs, which are lethal if unrepaired and initiate the recruitment of DNA repair processes to prevent death[52,53]. As of today, three major DNA repair pathways have been described in prokaryotes: homology-directed repair (HDR), non-homologous end joining (NHEJ) and alternative end joining (AEJ)[54].

Mediating the high-fidelity repair of DSBs in bacteria, HDR is accepted to be the main mechanism by which most prokaryotes circumvent death by DSBs[55,53], and it requires an intact copy of the genetic region to repair[56]. While several enzymes are involved in this type of repair, the RecBCD complex is generally considered its main actor, processing DSBs and generating DNA ends with a 3' extension that is recognized by the RecA recombinase, in turn, responsible for promoting strand invasion into the intact DNA template, DNA repair taking place and the initial sequence being restored thereafter[57]. In scenarios whereby an intact DNA template is not available, two other DNA repair pathways can act to prevent cell death. Non-homologous end joining (NHEJ), on the one hand, is able to protect DNA ends via the ring-like structure-forming Ku protein[58], and process and repair them by action of the multi-functional ligase LigD[59] often introducing mutations at the repair site[60]. Alternative end joining (AEJ), on the other hand, depends on the RecBCD complex to process DNA ends until short homologies of varying length are exposed and can hybridize, the following ligation being executed by the DNA ligase LigA[61], frequently resulting in genomic deletions ranging from a few to several dozen kilobases[61].

From transposon-based methods[62], bacteriophage-derived techniques[63] and homologous recombination-based systems[64], a myriad of tools are currently available and have been successfully employed to genetically engineer a multitude of organisms. Systems relying on homologous recombination are often chosen to perform precise genome modifications in bacteria, but they have a few drawbacks. To begin with, the introduction of exogenous DNA templates for recombination in bacteria necessitates the bacterium to have rather proficient genetic accessibility, as the introduced DNA templates often need to be of multiple kilobases for the recombination efficiency to be sufficient and transformation efficiency is generally impaired with increasing plasmid sizes[65]. Additionally, a genome of interest must be characterized to design proper DNA templates for its editing, which is not always possible with novel species, and not all microbial species have efficient homologous recombination facilitating this type of repair.

The use of non-templated DNA repair pathways such as NHEJ and AEJ can be a useful alternative in cases where homologous recombination is not efficient enough[54]. The development of class 1 CRISPR systems for genome editing creates opportunities to explore the effect of big deletions in the genome of bacteria[66]. Along these lines, the field of genome minimization, also broadly known as top-down synthetic biology, aims to strip down genomes from non-essential genes and can be of great use to understand aspects related to fundamental genetics, its ultimate goal being the generation of microbial cell factories with more efficient metabolisms tailored to produce specific compounds of biotechnological interest.

1.5 Microbial engineering to fight global challenges

While an invaluable tool to propel humanity towards a growth-based economy, unregulated technological progress tends to come at a cost. The proliferating global population and its habits have acclimated to an unsustainable use of the planet’s resources, resulting among others in the depletion of big part of the existing fossil fuels and in the deforestation of a considerable part of the rainforests of our planet. Altogether, this has promoted climate change and global warming, which tragically but steadily propel our ecosystems into a regrettably anticipated collapse.

Decoupling conventional growth from the emission of greenhouse gases is vital to mitigate and remodel the impact of our species on the environment[67].

In the hopes to transition to a sustainable economy powered by biotechnology, several strategies have been developed to replace the synthesis of petroleum-based chemicals by that of biocatalysis. The engineering of microbial cell factories can help uncover biological processes by which renewable feedstocks can be turned into several products, from pharmacological assets to commodity chemicals and biofuels[68,69]. While several products are already being produced by microbes today, many of the initial efforts have been focused at the production of compounds with high market value[70,71], leaving industrial biotechnology yet to achieve the microbial production of many bulk chemicals, the production of which still depends on oil-based chemistry. The development of sustainable microbial cell factories requires the careful selection of a microbial host to carry the biological processes of interest, and various microbial species have been employed so far.

First described by Theodor Escherich in 1885[72], *Escherichia coli* is arguably one of the best characterized organisms on the planet and has become the go-to microorganism to perform routinary molecular biology work in most laboratories[73,74]. Thanks to the outstanding advances in the field, *E. coli* and its metabolism have been successfully engineered to perform previously unimaginable bioconversions, such as the synthesis of sugar from CO_2[75], a greenhouse gas, or the assimilation of formate, a sustainable feedstock, into biomass[76] for the production of different chemicals. As versatile as it is, however, *E. coli* has limitations that cannot truly be overcome through its genetic and metabolic engineering, which implores the exploration of different microorganisms tailored to specific interesting biological processes.

Along these lines, another bacterium that has drawn great interest in the biotechnological field is *Rhodobacter sphaeroides*, a purple photosynthetic bacterium. Belonging to alphaproteobacterial, *R. sphaeroides* can fix nitrogen and excels at anaerobic phototrophy but can also grow heterotrophically via fermentation and both aerobic and anaerobic respiration. Its versatile metabolism has made it not only a model organism for research on photosynthesis[77,78], stress regulation[79,80] and chemotaxis[81,82], but also for the production of terpenes[83,84] and has become a model microbial cell factory, already being used for the production of several compounds[85]. The untapped potential of any microorganism, however, will be as great as the tools available to engineer it. It is therefore imperative to critically assess the state-of-the-art of the genome editing tools at hand to ensure that the best strategies are followed to carry out a specific metabolic engineering feat.

1.6 Thesis outline

This thesis provides an overview of the major topics and findings in the fields of gene expression and CRISPR-based template-free genetic engineering, exploring various strategies relevant for the fields of gene expression, genetic engineering, and synthetic biology.

In **Chapter 2**, recent advances of the state-of-the-art principles of gene expression are reviewed. Dissecting the influences of both protein-coding and non-coding sequences on protein production, the molecular mechanisms governing transcription, mRNA decay and translation are discussed. While these factors are all but orthogonal, new technological breakthroughs in high-throughput analyses and molecular biology enable the update of the strategies currently employed to design genes and genetic constructs to optimize protein production.

In **Chapter 3**, a strategy combining a constitutive promoter with a library of BiCistronic Design (BCD) elements is developed and applied to produce functional membrane proteins in *Escherichia coli*. Exploiting tuned translation initiation, the method enables inducer-free, functional membrane protein expression of proteins relevant for synthetic biology applications, as well as proteins previously proven to be hard to express.

In **Chapter 4**, the two known prokaryotic template-independent DNA repair pathways are reviewed, with a focus on their biotechnological application in combination with CRISPR-Cas tools. Through the exploitation of either the non-homologous end joining (NHEJ) or the alternative end joining (AEJ) DNA repair pathways, different strategies for the genetic engineering of prokaryotes are discussed.

In **Chapter 5**, a CRISPR-Cas9-based method for the non-templated genome editing of *Rhodobacter sphaeroides* is presented. Revealing molecular details on the NHEJ and HDR DNA repair pathways, several mutant strains lacking one or multiple genes involved in the DNA repair pathways are assessed. Additionally, issues arising from the use of 5-fluorouracil as a counter-selecting agent are examined, specifically the spontaneous appearance of microhomology-flanked deletions and insertions in the experimental setting.

In **Chapter 6**, a method based on a type I-C CRISPR system combined with bacterial conjugation is developed and assessed to perform the iterative genome minimization of *E. coli* MG1655. Exploiting the cycling of different antibiotics, the technique is reported to mediate genomic deletions of up to 110 kb in the bacterium. However, the protocol is not able to generate strains with multiple mutations as initially designed. With some tweaking, the method will likely be able to mediate the generation of strains carrying several large deletions in a record amount of time, speeding up efforts in the top-down synthetic biology field.

In **Chapter 7**, a summary of the thesis is outlined discussing the major achievements and reflections therein. Lastly, potential future directions stemming from research included in this manuscript and from late developments in the field are discussed.

Chapter 2. The ongoing quest to crack the genetic code for protein production

Thijs Nieuwkoop[1], **Max Finger-Bou[1]**, John van der Oost[1], Nico J. Claassens[1]

[1]Laboratory of Microbiology, Wageningen University and Research, Wageningen, the Netherlands

Adapted from publication:

Nieuwkoop, T., **Finger-Bou, M.**, van der Oost, J., & Claassens, N. J. (2020). The ongoing quest to crack the genetic code for protein production. *Molecular cell*, *80*(2), 193-209.

2.1 Abstract

Understanding the genetic design principles that determine protein production remains a major challenge. Although the key principles of gene expression were discovered 50 years ago, additional factors are still being uncovered. Both protein-coding and non-coding sequences harbor elements that collectively influence the efficiency of protein production by modulating transcription, mRNA decay, and translation. The influences of many contributing elements are intertwined, which complicates a full understanding of the individual factors. In natural genes, a functional balance between these factors has been obtained in the course of evolution, whereas for genetic-engineering projects, our incomplete understanding still limits the optimal design of synthetic genes. However, notable advances have recently been made, supported by high-throughput analysis of synthetic gene libraries as well as by state-of-the-art biomolecular techniques. We discuss here how these advances further strengthen the understanding of the gene expression process and how they can be harnessed to optimize protein production.

2.2 Introduction

The biosynthesis of proteins is one of the core processes in living cells, as well as in many biotechnological applications. It has already been 50 years since Francis Crick proposed the central dogma of molecular biology[16], explaining how DNA is transcribed to mRNA, which is then translated to protein. A characteristic feature of the conversion of the information stored in the nucleotide building blocks of DNA and mRNA into the amino acid building blocks of proteins is the redundancy in the number of codons on the nucleotide level. Although there are 64 unique codons (nucleotide triplets), only 20 different amino acids make up proteins in most organisms. This redundancy gives astronomical numbers of codon combinations to encode the same amino acid sequence, e.g., the medium-size green fluorescent protein (GFP, 238 amino acids) can be encoded by 3 x 10^{110} different open reading frames (ORFs).

However, different sequences encoding an identical protein sequence can lead to dramatic variations in protein production levels, and sometimes even lead to differences in protein folding and functionality[37,86,87] (Figure 1). Apart from ORFs, non-coding regions with potential regulatory functions, such as promoters and untranslated regions (UTRs; Figure 1) add a vast sequence space. As the design principles of both the coding and non-coding sequences are only partly known, the design of synthetic genes for expression is still a major challenge.

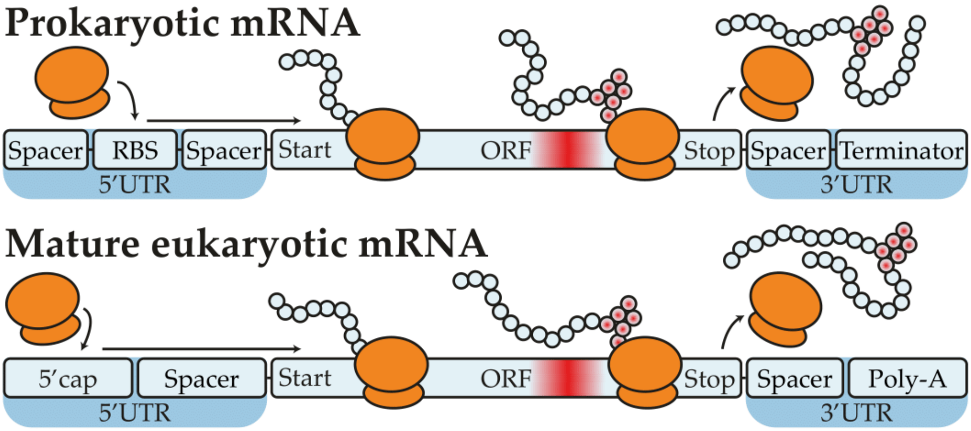


Figure 1: Schematic overview of a prokaryotic and mature eukaryotic mRNA being translated by ribosomes. *RBS, ribosome binding site; ORF, open reading frame; 5'/3'UTR, 5'/3' untranslated region. The co-translational folding phenomenon is indicated with a red gradient in the mRNA and the associated amino acids.*

Already, since the early days of gene sequencing in the 1980s, a bias has been recognized in the codon usage of highly expressed native genes; particular synonymous codons (i.e., different codons encoding the same amino acid) were observed to be used more frequently than others. This notion led to the formulation of the Codon Adaption Index (CAI)[31], and it was postulated that the codon bias within highly expressed genes allowed for more-efficient translation. An underlying hypothesis to this observation is that the (amino-acid-charged) cognate tRNAs for these frequent codons are more abundant and that they are more-efficient decoders during ribosomal protein biosynthesis[33,88]. In recent decades, the advent of high-throughput sequencing technologies has revealed more codon usage signatures varying across organisms, tissue types, and genes[32].

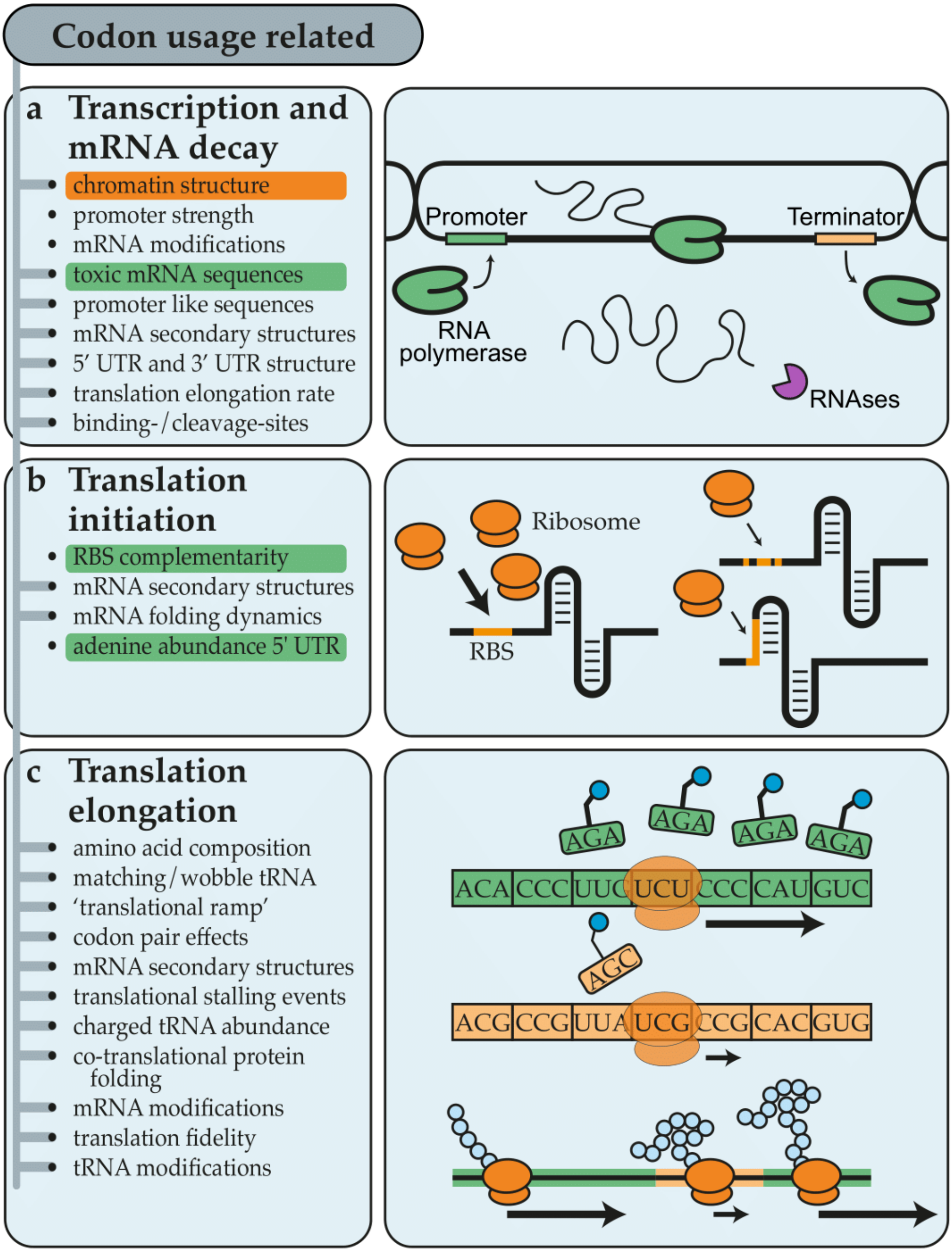


Figure 2: An overview of reported factors involved in protein production. *Factors at the level of (a) transcription and mRNA decay, (b) translation initiation, and (c) translation elongation. Factors that can be related to codon usage are connected by the gray bar. Factors*

that have only been well-described in eukaryotes (orange) or prokaryotes (green) are highlighted; other factors have been observed in both domains of life.

Following these observations, several types of codon bias and mechanistic explanations were introduced[28]. The current view on codon usage is that it is related to a complexity of factors. The weight of those factors varies depending on the context, which includes the type of organism, tissue, or compartment; physiological control (e.g., pathway or growth phase); or even the position within an ORF[28,32]. It became clear that the notion of frequent versus rare codons, similar to good versus bad codons for protein production, is an oversimplification of biological reality. Consequently, codon optimization algorithms, which are all based on simplified assumptions and codon indices[34], cannot warrant successful heterologous protein production [35]. Because codon choices are related to diverse mechanisms and regulatory processes, we prefer to use the term "codon optimality" only when a range of factors acting at different levels of the expression process have been taken into account.

A couple of years ago, we reviewed the effect of codon usage within the ORF on expression[28]. Impressive advances have been made in the field since then, because, on one hand, of the technical advances, including high-throughput analyses of large synthetic gene libraries[89] and, on the other hand, because of innovative molecular biology approaches that unravelled additional details of transcription, translation, and protein folding[37,86,90]. These studies contributed to a further understanding of some of the factors involved and have also revealed relevant interactions among them.

Here, we provide a timely overview of the field of gene expression, discussing relevant features both in the regulation of non-coding regions and in ORFs. As transcription, mRNA decay, and translation (initiation and elongation) all have important roles in controlling protein production, we discuss all these stages (Figure 2). Furthermore, we highlight key controversies and knowledge gaps in the field and propose potential avenues to resolve these. Lastly, we discuss how our relatively poor understanding of optimal gene designs is a major limitation for biotechnology and synthetic biology. We examine how emerging tools and approaches can aid in overcoming challenges for engineering protein production.

2.3 Transcription and mRNA decay

2.3.1 Transcription initiation

The first step in protein production is the transcription of DNA to mRNA by RNA polymerase (RNAP). Synthesis rates of mRNA are mediated by the binding affinity of RNAPs and related transcriptional factors with the promoter sequences; other factors, such as chromatin structures in eukaryotes, also have a role[30]. In addition, the transition from transcription initiation to transcription elongation is important in determining mRNA synthesis rates. After the RNAP is bound, DNA is unwound, and an open complex is formed. During the open complex configuration, the first short RNA stretch is transcribed, and then, the RNAP either moves on to transcribe the full mRNA (promoter escape) or the initiation is aborted. Several promoter sequence features, for example, the length and nucleotides in the bacterial discriminator region (±4–7 bp upstream of the transcription start), determine the efficiency of the promoter escape[91,92]. Promoter sequence regions, as well as transcription initiation and elongation factors involved in promoter escape, are reviewed in more detail elsewhere[93,94]. Although most of the key principles of transcription initiation and promoter escape are known, models to predict promoter strengths from sequences are still under development.

Recently, several groups investigated promoter properties and design constraints by expressing some reporter genes from libraries with randomized promoter sequences. Some studies in *Escherichia coli* reported that, of all fully randomized promoter sequences, 7-10% resulted in detectable expression[40,95]. Furthermore, it was found during laboratory evolution of random sequences in *E. coli* that 60% of those sequences became functional promoters with only one mutation[95]. Functional promoters in *E. coli* were generally observed to have at least a canonical -10 or -35 motif for binding the RNAP-sigma subunit, which occurs relatively frequently in DNA sequences by chance. Another study randomized the yeast -90 to -170 promoter region, whereas the consensus TATA region was kept constant, which resulted in detectable expression for 83% of the sequences[29].

The increasing data on characterized (random) promoters has also been used to create predictive models. Such *in silico* predictions have been successful for predicting promoter strengths of yeast, by modelling the transcription factor binding sites and their accessibility[29,96]. However, the generation of predictive models for *E. coli* based on a set of fully randomized and native promoters by machine learning was still unsuccessful[40]. This may be explained by the diverse sigma-factor-type promoters that are included in the training set. A previous study that performed machine learning and regression only on sigma-70 “household” promoters in *E. coli* did result in good predictive models[97].

Apart from the influence of promoter regions on transcription, it was observed in some eukaryotes that the codon or nucleotide usage within an ORF might also affect transcription rates[98–100]. Proposed mechanisms through which nucleotide composition or codons could modulate transcriptional activity are related to histone modifications or the influence of GC-content on transcription elongation rates.

2.3.2 mRNA decay

All cells harbour several endo- and exo-ribonucleases that are involved in degrading mRNA, providing additional control over mRNA levels and protein production[101]. Furthermore, ribonucleases can clean up non-functional RNAs, e.g., from accidental transcription. The dynamics between mRNA transcription and mRNA decay result in a wide range of mRNA half-lives, serving as one of the key factors for protein production[102–104].

One of the factors modulating mRNA stability is the presence of structural elements in their untranslated regions. Secondary structures and sequences of UTRs can influence mRNA decay rates, especially in bacteria[39]. Recently an increasing number of studies demonstrated the important role of the 3'UTR region in controlling mRNA decay[105,106]. For the 5' UTR, it is harder to determine the effect of the sequence itself on mRNA stability because that region also has a key effect on translation initiation. In eukaryotes, 5' caps and 3' poly-A tails (Figure 1) are the primary features of the UTR regions that protect mRNAs from degradation[107].

Diverse, alternative polyadenylation mechanisms in eukaryotes are activated by different signals in 3'UTR sequences and lead to differing poly-A tails and 3'UTR lengths; this region is highly interactive with RNA binding proteins, microRNA and long noncoding RNAs. These interactions and the 3'UTR length influence mRNA stability and decay, but also influence mRNA translation, as extensively reviewed elsewhere[108].

In the past decades, it has been suggested that the translation process may influence mRNA stability in yeast, as reviewed previously[32]. More recently, this connection gained additional attention in extensive studies in a range of eukaryotes, which all clearly demonstrated a positive correlation between the presence of certain codons in ORFs and the stability of the corresponding mRNAs[104,109–117]. In particular, specific codons are observed to be more abundant in mRNAs with a longer half-life. This observation was captured by a newly proposed codon index, the codon stability coefficient (CSC), which can be calculated for each codon as the correlation coefficient between the codon frequency in transcripts and their mRNA half-life[104] (Figure 3a). In several studies, it was found that this coefficient correlates moderately with the tRNA availability index (tAI). The latter index is based on the gene copy number of tRNAs available to decode a certain codon[88,104]. The observation that codons leading to high mRNA stability seem related to more-abundant tRNAs, remarkably suggests that the translational process may influence the stability of mRNAs. This was further supported by experiments that compared the mRNA stability with and without blocking the translation process[109,118]. These experiments showed that when translation is inhibited, the mRNA half-life times are reduced, especially for transcripts with high "codon optimality."

On top of codon identity, a link is also suggested between amino acid identity and mRNA decay. A few amino acids are also specifically correlated to more or less stable mRNAs[109,111,116,118]. It is hypothesized that for these amino acids' higher or lower intracellular concentrations influence the amount of available tRNAs for translating those amino acids and hence influence translation elongation rates and consequently mRNA stability. In summary, several lines of evidence suggest that faster translation elongation leads to higher mRNA stability.

A potential molecular mechanism connecting translation elongation rates to mRNA decay has recently been unraveled (Figure 3b). Clear evidence was found in yeast that the de-adenylating Ccr4-Not complex directly interacts with ribosomes that are not loaded with a new tRNA in their A-site[90]. Hence, this complex can sense slow-moving ribosomes and then triggers de-adenylating of the poly-A tail; after which, the RNA helicase Dhh1p activates de-capping, eventually resulting in mRNA decay[115,119,120].

A link between codon usage and mRNA stability was also suggested for the bacterium *E. coli* to have a major role in protein production efficiency[102]. This study focused on expression data from a large set of plasmid-encoded heterologous genes transcribed by T7 RNAP. So far, no genome-wide analyses are available on such correlations in bacteria for native gene expression.

In relation to that, it is interesting to note here that recent structural studies in *E. coli* and *Mycoplasma pneumonia* clearly show that the RNAP complex can be linked to ribosomes in a so-called expressome, which leads to the coupling of transcription elongation to the translation process[121]. However, it was also recently reported that this coupling is not present in all bacteria because it was demonstrated in *Bacillus subtilis* that its RNAP moves faster than its ribosomes, in so-called runaway transcription[122]. The consequences of the presence and absence of this mechanism in different bacteria for the influence of codon usage and translation elongation on transcription deserve further analysis.

Lastly, another mRNA-mediated mechanism was discovered in *E. coli*, in which specific heterologous sequences of the mRNA appear to be toxic to the bacterial cells. It is not uncommon that the expression of heterologous proteins causes growth retardation in the expressing host, usually related to a protein production burden. However, a recent study surprisingly demonstrates that the growth retardation for specific heterologous mRNAs still happens when translation is blocked[123]. It is hypothesized that specific mRNA secondary structures cause toxic effects in the cell via a yet unknown mechanism.

Overall, our understanding of control mechanisms that determine mRNA concentrations is increasing. It is clear that mRNA abundance is affecting the downstream translational process and, remarkably, also vice versa translational processes seem to exert control on mRNA levels.

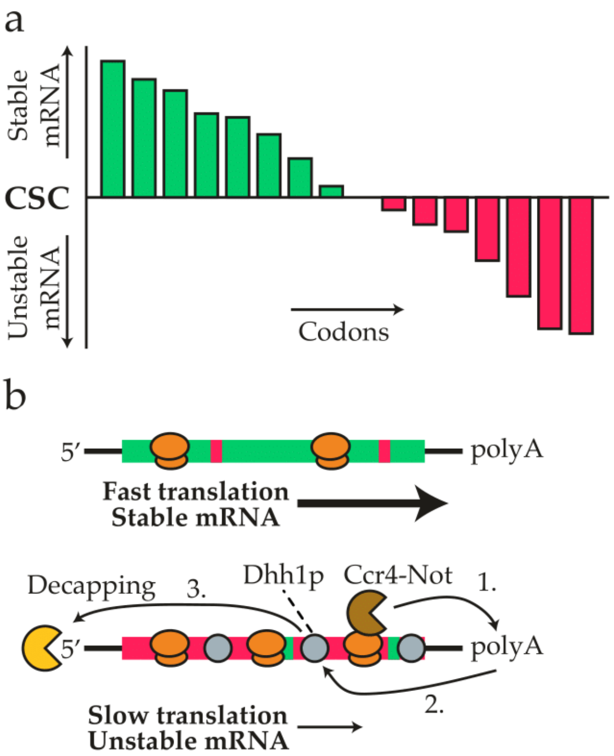


Figure 3: Codon Usage and Translation Elongation Are Related to mRNA Stability in Several Eukaryotes. *(a) A schematic representation of a codon-stabilization coefficient (CSC) plot, based on recent studies in several eukaryotes, e.g., Presnyak* et al.*[104]. Bars for each codon represent the correlation between the codon frequency in the transcripts and the*

half-life of the transcripts. Positive correlations (green) indicate codons that are more abundant in mRNAs with a longer half-life time, whereas negatively correlated codons (red) are overrepresented in less-stable mRNAs. For illustrative purposes, only a few codons are depicted; in a real plot, the CSC value for all 61 amino-acid-encoding codons would be shown. (b) mRNAs with more codons with a high, positive CSC value (green) are observed to be translated faster by the ribosomes because, for example, those codons have more abundant cognate tRNAs. In the eukaryotic model organism yeast, a molecular mechanism has been elucidated that can explain the connection between slowly translated mRNAs and mRNA decay rates. The de-adenylating Ccr4-Not complex can directly interact with ribosomes that are not loaded with a new tRNA in their A-site[90]*. Likely, this complex senses slow-moving ribosomes and then triggers de-adenylating of the poly-A tail, and next the RNA helicase Dhh1p activates de-capping and subsequent mRNA decay.*

2.4 Translation initiation

For transcripts to be translated into protein, ribosomes need to associate with the 5'UTR of the mRNA and start translating the ORF from the start codon. The translation initiation process is considered one of the most influential steps in translation efficiency.

In prokaryotes, it is generally assumed that translation initiation begins when the 30S ribosomal subunit recognizes a ribosome binding site (RBS) in the 5'UTR. The RBS usually contains a Shine-Dalgarno (SD) sequence, which has high complementarity to the 3' end of the 16S rRNA of the 30S ribosomal subunit, the so-called anti-Shine-Dalgarno sequence (aSD)[124]. In eukaryotes, the ribosome binds the 5' cap or an internal ribosome entry site (IRES) and usually translation initiation is further controlled by a Kozak sequence[125], a motif surrounding the start codon with a relatively high abundance of adenines[126]. However, because most recent studies on translation initiation used *E. coli* as a model, we mostly discuss prokaryotic translation initiation. For detailed insights on translation initiation and the 5'UTR in eukaryotes, we refer to other recent reviews[126,127].

Numerous studies, mostly investigating heterologous protein production in *E. coli*, have found that strong mRNA secondary structures around the

RBS/SD region severely hamper translation initiation[102,89,128,129]. The mRNA folding in this region is also regularly observed to be influenced by the codon usage at the start of the ORF. A recent study aimed to quantify the influence of mRNA secondary structures more accurately by designing strong RNA hairpins in the 5'UTR region of a reporter protein. Although secondary structures located far from the SD only result in less than 2-fold repression of translation, secondary structures close to the SD were shown to repress translation more than 100-fold; the repression levels are proportional to the free energy needed to unfold the RNA hairpins[130]. Furthermore, a study that introduced synonymous codon mutations throughout ORFs of two native *E. coli* genes revealed that, especially mutations leading to relatively strong, predicted mRNA secondary structures that include the RBS, result in significantly decreased protein production levels[131].

Although most studies base their mRNA structure predictions on *in silico* folding energy models, some recent studies have applied transcriptome-wide *in vivo* experiments to determine mRNA secondary structures. Experimental high-throughput measurements of mRNA secondary structures can be performed by cell-permeable chemicals that react selectively with non-paired RNA bases, e.g., SHAPE probes that acylate 2' hydroxyl groups of unpaired nucleotides (SHAPE-MaP)[132] or dimethyl sulfate that modifies unpaired adenine and cytosine residues (DMS-seq)[133]. As the next step, cDNA is generated from the chemically modified RNAs, and next-generation DNA sequencing allows for mapping of the modifications in non-structures regions and, hence, allows the elucidation of non-structured and structured mRNA regions. One of these studies, based on SHAPE-MaP in *E. coli*, demonstrated that the translation efficiency of native genes is, in large part (40%), determined by mRNA structures covering the RBS[134].

The improved resolution of mRNA structure measurements also allowed the study of two alternative models for translation initiation: the equilibrium model and the kinetic model. In the equilibrium model, the ribosome, once bound, remains and creates a new equilibrium mRNA secondary structure. In the kinetic model, however, there is a continuous competition between the unfolding and refolding of the mRNA and association and dissociation of the ribosome. Experimental data, as well as a theoretical biophysical approach, now suggest the kinetic model best explains translation initiation in *E. coli*[134,135]. This also allows for "ribosome drafting" in some highly translated mRNAs, a mechanism in which successive ribosomes bind an mRNA faster than the mRNA can refold.

In contrast with the ribosome drafting mechanism, in eukaryotes, it was observed that ribosome clearance around the translation initiation site is required for high-expressing genes. It is suggested that codons directly after the start codon need to mediate relatively fast translation elongation to free up space for the next ribosome to initiate translation[136].

Although it is generally accepted that SD-aSD interaction is the main player involved in prokaryotic ribosome loading, new findings hint at alternative mechanisms regulating ribosome recruitment and translation initiation. Several bacterial species, for example, *Flavobacterium johnsoniae*, naturally lack SD sequences. In this species, it was observed that at some key nucleotide positions upstream of the start codon (-3, -6, -13, and -23), the presence of adenine nucleotides is a positive determinant for translation initiation[137]. The molecular basis for this observation is currently not known but, as the authors state, it seems reminiscent of the eukaryotic Kozak sequence, which also shows a preference for adenine at position −3. Furthermore, some recent *E. coli* studies on native and reporter gene expression report an enrichment in adenines at sites mostly upstream, or shortly downstream of the start codon for well-expressed genes[138,139]. It was demonstrated experimentally that these A-rich sequences contribute to the identification of translational start sites, suggesting that these adenines could be highly conserved as an alternative mechanism for start site selection in bacteria[139].

2.5 Translation elongation

2.5.1 Codon usage and translation rates

After successful initiation, ribosomes continue with translation elongation, i.e., the sequential decoding of the codons of the mRNA to synthesize the corresponding amino acid sequence. The effect of codon usage during translation elongation has been extensively studied by multiple methods, however, often leading to contrasting conclusions. A popular hypothesis is that codon usage controls the speed of ribosomal translation elongation. The underlying assumption is that translating ribosomes slow down when they encounter "sub-optimal" codons, e.g., codons that are decoded by less-abundant (amino-acid-loaded) cognate tRNAs or by lower-affinity-matching tRNAs through wobble base-pairing.

A decade ago, the ribosome profiling technique was developed to monitor translation elongation rates in a high-throughput manner[140]. This approach is based on the high-throughput sequencing of ribosome-protected mRNA fragments, providing a snapshot of ribosome density throughout the transcriptome. Initially, differences in experimental ribosome profiling protocols and subsequent data analysis led to conflicting conclusions on whether translation elongation speeds are influenced by codon usage or not[141–144]. However, in recent years ribosome profiling protocols and data analysis were refined, e.g., by the use of flash freezing to stall translation, instead of the use of cycloheximide[145]. Improved protocols led to a better consensus that codon usage may influence the translation elongation speed, but that this effect is rather weak and that a multitude of other factors are also involved[32].

Recently, considerably more sensitive approaches that use cell-free translation systems[37,146] and *in vivo* imaging of nascent polypeptide synthesis[147,148] have been established. These methods all confirmed that heterologous mRNAs with "optimal" codon usage are translated faster. However, these studies monitored the strong contrast between synthetic genes that were designed to have almost only optimal codons with non-optimized genes. Within natural genes, which often have fluctuating use of optimal codons along the ORF, translational speed differences are generally more subtle.

It was also demonstrated for eukaryotic translation, both *in vivo* and *in vitro*, that rare codons sometimes not only slow down translation, but they can even stall part of the elongating ribosomes, leading to premature translation termination[146,149,150].

2.5.2 Does an mRNA secondary structure influence translation elongation?

Besides the influence of codon usage on translation speed, the mRNA secondary structure within an ORF was also suggested as influencing translation elongation. However, until recently, it was hard to verify that hypothesis because only rough *in silico* predictions of mRNA folding energy were available to estimate mRNA structures. However, the aforementioned development of several experimental protocols allows for probing RNA structure *in vivo* at a transcriptome-wide scale. Two studies in this field used different methods to both reach the conclusion that translating ribosomes in *E. coli* dissolve RNA secondary structures[134,151], which is in line with the demonstration that the *E. coli* ribosome exhibits helicase activity[152].

Apart from that finding, the DMS-seq analysis by Burkhardt *et al.*[151] reported a strong correlation between mRNA secondary structures in an ORF and its translation elongation efficiency, suggesting that at least some of those structures can still be an obstacle for translating ribosomes. In contrast, the SHAPE-MaP analysis by Mustoe *et al.*[134] could not confirm that correlation. Hence, despite advances in *in vivo* RNA structure mapping, it remains unclear to what extent mRNA structures influence translation-elongation rates. Refinement and application of these methods throughout multiple organisms are required to clarify this matter.

2.5.3 Co-translational folding mediated by the ORF sequence

For a few specific proteins, single-molecule approaches have been used to accurately monitor translation elongation rates and related co-translational protein-folding processes. In some cases, it was clearly shown that the slow-down of translation elongation is crucial to facilitate proper co-translational folding of the nascent protein[37,86].

Similarly, it has been demonstrated *in vivo* for some eukaryotes that codon usage is crucial for the folding and functionality of some circadian clock proteins, especially for the unstructured domains of these proteins. When the sub-optimal codon usage in unstructured regions of these circadian clock genes, as well as in a luciferase reporter gene, was changed to a more-optimal codon usage, the *in vivo* functionality of these proteins was compromised [87,146,153,154]. This folding hypothesis is further supported by broad bioinformatic analyses of genes from several organisms, based on which correlations are reported between less-optimal codons in unstructured regions in between more-structured protein domains[154,155]. Despite the fact that these unstructured domains do not form defined structures (alpha helices or beta sheets), they seem to have certain folds (e.g., coils) that can be essential for their functionality. These studies suggest that translation slows down to facilitate folding either of these unstructured domains themselves or at structural junctions between structured and unstructured domains.

However, a broader analysis of clusters of rare codons throughout many genomes in all domains of life challenges this observation of rare codons within unstructured domains[156]. That study, in fact, reports an enrichment of rare codons within structural domains, suggesting that translational slow-downs may be specifically relevant for the folding of smaller structural sub-elements. As an example, they show conservation of rare codon clusters for two proteins at the same “structural” positions throughout different organisms. Providing such comparative analyses for more proteins, as well as performing functional experiments on these, could strengthen the proof that sub-optimal codons are also relevant within structural protein domains.

Overall, there is clear case-based evidence on the effects of codon bias and translational speed on co-translational folding for some specific proteins. However, interpretation of these effects on a genome-wide scale is complicated, given the limited understanding of the genetic features determining the translational speed and the subjective definitions of optimal and non-optimal codons. Furthermore, determining the relevance of the coding sequence on protein folding is challenging, as it is currently not possible to experimentally determine protein structures or folding processes in a high-throughput manner.

2.5.4 Translation effects at the start of the ORF

Another frequently reported and heavily debated observation is the slower translation at the 5' end of an ORF. Some evidence for this has been based on ribosome profiling data and the higher frequency of rare codons in the first part of the ORF [157,158]. A main hypothetical explanation for the presence of a so-called translational ramp at that location is the distancing between ribosomes to prevent detrimental ribosomal collisions. Still, there are alternative explanations for the observed codon bias at the 5' of ORFs. A key alternative hypothesis is that a strong selection against mRNA secondary structures at the 5' end to facilitate translation initiation of highly expressed genes is more important than the selection pressure for well-translated codons in that region of the ORF.

Interestingly, several studies that randomized synonymous codons in *E. coli*, usually for GFP as a reporter protein, found strong correlations between protein production and reduced mRNA secondary structures around the 5' end of the ORF[128,129,159]. A recent study tried to resolve the factors in the 5' end of the ORF in a more systematic way by designing >200,000 different N-terminal tags for 32 codons, followed by a GFP reporter gene[89]. Several factors were varied in the N-terminal library design, including the presence of different-strength translational ramps, as well as the presence of mRNA secondary structures at different positions. Although no correlation was detected between translational ramps and expression, that study did demonstrate a major role of mRNA structural elements in RBS availability and, consequently, in overall protein production. However, as the authors admit, the conclusion that the presence of a translational ramp could not be detected in that study might have been the result of non-optimal design. Although it remains unclear to what extent translations ramps influence expression levels, it was demonstrated recently that a ramp can decrease the resource costs of expression[160], likely by preventing ribosome jamming and translational abortion events[158].

2.5.5 Other factors observed at the translational level

Apart from the effect of single codons on translational dynamics, it was observed previously that specific codon pairs might also influence translational processes[161,162]. In yeast, ribosomal stalling has been reported for a small subset of codon pairs, mostly when they occur in a specific order[163]. Recently, a mechanistic explanation for that observation was found. It was determined that interactions of specific codons pairs with their tRNAs, mostly involving wobble-base pairing, induce certain conformational changes in the ribosomes that lead to stalling[164].

The use of sub-optimal pairs of codons has also been proposed as a strategy to create live-attenuated viruses for vaccine development. However, there has been a lively debate about whether the decreased expression of those viruses in eukaryotic host cells should be attributed to suboptimal codon pairs or, alternatively, to sub-optimal dinucleotide pairs[165]. A recent study that aimed to disentangle the effects of dinucleotide bias and codon-pair bias in virus attenuation concluded that sub-optimal codon pairs primarily caused the decreased translational efficiency[166]. That study shows that the influence of sub-optimal codon pairs can, at least partly, be related to decreased mRNA stability, in line with the previously discussed correlation between codon usage, translation efficiency, and mRNA stability in eukaryotes.

In bacteria, the presence of SD-like sequences within ORFs was previously suggested to result in a slowdown of the translation-elongation process[167]. However, that observation was later toned down in a re-evaluation of ribosome-profiling data, which concluded that SD-like sequences have little or no effect on translational pausing[168]. Recently, a bioinformatical analysis studying the evolutionary conservation of those SD-like sequences in ORFs of several bacterial species, concluded that they are less conserved than would be expected by random chance[169]. This suggests a negative evolutionary selection against SD-like sequences, hinting at a potential decrease in fitness caused by the presence of those sequences within ORFs, possibly because they could induce mistranslation or erroneous frameshifting. In conclusion, it seems that SD-like sequences are not frequently used in nature because of detrimental by-effects on translation and that they do not have a major role in controlling translation elongation rates.

Another recent study has revealed an interesting effect of certain short amino acid motifs on translation elongation. That study focused on mutating codons at positions 3, 4, and 5 of a GFP reporter in *E. coli* and allowed non-synonymous mutations[170]. They identified specific amino acid motifs at the start of the ORF that lead to high translation efficiency, independent of specific codons or mRNA structures. At the same time, they identified detrimental amino acid motifs in the 5' region of the ORF, which can cause pausing of the translation and lead to increased translational abortion. This observation was explained by specific interactions of the nascent peptide motif with the ribosome exit tunnel that could lead to ribosomal stalling and drop-off.

There are more reports of specific peptide motifs that cause stalling or translational slowdown, likely via interactions in the ribosome exit tunnel. Motifs such as poly-proline sequences can slow down or stall translation in organisms throughout all domains of life[171,172]. In addition, it was observed in *E. coli* that four specific amino acid triplets completely stalled translation and were avoided within its proteome[173]. It is good to realize that both in evolution and in synthetic biology approaches, the flexibility to evolve or design acceptable changes in amino acid sequences, without altering residues that are critical for protein functionality, may sometimes result in improved translation efficiency.

Furthermore, translational speed can be influenced by the modifications of mRNA and tRNAs. It is well established that the great diversity of tRNA modifications, especially modifications of ribonucleotides in anticodon regions, can have a major effect on translation rates and fidelity[174–176]. Recently, it was also observed that modifications of mRNA, e.g., N6-methyl-adenosine and N4-acetylcytidine, influence translation elongation and mRNA decay in both eukaryotes and bacteria[177–179].

2.5.6 Translational fidelity versus translation rate and translation termination

Apart from governing translational speed, ORF sequence features such as codon usage have been postulated to govern translational fidelity. Even though support for this theory has been provided by bioinformatic analyses[180], only very recently has experimental evidence for this hypothesis been obtained. Using a “deep proteomics” approach, translational errors have been identified in the proteomes of *E. coli* and *Saccharomyces cerevisiae*[181]. That study revealed that translation errors are relatively abundant, occurring on average once every 1,000 amino acids. Transcriptional error rates occur much less frequently, at about 1 in 25,000 nucleotides[182].

Both the misloaded tRNAs and tRNA-codon mispairing can cause translation errors, but the latter error is more abundant. In that case, wrong amino acids are delivered by near-cognate tRNAs, which have only one mismatch between codon and anti-codon[181]. Interestingly, the effect of mistranslation events is probably reduced because the genetic code has evolved such that these near-cognate tRNAs often deliver amino acids with similar chemical properties. Some codons are more sensitive to mistranslation than others, and that pattern was relatively similar both in yeast and in *E. coli*, suggesting that evolutionarily conserved mechanisms or universal chemical interactions lead to occasional mistranslation.

The same study also demonstrated a negative correlation between translation speed and translation fidelity, suggesting a trade-off between optimizing coding sequences for translational speed and fidelity. This fidelity theory (slowdowns to reduce translational errors) is an interesting alternative explanation for the aforementioned occurrence of "slow" codons in structurally important regions, which, in many reports, is explained by the co-translational folding theory[37,86].

Frameshifting during translation has an even bigger effect on protein function than amino acid misincorporation because the downstream sequence is completely mistranslated. However, the operation of ribosomes and their translation elongation factors seems to limit frameshifting. Recently, another mechanism for frameshift fidelity was observed in human cells[183]. It was suggested that periodic pairing of certain "sticky codons" on the mRNA with complementary triplets in the rRNA, near the exit of the ribosomal mRNA channel, helps to prevent frameshifting. That conclusion was supported by the substitution of sticky codons by synonymous counterparts, which led to a 4-fold increase in frameshifting, as well as mutating the complementary triplet at the exit of the ribosomal mRNA channel, which also influenced the frameshifting rate. Finally, it seems that these sticky codons are naturally underrepresented in a non-coding frame in eukaryotic genomes, which may be to prevent accidental frameshifting[183]. This mechanism deserves further analysis throughout different types of organisms and may cause certain codon preferences to limit frameshifting.

At the end of the translation-elongation process, the ribosome encounters a stop codon, and upon binding of a release factor (a protein mimic of a tRNA), the translation is ended, and the ribosome is released from the mRNA. However, in rare cases, translation read through happens, generally leading to the synthesis of non-functional proteins. If such a read-through event takes place, the ribosomes either encounter an in-frame stop codon within the 3'UTR or they get stalled at the end of the mRNA[172]. These read-through proteins are generally degraded co- or post-translationally[184]. Some organisms may prevent translational read through by using tandem stop codons, which are, for example, observed more frequently in the 3'UTR of ciliates[185].

2.6 The interaction between different factors

2.6.1 Cooperative and counteracting features

As discussed, distinct factors are involved in different steps of the gene-expression process, and they interact with each other in multiple ways. Some factors in the protein-production process act in a cooperative fashion. As a remarkable example of that, the translation-elongation efficiency and mRNA stability in eukaryotes have been demonstrated to be mechanistically linked, leading to positive feedback between translation elongation and mRNA stability[90,119]. However, other sequence features may also influence each other negatively. For example, a high-affinity SD sequence and well-translated codons in the 5' region of the ORF could form a base pair and, consequently, form undesired mRNA secondary structures that hamper efficient translation initiation. These counteracting and cooperative features complicate the evaluation of individual factors.

Several studies have attempted to reveal new factors and to disentangle their connections in recent years. Many of those studies applied randomized or systematically designed reporter gene-variant libraries of GFP in *E. coli*[89,128,129,160]. The consensus of those studies is that gene expression is significantly affected by strong (predicted) mRNA secondary structures in the 5'UTR and the 5' region of the ORF. However, a large part of the variation in expression levels in those studies is explained by a range of other factors, and a substantial part of the observed fluctuations cannot be explained at all. Furthermore, it is not certain that those studies properly reflect features that are relevant to native genes. Nevertheless, a number of recent studies on native gene expression in *E. coli* also suggest that mRNA structures and associated RBS availability are key factors that determine the expression rate of natural genes[134,159].

A combination of different experimental approaches to study native gene expression was recently performed in yeast, integrating multiple omics data and measurements of mRNA and protein half-life times[103]. The latter is an often overlooked factor because proteins with shorter half-lives need to be translated at higher levels to sustain sufficient protein levels. That study found large differences in protein yield per mRNA, varying up to 400-fold among some proteins, suggesting an important role in the efficiency of the translation processes. However, when accounting for all proteins, translation-elongation efficiency only explained 15% of the protein abundance observed, whereas mRNA abundance was the most important explanatory factor for protein levels (explaining 61%). A large study on a diverse set of heterologous proteins in *E. coli* also reported mRNA abundance as the main predictor for protein abundance[102]. However, it is important to realize that mRNA abundance can also be influenced by translation efficiency.

