TU Delft

Delft University of Technology

Document Version
Final published version

Citation (APA)
van den Brink, M. (2026). *Laboratory evolution of synthetic cells*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:5826ba2c-4be8-4b31-bc99-70741b819303

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

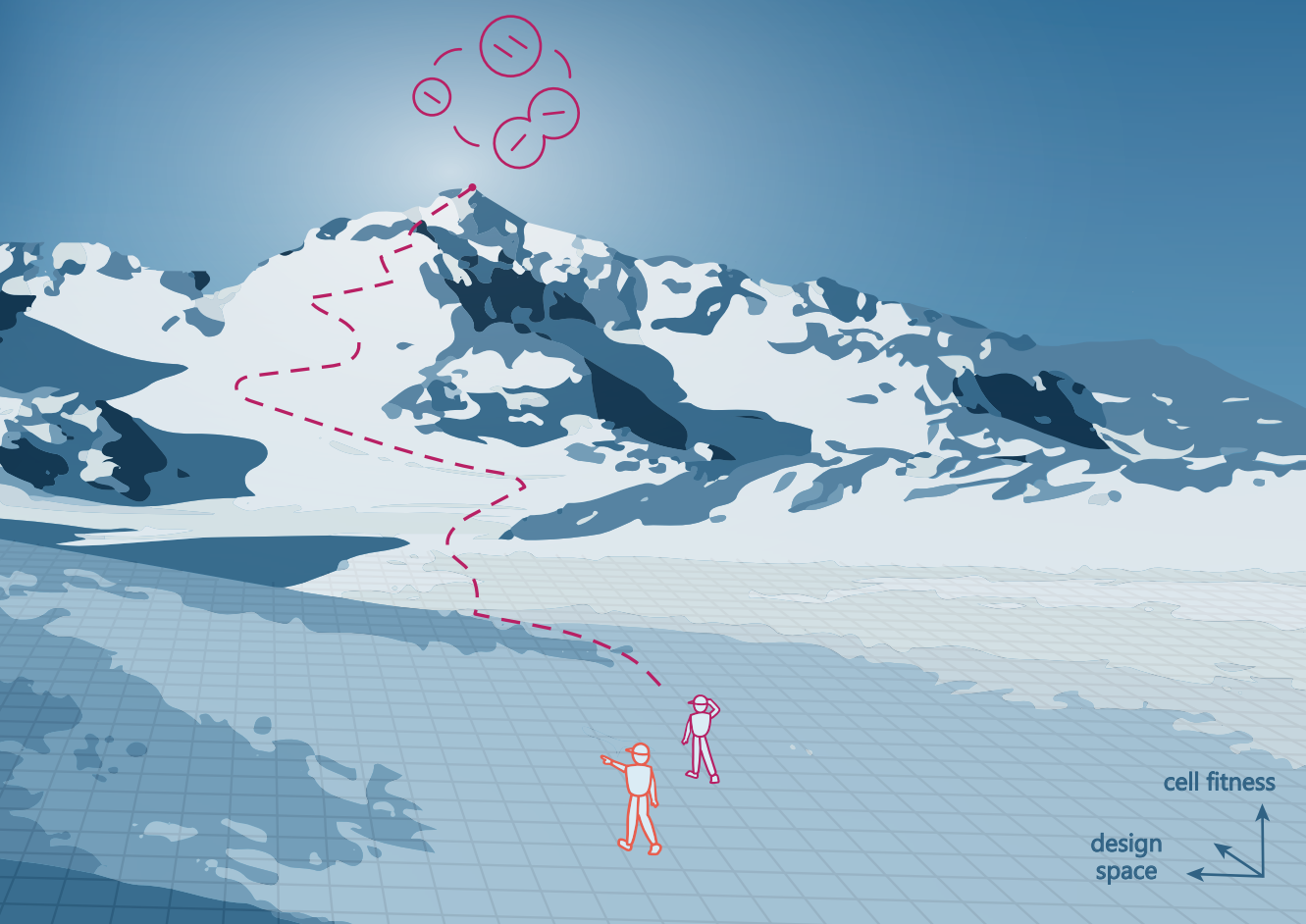
Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

Laboratory evolution of synthetic cells

Marijn van den Brink



Laboratory evolution of synthetic cells

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus,
Prof. dr. ir. H. Bijl,
chair of the Board for Doctorates
to be defended publicly on
Tuesday 22 September 2026 at 12:30

by

Marijn VAN DEN BRINK

This dissertation has been approved by the (co)promoters.

Composition of the doctoral committee:

Rector Magnificus	chairperson
Prof. dr. C.J.A. Danelon	Toulouse Biotechnology Institute, France, promotor
Prof. dr. G.H. Koenderink	Delft University of Technology, promotor
Prof. dr. N.J.H.P. Claassens	Wageningen University & Research, copromotor

Independent members:

Prof. dr. ir. S.J. Tans	Delft University of Technology/AMOLF
Dr. M.P. Chien	Erasmus University Medical Center
Dr. ir. L. Laan	Delft University of Technology
Dr. Y. Rondelez	ESPCI Paris/CNRS, France
Prof. dr. M.E. Aubin-Tam	Delft University of Technology, *reserve member*

The research in this thesis was performed at the Department of Bionanoscience at the Delft University of Technology, The Netherlands. The project received financial support from the "BaSyC – Building a Synthetic Cell" Gravitation grant of the Netherlands Ministry of Education, Culture and Science (OCW) and the Netherlands Organisation for Scientific Research (NWO).