2.6.2 Influence of gene designs on resource consumption and growth

An important, overarching aspect for protein production is the high metabolic costs associated with transcription and translation processes. Those additional costs include “materials”, such as demands for ATP, nucleotides, and amino acids, but also the extra demand for the transcriptional and translation factors, such as RNAPs and ribosomes. There is an evolutionary pressure on the genome in general, and the architecture of genes and their regulation in particular, to reduce metabolic costs to optimize cellular fitness. Within synthetic-biology applications, the reduction of energy and resource requirements is of importance for gene design.

Hence, recent efforts studied growth parameters of microbial cells harbouring codon-variant libraries of reporter genes (e.g., GFP) or of a growth-essential gene. The relative fitness of different variants was recorded, either by measuring growth curves for individual strains or by performing competition experiments between them[89,159,160]. One of the main conclusions is that, especially for highly expressed genes, a high level of protein produced per mRNA is a resource-efficient way for high expression. So, even though, in nature, high mRNA levels are typically correlated to high expression, boosting expression solely by high mRNA levels is not the best strategy. Extremely abundant mRNAs potentially imply excessively high transcription costs or may sequester excessive amounts of ribosomes from the limited pool. In contrast, we note that the strategy to keep mRNA levels low and, rather, to couple it to highly efficient translation can increase the cell-to-cell variability in mRNA and protein concentrations[186]. Thus, to achieve both high resource efficiency and low cell-to-cell expression variability, nature and synthetic biologists need to properly tune the translation efficiency per mRNA.

2.7 Biotechnological challenges and opportunities for gene design

Innovations in DNA synthesis and genetic engineering have tremendously accelerated the capacity to express synthetic genes. However, based on data from consortia aiming to resolve large numbers of protein structures, it is estimated that only about one-half of the attempts for heterologous protein production led to successful expression[35]. In practice, in molecular biology and synthetic biology projects, the expression of synthetic genes regularly leads to sub-optimal production or problematic growth because of the excessive expression burdens.

2.7.1 Limitations of codon optimization algorithms

Synthetic genes for heterologous protein production are typically designed with codon-optimization algorithms, which generally optimize a particular ORF, adapting it to a codon-usage index of the expression host[35]. Those codon indices are frequently determined with either the codon usage within a set of highly expressed reference genes (e.g., CAI) or the tRNA copy numbers (e.g., tAI) in the host cell. Some academic and commercial algorithms also take alternative parameters into account, such as GC content and avoidance of certain regulatory motifs, such as SD sequences or repeats[187]. Only a few algorithms additionally aim to minimize mRNA secondary structures[187], even though the folding in the translation-initiation region, certainly in prokaryotes, is a key determinant of expression. A promising exception is the novel 31C-FO algorithm, which aims to minimize mRNA folding of the 5'UTR and the first 48 bases of the ORF[102]. At the same time, that algorithm optimizes codon usage by only including 31 codons that are correlated to high expression in *E. coli*. That algorithm was reported to lead to successful expression of several proteins by Boël *et al.*[102] but has not been reported in other studies yet, and no easy tool for that algorithm is available so far.

Generally, the features involved in gene expression, individually or in concert with others, are still not understood in sufficient detail to compose robust optimization algorithms for relevant host organisms. Multi-parameter algorithms, such as EuGene or DNA-Tailor (D-Tailor)[187], typically leave the setting of specific objectives up to the users, which, in practice, is hard to decide upon, given the unknown weight of the different factors. Furthermore, it has been shown that a so-called design-of-experiments approach, which systematically varies multiple factors, is no guarantee for successful expression because not all relevant factors are known yet or are not known in sufficient detail[89].

In addition to expression levels, proper protein folding is important for the functional production of proteins. The accumulating evidence on the role of codon usage in protein folding led to several approaches that aimed to include the translation-speed landscape to accommodate the folding of structural elements. For example, codon-harmonization algorithms have been proposed to tackle this issue[37,188]. These algorithms have as their objective to copy the *native*-codon-usage landscape of a gene-of-interest (distribution of rare and frequent codons in the organism from which the gene originated natively) into a *heterologous*-codon-usage landscape (similar distribution of rare and frequent codons in the context of the expression host). However, codon harmonization does not always give the best expression levels in *E. coli* when comparing the production levels of codon-harmonized gene variants with native genes or CAI-codon-optimized genes for some membrane proteins[38].

In some studies, sub-optimal codons or SD-like sequences have been included in ORFs to slow down translation in between structural domains, which was reported to improve protein solubility in a few cases[189,190]. That, however, requires laborious, detailed studies to determine exactly the position and strength of the required translation pauses to optimize the folding of a specific protein. Furthermore, there is no full understanding yet on the role of the coding-sequence features for translational speed; this all restrains robust design approaches for proper folding of proteins.

In summary, improving heterologous protein production by codon-optimization algorithms often remains a trial-and-error approach. Success rates can be increased by testing multiple different codon-optimized variants, but that also increases experimental labour and costs.

2.7.2 UTR optimization strategies

In numerous studies, the 5'UTR has been identified as a critical region that determines translation-initiation efficiency in protein production. As discussed, few of the available codon-optimization algorithms take the 5'UTR into account and do not have integrated functionality to avoid detrimental mRNA structures in the translation-initiation region. Nonetheless, some specific tools have been developed to design optimized 5'UTR regions for bacterial protein production, which generally try to design 5'UTRs to have strong and accessible RBSs, taking into account the downstream ORF region. Hereto, these tools have used *in silico* mRNA folding energy calculations[191–193]. Despite their wide use and relatively successful predictions, they still suffer from the limited reliability of *in silico* RNA structural predictions. Recently emerging experimental tools for measuring *in vivo* RNA folding may become helpful to assess the validity of computational predictions[132,133].

Alternatively, standardized 5'UTR modules have been employed for robust gene expression, for example, by using combinations of well-expressed 5'UTRs and N-terminal tags[194]. In addition, bicistronic RBS modules have proven highly useful because these modules partly uncouple translation-initiation efficiencies from the ORF sequence[89,195]. These bicistronic design elements (BCDs) have been shown to allow for tuned and improved expression levels in *E. coli* and *Corynebacterium glutamicum*[36,43,196].

The initiation mechanisms in the 5'UTR in eukaryotes seem more diverse and complicated than do those for prokaryotes. However, recent studies have shown that the 5'UTR sequence has great potential for tuning the expression in eukaryotes, such as *S. cerevisiae* or Chinese hamster ovary-S (CHO-S) cells[197–199]. These studies provided modular 5'UTRs designs that work relatively well with low-context dependence on the downstream ORF. One of the key factors that improve the performance of those 5'UTR is the reduction of mRNA secondary structures in that region.

The 3'UTR is less studied in relation to expression efficiency, but it has also been reported to influence mRNA stability and transcription termination efficiency, thereby modulating expression efficiency. Examples of 3'UTR engineering in bacteria are scarce, so far. For yeast and human cell lines, some short synthetic 3'UTR modules have been developed that relatively robustly increase expression for multiple genes throughout multiple species but also seem partly dependent on the upstream ORF sequence[200,201].

An important part of the influence of 5'UTRs and 3'UTRs on protein production is explained by their roles in mRNA stability. An alternative, promising approach to improve mRNA stability for protein production, is through the circularization of mRNAs, which also occurs in nature. Synthetic circular mRNAs can, for example, be generated by harnessing the mechanism of self-splicing introns[202,203]. A recent surge of research in this field showed promising applications for protein production driven by synthetic, circular mRNA transcripts in eukaryotes. Because canonical-eukaryotic translation initiation relies on the 5' cap, alternative translation-initiation mechanisms, such as IRES or N^6-methyladenosine modifications, are required to ensure sufficient translation initiation in circular mRNAs. Furthermore, it has been proposed that the translation of circular mRNAs can be increased by creating an infinite ORF, by removing the stop codon of the ORF[202]; the same ribosomes will repeatedly translate the same sequence, leading to a multimeric protein, and individual functional proteins can be produced by introducing protease cleavage or self-cleavage sites in the polypeptide. A recent review elaborates in great detail on the developments of engineering of circular RNA[204].

2.7.3 Randomization, smart selection, and machine learning

The number of studies that randomly vary sequences in promoters, 5'UTRs, and the start of the ORF have steadily increased, mostly for GFP. This randomization approach may also be relevant for optimizing or fine-tuning the production of more biotechnologically relevant proteins (Figure 4). However, unlike expression levels of reporter proteins, levels of most proteins of interest are generally hard to screen with sufficient throughput from large randomized libraries. Still, some well-expressing modules identified in reporter-based screens, e.g., promoters or 5'UTR-N-terminal tag peptide combinations, have been used successfully for the optimized production of other proteins.

When randomly optimizing the coding sequence, or at the junction of the 5'UTR and coding sequence, novel approaches are required to screen for well-expressed gene variants in the case of non-reporter proteins. One simple approach is to fuse the protein of interest to a reporter protein, but such a fusion frequently distorts the function of the protein of interest. An alternative method, not based on a protein fusion, was recently established by translational coupling of the protein of interest to a selectable antibiotic-resistance reporter. This so-called TARSyn system was demonstrated for the high-throughput selection of optimized 5'UTR:ORF junctions for the expression of antibody proteins in *E. coli*[205] (Figure 4). We consider the development of selection and screening systems of well-expressed “randomized” sequences to be a very promising avenue for further exploration.

Alternatively, data collected from large-scale randomization studies on reporter proteins or growth-selectable markers may help to generate better predictive algorithms (Figure 4). These large-scale data could serve as training sets for machine learning. Different types of machine learning can be employed to generate more reliable algorithms to improve the design of synthetic genes[206].

A recent, innovative study that used machine learning focused on predicting the influence of different 5'UTR sequences in *E. coli*[207]. The study developed an innovative reporter system, based on a recombinase protein, to quantify the expression from a large library of randomized 5'UTR sequences (Figure 4). At a certain expression level, that site-specific recombinase flips a DNA sequence, which is located directly next to the 5'UTR on the same plasmid. Subsequent, high-throughput sequencing of short DNA fragments that contain both the 5'UTR and the potentially flipped DNA sequence gives information on both 5'UTR genotype and related expression phenotype, which provided data on the recombinase expression from 300,000 different 5'UTRs, which were fed into machine learning. The analysis, surprisingly, revealed that, rather than mRNA secondary structures, the presence and positioning of the SD are most important for high protein production in this case, possibly because the 5' end of the recombinase ORF was unlikely to form strong mRNA structures with any UTR. The machine-learning approach was used to develop a new 5'UTR design algorithm that Höllerer *et al.*[207] report outperformed currently available algorithms, which are mostly based on biophysical models. However, this algorithm has not yet been tested for ORFs other than the recombinase ORF in that study.

Likewise, successful 5'UTR prediction algorithms based on multiple regression or machine learning approaches have been developed for yeast[197,208,209]. Such big-data analyses, based on randomized sequence libraries, seem a promising road toward better predictive algorithms for robust regulation of synthetic genes.

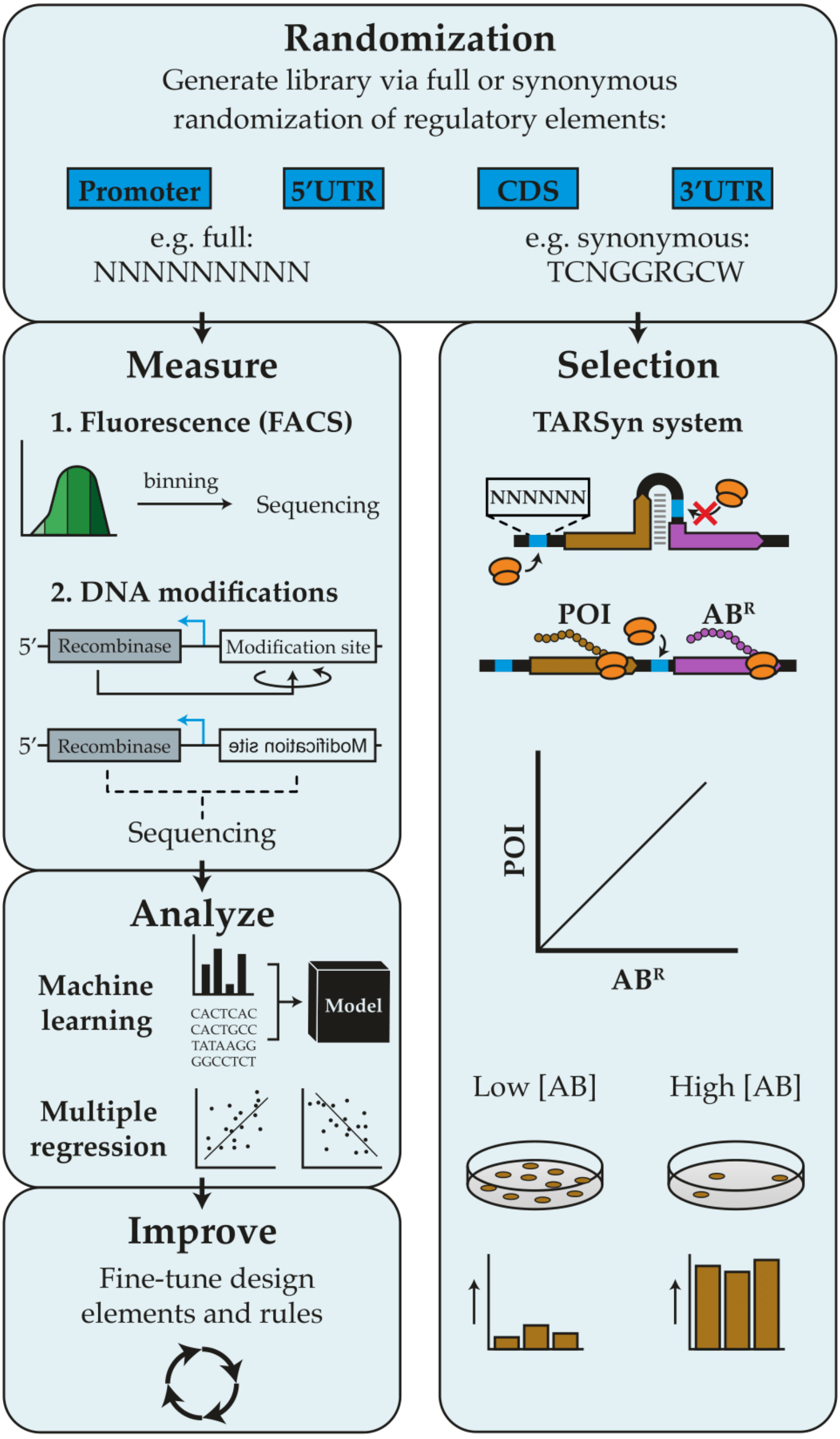

Randomization
Generate library via full or synonymous randomization of regulatory elements:
Promoter
5′UTR
CDS
3′UTR
e.g. full:
NNNNNNNNN
e.g. synonymous:
TCNGGRGCW
Measure
1. Fluorescence (FACS)
binning
Sequencing
2. DNA modifications
5′
Recombinase
Modification site
5′
Recombinase
Modification site
Sequencing
Selection
TARSyn system
NNNNNN
POI
AB^R
POI
AB^R
Low [AB]
High [AB]
Analyze
Machine learning
CACTCAC
CACTGCC
TATAAGG
GGCCTCT
Model
Multiple regression
Improve
Fine-tune design elements and rules

Figure 4: Overview of a typical workflow randomizing gene regulatory and ORF sequences. *After randomization of genetic regulatory sequences or (part of) the codons of an ORF, the protein production by the resulting (large) variant library can be measured and binned according to fluorescence levels by fluorescent activated cell sorting (FACS). As an alternative to fluorescent-reporter proteins, a DNA-modifying enzyme can be used as a reporter because its expression can be assessed by high-throughput sequencing of modifications in the DNA. The latter approach was demonstrated for the expression of a randomized 5'UTR library mediating the expression of a recombinase that flips a nearby DNA modification site. In the same single-sequencing read, the 5'UTR variant can be identified and whether the site was flipped or not, allowing high-quality, large-scale data on expression levels[207]. Analysis of generated large-scale data is typically performed by multiple regression analysis and, recently, by machine-learning algorithms. Next, understanding of expression levels can be further improved by correlations or rules derived from the analysis, and expression could be further studied during next-iteration rounds in which randomized sequence space can be limited, based on the results of the previous iterations. As an alternative to the learning cycle, a direct-selection system can be used for the selection of high-expressing variants. For example, the so-called TARSyn system allows for the selection of high-expressing clones based on antibiotic resistance[205]. The expression of a (non-reporter) protein of interest is translationally coupled to downstream antibiotic resistance, allowing for easy selection for high expression under high antibiotic concentrations.*

2.8 Conclusions

Despite significant efforts to elucidate the effect of codon usage and other gene features on protein production, it is still not completely understood. During the past decade, genome, transcriptome, proteome, and translatome (ribosome profiling) data became increasingly available. Bioinformatic analysis of those data has provided relevant insights into coding features and their relation to protein production. Recently, such analyses, combined with half-life measurements of mRNA, led to the discovery that optimal translation of an mRNA increases its stability in eukaryotes. However, many factors and their relevance are still unclear and require further investigation and, possibly, new experimental approaches.

One of the key knowledge gaps is the role of mRNA secondary structures, which is suggested to have a pivotal role in translation initiation and elongation, but its true effect is still unsettled. Recently, emerging protocols enabled the generation of transcriptome-wide *in vivo* mRNA structural data. However, groups using such methods report partly contradicting results for the role of mRNA secondary structures on translation-elongation efficiency[134,151]. Further refinement and validation of those protocols are required to improve the understanding of mRNA structures on translation. Another poorly explored territory is the influence of the ORF's codon sequence on co-translational folding and fidelity. Bioinformatic analysis of genome and translatome data suggested important roles for translation speed on protein folding, at least for some proteins. Detailed molecular studies focusing on some specific proteins have confirmed that codon usage has a crucial role in folding. However, data and protocols to test this hypothesis experimentally for larger sets or proteins or on a proteome-wide scale are lacking.

A general limitation of studying genetic features within native genes (in a certain organism or under certain conditions) is the complexity in detecting "weak signals" from relevant factors within sequences that underwent optimization during millions of years of evolution. Alternative approaches, based on synthetic gene libraries, represent strong complementary methods in which many variants for a single gene can be generated to probe relevant factors. However, these "controlled" studies have, so far, been able to provide generic explanations for variable protein production levels only to some extent and are mostly based on correlating expression with known factors. In addition, those studies have mostly focused on a few highly expressed reporter proteins (mostly GFP), which may make conclusions biased.

Machine-learning approaches may help to further elucidate unknown features and factors in a more unbiased way. Such approaches have recently been applied to analyze expression data from randomized synthetic libraries of promoters and 5'UTRs. Such approaches may be promising for developing better predictive algorithms. However, large datasets are required for machine-learning algorithms to generate predictive models, and machine learning does not necessarily lead to increased biological understanding because, sometimes, such machine-learning approaches generate a predictive “black box.”

The limited understanding of the fundamental rules in protein production remains a significant challenge for its applications. Problems in synthetic gene design are regularly observed for tuning and optimizing production of biotechnological or medical relevance. These challenges become even more pressing for synthetic biologists trying to construct designer genomes, which require tuning of many synthetic genes simultaneously.

Specific methods have been proposed that can, to some extent, increase the predictability of synthetic gene design. Typically, commercial or academic codon optimization algorithms are used to design ORF regions for heterologous expression, often with limited success, which is not surprising given the current knowledge gaps. However, promising design and randomization approaches have been established regarding the engineering of the highly influential region comprising the 5'UTRs and the first few codons of an ORF.

Overall, both for the understanding of the fundamental natural principles of gene design and expression and for diverse applications, there remains a need to delve further into the outstanding questions in this field. Despite the impressive recent progress, further refinement of recently launched techniques, as well as the development of new experimental and computational approaches, will be essential to address key questions that have intrigued many biologists for decades.

2.9 Acknowledgements

T.N., M.F.-B., N.J.C., and J.v.d.O. acknowledge the support of the Netherlands Organisation for Scientific Research (NWO) via the Gravitation Project BaSyC (024.003.019) and Spinoza (SPI 93-537), awarded to J.v.d.O. In addition, N.J.C. acknowledges support from his NWO Veni fellowship (VI.Veni.192.156).

Chapter 3. BiCistronic Design-based continuous and high-level membrane protein production in *Escherichia coli*

Max Finger-Bou[1,*], Nico J. Claassens[1,*], Bart Scholten[1], Frederieke Muis[1], Jonas J. de Groot[1], Jan-Willem de Gier[2], Willem M. de Vos[1,3], and John van der Oost[1]

[1]Laboratory of Microbiology, Wageningen University and Research, Wageningen, the Netherlands

[2]Department of Biochemistry and Biophysics, Center for Biomembrane Research, Stockholm University, SE-106 91, Stockholm, Sweden

[3]Human Microbiome Research Program, Faculty of Medicine, University of Helsinki, Haartmaninkatu 3, FI-00014, Helsinki, Finland

Adapted from publication:

Finger-Bou, M.*, Claassens, N. J.*, Scholten, B., Muis, F., De Groot, J. J., de Gier, J. W., de Vos, W. M. & Van Der Oost, J. (2019). Bicistronic design-based continuous and high-level membrane protein production in Escherichia coli. *ACS synthetic biology*, *8*(7), 1685-1690.

*: contributed equally

3.1 Abstract

Escherichia coli has been widely used as a platform microorganism for both membrane protein production and cell factory engineering. The current methods to produce membrane proteins in this organism require the induction of target gene expression and often result in unstable, low yields. Here, we present a method combining a constitutive promoter with a library of BiCistronic Design (BCD) elements, which enables inducer-free, tuned translation initiation for optimal protein production. Our system mediates stable, constitutive production of bacterial membrane proteins at yields that outperform these obtained with *E. coli* Lemo21(DE3), the current gold standard for bacterial membrane protein production. We envisage that the continuous, fine-tuneable and high-level production of membrane proteins by our method will greatly facilitate their study and their utilization in engineering cell factories.

3.2 Introduction

High-level heterologous production of membrane proteins in *E. coli* and other hosts has proven challenging, especially due to oversaturation of the membrane protein biogenesis machinery[210]. Common systems for recombinant protein production, such as those based on the strong T7 promoter, often lead to jamming of chaperones and membrane translocation systems, consequently making it impossible to produce correctly-folded membrane proteins at high levels[210,211].

Several *E. coli* strains and expression systems have been developed to improve the production of especially bacterial membrane proteins. Commonly used systems include the *E. coli* Walker strains (C41(DE3),C43(DE3))[212], *E. coli* BL21-AI[213], and the more recently developed *E. coli* Lemo21(DE3)[214–216]. These systems rely on downregulating the levels of T7 RNA polymerase (T7RNAP), consequently reducing expression rates to better accommodate translocation and folding of membrane proteins[217]. Particularly, *E. coli* Lemo21(DE3) has been constructed for the fine-tuning of transcription through an indirect control of T7RNAP activity through L-rhamnose-inducible production of its inhibitor, T7-lysozyme (LysY)[214–216]. The system has proven successful and has been recently streamlined into a one-plasmid system named pReX[218], but it requires the properly timed addition of two different inducer compounds: L-rhamnose and the expensive IPTG (isopropyl β-D-1-thiogalactopyranoside). Additionally, to date, neither Lemo21(DE3) nor any other currently available system has been demonstrated successfully for long-term (>24 hours) continuous production of membrane proteins in *E. coli*. Realizing inducer-free, stable continuous production remains a major challenge relevant to many synthetic biology applications. This includes, for example, the heterologous production of transporter membrane proteins in microbial cell factories to be used in (continuous) production processes, or signal-transduction membrane proteins in strains that need to function as a biosensor over a long time without induction.

Another limitation in heterologous protein production relates to Ribosome binding site (RBS) accessibility[219–221]. Several proteins, including membrane proteins, have been categorized as ‘difficult-to-produce’ due to strong mRNA secondary structures in the 5’ untranslated region (5’-UTR) and in the start of the coding sequence (CDS), which impede the proper translation initiation at the RBS[222,223]. Some efforts aimed at resolving such structures rely on fusing well-expressed short peptide tags to membrane proteins[222,224,225]. However, N-terminal fusion peptides can affect protein stability, structure and function[222]. Furthermore, translation initiation has been successfully improved by randomly mutating nucleotides around the start codon[222,223,226], which enhanced the production of several ‘difficult-to-produce’ membrane proteins, but this method requires a high-throughput screening or selection approach.

3.3 Results and discussion

To overcome the limitations of state-of-the-art membrane protein production, we explored an alternative method for protein expression: the so-called BiCistronic Design elements (BCDs)[227]. The system is based on a constitutive promoter and tuning by using two RBSs that are translationally coupled. The first RBS mediates strong translation initiation of a short leader peptide, while the second RBS, which is located within the leader peptide’s CDS and drives the translation of the protein of interest, has a tuneable strength (Figure 1). It is hypothesized that the intrinsic helicase activity of the ribosomes translating the leader peptide can unwind potential secondary structures of the mRNA, thereby eliminating translation initiation problems near the second RBS and the start of the target CDS[228].

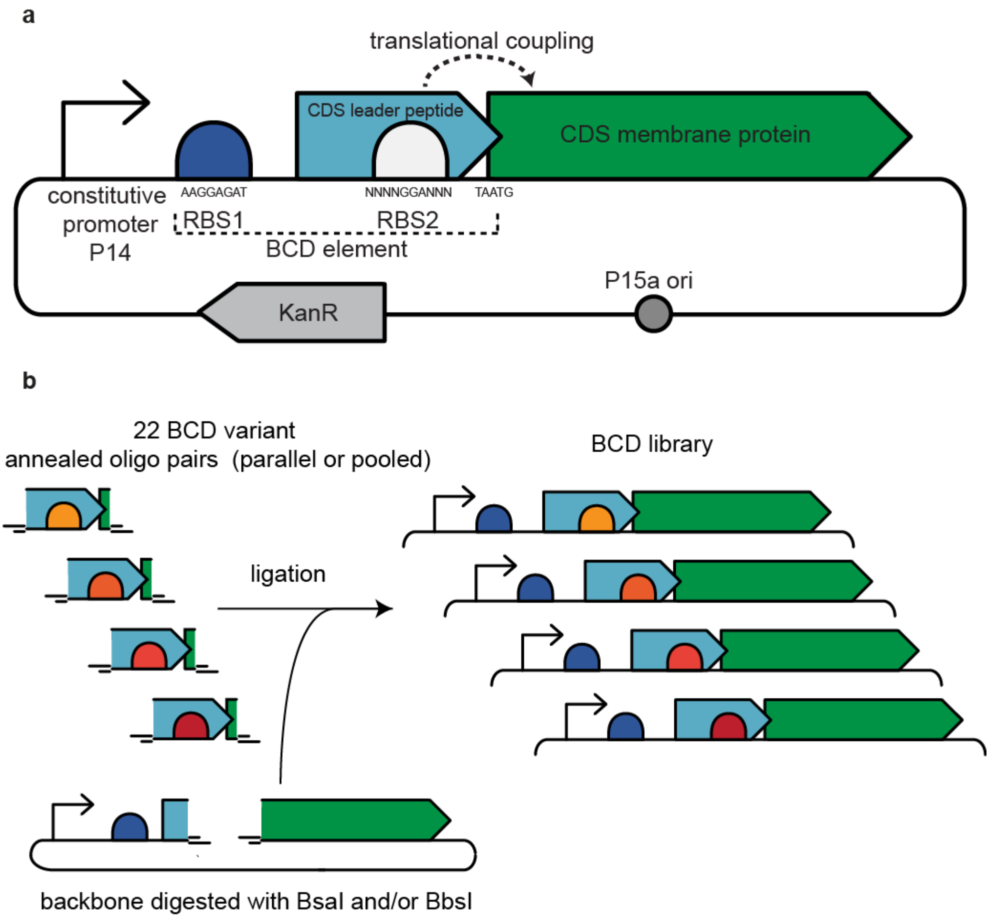


Figure 1: Expression vector design and assembly. *(a) The standard expression vector contains a medium-strength constitutive promoter, RBS1, which allows for strong translation initiation of a leader peptide, and a translationally coupled, variable RBS2, mediating translation initiation of the Coding Sequence (CDS) of the membrane protein of interest[227]. (b) Vectors are assembled with different BCD elements. First, the vector is amplified by PCR, subsequently it is digested by type IIS restriction enzymes. The latter allows for seamless assembly with a library of annealed oligo pairs (encoding the different BCD variants), which have overhangs complementary to the digested vector.*

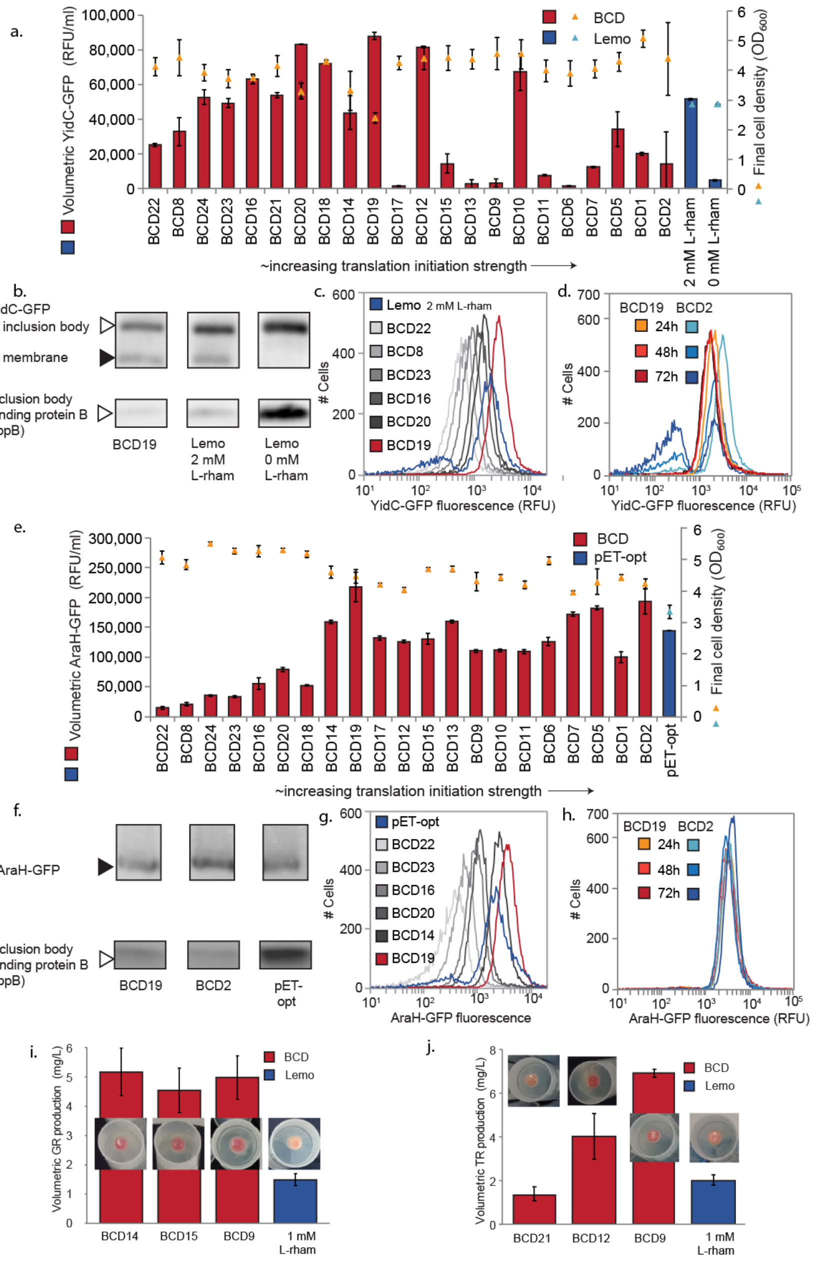


Figure 2: Production of YidC-GFP, AraH-GFP and rhodopsins by BCD elements and comparison to state-of-the art systems. *(a) Volumetric YidC-GFP production based on*

whole-cell fluorescence measurements and final growth yields for the BCD constructs, and a comparison to Lemo21(DE3)-based production at optimized (2 mM L-rhamnose) and non-optimized (0 mM L-rhamnose) conditions. BCD variants are ordered in the X-axis based on previously reported translation initiation strength[227]. (b) Western blots performed with anti-his-tag antibody (upper panels) to visualize both inclusion body and well-folded YidC-GFP-his, and anti-IbpB[229,230] (lower panels) to visualize inclusion body binding protein B. (c) Single-cell production of YidC-GFP analysed by flow cytometry for several increasing strength BCD elements and optimized Lemo21(DE3). (d) Single-cell production of YidC-GFP by BCD19 and BCD2 in a 72h stability experiment. (e) Volumetric AraH-GFP production based on whole-cell fluorescence measurements and final growth yields for the BCD constructs, and a comparison to pET-opt-AraH-GFP[223]. BCD variants are ordered in the X-axis based on previously reported variant strength[227]. (f) Western blots performed with anti-his-tag antibody (upper panels) to visualize both AraH-GFP-his, and anti-IbpB[229,230] (lower panels) to visualize inclusion body binding protein (g) Single-cell production of AraH-GFP analysed by flow cytometry for several increasing strength BCD elements and pET-opt-AraH-GFP (h) Single-cell production of AraH-GFP by BCD19 and BCD2 in a 72h stability experiment. (i) Volumetric Gloeobacter Rhodopsin (GR) and (j) Thermophilic Rhodopsin (TR) production determined by spectroscopy, and pictures of red-pigmented pellets. All cultivations were performed in 10 mL medium in 50 mL tubes, for YidC-AraH and AraH-GFP at 30ºC, for rhodopsins GR and TR at 37ºC, and for pET-opt-AraH-GFP in E. coli BL21(DE3) pLysS at 25ºC (as optimized for in original work). BCD-based production was measured after 22 hours of cultivation, while Lemo21(DE3) and pET based production was measured after 22 hours of induction. Whole-cell fluorescence or rhodopsin quantification data are based on at least three biological replicates. For 72h stability experiments E. coli BL21(DE3) harboring BCD vectors were re-inoculated 1:50 into fresh LB kanamycin medium every 24 hours. RFU: Relative Fluorescence Units; OD_{600}: Optimal Density 600 nm.

In this work, we employ a library of translational coupling elements to tune and optimize the constitutive production of several bacterial membrane proteins. A medium-strength, constitutive promoter (P14) and 22 variable-strength translational coupling elements (BCDs) were selected from the work of Mutalik *et al.*; the BCD elements can be inserted seamlessly in the expression vector using a simple Golden Gate-based cloning method[231] (Figure 1).

We tested our system for the expression of four different membrane proteins: YidC, AraH and two rhodopsins. YidC is a membrane-translocation chaperone in *E. coli* that has been frequently used as a model for studying membrane protein production[229,230,232]. AraH is the integral membrane component of the *E. coli* arabinose ABC transporter, which is considered a ‘difficult-to-produce’ protein because of its translation initiation limitations in a typical pET vector[222,223]; GR is a proton-pumping rhodopsin photosystem from the cyanobacterium *Gloeobacter violaceus*, and TR is a thermophilic rhodopsin from *Thermus thermophilus*.

For the two first proteins, YidC and AraH, GFP was fused to their C-terminus, rendering YidC-GFP and AraH-GFP. All the 22 BCD variants were cloned separately into expression vectors carrying YidC-GFP and AraH-GFP. We then estimated levels of membrane inserted YidC and AraH by measuring fluorescence levels. C-terminal GFP only folds properly and results in fluorescence when membrane proteins fused to the GFP are integrated into the membrane and not end up in inclusion bodies, hence providing a quantitative approximation for membrane-embedded expression[233,234]. Expression of YidC-GFP and AraH-GFP by different BCD elements was ranked in order of the translation initiation strengths previously observed for each BCD element in the work by Mutalik *et al.*, in order to allow for a systematic analysis of their fluorescence (Figure 2a,2e).

For both fusion proteins, the tested BCD variants resulted in a range of GFP fluorescence-signals, suggesting different levels of functional membrane protein production. In the case of YidC-GFP, a rough pattern was observed considering the correlation between the fluorescence and the expected translation initiation strength of the different BCD constructs. BCD elements up to BCD19 generally resulted in increased levels of production, whereas elements stronger than BCD19 mostly resulted in lower production levels (Figure 2a). Some of the strongest translation initiation variants resulted in negligible and/or highly irreproducible production levels. For example, in some replicate cultures the strong BCD2 gave high expression but in several other cultures expression was completely absent (ranging from no expression to 70,000 RFU/mL). When comparing BCD-based expression to optimized Lemo21(DE3)-based expression of YidC-GFP, the latter gave rise to an emerging non-producing subpopulation of cells after 22h, as indicated by flow cytometry analysis, whereas the production by the highest-producing BCD19-YidC-GFP remained homogeneous (Figure 2c). In agreement with this observation, Western blot analysis revealed that the formation of YidC-GFP in inclusion bodies was reduced for the BCD19 versus optimized Lemo21(DE3)-expression (Figure 2b). Moreover, the medium-strength BCD19 yielded approximately twice as much production per cell than *E. coli* Lemo21(DE3), which was previously proven to be a superior production system for YidC-GFP over other commonly used systems such as *E. coli* C41(DE3) and C43(DE3)[215].

In the case of AraH-GFP, a similar trend was observed, *i.e*. increasing levels of fluorescence were measured up to BCD19. However, unlike for YidC-GFP, no large decrease or unstable AraH-GFP production was observed for stronger BCD elements, and the strongest BCD2 produced at similar high levels as BCD19 (Figure 2e). AraH-GFP production by BCD19 and BCD2 was compared to the production by a previously optimized pET vector (pET-opt-AraH-GFP), which was obtained by screening a large library of pET vectors with mutations around the start codon[223]. The volumetric production by both BCDs was found to be higher than that of pET-opt-AraH-GFP at 30ºC, mainly due to a higher biomass yield (Figure 2e, see Figure S2 for production at 25ºC). Production by the BCDs also resulted in less inclusion body binding protein than pET-based production, while non-folded protein could not be well detected in Western blot for any construct (Figure 2f). Additionally, production by pET-opt-AraH-GFP resulted in a small emerging non-producing population of cells after 22 hours of cultivation, as revealed by flow cytometer analysis, while production by all BCD elements was still fully homogeneous (Figure 2g). Notably, by screening a library of only 22

BCD variants, we were able to find clones whose production was comparable in production (per cell) or even better (per volume) than that of the previously optimized pET-opt-AraH-GFP, which required the high-throughput screening of a large library of variants ($1.6 \cdot 10^4$) through fluorescence automated cell sorting[223].

To further estimate the production stability of the medium-strength, high-producing BCD19 and the strongest translation initiation element BCD2, we assessed production per cell by flow cytometry during longer serial cultivation experiments up to 72 hours. Remarkably, BCD19-YidC-GFP and BCD19-AraH-GFP result in stable homogenously producing populations, even after 72 hours, which was also the case for BCD2-AraH-GFP (Figure 2d,h, Figure S3). For BCD2-YidC-GFP, however, only 2 out of 4 pre-cultures prepared for the stability experiment maintained their initial production level, despite the fact that the colonies used for initial inoculation were selected on basis of high fluorescence. This demonstrated again an instable production phenotype for the strong BCD2 with YidC, as observed before for the whole-cell fluorescence measurements. The two stably producing pre-cultures were further inoculated for the long-term experiment in fresh medium and their production decreased over the course of time (Figure 2d, Figure S3). The strong translation initiation of YidC-GFP driven by BCD2 seemed to stress the cells, favouring the emergence of non-expressing cells in the population (Figure 2d), which may have been caused by suppressing mutations in plasmids or the genome (not further characterized).

The BCD system was further employed to optimize the production of bacterial rhodopsin proteins. These simple membrane-bound light-harvesting energy systems can be employed for light-driven cell-factories[235] or optogenetic regulation[236]. When properly folded in the cytoplasmic membrane, rhodopsins are known to bind the retinal pigment, leading to red pigmentation of the host cells[237,238]. Rather than cloning all the BCDs in parallel, this time the 22 BCD variants were pooled, cloned into the vectors containing GR or TR, and transformed as a library. Clones of the resulting transformation were then randomly picked and grown in 96-well plates with retinal (Figure S4). For both GR and TR, 11 clearly red-pigmented pellets were identified out of 89 and 96 screened clones, respectively. For each of the rhodopsin proteins, three intensely red clones were selected for further characterization. In the case of GR, one of the selected clones contained BCD14 and the other two carried BCD15. For TR, BCD9, BCD12 and BCD21 were identified. All these variants are in the medium-strength range of the BCD system.

Production of GR and TR by the best rhodopsin-producing BCD variants was scaled up from deep-well plates to 10 mL cultures in 50 mL tubes. While all elements rendered very similar levels of GR expression (Figure 2i), the production of TR significantly differed from one BCD to another; BCD9 performed significantly better than BCD12, which produced more than BCD21 (Figure 2j). These three variants performed very similarly when grown in deep-96-well plates. while the results of the production assays in 50 mL tubes are significantly different. This indicates that results are not always comparable when scaling-up from deep-well plates (0.5 mL culture) to test tubes (10 mL culture). The previously generated libraries of YidC-GFP and AraH-GFP were also grown in deep-well plates (0.5 mL culture) and their performance was compared to that of 10 mL cultures in 50 mL tubes (Figure S5), confirming differences in performance depending on culture conditions. This reflects the common challenge of optimizing and scaling up recombinant production. However, the optimization for the BCD system in up scaled conditions is quite feasible given the limited number of tuning variants that need to be tested.

We then compared the production of GR and TR to the levels produced by the Lemo21(DE3) system, which was optimized in tubes. Compared to rhodopsin production by the Lemo21(DE3) system, the best performing BCD variants for GR and TR resulted in at least 2-fold and 3-fold higher volumetric rhodopsin production, respectively (Figure 2i,j).

The here employed BCD system outperforms the membrane protein productivities of previously established approaches. Moreover, the BCD system provides stable membrane protein production for at least 72 hours, a stability never reported to date. While most current systems for membrane protein production are limited to specific *E. coli* strains, the applied P14 constitutive promoter allows our system to be generally applicable in all *E. coli* strains. By applying the principles of our method and potentially including different promoters and/or BCDs, our system is likely feasible for bacterial membrane protein production in other bacterial hosts as well. The BCD system may also be applicable for producing certain eukaryotic membrane proteins in *E. coli,* although expression of eukaryotic proteins in bacterial hosts may lead to issues that cannot be solved just by tuning expression strength and tackling RBS-accessibility, such as glycosylation or the requirement of eukaryotic-like membrane lipids[210]. Overall, it is anticipated that the here described approach for bacterial membrane protein production will be useful for many future studies, ranging from biochemical characterization to cell factory engineering.

3.4 Materials and methods

3.4.1 Strains and culture conditions

E. coli strains (Table 1) were cultured in liquid Lysogeny Broth (LB) or LB-agar with appropriate antibiotics, kanamycin (50 µg/mL) and/or chloramphenicol (35 µg/mL). For rhodopsin production, all-trans retinal (Sigma-Aldrich) was added to cultures from a 20 mM ethanol stock to a final concentration of 20 µM and re-added once after 2-4 hours of cultivation or induction to compensate for degradation of the unstable pigment. Long-term stability experiments were performed over 72 hours by diluting 4 replicate cultures 1:50 every 24 hours.

Table 1: E. coli strains used in this study.

Strain	Description	Origin
E. coli DH5α	F– λ– fhuA2 Δ(argF-lacZ)U169 phoA glnV44 Φ80 Δ(lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17	NEB
E. coli BL21(DE3)	F– λ– fhuA2 [lon] ompT gal (λ DE3) [dcm] ∆hsdS	NEB
E. coli Lemo21(DE3)	F– λ– fhuA2 [lon] ompT gal (λ DE3) [dcm] ∆hsdS/ pLemo(CamR)	NEB
E. coli BL21(DE3) pLysS	F– λ– fhuA2 [lon] ompT gal (λ DE3) [dcm] ∆hsdS/pLysS (CamR)	Promega

E. coli BL21(DE3) was used to express all the BCD constructs. Tube cultures were performed in 10 mL LB medium in 50 mL Greiner tubes. 1% or 2% overnight pre-cultures were used to start the cultures and these were grown for 22 hours at 250 rpm, at 30°C for AraH-GFP and YidC-GFP, and 37°C for GR and TR. Pre-cultures (100 µL/well) and expression cultures (500 µL/well) for library screening of *E. coli* BL21(DE3) were performed in MASTERBLOCK® Deep 96-Well plates (Greiner). Plates were kept in a plastic bag with a humidified atmosphere to prevent evaporation and grown for 22 hours at the defined temperature and 900 rpm.

E. coli Lemo21(DE3) was used to produce YidC, as well as GR and TR. Fresh colonies were used to inoculate pre-cultures, as the use of re-streaked glycerol stocks was reported to lead to severely reduced recombinant protein production in LEMO[216]. Overnight pre-cultures were used to inoculate 1%-2% in 10 mL LB medium in 50 mL Greiner tubes with different L-rhamnose concentrations (0, 50, 100, 500, 1000 and 2000 μM). At an OD_{600} of 0.35-0.45, cells were induced with IPTG (isopropyl β-D-1-thiogalactopyranoside) at a concentration 0.4 mM, and for rhodopsin production all-trans retinal was added as well, and re-added 2-4 hours after induction. After induction, cells were grown for 22 hours (30°C for YidC-GFP/AraH-GFP 37°C for rhodopsins, 250 rpm) and then harvested for production-level analysis.

E. coli BL21(DE3) pLysS was used for the production of AraH-GFP from pET28(+)-opt-AraH-GFP-his. The strain was cultured as previously published[239]. In short, 2% overnight pre-culture was inoculated in 10 mL LB in 50 mL tubes, and incubated at 37°C, 180 rpm until OD_{600} of 0.25-0.35 was reached. Then the cultures were induced with 1.0 mM IPTG and incubated for 5 and 22 hours at 25°C or 30°C at 180 rpm.

3.4.2 Plasmid construction

Plasmids used for our study can be found in Supplementary Table 2, oligos used are in Supplementary Table 3. PCR products were generated using either Phusion polymerase (ThermoFisher) or Q5 polymerase (NEB) according to manufacturer’s protocols. Assemblies were mostly performed using type IIS restriction enzymes BsaI or BbsI (NEB) and subsequent ligation by T4 ligase (NEB). When appropriate PCR samples were treated with DpnI (NEB) to digest template DNA. For the construction of some plasmids Gibson Assembly was performed using the NEBuilder® HiFi DNA Assembly (NEB) according to the manufacturer’s protocol. Agarose gel DNA purifications and DNA purifications were respectively performed using ZymocleanTM Gel DNA Recovery Kit and DNA Clean & Concentrator (Zymo Research). Assemblies were generally confirmed by Sanger sequencing (GATC Biotech).

For construction of the BIOFAB-P14 backbone plasmid for YidC-GFP, *yidc-gfp* was amplified from pET28a(+)-YidC-GFP (BG7342,BG7343) and assembled into a BIOFAB-P14 vector pFAB3913 (amplified from pFAB3913 by BG7338,BG7339) via Gibson Assembly. pFAB3913 was a kind gift of Drew Endy and obtained from Addgene (Plasmid #47816)[195]. For the construction of the BIOFAB-P14 assembly with AraH-GFP, *arah-gfp-his* was amplified from pET28a(+)-AraH-GFP-his (BG8448,BG7937), and together with one BCD oligo pair, introduced in the PCR amplified BIOFAB-P14 backbone from pFAB3913 (BG7933, BG7934) by three part ligation. pET28a(+)-AraH-GFP was a kind donation of Daniel Daley[240]. The gene encoding GR was codon-optimized for *E. coli* (Supplementary data 1) and synthesized by GeneArt for insertion, initially for another study together with an N-terminal GFP fusion, into BglBrick vector pBbE0A (Addgene # 35372)[241]. GR-GFP was then PCR amplified from pBbE0A-GR-GFP (BG5971, BG5972) for introduction into BIOFAB standard plasmids pFAB3913 by Golden Gate assembly. As the GFP fusions at the C-terminus of GR were hampering proper retinal incorporation into GR, those were removed from the constructs using phosphorylated primers BG6162 and BG6163 and subsequent plasmid recircularization with T4 ligase. TR was codon-harmonized using our public Codon Harmonizer Tool (http://codonharmonizer.systemsbiology.nl/) [188,242] (Supplementary data 1), subsequently synthesized and cloned into pGFPe by GeneArt. For construction of the BIOFAB-P14 plasmid for TR, TR was amplified from pGFPe-TR (BG7340, BG7337) and introduced in a BIOFAB-P14 backbone (pFAB3913 amplified from BG7339, BG7341) by Gibson assembly.