Keywords: synthetic biology, synthetic cell, cell-free gene expression, *in vitro* directed evolution, combinatorial DNA libraries

Printed by: ProefschriftMaken
Cover by: Marijn van den Brink
Copyright © 2026 by Marijn van den Brink
ISBN: 978-94-6384-998-2

An electronic version of this dissertation is available at: http://repository.tudelft.nl

Table of Contents

Summary **5**

Samenvatting **9**

Chapter 1 **13**
Introduction

Chapter 2 **31**
MOSAIC: A highly efficient, one-step recombineering approach to plasmid editing and diversification

Chapter 3 **61**
Combinatorial optimization of protein systems in synthetic cells

Chapter 4 **117**
Image-based phenotypic sorting of synthetic cells

Chapter 5 **157**
Co-optimization of the genetic and molecular composition of a synthetic cell

Chapter 6 **179**
Perspective on laboratory evolution of synthetic cells

Acknowledgements **203**

Curriculum vitae **207**

List of publications **208**

Summary

The aim of this thesis was to explore methodologies that accelerate the building of a minimal, synthetic cell. This was achieved by building targeted combinatorial DNA libraries, subjecting these to various screening and selection methods, and exploring nanopore sequencing analysis and visualization methods for sequence-function mapping.

Throughout billions of years of evolution, even the simplest cells accumulated extraordinarily complex genetic architectures, spatial organizations and molecular interactions. This complexity makes it difficult to disentangle the role of each element in the cell. What if we could reconstruct a cell but leave out all unessential complexity? A minimal version of a cell, composed of well-characterized parts, would give insight into which cellular processes are essential for life and how interacting genes and proteins give rise to a cell's behaviors. This is the purpose of building a minimal, synthetic cell capable of self-replication, self-maintenance and evolution.

Building a cell without a blueprint requires constantly going back and forth between prototyping and learning. To accelerate this research methodology, we postulate that the rational approach to prototyping must be complemented with randomness-leveraging methods with high exploratory powers. Directed evolution is such a method: a widely developed protein engineering method that can evaluate many prototypes in parallel. In our lab, the molecular framework of synthetic cells consists of liposomes capable of gene expression due to the encapsulation of DNA and a purified cell-free expression system (PURE system). **Chapter 1** of this thesis **outlines the current capabilities of directed evolution methods** for the engineering of such *in vitro* molecular frameworks. This outline shows that directed evolution has been well established for optimizing single proteins but has barely been applied to the engineering of multi-protein systems *in vitro*. We identify the following challenges in applying directed evolution for building a synthetic cell: (i) the vastness of the design space, (ii) selecting the optimization targets, and (iii) the lack of spatiotemporal screening and sorting methods for vesicles that exhibit time-dependent morphological or localization features. The next chapters will address these challenges.

Chapter 2 introduces a new protocol for the **construction of targeted combinatorial plasmid libraries** named MOSAIC. Traditional protocols for plasmid editing that rely on DNA amplification and/or cloning steps can be laborious when mutating many sites on one DNA template. Therefore, we adapted an *Escherichia coli* genome editing method, capable of introducing mutations at multiple sites at once, for a new purpose: mutating plasmids. A target plasmid and DNA oligos harboring the desired mutations were co-electroporated into *E. coli* expressing the bacteriophage-derived single-stranded DNA annealing protein CspRecT. During plasmid replication, CspRecT mediates the annealing of the oligos to the plasmid DNA

at the replication fork, leading to the incorporation of mutations (mismatches, insertions or deletions) with 50-90% editing efficiency per target site. A combinatorial library was constructed using an oligo library targeting four sites at once. The library was characterized by long-read nanopore sequencing, applying quality control to the regions of interest.

Chapter 3 studies the **impact of tuning protein translation initiation rates** on the activity of DNA-encoded protein systems. Using MOSAIC, targeted combinatorial libraries were constructed for a two-gene self-replicator DNA and a four-gene phospholipid synthesis pathway. For the DNA replicator, the best-performing variants of a 156-variant ribosomal binding site (RBS) library were selected as the result of their superior self-amplification. For the phospholipid synthesis pathway, two libraries – one RBS library of 10^4 variants and one library of 10^8 variants in which the first six codons of the four genes were mutated – were screened and the top variants sorted by fluorescence-activated cell sorting (FACS). Nanopore sequencing enabled the reconstruction of experimental fitness landscapes mapping DNA sequences with system's level activities. These landscapes provided insights into the optimal RBS strengths and base substitutions and revealed which genes were most impactful in modulating the functionality of the protein systems. In addition, the combinatorial DNA library approach allowed for studying the effects of combinations of mutations. The theoretical fitness of combinatorial variants was very predictive for the DNA replicator library, indicating the absence of epistatic interactions between the RBSs of the two replication genes. The combinatorial variant fitness was less predictive for the larger phospholipid synthesis pathway libraries. Taken together, we exemplified how combinatorial testing can be expanded from single proteins to multiprotein systems, and we found that tuning protein translation initiation rates was effective in modulating DNA replication and phospholipid synthesis at the system's level.

Chapter 4 describes a method for the **image-based phenotypic sorting** of synthetic cells. This method explores the phenotypes of gene-expressing vesicles within nonclonal populations using microscopy. The desired variants are marked by photoactivating the fluorescence of an encapsulated protein and subsequently sorted by FACS. By automating optical microscopy and machine learning-assisted image analysis, we demonstrated the screening of up to 10^5-10^6 vesicles per day, of which 100-1000 were sorted with 80% precision. Using a reconstituted bacterial protein oscillator – the Min system – as a case study, we showed that vesicles can be selected based on fluorescence intensity, protein localization, membrane morphology and dynamic behaviors, and that their phenotype can be linked to genetic content. This approach greatly expands the set of phenotypes that can inform sorting decisions and is therefore key to the evolutionary engineering of synthetic cells.