For the construction of the BCD variant libraries, the above constructed BIOFAB-P14 plasmids were PCR amplified to introduce BsaI or BbsI type IIs restriction sites (BG7784, BG7785 for YidC-GFP and BG8448, BG7934 for AraH-GFP, BG7335, B7336 for GR and BG7505, BG7506 for TR). For the BIOFAB-P14 plasmid, YidC-GFP BbsI was used, as BsaI gave digestion issues. The BsaI/BbsI-digested PCR products were subsequently assembled one by one with the annealed, phosphorylated oligo pairs (BG7291-BG7334) for each of the BCD variants (YidC-GFP, AraH-GFP), or with the complete pool of all BCD variants (GR, TR). The oligos used hereto were first phosphorylated using T7-polynucleotide kinase (NEB) according to the manufacturer’s protocol. Then, pairs of oligos for each BCD variants were annealed, by heating to 95°C for 3 min and gradual cooling to RT in 30 min. The annealed oligo pairs were stored for long-term usage, both individually and pooled. 0.5 µl of individual or pooled annealed oligos (10 ng/µL) was ligated with ~100 ng BsaI/BbsI-digested, dephosphorylated PCR products. Those ligation mixes were transformed to directly to *E. coli* BL21(DE3) (YidC-GFP, AraH-GFP) or first through *E. coli* DH5α for TR and GR. For AraH-GFP clones with BCD21 were not obtained in the first attempts, and hence not further pursued as this low-strength BCD is probably not so relevant. For YidC-GFP and ArH-GFP at least three colonies for each BCD variant were picked and sequence verified, for some more colonies had to be picked to obtain at least three correct replicate clones.

The pET28a(+)-AraH-GFP vector with an optimized sequence at the 5'UTR:CDS junction was reconstructed based on the optimal junction found sequence before[239] by PCR amplification of pET28(+)-AraH-GFP by phosphorylated primers (BG8565,BG8566) and recirculation by T4 ligase.

For validating membrane protein production of GR and TR in *E. coli* Lemo21(DE3), pET28a(+)-expression vectors were generated. GR and TR genes were subcloned into the pET28a(+)-derived pGFPe and the C-terminal GFP fusion was removed by PCR amplification with phosphorylated primers (BG6696, BG6697 for GR, BG6696, BG8454 for TR) and recircularization by T4 ligase.

Table 2: Plasmids used in this study.

Plasmid name	***Antibiotic marker***	***Origin of replication***	***Important components***	***Reference***
pFAB-P14-BCD#-YidC-GFP	Kan	p15a	P14, varying BCDs, YidC-GFP-his	This work
pFAB-P14-BCD#-AraH-GFP-his	Kan	p15a	P14, varying BCDs, AraH-GFP-his	This work
pFAB-P14-BCD#-GR-his	Kan	p15a	P14, varying BCDs, GR-his	This work
pFAB-P14-BCD#-TR-his	Kan	p15a	P14, varying BCDs, TR-his	This work
pET28a(+)-YidC-GFP	Kan	pBR322	P_{T7}, YidC-GFP fusion	[214]
pET28(+)-opt-AraH-GFP-his	Kan	pBR322	P_{T7}, opt junction, AraH-GFP-his	Reconstructed as published[239]
pET28a(+)-GR-his	Kan	pBR322	P_{T7}, GR-his	This work
pET28a(+)-TR-his	Kan	pBR322	P_{T7}, TR-his	This work
pFAB3913	Kan	P15a	P14,BCD9,RFP	[195]
pLemo	Cam	P15a	PrhaBAD, lysY	[243]
pGFPe	Kan	pBR322	pET28a(+) derived for N-term GFP-his fusions	[244]
pGFPe-AraH (pET28(+)-AraH-GFP-his)	Kan	P15a	P_{T7}, AraH-GFP-his fusion	[240]
pBbE0A-GR-GFP	Amp	colE1	GR-GFP-his fusion protein	This work
pGFPe-GR	Kan	pBR322	P_{T7}, GR-GFP-his fusion	[242]
pGFPe-TR	Kan	pBR322	P_{T7}, TR-GFP-his fusion	GeneArt/this work

3.4.3 Whole-cell GFP quantification

Production of YidC-GFP and AraH-GFP was estimated using whole-cell GFP fluorescence as described before[245]. In short, 1 mL of culture (or 0.5 mL from deep-well cultivations) was spun down (10 min, 13,000xg, 4^{o}C) and the pellet was resuspended in ice-cold 100 µL PBS (phosphate buffer saline) and incubated at 4°C for at least 1 hour for further maturation of GFP. After this, suspensions were centrifuged (10 min, 13,000xg, 4^{o}C) and resuspended in 100 µL PBS, then they were transferred to a black 96-well plate with transparent bottoms (Greiner). Fluorescence was directly measured using excitation at 485 nm and emission at 512 nm at a constant gain value (75) (BioTEK SynergyMX).

3.4.4 In-gel fluorescence assay

To validate if the GFP signal originated from full-length fusions of YidC-GFP and AraH-GFP, an in-gel fluorescence assay was performed. Cultures were centrifuged for 5 minutes at 13,000xg and the pellets were stored at -20°C. After thawing, pellets were resuspended to an estimated final concentration of 5 µg protein/µL in 50 mM kPi buffer (pH 7.5) (assuming 150 mg protein/L for OD_{600} of 1). This buffer was supplemented with 1 mM $MgSO_4$, 10% glycerol, 1 mM EDTA, 0.1 mg/mL DNase and 10 mg/mL lysozyme. Cells were lysed for 1 hour at 300 rpm at room temperature and stored at -20°C for later analysis. 4x Laemmli buffer (Biorad) was added to the cell lysate, incubated for 5 minutes at 37°C (and not higher to prevent denaturation of folded GFP). After incubation and right before loading in gel, the samples were shortly sonicated with three 0.1 ms pulses (Bandelin SONOPLUS HD 3100) to reduce sample viscosity. Twenty-five µL of diluted sample (containing 6.25 µg protein in all cases except for AraH-GFP dry blotting, where 93.5 µg were used) were loaded and run on a 10% Mini-PROTEAN® TGX™ protein gel (Biorad) in Tris-Glycine-SDS buffer (25 mM, 250 mM and 0.1% respectively). In-gel fluorescence was then imaged with a Syngene G-box using a 525nm filter.

3.4.5 Western blot

After imaging in-gel fluorescence, the proteins were transferred overnight from the gel to 0.2 micron PVDF membranes by wet transfer in a tank blotting system (Bio-Rad) at 70 mA, or for AraH-GFP blotting with anti-his by dry transfer (iBlot2® Dry Blot system, ThermoFisher) using standard settings. After transfer, the membranes were blocked in PBST (3% BSA) for 1 hour, washed twice in PBST, incubated with 6x-His Tag Mouse Monoclonal Primary Antibody (3D5) (ThermoFisher) or IbpB rabbit antiserum (1/5000)[214,243] for 2 hours, washed three times in PBST and finally incubated for 1 hour with either Goat anti-Mouse IgG (H+L) Secondary Antibody-HRP or Goat anti-Rabbit IgG (H+L) Secondary Antibody-HRP (ThermoFisher) (1/20000), respectively. Blots were developed using SuperSignal West Pico PLUS Substrate (ThermoFisher) following manufacturer's instructions and imaged in a Syngene G-box.

3.4.6 Flow cytometry

Samples for flow cytometry were washed in PBS, diluted 10,000 times in 1 mL PBS ($<10^6$ cells per mL), supplemented with 2 µL of 0.2 mM FM4-64 dye (ThermoFisher) to stain all cells and discern them from debris, and left on ice for 30 minutes, as performed before.[246] Single-cells were analyzed for GFP fluorescence, capturing 10,000 events for each sample by an BD Accuri C6 Flow Cytometer. Data were processed using BD Accuri C6 software.

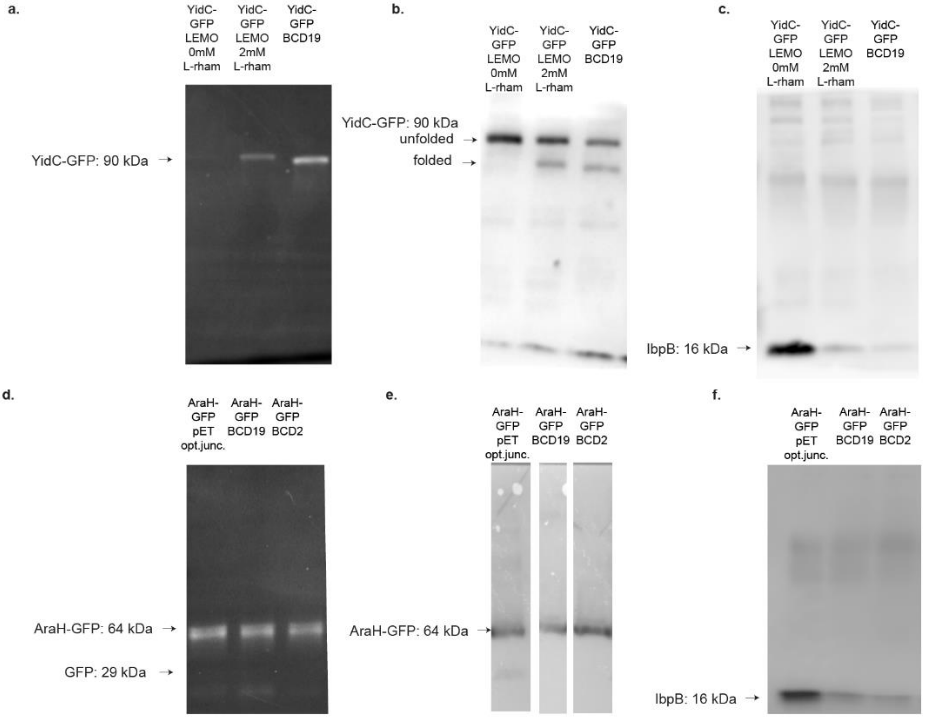
3.4.7 Rhodopsin quantification

Rhodopsin quantification was adapted from a previous method[247]: 10 mL of culture were resuspended in 295 µL extraction buffer, and frozen for at least 1 hour to increase lysis efficiency. Cells were thawed and additional 295 µL extraction buffer was added, supplemented with 6 mg/mL lysozyme and 0.4 mg/mL DNase. For cell lysis, this suspension was incubated at room temperature for 30 minutes. Rhodopsins were extracted from the crude cell extract by addition of 2.5% (w/v) dodecyl-maltoside (DDM, Sigma) and incubation at 180 rpm for 24-48 hours in the dark. The extraction for GR was performed at room temperature, and to increase the extraction efficiency of TR, its extraction was performed at 65°C. After extraction the mixture was spun down to check for the color of the pellet, and if a colorless pellet was obtained the supernatant fraction was used for spectroscopic quantification. 200 µL of supernatant was transferred to a transparent flat bottom 96-well plate (Greiner) and the absorption spectrum (300-700 nm) was measured (Synergy MX BioTek). Next, 0.1 M hydroxylamine was added to bleach the retinal from the rhodopsin for 1 hour in dark at room temperature with gentle shaking. Then, the absorption spectrum was measured again. The difference absorption spectrum could be generated, and from differential absorption at 540 nm (GR) or 525 nm (TR), the molar rhodopsin concentration was determined, assuming an extinction coefficient of 50,000 (M cm^{-1}) for both rhodopsins, and was converted to mg/L based on rhodopsin molecular weights.

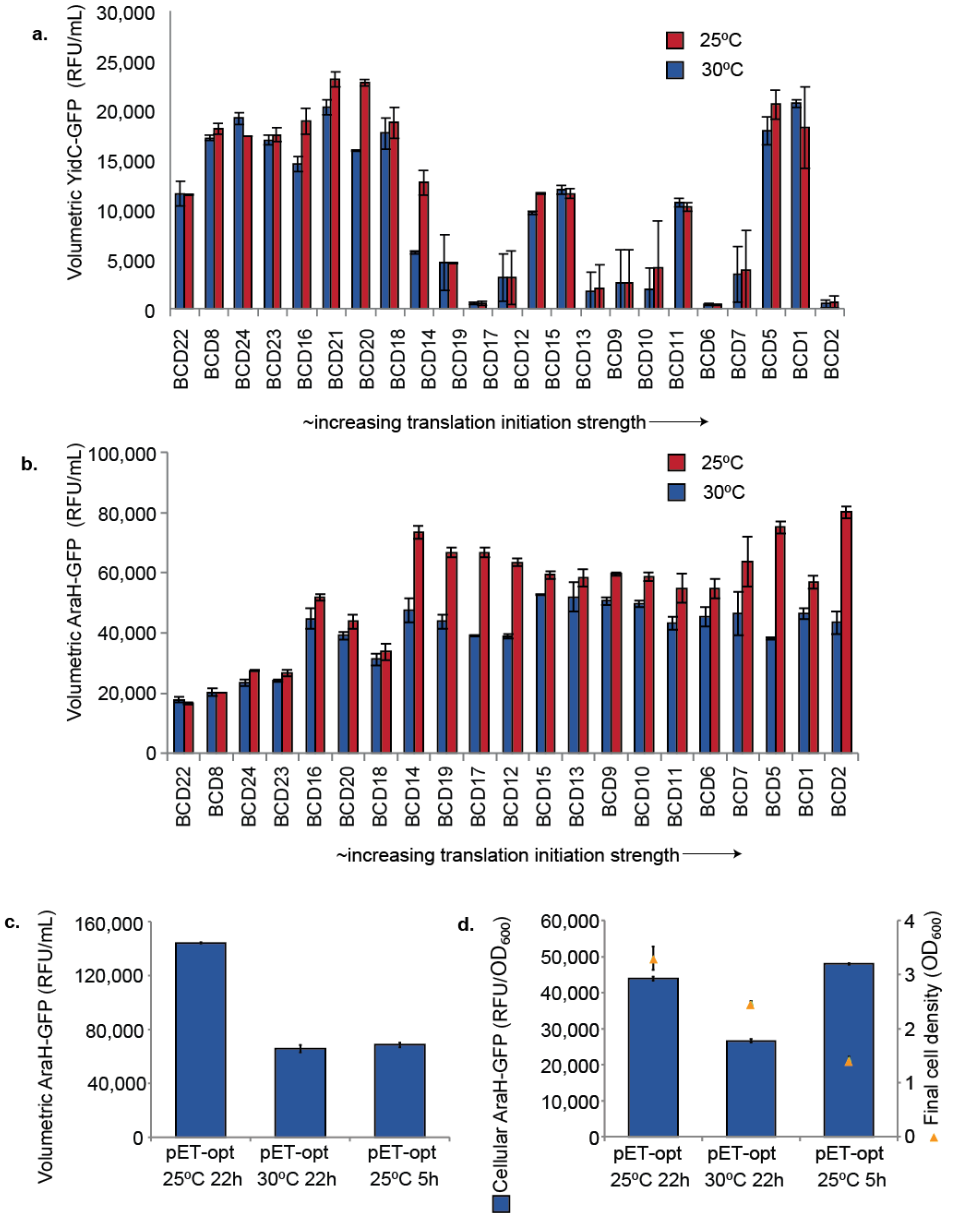
3.5 Acknowledgements

We thank Wendy van der Vliet and Snezana Gegic for their kind assistance with the flow cytometry experiments. We thank Daniel Daley for sharing pGFPe-AraH-his. All BIOFAB plasmids were a kind gift of Drew Endy. We gratefully acknowledge financial support from the Wageningen University IP/OP Systems Biology (grant KB-17-003.02-024) and the Netherlands Organization for Scientific Research (NWO) via the Spinoza and SIAM Gravitation Grant (Project 024.002.002) to WMdV, the BaSyC Graviation grant to JvdO (024.003.019) and the Rubicon grant to NC (019.163LW.035). The funders had no role in study design, data collection and analysis, or preparation of the manuscript.

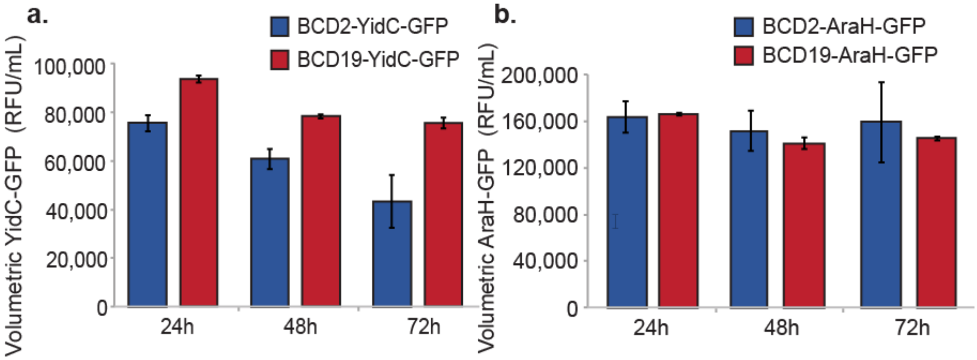
3.6 Supplementary data



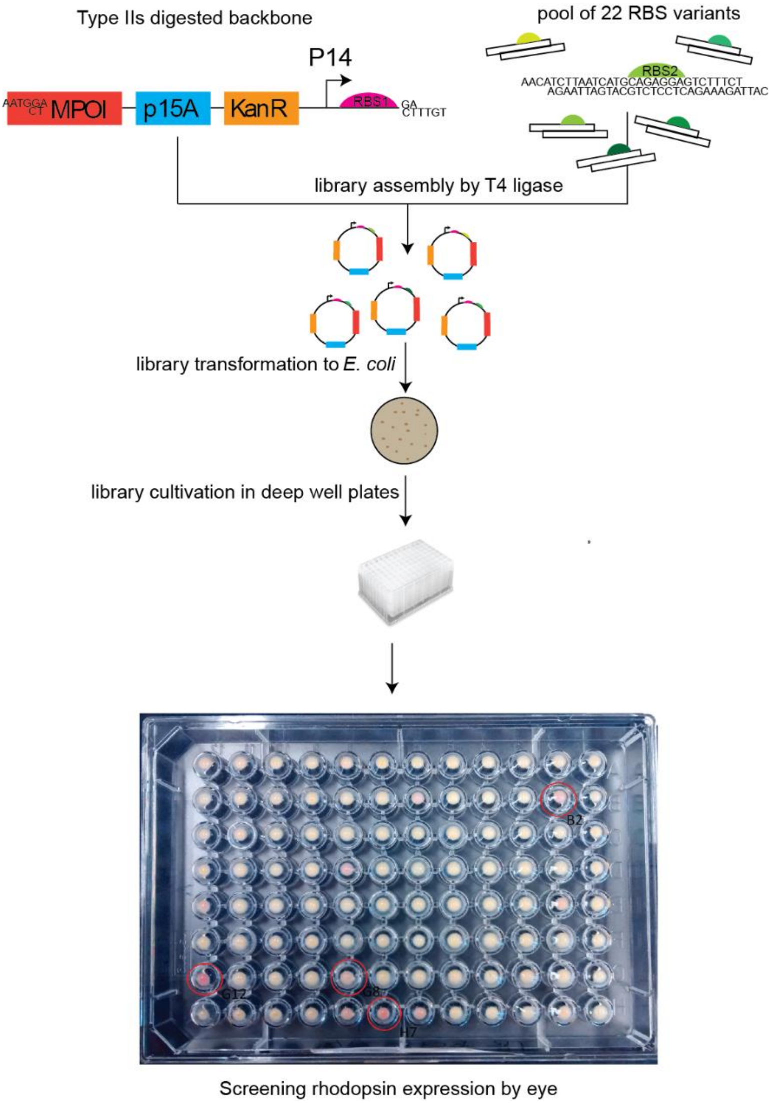
Supplementary Figure 1: In-gel fluorescence and complete Western blots. *(a) The fluorescent signal from YidC-GFP cultures completely originated from a single-sized protein product, corresponding to full-length YidC-GFP, as checked for production from some BCD variants and Lemo21(DE3). (b) Western blot performed with anti-his-tag antibody to visualize both inclusion body and well-folded YidC-GFP-his. (c) Western blot performed with anti-IbpB[229,243] to visualize inclusion body binding protein B for YidC-GFP expression. (d) The fluorescent signal from AraH-GFP cultures comes predominantly from full size AraH-GFP, and only a very minor fraction from a small product, probably loose GFP, as confirmed for two BCD variants and production from the pET-opt-AraH-GFP vector. (e) Western blot performed with anti-his-tag antibody to AraH-GFP-his (f) Western blot performed with anti-IbpB[229,243] to visualize inclusion body binding protein B for AraH-GFP expression. In-gel fluorescence and Western blots were performed on samples obtained from representative cultures, cultivated for 22 hours at 30ºC (or 25ºC for pET-opt-AraH-GFP).*



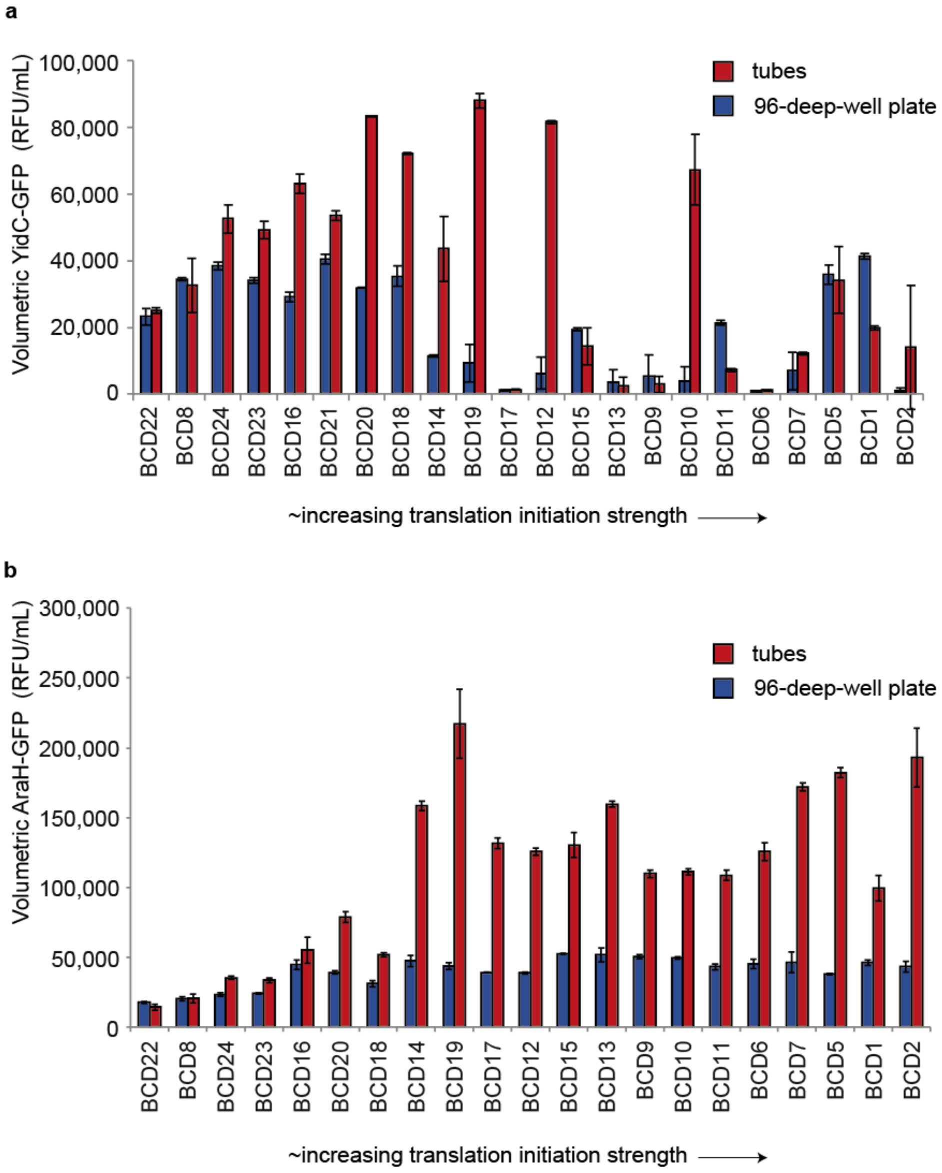
***Supplementary Figure 2: Production of YidC-GFP** (a) **and AraH-GFP** (b) **from the BCD variants was compared at 25 °C and 30°C after cultivation in 96-deep-well plates for 22h**. Whole-cell fluorescence for pET-opt-AraH-GFP was also measured per volume (c) and normalized per OD_{600} (c) at 25ºC for 5 hours (as in original publication of optimized vector[239]) and at 30ºC for 5 hours and at 25ºC and after 22h of induction.*



***Supplementary Figure 3: Volumetric production of YidC-GFP** (a) **and AraH-GFP** (b) **during 72h stability experiment.** E. coli BL21(DE3) harbouring BCD vectors were re-inoculated 1:50 into fresh LB kanamycin medium every 24 hours. Volumetric production data come from 4 biological replicate cultures, except for BCD2-YidC-GFP, where only 2 pre-cultures were available that still had an initial high production level.*


Type IIs digested backbone
pool of 22 RBS variants
P14
AATGGA
CT
MPOI
p15A
KanR
RBS1
GA
CTTTGT
RBS2
AACATCTTAATCATGCAGAGGAGTCTTTCT
AGAATTAGTACGTCTCCTCAGAAAGATTAC
library assembly by T4 ligase
library transformation to E. coli
library cultivation in deep well plates
B2
G12
G8
H7
Screening rhodopsin expression by eye

Supplementary Figure 4: Workflow for pooled cloning and visual screening for high-producing rhodopsin clones. *The plate with cell pellets depicted contains the BCD-TR library from which B2, G12 and H7 (all encircled red) were selected for sequence analysis and tube cultivation.*



Supplementary Figure 5: Comparing cultivation of BCD strains in tubes versus 96-deep-well plates. *All BCD variants for YidC-GFP (a) and AraH-GFP (b) were cultivated in 50 mL tubes (10 mL medium) and 96-deep-well plates (0.5 mL medium). Volumetric, membrane-integrated production was measured and normalized using the same method; this demonstrated different performance dependent on the cultivation conditions. All cultures were grown at 30ºC and data for tubes are from at least 3 biological replicates and for plates from 2 biological replicates.*

3.6.1 Codon-optimized GR + his

ATGCTGATGACCGTTTTTAGCAGCGCACCGGAACTGGCACTGCTGGG
TAGCACCTTTGCACAGGTTGATCCGAGCAATCTGAGCGTTAGCGATAG
CCTGACCTATGGTCAGTTTAATCTGGTGTATAACGCATTTAGCTTTGCC
ATTGCAGCAATGTTTGCAAGCGCACTGTTTTTTTTCAGCGCACAGGCA
CTGGTTGGTCAGCGTTATCGTCTGGCCCTGCTGGTGAGCGCAATTGTT
GTTAGCATTGCAGGCTATCATTATTTCCGCATTTTCAATAGCTGGGATG
CAGCATATGTTCTGGAAAATGGTGTTTATAGTCTGACCAGCGAGAAAT
TCAATGATGCCTATCGTTATGTTGATTGGCTGCTGACCGTTCCGCTGC
TGCTGGTTGAAACCGTTGCAGTTCTGACCCTGCCTGCAAAAGAAGCAC
GTCCTCTGCTGATCAAACTGACCGTTGCAAGCGTTCTGATGATTGCAA
CCGGCTATCCGGGTGAAATTAGTGATGATATTACCACCCGTATTATTT
GGGGCACCGTTAGCACCATTCCGTTTGCATATATTCTGTATGTTCTGT
GGGTTGAACTGAGCCGTAGCCTGGTTCGTCAGCCTGCCGCAGTGCAG
ACCCTGGTGCGTAATATGCGTTGGTTACTGCTGCTGAGCTGGGGTGTT
TATCCGATTGCATATCTGCTGCCGATGCTGGGTGTGAGCGGCACCAG
CGCAGCAGTTGGTGTTCAGGTTGGTTATACCATTGCAGATGTTCTGGC
CAAACCTGTTTTTGGTCTGCTGGTTTTTGCAATTGCCCTGGTTAAAACC
AAAGCAGATCAAGAAAGCAGCGAACCGCATGCAGCAATTGGTGCAGC
AGCAAATAAAAGCGGTGGTAGCCTGATTAGCCACCACCACCACCACC
ACTAA

3.6.2 Codon-harmonized TR + his

ATGCGGATGTTACCCGAACTGAGCTTTGGAGAATATTGGTTAGTCTTT
AACATGCTGAGCCTGACCATTGCGGGCATGTTAGCGGCGTTTGTCTTT
TTTCTGTTAGCTCGGAGCTATGTGGCGCCGCGTTATCATATTGCGCTG
TATCTGAGCGCGCTGATTGTCTTCATTGCGGGCTATCATTATTTAAGGA

TTTTCGAAAGCTGGGTGGGCGCGTATCAGTTACAGGATGGCGTATATG
TGCCCACTGGCAAACCGTTTAACGATTTTTATCGTTATGCGGATTGGC
TGCTGACCGTGCCGTTACTGCTGTTAGAACTGATTTTAGTCCTAGGTC
TTACCGCTGCGCGTACCTGGAACCTAAGCATTAAACTTGTGGTGGCGT
CAGTCTTAATGTTAGCGCTTGGCTATGTGGGAGAAGTGAACACTGAAC
CGGGACCGCGGACCTTATGGGGCGCGTTAAGCAGCATACCGTTTTTT
TATATTCTGTATGTGCTGTGGGTGGAATTAGGTCAGGCGATTCGCGAA
GCTAAATTTGGTCCGCGGGTGTTAGAATTATTAGGTGCGACCCGTCTG
GTCCTGTTAATGAGCTGGGGTTTTTATCCGATTGCGTATGCGTTAGGT
ACCTGGCTGCCGGGAGGCGCTGCGCAGGAAGTGGCGATTCAGATAG
GTTATAGCCTTGCTGATTTAATTGCGAAACCGATTTATGGTTTATTAGT
CTTTGCGATTGCGCGCGCGAAAAGCCTGGAAGAAGGTTTTGGTGTGG
AAGCTAAAGCGGCGTTAGAGCACCACCACCACCACCACTAA

Chapter 4. CRISPR with a happy ending: non-templated DNA repair for prokaryotic genome engineering

Max Finger-Bou[1], Enrico Orsi[2], John van der Oost[1], Raymond H.J. Staals[1]

[1]Laboratory of Microbiology, Wageningen University and Research, Wageningen, the Netherlands

[2]Bioprocess Engineering, Wageningen University and Research, Wageningen, the Netherlands

Adapted from publication:

Finger-Bou, M., Orsi, E., van der Oost, J., & Staals, R. H. (2020). CRISPR with a happy ending: Non-templated DNA repair for prokaryotic genome engineering. *Biotechnology Journal*, *15*(7), 1900404.

4.1 Abstract

The exploration of microbial metabolism holds enormous power to help society shift towards an environmentally sustainable economy and to tackle a plethora of problems related to the burdens of human consumption. Microbial cell factories have the potential to catalyze a wide range of processes which are currently either unavailable, unsustainable and/or inefficient. The metabolism of microorganisms can be optimized and further expanded using genetic engineering tools, like the CRISPR-Cas systems. These tools have revolutionized the field of biotechnology, as they greatly facilitate the genetic optimization of organisms from all domains of life. These and other nucleases mediate double-strand DNA breaks, which must be repaired to prevent cell death. These breaks can be repaired in prokaryotes through either homologous recombination, when a DNA repair template is available, or through template-independent end joining, of which two major pathways are known. These end joining pathways depend on different sets of proteins and mediate DNA repair with different outcomes. Understanding these DNA repair pathways can therefore be advantageous to steer the results of genome engineering experiments. In this review, we discuss different strategies for the genetic engineering of prokaryotes through the exploitation of either the non-homologous end joining (NHEJ) or the alternative end joining (AEJ) DNA repair pathways, which are independent of exogenous DNA repair templates.

4.2 Introduction

Climate change, growing world population and scarcity of resources are issues gaining increasing attention from society and from the research community. Consequently, developing strategies for decoupling economic growth from the emission of greenhouse gases has become a pressing issue[67]. Biotechnology, at the core of the emerging concept of bio-economy, aims to facilitate the replacement of petroleum-based chemical synthesis by biocatalysis in which microbial cell factories convert renewable feedstocks into a wide range of products[68,69,248]. These sustainable products can be high value molecules such as pharmacologically active compounds[249], but also cheaper commodity chemicals[250] and biofuels[251]. While harnessing microbial production for the pharmaceutical and nutraceutical sectors is already a reality[70,71,252], the replacement of many petroleum-based commodity chemicals with their greener counterparts is still to be realized. The available microbial engineering tools, as well as our understanding about their precise molecular workings, are still often the bottleneck for the optimization of microbial cell factories. In cases where homologous recombination is inefficient, template-independent DNA repair represents an attractive alternative for prokaryotic genome editing. In this mini-review, we provide a concise summary of the two known prokaryotic template-independent end joining pathways, and we elaborate on different strategies to employ CRISPR-Cas systems and other nucleases in combination with these native or heterologously expressed DNA repair pathways. We hope our work will encourage researchers to explore the emerging field of non-templated prokaryotic engineering.

4.3 CRISPR-Cas systems, user-friendly tools for microbial engineering

Although several strategies have been applied for microbial genome engineering with moderate success before the rise of the CRISPR-Cas tools, there are certain drawbacks associated with their use. Allelic exchange, based on the introduction of positive and/or negative markers through homologous recombination, is often reported to carry high levels of false positive mutants, particularly in the case of negative selection markers[253,254]. An elegant alternative to introduce point mutations as well as big deletions and insertions is recombineering, which typically exploits bacteriophage proteins to mediate genome manipulations[255]. While these enzymes increase the efficiency of *in vivo* recombination, the heterologous expression of bacteriophage proteins required for recombineering can be cumbersome. Moreover, this approach necessitates the generation of a recombination template for every gene to be mutated[256], which is a bottleneck in large mutagenesis experiments. Whereas selection can be carried through positive and/or negative markers, scar-less mutations require multiple rounds of crossover recombination, which often complicates experiments as it involves the screening of many colonies and it often entails high rates of false positive mutants[257].

Additionally, numerous efforts have been focused at engineering sequence-specific endonucleases. In particular, zinc finger nucleases (ZFNs)[44] and transcription activator-like effector nucleases (TALENs)[45] have been extensively exploited for genome editing. These synthetic protein-complexes recognize specific nucleic acid sequences by protein-DNA interactions. Reprogramming their target specificity involves time-consuming protein engineering, which is inconvenient when multiple sequences need to be targeted. Alternatively, homing mega-nucleases, like I-SceI, can be repurposed to drive targeting of their recognition sites, but this requires the presence or introduction of a restriction site sequence in the DNA to be targeted[258], thereby limiting its convenience.

The discovery of an adaptive immunity system in prokaryotes[46,48,49,259], namely clustered regularly interspaced short palindromic repeats (CRISPRs) and its associated proteins (Cas), has shaken the grounds of genome editing. Unlike the previously engineered endonucleases, the different Cas effector proteins are guided to the target nucleic acids by small RNA molecules[259]. As such, CRISPR-Cas systems are much easier to reprogram towards targeting other DNA sequences than previous genome engineering technologies[257].

Given its advantages over other traditional mutagenesis methods, the RNA-guided DNA endonuclease activity of CRISPR-Cas systems has rapidly become the standard tool for modern genome editing, initially by allowing for easy counter-selection after homologous recombination experiments[51]. New variants are still being discovered and added to the prokaryotic CRISPR toolbox[260], such as RNA-guided RNA endonucleases[261], thermophilic Cas proteins variants[262,263] or RNA-guided DNA integrases[264,265]. Many exciting synthetic variants have been engineered to allow for novel functions such as transcriptional regulation[266,267], DNA nicking[268], *in vivo* base editing[269], novel fusions of different nucleases[270] and protein chimeras that edit by reverse transcription of a prime editing guide RNA[271].

Despite the emergence of novel functionalities in the CRISPR-Cas toolkit, the main feature of these and other nucleases is the generation of target-specific double-strand DNA breaks to initiate a certain edit at the target site. This type of DNA break is lethal if left unrepaired; DNA repair mechanisms are therefore key for cell survival, but also for the introduction of the desired mutations.

4.4 Non-templated DNA repair in prokaryotes

When a copy of the broken DNA is available, prokaryotes can repair double-strand DNA breaks with accuracy[53]. For most prokaryotes, a repair template is only available during the logarithmic phase of growth, when more than one copy of the chromosome is present in the cell during replication[272]. In bacteria, chromosomal breaks can be recognized by multi-subunit helicase-nuclease complexes which process DNA ends and drive homologous recombination[55]. For instance, the well-studied RecBCD complex recognizes and processes DNA ends, and loads multiple RecA proteins onto resected, single stranded DNA, which facilitates homologous recombination[55], leading to the accurate repair of the chromosomal DNA break.

Naturally, when no repair template is available, cells rely on their intrinsic ability to join DNA ends. Understanding the mechanisms that drive these alternative repair pathways is key to embracing the full potential of template-independent genome editing[273]. Microbes rely on different sets of enzymes which protect, process and ligate DNA ends[60,274]. To date, two natural pathways have been described in bacteria: non-homologous end joining (NHEJ)[275,276] (Figure 1A) and alternative end joining (AEJ; also referred to as microhomology-mediated repair, MMEJ)[59,277] (Figure 1B).

NHEJ is thought to be active in all eukaryotes[278], a sub-set of bacteria[279] and archaea[280]. Very well studied in humans, NHEJ involves more than ten proteins, but the core consists of the heterodimer Ku70/Ku80 and the ligase IV, as extensively reviewed elsewhere[278,281]. In contrast, the mechanisms governing DNA repair in absence of a repair template have remained more elusive in prokaryotes. Being firstly predicted through *in silico* analyses, the prokaryotic NHEJ machinery was suggested to be considerably simpler than its eukaryotic counterpart, with only two proteins predicted to intervene, Ku and LigD[282,283]. More frequent in bacteria, NHEJ in archaea is however considered to be rare, as not many species have the Ku protein[280]; so far, the full, canonical NHEJ system has only been described in the archaeal species *Archaeoglobus fulgidus*[282] and in *Methanocella paludicula*[284].

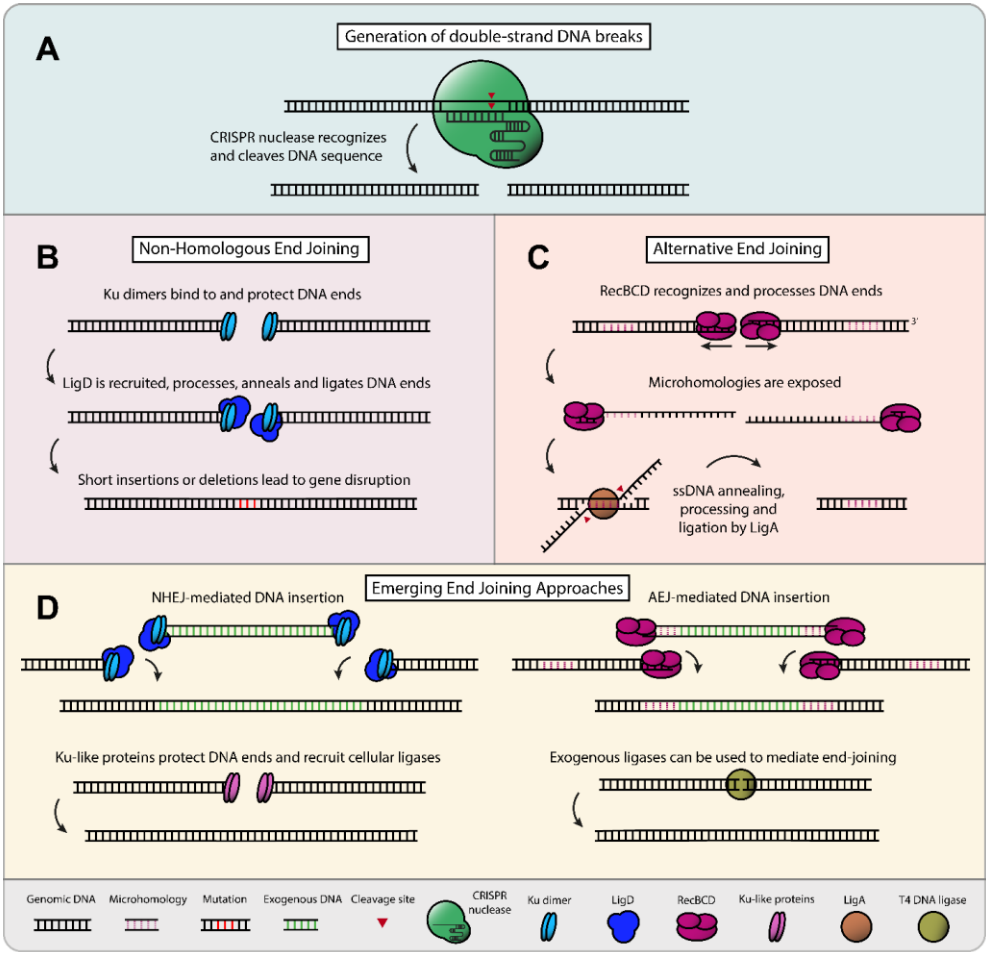

A
Generation of double-strand DNA breaks
CRISPR nuclease recognizes and cleaves DNA sequence
B
Non-Homologous End Joining
Ku dimers bind to and protect DNA ends
LigD is recruited, processes, anneals and ligates DNA ends
Short insertions or deletions lead to gene disruption
C
Alternative End Joining
RecBCD recognizes and processes DNA ends
3'
Microhomologies are exposed
ssDNA annealing, processing and ligation by LigA
D
Emerging End Joining Approaches
NHEJ-mediated DNA insertion
AEJ-mediated DNA insertion
Ku-like proteins protect DNA ends and recruit cellular ligases
Exogenous ligases can be used to mediate end-joining
Genomic DNA
Microhomology
Mutation
Exogenous DNA
Cleavage site
CRISPR nuclease
Ku dimer
LigD
RecBCD
Ku-like proteins
LigA
T4 DNA ligase

Figure 1: Schematic representation of a genome engineering experiment where the genome is targeted by a nuclease and different DNA repair pathways can mend the

double-strand DNA break. *(A) A CRISPR nuclease recognizes and cleaves a target sequence, generating a double-strand DNA break (not to scale). (B) In non-homologous end joining, Ku dimers bind to DNA ends and protect them against the effect of cellular nucleases, and they recruit LigD (or multiple subunits) which then processes, anneals and ligates DNA ends, often generating short insertions or deletions which induce frameshift mutations, leading to gene disruption. (C) In alternative end joining (also named microhomology mediated repair), RecBCD (or other complexes) recognize and resect DNA ends, when short stretches of microhomology (1-9 nt) are exposed, which then allow the annealing of the processed ssDNA. The protruding ssDNA ends are digested by cellular nucleases and the junctions are sealed by cellular DNA ligase A. (D) Novel approaches are emerging where both end joining mechanisms are used to insert exogenous DNA in the genome, Ku-like proteins like Gam from phages μ and λ, and other ligases like phage T4 DNA ligase are also being used to mediate synthetic end joining.*

Just like their eukaryotic homologs, prokaryotic Ku proteins form a ring-like structure that encloses broken DNA ends, protects them from the activity of cellular exonucleases and recruits LigD[60]. LigD, in turn, is a multidomain protein with nuclease, polymerase and (ATP-dependent) ligase activities, organization of which varies between species, and it can also be present as a holoenzyme made of subunits[279,285]. Upon its recruitment by Ku, LigD processes the DNA ends with its nuclease and polymerase activities and ligates them in an ATP-dependent manner[286,287] (Figure 1A). While the proteins are known to be essential during stationary phase in irradiated cultures and in spores during desiccation, they can be knocked out without apparent detrimental effects to cellular fitness under normal growth conditions[272,288].

Despite Ku and LigD being thought to be the only proteins able to fix double-strand DNA breaks and prevent cellular death[274], a Ku- and LigD-independent end joining pathway has been described in *Escherichia coli* by Chayot *et al*[61]. Named alternative end-joining (AEJ), it was proven to mediate plasmid and genomic DNA recircularization *in vivo* without neither Ku nor LigD. A characteristic feature of AEJ is the large reliance on microhomologies (1-9 nucleotides), which are exposed due to the action of the RecBCD complex and enable DNA end annealing and ligation by the NAD-dependent DNA ligase A (LigA)[61] (Figure 1B).

4.5 Strategies for non-templated prokaryotic genome engineering

In Table 1, a summary is given of the different strategies successfully applied for non-templated prokaryotic engineering, including the edited species, the used endonucleases, the origin of the DNA repair mechanisms employed and an outline on the observed mutations and the presence of microhomologies.

Table 1: Summary of the published studies to date where CRISPR and other endonucleases are employed for prokaryotic genome editing in combination with either native or heterologously expressed DNA repair systems. *Abbreviations: Spy,* Streptococcus pyogenes*; Fn,* Francisella novicida*; Sca,* Streptomyces carneus*; Mtu,* Mycobacterium tuberculosis*; Msm,* Mycobacterium smegmatis*; Bsu,* Bacillus subtilis*; Mpa,* Methanocella paludicola*; Sda,* Streptomyces daghestanicus*; Pat,* Pectobacterium atrosepticum*; Ppu,* Pseudomonas putida*; Eco,* Escherichia coli*.*

Edited species	**Endonuclease**	**DNA repair**	**Observed mutations**	**(Micro)homologies**	**Reference**
Mycobacterium smegmatis	I-SceI	Native	Up to 221 bp	Yes	[289]
Escherichia coli	I-SceI, 2 sites	Native	Up to 12.3 kb deletions	Yes	[61]
		Native	Insertion of about 1kb	Yes	
Pectobacterium atrosepticum	Native (Type I-F CRISPR)	Native	40.2 kb deletion	Yes	[290]
Sterptomyces coelicolor	SpyCas9, 1 gRNA	Native	Small insertions (1/4 colonies) Up to 37 kb deletions (3/4 colonies)	Not reported	[291]
		ScaLigD	Small indels (16/16 colonies)	Not reported	
E. coli	SpyCas9 (nickase), 1 gRNA	Native	1 kb deletion	Yes	[292]
	SpyCas9 (nickase), 2 gRNA	Native	36 – 97 kb deletions	Yes	
E. coli	SpyCas9, 1 gRNA	Native	Up to 27.8 kb deletions	Yes	[53]
		MtuKuLigD	9 – 298 bp deletions	Yes	
		μGam	Not reported	Not reported	
E. coli	SpyCas9, 1 gRNA	MtuKuLigD	10 – 267 bp deletions	Yes	[293]
	SpyCas9, 2 gRNA	MtuKuLigD	Up to 17 kb deletions	Not reported	

Edited species	Endonuclease	DNA repair	Observed mutations	(Micro)homologies	Reference
E. coli	SpyCas9, 1 gRNA	BsuKuLigD	13 – 172 bp deletions	Yes	[294]
	SpyCas9, 1 gRNA	MtuKuLigD	10 – 26 bp deletions	Yes	
	SpyCas9, 1 gRNA	MsmKuLigD	13 – 37 bp deletions	Yes	
	SpyCas9, 2 gRNA	MsmKuLigD	Up to 123 kb deletions	Yes	
Methanosarcina acetivorans	SpyCas9, 1 gRNA	MpaKuLigD (*: subunits expressed instead of holoenzyme)	75 bp – 2.7 kb deletions	Yes	[295]
	SpyCas9, 2 gRNA	MpaKuLigD* (*: subunits expressed instead of holoenzyme)	1.3 kb deletions	Yes	
Mycobacterium smegmatis	FnCpf1, 1 gRNA	Native (MsmKuLigD)	Up to 1.5 kb deletions	Not reported	[296]
S. coelicolor	FnCpf1, 1 gRNA	MsmKuLigD	409 – 1624 bp deletions	Not reported	[297]
	FnCpf1, 2 gRNA	MsmKuLigD	Up to 28 kb deletions	Not reported	
E. coli	I-SceI	µGam, EcoLigA		Yes	[298]
Sinorhizobium meliloti	I-SceI	Native (SmeKuLigD)	Plasmid deletions of up to 994bp	Not reported	[299]
		Native (SmeKuLigD)	Insertion of 1.3 kb resistance cassette	Not reported	
		Native (SmeKu3-4)	Up to 343 bp deletion in chromosome	Not reported	
E. coli	SpyCas9, 1 gRNA	MtuKuLigD	Not reported	Not reported	[300]
	SpyCas9, 1 gRNA	T4 ligase	Up to 35 kb deletions	Yes	
	SpyCas9, 1 gRNA	T4 ligase, λGam	Up to 35 kb deletions	Not reported	
E. coli	xCas9-3.7, 1 gRNA	Native	Up to 1.7 kb deletions	Yes	[301]
	xCas9-3.7, 2 gRNA	Native	Up to 83 kb deletions	Yes	
Pseudomonas aeruginosa	Native, CRISPR type I-C	Native	7 – 424 kb deletions	Yes	[302]
Pseudomonas syringae	Pae CRISPR type I-C	Native	55 – 101 kb deletions	Yes	
E. coli	Pae CRISPR type I-C	Native	17 – 106 kb deletions	Yes	

Both native and heterologous DNA repair systems have been successfully employed for template-independent genome engineering of archaea and several bacteria. Although conventional engineering methods are still being used (e.g. I-SceI[61,289,298,299]), the class 2 type II CRISPR-Cas9 system from *Streptococcus pyogenes* is nowadays the most frequently used tool for the prokaryotic genome engineering[53,291–295,300]. Additionally, other CRISPR nucleases are gaining popularity, such as Cas12a from *Francisella novicida*[296,297,303], which has different properties than Cas9[304], and other newly engineered CRISPR-Cas variants[301]. Class 1 CRISPR systems have also been used[290,302] but are far less usual.

Ku and LigD have been, for long, the go-to proteins to transplant template-independent DNA repair pathways. However, the use of different ligases or Ku-like proteins is surging. Gam proteins from bacteriophage μ[298] and λ[300] have been successfully demonstrated to protect linear DNA ends from degradation by cellular nucleases. Interestingly, μGam has been proven to promote binding of DNA ligase A to DNA ends[298], and the ligase from bacteriophage T4 has been used as a single component NHEJ protein, increasing the surviving rates of Cas9-targeted bacteria[300]. Non-canonical end joining proteins prove therefore a useful addition to the molecular toolbox of non-templated genome engineering.

Altogether, the outcomes of genome editing with CRISPR-Cas systems or other endonucleases can be roughly grouped in three categories: gene inactivation, gene insertion and genome minimization.

4.5.1 Gene inactivation

The most straightforward approach to take advantage of the error-prone nature of NHEJ is to generate small indels in protein-encoding genes, causing frameshifts or nonsense mutations, which disrupt the function of genes. An efficient NHEJ pathway in combination with CRISPR-Cas targeting can provide an excellent platform to perform high-throughput functional genomics in different organisms without the need of templates for homologous recombination[305].

4.5.2 DNA insertion

It was demonstrated in *E. coli* that antibiotic resistance cassettes can be acquired through AEJ, independently of Ku and LigD[61], allegedly relying on the action of the essential LigA and other cellular components. Additionally, it was recently proven that Ku and LigD can also mediate acquisition of DNA through classical NHEJ[299]. Further characterization of the different DNA repair mechanisms will likely be needed to make DNA insertion a reliable feature of engineering through end-joining pathways.

4.5.3 Genome minimization

Whereas the use of CRISPR technologies has been mostly targeted at small scale genome engineering, DNA repair pathways independent of exogenous template offer the possibility of minimizing bacterial genomes with ease, opening up opportunities for both fundamental and applied biotechnological research[306].

Both NHEJ and AEJ have been successfully used to generate large deletions in prokaryotes (Table 1). In these cases, the extent of the introduced genomic deletion seems determined by the presence of essential genes in proximity of the targeted locus[61], highlighting the role of DNA repair in mediating large-scale genome rearrangements. Notably, the relatively simple CRISPR type I-C system was recently shown to mediate deletions of up to 424 kb in *Pseudomonas aeruginosa* using only one guide, allegedly facilitated by AEJ[302].

4.6 Native vs heterologous DNA end joining for prokaryotic engineering

Whether it is HDR, AEJ, NHEJ or a combination thereof, all microorganisms have at least one of these DNA repair systems. It is therefore possible, theoretically, to apply CRISPR-Cas tools without the heterologous expression of any exogenous DNA repair pathway. In practice, however, whether native DNA repair systems are adequate to genetically engineer a microorganism depends on two aspects: the type of native DNA repair system and the specific genetic engineering goal.

The spectra of mutations mediated by the different types of DNA repair differ substantially from NHEJ to AEJ. While repair mediated by NHEJ is typically associated with short, unpredictable nucleotide insertions or deletions[279], repair by AEJ can generate deletions from dozens of nucleotides[61] to hundreds of kb[302] and is driven by microhomology. Microhomologies can be computationally predicted[307], which can be instrumental in establishing AEJ as a valuable tool for prokaryotic genome engineering.

Depending on the specific genetic engineering goal, a DNA repair pathway might be more favorable than the other. Small mutations introduced by NHEJ can be very useful to disrupt the function of several genes in iterative cycles or multiplex genome editing[308], as well as to carry functional genomics studies[305]. On the other hand, large genomic deletions driven by AEJ can facilitate the genome minimization of prokaryotes and help unravel the poorly understood secrets of the noncoding genome[309].

4.7 Future directions

A few challenges are to be faced when it comes to the development of the field. It has been shown that interactions occur between prokaryotic DNA repair pathways and certain CRISPR-Cas systems[310], further computational analysis suggesting that there might exist incompatibility between some DNA repair mechanisms and certain biochemical properties of CRISPR immunity[311]. Additionally, it is not easy to currently distinguish between the outcomes of repair by NHEJ and AEJ without intensive sequence analysis, which prevents some studies from having clear conclusions about the DNA repair pathways involved in specific experiments. Further study and elucidation of these phenomena in combination with computational tools to predict outcomes of DNA repair will likely facilitate the establishment of better rules and principles for non-templated prokaryotic genome engineering. Despite these caveats, novel combinations of CRISPR-Cas nucleases and DNA processing/repair enzymes are expected to flourish in the near future, providing the field of genome engineering with unprecedented power.

4.8 Acknowledgements

The Netherlands Organization for Scientific Research (NWO) is acknowledged for financial support through a Gravitation grant (BaSyC, 024.003.019) to M.F.B and J.v.d.O, and a VENI grant (016.Veni.171.047) to R.H.S. E.O. acknowledges the public-private NWO-Green Foundation for sustainable production and supply chains in agriculture and horticulture for their financial support (870.15.130, 2015/05279/ALW).

Chapter 5. Elucidating non-homologous and alternative end joining in *Rhodobacter sphaeroides*

Max Finger-Bou[1], Enrico Orsi[2], Ioannis Mougiakos[1], Gijs Bastian[1], Leonardo Morini[1], Victor Pool[1], Alberto de Maria[1], Mohammad Rifqi Ghiffary[1], Ruud A. Weusthuis[2], Richard van Kranenburg[1,3], Nico J. Claassens[1], Raymond H.J. Staals[1], John van der Oost[1]*

[1]Laboratory of Microbiology, Wageningen University and Research, Wageningen, the Netherlands

[2]Bioprocess Engineering, Wageningen University and Research, Wageningen, the Netherlands

[3]Corbion N.V., Gorinchem, the Netherlands

5.1 Abstract

We recently described efficient homologous recombination-based genome editing of *Rhodobacter sphaeroides* by using the CRISPR-associated Cas9 nuclease. In the present study, we describe a method that allows using the CRISPR-Cas9 tool for non-templated genome editing in *R. sphaeroides*. During these efforts, we have revealed molecular details on the non-homologous end joining (NHEJ) as well as the homology-directed repair (HDR) DNA repair pathways in the bacterium. We reveal the involvement in mutagenic repair of an uncharacterized gene (*RSP_2678*), located directly downstream of *ligD*. Furthermore, we conclude that RecA is responsible for the high-fidelity repair of Cas9-induced DNA breaks in the bacterium, and that it competes with several NHEJ proteins. Moreover, evidence is presented for the involvement of alternative end joining in *R. sphaeroides*. Lastly, our results indicate that 5-fluorouracil, used in this and other studies to facilitate counter-selection of the *upp* gene, encoding the uracil-phosphoribosyl-transferase, has mutagenic properties that trigger DNA repair and thus complicates the analysis of mutagenesis studies.

5.2 Introduction

The addition of reprogrammable DNA endonucleases to the toolbox for genetic engineering has tremendously stimulated the field of biotechnology. The early notable developments were based on zinc finger nucleases (ZFNs)[1] and transcription activator-like effector nucleases (TALENs)[2], as for the first time they enabled biologists to induce DNA double-stranded breaks (DSBs) at specific loci of interest in genomes of organisms across all domains of life. More recently, clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR associated (Cas) proteins, which encode for prokaryotic adaptive immune systems[3–6], were discovered and soon repurposed as an extremely versatile genome editing tool[7–10].

As they are considerably easier to reprogram than ZFNs or TALENs, CRISPR-Cas tools have become the preferred programmable endonuclease for genome editing[11,12]. These are generally employed to generate DNA double-strand breaks (DSBs), which have been described to be lethal in prokaryotes when left unrepaired[13,14] and induce the recruitment of DNA repair pathways to prevent cell death. Three distinct DNA repair pathways have been described in prokaryotes: homology-directed repair (HDR), non-homologous end joining (NHEJ) and alternative end joining (AEJ)[15].

HDR is accepted to be the main mechanism by which most bacteria repair DSBs[16,14]. HDR mediates the high-fidelity repair of DSBs in a bacterial genome and requires the presence of an intact second copy of the genome, which generally is available during the exponential phase of growth and in polyploid prokaryotes[17]. During HDR, DSBs are recognized and processed by different enzymes such as the RecBCD complex to generate DNA ends with a 3' extension onto which multiple units of the RecA recombinase protein are loaded, facilitating strand invasion into the intact template DNA molecule, after which DNA is repaired, restoring the initial sequence[18].

In contrast, when an extra copy of the chromosome is not available, as is the case for example during spore germination[19] or during the stationary phase of growth[20], prokaryotes seem to rely mainly on two different template-independent repair pathways to fix DNA DSBs: non-homologous end joining (NHEJ) and/or the more recently described alternative end-joining (AEJ)[15]. Prokaryotic NHEJ pathways typically consist of two proteins: Ku and LigD[21,22]. Ku proteins assemble in homo- or heterodimers[23] forming a ring-like structure which recognizes and binds DNA ends, preventing their degradation by cellular exonucleases[24]. The other component of NHEJ is usually LigD, a multi-functional complex enzyme containing ATP-dependent DNA ligase, phosphatase and polymerase domains[25], the architecture of which differs across species[22], and is responsible for the processing and ligation of DNA ends[26]. Upon recognition of a DSB, Ku dimers bind to the DNA ends and recruit LigD, mediating error-prone DSB repair[27], which frequently results in small insertions and deletions (indels), as well as point mutations at the repair site[15,21,22].

On the other hand, in the currently accepted model for AEJ, also termed microhomology-mediated end joining (MMEJ), helicase-nucleases such as the RecBCD complex recognize and process DSBs until microhomologies (homologies of 1-9 nucleotides as initially described) are revealed in the generated 3' ends, allowing DNA ends to hybridize and be ligated by the NAD^+-dependent DNA ligase LigA[28]. Because of the processing carried out by the RecBCD exonuclease to find microhomologies, AEJ is accompanied by bidirectional deletions flanked by microhomologies. Given that the main functions of LigA and RecBCD take place in DNA replication[29] and HDR[30], respectively, rather than a concrete, specific DNA repair pathway, AEJ seems to be the result of a stochastic interaction of two otherwise orthogonal cellular processes. Because these two end-joining pathways do not require an exogenous DNA template to mediate their mutagenic repair, they can be useful for disrupting genes in microorganisms with limited genetic toolboxes and in bacteria with inefficient homologous recombination machineries[28,31,32].

Rhodobacter sphaeroides is a gram negative, purple non-sulfur bacterium belonging to α-proteobacteria that has gained biotechnological attention during recent years due to its extremely versatile metabolism[33]. Largely studied for photoheterotrophic hydrogen production[34,35] and chemoheterotrophic biosynthesis of terpenes[36–39], the bacterium has also become a model for the study of photosynthesis[40–43], chemotaxis[44–46] and regulation of different types of stress[47–51]. Two CRISPR tools have recently been added to the genome editing toolbox of *R. sphaeroides*: a homologous recombination-based tool[52] and a base editor[53]. Additionally, a tool based on CRISPR-Cas12a has recently been developed for *R. capsulatus*, a close relative of *R. sphaeroides*, enabling the template-free multiplex genome editing of two genes and also the downregulation of two reporter genes through CRISPR interference[54].

A gene frequently used to facilitate screening of mutants in *R. sphaeroides* is the *upp* gene[52], encoding the uracil-phosphoribosyl-transferase (UPRTase) enzyme (EC 2.4.2.9). Belonging to the pyrimidine biosynthesis pathway, it mediates the conversion of uracil into uridine monophosphate (UMP). The UPRTase enzyme can also convert 5-fluorouracil (5-FU) into 5-fluoro-UMP, which is metabolized into 5-fluoro-dUMP and causes the irreversible inhibition of the thymidylate synthase, inducing bacterial death by thymine depletion[55]. The counter-selectable nature of the gene makes it a convenient marker to study mutagenesis in several bacterial species and it has been successfully employed in, among others, *Bacillus subtilis*[56], *Pseudomonas putida*[57] and *R. capsulatus*[54].

In one of these studies[52], carried out in our laboratory, *R. sphaeroides* was engineered using CRISPR-Cas9, which appears to trigger HDR and facilitate the recombination of a DNA repair template into the genome of the bacterium. The provided repair template included a truncated copy of the *upp* gene, enabling the selection of the recombinants in 5-FU. Interestingly, throughout the completion of the study, a small number of clones were found to overcome 5-FU selection without the presence of the aforementioned DNA repair template, hinting at the presence of a functional end-joining repair system independent of HDR. In this study, we aimed at developing a template-independent Cas9-based genome editing tool and to elucidate the DNA repair pathways of *R. sphaeroides*.

5.3 Results and discussion

5.3.1 Low-frequency mutation of the *upp* gene confers *R. sphaeroides* with 5-FU resistance

A popular approach to characterize end joining DNA repair pathways in bacteria involves the transformation of linearized plasmids with different ends[23,26,28,58]. Currently, however, the only available method for introducing plasmids in *R. sphaeroides* is biparental conjugation, through which the transformation of linear plasmids with different DNA ends is not possible. Given the limited genetic accessibility of this bacterium, we chose to study the deactivation of its *upp* gene (*RSP_1598*) (Figure 1A) following the conjugation of a plasmid carrying a *upp*-targeting CRISPR-Cas9 system. We modified a plasmid employed in our previous work[52] to constitutively express a harmonized version of the CRISPR-associated SpCas9 nuclease and a single-guide RNA (sgRNA) targeting the *upp* gene of *R. sphaeroides* (pBBR_Cas9_*upp*) in absence of a repair template, using *E. coli* S17-1 to conjugate it into *R. sphaeroides* (Figure 1B).

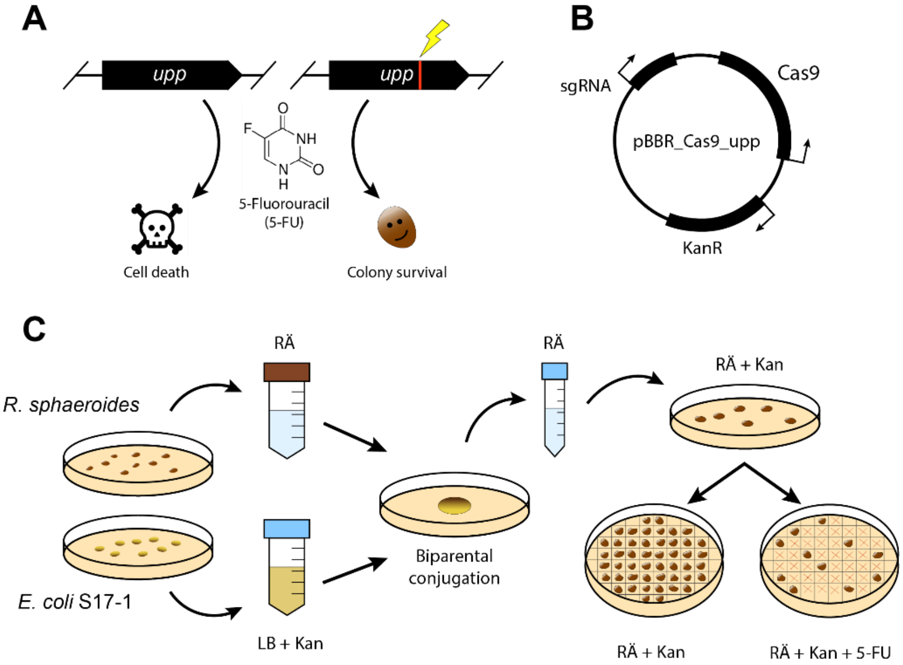


Figure 1: Overall experimental design of the study. *(A) Our selection scheme was based on the resistance against 5-fluorouracil obtained by mutation of the upp (RSP_1598) gene. (B) Our plasmid design of pBBR_Cas9_upp, including a constitutively expressed Cas9 protein and a upp-targeting single-guide RNA (sgRNA), as well as a kanamycin resistance gene (KanR). (C) In our experiments, R. sphaeroides was conjugated with pBBR_Cas9_upp using E. coli S17-1 as a donor strain. A mix of the cells was incubated for several hours, after which the cells were resuspended, plated onto RÄ supplemented with kanamycin and incubated for 48 hours. The resulting colonies were restreaked onto RÄ minimal media with and without 5-FU to score the percentage of colonies resistant to 5-FU.*

To investigate whether Cas9-induced DSBs mediated the disruption of the *upp* gene through non-templated DNA repair, *R. sphaeroides* transconjugants were first isolated by plating onto selective RÄ-Kan50, from which single colonies were picked and re-streaked in parallel onto agar RÄ-Kan50 plates with and without 5FU (Figure 1C).

Corroborating our previous results[52], conjugation of the *upp*-targeting plasmid resulted in a drop in conjugation efficiency of more than 3 orders of magnitude compared to that of the non-targeting (NT) control plasmid (RÄ-

Kan50 plates; Figure 2A), reflecting efficient Cas9-mediated genome targeting.

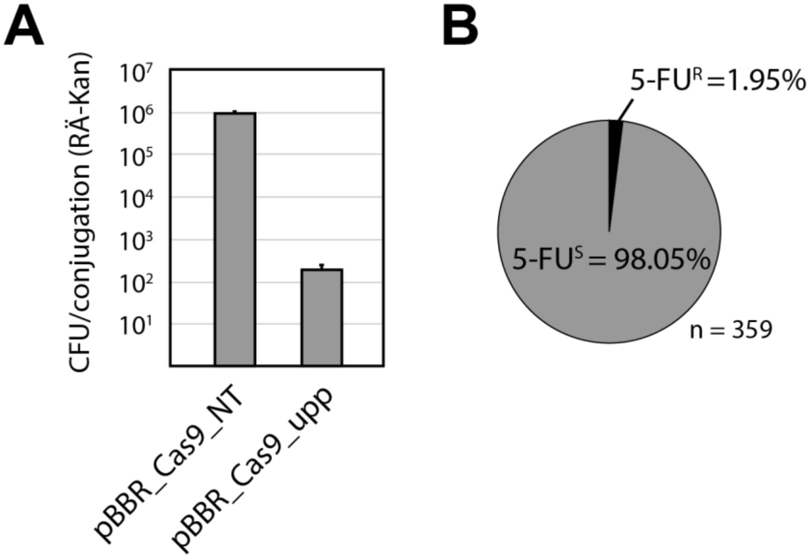


Figure 2: *(A) CFUs scored per conjugation obtained in RÄ minimal medium supplemented with kanamycin after conjugation of R. sphaeroides with the negative control non-targeting plasmid (pBBR_Cas9_NT) compared to that of the upp-targeting plasmid (pBBR_Cas9_upp). (B) Percentage of transconjugant colonies resistant and sensitive to 5-FU (5-FU^R and 5-FU^S, respectively) obtained after restreaking the first transconjugants.*

Strikingly, upon re-streaking several kanamycin-resistant (Kan^R) colonies conjugated with the *upp*-targeting Cas9 system, only approximately 2% were resistant to 5-FU (5-FU^R) (Figure 2B), suggesting that the majority still carried a functional *upp* gene despite the constitutive expression of both Cas9 and the *upp*-targeting gRNA. Colonies carrying a non-targeting gRNA, however, were never able to grow on 5-FU within 48 hours after being restreaked, indicating that Cas9-mediated DSBs were the main drivers of mutagenesis of the *upp* locus during the first 48 hours of incubation.

5.3.2 *Rhodobacter spharoides* is able to survive Cas9 targeting

We first sought to understand why cells of Kan^R colonies were able to withstand the targeting of their *upp* gene while being sensitive to 5-FU (5-FU^S). Sequencing of the *upp* gene of a handful of these colonies revealed that neither the *upp* ORF nor its promoter had mutations in the analyzed clones, which led us to question whether the plasmid-encoded editing system could have been mutated spontaneously, potentially releasing the clones from Cas9-mediated selective pressure. Sequencing of the editing plasmids of these colonies revealed that, in fact, none of these carried mutations, indicating that many *R. sphaeroides* colonies were somehow able to cope with Cas9-mediated genome targeting without the target sequence (protospacer) in their *upp* gene becoming disrupted.

It has been demonstrated that DSBs mediated by relatively weak spacers (i.e. with low targeting efficiency) can be repaired by the RecA-dependent HDR pathway[14]. We thus hypothesized that targeting mediated by our spacer was not efficient enough to outcompete the native high-fidelity repair driven by HDR. Most colonies growing on kanamycin might therefore have been able to grow under Cas9-targeting conditions due to HDR, but died when plated onto 5-FU because, upon disruption by Cas9, their *upp* gene could be continuously repaired with fidelity, thereby maintaining the clone’s sensitivity to 5-FU.

To pinpoint whether the low frequency of mutagenesis in our experiments had a genetic basis, we generated several mutant strains, which lacked one or combinations of different genes presumed to participate in the different repair pathways present in the bacterium. As such, *recA*-deficient (*RSP_0452*) strains were generated to assess the potential role of HDR, and *ku*-deficient (*RSP_0523* and/or *RSP_0524*) and *ligD*-deficient (*RSP_2679*) strains were created to assess the participation of NHEJ. Lastly, we also generated strains lacking an ORF directly downstream of LigD (*RSP_2678*), likely in the same operon, which to our knowledge has neither been described nor characterized.

RecA is the main responsible for survival in Cas9-targeting conditions, and it competes with NHEJ

Colony counts were obtained in plasmid-selecting media to compare the conjugation efficiencies of the generated mutant strains following conjugation of the *upp*-targeting plasmid (Figure 3). The only statistically significant difference was that of the Δ*recA* strain, with a conjugation efficiency that dropped to 25% of the wild type's (Figure 3A) (unpaired t-test, p-value = 0.0208) (Supplementary Data), suggesting that RecA was, as suspected, the main enabler of *R. sphaeroides*' growth under Cas9-targeting conditions. Interestingly, deletion of NHEJ genes onto strains lacking *recA* partially restored conjugation efficiencies back to those of the wild-type strain (Figure 3C). These results suggest an interesting epistatic interaction between HDR and NHEJ in *R. sphaeroides*.

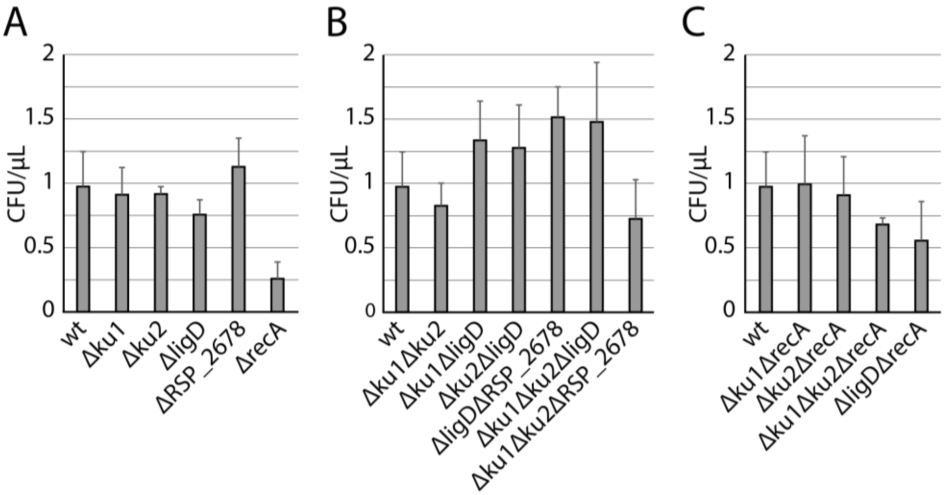


Figure 3: Conjugation efficiencies stated as CFU per µL of final conjugation mix resulting from conjugation a plasmid targeting the* upp *gene. *CFUs per µL of the different generated mutant strains are compared that of the wild-type strain in RÄ-Kan50. (A) Single mutants, (B) double and triple NHEJ mutants, (C) double and triple mutants lacking* recA *and NHEJ genes.*

As NHEJ is known to be an error-prone DNA repair pathway, the deletion of NHEJ genes was expected to have a negative effect on the conjugation efficiency of *R. sphaeroides* in this experiment, since Cas9-mediated DNA DSBs repaired through NHEJ should result in a modified protospacer, facilitating the escape from Cas9-targeting. Surprisingly, the individual deletion of NHEJ-associated genes did not seem to affect conjugation efficiencies significantly (Figure 3B). Strikingly, however, strains lacking a combination of *ligD* and any of the *ku* genes or the *RSP_2678* gene appeared to display slightly higher conjugation efficiencies compared to the wild-type strain. Specifically, the double knockout strain *ΔligDΔRSP_2678* had an approximately 50% higher conjugation efficiency than the wild-type strain with an almost significant statistical difference (unpaired t-test, p-value = 0.072). Despite the weak statistical strength of these results, it is tempting to speculate and hypothesize the presence of a NHEJ complex that requires *ligD* and either *ku1*, *ku2* or *RSP_2678*.

Altogether, the conjugation efficiencies in this experiment suggest that HDR has a bigger effect than NHEJ in repair of Cas9-induced DNA breaks in the wild-type strain. When only RecA is missing, HDR cannot take place as efficiently as in the wild-type strain and the NHEJ complex takes over, which is not as efficient in DNA repair as HDR, resulting in a drop of conjugation efficiency. However, when both RecA and any protein involved in NHEJ are missing, conjugation efficiencies are restored to levels comparable to those of the wild-type strain, hinting at the possibility of another DNA repair pathway being present and taking over when both HDR and NHEJ are impaired.

Table 1: Analysis of mutations in the different studied strains.

		Small mutations						MH-flanked mutations				
		In protospacer			Outside of protospacer			Spacer KO		Spacer intact		
	N	Ins	Del	Pt mut	Ins	Del	Pt mut	Ins	Del	Ins	Del	upp intact
wt	23	4	4		1	4	5				5	
ku1	13		1		1						4	7
ku2	13		1			2			1		4	5
ligD	16	3	2	1	2	1			1		4	2
rsp2678	6											6
recA	8	2	1			2					3	
ku12	19	3			3	4	3		2		4	
ku1ligD	7			1					1			5
ku2ligD	7	1	1						1			4
ku12ligD	8		1				1					6
rsp2678ku12	6	3										3
rsp2678ligD	8	1										7
recAku1	8	1			1		1				5	
recAku2	8		2		2	1	1				2	
recAku12	7		2			1			1	1	2	
recAligD	7	1					2			1	3	

Table 2: Percentages of mutations in the different studied strains. Data from 164 colonies is included in this dataset.

	Small mutations		MH-flanked mutations		
	In protospacer	Outside of protospacer	spacerKO	spacerNoKO	intact upp
wt	34.8	43.5	0.0	21.7	0.0
ku1	7.7	7.7	0.0	30.8	53.8
ku2	7.7	15.4	7.7	30.8	38.5
ligD	37.5	18.8	6.3	25.0	12.5
SB	0.0	0.0	0.0	0.0	100.0
recA	37.5	25.0	0.0	37.5	0.0
ku12	15.8	52.6	10.5	21.1	0.0
ku1ligD	14.3	0.0	14.3	0.0	71.4
ku2ligD	28.6	0.0	14.3	0.0	57.1
ku12ligD	12.5	12.5	0.0	0.0	75.0
SBku12	50.0	0.0	0.0	0.0	50.0
SBligD	12.5	0.0	0.0	0.0	87.5
recAku1	12.5	25.0	0.0	62.5	0.0
recAku2	25.0	50.0	0.0	25.0	0.0
recAku12	28.6	14.3	14.3	42.9	0.0
recAligD	14.3	28.6	0.0	57.1	0.0

5.3.3 RecA-mediated HDR and interactions with other NHEJ proteins

Sequencing of several of the generated mutant strains growing in 5-FU revealed colonies with an entirely intact *upp* gene, including its promotor sequence. The mechanisms enabling resistance to 5-FU with an intact *upp* gene are discussed later.

Contradicting our expectations, upon sequencing, none of the wild-type *R. sphaeroides* colonies were found to carry an intact *upp* gene (Table 1 and

Table 2), suggesting that, while the deletion of *recA* had a negative impact on conjugation efficiency (Figure 3C), it did not seem able to mediate the high-fidelity repair of the *upp* gene on a wild-type background. Remarkably, however, the individual deletion of *ku1*, *ku2*, *ligD* or *RSP_2678* increased the number of transconjugant colonies found to carry an intact *upp* gene, further suggesting interactions between the NHEJ proteins and the mechanisms enabling high-fidelity repair of the *upp* gene. The stacking of deletions of NHEJ genes rendered, almost in all cases, higher numbers of transconjugant colonies with an intact *upp* gene, the only exception being the double knockout strain *Δku1Δku2*, which strikingly was never found to carry an intact *upp* gene.

While most of the generated knockout strains were found to carry an intact *upp* gene with varying incidence, none of the colonies for any of the strains lacking *recA* were found to carry an intact *upp* gene, indicating that *recA* was, at least partially, responsible for the high-fidelity repair of the *upp* gene.

These results suggest interesting interactions between HDR and NHEJ; while *recA* seemed to be mainly responsible for the lack of mutations in the *upp* gene, it seemed to necessitate the deletion of NHEJ proteins to mediate high-fidelity repair in the wild-type genetic background, except in *Δku1Δku2*. Surprisingly, while several colonies of the single mutants *Δku1* and *Δku2* were found to carry an intact *upp* gene (53.8% and 38.5%, respectively), all the analyzed *Δku1Δku2* colonies were found to carry mutations in the gene.

5.3.4 Inactivation of the *upp* gene by NHEJ-like mutations

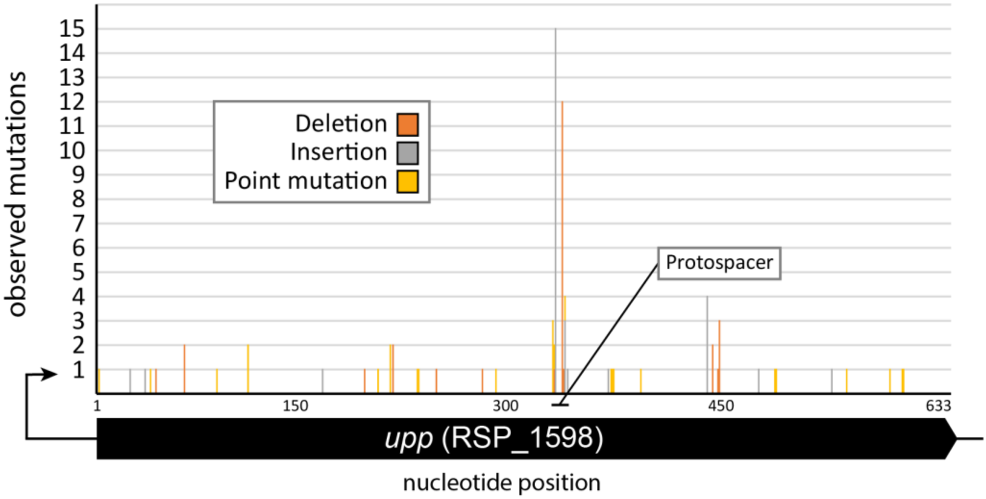


Figure 4: Distribution of NHEJ-like mutations throughout the upp gene.

While most NHEJ-like mutations (small indels and point mutations) were found in the targeted protospacer of many of the sequenced strains, these mutations were also found throughout the *upp* gene of several colonies without disrupting the protospacer (Figure 4, Table 1 and Table 2). As Cas9-mediated DNA DSBs is highly unlikely to trigger mutations outside of the protospacer, the ubiquitous distribution of mutations calls for caution, as it indicates an unforeseen cause of mutagenesis in our experiments. Provided the presence of another mutagen, we cannot rule out that at least part of the mutations observed in the protospacer were not caused by Cas9-mediated targeting. Given that our CRISPR system mediated a drop in conjugation efficiency of three orders of magnitude, the enrichment of mutations in the protospacer might be a product of counter-selection.

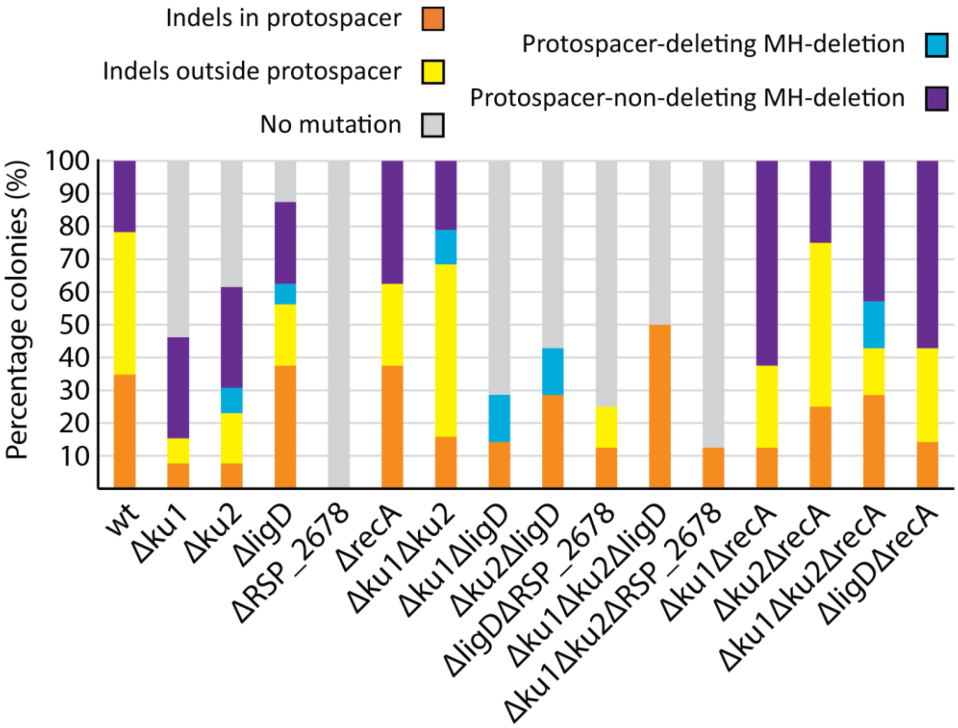


Figure 5: Summary of the different types of mutations found after sequencing the **upp** ***gene of 5-FU survivors of all the different studied strains.***

While 35% (8/23) wild-type Kan^R, 5-FU^R colonies of *R. sphaeroides* carried small insertions or deletions in the protospacer characteristic of NHEJ, only 8% (1/13) *Δku1*, 8% (1/13) *Δku2* and 16% (3/20) of *Δku1Δku2* colonies carried this type of mutations (Table 1 and Table 2), suggesting at least a partial involvement of the genes in the DNA repair pathway. Remarkably, 38% (6/16) of *ΔligD* colonies carried this type of mutation in the protospacer, indicating that the multifunctional ligase is not involved in the observed mutagenesis. Intriguingly, while only 8% (1/13) of the single mutants *Δku1* and *Δku2* carried NHEJ-like mutations, additional deletion of *ligD* seemed to partially restore the efficiency of NHEJ to 14% (1/7) for *Δku1ΔligD* and 29%

(2/7) for *Δku2ΔligD*, illustrating complex interactions between the proteins involved in DNA repair in *R. sphaeroides*.

The single deletion of the uncharacterized ORF directly downstream of LigD, *RSP_2678*, resulted in colonies bearing no mutations whatsoever in their *upp* gene (Table 1 and Table 2). Mutants lacking *RSP_2678* and *ku1ku2* or *ligD* were also found to carry high rates of an intact *upp* gene (50% and 87.5%, respectively). Altogether, these results suggest that *RSP_2678* might be involved in mutagenic DNA repair, although the exact mechanism governing its role can only be clarified through further research.

Additionally, the fact that NHEJ-like mutations are still found in absence of *ligD*, which is assumed to be the enzyme in charge of processing and ligation of DNA ends, suggests either that there might have been another ligase active in the tested conditions, or the observed mutations were not caused by Cas9-triggered NHEJ repair.

5.3.5 Inactivation of the *upp* gene by microhomology-flanked deletions

Analysis of the amplified *upp* loci of several Kan^R, 5-FU^R colonies revealed deletions ranging from 11 to 1511 bp in the *upp* gene (Table 1 and Table 2, Supplementary Table 1). However, several of the sequenced colonies revealed microhomology-flanked deletions which were not disrupting the protospacer (Figure 6A). The deletions were in all cases flanked by 2-15 bp of microhomology (Figure 6B, Supplementary Table 1).

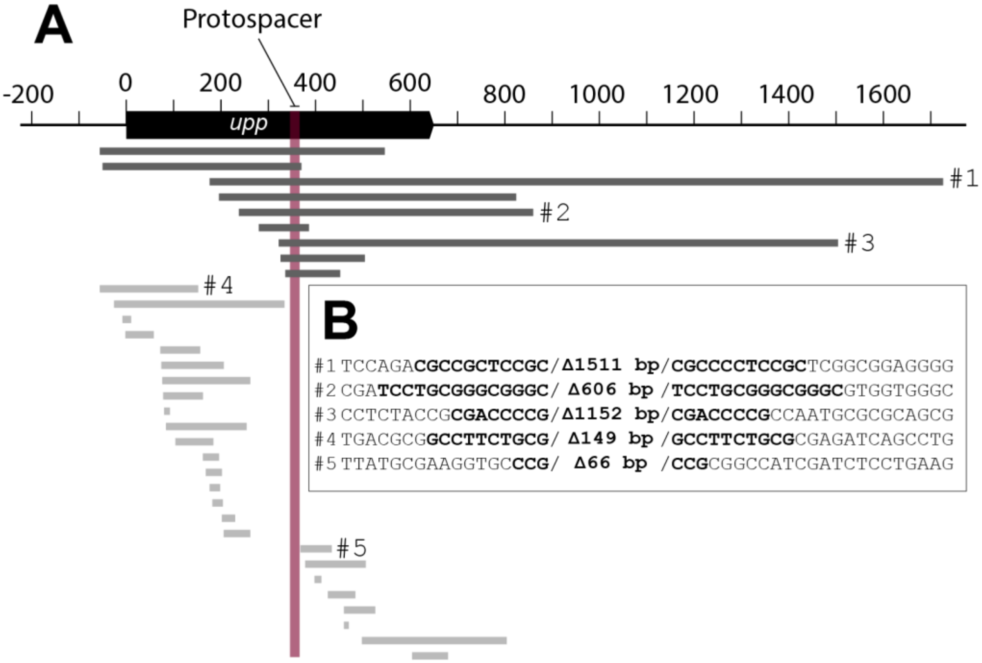

A
Protospacer
-200
0
200
400
600
800
1000
1200
1400
1600
upp
#1
#2
#3
#4
#5
B
#1 TCCAGA**CGCCGCTCCGC**/**Δ1511 bp**/**CGCCCCTCCGC**TCGGCGGAGGGG
#2 CGA**TCCTGCGGGCGGGC**/ **Δ606 bp** /**TCCTGCGGGCGGGC**GTGGTGGGC
#3 CCTCTACCG**CGACCCCG**/**Δ1152 bp**/**CGACCCCG**CCAATGCGCGCAGCG
#4 TGACGCG**GCCTTCTGCG**/ **Δ149 bp** /**GCCTTCTGCG**CGAGATCAGCCTG
#5 TTATGCGAAGGTGC**CCG**/ **Δ66 bp** /**CCG**CGGCCATCGATCTCCTGAAG

***Figure 6**: (A) Distribution of microhomology-flanked deletions throughout of the upp gene, including data from several of the studied strains. (B) A selection of some of the deletions, with the microhomologies flanking the deletions highlighted in bold. Note that in every deletion one of the two sequences with microhomology is deleted.*

These microhomology-flanked deletions were found in all knockout strains except in strains missing *RSP_2678* and in the triple knockout strain *Δku1Δku2ΔligD* (Table 1 and Table 2, Supplementary Table 1). Despite the small number of colonies analyzed, it is tempting to speculate a role of both *RSP_2678* and the NHEJ complex Ku1Ku2LigD in the generation of microhomology-flanked deletions.

MH-flanked deletions revealed a correlation between microhomology GC content and deletion length

To elucidate whether microhomology-flanked deletions occur with any observable pattern, regression analysis between deletion length, microhomology GC content and microhomology length was performed using data from all the studied strains.

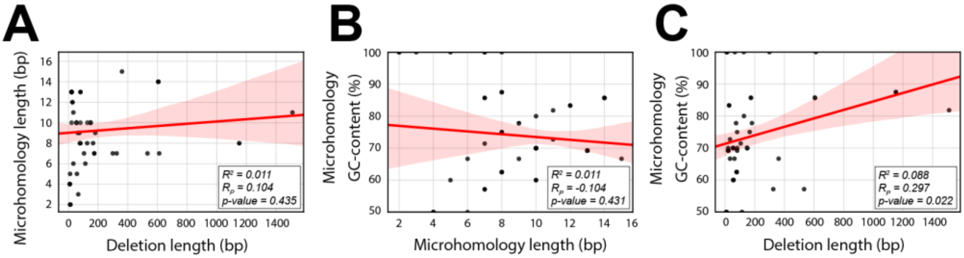


Figure 7: Analysis performed to uncover potential correlations between microhomology length, microhomology GC-content and deletion length.

While no correlation was found to exist between microhomology length and deletion length or microhomology GC-content, the analysis revealed a weak correlation between deletion length and the microhomology GC-content (Pearson correlation coefficient = 0.297; p-value = 0.022) (Figure 7C), suggesting that the stable interaction between sequences with microhomologies is crucial for the joining of DNA regions distant to the DNA DSB site. Notably, several of the analyzed deletions occurred exactly at the same spot and were flanked by the same regions of microhomology in different colonies (Supplementary File 1), suggesting that certain regions of the genome may act as microhomology-mediated repair hotspots.

5.3.6 Microhomology-flanked duplications point at replication mechanisms governing microhomology-flanked mutagenesis

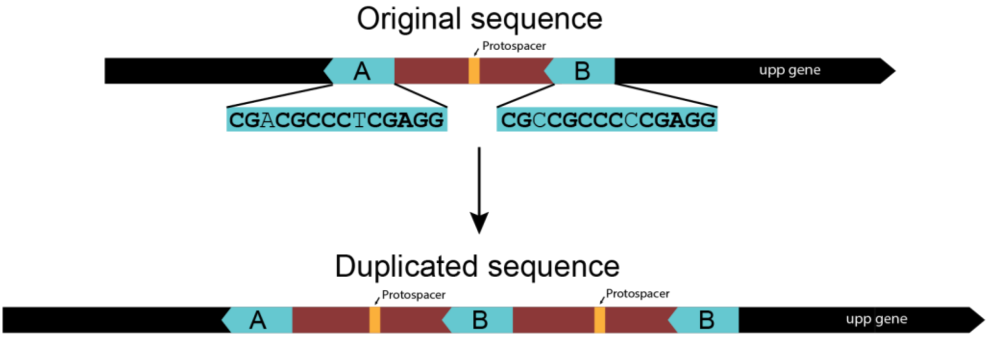


***Figure 8: Duplication of 126 bp in the upp gene in a Δku1Δku2ΔrecA strain**. A sequence flanked by almost perfect direct repeats is duplicated, resulting in a disrupted upp gene while maintaining the integrity of the (now double) protospacer.*

In two cases, rather than small indels or microhomology-flanked deletions, the colony-derived genomes appeared to carry microhomology-flanked duplications. The first, a duplication of 126 bp in *Δku1Δku2ΔrecA*, was seen to be mediated by an almost perfect direct repeat of 14 bp (Figure 7, Supplementary Table 1). The second, of 54 bp and found in *ΔligDΔrecA*, was mediated by a perfect direct repeat of 10 bp.

Such duplications are, to our knowledge, rarely observed, if at all, in Cas9-targeting experiments, and have never been reported to be a product of neither NHEJ nor AEJ in prokaryotes. Despite our system mediating a drop in conjugation efficiency of three orders of magnitude, we observed several colonies withstand the targeting of Cas9 without the *upp* protospacer being disrupted. A duplication event, as shown above, resulted in DNA sequences that included not one, but two intact protospacers mediating negative selective pressure. We therefore deem such an event to be extremely rare, as shown by the very limited occurrence thereof in our dataset (only two cases out of several hundreds of colonies sequenced from the start of this project). We hypothesize that such an event might be mediated by replication slippage[59], likely triggered after replication fork stalling[60], as discussed in more detail later.

5.3.7 5-FU-resistant *upp* mutants can arise independently of Cas9-targeting

Several of the NHEJ-like and microhomology-flanked deletions occurred without affecting the *upp* protospacer, suggesting that Cas9-targeting might not have been involved in the generation of these mutations. We thus hypothesized that mutation of the *upp* locus might have occurred after the colonies were restreaked onto RÄ-5-FU, rather than being strictly induced by Cas9 on the RÄ-Kan50 plate. To confirm whether another source of mutagenesis was present in our experiments, we repeated the experiment including a non-targeting guide RNA. Additionally, we sought out to study the chronological occurrence of mutations, sequencing colonies in Kan50 before and after being restreaked onto plates with 5-FU (Table 3).