Interdependencies between the DNA and molecular hardware of synthetic cells (the cell-free expression system) suggest that a co-optimization approach is likely key to engineering synthetic cells. **Chapter 5** describes the design of a **co-optimization strategy exploiting**

DNA barcoding to identify the molecular composition. Here, the *molecular composition* refers to the components of the cytoplasm (mainly the cell-free expression system) and the membrane. We estimated that this strategy should enable the screening of libraries of 10^3-10^4 DNA variants in combination with 30-100 molecular variants by FACS per day. Using proof-of-concept experiments with a wild-type DNA template, we validated that the barcodes do not bias variant enrichment and allow the retrieval of well-performing molecular compositions. Streamlining the preparation of liposomes with distinct genetic and molecular contents and exploiting active learning to bridge the gap between the large search space and the relatively low experimental capacity are the next steps to further demonstrate the feasibility of this method.

Finally, **Chapter 6** describes the **future perspectives** on laboratory evolution of synthetic cells by identifying three methodological improvements. First, sorting should not be limited to the top performing variants only, but (quantitative) data should also be gained on lower-fitness variants. Second, synthetic cell characterization could be expanded from DNA sequencing to multiomics analysis. Third, DNA recovery from sorted liposomes could be optimized to enable more accurate quantification of variant frequencies. These improvements will strengthen our knowledge gained from each prototyping experiment, thereby reducing the number of engineering rounds required to reach the final goal, and eventually increasing the success rate of building a synthetic cell.

Taken together, this thesis exemplifies how targeted combinatorial testing can be applied to the optimization of *in vitro* reconstituted multiprotein systems. In future research, the directed evolution methods explored in this work could be further exploited to enable the evolutionary engineering of larger synthetic genomes, and eventually an entire artificial cell.

Samenvatting

Het doel van deze thesis was het onderzoeken van methodologieën die het bouwen van een minimale, synthetische cel versnellen. Dit werd bereikt door het maken van gerichte combinatorische DNA libraries, door deze te onderwerpen aan verschillende screening- en selectiemethoden, en door nanopore sequencing analyses en visualisatiemethodes te verkennen die DNA sequentie aan systeemfunctie koppelen.

Gedurende miljarden jaren evolutie hebben zelfs de eenvoudigste cellen buitengewoon complexe genetische architectuur, ruimtelijke organisaties en moleculaire interacties opgebouwd. Deze complexiteit maakt het moeilijk om de rol van elk afzonderlijk element in de cel te ontdekken. Wat als we zelf een cel zouden kunnen maken waarbij we alle niet-essentiële complexiteit weglaten? Een minimale versie van een cel, bestaande uit goed gekarakteriseerde onderdelen, zou inzicht kunnen geven in welke cellulaire processen essentieel zijn voor leven en hoe interacties tussen genen en eiwitten leiden tot het uiteindelijke gedrag van een cel. Dat is het doel van het bouwen van een minimale, synthetische cel die in staat is zichzelf te repliceren, te onderhouden en te evolueren.

Het bouwen van een cel zonder dat een bouwtekening beschikbaar is, vereist het voortdurend schakelen tussen het testen van prototypes en daarvan leren om weer nieuwe prototypes te kunnen ontwerpen. Om deze onderzoeksmethodiek te versnellen, stellen we dat de rationele aanpak van prototypering moet worden aangevuld met methoden die grote aantallen prototypes, vaak (gedeeltelijk) willekeurig gegenereerd, tegelijk kunnen testen. *Directed evolution* (ook wel gedirigeerde evolutie of laboratoriumevolutie genoemd) is zo'n methode: een breed ontwikkelde techniek voor het optimaliseren van eiwitten, die enorm veel prototypes tegelijk kan testen. In ons lab bestaat het moleculaire framework van synthetische cellen uit liposomen waarin genen tot expressie komen doordat de liposomen DNA en een gezuiverd cel-vrij expressiesysteem (PURE system) bevatten. **Hoofdstuk 1** van deze thesis **bespreekt de huidige mogelijkheden van *directed evolution* methoden** voor het engineeren van zulke *in vitro* moleculaire frameworks. Dit overzicht laat zien dat *directed evolution* vooral is toegepast voor het optimaliseren van individuele eiwitten, maar nauwelijks voor het engineeren van *in vitro* systemen bestaande uit meerdere eiwitten. We identificeren de volgende uitdagingen bij het toepassen van *directed evolution* voor het bouwen van een synthetische cel: (i) de omvang van de ontwerpruimte, (ii) het kiezen van optimalisatie-targets en (iii) het ontbreken van spatiotemporele screening- en sorteermethoden voor liposomen die tijdsafhankelijke, morfologische of localisatie-eigenschappen vertonen. De volgende hoofdstukken laten zien hoe deze uitdagingen kunnen worden aangegaan.

Hoofdstuk 2 introduceert een nieuw protocol voor het **construeren van gerichte combinatorische plasmide libraries**, genaamd MOSAIC. Traditionele protocollen voor het

muteren van plasmiden, die afhankelijk zijn van DNA-amplificatie en/of kloneringsstappen, kunnen arbeidsintensief zijn wanneer veel plaatsen op één DNA-molecuul moeten worden gemuteerd. Daarom hebben we een methode voor het muteren van het *Escherichia coli* genoom – die meerdere loci tegelijk kan muteren – aangepast voor een nieuw doel: het muteren van plasmiden. Een target plasmide en DNA oligo's met de gewenste mutaties werden samen via elektroporatie ingebracht in *E. coli* cellen die het enkelstrengs DNA-annealingeiwit CspRecT (afkomstig van een bacteriofaag) tot expressie brengen. Tijdens replicatie van het plasmide medieert CspRecT de annealing van de oligo's aan het plasmide-DNA in de replicatievork, wat leidt tot de incorporatie van mutaties (mismatches, inserties of deleties) met een efficiëntie van 50-90% per mutatie-plek. Een combinatorische DNA library werd geconstrueerd met behulp van een oligo library die vier posities tegelijk target. Deze library werd gekarakteriseerd met long-read nanopore sequencing, waarbij kwaliteitscontrole werd toegepast op de gemuteerde regio's.