As mentioned earlier, upon restreaking from RÄ+Kan50 onto RÄ+Kan50+5-FU, none of the colonies carrying a NT gRNA was able to grow within 48 hours of incubation, whereas a small fraction of the colonies carrying a targeting sgRNA was quickly able to grow in RÄ+Kan50+5-FU. However, for this experiment, *R. sphaeroides* was incubated for between 72 and 96 hours, and small colonies were observed to appear in the presence of 5-FU when the strains carried a non-targeting plasmid, suggesting that mutagenesis of the *upp* gene occurred independently of Cas9-mediated targeting in media with 5-FU. An analyzed colony carrying the non-targeting plasmid and with an intact *upp* gene in RÄ+Kan50 was seen to generate two small colonies with different large deletions flanked by microhomologies (RÄ+Kan50+5-FU), further supporting that mutagenesis of the *upp* gene occurred spontaneously in media with 5-FU.

Table 3: Colonies carrying non-targeting (NT) or upp-targeting plasmids were plated first on Kan50 and then restreaked onto Kan50+5-FU*. Both the parental colony (in Kan50) and two daughter colonies (in Kan50+5-FU) were sequenced, and the resulting sequences are shown in the table below.*

Strain	Spacer	Plate	Mutation	MH length	Del length	Del position	Spacer state
wt	NT	Kan50	No mutation	-	-	-	No spacer
		Kan50+5FU	MH-del	13	174	295	
		Kan50+5FU	MH-del	8	78	102	
ku1ku2	upp	Kan50	No mutation	-	-	-	Intact
		Kan50+5FU	Spont. deletion	-	1	460	Intact
		Kan50+5FU	MH-del	3	66	358	Intact
wt	upp	Kan50	No mutation	-	-	-	Intact
		Kan50+5FU	Spont. deletion	-	1	461	Intact
wt	upp	Kan50	NHEJ (6bp ins)	-	-	-	Disrupted
		Kan50+5FU	NHEJ (6bp ins)	-	-	-	Disrupted
wt	upp	Kan50+5FU	MH-del	13	174	295	Disrupted
wt	upp	Kan50	No mutation	-	-	-	Intact
		Kan50+5FU	MH-del	13	174	295	Disrupted
		Kan50+5FU	MH-del	13	174	295	Disrupted
wt	upp	Kan50	No mutation	-	-	-	Intact
		Kan50+5FU	NHEJ (1bp subs)	-	-	-	Disrupted
		Kan50+5FU	NHEJ (1bp subs)	-	-	-	Disrupted

Upon observing the appearance of spontaneous, slow-growing mutants in media with 5-FU, we also incubated strains carrying targeting plasmids for a longer period of time. While only a small percentage of colonies was initially observed to withstand 5-FU within 48 hours, small colonies also observed to appear when the plates were incubated for up to 96 hours. Sequencing of parental colonies and their late-appearing restreaked descendant colonies revealed different mutations in their *upp* gene despite having originated from the same colony, indicating that mutagenesis in these cases occurred only upon restreaking onto 5-FU-containing plates.

Only one of the parental colonies was observed to share mutations with its daughter colony on 5-FU, reflecting that Cas9-mediated targeting is also able to generate mutations attributable to NHEJ.

Interestingly, one of the microhomology-flanked deletions observed in the non-targeting control was also found in two other, independent colonies growing on 5-FU with a targeting plasmid. This further supports that there are certain hotspots for microhomology-mediated rearrangements. Altogether, the types of observed mutations caused by Cas9 are indistinguishable from those caused by 5-FU. It is therefore likely that 5-FU triggers DNA repair mechanisms that are shared with those triggered by Cas9-mediated DNA breaks.

5.3.8 Survival of colonies with an intact *upp* gene in 5-FU

Because 5-FU has frequently been used as a counter-selective agent to select for mutations in the *upp* gene, we were surprised to find so many colonies able to grow in presence of the compound while not having a mutated *upp* locus. In cells with an intact *upp* gene, the UTMase enzyme will metabolize 5-FU into FdUMP, which will irreversibly inhibit the thymidylate synthase (TS) enzyme[61]. Inhibition of this enzyme, in turn, is thought to block *de novo* production of dTMP[62]. However, mutation of the *pyrR* repressor in *Mycobacterium tuberculosis* has been shown to drive the upregulation of the *pyr* operon, increasing *de novo* production of UMP and thereby rescuing the cells from FdUMP toxicity[63]. 5-FU was also seen in *M. tuberculosis* to modestly induce the expression of *recA* and *radA* genes, suggesting the presence of DNA DSBs, which were presumed to be mediated by action of the uracil DNA glycosilase and AP endonuclease enzymes upon recognition of DNA with incorporated 5-FU.

Via either any of these routes or via an uncharacterized mechanism, we observe that the growth of wild-type *upp* colonies is impaired when streaked onto 5-FU, likely indicating temporary inhibition of the TS enzyme. In such scenario, the thymine pools in the cell are very low, which likely blocks replication. Under these circumstances, any mutation in their *upp* gene would generate a subpopulation of cells whose active TS levels are increased over time after functional enzymes are produced, progressively resulting in decreasing concentrations of FdUMP.

While the system has proven useful in homologous recombination-based genome editing, the possibility of cells to escape 5-FU mediated death through different mechanisms puts in question the reliability of the counter-selection method system, especially in experiments where stringent selection is required.

5.3.9 5-FU-induced mutagenesis

Despite the broad use of 5-FU and other similar compounds for screening and selection in biotechnology and genetic engineering, our results indicate a potential mutagenic role of the compound in our experiments. Given that microhomology-flanked deletions and small indels were found in strains lacking a targeting Cas9 plasmid, we explored the available literature for mechanisms through which 5-FU could be causing these types of mutations in our setup.

In our experiments, deactivation of the thymidylate synthase could drive a low concentration of cellular dTMP, which in turn could lead to replication fork stalling and break-induced replication[60]. During this process, a newly synthesized DNA strand separates from its template and anneals to a sequence with short homology either upstream or downstream of its *bona fide* template, resulting in either the deletion or insertion of repeated sequences, respectively[64]. Additionally, double-stranded DNA breaks can take place at stalled replication forks, which can be processed and ligated generating microhomology-flanked deletions[65], identical to those generated by AEJ.

Alternatively, during replication slippage, a nascent DNA strand can be displaced from its original template and misaligned to a repeated sequence, resulting in deletions or duplications[66]. In a similar subsequent study, RecA was reported to be involved in the deletion of a sequence flanked by a fully homologous 101 bp repeat, in a model called sister chromosome exchange, whereby a replicating strand with two repeats would originate two nascent strands, one left with one repeat and the other left with three[67,68]. The model for replication slippage has been used to explain deletions and insertions of

1 nucleotide[59], which are also common in our sequencing results, as well as in the NHEJ literature.

Additionally, FdUMP, produced when 5-FU is metabolized by the cell, has been reported to be incorporated into genomic DNA, albeit in human cell lines. Upon incorporation of FdUMP into the genome, uracil DNA glycosylases can recognize and cleave it, initiating the base-excision repair (BER) pathway[69], which begins by nicking the DNA, a process reportedly able to mediate long genome deletions flanked by homologies in *E. coli*[70].

5.4 Conclusions

5.4.1 Flawed experimental design to study Cas9-mediated mutations due to 5-FU genotoxicity

The mutations in the *upp* gene of several of the sequenced colonies were distributed between the promoter and the end of the coding sequence, rather than localized only within the protospacer. Our experiments with non-targeting plasmids, as well as our monitoring of the chronological appearance of mutations, suggest that Cas9-mediated genome targeting is not the only element driving mutagenesis in our experimental setup. Having noted that the literature broadly supports the genotoxic nature of 5-FU, and having found identical mutations caused independently by 5-FU and Cas9, we can conclude that the DNA repair pathways triggered by 5-FU and Cas9-targeting converge.

5.4.2 Lessons learned about DNA repair in *R. sphaeroides*

Likely the clearest result we have obtained throughout this study is the finding that no colonies lacking *recA* were observed to carry an intact *upp* protospacer. This indicates that *recA* is needed in HDR to prevent chromosomal mutations and rearrangements. Interestingly, also colonies of the double knockout strain missing both *ku1* and *ku2* were never found to carry an intact *upp* protospacer, just like the wild-type and all strains missing *recA*. While we cannot currently explain the differences in mutagenesis observed in strains missing either *Δku1* or *Δku2* and the double mutant *Δku1Δku2*, it has been shown that Ku proteins can assemble in heterodimers, mediating slightly different functions than that of its homodimers[23]. Further experiments are required to dissect the convoluted interactions between these DNA repair pathways in *R. sphaeroides*. Modifying the experimental design will be crucial for a proper assessment of the effect of individual mutagens.

Furthermore, as we find protospacer-disrupting NHEJ-like mutations in strains missing *ligD*, it is likely that LigD is not strictly needed for this type of repair. Further bioinformatic analysis revealed a gene coding for another multifunctional ligase in the genome of *R. sphaeroides*, annotated as *lig2* in the Uniprot database (*RSP_2413* / Q3J3R0), albeit with a different domain structure than the studied *ligD*. The generation of knockout strains lacking *lig2* in combination with other putative NHEJ proteins will be imperative to finally elucidate the DNA repair pathway in the bacterium.

Interestingly, none of the colonies for the strains *Δku1Δku2ΔligD*, *Δku1Δku2ΔRSP_2678* or *ΔligDΔRSP_2678* were found to have microhomology-flanked mutations. Although the number of successfully sequenced colonies for these strains is low (8, 6 and 8, respectively), these results may suggest that the components of the NHEJ system can form different complexes that have different functions depending on their subunit composition, as has been found in *Sinorhizobium meliloti*[23].

Lastly, strains missing *RSP_2678* were observed to carry an intact *upp* protospacer much more frequently than the other strains, which suggests a potential involvement of the encoded protein in mutagenic DNA repair. The number of colonies that we were able to sequence successfully was rather limited, but these preliminary results indicate that *RSP_2678* is actively involved in DNA repair.

5.4.3 Mechanisms governing microhomology-flanked mutations

The fact that we found two colonies containing microhomology-flanked duplications (that is, duplications flanked by direct repeats) strongly suggests that faulty replication is responsible for this type of mutagenesis. Such events can also result in the deletion of sequences between repeats, which in we observe much more often in our dataset likely due to Cas9-mediated counter-selection.

5.4.4 Further remarks on using 5-FU for genetic engineering

Both 5-FU or 5-FOA are commonly used to counter-select cells with wild-type genotypes for the *upp* and *pyrF* genes, respectively. In such experimental framework, the cause of toxicity often quoted is thymine depletion. Thymine, one of the four bases of DNA, can also be derived through methylation of uracil. In most experiments involving the generation of *pyr* mutants, knockouts are considered uracil auxotrophs and as such they are supplemented with uracil to enable growth.

In the study preceding the present, however, *upp* knockouts of *R. sphaeroides* were not supplemented with uracil[52]. Despite the lack of supplementation, it was not observed that the bacterium had become auxotrophic for uracil. The fact that supplementation was not required for its growth suggests that its metabolism can adapt to overproduce pyrimidine precursors and specifically UMP, which has been proven to confer with resistance against 5-FU[63]. Although we did not characterize the exact mechanism for such resistance in *R. sphaeroides*, our results illustrate that the toxicity mediated by 5-FU can be circumvented, making the compound unreliable for stringent counter-selection.

5.5 Materials and methods

5.5.1 Bacterial strains and growth conditions

Derived from *R. sphaeroides* ATCC® 35053™, *R. sphaeroides* 265-9c, herein referred to as wild-type, was used to generate all the mutant strains employed in this study. All of the *R. sphaeroides* strains were grown on RÄ minimal medium, either liquid or 1.5% w/v agar. RÄ medium contained (per liter): 3 g malic acid, 0.2 g $MgSO_4·7H_2O$, 1.2 g $(NH_4)_2SO_4$, 0.07 g $CaCl_2·2H_2O$, 1.5 mL of microelements stock solution, 2 mL of vitamin stock solution and 5 mL of phosphate buffer. In case of RÄ agar medium, 15 g/L agar was added. The microelements solution contained: 0.5 g/L Fe(II)-Citrate, 0.02 g/L $MnCl_2·4H_2O$, 0.005 g/L $ZnCl_2$, 0.0025 g/L KBr, 0.0025 g/L KI, 0.0023 g/L $CuSO_4·5H_2O$, 0.041 g/L Na_2MoO, 0.005 g/L $CoCl_2·6H_2O$, 0.0005 g/L $SnCl_2·2H_2O$, 0.0006 g/L $BaCl_2–2H_2O$, 0.031 g/L AlCl, 0.41 g/L H_3BO, 0.02 g/L EDTA. The vitamin solution contained: 0.2 g/L nicotinic acid, 0.4 g/L thiamine HCl, 0.008 g/L biotin, 0.2 g/L nicotinamide. The phosphate buffer contained 0.6 g/L KH_2PO_4 and 0.9 g/L K_2HPO_4.

E. coli DH5α was used to clone and amplify all plasmid constructs, while *E. coli* S17-1 was used as the donor strain for biparental conjugation of *R. sphaeroides*. Both *E. coli* strains were grown in LB medium, either liquid or 1.5% w/v agar, and were made chemically competent and transformed by heat-shock treatment following a protocol described elsewhere[71]. LB contained (per liter): 10 g tryptone, 10 g NaCl and 5 g yeast extract.

5.5.2 *R. sphaeroides* knockout strains

All the knockout strains were derived from *R. sphaeroides* 265-9c following the protocol developed in our previous work[52]. Homologous recombination-based editing plasmids were built encoding a *Streptococcus pyogenes* Cas9 nuclease codon-harmonized to match *R. sphaeroides*' codon usage, a single-guide RNA (sgRNA) module and two homology arms of 1 kb each. Several mutant strains were generated during our previous study and their development is described in another document[312]. The homology arms were designed to substitute 150 bp of each gene for three tandem stop codons, followed by either a BamHI or an EcoRI restriction site. To screen for knockout mutants, primers were designed to amplify the locus of interest binding outside of the homology arms, and colony PCRs were carried using Q5 High-Fidelity 2X Master Mix (New England Biolabs) supplemented with 3% v/v DMSO and using *R. sphaeroides* conjugant colonies as a template. The amplified DNA fragments were then purified using the DNA Clean & Concentrator kit (Zymo Research), digested with either BamHI or EcoRI (New England Biolabs) and analyzed by gel electrophoresis to confirm the successful deletion of the desired genes.

5.5.3 Plasmid construction

All plasmids were built through Gibson assembly[72] using the HiFi DNA Assembly Master Mix (New England Biolabs). DNA fragment amplification for plasmid construction was performed using Q5 High-Fidelity 2X Master Mix (New England Biolabs) using either the pBBR_Cas9_NT[52] plasmid or *R. sphaeroides*' genome (supplemented with 3% v/v DMSO) as a template.

After being run by gel electrophoresis, the amplified DNA products were purified using the Zymoclean Gel DNA Recovery Kit (Zymo Research). Once transformed, plasmids were extracted from *E. coli* DH5 α using the GeneJET Plasmid Miniprep Kit (Thermo Fisher).

5.5.4 Biparental conjugation of *R. sphaeroides* and plasmid curing

R. sphaeroides cultures were inoculated from frozen glycerol stocks into 10 mL of RÄ media and grown at 30ºC, 200 rpm. After 48 hours of incubation, 200 µL of grown culture were reinoculated into 10 mL of fresh RÄ medium and incubated in the same conditions for 24 hours. Donor *E. coli* S17-1 strains were inoculated from frozen stocks into LB supplemented with 50 µg/mL kanamycin and incubated overnight at 37ºC, 250 rpm. Cultures of *E. coli* S17-1 were then diluted 30 times in LB liquid media and incubated in the same conditions until they reached an OD600 of 1. Then, 1 mL of culture was washed twice in RÄ medium and mixed with 1 mL of R. sphaeroides culture. The mixture was then spun down (1 minute at maximum speed) and the pellet was resuspended in 100 µL of RÄ. The mixture was then transferred on top of a sterile 0.22 µm, 47 mm diameter nitrocellulose filter agar plate and the conjugation mixture was harvested from the filter by gently scraping it with a sterile loop, resuspending the mixture in 2 mL of RÄ medium. Different dilutions were plated onto RÄ agar plates supplemented with 50 µg/mL kanamycin.

5.5.5 5-FU screening

For selection of *upp* mutants, *R. sphaeroides* colonies resulting from conjugation of pBBR_Cas9 plasmids were picked and restreaked onto RÄ agar plates supplemented with 50 µg/mL kanamycin and 100 µg/mL 5-FU (using a 50 mg/mL 5-FU stock solution prepared in dimethyl sulfoxide). The *upp* locus of surviving colonies was amplified by PCR using the Q5 High-Fidelity 2X Master Mix (New England Biolabs) (with 3% v/v DMSO), purified using the DNA Clean & Concentrator kit (Zymo Research) and analyzed through Sanger sequencing (Macrogen Europe B.V.).

5.5.6 *upp* mutation analysis

All the reads obtained through Sanger sequencing were analyzed computationally using a custom Python script. Microhomologies were annotated as such when they were consisting of two or more nucleotides and gaps of up to one nucleotide were permitted.

5.5.7 Primers

Table 2: Primers used in this study.

BG20060	Rse_upp_4kb_rv	AGGACTCCGTCGACGGC	
BG20061	Rse_upp_4kb_fw	CGACAGCGTGCTTCAGATCG	
BG20273	Rse_Ku1_KOSeq_fw	GTCGAGGCGCAGATGCTGG	
BG20274	Rse_Ku1_KOSeq_rv	GCAAGGCTCCGCCGGAG	
BG20275	Rse_Ku2_KOSeq_fw	CCGAGATCAGGGTGAAGGCCG	
BG20276	Rse_ku2_KOSeq_rv	GCCACACTCTCGCGCAGC	
BG20277	Rse_ligD_KOSeq_fw	CATTCCTGCCAGGGTCGGC	
BG20278	Rse_ligD_KOSeq_rv	CCTCGCCGTCGATCAGCG	
BG20989	Rse_recA_Seq_Fw	GGGATGGACGGACCGACATG	
BG20990	Rse_recA_Seq_Rv	GCGTCCGACTTGAAATAGTCGCCG	
BG20488	Rse_smallB_Seq_Fw	GTGCGGAAGAAGGGGACGAAGAC	
BG20489	Rse_smallB_Seq_Rv	GTTTCCGGTCAGACGCCCAC	
BG20462	Rsp265_KO_A_ori_Fw	aggcggtttgcgtattgggc	
BG20463	Rsp265_KO_A_ori_Rv	gcctgaatggcgaatggaaattgtaagcg	
BG20464	Rsp265_KO_B_HA_smallboi_Fw	acgcttacaatttccattcgccattcaggcAGACGCCCCCTCACCCG	
BG20465	Rsp265_KO_B_HA_smallboi_Rv	GCGGATCCTCATCATCATTTCGTCATGGTTTCCCTCCTGCC	
BG20466	Rsp265_KO_C_HA_smallboi_Fw	GAAACCATGACGAAATGATGATGAGGATCCGCACCGGGCATCGGC	
BG20467	Rsp265_KO_C_HA_smallboi_Rv	cagctggcgtaatagcgaagaggcccgcacATCGCCCTGCCCATGATCAGC	
BG20468	Rsp265_KO_D_sgRNA_Fw	gtgcgggcctcttcgctattacg	
BG20469	Rsp265_KO_D_sgRNA_Rv-RSP2678_1	tgctatttctagctctaaaacTGCGCTGCGCGATCAGCCGCaaccagcgatcccgtccgc	
BG20470	Rsp265_KO_D_sgRNA_Rv- RSP2678_2	tgctatttctagctctaaaacGGTGGCGCTGGCAGGCAGCGaaccagcgatcccgtccgc	
BG20471	Rsp265_KO_D_sgRNA_Rv- RSP2678_3	tgctatttctagctctaaaacGGGTCCTCCTGATGGTCCTGaaccagcgatcccgtccgc	
BG20472	Rsp265_KO_E_Cas9	gttttagagctagaaatagcaagttaaaataaggctagtccgttatcaac	
BG20473	Rsp265_KO_E_Cas9	ttatgcatgcgcccaatacgcaaacc	

Several knockout strains were constructed using plasmids previously built during the completion of another PhD thesis (ISBN: 9789463434096, 9463434097).

5.6 Supplementary data

Sequencing data can be downloaded at the following link:

https://data.4tu.nl/private_datasets/ThZrIIeG7msJHS6mRx1m4kgcVT8jZ--nspyKTHBGn5M

5.7 Acknowledgements

The Netherlands Organization for Scientific Research (NWO) is acknowledged for financial support through a Gravitation grant (BaSyC, 024.003.019) to M.F.-B, N.J.C. and J.v.d.O and a VIDI grant (VI.Vidi.203.074) to R.H.J.S. In addition, N.J.C. acknowledges support from his NWO Veni fellowship (VI.Veni.192.156). E.O. and R.A.W. were supported by the public-private NWO-Green Foundation for sustainable production and supply chains in agriculture and horticulture for their financial support (870.15.130, 2015/05279/ALW). I.M. acknowledged support from the Dutch

Technology Foundation STW, part of the NOW, and R.v.K. is employed by Corbion.

Chapter 6. Exploring iterative genome minimization of *Escherichia coli* with a type I-C CRISPR-Cas tool

Max Finger-Bou[1], Laura Pol[1], Gijs Bastian[1], Nico J. Claassens[1], John van der Oost[1], Raymond H.J. Staals[1]

[1]Laboratory of Microbiology, Wageningen University and Research, Wageningen, the Netherlands

6.1 Abstract

Genome minimization aims at creating a simplified version of an organism's genome while maintaining its biological functions or a specific part thereof. Apart from gaining fundamental insights on genome organization and gene functionality, a minimal genome may result in more efficient microbial factories for biotechnology. *Escherichia coli* is an attractive model for genome minimization efforts due to its well-characterized genome and the plethora of tools available for its genetic engineering. Making use of such a toolbox, several studies have reported the successful deletion of parts of the *E. coli* genome, resulting in improved growth rate/yield as well as in enhanced formation of desired products. The development of novel class 1 CRISPR-Cas genome editing tools presents a great opportunity to further advance the field, as it facilitates the generation of very large deletions. In this study, we designed and assessed a novel genome minimization method based on bacterial conjugation. We attempted to generate consecutive deletions in the genome of *E. coli* utilizing a type I-C CRISPR-Cas system. Although we were not able to stack deletions, we demonstrated that the method successfully delivered single deletions of up to 110 kb.

6.2 Introduction

"In an environment that is free from stress and provides all necessary nutrients, what would constitute the simplest free-living organism?", wrote John Glass in one of the several research articles devoted to discussing and expanding our knowledge on engineering a cell with a minimal genome[313–315]. Having become an established field of research in the last 20 years, genome minimization aims to create a simplified version of an organism's genome while maintaining its core (essential) biological functions. Among the several potential applications, perhaps the most promising are the creation of more efficient biotechnological microbial cell factories and the characterization and engineering of simpler synthetic genomes with less genetic elements and streamlined regulatory networks[313,316–318].

A microorganism that has received significant attention in the field is *Escherichia coli*. Being extensively used as the bacterial workhorse for research in molecular biology and genetics, the genome of this model organism is one of the best characterized to date[73,74]. Additionally, the many genetic tools available to engineer the bacterium make it a prime candidate for genome minimization efforts, as demonstrated extensively[319,256,320–325,66]. Examples of successful genome minimization for improved product biosynthesis in the bacterium are the deletion of 22% (1.03 Mb) from the genome of the *E. coli* strain W3110, resulting in improved growth in minimal medium and increased production of L-threonine[321,322], and the engineering of strain MG1655, whose heterologous yield of enhanced green fluorescent protein (eGFP) was increased by 44% after several deletions[326].

The discovery of clustered regularly interspaced short palindromic repeats (CRISPR) has shaken the grounds of genome editing[50,327]. Recent additions to the CRISPR genome editing toolbox have opened up interesting possibilities for genome minimization. Specifically, the development of class 1 CRISPR-Cas tools, a hallmark of which is that they make use of an RNA-guided, multi-protein effector complex, consisting of several CRISPR-associated (Cas) proteins. These CRISPR-Cas effector complexes use their bound CRISPR-RNA (crRNA) to bind to a complementary target DNA sequence (protospacer). Successful protospacer binding, which also requires the recognition of a protospacer-adjacent motif (PAM), results in the recruitment of the processive Cas3 nuclease. While this nuclease has been shown to generate large uni-directional deletions in human cells[270,328,329], this nuclease is known to generate large bi-directional deletions in prokaryotes making it very interesting for the field of prokaryotic genome minimization[66].

6.3 Results and discussion

6.3.1 Assessing the efficiency of Cascade-Cas3 genome minimization in *E. coli*

In an attempt to reproduce the reported efficiency of the tool, we first sought out to reproduce the experiments performed by Csörgő *et al.* in *E. coli*[66]*.* In their study, *E. coli* K-12 MG1655 competent cells were transformed by electroporation with a plasmid encoding the type I-C CRISPR system (targeting either the *pdeL* gene, about 30kb upstream of *lacZ,* or the *lacZ* gene), recovered for 1h and plated with selection for the plasmid. Next, colonies were grown overnight under plasmid-selective conditions. Single colonies were then grown overnight in liquid media, and the resulting cultures were screened on plates containing the *lac*-operon inducer IPTG and the LacZ synthetic substrate X-gal to perform white/blue screening. This allows for detecting genomic deletions that encompass the beta-galactosidase gene (*lacZ*), as colonies with the intact *lacZ* gene turn blue, while mutants appear as white. In our experiment, the colonies resulting from electroporation were grown overnight in liquid medium in parallel with or without the L-rhamnose inducer to assess the effect of inducing the expression of the CRISPR system earlier in the workflow on the efficiency of genome deletions.

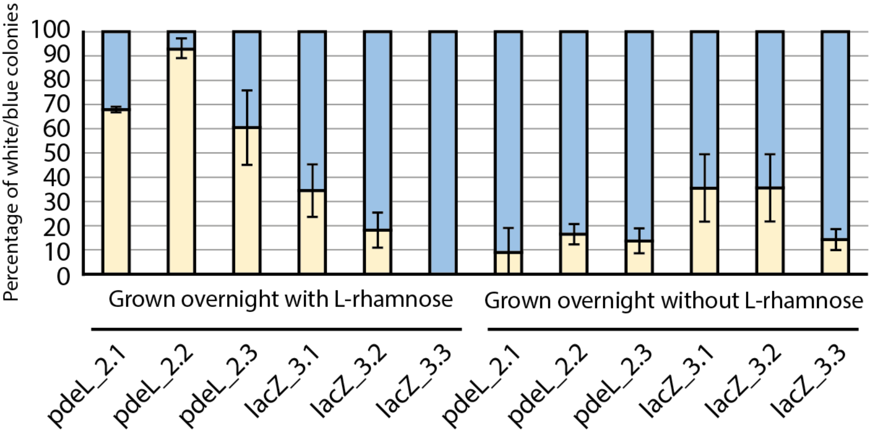


Figure 1: Genome editing efficiencies of the pdeL_2 and lacZ_3 plasmids. *Graph plotting the percentage of white and blue colonies scored after repeating the experiment described in Csörgő et al. Cells were transformed with a plasmid encoding a minimal type I-C CRISPR-Cas system under the control of a L-rhamnose-responsive promoter and a guide RNA targeting either pdeL (pdeL_2) or lacZ (lacZ_3). Cells were grown overnight without (original protocol, right) or with induction (left) after which genome deletion efficiencies were scored by blue/white screening. Per plasmid, three different colonies were picked, grown and studied, and calculations were made using two technical replicates per culture.*

In contrast to the results obtained by the authors in the original study (personal communication), growing the strains overnight in liquid inducing media rendered a higher percentage of edited (white) colonies compared to growing them without induction before plating (Figure 1). In the study carried by Csörgő *et al*., the crRNAs targeting *pdeL* and *lacZ* rendered percentages of white colonies of 82-85% and 51-90%, respectively[66], whereas our repetition resulted in efficiencies of only 9-16% and 14-36% (Figure 1, right). However, when grown overnight under induction, we obtained editing efficiencies of 60-93% (*pdeL*) and 0-35% (*lacZ*). In conclusion, we were able to obtain similar editing efficiencies than those reported in the original study, although a slight modification to the protocol was required for the system to render the highest efficiency in our experimental setting.

6.3.2 Devising the non-essential regions in *E. coli*

After establishing an efficient protocol for introducing large-scale deletions using the minimal type I-C CRISPR-Cas system, we set out to generate deletions as large as possible in several different genomic loci. To this end, we first screened the literature for studies that introduced deletions in the genome of *E. coli* to identify its genomic non-essential regions (NERs). Taking together the results of three independent studies[320,325,66] and setting the minimum region size to 10 kb, we defined 17 large NERs ranging from 10.2 kb to 123.1 kb (Table 1) and designed crRNAs to target them.

Table 1: Summary of the deletions generated in the aforementioned studies, which we used to craft our non-essential region (NER) targeting scheme.

Strain	Deletion size(s)	Coordinates MG1655	NER	NER size	NER coordinates
D1,2,3 **No.3 (pdeL)** **MD1** **MD43**	33 kb 110 kb 62 kb 45 kb	258167 – 291346 270603 – 381261 262738 - 324634 331590 – 376540	NER 1	123.1	258167 - 381261
MD12	21 kb	564278 – 585331	NER 2	21.1	564278 - 585331
MD36	11 kb	1085330 - 1096545	NER 3	11.2	1085330 - 1096545
MD26	12 kb	1128620 - 1140210	NER 4	11.6	1128620 - 1140210
MD11	26 kb	1196360 - 1222299	NER 5	25.9	1196360 - 1222299
D5 **MD2**	71 kb 82 kb	1397236 - 1467913 1398350 - 1480279	NER 6	83.0	1397236 - 1480279
MD8	25 kb	1625542 - 1650785	NER 7	25.2	1625542 - 1650785
MD27 **D7,8,9**	16 kb 16 kb	1960590 - 1977353 1962083 - 1978502	NER 8	17.9	1960590 - 1978502
MD28	27 kb	1995136 - 2021702	NER 9	26.6	1995136 - 2021702
MD5 **D10** **MD22**	14 kb 36 kb 36 kb	2064329 - 2078615 2066704 - 2102294 2099420 - 2135740	NER 10	71.4	2064329 - 2135740
MD37	12 kb	2163175 - 2175232	NER 11	12.1	2163175 - 2175232
D11	10 kb	2466369 - 2476583	NER 12	10.2	2466369 - 2476583
MD4	35 kb	2754181 - 2789271	NER 13	35.1	2754181 - 2789271
MD10	26 kb	3108702 - 3134399	NER 14	25.7	3108702 - 3134399
MD6	16 kb	3451950 - 3467875	NER 15	15.9	3451950 - 3467875
D12	38 kb	4285317 - 4323260	NER 16	37.9	4285317 - 4323260
MD9 **MD29**	53 kb 42 kb	4494698 - 4547733 4553513 - 4595035	NER 17	100.3	4494698 - 4595035
			667 kb = 14.37% of genome of *E. coli* MG1655		

As a proof of concept, we opted to first target the six largest NERs, namely NERs 1 (123.1 kb), 6 (83 kb), 10 (71.4 kb), 13 (35.1 kb), 16 (37.9 kb) and 17 (100.3 kb). We decided to target two sites per NER, one located towards approximately at 1/3 of the length of the NER (pNER#.1), and the other approximately at 2/3 of the NER (pNER#.2). Based on the original plasmid pCas3cRh (gentamycin resistance-containing) and to enable for later antibiotic cycling, two additional variants were generated with either a tetracycline or a kanamycin resistance cassette to later include the spacers as elaborated below (Table 2).

Table 2: Spacer design to perform the proof-of-concept genome minimization experiment.

Plasmid	Protospacer (5’ -> 3’)	Antibiotic marker
pCas3cRh	N/A	Gen / Tet / Kan
pNER1.1	CAGATCTAAGCTGTCTTGGCAGAACTGTGGAGGA	Gen
pNER1.2	CACCACCAGCATCCCTTTCTCGTTGACGCCACAC	Gen
pNER6.1	GTAATGCACCATCTTATCTCTCCCCTTAACGCCG	Tet
pNER6.2	TGGGCAATAAAGTTGCTGTGGGTGACTTTCACGG	Tet
pNER10.1	AGCGTGGTATCCAGTGCGTGACCATGTACAGTCT	Tet
pNER10.2	TCGGTAAGTTTGTTCAGTTGACGCAGCTGTTCCG	Tet
pNER13.1	GTACTTCCGTGTCGAGCAGTTCGAAATGCTGCAA	Kan
pNER13.2	TCCATCCACAATCACGTGGCAGACAATGTGTTGC	Kan
pNER16.1	TGGTTTCGGTTCATGCGTTCGCGCTGGATAACGT	Kan
pNER16.2	GATGGCGTTGATCATCTTCTCCAGATTCTCCAGC	Kan
pNER17.1	AAGACGTCCAGACATTTCTACGGCCTTAATAGGT	Gen
pNER17.2	TACCGTTTTATAACCGAGACTACGCACCACCAGT	Gen

6.3.3 A novel conjugation protocol for fast iterative genome minimization

The transformation of plasmids into *E. coli* is typically conducted by chemical transformation (using heat-shock) or by electroporation. Although these methods often provide reliably high transformation efficiencies, the preparation of competent cells can be laborious, especially in the context of making multiple NER deletions. In an experiment where strains need to undergo multiple consecutive cycles of genome editing, having to prepare competent cells every round of mutagenesis would slow down the process.

In an attempt to design a fast method for iterative genome editing, we drafted a protocol by which we subjected *E. coli* MG1655 to consecutive rounds of genome minimization by bacterial conjugation, using *E. coli* ST18 as a donor strain[330]. One of the promises of genome minimization involves enhancing the fitness of an organism by eliminating negative fitness costs associated with carrying redundant and/or non-essential gene content. Therefore, if a population of bacteria contains cells with deletions that improve their fitness, their relative abundance in a population will increase over time.

To increase the potential genetic variation obtained in each round of genome editing, we decided to pool different *E. coli* ST18 strains carrying a variety of plasmids. Up to 4 strains with different plasmids were pooled, all carrying the same antibiotic resistance marker and each targeting a different (region) of up to two NERs (for example, pNER1.1, pNER1.2, pNER17.1, and pNER17.2 were pooled and selected for in gentamycin) and conjugated them simultaneously into MG1655 recipient cells. The different transconjugants were then grown overnight in liquid media while inducing the CRISPR system. The resulting culture, made up of cells potentially carrying different deletions caused by the four pooled targeting plasmids, was then used for a subsequent round of conjugation. Here, *E.coli* ST18 donor cells containing a different set of NER-targeting plasmids were used, but with a different antibiotic selection marker to enable selection of the newly conjugated plasmid over the old one. The full cycle was carried out a total of three times, implying that three conjugations were performed, each of which involved four NER-targeting plasmids, and three selection rounds were carried out for every culture (Table 3). To assess whether the order in which the selection for the different antibiotics had any effect, the cycle was performed in parallel with all of the antibiotics, and the order of cycling we used was tetracycline after gentamycin, kanamycin after tetracycline, and gentamycin after kanamycin.

Table 3: Plasmid conjugation scheme listing the three tested antibiotic cycles.

	Round 1	Round 2	Round 3
Antibiotic cycle 1 (TKG) (Tet → Kan → Gen)	pNER6.1 pNER6.2 pNER10.1 pNER10.2	pNER13.1 pNER13.2 pNER16.1 pNER16.2	pNER1.1 pNER1.2 pNER17.1 pNER17.2
Antibiotic cycle 2 (KGT) (Kan → Gen → Tet)	pNER13.1 pNER13.2 pNER16.1 pNER16.2	pNER1.1 pNER1.2 pNER17.1 pNER17.2	pNER6.1 pNER6.2 pNER10.1 pNER10.2
Antibiotic cycle 3 (GTK) (Gen → Tet → Kan)	pNER1.1 pNER1.2 pNER17.1 pNER17.2	pNER6.1 pNER6.2 pNER10.1 pNER10.2	pNER13.1 pNER13.2 pNER16.1 pNER16.2

6.3.4 Genotypic profiling of strains undergoing iterative genome minimization

In order to monitor the genotypes present in the population of each tube undergoing the genome minimization experiment, we assessed whether we could analyze the bacterial culture directly by a multiplex loci PCR (MLPCR) reaction involving 12 primer pairs (each amplifying a region comprising the protospacer within the NERs) to assess the deletion efficiency of the targeted NERs (Figure 2; note that only 9 strong and 2 faint bands are seen in the wild type control, as one of the primer pairs never generated an amplicon).

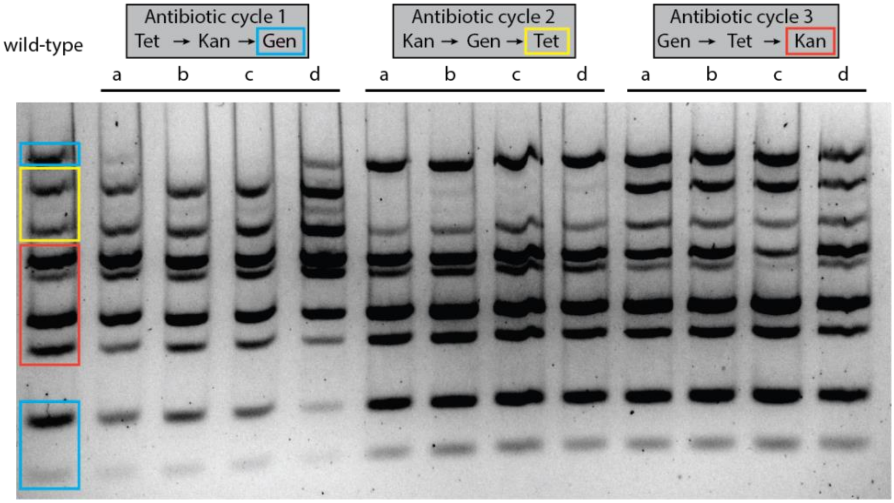


***Figure 2: Agarose gel electrophoresis of the MLPCRs performed**, **the template in the reactions being one of the four cultures per cycle (a, b, c, d) after having undergone the genome editing cycle shown on top.** The colored rectangles point at the NERs targeted by the last plasmids conjugated in the cycle, blue being NERs 1 and 17, yellow being NERs 6 and 10, and red being NERs 13 and 16.*

Several of the cultures used to assess the presence of deletions through MLPCR seemed to lack one or more amplicons, indicative of a deletion in these regions (Figure 2). Remarkably, the regions deleted coincided with the plasmids used in the last round of conjugation, possibly indicating that the last plasmids conjugated in the strains might be responsible for most of the deletions in the populations. Except for the genome editing cycles finishing with kanamycin resistance-encoding plasmids, these results reflected a potential success in generating deletions through our protocol (Figure 2).

To investigate the exact nature of some of these deletions, the 3 cultures were plated in parallel on plates containing the inducer and the antibiotic they had been selected for in the last round of conjugation. After overnight incubation of the plates, we picked six colonies for each of the 3 cultures and assayed them through MLPCR to decide which ones might be good candidates for whole-genome sequencing.

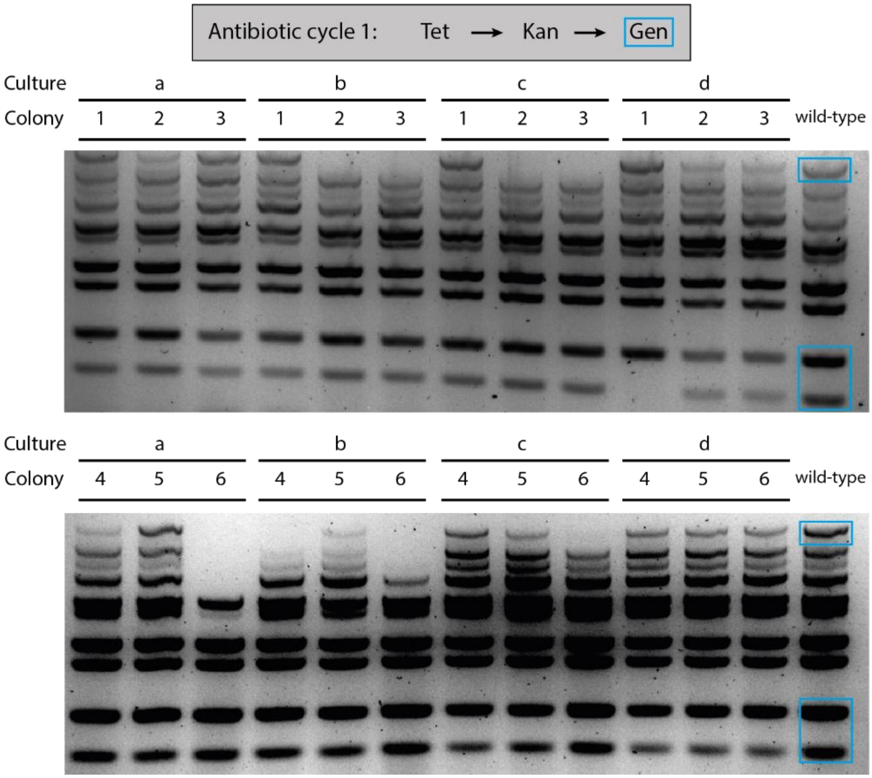


Figure 3: Agarose gel electrophoresis of the MLPCRs performed on six individual colonies from the four different cultures (a, b, c, and d) having followed the genome editing cycle 1 (tetracycline to kanamycin to gentamycin; TKG).

Several of the colonies resulting from the tetracycline, kanamycin, gentamycin (TKG) cycle showed a promising genotype (i.e. indicative of deletion of one or multiple NERs) according to the MLPCR results (Figure 3). Colonies b2, b3, c2, c3 and c6 seemed to lack the upper band, potentially reflecting the successful editing of the NER 1. Only one of the colonies, d1, seemed to lack the band for NER 17. Interestingly, a few colonies seemed to miss multiple bands. Specifically, colony a6 missed bands for NERs 1, 6, 10 and 13, whereas colony b6 missed bands for NERs 1, 6 and 10.

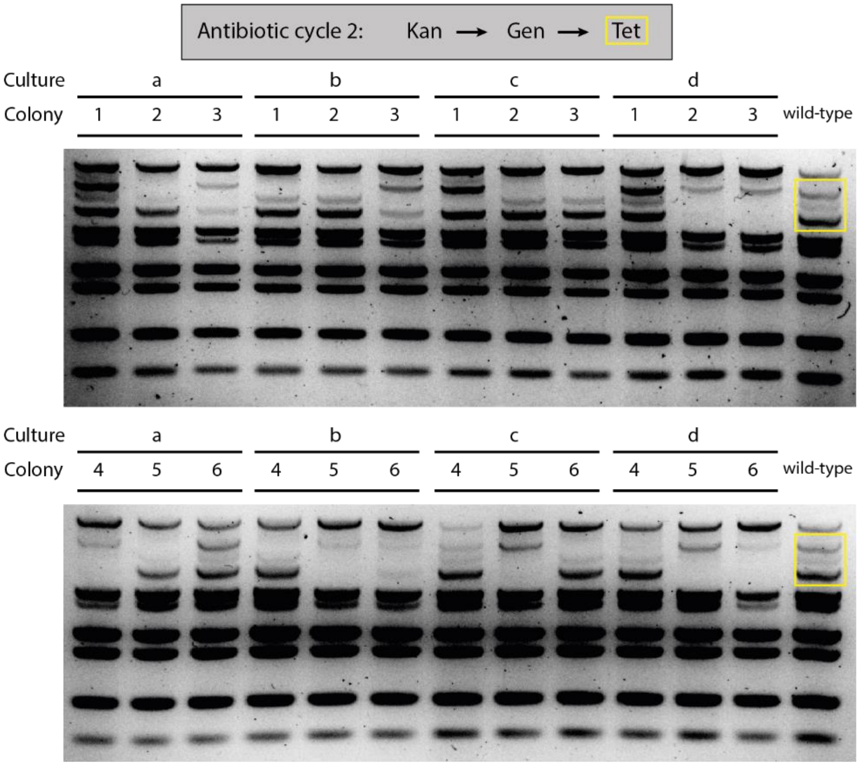

Antibiotic cycle 2: Kan → Gen → Tet
Culture a b c d
Colony 1 2 3 1 2 3 1 2 3 1 2 3 wild-type
Culture a b c d
Colony 4 5 6 4 5 6 4 5 6 4 5 6 wild-type

Figure 4: Agarose gel electrophoresis of the MLPCRs performed on six individual colonies from the four different cultures (a, b, c, and d) having followed the genome editing cycle 2 (kanamycin to gentamycin to tetracycline; KGT).

Based on MLPCR, the strains that were subjected to the KGT cycle seemed to render relatively good success after examining the results of the MLPCRs (Figure 4). The majority of the colonies seemed to have amplicon patterns different to the wild type, suggesting that this genome editing cycle worked.

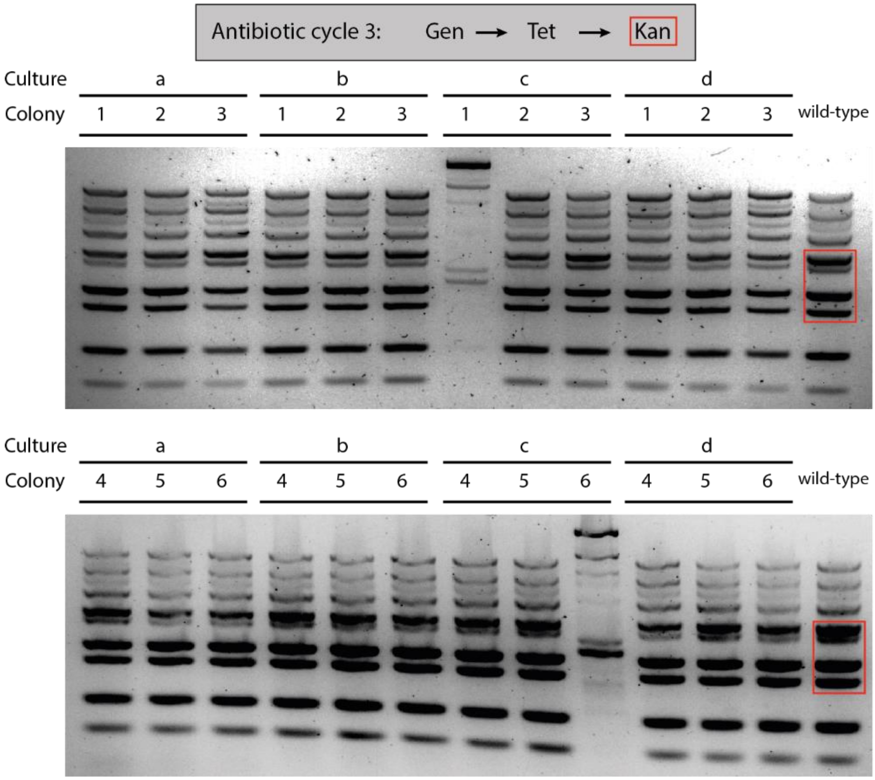


Figure 5: Agarose gel electrophoresis of the MLPCRs performed on six individual colonies from the four different cultures (a, b, c, and d) having followed the genome editing cycle 3 (gentamycin to tetracycline to kanamycin; GTK).

Most of the genotyped colonies belonging to the GTK cycle exhibited the same bands as the wild-type strain, indicating that this cycle was not successful (Figure 5). Colonies c1 and c6 showed completely different band patterns; after further investigation through Sanger sequencing of their 16S rRNA gene we determined that these colonies were in fact not *E. coli* and rather a contamination of *Cupriavidus gilardii*, probably acquired during the experiment.

A set of three promising colonies (TKGb6, KGTb5 and KGTb6) was selected and grown, after which their genomes were isolated and sequenced via Nanopore sequencing.