Hoofdstuk 3 onderzoekt de **impact van het veranderen van eiwittranslatie-initiatie snelheden** op de activiteit van DNA-gecodeerde eiwitsystemen. Met behulp van MOSAIC werden gerichte combinatorische libraries geconstrueerd voor een DNA-zelfreplicator bestaande uit twee genen en een fosfolipidesyntheseroute bestaande uit vier genen. Voor de DNA-replicator werden de best presterende varianten geselecteerd uit een library van 156 varianten met verschillende ribosoombindingsplaats (RBS) sequenties. Selectie gebeurde op basis van hun superieure zelfamplificatie. Voor de fosfolipidesyntheseroute werden twee libraries – een RBS-library van 10^4 varianten en een library van 10^8 varianten waarin de eerste zes codons van de vier genen waren gemuteerd – gescreend en de topvarianten gesorteerd via fluorescence-activated cell sorting (FACS). Nanopore sequencing maakte het mogelijk om fitnesslandschappen te reconstrueren waarin DNA-sequenties worden gekoppeld aan de activiteit van de eiwitsystemen. Deze landschappen boden inzicht in optimale RBS-sterktes en basensubstituties, en onthulden welke genen de functionaliteit van de eiwitsystemen het meest konden moduleren. Omdat de DNA libraries combinatorisch waren, konden we ook combinaties van mutaties bestuderen. De theoretische fitness van combinatorische varianten was zeer voorspellend voor de DNA-replicator library, wat wijst op het ontbreken van epistatische interacties tussen de RBS'en van de twee replicatiegenen. De voorspelbaarheid was lager voor de grotere fosfolipidesyntheseroute libraries. Al met al hebben we laten zien hoe combinatorisch testen kan worden uitgebreid van individuele eiwitten naar systemen bestaande uit meerdere eiwitten, en dat het veranderen van translatie-initiatie snelheden een effectieve manier is om DNA-replicatie en fosfolipidesynthese te moduleren op systeemniveau.

Hoofdstuk 4 beschrijft een methode voor het **sorteren van synthetische cellen op basis van hun fenotypes verkregen via foto's**. Deze methode onderzoekt de fenotypes van liposomen binnen niet-klonale populaties via microscopie. De gewenste varianten worden gemarkeerd door de fluorescentie van een eiwit te activeren met UV licht. Vervolgens worden de

gemarkeerde liposomen gesorteerd door FACS. Door het automatiseren van optische microscopie en *machine learning*-geassisteerde analyse van de foto's, demonstreerden we het screenen van 10^5-10^6 liposomen per dag, waarvan 100-1000 werden gesorteerd met een precisie van 80%. Aan de hand van een *in vitro* gereconstitueerde bacteriële eiwitoscillator, het Min-systeem, toonden we aan dat liposomen geselecteerd kunnen worden op basis van fluorescentie-intensiteit, eiwitlocalisatie, membraanmorfologie en dynamische fenotypes, en dat het fenotype kan worden gekoppeld aan de genetische inhoud van de liposomen. Al met al vergroot deze methode het scala aan fenotypes op basis waarvan liposomen kunnen worden gesorteerd en is daarom essentieel voor laboratoriumevolutie van synthetische cellen.

Door onderlinge afhankelijkheden tussen het DNA en de moleculaire hardware van synthetische cellen (het expressiesysteem) is een co-optimalisatiestrategie waarschijnlijk essentieel voor het bouwen van een synthetische cel. **Hoofdstuk 5** beschrijft een **co-optimalisatiestrategie** die door middel van **DNA-barcoding** de moleculaire samenstelling identificeert. De moleculaire samenstelling verwijst hier naar de componenten van het cytoplasma (bestaande voornamelijk uit het expressiesysteem) en het membraan. We schatten dat deze strategie het mogelijk zal maken om met FACS in één dag een library van 10^3-10^4 DNA-varianten te screenen in combinatie met 30-100 moleculaire varianten. Via proof-of-conceptexperimenten met wildtype DNA valideerden we dat de barcodes geen bias veroorzaken in het verrijken van topvarianten en dat ze de goed presterende moleculaire samenstellingen kunnen identificeren. Een gestroomlijnde productiemethode van liposomen met verschillende genetische en moleculaire inhoud, en het toepassen van *active learning* om de grote variabele ruimte en de relatief beperkte experimentele capaciteit te overbruggen, zullen de bruikbaarheid van de ontwikkelde strategie in de toekomst kunnen aantonen.

Tot slot beschrijft **Hoofdstuk 6** de **toekomstige perspectieven** op laboratoriumevolutie van synthetische cellen door drie methodologische verbeteringen uit te lichten. Ten eerste zullen niet alleen de best presterende varianten, maar ook varianten met lagere fitness gesorteerd moeten worden om hun (kwantitatieve) eigenschappen te analyseren. Ten tweede zou de karakterisering van synthetische cellen kunnen worden uitgebreid van alleen DNA-sequencing naar *multi-omics* analyses. Ten derde zou DNA amplificatie uit gesorteerde liposomen kunnen worden geoptimaliseerd voor een nauwkeurigere kwantificatie van variantfrequenties (fitness). Deze verbeteringen zullen de kennis die uit elke prototypingronde wordt verkregen vergroten, waardoor het aantal engineeringrondes dat nodig is om het einddoel te bereiken afneemt en de kans op het succesvol bouwen van een synthetische cel toeneemt.

Samenvattend laat deze thesis zien hoe gericht combinatorisch testen kan worden toegepast voor de optimalisatie van *in vitro* gereconstitueerde systemen die uit meerdere genen bestaan. In toekomstig onderzoek kunnen de methoden die verkend zijn in deze thesis worden uitgebreid naar de laboratoriumevolutie van grotere synthetische genomen en uiteindelijk naar die van een volledige synthetische cel.

1

Introduction

Can we create a minimal version of a cell, composed of well-characterized parts, to reveal which cellular processes are essential for life and how complex interaction networks give rise to biological function? In this chapter, we introduce this ambitious engineering goal and outline the molecular framework of such a minimal synthetic cell. Research methodologies are discussed that could accelerate its construction, highlighting directed evolution as a promising technique due to its high exploratory power. Although directed evolution methods have been well-established for optimizing single proteins, they have barely been applied to the engineering of multi-protein systems *in vitro*. This brings challenges in applying these methods for building a synthetic cell. After outlining these challenges, we give an overview of the chapters in this thesis addressing these challenges.

It is remarkable to realize that, over a time span of 3.5 to 4 billion years, the descendants of the last universal common ancestor (LUCA) diversified into multiple domains of life: the bacteria, archaea and eukaryotes. When looking at the onions growing in my vegetable garden, the bacteria in my yoghurt, and the cat on my lap, I am already amazed by the large diversity that life forms can take. Yet, all current cellular life forms share the same basis of a DNA genome, transcription-translation of DNA to RNA to proteins, and DNA replication. With each round of replication, changes can accumulate in the DNA sequence that slowly expand the diversity of cellular life.

Throughout the billions of years of evolution, even the simplest cells have built up extraordinary complexity through interactions between the cell's components and with the environment. Factors such as gene redundancy, pleiotropy, gene fusion, functional promiscuity and protein moonlighting have shaped complex networks that drive cellular functions (**Box 1**). Consequently, classical ways to study cells using single-gene inactivation studies (e.g., gene knockout or mutagenesis studies) often do not uncover the precise roles of genes[2,3]. Even for the organism with the smallest genome known among autonomously replicating organisms found in nature, *Mycoplasma genitalium* (580 kb, 483 protein-coding genes), and for the organisms with synthetic minimized genomes of *Mycoplasma mycoides*, *JCVI-syn3.0* (531 kb, 473 genes) and *JCVI-syn3A* (543 kb, 493 genes), the functions of about 15 to 30% of the genes essential for cell growth remained unknown after extensive mutagenesis studies[4–7]. Moreover, the reconstitution of several protein machineries from *Escherichia coli* in synthetic environments revealed protein functions that were hidden in the natural context of the cell[8]. We could question whether these 'hidden functions' could be relics from evolutionary history and physiologically irrelevant to the modern cell, or instead they are important to understand biological function, proposing a radical undermining of the traditional – too simplistic – view on well-defined roles of proteins in the cell[8].

Box 1 Examples of complexity observed in modern cells.

Interactions between the cell's components
E.g., protein-protein interactions constituting signaling pathways and gene regulation.

Gene redundancy
Multiple genes perform the same function (they rescue function after gene loss).

Pleiotropy
A single gene influences multiple phenotypic traits, either due to a single activity that influences many processes, or multiple activities each affecting a different process.

Gene fusion (multidomain proteins)
Leads to proteins with multiple domains each having a distinct evolutionary history.

Functional promiscuity
Single protein performing multiple tasks (e.g., catalytic activities or bindings). E.g., an enzyme can catalyze the conversion of multiple different substrate molecules.

Protein moonlighting
Single protein performing multiple tasks that are totally different from each other, e.g., through multiple distinct binding sites. The definition of protein moonlighting excludes multifunctional proteins that result from gene fusion[1].

The above examples suggest that studying biological function in native cellular environments leaves questions about the roles of genes and proteins unanswered. This can make us wonder: are all cell complexities really needed to understand the *essence* of a biological cell?

An engineering approach to study biology

Fundamental architectures of the cell – for instance genetic circuits and metabolic networks – that underly many important cellular functions have been studied by an **engineering approach**, where the molecular circuitry is described in terms of only its essential parts – the interacting genes and proteins – and how they give rise to the behavior of the whole[9]. Such an approach can be performed in relative isolation of all unessential complexity by building the simplest system comprised of well-characterized parts to realize a target function. For example, by designing a **range of putative** genetic circuits or metabolic pathways, we can deduce design principles for how these networks are wired in cells. Noor *et al.* used all known enzyme classes found across diverse organisms that operate on carbohydrates – each class representing one type of reaction – to generate *in silico* all possible reactions that can convert metabolic input molecules (sugars) to output molecules (the precursors to generate biomass)[10]. By **comparing the putative networks** to the network used in ***E. coli***, they found that many parts of the *E. coli* metabolic pathway could have been shorter. However, they found that the *in vivo* network follows a specific **design principle** as a minimal walk between a specific subset of the metabolites. In other words, this subset of metabolites can be converted into each other through the shortest possible enzymatic paths. Interestingly, these metabolites were found to be direct precursors for biomass production (e.g., nucleotides, amino acids, lipids), while the metabolites not connected to each other via the shortest paths were not direct biomass precursors. Instead, they acted as connector metabolites. This example shows that an engineering approach can (1) **extract design principles** and (2) **define the essential components** to achieve this "minimal walk" design principle.

Building a minimal synthetic cell

When extending this approach to the complete set of biological functions essential to a living cell, well-characterized minimal synthetic cells may ultimately define the design principles and minimal set of components essential to life. While synthetic cells on their own may not accurately mimic how complex biological cells work, **comparing real life forms to *potential* forms of life** may give insight into the architecture of biological cells, as illustrated with the metabolic network example in the previous paragraph. This would mean a shift in the field of biology from *observations* of life to *natural laws* that define what can and cannot be life[9].