6.3.5 Sequencing results suggest that MLPCR is not a good proxy to assess multiple deletions

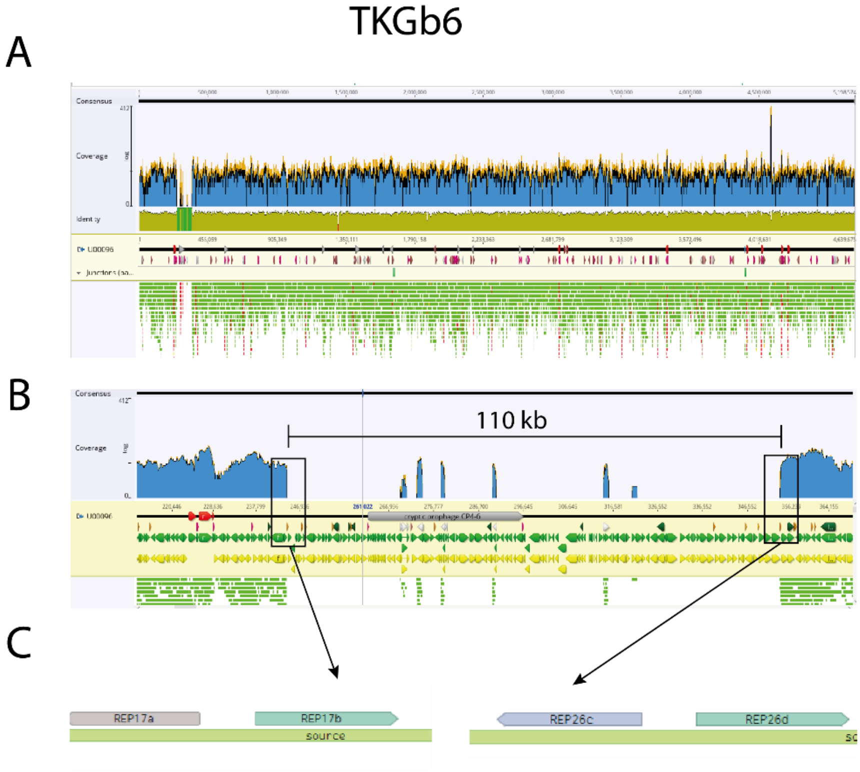


Figure 6: Nanopore sequencing results for the TKGb6 colony visualized in the Geneious software. *(A) Mapped reads (green) aligned to the genome of* E. coli *MG1655, producing a coverage map (blue). (B) Zoomed in region with a drop in coverage reflecting a deletion of 110 kb. The coverage islands between the deletion junction are regions mapped incorrectly due to genomic repetitive sequences. (C) Deletion junction found between REP elements 17 and 26.*

In the case of the colony TKGb6, the MLPCR results were indicative of deletions in NERs 1, 6 and 10 (Figure 3). However, only a very large deletion of 110 kb was found in the genome of the strain, as reflected by the drop in coverage, located in the NER 1 (Figure 6). TKG strains were last selected for gentamycin resistance, and NER 1 is targeted by a gentamycin resistance-encoding plasmid, which supports our previous observation that the last antibiotic used in the cycle might be correlated with the mutations observed in the studied strains. This specific NER1 deletion has its boundaries within bacterial interspaced mosaic elements (BIMEs)[331], in this case between the repetitive extragenic palindrome (REP) elements 17 and 26 (Figure 6C).

Deletion junctions occurring between REP elements have often been reported in similar experiments in *E. coli*[53,66,325,332]. These elements have been shown to bind DNA gyrase[333] and DNA polymerase I[334], and cleavage by Cas3 has been reported to collocate with protein roadblocks, that is, proteins bound to DNA that impede the DNA reeling by Cas3, thereby causing Cas3 to generate pulling forces strong enough to cause a DNA break[335]. It is therefore likely that Cas3 reels one of the DNA ends until it finds protein roadblocks at, for example, a region with REP elements with proteins bound to it, and then a single or double strand break is created, inciting DNA repair. Alternatively, DNA processing enzymes might also cleave DNA upon encountering a protein roadblock, resulting in frequent cleavage at REP elements. In any case, the ubiquity of these DNA elements in deletion junctions found in this type of genome editing experience reflects that these are hotspots for DNA repair, possibly through alternative end joining (AEJ), also known as microhomology-mediated end joining (MMEJ)[61].

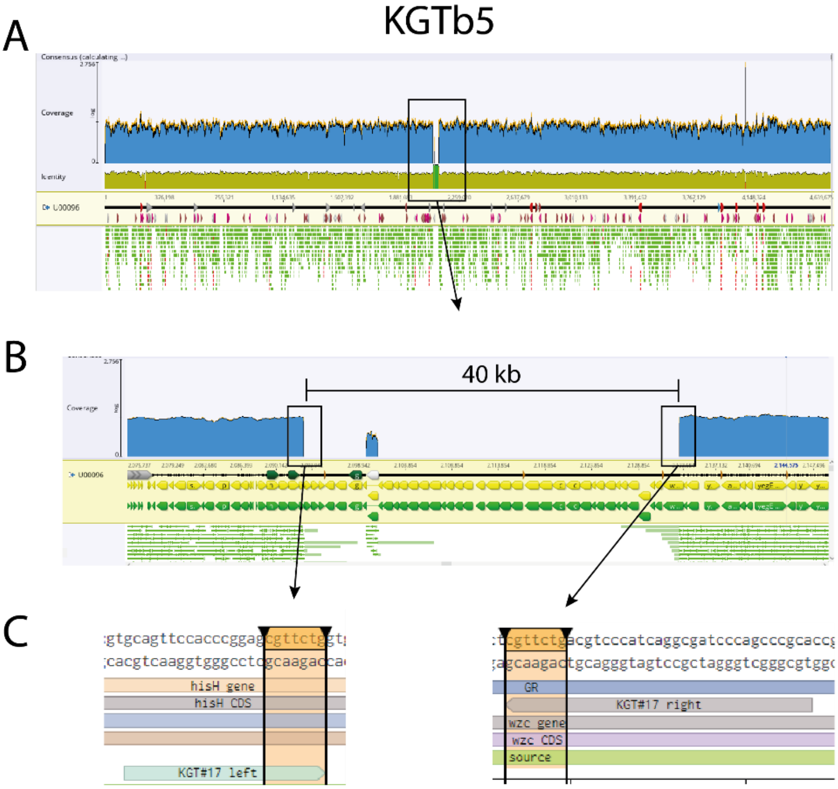

A
KGTb5
Consensus (calculating ...)
Coverage
Identity
U00096
B
40 kb
Coverage
U00096
C
gtgcagttccacccggagcgttctggt
cacgtcaaggtgggcctcgcaagacca
hisH gene
hisH CDS
KGT#17 left
tcgttctgacgtcccatcaggcgatcccagcccgcaccg
agcaagactgcagggtagtccgctagggtcgggcgtggc
GR
KGT#17 right
wzc gene
wzc CDS
source

Figure 7: Nanopore sequencing results for the KGTb5 colony visualized in the Geneious software. *(A) Mapped reads (green) aligned to the genome of* E. coli *MG1655, producing a coverage map (blue). (B) Zoomed in region with a drop in coverage reflecting a deletion of 40 kb. The coverage island between the deletion junction is a region mapped incorrectly due to genomic repetitive sequences. (C) Deletion junction found to have 7 bp of microhomology.*

As for the colony KGTb5, a deletion of 40 kb was found within the NER 10 (Figure 7). In this case, 7 bp of microhomology were observed at the deletion junction, suggesting alternative end joining (AEJ) to be responsible for the DNA repair leading to this deletion. Further, NER 10 is targeted by plasmids encoding for tetracycline resistance, which were the last plasmids conjugated into the KGT strains, reinforcing the notion that the presented protocol is only able to generate deletions at the last round of the genome editing cycle.

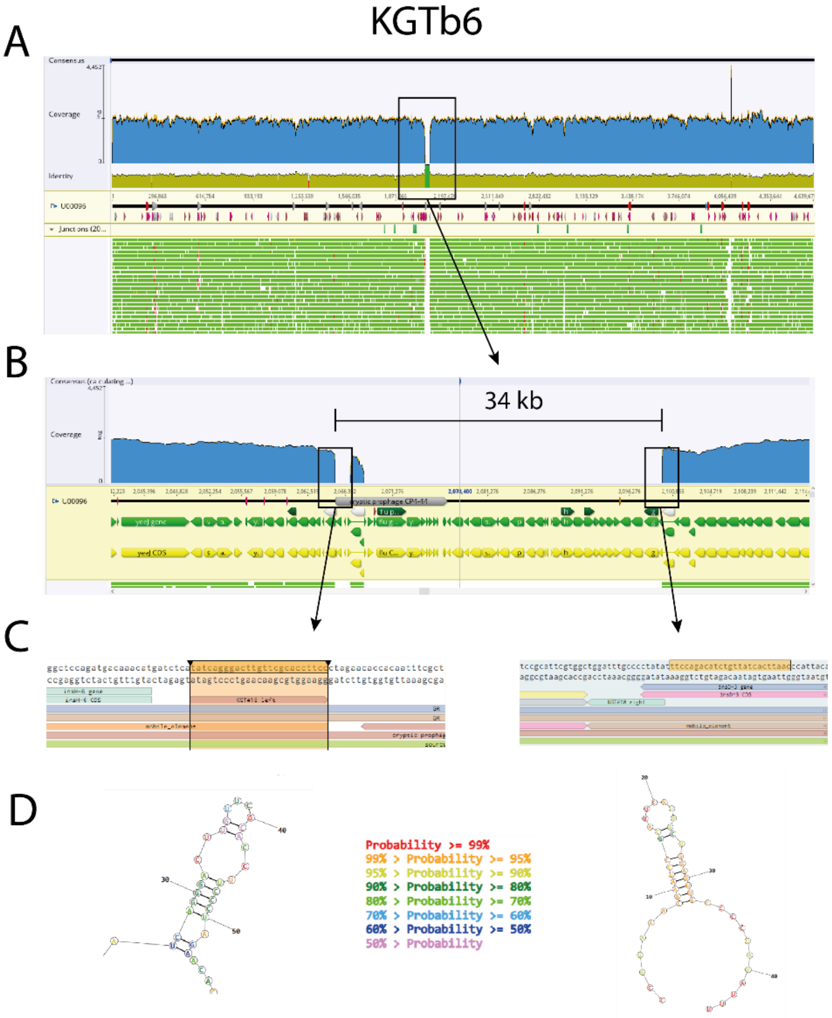


Figure 8: Nanopore sequencing results for the KGTb6 colony visualized in the Geneious software. *(A) Mapped reads (green) aligned to the genome of* E. coli *MG1655, producing a coverage map (blue). (B) Zoomed in region with a drop in coverage reflecting a deletion of 34 kb. (C) Deletion junction found between insertion elements. (D) The sequences at the deletion junction are predicted to have a stem loop secondary structure when single stranded.*

The colony KGTb6 was found to carry a deletion of 34 kb in the NER 10 (Figure 8). In this case, no obvious sequence homology was found at the deletion junction. However, analyzing the flanking sequences with a nucleic acid secondary structure online prediction tool[336] suggested a strong probability of these sequences having a stem loop type of secondary structure, which is also common to REP elements[331], insertion sequences and other transposable elements often found at deletion junctions.

Taken together, our Nanopore sequencing results confirm our genome editing protocol to successfully mediate large deletions in the genome of *E. coli*. However, the stacking of deletions was not successful in the tested experimental conditions, as only one deletion was found per sequenced colony. The deletion in all the cases was generated by targeting of one of the plasmids in the last round of iterative conjugations, suggesting an unexpected problem arising during our protocol.

6.3.6 Plasmid maintenance as a potential explanation for the ineffective editing of multiple loci

To understand what could have gone wrong in our experimental setup, we first thought of plasmid maintenance as a potential cause for the inconsistency between our MLPCRs and the actual sequencing results. For our consecutive rounds of genome editing to work, as they all use plasmids with the same origin of replication, it is important that the plasmids conferring resistance to a certain antibiotic are cured during the growth phase where the antibiotic selection is not included, ideally before the new plasmid (with a different antibiotic resistance marker) is introduced. In our experimental design, all plasmids are targeting the genome and thus their maintenance is expected to carry a significant fitness cost. This fitness cost would be eliminated if deletions are generated at the target sites, while otherwise the plasmids might be lost in the next round of genome editing upon incubation with a different antibiotic.

To assess whether (unsuccessful) plasmid curing was the reason for the absence of multiple deleted NERs in our experiments, we assessed plasmid curing by scoring the total number of CFUs found on plates supplemented with an antibiotic corresponding with either the last plasmid the cells were transformed with, or the very first (Table 1).

Table 4: Results of the plasmid curing experiments. *CFUs were scored by plating cultures of the indicated strains on a single plate containing the antibiotics that were used in either the first (red) or last round (green) of conjugation. Numbers in the left half of the table were obtained after conjugation of non-targeting plasmids, while that of the left half were obtained after conjugation of the targeting plasmids used in the genome editing cycle*

Non-targeting plasmids				Targeting plasmids			
Strain	pNER0_Tet	pNER0_Kan	pNER0_Gen	Strain	pNER_Tet	pNER_Kan	pNER_Gen
TKG	Round 1	Round 2	Round 3	TKG	Round 1	Round 2	Round 3
KGT	Round 3	Round 1	Round 2	KGT	Round 3	Round 1	Round 2
GTK	Round 2	Round 3	Round 1	GTK	Round 2	Round 3	Round 1

Strain	Tet3	Kan50	Gen15	Strain	Tet3	Kan50	Gen15
TKGa	0		1	TKGa	6		134
TKGb	0		0	TKGb	0		0
TKGc	3		0	TKGc	0		16
TKGd	2		0	TKGd	0		6
KGTa	230	64		KGTa	228	52	
KGTb	110	8		KGTb	13	286	
KGTc	288	696		KGTc	34	7	
KGTd	451	1189		KGTd	3	0	
GTKa		2	6	GTKa		>2000	23
GTKb		17	66	GTKb		700	126
GTKc		28	200	GTKc		>1000	>1000
GTKd		0	30	GTKd		>1000	0

The numbers in Table 4 reflect substantial differences between strains regarding the recalcitrance of the plasmids used. For the TKG cultures, most of them had successfully cured the plasmid they were conjugated with in the first round. For the KGT and GTK strains, this was not so evident, as roughly similar CFUs were obtained when plating on the first and last round’s antibiotic The trends differed between biological replicates, and the CFU counts were not high enough in several of the samples due to plating too diluted cultures. Altogether, the results seem to indicate that spontaneous plasmid curing is a stochastic process and might depend on the order of introduction of the plasmids used, as curing in the TKG cycle seemed more efficient when compared to the other cycles.

Reflecting on our proof-of-concept experiment, in every round of editing, plasmids newly conjugated into MG1655 might be subjected to inconsistent patterns of maintenance, rendering the cycle unsuccessful.

6.4 Conclusions

6.4.1 Evaluation of the iterative conjugation protocol

The MLPCRs on the cultures and on single colonies seemed indicative of multiple deletions being present in the screened cultures and strains. After genome sequencing of individual clones, however, we observed only one deletion to be present per clone. The deleted NER seemed to coincide with the region targeted by the plasmids in the last round of the iterative conjugation cycle. After assessing the plasmid curing efficiency, we concluded that plasmid maintenance and curing seemed to follow very stochastic patterns, making the protocol unreliable for the generation of strains with multiple deletions. However, we did confirm that the generation of large deletions is possible through conjugation, and changes in the methodology might make a similar protocol suitable for stacking deletions and generating strains with a minimized genome.

We suggest that the protocol needs to be adapted to generate strains with multiple deletions. To ensure the generation of strains with multiple deletions, transconjugant colonies should be grown in LB agar with an appropriate antibiotic. Subsequently, rather than using whole liquid cultures for the consecutive conjugations, single colonies of the potential mutants should be first characterized via MLPCR, then grown overnight without antibiotics, and only then they should serve as recipient for the next conjugation cycle. While this would likely result in strains with multiple deletions, an extra day or two per round would be needed, implying almost a doubling of the required time (from approximately 4 working days to at least 7 working days). Additionally, by doing this, one would be limiting the possible genetic space explored in each round of editing. Alternatively, the use of a temperature-sensitive origin of replication[337] could enable the experiment to be performed while more stringently ensuing plasmid curing during the growth of the cultures prior to each conjugation.

6.4.2 Role of repetitive genomic elements in DNA repair

So far, as mentioned above, several independent groups have reported the generation of big deletions in the genome of *E. coli* via targeting its genome with different genome editing tools. While the endonucleases used throughout these studies vary and differ considerably with respect to their mechanisms of action. Nevertheless the deletion junctions are often found in the same places, namely between REP elements, prophages and other transposable sequences. Altogether, this reflects that mutagenesis in these cases is not performed by the tools, but rather by the DNA repair mechanisms able to ligate the chromosome after DNA breaks occur.

6.4.3 Subpopulation fitness might impede the stacking of multiple deletions in our protocol

The consequences of deleting big regions in the genome of microorganisms often are deleterious for growth. It is therefore likely that, during the course of our experiments, cells with multiple deletions in their genome have a reduced fitness compared to cells in which this locus is still intact. The latter will be able to outgrow the mutants, rendering the fraction of multiple mutants in a given culture very low. This would likely drive the efficiency of our protocol to be very low at stacking mutations, which seemed to be the case. More stringent genome editing tools might thus be effective at ensuring that editing occurs at every step of the protocol, as it seems that the CRISPR tool we employed was not stringent enough to prevent cells without mutations to continue being present in the populations used in our experiments.

Additionally, effective polyploidy[56] (that is, *E. coli* having more than one plasmid at the same time during fast exponential phase) might be another reason by which cells bypass having multiple deletions. While one or multiple copies of a chromosome might be edited within a cell while grown under selection for targeting of a certain locus, once these cells are selected for the next antibiotic in the cycle, cells originating from the polyploid cell that now carry an intact copy of the chromosome will likely have a considerable growth advantage, which will ultimately result in cells only maintaining one deletion in their genome after having undergone our genome editing cycle.

6.5 Materials and methods

6.5.1 Bacterial strains and growth conditions

Escherichia coli XL1-Blue was kindly provided by the Bondy-Denomy lab from UCSF, and was used for the amplification of the *pdeL* and the *lacZ* targeting plasmids, as well as for the cloning of all other generated plasmids. *E. coli* DH5α was routinely used for plasmid cloning and maintenance. *E. coli* ST18[330] was used for the transformation of plasmids into MG1655 through conjugation. Genome minimization experiments were performed using *E. coli* MG1655. For preparation of chemically competent cells, *E. coli* DH5a, ST18 and K-12 MG1655 were prepared following an elsewhere described protocol[338] and transformed using heat-shock. All strains were grown in either liquid or 1.5% agar w/v Luria-Bertrani (LB) medium at 37ºC.

The medium was supplemented with gentamycin (15 µg/mL), tetracycline (5 µg/mL) or kanamycin (50 µg/mL) for plasmid maintenance, with 0.1% L-rhamnose for plasmid induction and 40 µg/mL 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal) and 0.5 mM isopropyl β-d-1-thiogalactopyranoside (IPTG) for white/blue colony screening. A full overview of the plasmid editing system is provided in the original study[66].

6.5.2 Plasmid construction and sgRNA design

The plasmids pCas3cRh, pCas3cRh_pdeL2 and pCas3cRh_lacZ3 were kindly provided by the Bondy-Denomy lab from UCSF. To generate the rest of the plasmids, first, amplified pCas3cRh was purified using the GeneJET Plasmid Miniprep Kit (Thermo Fisher) and digested with Bsa-I restriction enzyme. After heat-inactivation, the linearized plasmids were ran in 1% agarose electrophoresis gels and purified using the Zymoclean Gel DNA Recovery Kit (Zymo Research). Targeting spacers were designed using a custom Python script, using a 5'-TTC PAM and being 34 nt long. Oligonucleotides were ordered including specific overhangs to facilitate their ligation to the plasmid. These oligonucleotides were first phosphorylated with PNK and then annealed by mixing them in 1:1 ratio and cooling them from 95ºC to 12ºC in 45 minutes. The resulting annealed spacer cassettes were then diluted 1:500 in MQ, and 5 µL cloned into 150 ng of the previously purified linear pCas3cRh using the T4 DNA Ligase (New England BioLabs) by incubation at room temperature for 10 minutes and following heat inactivation for 10 minutes at 65ºC. 1-5 µL of the resulting ligated product were used in heat-shock transformations as described above.

6.5.3 White/blue colony screening

To study the disruption of the *lacZ* gene, *E. coli* MG1655 was chemically transformed with 100 ng of pCas3cRh, pCas3cRh_pdeL2 and pCas3cRh_lacZ3. The transformants were plated onto LB agar plates supplemented with gentamycin and incubated overnight. The resulting colonies were then resuspended in 1 mL liquid LB medium, and 100 µL of the resuspension were added to 10 mL liquid LB medium with gentamycin, with and without 0.1% L-rhamnose, grown overnight, and finally plated in LB agar with gentamycin, L-rhamnose, IPTG and X-gal. To understand the editing efficiency of the different plasmids respective to the *lacZ* gene, the percentage of white colonies was calculated per plasmid, using three transformation experiments, with and without induction of the CRISPR system.

6.5.4 Conjugation of *E. coli* strains

E. coli ST18, auxotroph for 5-aminolevulinic acid (5-ALA)[330], was used as a donor strain to conjugate the recipient strain *E. coli* MG1655. Donor and recipient strains were grown overnight at 37ºC with appropriate supplementation after which 500 µL of each were put together on a 1.5 mL microcentrifuge tube. The tube was then spun down, the supernatant discarded, and the pellet was resuspended in 100 µL of fresh LB medium without antibiotic and with 5-ALA. Of the resulting mix, 25 µL were placed on an LB agar plate and incubated at room temperature for 5 hours to facilitate conjugation. Afterwards, the dry mix of cells was resuspended in 100 µL of fresh LB medium and different dilutions of this mixture were used for plating LB plates with appropriate antibiotics for selection. The plates were incubated overnight at 37ºC and collected for analysis or further experimentation during the morning.

6.5.5 Multiplex loci PCR (MLPCR)

Inspired by the design employed to perform the so-called multiplex allele-specific PCR[339], we designed the herein called multiplex loci PCR (MLPCR) in which 12 primer pairs were designed to give information about the presence or absence of the targeted NERs, having designed 2 amplicons per NER. In order to minimize the chances of primer dimers and other amplification artifacts, we used the online tool Multiple Primer Analyzer from ThermoFisher Scientific to modify the primer pairs until no primer interactions were predicted.

The MLPCRs were performed using the Q5 polymerase (Q5 High-Fidelity 2x Master Mix, New England BioLabs). After some experimentation, we found the best results using an annealing temperature of 63ºC and a concentration of 0.83 µM per primer. After three rounds of troubleshooting and redesigning the primers, we obtained a mix that rendered 11 out of the 12 desired bands (Figure 8), which we deemed suitable for the first proof of concept experiment. Individual colonies and liquid cultures were used as a template for MLPCR.

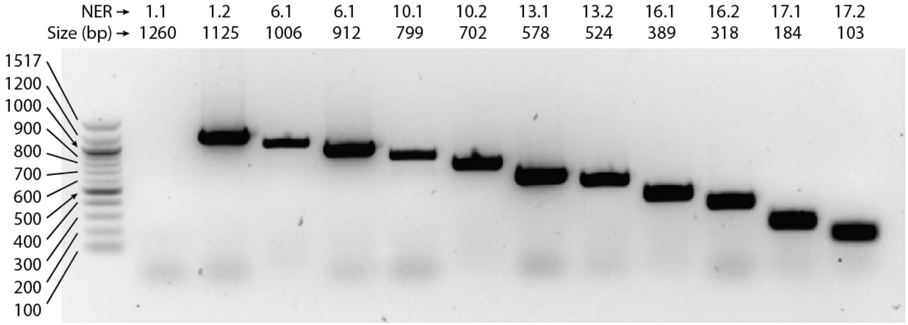


Figure 9: Agarose gel electrophoresis results of the individual bands designed for the MLPCR. Note that the first band (1.1, of 1260 bp) was never successfully amplified.

6.5.6 Sequencing and analysis of deletion strains

Cultures of *E. coli* MG1655 and its derived deletion strains were grown in LB medium (10 g/L bacto tryptone, 10 g/L NaCl, 5 g/L yeast extract) at 37ºC for 20 hours at 200 rpm and then pelleted by centrifugation. Genomic DNA was extracted using a DNeasy blood and tissue kit (Qiagen) and cleaned and concentrated using AMPure XP beads (Beckman Coulter). Long-read sequencing data was obtained using the Nanopore MinION Mk1C platform, using the SQK-LSK109 library kit, the rapid barcoding kit SQK-RBK004, the flonge sequencing expansion (EXP-FSE001), the flow cell priming kit (EXP-FLP002) and a MinION R9.4.1 flow cell. Nanopore data was base called using Guppy (Oxford Nanopore Technologies) and demultiplexed using Deepbinner v0.2.0[340]. Nanopore FAST5 reads were converted to FASTQ files using Poretools v0.6.0[341], discarding reads shorter than 2000 bp. The first 50 bp and last 20 bp of each Nanopore read were trimmed, and reads with average quality scores lower than 9 were removed using NanoFilt v2.6.0[342]. The processed Nanopore reads were aligned onto the genome of *E. coli* K12 MG1655 (accession NC_000913) and final assemblies were performed using Minimap2[343]. Visualization of the assemblies was done using the Geneious bioinformatics software.

6.5.7 Plasmid curing experiments

To score plasmid curing, the conjugation protocol was followed as described above and 10 μL of resuspended conjugation mixture were added to 10 mL of fresh LB medium with L-rhamnose and the appropriate antibiotic following the iterative conjugation cycle as also described above. The inoculated cultures were then grown overnight at 37ºC and then used to conjugate the next plasmids in the editing cycle. Once the last batch of plasmids was conjugated into the strains, 10 μL of the resuspended conjugation mixture were added to 10 mL of fresh LB medium with L-rhamnose and the appropriate antibiotic. Then, the resulting cultures were plated onto LB agar medium with the appropriate antibiotic for selection for the last plasmid of the cycle and L-rhamnose and incubated overnight at 37ºC.

6.6 Acknowledgements

We are very grateful to Bálint Csörgő and Joe Bondy-Denomy for the plasmids and the discussions throughout the duration of this study. We thank Simon Jackson from the University of Otago for his guidance regarding the Nanopore scripts. We acknowledge financial support from the Dutch Research Council (NWO) Gravitation grant BaSyC [024.003.019 to M.F.-B., N.J.C., R.H.J.S and J.v.d.O.], and a VIDI grant (VI.Vidi.203.074) to R.H.J.S., and Spinoza prize [SPI 93-537] to J.v.d.O..

6.7 Supplementary data

6.7.1 NER spacers

Table 5: Spacers designed to target non-essential regions (NERs) in the genome of **E. coli** ***MG1655.***

NER	Spacer sequence (5’→3’)
1.1	cagatctaagctgtcttggcagaactgtggagga
1.2	caccaccagcatccctttctcgttgacgccacac
6.1	gtaatgcaccatcttatctctccccttaacgccg
6.2	tgggcaataaagttgctgtgggtgactttcacgg
10.1	agcgtggtatccagtgcgtgaccatgtacagtct
10.2	tcggtaagtttgttcagttgacgcagctgttccg
13.1	gtacttccgtgtcgagcagttcgaaatgctgcaa
16.1	tccatccacaatcacgtggcagacaatgtgttgc
16.2	tggtttcggttcatgcgttcgcgctggataacgt
17.1	gatggcgttgatcatcttctccagattctccagc
17.2	aagacgtccagacatttctacggccttaataggt

6.7.2 Primers for MLPCR

Table 6: Primers designed to carry the multiplex loci PCR (MLPCR).

Primer name		Sequence (5´→3’)	Amplicon length (bp)	Annealing temperature
BG24255	MLPCR_NER1.1_Fw2	GATATGTACTCCATGAGTACCTTG	1260	63ºC
BG28647	MLPCR_NER1.1_Rv3	CAGGCTCTACTCTTGTCAAC		
BG24123	MLPCR_NER1.2_Fw	TGTCGATTTCGCGTCG	1125	
BG24124	MLPCR_NER1.2_Rv	TGCGAGACAAACCAGAC		
BG24125	MLPCR_NER6.1_Fw	ACTTCTCCTTGCCGTAAC	1006	
BG24126	MLPCR_NER6.1_Rv	CCCCCACCAATTCAGATAC		
BG24256	MLPCR_NER6.2_Fw2	AATAGTGGCAGTGTAACAGG	912	
BG24257	MLPCR_NER6.2_Rv2	AACCGTATTGGGCTTCG		
BG24258	MLPCR_NER10.1_Fw2	GGTCTTCAGAGAGTGAACC	799	
BG24130	MLPCR_NER10.1_Rv	CGCCATTTCCCATGGATA		
BG24131	MLPCR_NER10.2_Fw	CCAGTAATATTCGCCGCT	702	
BG24132	MLPCR_NER10.2_Rv	GCTGCCCAAACACAAC		
BG24133	MLPCR_NER13.1_Fw	GAGCTTCAAATCATTGTTGAGG	578	
BG24259	MLPCR_NER13.1_Rv2	GGAACATCAAAGGTCGCG		
BG24135	MLPCR_NER13.2_Fw	GATATTTTGCGTTCCGCTG	524	
BG24260	MLPCR_NER13.2_Rv2	ACACGACGATTGAAGGC		
BG24261	MLPCR_NER16.1_Fw2	GATCATCCTCGGTTCCTG	389	
BG24262	MLPCR_NER16.1_Rv2	TTAACATTTCCCGCGATAACC		
BG24139	MLPCR_NER16.2_Fw	GGTCCATCAGCGACAC	318	
BG24140	MLPCR_NER16.2_Rv	TGGATAAATTCTTCACCCCG		
BG24141	MLPCR_NER17.1_Fw	TGAAGATCACGCCCATC	184	
BG24263	MLPCR_NER17.1_Rv2	ATAGCGCATGATGGCTG		
BG24143	MLPCR_NER17.2_Fw	AGCAGGCAACGATTCC	103	
BG24144	MLPCR_NER17.2_Rv	GCCGGCAGAATAAACTCG		

Chapter 7. Summary and discussion

7.1 Thesis summary

An abundance of new information has been produced by the research fields of molecular biology and genome engineering in recent years. In thesis I review the most relevant new findings and the state-of-the-art, as well as experimental studies, of two specialized fields: protein expression and CRISPR-based template-free genome editing.

In Chapter 1, the present thesis is introduced, starting with some examples of historical early uses of biotechnology for the benefit of humanity. A brief history of molecular biology is outlined, emphasizing the developments that arguably had the biggest impact on the field. The findings that laid foundation for the fields of recombinant protein production and microbial genome editing are discussed, noting the potential of microbial metabolism to solve global challenges, and tracing the relevance of DNA repairs in that context.

In Chapter 2, a review of several recent studies undertaken to improve protein production is provided. Progress in the generation and analysis of big data is helping to provide a more comprehensive overview of the many factors involved in this process. Specifically, in eukaryotes, there appears to be a strong correlation between mRNA stability and translation elongation rates. Transcripts with fast translation rates might be more stable due to the higher load of ribosomes protecting it from degradation effected by RNA-degrading enzymes. Additionally, novel tools are being used to investigate translation initiation, enabling the experimental determination of RNA secondary structures *in vivo*, which are much more valuable than *in silico* predictions, as they will help to better understand the mechanistic consequences of secondary structures in translation elongation. Altogether, given that every sequence feature affects several factors simultaneously, it is often difficult or impossible to distinguish the specific effects of particular sequence features in the overall efficiency of protein production. The effect of codon usage on translation elongation, for instance, is primarily obscured by changes in translation initiation surfaced by changes in secondary structures following codon modifications. Machine learning methods may prove a useful tool to reveal these effects and their dynamics. While this may lead to improved predictability of sequence features, the downside is that such approaches may not necessarily contribute to a better understanding of the underlying biology, due to their "black box" nature, ultimately enabling researchers to optimize the production of proteins without fully understanding all the individual factors that enable it.

In Chapter 3, we demonstrate the successful optimization of membrane protein expression in *E. coli* by using a translational-tuning system. The Bicistronic Design (BCD) elements leverage the translational coupling of a gene encoding a short leader peptide and a specific gene of interest. Combining two RBS sites in the same mRNA, the first RBS prevents the formation of translation initiation-hindering secondary structures and enables the accessibility of the second RBS, which can vary in strength. A standardized library comprising BCD elements proves a succinct method to fine-tune the production of membrane proteins. This approach involves an easy library assembly step followed by microtiter plate analysis, facilitating the efficient selection of high-producing clones. Application of the BCD approach to the four membrane proteins tested in this study resulted in increased protein levels, from 3 to 7-fold higher than those obtained with two other recently developed methods for optimizing membrane protein production. The inducer-free, constitutive, and high-level production of functional membrane proteins presented in this chapter may find broad utility in membrane protein purification studies as well as in synthetic biology projects that involve membrane proteins.

In Chapter 4, a review is given on the most relevant developments of the prokaryotic DNA repair field, specifically template-free repair (that is, without the requirement of exogenous DNA templates for recombination), with a focus on CRISPR-based genome engineering. A swift summary on the development of nuclease-based genome engineering tools is given, finishing with how CRISPR tools quickly revolutionized the field. The two prokaryotic non-templated DNA repair pathways, non-homologous end joining (NHEJ) and alternative end joining (AEJ), are explained and the proteins and mechanisms involved therein dissected. Different strategies making use of template-free DNA repair pathways and CRISPR-Cas tools are then discussed, including ubiquitously used gene inactivation, sparse DNA insertion and the increasingly popular genome minimization. Some considerations are noted regarding whether NHEJ or AEJ are preferred; NHEJ typically generates smaller indels, which are attractive for gene inactivation, while AEJ requires microhomologies and can generate deletions of more than 100 kb, making it not only suitable for large gene inactivation but also preferred for genome minimization efforts. Lastly, a few words on future directions are given, mostly highlighting the need for further studies being able to fully distinguish outcomes of NHEJ and AEJ for the full characterization of prokaryotic non-templated DNA repair pathways.

In Chapter 5, the theory discussed in Chapter 4 is put into practice. Throughout the completion of a previous study in *Rhodobacter sphaeroides*, it was noted that certain colonies were able to escape CRISPR-Cas9-targeting without the presence of an exogenous DNA repair template. After some preliminary tests, it was confirmed that NHEJ was most likely the responsible for the survival of those clones. In this chapter, we attempt to develop a CRISPR-Cas9-based tool independent of exogenous DNA templates, that is, for template-free genome editing. We generate several mutant strains lacking one or combinations of the several genes putatively involved in template-free DNA repair in the bacterium. Using these mutant strains, we confirm that HDR is responsible, thanks to the RecA protein, for the high-fidelity repair of the chromosome. Lastly, we reassess the use of 5-fluorouracil (5-FU) for screening of *upp* mutants in genetic engineering experiments, and we conclude that it is not a suitable compound given its genotoxic properties, which make the assessment of mutations caused by other agents such as CRISPR-Cas9 very hard. We hope our findings to be useful for further investigation of DNA repair pathways in *R. sphaeroides*, as well as for informing future genetic screenings in bacteria that depend on selection schemes like that these based on *upp* mutants.

In Chapter 6, we develop an experimental method to minimize the genome of *E. coli* MG1655 through conjugation with *E. coli* strain ST18. Using a novel type I-C CRISPR system involving the Cas3 nuclease, we first confirmed that the tool can be used in our laboratory to generate deletions of the *lacZ* gene, successfully mediating efficient genome editing of the region. Then, we review the literature for recent genome minimization efforts in the bacterium, and we combine information regarding deletions that have been made into a list of 17 non-essential regions (NERs) bigger than 10 kb. As a proof of concept, we focus on the biggest 6 NERs, and we design 12 plasmids encoding for the CRISPR system targeting each of the selected NERs twice, and we use 3 different antibiotic resistance markers to be able to carry fast iterative genome editing rounds, with the aim of performing up to 3 different large deletions in 4 working days. Further, we develop a multiplex loci PCR (MLPCR) through which we can successfully monitor 9-11 loci simultaneously in a single PCR reaction. Lastly, we select three promising colonies, we isolated their genome and sequenced it through Nanopore sequencing. While we prove the system to work and generate deletions of up to 110 kb, the stacking of deletions was not possible through our method, possibly due to plasmid recalcitrance and stochastic plasmid curing. With some tweaking, we expect our study to be the basis of successful strategies for iterative genome minimization of *E. coli* and other bacterial species that can be conjugated.

7.2 General discussion

7.2.1 Exploring the allure of protein expression engineering

Over the past few decades, significant breakthroughs have continued to propel the field of biotechnology forward. Cutting-edge protein expression tools enable the successful expression of functional proteins not only in synthetic biology laboratory environments, but also in industrial production settings, allowing for the production of functional recombinant proteins at diverse scales with higher efficiency and purity. These advances pave the way for the development of novel therapies, molecular diagnostic tools, industrial products, and groundbreaking leaps in bottom-up synthetic biology. Some of the many useful outcomes stemming from research on protein expression optimization are the use of codon deoptimization to generate attenuated viruses for the development of vaccines[344] including SARS-CoV-2[345], and novel examples of BCD-based protein production systems have already unfolded, enabling the production of recombinant proteins in *Corynebacterium glutamicum*[196] and production of the antibiotic aurachin D in *E. coli*[346].

At present, the field has a relatively good understanding of several crucial factors influencing protein expression, such as the importance of mRNA secondary structures and codon bias on translation initiation[347] and elongation[348], respectively. All this knowledge can be gathered to develop relatively well-informed algorithms[349] to design DNA sequences predicted to render the highest protein production levels. In the end, however, a DNA sequence is not an all-encompassing element that can predictably render a specific protein titer. The cellular and physiological context of any DNA sequence, as well as the growth conditions of its host organism, the size of the bioreactor where its cultivated, and plethora of other factors will have an variable impact on the final levels of expression. While some tools such as the BCD system facilitate the robust scale-up of protein expression, as we demonstrate in Chapter 3, the empirical quantification of protein titers given by a library of DNA sequences is still an imperative to successfully carry projects that depend on sufficient levels of functional protein expression.

7.2.2 Less is not always more

The idea that reducing the size of a genome can lead to improved fitness has captivated scientists for years, as it is easy to imagine how hundreds or thousands of unexpressed genes toll cells with a replicative burden that never pays off. Many benefits have certainly been associated with genome minimized strains; higher transformation efficiencies[320], faster growth[321,325] and improved production yields[325,322] are reported as the main benefits driving the need to perform genome minimization for the optimization of microbial cell factories.

However, numerous studies have concluded that enhancing an organism's fitness simply by decreasing its genome size it is exceedingly difficult to achieve. As researchers continue delving further into the intricacies of genome minimization, a complex reality emerges, challenging the notion that "less is more" in all cases. Generating and stacking large genomic deletions often results in unwanted negative consequences to cellular fitness such as slower growth rate, abnormal cell division, increased sensitivity to stress and decreased feedstock uptake[320,350–353], which counter the initial expectations and aims of the process.

The duality in the outcomes of genome minimization can be attributed to the fact that genetic elements deemed non-essential can unexpectedly play crucial roles in various biological processes. Additionally, several complex epistatic interactions can occur between seemingly non-essential genes, rendering an otherwise efficient process obsolete when one of the parts involved in the system is eliminated. In all, it would be wise to caution against a full adherence to the belief that reducing the genome size will automatically confer benefits. While genome minimization undoubtedly holds promise and has provided valuable insights, especially for the field of bottom-up synthetic biology, it is essential to recognize that the relationship between genome size and fitness is far from linear. In the intricate biological realm, the quest for understanding the delicate balance between genomic complexity and functionality continues to challenge the reductionistic notion that smaller genomes are better.

The development of genome minimization tools is a worthwhile investment, but they are one of the many available methods to elucidate the function and dynamics of biological systems, as well as enabling the exploration of the potential genetic space for more desirable phenotypes. The more knowledge is generated about the function and interaction of unknown genes, however, the closer we are to leaving no (molecular) stone unturned. As we learn more about molecular genetics, the same goals of successful genome minimization projects can be achieved through a combination of rational engineering, stringent rational selection systems and adaptive laboratory evolution.

7.2.3 Future scrutinization of prokaryotic DNA repair pathways

Due to the implications of DNA repair pathways in human cancer, both NHEJ[354] and AEJ[355] have been extensively studied in eukaryotes, their intricate pathways being much better characterized than that of prokaryotes. Throughout the completion of this thesis and while examining the literature on the topic, several inconsistencies have stood out in the endless sea of articles devoted to studying these pathways.

The prokaryotic NHEJ pathways was first predicted by bioinformatic analysis in 2001, when homologs of the eukaryotic pathway were found in *Streptomyces coelicolor*[282]. Since then, several studies and literature reviews have stipulated that NHEJ generally results in small indels, while AEJ, only described in 2010[61], is now accepted to generate larger deletions accompanied by microhomologies. Paradoxically, several studies performed between the discovery of prokaryotic NHEJ and that of AEJ are aimed at observing the outcomes of specifically NHEJ in mutagenesis and DNA repair. Interestingly, several of these studies, which employ either natively or heterologously expressed Ku and LigD proteins, report mutations allegedly generated by NHEJ to be accompanied by microhomologies[54], which are now accepted to be a result of AEJ.

While it is obvious that both pathways cannot be fully orthogonal and independent, studies designed to fully elucidate the exact contributions of NHEJ and AEJ to prokaryotic DNA repair will be instrumental in providing models to explain their interactions. Techniques such as single-molecule tracking[356,357] can be extremely helpful to determine the specific actions that the different enzymes take in the overall process, helping researchers in the field better design genome editing experiments.

7.2.4 The curious case of GMO legislation

Written by Hans Jonas in 1979, *The Imperative of Responsibility* presents strict ideas regarding the effects of technology on the magnitude of impact of human action and invites to "*[…] not compromise the conditions for an indefinite continuation of humanity on Earth"*. Soon becoming what we know today as the precautionary principle, the epistemological approach to innovations advocates for taking preventive measures in the face of uncertain risks to human health or the environment. While it does not call for a complete ban on emerging technologies, it emphasizes careful evaluation, risk assessment, and consideration of alternatives with less associated risk[358]. The principle has been applied successfully to implement several emerging technologies.

While many were always eager to use it since its invention, the development of the internet was not free of controversy, as there were and have always been risks associated to it. Rather than imposing a ban on its world-wide application, precautionary steps were taken, such as the development of data protection laws, encryption technologies, and content moderation policies. These measures aimed to mitigate potential risks while allowing the transformative power of the internet to be harnessed, thereby enabling it to thrive and become an essential tool for communication, information sharing, and global connectivity, while continuously addressing the need for user safety and protection[359,360].

A similar example can be drawn from the development of vaccines; thanks to rigorous testing, exhaustive clinical trials and strict regulatory frameworks, vaccines have been widely adopted as a live-saving technology[361]. While several myths and controversies have been constructed by different groups for different reasons, these have been successfully disproven by science on multiple occasions[362], enabling the development of vaccines to successfully eradicate infectious agents like smallpox and rinderpest[363], as well as limiting the spread and severity of viruses like hepatitis B, human papilloma[364] and the more recent SARS-CoV-2[365]. Future developments in the field are expected to assist protecting humanity against more complex infections like tuberculosis, HIV and malaria[363].

Flourishing from the advent of recombinant DNA technologies, the development of genetically modified organisms (GMOs) and its embedding into the global legal framework poses an interesting case. While the precautionary principle emphasizes careful evaluation and risk assessment, the interpretation and application of this principle in the legislation and regulation of GMOs seems exceptional[366]. Rather than being based on scientific evidence, the relevant regulatory frameworks often seem based on hypothetical risks associated with the technology, being shaped by public concerns independent on the actual scientific consensus on the safety of genetically modified crops. Discussing the risks of GMOs, like any other emerging technology, with non-experts often results in the following infallible argument being wielded: it is impossible to understand all the risks associated to the use of the new technology, and its application should therefore be prevented.

Whereas the regulatory “freeze” on GMO technologies seems at least partially justified, it is so for the wrong reasons. While there is enough knowledge to accurately assess the risk of using certain GMOs for specific uses, it is virtually impossible to predict how any technology will affect our planet and its ecosystems when technological developments are embedded in a morally frugal capitalistic system that advocates for a minimally (if at all) regulated free market. The consequences of growth-oriented market practices can indeed be catastrophic for fragile ecosystems and societies, and governments and legislating entities should focus on ensuring that economic profit does not prevail over human rights[367].

Until policies and regulatory frameworks are developed reflecting the scientifically backed risks of GMOs and protecting society from their abuse in the market, genetic and metabolic engineers are forced to use strategies to circumvent the legislation which arbitrarily determines what is and what is not a GMO. While it obviously leads to genetic modifications, legislations often are permissive regarding the engineering of organisms through the application of chemical mutagens as well as through the use of physical mutagens[368]. Interestingly, data from chapters 5 and 6 included in this thesis support the idea that, following the application of both genome editing tools and chemical mutagens, mutagenesis is caused by non-templated DNA repair pathways like NHEJ and AEJ, which are naturally occurring molecular systems that regulate the occurrence of mutations in organisms throughout the entire realm of life.

The European Union currently considers GMOs "*organisms in which the genetic material (DNA) has been altered in a way that does not occur naturally by mating or natural recombination*". Given that aggressive physical mutagenic agents like gamma irradiation are commonly used within the legal framework of plant breeding in Europe, it is obvious that the previous definition referring to *natural recombination* includes large genomic aberrations caused by man-made radiation. It is therefore plausible to propose the use of either transiently expressed or ribonucleoprotein complexes of CRISPR-Cas tools to generate genetically modified organisms which technically are not GMOs, appealing to the fact that the mechanisms included in the definition of *natural recombination* driven by conventional (and well-accepted) chemical or physical mutagenesis are, in fact, the very same that govern the nature of the mutagenesis facilitated by programmable endonucleases like CRISPR-Cas9, among many others. One can only hope that the appropriate regulatory bodies will eventually reassess their judgement on these technologies, perhaps even developing policies and laws that prevent big corporations from completely dominating the market in different industries. In the current framework, however, the leeway given to genetic engineering technologies is severely limited, and the field will have to do with considerably restricted applications thereof.

7.2.5 Cellular agriculture and precision fermentation to change the mid future

About 100 years ago, the production of 200 grams insulin through the processing of 2 tons of porcine pancreas was a revolutionary technological breakthrough for human health. The same concept, however, would be considerably harder to justify morally today. Technological advances for a more sustainable production of food, materials and compounds provide humanity with not only an opportunity, but with the responsibility of choosing the right process to obtain a specific result. The hour has come for humanity to reassess the legitimacy of strongly rooted supply chains and food systems, particularly those controlled by corporate behemoths resembling monopolies and bolstered by formidable lobbying power. While acknowledging the undeniable significance of conventional agricultural practices, it is imperative to recognize that the unregulated industrialized production of food and feed imposes irreparable repercussions on our planet.

While fermentation technologies have existed for decades, a distinct subset is now emerging under the name of precision fermentation. Leveraging the vast power of microorganisms to produce specific molecules of interest in a controlled and precise manner, it offers the opportunity to move away from conventional production methods for animal-based products and to start producing them in a more ethical and sustainable way. A case similar to that of insulin can be found in the production of collagen. Being the main component of connective tissue, collagen is normally obtained from animal sources like meat and fish, and it finds wide use in the fitness and cosmetics industry. Thanks to precision fermentation, collagen is now available through the animal-free microbial production[369].

Applying the same critical reassessment of our processes, the validity of the current systems for food production must be put into question, and more sustainable alternative methods of production should be developed to satisfy the needs of the fanciful current global population.

7.2.6 One does not simply dictate individual desires: nurturing consumer choices responsibly

The biggest contributor to the emission of greenhouse gasses priming climate change is the abuse of fossil fuels[370]. However, greenhouse gas emissions from the agricultural sectors are estimated to be higher than a fifth of global total emissions, livestock production accounting for nearly 80% of it[371]. While consensus is questionable regarding the growth trends of the global population, it is clear that the demand for meat and animal products is raising in developing countries[372]. Rather than being used to feed more than 3 billion people, over one-third of the total world cereal use is estimated to be used as livestock feed[373] and is grown on arable land that could otherwise be farmed for the production of vegetables to be fed directly to humans or for ecological restoration and preservation. While there are merits in using animals to maintain specific ecosystems, such as minimizing the risk of wildfires and contributing to natural nutrient cycles, the production of animals to feed humans is a thermodynamical tragedy: in any food chain, every step distancing the primary producers from the final trophic level is bound for energy loss, making the allocation of solar energy, farmable land and nutrients into produce to feed livestock a shortsighted ecological mishap.

Having refined the art of embracing cognitive dissonance[374], meat consumers are extremely unlikely to stop consuming animal products because scientists warn about its severe consequences on the environment or even on human health. However, by developing an alternative products that satisfy the needs of the consumer, scientists can nudge customers to do two very important things: to think about the moral and environmental implications of consuming animal products over an alternative, and to actually stop participating in the animal farming industry provided that an alternative is available and preferred. With that in mind, several meat substitute products have been developed throughout history. A meat analog that by 1930 had gathered considerable success was Protose, made from peanuts and wheat gluten[375]. A more recently developed product is Quorn, which derives single-cell mycoprotein from *Fusarium venenatum* into several analog products. Several other commercially successful plant-based meat substitutes have been developed, like the Beyond Burger or the Impossible Burger, the latter containing heme protein, recombinantly produced in yeast, giving the analog burger a very similar appearance to real meat, fostering the acceptance of these products within the animal meat-consuming customer base.

Likely the currently most groundbreaking development in the meat analogue industry is the birth of several companies aiming to produce meat cultured *in* *vitro*, a form of cellular agriculture. Estimated to have costed over $300,000, the first *in vitro* cultured beef burger patty was created in Maastricht University, starting a trend that has seen the technology evolve very quickly, to the point that several start-ups are being developed globally attempting to earn a share of what might be a very profitable market. By combining the *in vitro* culturing of different cell lines and 3D printing technologies, animal-free versions of different types of meat products can be developed. While scalability, overall energetic efficiency and especially the cost of lab grown meat are still to be demonstrated, its appearance in the market is a brilliant technological development at the interface of biotechnology and consumer behavior that promotes discussion and primes consumers to reassess their consumption habits.

7.2.7 Other futuristic developments for the animal-free production of animal products

The first example of commercialized animal-free animal protein for the food industry is that of a whey protein produced for ice-cream by Perfect Day, which has since promoted the appearance of other companies trying to achieve similar feats[376]. While a culinary miracle for some, cheese is simply the result of the coagulation of casein, a protein naturally occurring in mammalian milk. It would seem that the current organisms in which caseins are produced for the animal-free production of cheese are the yeast *Pichia pastoris*, and other fungi from the genera of *Aspergillus* and *Trichoderma*.

While perhaps less ostentatious than the development of *in vitro* meat products, the entrance of animal-free cheese and dairy products into the market seems like a more feasible attempt at making a more sustainable product almost identical to the conventional. Animal-free dairy poses additional advantages versus their traditional counterparts, most notably the suppression of the need for the products to contain lactose, a sugar infamous for causing mild to severe issues associated with its intolerance. Another door opened up with the microbial production of casein and its derivates is the production of dairy products from any animal that has been sequenced; there currently are avid lovers of cow, goat or sheep cheese, but in 50 years from now these might have been overshadowed by the unanticipated richness of the flavors of giraffe, blue whale or even woolly mammoth cheeses.

A similar phenomenon is occurring with eggs and egg-derived products, broadly used in the food and other industries. Onego, a Finnish start-up, has started producing animal-free egg proteins to provide an ethical and sustainable alternative to the practices of the poultry industry by using the fungus *Trichoderma reesei* as a microbial cell factory[377].

Fascinating biotechnological advances are seen in the fields of protein expression optimization and genome engineering, parallel to other breakthroughs taking place within technologies like artificial intelligence. However, their astounding potential to mitigate and solve global challenges is futile when their application disregards the underlying problems that lead to climate change, predominantly driven by corporate actions. Merely utilizing these technologies without addressing the root causes for the present global crisis will yield limited impact. Should regulatory agencies and governments keep failing to effectively monitor and control the application of these technologies and the use of our resources, our planet and its delicate ecosystems will continue to deteriorate, thrusting humanity into an unprecedented crisis, one that can only be surmounted by radically transforming our approach to growth.

References

1. Braidwood, R. J. The Agricultural Revolution. *Sci. Am.* **203**, 130–152 (1960).
2. Liu, L. *et al.* Fermented beverage and food storage in 13,000 y-old stone mortars at Raqefet Cave, Israel: Investigating Natufian ritual feasting. *J. Archaeol. Sci. Rep.* **21**, 783–793 (2018).
3. McGovern, P. E. *et al.* Fermented beverages of pre- and proto-historic China. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 17593–17598 (2004).
4. Salque, M. *et al.* Earliest evidence for cheese making in the sixth millennium BC in northern Europe. *Nature* **493**, 522–525 (2013).
5. Dahm, R. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Hum. Genet.* **122**, 565–581 (2008).
6. Hershey, A. D. & Chase, M. Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage. in *Die Entdeckung der Doppelhelix: Die grundlegenden Arbeiten von Watson, Crick und anderen* (ed. Nickelsen, K.) 121–139 (Springer, 2017). doi:10.1007/978-3-662-47150-0_3.
7. Franklin, R. E. & Gosling, R. G. Molecular Configuration in Sodium Thymonucleate. *Nature* **171**, 740–741 (1953).
8. Wilkins, M. H. F., Stokes, A. R. & Wilson, H. R. Molecular Structure of Nucleic Acids: Molecular Structure of Deoxypentose Nucleic Acids. *Nature* **171**, 738–740 (1953).
9. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738 (1953).
10. Lehman, I. R., Bessman, M. J., Simms, E. S. & Kornberg, A. Enzymatic synthesis of deoxyribonucleic acid. I. Preparation of substrates and partial purification of an enzyme from Escherichia coli. *J. Biol. Chem.* **233**, 163–170 (1958).
11. Crick, F. *What mad pursuit*. (Basic Books, 2008).
12. Crick, F. H. C., Barnett, L., Brenner, S. & Watts-Tobin, R. J. General Nature of the Genetic Code for Proteins. *Nature* **192**, 1227–1232 (1961).
13. Sarabhai, A. S., Stretton, A. O. W., Brenner, S. & Bolle, A. Co-Linearity of the Gene with the Polypeptide Chain. *Nature* **201**, 13–17 (1964).
14. Brenner, S., Jacob, F. & Meselson, M. An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis. *Nature* **190**, 576–581 (1961).

15. Gros, F. *et al.* Unstable Ribonucleic Acid Revealed by Pulse Labelling of Escherichia Coli. *Nature* **190**, 581–585 (1961).

16. Crick, F. Central Dogma of Molecular Biology. *Nature* **227**, 561–563 (1970).

17. Meselson, M. & Yuan, R. DNA Restriction Enzyme from E. coli. *Nature* **217**, 1110–1114 (1968).

18. Cohen, S. N., Chang, A. C. Y., Boyer, H. W. & Helling, R. B. Construction of Biologically Functional Bacterial Plasmids In Vitro. *Proc. Natl. Acad. Sci. U. S. A.* **70**, 3240–3244 (1973).

19. Brock, T. D. & Freeze, H. Thermus aquaticus gen. n. and sp. n., a Nonsporulating Extreme Thermophile. *J. Bacteriol.* **98**, 289–297 (1969).

20. Chien, A., Edgar, D. B. & Trela, J. M. Deoxyribonucleic acid polymerase from the extreme thermophile Thermus aquaticus. *J. Bacteriol.* **127**, 1550–1557 (1976).

21. Rabinow, P. *Making PCR: a story of biotechnology*. (University of Chicago Press, 1996).

22. Saiki, R. K. *et al.* Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**, 1350–1354 (1985).

23. Mullis, K. B. & Faloona, F. A. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol.* **155**, 335–350 (1987).

24. Saiki, R. K. *et al.* Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**, 487–491 (1988).

25. Wendt, D. Two tons of pig parts: Making insulin in the 1920s. *National Museum of American History* (2013).

26. Goeddel, D. V. *et al.* Expression in Escherichia coli of chemically synthesized genes for human insulin. *Proc. Natl. Acad. Sci.* **76**, 106–110 (1979).

27. Nieuwkoop, T., Finger-Bou, M., van der Oost, J. & Claassens, N. J. The Ongoing Quest to Crack the Genetic Code for Protein Production. *Mol. Cell* **80**, (2020).

28. Quax, T. E. F., Claassens, N. J., Söll, D. & van der Oost, J. Codon Bias as a Means to Fine-Tune Gene Expression. *Mol. Cell* **59**, 149–161 (2015).

29. de Boer, C. G. *et al.* Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat. Biotechnol.* **38**, 56–65 (2020).

30. Lenstra, T. L., Rodriguez, J., Chen, H. & Larson, D. R. Transcription Dynamics in Living Cells. *Annu. Rev. Biophys.* **45**, 25–47 (2016).

31. Sharp, P. M. & Li, W.-H. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295 (1987).

32. Hanson, G. & Coller, J. Translation and Protein Quality Control: Codon optimality, bias and usage in translation and mRNA decay. *Nat. Rev. Mol. Cell Biol.* **19**, 20–30 (2018).

33. Ikemura, T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2**, 13–34 (1985).

34. Bourret, J., Alizon, S. & Bravo, I. G. COUSIN (COdon Usage Similarity INdex): A Normalized Measure of Codon Usage Preferences. *Genome Biol. Evol.* **11**, 3523–3528 (2019).

35. Parret, A. H., Besir, H. & Meijers, R. Critical reflections on synthetic gene design for recombinant protein expression. *Curr. Opin. Struct. Biol.* **38**, 155–162 (2016).

36. Nieuwkoop, T., Claassens, N. J. & van der Oost, J. Improved protein production and codon optimization analyses in Escherichia coli by bicistronic design. *Microb. Biotechnol.* (2018) doi:10.1111/1751-7915.13332.

37. Buhr, F. *et al.* Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. *Mol. Cell* **61**, 341–351 (2016).

38. Claassens, N. J. N. J. *et al.* Improving heterologous membrane protein production in Escherichia coli by combining transcriptional tuning and codon usage algorithms. *PloS One* **12**, e0184355 (2017).

39. Mohanty, B. K. & Kushner, S. R. Regulation of mRNA Decay in Bacteria. *Annu. Rev. Microbiol.* **70**, 25–44 (2016).

40. Urtecho, G. *et al.* Genome-wide Functional Characterization of Escherichia coli Promoters and Regulatory Elements Responsible for their Function. 2020.01.04.894907 Preprint at https://doi.org/10.1101/2020.01.04.894907 (2020).

41. Wagner, S. *et al.* Consequences of Membrane Protein Overexpression in Escherichia coli*. *Mol. Cell. Proteomics* **6**, 1527–1550 (2007).

42. Schlegel, S., Hjelm, A., Baumgarten, T., Vikström, D. & de Gier, J.-W. Bacterial-based membrane protein production. *Biochim. Biophys. Acta BBA - Mol. Cell Res.* **1843**, 1739–1749 (2014).

43. Claassens, N. J. *et al.* Bicistronic Design-Based Continuous and High-Level Membrane Protein Production in Escherichia coli. *ACS Synth. Biol.* **8**, 1685–1690 (2019).

44. Urnov, F. D., Rebar, E. J., Holmes, M. C., Zhang, H. S. & Gregory, P. D. Genome editing with engineered zinc finger nucleases. *Nat. Rev. Genet.* **11**, 636–646 (2010).

45. Voytas, D. F. & Bogdanove, a. J. TAL Effectors: Customizable Proteins for DNA Targeting. *Science* **333**, 1843–1846 (2011).

46. Barrangou, R. *et al.* CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science* **315**, 1709–1712 (2007).

47. Brouns, S. J. J. *et al.* Small Crispr Rnas Guide Antiviral Defense in Prokaryotes. *Science* **321**, 960–964 (2008).

48. Horvath, P. & Barrangou, R. CRISPR/Cas, the immune system of Bacteria and Archaea. *Science* **327**, 167–170 (2010).

49. Garneau, J. E. *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67–71 (2010).

50. Mougiakos, I., Bosma, E. F., de Vos, W. M., van Kranenburg, R. & van der Oost, J. Next Generation Prokaryotic Engineering: The CRISPR-Cas Toolkit. *Trends Biotechnol.* **34**, 575–587 (2016).

51. Mougiakos, I., Bosma, E. F., Ganguly, J., van der Oost, J. & van Kranenburg, R. Hijacking CRISPR-Cas for high-throughput bacterial metabolic engineering: advances and prospects. *Curr. Opin. Biotechnol.* **50**, 146–157 (2018).

52. Xu, T. *et al.* Efficient genome editing in clostridium cellulolyticum via CRISPR-Cas9 nickase. *Appl. Environ. Microbiol.* **81**, 4423–4431 (2015).

53. Cui, L. & Bikard, D. Consequences of Cas9 cleavage in the chromosome of Escherichia coli. *Nucleic Acids Res.* **44**, 4243–4251 (2016).

54. Finger-Bou, M., Orsi, E., van der Oost, J. & Staals, R. H. J. CRISPR with a Happy Ending: Non-Templated DNA Repair for Prokaryotic Genome Engineering. *Biotechnol. J.* **1900404**, e1900404 (2020).

55. Wigley, D. B. Bacterial DNA repair: Recent insights into the mechanism of RecBCD, AddAB and AdnAB. *Nat. Rev. Microbiol.* **11**, 9–13 (2013).

56. Sun, L. *et al.* Effective polyploidy causes phenotypic delay and influences bacterial evolvability. *PLOS Biol.* 1–24 (2018) doi:10.1101/226654.

57. McGrew, D. A. & Knight, K. L. Molecular Design and Functional Organization of the RecA Protein. *Crit. Rev. Biochem. Mol. Biol.* **38**, 385–432 (2003).

58. Pitcher, R. S. *et al.* Mycobacteriophage Exploit NHEJ to Facilitate Genome Circularization. *Mol. Cell* **23**, 743–748 (2006).

59. Aniukwu, J., Glickman, M. S. & Shuman, S. The pathways and outcomes of mycobacterial NHEJ depend on the structure of the broken DNA ends. *Genes Dev.* **22**, 512–527 (2008).

60. Bowater, R. & Doherty, A. J. Making ends meet: Repairing breaks in bacterial DNA by non-homologous end-joining. *PLoS Genet.* **2**, 93–99 (2006).

61. Chayot, R., Montagne, B., Mazel, D. & Ricchetti, M. An end-joining repair mechanism in Escherichia coli. *Proc. Natl. Acad. Sci.* **107**, 2141–2146 (2010).

62. Emond, S. *et al.* Accessing unexplored regions of sequence space in directed enzyme evolution via insertion/deletion mutagenesis. *Nat. Commun.* **11**, 3469 (2020).

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

64. Pelicic, V., Reyrat, J. M. & Gicquel, B. Generation of unmarked directed mutations in mycobacteria, using sucrose counter-selectable suicide vectors. *Mol. Microbiol.* **20**, 919–925 (1996).

65. Szostková, M. & Horáková, D. The effect of plasmid DNA sizes and other factors on electrotransformation of Escherichia coli JM109. *Bioelectrochem. Bioenerg.* **47**, 319–323 (1998).

66. Csörgő, B. *et al.* A compact Cascade–Cas3 system for targeted genome engineering. *Nat. Methods* (2020) doi:10.1038/s41592-020-00980-w.

67. Lopes, M. S. G. Engineering biological systems toward a sustainable bioeconomy. *J. Ind. Microbiol. Biotechnol.* **42**, 813–838 (2015).

68. Aguilar, A., Twardowski, T. & Wohlgemuth, R. Bioeconomy for Sustainable Development. *Biotechnol. J.* **14**, 1800638 (2019).

69. Tylecote, A. Biotechnology as a new techno-economic paradigm that will help drive the world economy and mitigate climate change. *Res. Policy* **48**, 858–868 (2019).

70. Pollard, D. J. & Woodley, J. M. Biocatalysis for pharmaceutical intermediates: the future is now. *Trends Biotechnol.* **25**, 66–73 (2007).

71. Sauer, M. & Mattanovich, D. Construction of microbial cell factories for industrial bioprocesses. *J. Chem. Technol. Biotechnol.* **87**, 445–450 (2012).

72. Lim, J. Y., Yoon, J. W. & Hovde, C. J. A Brief Overview of Escherichia coli O157:H7 and Its Plasmid O157. *J. Microbiol. Biotechnol.* **20**, 5–14 (2010).

73. Blattner, F. R. *et al.* The complete genome sequence of Escherichia coli K-12. *Science* **277**, 1453–1462 (1997).

74. Durfee, T. *et al.* The complete genome sequence of Escherichia coli DH10B: Insights into the biology of a laboratory workhorse. *J. Bacteriol.* **190**, 2597–2606 (2008).

75. Antonovsky, N. *et al.* Sugar Synthesis from CO2 in Escherichia coli. *Cell* **166**, 115–125 (2016).

76. Yishai, O., Goldbach, L., Tenenboim, H., Lindner, S. N. & Bar-Even, A. Engineered Assimilation of Exogenous and Endogenous Formate in *Escherichia coli*. *ACS Synth. Biol.* acssynbio.7b00086 (2017) doi:10.1021/acssynbio.7b00086.

77. Yeates, T. O., Komiya, H., Rees, D. C., Allen, J. P. & Feher, G. Structure of the reaction center from Rhodobacter sphaeroides R-26: membrane-protein interactions. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 6438–6442 (1987).

78. Kiley, P. J. & Kaplan, S. Molecular genetics of photosynthetic membrane biosynthesis in Rhodobacter sphaeroides. *Microbiol. Rev.* **52**, 50–69 (1988).

79. Dufour, Y. S., Landick, R. & Donohue, T. J. Organization and Evolution of the Biological Response to Singlet Oxygen Stress. *J. Mol. Biol.* **383**, 713–730 (2008).

80. Berghoff, B. A., Glaeser, J., Sharma, C. M., Vogel, J. & Klug, G. Photooxidative stress-induced and abundant small RNAs in Rhodobacter sphaeroides. *Mol. Microbiol.* **74**, 1497–1512 (2009).

81. Armitage, J. P. & Macnab, R. M. Unidirectional, intermittent rotation of the flagellum of Rhodobacter sphaeroides. *J. Bacteriol.* **169**, 514–518 (1987).

82. Armitage, J. P. & Schmitt, R. Bacterial chemotaxis: Rhodobacter sphaeroides and Sinorhizobium meliloti - Variations on a theme? *Microbiology* **143**, 3671–3682 (1997).

83. Orsi, E. *et al.* Metabolic flux ratio analysis by parallel 13C labeling of isoprenoid biosynthesis in Rhodobacter sphaeroides. *Metab. Eng.* **57**, 228–238 (2020).

84. Orsi, E. *et al.* Functional replacement of isoprenoid pathways in Rhodobacter sphaeroides. *Microb. Biotechnol.* **13**, 1082–1093 (2020).

85. Orsi, E., Beekwilder, J., Eggink, G., Kengen, S. W. M. & Weusthuis, R. A. The transition of Rhodobacter sphaeroides into a microbial cell factory. *Biotechnol. Bioeng.* **118**, 531–541 (2021).

86. Kim, S. J. *et al.* Translational tuning optimizes nascent protein folding in cells. *Science* **348**, 444–448 (2015).

87. Zhou, M. *et al.* Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature* **495**, 111–115 (2013).

88. dos Reis, M., Wernisch, L. & Savva, R. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. *Nucleic Acids Res.* **31**, 6976–6985 (2003).

89. Cambray, G., Guimaraes, J. C. & Arkin, A. P. Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in Escherichia coli. *Nat. Biotechnol.* **36**, 1005–1015 (2018).

90. Buschauer, R. *et al.* The Ccr4-Not complex monitors the translating ribosome for codon optimality. *Science* **368**, (2020).

91. Henderson, K. L. *et al.* Mechanism of transcription initiation and promoter escape by *E . coli* RNA polymerase. *Proc. Natl. Acad. Sci.* **114**, (2017).

92. Winkelman, J. T. *et al.* Multiplexed protein-DNA cross-linking: Scrunching in transcription start site selection. *Science* **351**, 1090–1093 (2016).

93. Lee, J. & Borukhov, S. Bacterial RNA Polymerase-DNA Interaction—The Driving Force of Gene Expression and the Target for Drug Action. *Front. Mol. Biosci.* **3**, (2016).

94. Wade, J. T. & Struhl, K. The transition from transcriptional initiation to elongation. *Curr. Opin. Genet. Dev.* **18**, 130–136 (2008).

95. Yona, A. H., Alm, E. J. & Gore, J. Random sequences rapidly evolve into de novo promoters. *Nat. Commun.* **9**, (2018).

96. Levo, M. *et al.* Systematic Investigation of Transcription Factor Activity in the Context of Chromatin Using Massively Parallel Binding and Expression Assays. *Mol. Cell* **65**, 604-617.e6 (2017).

97. Urtecho, G., Tripp, A. D., Insigne, K. D., Kim, H. & Kosuri, S. Systematic Dissection of Sequence Elements Controlling σ70 Promoters Using a Genomically Encoded Multiplexed Reporter Assay in Escherichia coli. *Biochemistry* **58**, 1539–1551 (2019).

98. Fu, J., Dang, Y., Counter, C. & Liu, Y. Codon usage regulates human KRAS expression at both transcriptional and translational levels. *J. Biol. Chem.* **293**, 17929–17940 (2018).

99. Newman, Z. R., Young, J. M., Ingolia, N. T. & Barton, G. M. Differences in codon bias and GC content contribute to the balanced expression of TLR7 and TLR9. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E1362–E1371 (2016).

100. Zhou, Z. *et al.* Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc. Natl. Acad. Sci.* **113**, E6117–E6125 (2016).

101. Schmid, M. & Jensen, T. H. Controlling nuclear RNA levels. *Nat. Rev. Genet.* **19**, 518–529 (2018).

102. Boël, G. *et al.* Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature* **529**, 358–363 (2016).

103. Lahtvee, P.-J. *et al.* Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast. *Cell Syst.* **4**, 495-504.e5 (2017).

104. Presnyak, V. *et al.* Codon Optimality Is a Major Determinant of mRNA Stability. *Cell* **160**, 1111–1124 (2015).

105. Menendez-Gil, P. *et al.* Differential evolution in 3′UTRs leads to specific gene expression in Staphylococcus. *Nucleic Acids Res.* **48**, 2544–2563 (2020).

106. Zhao, J.-P., Zhu, H., Guo, X.-P. & Sun, Y.-C. AU-rich long 3′ untranslated region regulates gene expression in bacteria. *Front. Microbiol.* **9**, (2018).

107. Mugridge, J. S., Coller, J. & Gross, J. D. Structural and molecular mechanisms for the control of eukaryotic 5′–3′ mRNA decay. *Nat. Struct. Mol. Biol.* **25**, 1077–1085 (2018).

108. Tian, B. & Manley, J. L. Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.* **18**, 18–30 (2017).

109. Bazzini, A. A. *et al.* Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition. *EMBO J.* **35**, 2087–2103 (2016).

110. Burow, D. A. *et al.* Attenuated Codon Optimality Contributes to Neural-Specific mRNA Decay in Drosophila. *Cell Rep.* **24**, 1704–1712 (2018).

111. Forrest, M. E. *et al.* Codon and amino acid content are associated with mRNA stability in mammalian cells. *PLOS ONE* **15**, e0228730 (2020).

112. Harigaya, Y. & Parker, R. Analysis of the association between codon optimality and mRNA stability in Schizosaccharomyces pombe. *BMC Genomics* **17**, 895 (2016).

113. Hia, F. *et al.* Codon bias confers stability to human mRNAs. *EMBO Rep.* **20**, (2019).

114. Jeacock, L., Faria, J. & Horn, D. Codon usage bias controls mRNA and protein abundance in trypanosomatids. *eLife* **7**, e32496 (2018).

115. Mishima, Y. & Tomari, Y. Codon Usage and 3′ UTR Length Determine Maternal mRNA Stability in Zebrafish. *Mol. Cell* **61**, 874–885 (2016).

116. Narula, A., Ellis, J., Taliaferro, J. M. & Rissland, O. S. Coding regions affect mRNA stability in human cells. *RNA* **25**, 1751–1764 (2019).

117. de Freitas Nascimento, J., Kelly, S., Sunter, J. & Carrington, M. Codon choice directs constitutive mRNA levels in trypanosomes. *eLife* **7**, e32467 (2018).

118. Wu, Q., Zhao, H. & Zhang, Z. Long-term role of cooling the underlying permafrost of the crushed rock structure embankment along the Qinghai–Xizang railway. 1-12 (2019).

119. Radhakrishnan, A. *et al.* The DEAD-Box Protein Dhh1p Couples mRNA Decay and Translation by Monitoring Codon Optimality. *Cell* **167**, 122-132.e9 (2016).

120. Webster, M. W. *et al.* mRNA Deadenylation Is Coupled to Translation Rates by the Differential Activities of Ccr4-Not Nucleases. *Mol. Cell* **70**, 1089-1100.e8 (2018).

121. O'Reilly, F. J. *et al.* In-cell architecture of an actively transcribing-translating expressome. *Science* **369**, 554–557 (2020).

122. Johnson, G. E., Lalanne, J.-B., Peters, M. L. & Li, G.-W. Functionally uncoupled transcription–translation in Bacillus subtilis. *Nature* **585**, 124–128 (2020).

123. Mittal, P., Brindle, J., Stephen, J., Plotkin, J. B. & Kudla, G. Codon usage influences fitness through RNA toxicity. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 8639–8644 (2018).

124. Shine, J. & Dalgarno, L. The 3′-Terminal Sequence of Escherichia coli 16S Ribosomal RNA: Complementarity to Nonsense Triplets and Ribosome Binding Sites. *Proc. Natl. Acad. Sci.* **71**, 1342–1346 (1974).

125. Kozak, M. Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes. *Nucleic Acids Res.* **9**, 5233–5252 (1981).

126. Leppek, K., Das, R. & Barna, M. Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nat. Rev. Mol. Cell Biol.* **19**, 158–174 (2018).

127. De Nijs, Y., De Maeseneire, S. L. & Soetaert, W. K. 5′ untranslated regions: the next regulatory sequence in yeast synthetic biology. *Biol. Rev.* **95**, 517–529 (2020).

128. Goodman, D. B., Church, G. M. & Kosuri, S. Causes and Effects of N-Terminal Codon Bias in Bacterial Genes. *Science* **342**, 475–479 (2013).

129. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-Sequence Determinants of Gene Expression in Escherichia coli. *Science* **324**, 255–258 (2009).

130. Borujeni, A. E. *et al.* Precise quantification of translation inhibition by mRNA structures that overlap with the ribosomal footprint in N-terminal coding sequences. *Nucleic Acids Res.* **45**, 5437–5448 (2017).

131. Bhattacharyya, S. *et al.* Accessibility of the Shine-Dalgarno Sequence Dictates N-Terminal Codon Bias in E. coli. *Mol. Cell* **70**, 894-905.e5 (2018).

132. Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E. & Weeks, K. M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods* **11**, 959–965 (2014).

133. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**, 701–705 (2014).

134. Mustoe, A. M. *et al.* Pervasive Regulatory Functions of mRNA Structure Revealed by High-Resolution SHAPE Probing. *Cell* **173**, 181-195.e18 (2018).

135. Espah Borujeni, A. & Salis, H. M. Translation Initiation is Controlled by RNA Folding Kinetics via a Ribosome Drafting Mechanism. *J. Am. Chem. Soc.* **138**, 7016–7023 (2016).

136. Chu, D. *et al.* Translation elongation can control translation initiation on eukaryotic mRNAs. *EMBO J.* **33**, 21–34 (2014).

137. Baez, W. D. *et al.* Global analysis of protein synthesis in Flavobacterium johnsoniae reveals the use of Kozak-like sequences in diverse bacteria. *Nucleic Acids Res.* **47**, 10477–10488 (2019).

138. Komarova, E. S. *et al.* Influence of the spacer region between the Shine–Dalgarno box and the start codon for fine-tuning of the translation efficiency in Escherichia coli. *Microb. Biotechnol.* **13**, 1254–1261 (2020).

139. Saito, K., Green, R. & Buskirk, A. R. Translational initiation in E. coli occurs at the correct sites genome-wide in the absence of mRNA-rRNA base-pairing. *eLife* **9**, 1–19 (2020).

140. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* **324**, 218–223 (2009).

141. Charneski, C. A. & Hurst, L. D. Positively Charged Residues Are the Major Determinants of Ribosomal Velocity. *PLOS Biol.* **11**, e1001508 (2013).

142. Gardin, J. *et al.* Measurement of average decoding rates of the 61 sense codons in vivo. *eLife* **3**, e03735 (2014).

143. Li, G.-W., Burkhardt, D., Gross, C. & Weissman, J. S. Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. *Cell* **157**, 624–635 (2014).

144. Quax, T. E. F. *et al.* Differential Translation Tunes Uneven Production of Operon-Encoded Proteins. *Cell Rep.* **4**, 938–944 (2013).

145. Weinberg, D. E. *et al.* Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. *Cell Rep.* **14**, 1787–1799 (2016).

146. Yu, C.-H. *et al.* Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. *Mol. Cell* **59**, 744–754 (2015).

147. Chekulaeva, M. & Landthaler, M. Eyes on Translation. *Mol. Cell* **63**, 918–925 (2016).

148. Yan, X., Hoek, T. A., Vale, R. D. & Tanenbaum, M. E. Dynamics of Translation of Single mRNA Molecules In Vivo. *Cell* **165**, 976–989 (2016).

149. Yang, Q. *et al.* eRF1 mediates codon usage effects on mRNA translation efficiency through premature termination at rare codons. *Nucleic Acids Res.* **47**, 9243–9258 (2019).

150. Zhao, F., Yu, C. & Liu, Y. Codon usage regulates protein structure and function by affecting translation elongation speed in Drosophila cells. *Nucleic Acids Res.* **45**, 8484–8492 (2017).

151. Burkhardt, D. H. *et al.* Operon mRNAs are organized into ORF-centric structures that predict translation efficiency. *eLife* **6**, 1–23 (2017).

152. Takyar, S., Hickerson, R. P., Noller, H. F., Translocation, S. & Hall, C. mRNA Helicase Activity of the Ribosome. *Cell* **120**, 49–58 (2005).

153. Fu, J. *et al.* Codon usage affects the structure and function of the Drosophila circadian clock protein PERIOD. *Genes Dev.* **30**, 1761–1775 (2016).

154. Zhou, M., Wang, T., Fu, J., Xiao, G. & Liu, Y. Nonoptimal codon usage influences protein structure in intrinsically disordered regions. *Mol. Microbiol.* **97**, 974–987 (2015).

155. Pechmann, S. & Frydman, J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.* **20**, 237–243 (2013).

156. Chaney, J. L. *et al.* Widespread position-specific conservation of synonymous rare codons within coding sequences. *PLOS Comput. Biol.* **13**, e1005531 (2017).

157. Tuller, T. & Zur, H. Multiple roles of the coding sequence 5' end in gene expression regulation. *Nucleic Acids Res.* **43**, 13–28 (2015).

158. Tuller, T. *et al.* An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. *Cell* **141**, 344–354 (2010).

159. Kelsic, E. D. *et al.* RNA Structural Determinants of Optimal Codons Revealed by MAGE-Seq. *Cell Syst.* **3**, 563-571.e6 (2016).

160. Frumkin, I. *et al.* Gene Architectures that Minimize Cost of Gene Expression. *Mol. Cell* **65**, 142–153 (2017).

161. Buchan, J. R., Aucott, L. S. & Stansfield, I. tRNA properties help shape codon pair preferences in open reading frames. *Nucleic Acids Res.* **34**, 1015–1027 (2006).

162. Gutman, G. A. & Hatfield, G. W. Nonrandom utilization of codon pairs in Escherichia coli. *Proc. Natl. Acad. Sci.* **86**, 3699–3703 (1989).

163. Gamble, C. E., Brule, C. E., Dean, K. M., Fields, S. & Grayhack, E. J. Adjacent Codons Act in Concert to Modulate Translation Efficiency in Yeast. *Cell* **166**, 679–690 (2016).

164. Tesina, P. *et al.* Molecular mechanism of translational stalling by inhibitory codon combinations and poly(A) tracts. *EMBO J.* **39**, (2020).

165. Kunec, D. & Osterrieder, N. Codon Pair Bias Is a Direct Consequence of Dinucleotide Bias. *Cell Rep.* **14**, 55–67 (2016).

166. Groenke, N. *et al.* Mechanism of Virus Attenuation by Codon Pair Deoptimization. *Cell Rep.* **31**, 107586 (2020).

167. Li, G.-W., Oh, E. & Weissman, J. S. The anti-Shine–Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* **484**, 538–541 (2012).

168. Mohammad, F., Woolstenhulme, C. J., Green, R. & Buskirk, A. R. Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling. *Cell Rep.* **14**, 686–694 (2016).

169. Hockenberry, A. J., Jewett, M. C., Amaral, L. A. N. & Wilke, C. O. Within-Gene Shine–Dalgarno Sequences Are Not Selected for Function. *Mol. Biol. Evol.* **35**, 2487–2498 (2018).

170. Verma, M. *et al.* A short translational ramp determines the efficiency of protein synthesis. *Nat. Commun.* **10**, (2019).

171. Huter, P. *et al.* Structural Basis for Polyproline-Mediated Ribosome Stalling and Rescue by the Translation Elongation Factor EF-P. *Mol. Cell* **68**, 515-527.e6 (2017).

172. Wilson, D. N., Arenz, S. & Beckmann, R. Translation regulation via nascent polypeptide-mediated ribosome stalling. *Curr. Opin. Struct. Biol.* **37**, 123–133 (2016).

173. Navon, S. P. *et al.* Amino acid sequence repertoire of the bacterial proteome and the occurrence of untranslatable sequences. *Proc. Natl. Acad. Sci.* **113**, 7166–7170 (2016).

174. Chou, H.-J., Donnard, E., Gustafsson, H. T., Garber, M. & Rando, O. J. Transcriptome-wide Analysis of Roles for tRNA Modifications in Translational Regulation. *Mol. Cell* **68**, 978-992.e4 (2017).

175. Kimura, S., Srisuknimit, V. & Waldor, M. K. Probing the diversity and regulation of tRNA modifications. *Curr. Opin. Microbiol.* **57**, 41–48 (2020).

176. Nedialkova, D. D. & Leidel, S. A. Optimization of Codon Translation Rates via tRNA Modifications Maintains Proteome Integrity. *Cell* **161**, 1606–1618 (2015).

177. Arango, D. *et al.* Acetylation of Cytidine in mRNA Promotes Translation Efficiency. *Cell* **175**, 1872-1886.e24 (2018).

178. Choi, J. *et al.* N6-methyladenosine in mRNA disrupts tRNA selection and translation-elongation dynamics. *Nat. Struct. Mol. Biol.* **23**, 110–115 (2016).

179. Zhao, B. S., Roundtree, I. A. & He, C. Post-transcriptional gene regulation by mRNA modifications. *Nat. Rev. Mol. Cell Biol.* **18**, 31–42 (2017).

180. Drummond, D. A. & Wilke, C. O. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. *Cell* **134**, 341–352 (2008).

181. Mordret, E. *et al.* Systematic Detection of Amino Acid Substitutions in Proteomes Reveals Mechanistic Basis of Ribosome Errors and Selection for Translation Fidelity. *Mol. Cell* **75**, 427-441.e5 (2019).

182. Traverse, C. C. & Ochman, H. Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. *Proc. Natl. Acad. Sci.* **113**, 3311–3316 (2016).

183. Wan, J., Gao, X., Mao, Y., Zhang, X. & Qian, S.-B. A coding sequence-embedded principle governs translational reading frame fidelity. *Research* **2018**, (2018).

184. Arribere, J. A. *et al.* Translation readthrough mitigation. *Nature* **534**, 719–723 (2016).

185. Fleming, I. & Cavalcanti, A. R. O. Selection for tandem stop codons in ciliate species with reassigned stop codons. *PLOS ONE* **14**, e0225804 (2019).

186. Taniguchi, Y. *et al.* Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. *Science* **329**, 533–538 (2010).

187. Gould, N., Hendy, O. & Papamichail, D. Computational tools and algorithms for designing customized synthetic genes. *Front. Bioeng. Biotechnol.* **2**, (2014).

188. Angov, E., Hillier, C. J., Kincaid, R. L. & Lyon, J. A. Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host. *PLoS ONE* **3**, 1–10 (2008).

189. Hess, A.-K., Saffert, P., Liebeton, K. & Ignatova, Z. Optimization of Translation Profiles Enhances Protein Expression and Solubility. *PLOS ONE* **10**, e0127039 (2015).

190. Vasquez, K. A., Hatridge, T. A., Curtis, N. C. & Contreras, L. M. Slowing Translation between Protein Domains by Increasing Affinity between mRNAs

and the Ribosomal Anti-Shine–Dalgarno Sequence Improves Solubility. *ACS Synth. Biol.* **5**, 133–145 (2016).

191. Bonde, M. T. *et al.* Predictable tuning of protein expression in bacteria. *Nat. Methods* **13**, 233–236 (2016).

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

193. Salis, H. M., Mirsky, E. A. & Voigt, C. A. Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.* **27**, 946–950 (2009).

194. Ki, M.-R. & Pack, S. P. Fusion tags to enhance heterologous protein expression. *Appl. Microbiol. Biotechnol.* **104**, 2411–2425 (2020).

195. Mutalik, V. K. *et al.* Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods* **10**, 354–60 (2013).

196. Sun, M. *et al.* Enhanced production of recombinant proteins in Corynebacterium glutamicum by constructing a bicistronic gene expression system. *Microb. Cell Factories* **19**, 113 (2020).

197. Ding, W. *et al.* Engineering the 5′ UTR-Mediated Regulation of Protein Abundance in Yeast Using Nucleotide Sequence Activity Relationships. *ACS Synth. Biol.* **7**, 2709–2714 (2018).

198. Petersen, S. D. *et al.* Modular 5'-UTR hexamers for context-independent tuning of protein expression in eukaryotes. *Nucleic Acids Res.* **46**, e127 (2018).

199. Weenink, T., van der Hilst, J., McKiernan, R. M. & Ellis, T. Design of RNA hairpin modules that predictably tune translation in yeast. *Synth. Biol.* **3**, ysy019 (2018).

200. Cheng, J. K., Morse, N. J., Wagner, J. M., Tucker, S. K. & Alper, H. S. Design and Evaluation of Synthetic Terminators for Regulating Mammalian Cell Transgene Expression. *ACS Synth. Biol.* **8**, 1263–1275 (2019).

201. Curran, K. A. *et al.* Short Synthetic Terminators for Improved Heterologous Gene Expression in Yeast. *ACS Synth. Biol.* **4**, 824–832 (2015).

202. Perriman, R. & Ares Jr., M. Circular mRNA can direct translation of extremely long repeating- sequence proteins in vivo. *RNA* **4**, 1047–1054 (1998).

203. Wesselhoeft, R. A., Kowalski, P. S. & Anderson, D. G. Engineering circular RNA for potent and stable translation in eukaryotic cells. *Nat. Commun.* **9**, (2018).

204. Costello, A., Lao, N. T., Barron, N. & Clynes, M. Reinventing the Wheel: Synthetic Circular RNAs for Mammalian Cell Engineering. *Trends Biotechnol.* **38**, 217–230 (2020).

205. Rennig, M. *et al.* No Title. *ACS Synth. Biol.* (2017) doi:10.1021/acssynbio.7b00200.

206. de Jongh, R. P. H., van Dijk, A. D. J., Julsing, M. K., Schaap, P. J. & de Ridder, D. Designing Eukaryotic Gene Expression Regulation Using Machine Learning. *Trends Biotechnol.* **38**, 191–201 (2020).

207. Höllerer, S. *et al.* Large-scale DNA-based phenotypic recording and deep learning enable highly accurate sequence-function mapping. *Nat. Commun.* **11**, (2020).

208. Cuperus, J. T. *et al.* Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences. *Genome Res.* **27**, 2015–2024 (2017).

209. Decoene, T., Peters, G., De Maeseneire, S. L. & De Mey, M. Toward Predictable 5′UTRs in Saccharomyces cerevisiae: Development of a yUTR Calculator. *ACS Synth. Biol.* **7**, 622–634 (2018).

210. Schlegel, S., Hjelm, A., Baumgarten, T., Vikström, D. & de Gier, J.-W. Bacterial-based membrane protein production. *Biochim. Biophys. Acta* **1843**, 1739–49 (2014).

211. Wagner, S. *et al.* Consequences of Membrane Protein Overexpression in *Escherichia coli*. *Mol. Cell. Proteomics* **6**, 1527–1550 (2007).

212. Miroux, B. & Walker, J. E. Over-production of proteins in *Escherichia coli*: mutant hosts that allow synthesis of some membrane proteins and globular proteins at high levels. *J. Mol. Biol.* **260**, 289–98 (1996).

213. Narayanan, A., Ridilla, M. & Yernool, D. A. Restrained expression , a method to overproduce toxic membrane proteins by exploiting operator – repressor interactions. *Protein Sci.* **20**, 51–61 (2011).

214. Schlegel, S. *et al.* Optimizing Membrane Protein Overexpression in the Escherichia coli strain Lemo21(DE3). *J Mol Biol* **423**, 648–659 (2012).

215. Wagner, S. *et al.* Tuning Escherichia coli for membrane protein overexpression. *Proc. Natl. Acad. Sci.* **105**, 14371–14376 (2008).

216. Hjelm, A. *et al.* Chapter 24 Optimizing E. coli-Based Membrane Protein Production Using Lemo21(DE3) and GFP-Fusions. *Methods Mol. Biol.* **1033**,**.**

217. Schlegel, S., Genevaux, P. & de Gier, J.-W. De-convoluting the Genetic Adaptations of *E. coli* C41(DE3) in Real Time Reveals How Alleviating Protein Production Stress Improves Yields. *Cell Rep.* **10**, 1758–1766 (2015).

218. Kuipers, G. *et al.* The tunable pReX expression vector enables optimizing the T7-based production of membrane and secretory proteins in E. coli. *Microb. Cell Factories* **16**, (2017).

219. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-Sequence Determinants of Gene Expression in *Escherichia coli*. *Science* **324**, 255–258 (2009).

220. Cambray, G., Guimaraes, J. C. & Arkin, A. P. Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in *Escherichia coli*. *Nat. Biotechnol.* **36**, 1005–1015 (2018).

221. Nieuwkoop, T., Claassens, N. J. & van der Oost, J. Improved protein production and codon optimization analyses in *Escherichia coli* by bicistronic design. *Microb. Biotechnol.* **12**, 173–179 (2019).

222. Nørholm, M. H. H. *et al.* Improved production of membrane proteins in *Escherichia coli* by selective codon substitutions. *FEBS Lett.* **587**, 2352–8 (2013).

223. Mirzadeh, K. *et al.* Enhanced protein production in *Escherichia coli* by optimization of cloning scars at the vector:coding sequence junction. *ACS Synth. Biol.* **4**, 959–965 (2015).

224. Kim, H. S. *et al.* Translation levels control multi-spanning membrane protein expression. *PloS One* **7**, e35844 (2012).

225. Vazquez-Albacete, D. *et al.* An expression tag toolbox for microbial production of membrane bound plant cytochromes P450. *Biotechnol. Bioeng.* **114**, 751–760 (2017).

226. Rennig, M. *et al.* TARSyn: Tuneable antibiotic resistance devices enabling bacterial synthetic evolution and protein production. *ACS Synth. Biol.* **7**, 432–442 (2018).

227. Mutalik, V. K. *et al.* Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods* **10**, 354–360 (2013).

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

229. Schlegel, S. *et al.* Optimizing membrane protein overexpression in the *Escherichia coli* strain Lemo21(DE3). *J. Mol. Biol.* **423**, 648–59 (2012).

230. Wagner, S. *et al.* Tuning *Escherichia coli* for membrane protein overexpression. *Proc. Natl. Acad. Sci.* **105**, 14371–14376 (2008).

231. Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. *PLoS ONE* **3**, e3647 (2008).

232. Gialama, D. *et al.* Development of *Escherichia coli* strains that withstand membrane protein-induced toxicity and achieve high-level recombinant membrane protein production. *ACS Synth. Biol.* **6**, 284–300 (2017).

233. Gier, J. D. E. A scalable, GFP-based pipeline for membrane protein overexpression screening and purification. 2011–2017 (2011) doi:10.1110/ps.051466205.refolding.

234. Drew, D. E., Heijne, G. Von & Gier, J. L. De. Green fluorescent protein as an indicator to monitor membrane protein overexpression in *Escherichia coli*. *FEBS Lett.* **507**, 220–224 (2001).

235. Claassens, N. J., Volpers, M., Martins dos Santos, V. A. P., van der Oost, J. & de Vos, W. M. Potential of proton-pumping rhodopsins: engineering photosystems into microorganisms. *Trends Biotechnol.* **31**, 633–642 (2013).

236. Deisseroth, K. Optogenetics : 10 years of microbial opsins in neuroscience. **18**, 1213–1225 (2015).

237. Gourdon, P. *et al.* Optimized in vitro and in vivo expression of proteorhodopsin: A seven-transmembrane proton pump. *Protein Expr. Purif.* **58**, 103–113 (2008).

238. Martinez, A., Bradley, A. S., Waldbauer, J. R., Summons, R. E. & DeLong, E. F. Proteorhodopsin photosystem gene expression enables photophosphorylation in a heterologous host. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 5590–5595 (2007).

239. Mirzadeh, K. *et al.* Enhanced Protein Production in Escherichia coli by Optimization of Cloning Scars at the Vector−Coding Sequence Junction. **17**, 58 (2018).

240. Daley, D. O. Global Topology Analysis of the Escherichia coli Inner Membrane Proteome. *Science* **308**, 1321–1323 (2005).

241. Lee, T. *et al.* BglBrick vectors and datasheets: A synthetic biology platform for gene expression. *J. Biol. Eng.* **5**, 12 (2011).

242. Claassens, N. J. *et al.* Improving heterologous membrane protein production in Escherichia coli by combining transcriptional tuning and codon usage algorithms. *PloS One* **12**, e0184355 (2017).

243. Wagner, S. *et al.* Tuning *Escherichia coli* for membrane protein overexpression. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 14371–6 (2008).

244. Drew, D., Lerch, M., Kunji, E., Slotboom, D. & Gier, J. D. Optimization of membrane protein overexpression and purification using GFP fusions. *Nat. Methods* **3**, 303–313 (2006).

245. Hjelm, A. *et al.* Optimizing *E. coli* -Based Membrane Protein Production Using Lemo21(DE3) and GFP-Fusions. in *Membrane Biogenesis: Methods and*

Protocols (eds. Rapaport, D. & Herrmann, J. M.) 381–400 (Springer Science & Business Media, 2013). doi:DOI 10.1007/978-1-62703-487-6_24.

246. Zhang, Z. *et al.* High-level production of membrane proteins in *E. coli* BL21(DE3) by omitting the inducer IPTG. *Microb. Cell Factories* **14**, 142 (2015).

247. Engqvist, M. K. M. *et al.* Directed Evolution of *Gloeobacter violaceus* Rhodopsin Spectral Properties. *J. Mol. Biol.* **427**, 205–220 (2015).

248. Sheldon, R. A. Green and sustainable manufacture of chemicals from biomass: State of the art. *Green Chem.* **16**, 950–963 (2014).

249. Paddon, C. J. *et al.* High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature* **496**, 528–532 (2013).

250. Yim, H. *et al.* Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. *Nat. Chem. Biol.* **7**, 445–452 (2011).

251. Peralta-Yahya, P. P., Zhang, F., Del Cardayre, S. B. & Keasling, J. D. Microbial engineering for the production of advanced biofuels. *Nature* **488**, 320–328 (2012).