Molecular framework of a minimal synthetic cell

A minimal synthetic cell can be defined as a self-maintaining and self-replicating unit capable of evolution[11,12]. The "basic" cell could consist of an information storage molecule (e.g., **DNA**), a **metabolism** and a boundary (**cell membrane**) to separate the information storage molecule

and metabolism from the environment[12] (**Fig. 1**). The metabolism can be defined as all chemical processes to keep the cell from thermodynamic equilibrium, including **enzyme-catalyzed chemical reactions** to produce chemical energy (ATP), building blocks (amino acids), and macromolecules (proteins)[13]. Operating out of equilibrium, enabled by the consumption of energy from the environment, is key for a living cell, because it drives biochemical processes and sustains ordered biological structures[14].

1. DNA

The estimated number of genes required for a minimal synthetic genome lies between 150 and 200 genes, translating to an estimated genome size of 150-200 kb[15]. Genes were selected based on the simplest and/or best-characterized biological processes without (too severely) compromising self-replication efficiency and accuracy. Because there is no consensus on the definition of a minimal synthetic cell, different sets of genes have been considered. In one proposition, Forster and Church included 151 genes from *E. coli* and viruses (38 RNAs and 113 proteins, in total 113 kb)[16]. In a second example, Luisi *et al*. included 146-150 genes from *M. genitalium*[11]. While both gene sets promote **DNA, RNA and protein synthesis**, Forster and Church additionally include genes for RNA processing and modification, chaperones for protein folding, and post-translational modification of translation release factors, whereas Luisi *et al*. include genes for **cell growth** such that the cells do not rely on spontaneous growth by simple absorption of lipids from the environment. For both studies, **cell division** and **DNA segregation** are not catalyzed by protein-based mechanisms but rely on spontaneous fission and statistical distribution, respectively. Experiments should demonstrate whether additional genes for membrane deformation, membrane abscission, DNA segregation, energy regeneration, DNA repair, transport and/or RNA- and protein-modifying proteins are desirable for a minimal synthetic cell. Arguments for or against genetically encoding these processes include: cell autonomy versus intervention by the researcher, stochastic versus more controlled processes, and the robustness and the level of complexity of the resulting cell[17]. For example, reliance on stochastic processes for cell division could lead to a poor self-replication efficiency and result in only a fraction of the daughter cells being vital. This could limit the practical use of these synthetic cells for research purposes or for potential biotechnological and medical applications, e.g., the production of toxic-to-cell chemicals or drug delivery. On the other hand, such stochastic processes may mimic more closely the self-replication of early protocells, being relevant for origin-of-life studies.

Reconstituting genes from existing organisms is currently the most effective way to replicate biological functions *in vitro*, because designing entirely new, non-natural proteins with emergent functions is extremely challenging[18]. However, existing protein systems are possibly more complex than required for a minimal synthetic cell (**Box 1**) and may give rise to unexpected functionalities *in vitro*[8] and constrain biological design and evolvability[19]. Therefore, natural proteins with desired functions likely require optimization in the new

synthetic cell environment. For functions not present in existing proteins, *de novo* protein design could used[20].

2. Metabolism

Transcription and translation of the genes from the synthetic genome must be kickstarted, for instance by a cell-free expression system, whereafter the expressed genes should take over the functions of the initial expression system. *E. coli* cell extract is a widely used cell-free expression system that produces relatively high yields of protein (2-4 mg/mL) compared to other cell extracts[21]. While components supplemented to the extract, such as buffers, ions, energy-storing molecules, cofactors and tRNAs, can be controlled, the extract itself has a poorly defined composition and does not allow easy tuning of its components. Another type of cell-free expression system, Protein synthesis Using Recombinant Elements (**PURE**) **system**[22], is made from purified components which are mixed at defined concentrations. Its composition is relatively well-defined and tunable. PURE system contains all required components to enable transcription, aminoacylation, translation and energy regeneration. Most components from PURE system are derived from *E. coli*, while transcription is catalyzed by the T7 bacteriophage RNA polymerase. Challenges associated with the use of PURE system are its relatively low protein yield (0.1-1 mg/mL), the production of truncated proteins and its short lifetime in batch reactions compared to cell extracts[21,23]. These factors also challenge the expression of multiple proteins in the same PURE reaction.

3. Cell membrane

Different types of artificial cell boundaries have been explored to compartmentalize cell-free expression, including lipid membranes, polymersomes, hydrogels and coacervates[24]. Liposomes provide a lipid membrane with structural properties similar to natural membranes, including selective molecular diffusion and compatibility with the incorporation of proteins from biological cells. They also provide a confined femto- to picoliter-scale cytoplasm similar in volume to that of natural cells. Liposomes compatible with cell-free gene expression have been produced via several methods, such as lipid film hydration[25], water-in-oil droplet transfer across an oil-water interface covered by a phospholipid monolayer[26,27] and droplet-based microfluidics[28]. The bulk methods (hydration and emulsion transfer) are simple and low-cost but produce polydisperse liposome populations, while droplet microfluidics produce highly monodisperse liposome populations with a narrow size and protein expression level distribution[28] but require specialized equipment and expertise.

In the Danelon Lab (**Fig. 1**), ~2- to 12-µm liposomes are produced via gentle hydration of lipid films dried on glass beads to boost the liposome yield[29,30]. This method is compatible with the encapsulation of PURE system and with the expression of cytosolic proteins as well as the expression and spontaneous association/insertion of bacterial membrane proteins[31,32].