252. Schempp, F. M., Drummond, L., Buchhaupt, M. & Schrader, J. Microbial Cell Factories for the Production of Terpenoid Flavor and Fragrance Compounds. *J. Agric. Food Chem.* **66**, 2247–2258 (2018).

253. Hashimoto, J., Stevenson, B. & Schmidt, T. M. Rates and consequences of recombination between ribosomal RNA operons. *J. Bacteriol.* **185**, 966–972 (2002).

254. Ludu, J. S. *et al.* Genetic elements for selection, deletion mutagenesis and complementation in Francisella spp. *FEMS Microbiol. Lett.* **278**, 86–93 (2008).

255. Sharan, S. K., Thomason, L. C., Kuznetsov, S. G. & Court, D. L. Recombineering: A homologous recombination-based method of genetic engineering. *Nat. Protoc.* **4**, 206–223 (2009).

256. Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: The Keio collection. *Mol. Syst. Biol.* **2**, (2006).

257. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L. A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.* **31**, 233–239 (2013).

258. Stoddard, B. L. Homing endonuclease structure and function. *Q. Rev. Biophys.* **38**, 49–95 (2005).

259. Stan J. J. Brouns, Matthijs M. Jore, Magnus Lundgren, Edze R. Westra, Rik J. H. Slijkhuis, Ambrosius P. L. Snijders, Mark J. Dickman, Kira S. Makarova, Eugene V. Koonin, J. van der O. Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. *Science* 960–965 (2008).

260. Makarova, K. S. *et al.* Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* (2019) doi:10.1038/s41579-019-0299-x.

261. Abudayyeh, O. O. *et al.* C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* **353**, (2016).

262. Mougiakos, I. *et al.* Characterizing a thermostable Cas9 for bacterial genome editing and silencing. *Nat. Commun.* **8**, (2017).

263. Harrington, L. B. *et al.* A thermostable Cas9 with increased lifetime in human plasma. *Nat. Commun.* **8**, 1–7 (2017).

264. Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. *Nature* **21**, (2019).

265. Strecker, J. *et al.* RNA-guided DNA insertion with CRISPR-associated transposases. *Science* **364**, 48–53 (2019).

266. Bikard, D. *et al.* Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Res.* **41**, 7429–7437 (2013).

267. Gilbert, L. A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442 (2013).

268. Shen, B. *et al.* Efficient genome modification by CRISPR-Cas9 nickase with minimal off-target effects. *Nat. Methods* **11**, 399–402 (2014).

269. Gaudelli, N. M. *et al.* Programmable base editing of T to G C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).

270. Cameron, P. *et al.* Harnessing type I CRISPR–Cas systems for genome engineering in human cells. *Nat. Biotechnol.* 1–7 (2019) doi:10.1038/s41587-019-0310-0.

271. Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* (2019) doi:10.1038/s41586-019-1711-4.

272. Pitcher, R. S. *et al.* NHEJ protects mycobacteria in stationary phase against the harmful effects of desiccation. *DNA Repair* **6**, 1271–1276 (2007).

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

274. Brissett, N. C. & Doherty, A. J. Repairing DNA double-strand breaks by the prokaryotic non-homologous end-joining pathway. *Biochem. Soc. Trans.* **37**, 539–545 (2009).

275. Della, M. *et al.* Mycobacterial Ku and ligase proteins constitute a two-component NHEJ repair machine. *Science* **306**, 683–685 (2004).

276. Pitcher, R. S., Wilson, T. E. & Doherty, A. J. *New Insights into NHEJ Repair Processes in Prokaryotes*. *Cell Cycle* vol. 4 675–678 (Taylor & Francis, 2005).

277. Bertrand, C., Thibessard, A., Bruand, C., Lecointe, F. & Leblond, P. Bacterial NHEJ: A never ending story. *Mol. Microbiol.* (2019) doi:10.1111/mmi.14218.

278. Lieber, M. R. The mechanism of DSB repair by the NHEJ. *Annu. Rev. Biochem.* **79**, 181–211 (2011).

279. Shuman, S. & Glickman, M. S. Bacterial DNA repair by non-homologous end joining. *Nat. Rev. Microbiol.* **5**, 852–861 (2007).

280. White, M. F. & Allers, T. DNA repair in the archaea-an emerging picture. *FEMS Microbiol. Rev.* **42**, 514–526 (2018).

281. Chang, H. H. Y., Pannunzio, N. R., Adachi, N. & Lieber, M. R. Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nat. Rev. Mol. Cell Biol.* **18**, 495–506 (2017).

282. Aravind, L. & Koonin, E. V. Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system. *Genome Res.* **11**, 1365–1374 (2001).

283. Weller, G. R. *et al.* Identification of a DNA nonhomologous end-joining complex in bacteria. *Science* **297**, 1686–1689 (2002).

284. Bartlett, E. J., Brissett, N. C. & Doherty, A. J. Ribonucleolytic resection is required for repair of strand displaced nonhomologous end-joining intermediates. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 1–8 (2013).

285. Pitcher, R. S., Brissett, N. C. & Doherty, A. J. Nonhomologous End-Joining in Bacteria: A Microbial Perspective. *Annu. Rev. Microbiol.* **61**, 259–282 (2007).

286. Zhu, H. & Shuman, S. A primer-dependent polymerase function of Pseudomonas aeruginosa ATP-dependent DNA ligase (LigD). *J. Biol. Chem.* **280**, 418–427 (2005).

287. Nair, P. A., Smith, P. & Shuman, S. Structure of bacterial LigD 3’-phosphoesterase unveils a DNA repair superfamily. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 12822–7 (2010).

288. Moeller, R. *et al.* Resistance of bacillus subtilis spore DNA to lethal ionizing radiation damage relies primarily on spore core components and DNA repair, with minor effects of oxygen radical detoxification. *Appl. Environ. Microbiol.* **80**, 104–109 (2014).

289. Stephanou, N. C. *et al.* Mycobacterial nonhomologous end joining mediates mutagenic repair of chromosomal double-strand DNA breaks. *J. Bacteriol.* **189**, 5237–5246 (2007).

290. Vercoe, R. B. *et al.* Cytotoxic Chromosomal Targeting by CRISPR/Cas Systems Can Reshape Bacterial Genomes and Expel or Remodel Pathogenicity Islands. *PLoS Genet.* **9**, (2013).

291. Tong, Y., Charusanti, P., Zhang, L., Weber, T. & Lee, S. Y. CRISPR-Cas9 Based Engineering of Actinomycetal Genomes. *ACS Synth. Biol.* **4**, 1020–1029 (2015).

292. Standage-Beier, K., Zhang, Q. & Wang, X. Targeted Large-Scale Deletion of Bacterial Genomes Using CRISPR-Nickases. *ACS Synth. Biol.* **4**, 1217–1225 (2015).

293. Su, T. *et al.* A CRISPR-Cas9 Assisted Non-Homologous End-Joining Strategy for One-step Engineering of Bacterial Genome. *Sci. Rep.* **6**, 37895 (2016).

294. Zheng, X., Li, S. Y., Zhao, G. P. & Wang, J. An efficient system for deletion of large DNA fragments in Escherichia coli via introduction of both Cas9 and the non-homologous end joining system from Mycobacterium smegmatis. *Biochem. Biophys. Res. Commun.* **485**, 768–774 (2017).

295. Nayak, D. D. & Metcalf, W. W. Cas9-mediated genome editing in the methanogenic archaeon *Methanosarcina acetivorans*. *Proc. Natl. Acad. Sci.* **114**, 2976–2981 (2017).

296. Sun, B. *et al.* A CRISPR-Cpf1-Assisted Non-Homologous End Joining Genome Editing System of Mycobacterium smegmatis. *Biotechnol. J.* **13**, 1–10 (2018).

297. Li, L. *et al.* CRISPR-Cpf1-assisted multiplex genome editing and transcriptional repression in Streptomyces. *Appl. Environ. Microbiol.* **84**, 1–18 (2018).

298. Bhattacharyya, S. *et al.* Phage Mu Gam protein promotes NHEJ in concert with Escherichia coli ligase. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E11614–E11622 (2018).

299. Dupuy, P., Sauviac, L. & Bruand, C. Stress-inducible NHEJ in bacteria: function in DNA repair and acquisition of heterologous DNA. *Nucleic Acids Res.* **47**, 1335–1349 (2019).

300. Su, T. *et al.* The phage T4 DNA ligase mediates bacterial chromosome DSBs repair as single component non-homologous end joining. *Synth. Syst. Biotechnol.* **4**, 107–112 (2019).

301. Huang, C. *et al.* CRISPR-Cas9-assisted native end-joining editing offers a simple strategy for efficient genetic engineering in Escherichia coli. *Appl. Microbiol. Biotechnol.* **103**, 8497–8509 (2019).

302. Csörgő, B. *et al.* A minimal CRISPR-Cas3 system for genome engineering. *bioRxiv* 860999 (2019) doi:10.1101/860999.

303. Li, Z.-H., Liu, M., Lyu, X.-M., Wang, F.-Q. & Wei, D.-Z. CRISPR/Cpf1 facilitated large fragment deletion in *Saccharomyces cerevisiae*. *J. Basic Microbiol.* **58**, 1100–1104 (2018).

304. Swartjes, T., Staals, R. H. J. & van der Oost, J. Editor's cut: DNA cleavage by CRISPR RNA-guided nucleases Cas9 and Cas12a. *Biochem. Soc. Trans.* **0**, 1–13 (2019).

305. Ford, K., McDonald, D. & Mali, P. Functional Genomics via CRISPR–Cas. *J. Mol. Biol.* **431**, 48–65 (2019).

306. Martínez-García, E. & de Lorenzo, V. The quest for the minimal bacterial genome. *Curr. Opin. Biotechnol.* **42**, 216–224 (2016).

307. Bae, S., Kweon, J., Kim, H. S. & Kim, J. S. Microhomology-based choice of Cas9 nuclease target sites. *Nat. Methods* **11**, 705–706 (2014).

308. Adiego-Pérez, B. *et al.* Multiplex genome editing of microorganisms using CRISPR-Cas. *FEMS Microbiol. Lett.* **366**, 1–19 (2019).

309. Montalbano, A., Canver, M. C. & Sanjana, N. E. Review High-Throughput Approaches to Pinpoint Function within the Noncoding Genome. *Mol. Cell* **68**, 44–59 (2017).

310. Bernheim, A. *et al.* Inhibition of NHEJ repair by type II-A CRISPR-Cas systems in bacteria. *Nat. Commun.* **8**, 2094 (2017).

311. Bernheim, A., Bikard, D., Touchon, M. & Rocha, E. P. C. A matter of background: DNA repair pathways as a possible cause for the sparse distribution of CRISPR-Cas systems in bacteria. *Philos. Trans. R. Soc. B Biol. Sci.* **374**, (2019).

312. Mougiakos, I. Feel the burn: a collection of stories on hot'n'sharp DNA engineering. (Wageningen University, 2019). doi:10.18174/468570.

313. Glass, J. I. *et al.* Essential genes of a minimal bacterium. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 425–430 (2006).

314. Gibson, D. G. *et al.* Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. *Science* **329**, 52–56 (2010).

315. Hutchison, C. A. *et al.* Design and synthesis of a minimal bacterial genome. *Science* **351**, aad6253–aad6253 (2016).

316. Breuer, M. *et al.* Essential metabolism for a minimal cell. *eLife* **8**, (2019).

317. *Minimal Cells: Design, Construction, Biotechnological Applications*. (2020).

318. Xu, X. *et al.* Trimming the genomic fat: minimising and re-functionalising genomes using synthetic biology. *Nat. Commun.* **14**, 1984 (2023).

319. Yu, B. J. *et al.* Minimization of the Escherichia coli genome using a Tn5-targeted Cre/loxP excision system. *Nat. Biotechnol.* **20**, 1018–1023 (2002).

320. Pósfai, G. *et al.* Emergent properties of reduced-genome Escherichia coli. *Science* **312**, 1044–1046 (2006).

321. Mizoguchi, H., Mori, H. & Fujio, T. Escherichia coli minimum genome factory. *Biotechnol. Appl. Biochem.* **46**, 157–167 (2007).

322. Mizoguchi, H., Sawano, Y., Kato, J. & Mori, H. Superpositioning of Deletions Promotes Growth of Escherichia coli with a Reduced Genome. *DNA Res.* **15**, 277–284 (2008).

323. Yamamoto, N. *et al.* Update on the Keio collection of Escherichia coli single-gene deletion mutants. *Mol. Syst. Biol.* **5**, (2009).

324. Goodall, E. C. A. *et al.* The essential genome of Escherichia coli K-12. *mBio* **9**, 1–18 (2018).

325. Vernyik, V. *et al.* Exploring the fitness benefits of genome reduction in Escherichia coli by a selection-driven approach. *Sci. Rep.* **10**, 1–12 (2020).

326. Ziegler, M. *et al.* Engineering of a robust Escherichia coli chassis and exploitation for large-scale production processes. *Metab. Eng.* **67**, 75–87 (2021).

327. Oost, J. van der & Patinios, C. The genome editing revolution. *Trends Biotechnol.* **41**, 396–409 (2023).

328. Morisaka, H. *et al.* CRISPR-Cas3 induces broad and unidirectional genome editing in human cells. *Nat. Commun.* **10**, 5302 (2019).

329. Hatoum-Aslan, A. CRISPR-Cas3 Adds a Power Saw to the Toolbox for Human Genome Engineering. *CRISPR J.* **2**, 150–152 (2019).

330. Thoma, S. & Schobert, M. An improved Escherichia coli donor strain for diparental mating. *FEMS Microbiol. Lett.* **294**, 127–132 (2009).

331. Ton-Hoang, B. *et al.* Structuring the bacterial genome: Y1-transposases associated with REP-BIME sequences. *Nucleic Acids Res.* **40**, 3596–3609 (2012).

332. Meddows, T. R., Savory, A. P., Grove, J. I., Moore, T. & Lloyd, R. G. RecN protein and transcription factor DksA combine to promote faithful recombinational repair of DNA double-strand breaks. *Mol. Microbiol.* **57**, 97–110 (2005).

333. Espéli, O. & Boccard, F. In vivo cleavage of Escherichia coli BIME-2 repeats by DNA gyrase: genetic characterization of the target and identification of the cut site. *Mol. Microbiol.* **26**, 767–777 (1997).

334. Gilson, E., Perrin, D. & Hofnung, M. DNA polymerase I and a protein complex bind specifically to E.coli palindromic unit highly repetitive DNA: Implications for bacterial chromosome organization. *Nucleic Acids Res.* **18**, 3941–3952 (1990).

335. Hernandez, E. *et al.* Assembly and Translocation of a CRISPR-Cas Primed Acquisition Complex. *Cell* **175**, 934-946.e15 (2018).

336. Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**, 129 (2010).

337. Lauritsen, I., Porse, A., Sommer, M. O. A. & Nørholm, M. H. H. A versatile one-step CRISPR-Cas9 based approach to plasmid-curing. *Microb. Cell Factories* **16**, 135 (2017).

338. Green, R. & Rogers, E. J. Chemical Transformation of E. coli. *Methods* 3–6 (2014) doi:10.1016/B978-0-12-418687-3.00028-8.Chemical.

339. Gallagher, R. R., Li, Z., Lewis, A. O. & Isaacs, F. J. Rapid editing and evolution of bacterial genomes using libraries of synthetic DNA. *Nat. Protoc.* **9**, 2301–2316 (2014).

340. Wick, R. R., Judd, L. M. & Holt, K. E. Deepbinner: Demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks. *PLOS Comput. Biol.* **14**, e1006583 (2018).

341. Loman, N. J. & Quinlan, A. R. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics* **30**, 3399–3401 (2014).

342. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).

343. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

344. Lorenzo, M. M., Nogales, A., Chiem, K., Blasco, R. & Martínez-Sobrido, L. Vaccinia Virus Attenuation by Codon Deoptimization of the A24R Gene for Vaccine Development. *Microbiol. Spectr.* **10**, e00272-22 (2022).

345. Wu, X. *et al.* Optimization and deoptimization of codons in SARS-CoV-2 and the implications for vaccine development. 2022.09.03.506470 Preprint at https://doi.org/10.1101/2022.09.03.506470 (2022).

346. Kruth, S., Schibajew, L. & Nett, M. Biocatalytic production of the antibiotic aurachin D in Escherichia coli. *AMB Express* **12**, 138 (2022).

347. Nieuwkoop, T. *et al.* Revealing determinants of translation efficiency via whole-gene codon randomization and machine learning. *Nucleic Acids Res.* **51**, 2363–2376 (2023).

348. Lyu, X., Yang, Q., Zhao, F. & Liu, Y. Codon usage and protein length-dependent feedback from translation elongation regulates translation initiation and elongation speed. *Nucleic Acids Res.* **49**, 9404–9423 (2021).

349. Bhandari, B. K., Lim, C. S. & Gardner, P. P. TISIGNER.com: web services for improving recombinant protein production. *Nucleic Acids Res.* **49**, W654–W661 (2021).

350. Hashimoto, M. *et al.* Cell size and nucleoid organization of engineered Escherichia coli cells with a reduced genome. *Mol. Microbiol.* **55**, 137–149 (2005).

351. Iwadate, Y., Honda, H., Sato, H., Hashimoto, M. & Kato, J. Oxidative stress sensitivity of engineered Escherichia coli cells with a reduced genome. *FEMS Microbiol. Lett.* **322**, 25–33 (2011).

352. Westers, H. *et al.* Genome engineering reveals large dispensable regions in Bacillus subtilis. *Mol. Biol. Evol.* **20**, 2076–2090 (2003).

353. Sasaki, M., Kumagai, H., Takegawa, K. & Tohda, H. Characterization of genome-reduced fission yeast strains. *Nucleic Acids Res.* **41**, 5382–5399 (2013).

354. Lieber, M. R. The mechanism of human nonhomologous DNA end joining. *J. Biol. Chem.* **283**, 1–5 (2008).

355. Wang, H. & Xu, X. Microhomology-mediated end joining: new players join the team. *Cell Biosci.* **7**, 6 (2017).

356. Uphoff, S., Reyes-Lamothe, R., Garza de Leon, F., Sherratt, D. J. & Kapanidis, A. N. Single-molecule DNA repair in live bacteria. *Proc. Natl. Acad. Sci.* **110**, 8063–8068 (2013).

357. Rösch, T. C. *et al.* Single molecule tracking reveals spatio-temporal dynamics of bacterial DNA repair centres. *Sci. Rep.* **8**, 16450 (2018).

358. Jonas, H. *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*. (University of Chicago Press, 1984).

359. Kulesza, J. International Internet law. *Glob. Change Peace Secur.* **24**, 351–364 (2012).

360. Babu, M. *et al.* A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. *Mol. Microbiol.* **79**, 484–502 (2011).

361. Lombard, M., Pastoret, P.-P. & Moulin, A. M. A brief history of vaccines and vaccination. *Rev. Sci. Tech. OIE* **26**, 29–48 (2007).

362. Davidson, M. Vaccination as a cause of autism—myths and controversies. *Dialogues Clin. Neurosci.* **19**, 403–407 (2017).

363. Greenwood, B. The contribution of vaccination to global health: past, present and future. *Philos. Trans. R. Soc. B Biol. Sci.* **369**, 20130433 (2014).

364. Stanley, M. Tumour virus vaccines: hepatitis B virus and human papillomavirus. *Philos. Trans. R. Soc. B Biol. Sci.* **372**, 20160268 (2017).

365. Li, H., Wang, L., Zhang, M., Lu, Y. & Wang, W. Effects of vaccination and non-pharmaceutical interventions and their lag times on the COVID-19 pandemic: Comparison of eight countries. *PLoS Negl. Trop. Dis.* **16**, e0010101 (2022).

366. Anyshchenko, A. The Precautionary Principle in EU Regulation of GMOs: Socio-Economic Considerations and Ethical Implications of Biotechnology. *J. Agric. Environ. Ethics* **32**, 855–872 (2019).

367. Panayotou, T. 14. Economic Growth and the Environment. in *14. Economic Growth and the Environment* 140–148 (New York University Press, 2016). doi:10.18574/nyu/9781479862689.003.0018.

368. DIRECTIVE 2001/18/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL.

369. Cambrium Launches NovaColl^{TM}: First Micro-Molecular & Skin-Identical Collagen Ingredient for Personal Care Industry | GlobeNewswire by notified.

370. United Nations. Causes and Effects of Climate Change. *United Nations*.

371. McMichael, A. J., Powles, J. W., Butler, C. D. & Uauy, R. Food, livestock production, energy, climate change, and health. *The Lancet* **370**, 1253–1263 (2007).

372. OECD & Food and Agriculture Organization of the United Nations. Meat. in *OECD-FAO Agricultural Outlook 2021-2030* (OECD, 2021). doi:10.1787/cf68bf79-en.

373. Speedy, A. W. Global Production and Consumption of Animal Source Foods. *J. Nutr.* **133**, 4048S-4053S (2003).

374. Rothgerber, H. Meat-related cognitive dissonance: A conceptual framework for understanding how meat eaters reduce negative arousal from eating animals. *Appetite* **146**, 104511 (2020).

375. William Shurtleff and Akiko Aoyagi. Dr. John Harvey Kellogg and Battle Creek Foods.

376. Waltz, E. Cow-less milk: the rising tide of animal-free dairy attracts big players. *Nat. Biotechnol.* **40**, 1534–1536 (2022).

377. Melton, L. Egg without the chicken. *Nat. Biotechnol.* **40**, 812–812 (2022).

Acknowledgements

Anchored somewhere along the shores of Porto Cervo, I lay on the deck of my yacht, ostentatious, unnecessarily big and, regrettably, hypothetical. I force myself to ponder about these last years; not a small number of challenges of varying magnitude turned what I thought would be an intense, stimulating and just perhaps not too soul-crushing academic adventure into a remarkably testing journey from which the biggest lessons I take home, rather than about a specific scientific field, are about myself. A journey comprising several surprises including an interesting dive into the participatory councils of the university, as well as an unforeseen historical health crisis with a fascinating dystopian flair that taught me the affair is much better as a genre of fiction than as a factual experience. A journey with a curious sour taste comparable to chugging a large gulp of beer assuming it to be a regular pilsner while it actually is a lambic: it might leave us questioning what it is we like about beer, when it in fact doesn't quite get any more real than that. A journey I am, after all, very glad I started, but most importantly, that I am extremely happy to put an end to. However, I have been able to make it only thanks to the positive impact and influences that several people had on me. If it were not for many of you, there is no way I would have managed. I therefore thank you all, whether your name is somewhere below or not, for giving me the strength I didn't have when I needed it the most.

First, I would like to express my gratitude to **John**. Your patience and understanding have been key for me to develop a sustainable pace to write this thesis. While I wish we would have had the time and space to better discuss science in more detail and find projects to work on closer to both of our hearts, I feel lucky to have spent so much time in this lab, for it really is a

fantastic place to grow as a scientist and as a person. For always reminding us that *the most important thing is to have fun*, and for everything, thank you, John!

I must also take the time to thank you, **Raymond**, for your support and willingness to share some of your knowledge on bioinformatics with me. I would have loved to work closer with you on projects that require coding, I think we would have had great fun! But in any case, thank you for being there on both my PhD trips, for sharing your sense of humor so naturally with all the groups you interact with and for everything!

And a tad late as always, **Nico**! There are more things for which I must thank you than words I can muster in this paragraph. Having supervised me for my BSc thesis, my MSc internship and now my PhD, you have helped me grow tremendously throughout these years, both as a scientist and as a person. I admire your determination and resolve, and I am very grateful to you for showing me how fun research is when performed on a topic for which one is passionate about. Thank you for your support during the tough times and for everything, and I wish you the best for your scientific career in the coming years. I am very curious and excited to see how your scientific efforts develop, and I will make sure to stay tuned to see the developments myself!

I am very grateful to the thesis committee for taking the time to read through and evaluating my thesis: **Christophe Danelon**, **Joe Bondy-Denomy**, **María Suárez Diez** and **Richard A. Notebaart**.

Joep and **Thijs**, the OG BaSyC bois, thanks a lot for being there since the beginning and for being my paranymphs despite how busy you are right now. Building a synthetic cell with you two has been a privilege, as it has been being office mates and becoming friends. I have had a great time with you two, and I have missed you dearly since you graduated. I am looking forward to staying in touch in the future, and I wish you and your partners the best with your nearly due and your recent family incorporations!

Marijn, **Gijs**, **Leonardo**, **Victor**, **Sanne**, **Thomas** and **Laura**, I feel lucky to have had the opportunity to supervise and work with you all. Although our works together were certainly not without challenges, I have learned a lot from all the projects we worked on and from all our interactions. I wish you the very bright professional future I know you will all have!

Belén, **Ismael**, **Janneke**, **Joep**, **Jurre**, **Lorenzo** and **Thijs**, you were and likely will always be The best Office one could ask for. It was absolutely great to share a room with you! I enjoyed a lot our periodic dinners and playing the floor is lava to the best of our abilities, and I look forward to the day Thijs pays his cake debt! **Afonso**, **Ane** and **Carl**, you were great additions to the crew, and I wish we had spent more time together in the office.

Isabelle, **Laureen**, **Ricardo** and **Thomas**, thank you for accepting me into your office crew and treating me like one of your own! I had a lot of fun during the weeks I stayed with you guys, and I would have loved to have stayed even longer. **Christian**, **Eric**, **Tim**, **Olufemi** and **Vittorio**, I thank you too for having me in your office for another short while!

Michele P. and **Giulia B**., I am most thankful to the BaSyC consortium for making our paths cross. I thank you both for being among the most genuine people I have met these last years, and for all the laughs we had every time we met in a BaSyC meeting or in a conference. **Alberto B**., **Elisa G.**, **Daphne B**., I wish we had spent more time together during these years, but I am very happy to have met you through the consortium. **Christophe D**., I thank you for the brainstorms and the discussions we shared. I admire your critical spirit, and it is an honor to have you as one of my opponents. I wish you and your group the best for the future!

Adiini, **Afonso**, **Alex**, **Ane**, **Anneleen**, **Belén**, **Carina**, **Carl**, **Catarina**, **Charlotte**, **Christian**, **Costas**, **Despoina**, **Eric**, **Evgenios**, **Franklin**, **Guus**, **Hanne**, **Ioannis**, **Isabelle**, **Ismael**, **James**, **Janneke**, **Jeroen**, **Joep**, **John**, **Jorrit**, **Joyshree**, **Jurre**, **Lucía**, **Lorenzo**, **Maartje**, **Mamou**, **Mark**, **Melvin**, **Miguel**, **Mihris**, **Nico**, **Olufemi**, **Panos**, **Prarthana**, **Raymond**, **Ricardo**, **Rob**, **Servé**, **Stijn**, **Suzan**, **Teunke**, **Tijn**, **Tim**, **Thomas**, **Tom**, **Vittorio**, **Wen** and everyone I met during my time in BacGen plus everyone else I might be missing, a big thanks to everyone for the great atmosphere and the many remarkable shared moments.

Alberto, **Anastasia**, **Burak**, **Carrie**, **Catalina**, **Chen**, **Daan**, **Dani**, **Detmer**, **Diana**, **Emmy**, **Hugo**, **Irene**, **Isabelle**, **Ivette**, **Johanna**, **Jolanda**, **Kassi,** **Kate**, **Max**, **Michelle**, **Nancy**, **Patrick**, **Peter**, **Prokopis**, **Sara** ***C***., **Stephan** and **Sudarshan**, I am very grateful to have met you all. My time at MIB has been greatly improved by having you around! **Enrico**, I am happy to have stayed in touch for a while, scientifically speaking, and to have published a minireview together. I would have loved for our Rhodobacter ideas to succeed and for us to publish a very nice experimental paper together. I wish you the best wherever you go and whatever you decide to do, and I hope to stay in touch for many years! **Christos**, **Enrique**, **Luis**, **Lyon**, **María**, **Sabine**, **Sara B**. and **Wasin**, it has been a pleasure to share many years and experiences with you all. I admire you for your capacity to stay and work together in times of turmoil and make the most out of it with excellent results. **Belén**, **Carrie**, **Sharon**, **Jannie**, **Joep**, **Ivette**, **Catarina**, **Enrique**, **Janneke**, **Thijs** and **Patrick**, I am very happy to have been part of the group to religiously devour fries and more on Thursdays!

Dani, no sabes cuánto me alegro de que nos hayamos conocido en un ambiente tan variopinto como MIB. ¡Gracias por ser como eres, y muchísima suerte con tu nuevo grupo! **Stephan**, it’s been great to meet you and to share many fun times with you as of late! **Alberto**, **Daan**, **Dani**, **Eva F.C**., **Eva F.P**.,

Francesca, **Ivette**, **Jarrett**, **Jenna**, **Lot**, **Nico**, **Nohemi**, **Peter** and **Pocket**, I had a lot of fun with all of you and I look forward to the day we can all reconvene.

Anemoon, **Deli**, **Jenny**, **Marta**, **Morris**, **Riemer**, **Sophie**, **Sophieke**, **Sanne** and **Thomas**, it was truly a pleasure to take part in the supervision team during your time as WUR's iGEM team. Following the development of the team and being able to join you for the Jamboree in Paris were among the highlights of my PhD, so thank you all for that! **Despoina**, **Eric**, **Jurre**, **Lorenzo**, **Luis**, **Lyon**, **Nico**, **Niek**, **María**, **Raymond**, **Stijn** and ***Zoë***, it was great to be part of the supervision team with you. **María**, no tengo palabras para describir lo bien que me lo pasé *la noche final* en aquella rarísima azotea de París; ¡me alegro mucho de que acabáramos yendo juntos!

Annemerel and **Paula**, organizing the WPC PhD party 2019 with you was an absolute blast, I am still astonished at how well everything turned out. Thank you for everything! **Dani**, **Nancy**, **Reinier**, **Kassi**, **Panos**, it was great fun to organize the MIB seminar series with you all.

Caifang, **Catarina**, **Costas**, **Enrique**, **Giannis**, **Ivette**, **Lot**, **Nong** and **Ran**, I am very grateful to you for having organized such an amazing PhD trip (2019) to the East Coast of the USA. **Carrie**, **Catarina**, **Christos**, **Costas**, **Despoina**, **Hugo**, **Janneke**, **Jannie**, **Joep**, **Lyon**, **Mamou**, **Martha**, **Menia**, **Patrick**, **Prokopis**, **Rik**, **Sharon**, **Taojun**, **Thijs**, **Wasin** and **Wen**, thanks everyone for making it such a great experience, and many thanks to **Diana** and **Raymond** for being there with us!

Anastasia, **Carina**, **Despoina**, **Lorenzo**, **Lyon**, **María**, **Peter** and **Valentina**, it was an absolute pleasure to organize the PhD trip 2022 to the USA West Coast with you all! Organizing parties to raise funds for the trip was an amazing source of fun. **Architha**, **Costas**, **Enrique**, **Eric**, **Evgenios**, **Jurre**, **Kassi**, **Maaike**, **Marco**, **Marie-Luise**, **Marten**, **Maryse**, **Michelle**, **Miguel**, **Nancy**, **Raymond**, **Reinier**, **Ricardo**, **Sabine**, **Sara B.**, **Sara M.**, **Sonja**, **Suzan**, **Thomas**, **Marco** and **Timon**, thank you for being flexible during the tricky moments of the trips, and thank you for making the trip such a memorable time! **Eric**, it was amazing to spend an extra week with you traveling and enjoying the magic of the national parks. I am glad we got to know each other better and I feel lucky to now consider you a good friend!

Nico V., thank you for introducing us to D&D and standing us when we go full brainless. Thank you for having the patience to explain the consequences of casting Fireball indoors in a wooden building every time. Reed and Brocc will miss you forever! **Peter**, **Rob** and **Wen**, thank you too for providing our group with adventures that we will likely remember forever! **Daan**, **Thijs**, **Thomas**, **Jarrett**, **Jeroen** and **Ricardo**, it was an honor to battle along your side! Strahd will fall one day, of that much, I am sure.

Ivette, **Zak**, **Rodolfo**, I am so very nostalgic of the times we shared together! I hope we can meet soon and at the very least stay in touch to comment together on Zak's reproductive endeavors (you're going to be an amazing dad!) and other pressing philosophical matters. **Valentijn**, my good sir, I miss you enormously since you left MIB. I wish you and your partner the best with your tiny one. **Maarten K**., I will never forget our conversation about music over the lab sink and your question, "*are you familiar with improvisation*?". I'm happy we stayed in touch, and I wish you a very fulfilling life absolutely filled with music. **Simone**, meeting you in San Diego was lovely, thank you so much for bringing me to Panama 66, I loved it!

Anja, **Hannie**, **Heidi** and **Sarash**, thank you all for your time and help during all these years. **Belén**, **Iame**, **Ineke**, **Felix**, **Guus, Laura**, **Merlijn**, **Philippe**, **Rob**, **Sjon**, **Steven**, **Tom vdW**., **Tom S**., **Ton**, **Victor** and **Wim**, thank you for your support throughout the years and for making all of the great research going on in our floors possible! **Thijs E**., thank you for being a critical academic and for trying to make our department a better place for everyone.

My time in the WUR Council was an extremely interesting experience*, and I have* to thank everyone who encouraged me to channel my criticism to try and change things from within, as well as everyone with whom I interacted in the council and in its different committees; *I believe that* more people, like you, should *devote part of* their time to make people's life better. Special thanks to **Larissa**, whom it was always a pleasure to interact with, and **Clementine**: me siento tremendamente afortunado de haberte encontrado durante mi paso por el WURC. ¡No sé qué habría hecho sin ti, muchísimas gracias por todo! Everyone at the **WPC**, you're the group I admire the most, as you *work* on long-lasting projects with magnitudes *often* comparable to that of any PhD. It was also very stimulating to participate in **WGS** meetings, and I thank everybody therein for focusing on improving the life and experiences of PhD candidates at WUR. Special thanks to **Vesna** from VLAG, for your help during testing times but also for devoting yourself so much to the graduate school and its members! **Carina**, **Jolanda**, **Marie-Luise**, **Nico V.**, **Peter**, **Timon** and **Thijs**, it was an honor to join the MIB PhD board and work with you all to improve the working environment of our colleagues.

Alon, **Cecilia**, **Celia**, **Emilio**, **Leonardo**, **Loena**, **Miquel**, **Marta**, **Maura** and **Michele**, thanks to every single one of you for being how you are. Spending time with you is always amazing and a highlight that brightens my life even in the darkest times. I love you all!

Alex, **Angelo**, **Anna B**., **Anna V**., **Camila**, **Dona**, **Giulia**, **Guidoriccio**, **Helen**, **Inès**, **Janne**, **Javi**, **Jojo**, **Linda**, **Lucía**, **Luuk**, **Maarten**, **Massimo**, **Mathijs**, **Michele**, **Mick**, **Sophia**, **Yann** and the many other people I am forgetting. Thank you all for being amazing people and for making my years at Weppa an unforgettable adventure!

Adrià, mare meva, quin perill... estic molt content d'haver-te conegut! **Agata**, I am very happy that life put us together in that lab practical many years ago! I am convinced you will do fantastic in your new position and at whatever job the future might bring you. **Alejandro**, ha estat un plaer coneixer-te durant aquesta aventureta a Wageningen, et desitjo un futur fantàstic! **Alok**, I would have loved to celebrate my graduation with you and **Nick**, I am sure you both would have been celebrating with me until the very end. I am grateful to have met you both, as brief as it was. **Aravind**, I wish you the best wherever you are in the future and I hope to celebrate with you soon! **Debs** i **Tom**, espero que el futur ens porti a fer moltes més coses junts! Seguiu sent genials, siusplau! **Jordi**, **Ramon** y **Xabi**, siempre habéis sido ejemplos para mí en cuanto a cómo sobrevivir un doctorado en Droef. **Luc**, it was a pleasure working with you while I was more engaged with Droef affairs, and it has been a privilege to have you caring and doing so much work for the community. Thank you! **Maarten de G**., I wish we had gotten closer while you still lived at Droef, but I'm happy we still manage to see each other sometimes! I wish the best for you, **Patries** and **Evi**! **Manu**, I cannot thank you enough for reminding me that the sun had already been shining and for all the amazing times we had together. **Marina**, ¡me alegra muchísimo haberte conocido en la villa de Droef! Muchísimas gracias por tu hospitalidad en Sardeña, ¡me ha encantado verte! **Mattia**, it was an absolute pleasure to meet you during our *Plantsoen* times, between shots, laughter, and insane amounts of carbonara. I am glad to have kept in touch with you through the years! **Nico K**. I have missed having you around since you left Droef, I really hope by the time this booklet is out I have managed to visit you and jam together! **Pol E**. y **Gonzo**, ¡ha sido fantástico conoceros, hijos del metal! Espero que ho passis genial facis el que facis, Pol, y ¡estoy seguro de que te lo vas a gozar riquísimamente en el futuro próximo, Gonzo! **Tobi**, it has been great to share so many moments with you throughout these years, and I hope to be able to share many more in the future! **Vicky** and **Luca**, it has been a pleasure getting to know you both, I wish you guys the best for the future!

Cerro, **Espi**, **Gerard**, **Luis**, **Malu**, **Marc**, **Reca** y **Santos**, ¡muchísimas gracias por haberme adoptado en vuestro grupo! Difícilmente me lo paso mejor como cuando estoy con vosotros de barbacoa y de risas. **Ferran**, **Jordi**, **Kiru**, **Luis**, **Marc, Ponko**, **Raquel** y **Xavi**, ha sido un tremendo placer conoceros y viciarla con vosotros.

Alicia A., **A**. **Porcel**, **Clara S**., **Elisa G.**, **Enric C.**, **Ibai R**., **Joni de L**., **Jose C**., **Maria R**., **Miquel G**., **Nynke D**., you are light. I thank life for having brought me close to you, and I wish I could see you much more often than I do!

Special thanks to everyone whose theses I have checked to write mine with more confidence, and to **Alexandra Elkbayan** and all those who do the right thing because it is the right thing to do.

Eduard, estimat amic **Porexpant**, no puc estar-te més agraït per soportar-me durant el procés de fer-me la portada. Moltíssimes gràcies per tot el treball, estic enamorat del resultat final! Et desitjo un fantàstic futur amb la teva família i espero que ens veiem més freqüentment!

Finalmente, infinitos agradecimientos a mi familia por ser como son: geniales. **Rebe** y **Roberto**, me siento muy afortunado y orgulloso de tener a dos personas tan extraordinarias como vosotros en mi vida. **Ramón**, **Moon**, **Cates**, **Alicia**, **David**, **Micky**, **Kati** y **Angie**, me encantaría poder estar más presente en vuestras vidas, pero me alegro mucho de que nos vayamos haciendo mayores juntos aún en la distancia. **Sergio**, hubiera sido maravilloso poder celebrar este hito contigo, pues decidí estudiar lo que he estudiado en parte gracias a varias conversaciones que tuve contigo. **Padre**,

¡muchas gracias por haber despertado en mí la curiosidad por el fantástico mundo de la biología! **Madre**, ¡muchísimas gracias por tu apoyo incondicional y por estar ahí siempre que lo necesito, no sé qué haría sin ti!

About the title and cover

Being the fifth studio album released by the saxophone master John Coltrane, Giant Steps (1960) became one of the most influential jazz albums of all time, including several highly acclaimed compositions that are still highly regarded today. Giant Steps, the song giving the album its title, refers to the approach that Coltrane followed to compose and improvise solos during its recording; the chord progression of the song, outlined by the double bass, featured an unusual root movement by major thirds.

The title of this thesis, an academic *scherzo* in the widest sense and a playful nod to Coltrane's album, uses wordplay to emphasize two facts. On the one hand, because of the uncertain nature of research in the life sciences and specifically in molecular biology, the individual findings collected in the span of a PhD thesis are unlikely to dramatically push forward a specific scientific discipline, which can be a hard pill to swallow and yet it is part of the game. And, on the other hand, the experiments performed throughout this thesis, as well as any information discussed in the reviews included herein, belong to the realm of molecular biology, and as such they are part of an imperceptible, minuscule world, which is impossible to observe without the help of highly specialized and abstract technologies, making any potential progress in the field *very tiny*.

The cover of this thesis, designed with and crafted by Eduard Puig, represents the cover of a jazz album. It depicts a person playing an abstract DNA molecule as if it were a saxophone, being initially generated by artificial intelligence, somewhat reminiscent of Coltrane’s album cover of Giant Steps. Further, the cover style follows the overall design principles of the album covers belonging to the renowned Blue Note record label, which produced several greatly influential jazz albums, being principally associated with the subgenre of *hard bop* jazz, which is Max’s favorite subgenre.

List of publications

Mougiakos, I., Mohanraju, P., Bosma, E. F., Vrouwe, V., **Finger Bou, M.**, Naduthodi, M. I., ... & Van Der Oost, J. (2017). Characterizing a thermostable Cas9 for bacterial genome editing and silencing. *Nature communications*, *8*(1), 1647.

Finger-Bou, M.*, Claassens, N. J.*, Scholten, B., Muis, F., De Groot, J. J., de Gier, J. W., ... & Van Der Oost, J. (2019). Bicistronic design-based continuous and high-level membrane protein production in Escherichia coli. *ACS synthetic biology*, *8*(7), 1685-1690.

Finger-Bou, M., Orsi, E., van der Oost, J., & Staals, R. H. (2020). CRISPR with a happy ending: Non-templated DNA repair for prokaryotic genome engineering. *Biotechnology Journal*, *15*(7), 1900404.

Nieuwkoop, T., **Finger-Bou, M.**, van der Oost, J., & Claassens, N. J. (2020). The ongoing quest to crack the genetic code for protein production. *Molecular cell*, *80*(2), 193-209.

Claassens, N. J., Bordanaba-Florit, G., Cotton, C. A., De Maria, A., **Finger-Bou, M.**, Friedeheim, L., ... & Bar-Even, A. (2020). Replacing the Calvin cycle with the reductive glycine pathway in Cupriavidus necator. *Metabolic Engineering*, *62*, 30-41.

*: equal contribution

Overview of completed training activities

Discipline specific activities			
Event	**Organized by**	**Country**	**Year**
BaSyC kick-start training	BaSyC	NL	2018
NBC-18: Biotechnology in Harmony	NBV	NL	2018
WGS PhD Workshop Carousel 2018	WGS	NL	2018
1st BaSyC International Symposium	BaSyC	NL	2018
II GASB Conference	GASB	DE	2018
DutchBiophysics meeting	NWO	NL	2018
Fall meeting – General & Molecular Microbiology	KNVM	NL	2018
Biomimicry symposium	CODON/Alchimica	NL	2019
101st Dies Natalis: Innovation for nature conservation	WUR	NL	2019
BaSyC kick-start workshop	BaSyC	NL	2019
BaSyC Spring meeting	BaSyC	NL	2019
NBC-19: The Sound of Biotech*	NBV	NL	2019
Life2019	NWO	NL	2019
CRISPR 2019	Conférium	CA	2019
FEMS2019	FEMS	SCO	2019
III GASB Conference*	GASB	DE	2019
BaSyC Fall meeting	BaSyC	NL	2019
Bioinformatic Resources for Protein Biology	EMBL-EBI	GB	2020
BaSyC Spring Poster Session	BaSyC	online	2020
BaSyC Spring Pitches	BaSyC	online	2020
SynCell2020 Conference	SynCell2020	online	2020
Engineering with Evolution	Imperial College London	online	2020
BaSyC Summer School 2020	BaSyC	online	2020
IV GASB Conference	GASB	online	2020
5th Applied Synthetic Biology in Europe	ASBE	online	2020
BaSyC Fall Meeting 2020	BaSyC	online	2020
Nanopore Community Meeting 2020	Nanopore	online	2020

*: poster presentation; **: oral presentation

General courses

Event	Organized by	Country	Year
5th Annual Wageningen PhD Symposium**	WUR	NL	2018
Reviewing a Scientific Paper	WGS	NL	2018
Illustrator for Scientists	MIB	NL	2018
VLAG PhD week	VLAG	NL	2018
Making an Impact	WGS	NL	2018
Project and Time Management	WGS	NL	2019
Competence Assessment	WGS	NL	2019
Effective behaviour in your professional surroundings	WGS	NL	2019
6th Annual Wageningen PhD Symposium	WUR	NL	2019
Career Perspectives	WGS	NL	2021
Grant Application	BCF Courses		2022

Other activities

Event	Organized by	Country	Year
Preparing research proposal	BacGen (MIB)	NL	2018
Bacterial Genetics meetings	BacGen (MIB)	NL	2018-2022
BaSyC internal meetings	BacGen (MIB)	NL	2018-2020
PhD meetings	MIB	NL	2018-2022
PhD yearly initiative	BaSyC	NL	
PhD trip 2019 (as participant)	MIB & SSB	USA	2019
MIB Seminars (as organizer)	MIB	NL	2021-2022
PhD trip 2022 (as organizer)	MIB & SSB	USA	2022

*: poster presentation; **: oral presentation

About the author

Born in Barcelona, Spain on July 6th, 1993, Max spent most of his life in the coastal town of Calafell. Always fascinated by animals and other living beings, wrote his final high school paper on the topic of microbiology. During the following years, he became very intrigued about the worlds of molecular biology and genetics.

Following his curiosities, Max studied Agrifood Biotechnology at the University of Barcelona. During his BSc, he participated in a short exchange program and spent a month as a research assistant at the Department of Physiopathology at the University of Concepción, Chile. Under the supervision of Dr. Francisco Roa and Dr. Juan Carlos Vera†, he performed molecular cloning and immunocytochemistry techniques to study the role of the vitamin C transporter in liver cancer. Shortly after his return to Barcelona and avid to continue learning laboratory techniques, he joined the Department of Development Biology and Genomics, where he stayed for approximately four months as a part-time research assistant under the supervision of Dr. Marina Ruiz-Romero and Prof. Dr. Montserrat Corominas further learning molecular cloning techniques to ultimately study tissue regeneration and cancer using *Drosophila melanogaster*. To finish his degree and assisted by an Erasmus scholarship, he carried his BSc thesis at the Bacterial Genetics group at Wageningen University and Research, working on the development of genetic engineering tools for the expression of membrane proteins in *Escherichia coli* under the supervision of Dr. Nico Claassens and Prof. Dr. John van der Oost. He obtained his BSc in 2015 and, having had an amazing time in the Wageningen, he set out to return to the small student city after his graduation to continue his studies.

Approximately half a year later, Max returned to Wageningen to start his MSc in Cellular and Molecular Biotechnology. His MSc thesis took place, again, at the

Bacterial Genetics group at WUR, and was dedicated to developing CRISPR-based genetic engineering tools for thermophilic bacteria, under the supervision of Dr. Prarthana Mohanraju, Dr. Ioannis Mougiakos and Prof. Dr. John van der Oost. He then moved to Germany for half a year to carry his MSc internship at the Max Planck Institute of Molecular Plant Physiology, in Potsdam-Golm, Germany. At the Systems and Synthetic Metabolism research group and under the supervision of Dr. Nico Claassens and Dr. Arren Bar-Even†, he worked on the metabolic engineering of *Cupriavidus necator* to engineer a more efficient pathway for CO_2 assimilation while developing a CRISPR genome editing tool to engineer the bacterium.

After his MSc internship, he obtained his MSc (*cum laude*) and started his PhD at the Bacterial Genetics group, within the framework of the NWO’s Building a Synthetic Cell consortium. Under the supervision of Prof. Dr. John van der Oost, Dr. Raymond Staals and Dr. Nico Claassens, his PhD was focused on the engineering of protein expression, non-templated DNA repair and genome editing and minimization of bacteria. In addition, Max was elected as a PhD representative at the WUR Council for an academic year, where he joined several committees to work on pressing issues that affect especially the community of PhD candidates at WUR but also more general matters related to the growth and development of the university.

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

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

Cover design by Eduard Puig Sànchez

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