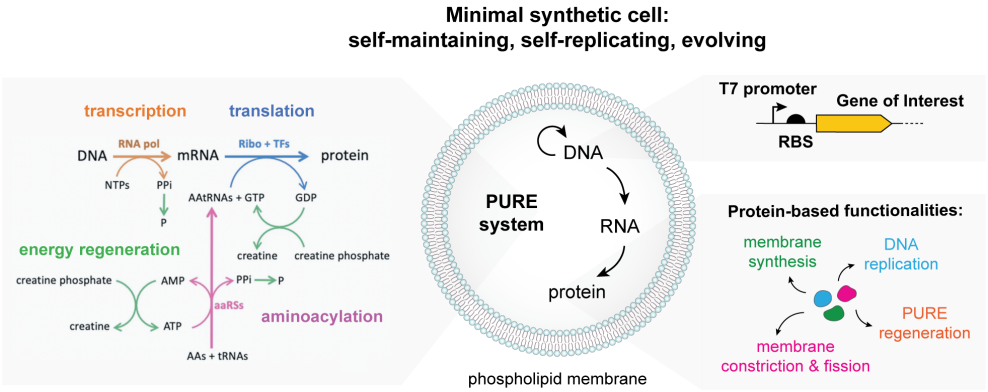


Fig. 1 Molecular framework of a minimal synthetic cell: the Danelon Lab approach. Minimal synthetic cells consist of liposomes, encapsulating DNA and PURE system. PURE system sustains four main types of reactions: transcription, translation, energy regeneration and aminoacylation. Gene expression by PURE system is controlled by regulatory elements such as T7 promoters and RBSs. We seek to express a minimal yet complete set of genes that enables self-maintenance, self-replication and evolution of the basic synthetic cell framework. Examples of envisioned cell functionalities that underly self-replication are the regeneration of the (protein-based) PURE components, DNA replication and membrane synthesis, constriction and fission. Schematic illustration of the biochemical reactions in PURE system (left) is adapted from Doerr *et al. Sci. Rep.* 11, 1898 (2021).

Research methodologies for building a minimal synthetic cell

The research methodology to building a minimal synthetic cell seems contradictory: we build a cell to understand how it can be wired. (This aligns with the famous quote from Richard Feynman: "What I cannot create, I do not understand"). "True" engineering of biological systems, which are not fully predictable systems, is not possible. Therefore, bio-engineering often requires optimization steps through slow processes including tinkering, trial-and-error and serendipity to achieve the engineering goals. The design process can be seen as employing a **cyclic iterative approach** where multiple *concepts* (i.e., design ideas whose viability is unknown) are translated into *knowledge* through testing their viability (**exploration**), and the gained knowledge is used to generate new design ideas (**exploitation**) (**Fig. 2a**)[33].

The **exploratory power** of the design process is the number of designs that can be tested per cycle multiplied by the number of cycles. While bio-engineering of a well-characterized system is feasible with a low exploratory power, solving bio-problems with a lack of *a priori* knowledge requires a research method with a high exploratory power. There exist different experimental methods spanning several orders of magnitude of experimental power (**Fig. 2b**), which will be discussed in the context of building a synthetic cell in the following sections.

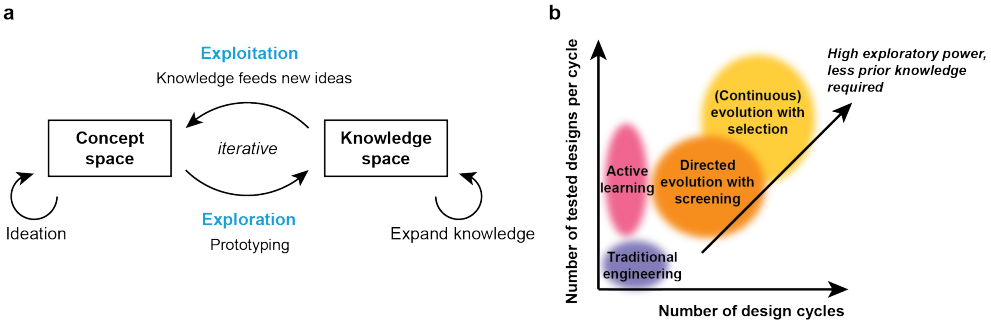


Fig. 2 Research methodology for bio-engineering. a The process of design. *Exploration* is building knowledge based on tested concepts; *exploitation* is using knowledge to constrain and guide ideas. **b** Research methods with different exploratory powers require different levels of prior knowledge for a successful engineering outcome. Figures are based on text and illustration from Castle *et al. Nat. Commun.* 15, 3640 (2024).

1. Traditional engineering

A biological module is a set of proteins, DNA, RNA and small molecules that together exert a specific function, that is separate from that of other modules[19]. Interactions between the components of a module give rise to the module's biological function. The modules can also be connected to other modules, upon which the modules can influence each other, or the modules are insulated, allowing biological processes to happen without influencing each other. Connections between modules can lead to higher-level functions, also referred to as **emergent properties**, which were not present in each of the individual modules and act at the system's level. Moreover, modularity in biological systems increases a system's robustness to perturbations, which allows the evolution of a system's parts and enables the integration of simpler parts into a more complex system[34]. For these reasons, a modular approach to building a synthetic cell could be a good starting point and has been widely adopted.

Many modules have already been reconstituted *in vitro* through transcription-translation of DNA templates, such as DNA replication[35–37], phospholipid synthesis[31,32,38,39], membrane remodeling and/or constriction[40–44] or partial regeneration of PURE system components[23,45–48]. These achievements have mostly been accomplished through a *traditional design approach*, employing an overall low exploratory power while building strongly on existing knowledge. Yet, many of these processes still require optimization. For example, with only a few percent increase in lipid content, phospholipid synthesis was not sufficient to observe growth of mother vesicles[32,38]. In addition, various strategies for controlled cell division, involving protein-directed membrane constriction completed by abscission, are still being explored[17,49–51]. As a last example, to enable complete PURE system regeneration, the current PURE system yield should be boosted with approximately two orders of magnitude and/or the PURE components should be modified or their relative stoichiometries optimized[48].

A cell is more than the sum of its parts

In addition to optimizing individual modules, a massive upcoming engineering challenge is the stepwise integration of modules. I here discuss some of the challenges. First, modules could lose activity or exhibit unintended behavior when combined with other modules. In a recent study on integrating a DNA replication system with a phospholipid synthesis pathway, the number of vesicles with active phospholipid synthesis was significantly reduced upon co-expressing the DNA replication module[52]. Possible engineering targets to restore activity include (i) improving robustness to changes in the environment (e.g., buffer, salt concentrations, pH, temperature); (ii) balancing gene expression within and across modules; (iii) removing unintended interactions between proteins, RNA, DNA, lipids and small molecules; and (iv) dealing with competition of multiple modules through the sharing of resources, for instance substrates or energy storage molecules[20]. Overall, conditions need to be found that are compatible with all modules.

Second, by introducing interactions between modules, emergent properties can be generated, as discussed earlier. For example, the combination of an FtsZ-based contractile ring and MinCDE-based cell pole-to-pole protein oscillators resulted in the equatorial positioning of the ring, which was promoted by the MinC-FtsZ interaction[44]. While in this example, the protein-protein interaction between the two systems is a naturally occurring interaction in the host cell *E. coli*, designing non-natural interactions or interactions between modules originating from different organisms may be more challenging. The engineering of emergent properties, for example self-organization, biological pattern formation, adaptability, autonomy, metabolism and evolution, could maybe be approached through the assets listed in **Box 1**, but remains an enormous challenge.

Another major challenge is the limited protein yield of PURE system and our inability to accurately model its processes to understand the limiting factors[53]. Moreover, we still rely on experimentally testing how regulatory sequences, coding sequences, and gene order affect gene expression in PURE system[54,55], as predictive tools developed to control protein production in *E. coli*[56] have not been extensively validated for *in vitro* expression. Lastly, integration may require spatial and temporal control over the different modules to establish a sustainable cell cycle[17]. Altogether, the optimization of the artificial cell system as a whole will become increasingly difficult with the size of the synthetic genome, where we can less and less rely on *a priori* knowledge. Therefore, the challenging tasks described above prompt us to shift from *traditional engineering* to methods with a higher exploratory power.

2. Directed evolution with screening

Randomness-leveraging strategies go beyond the constraints of rational design. In *directed evolution*, a widely developed protein engineering method[57], libraries of (semi-)random variants are created, genes are expressed, and the variants with the desired properties are selected (**Fig. 3a**). Typically, multiple of these cycles are iterated. In the context of synthetic

cells, these libraries could be a collection of different (a) **DNA sequences** or (b) **molecular compositions**. For the latter, the (concentrations of) the components of the "cytoplasm" (mainly the kickstarting expression system) or the membrane are varied.

(a) Optimize DNA sequences

Two main platforms have been developed to perform directed evolution in synthetic compartments: encapsulating and expressing DNA libraries either in phospholipid vesicles, also referred to as "liposome display", or in water-in-oil emulsion droplets[58]. Both platforms rely on encapsulating DNA libraries at a maximum of one variant per compartment to ensure a genotype-phenotype link. Generally, the function of the to-be-optimized gene is coupled to the expression of a fluorescent protein or is measured by a specific fluorescent reporter. The compartments expressing the genes are screened, and those with optimal function are sorted with high throughput using fluorescence-activated cell sorting (FACS)[59,60] or fluorescence-activated droplet sorting (FADS)[61].

Using liposome display, single genes have been optimized to obtain increased activity of an aminoacyl-tRNA synthetase[62], a multidrug transporter[63] or an α-hemolysin[64], and a 16S rRNA variant was obtained for the reconstitution of the 30S ribosome subunit without rRNA modification[65]. In these studies, the genes were randomly mutagenized by error-prone polymerase chain reaction (PCR) prior to encapsulation, resulting in 0-4 mutations per gene, and approximately 10^7 variants could be screened per day by FACS[60,66].

The droplet platform has been used more extensively[67]. Monodisperse emulsion droplets can be generated by mixing a reaction solution in oil and surfactants using microfluidic devices. Gan *et al.* specifically mutated the regulatory parts (promoter or ribosomal binding site (RBS); max. 10^6 mutants) of the *sfGFP* gene by performing PCR with degenerate primers[68]. They screened for high protein expression of the fluorescent sfGFP protein either directly on the microfluidic chip, or by creating double emulsions which are compatible with FACS sorting, leading to a higher throughput. Scheele *et al.* performed an ultra-high throughput screen (>10^7 variants per day) using a targeted library for a specific domain of a kinase protein[69]. By studying the interactions between multiple simultaneously mutated amino acids, they discovered a preferred sequence motif that promotes the kinase's phosphorylation activity.

The last study highlights the importance of combinatorial testing and the investigation of epistatic interactions. **Epistatic interactions** occur when the observed phenotype deviates from the expected phenotype when multiple mutations are combined. This leads to exceptionally high or low fitness values that are not directly predictable from analyzing each position in isolation (**Fig. 3b**). Thus, **combinatorial testing** can reveal DNA variants that would be hidden when testing each element in isolation and could give insight into the system's mechanistic properties (e.g., by revealing functional relationships between nucleotides, amino acids, genes and pathways).

The methods for screening emulsion droplets are advantageous because the droplets can be generated and screened directly on the chip, and the droplets have a high monodispersity compared to liposomes. However, these techniques are limited to the engineering of soluble proteins, while liposome display is also compatible with the expression of membrane proteins[66]. Thus, the liposome platform is ultimately favored for engineering synthetic cells.

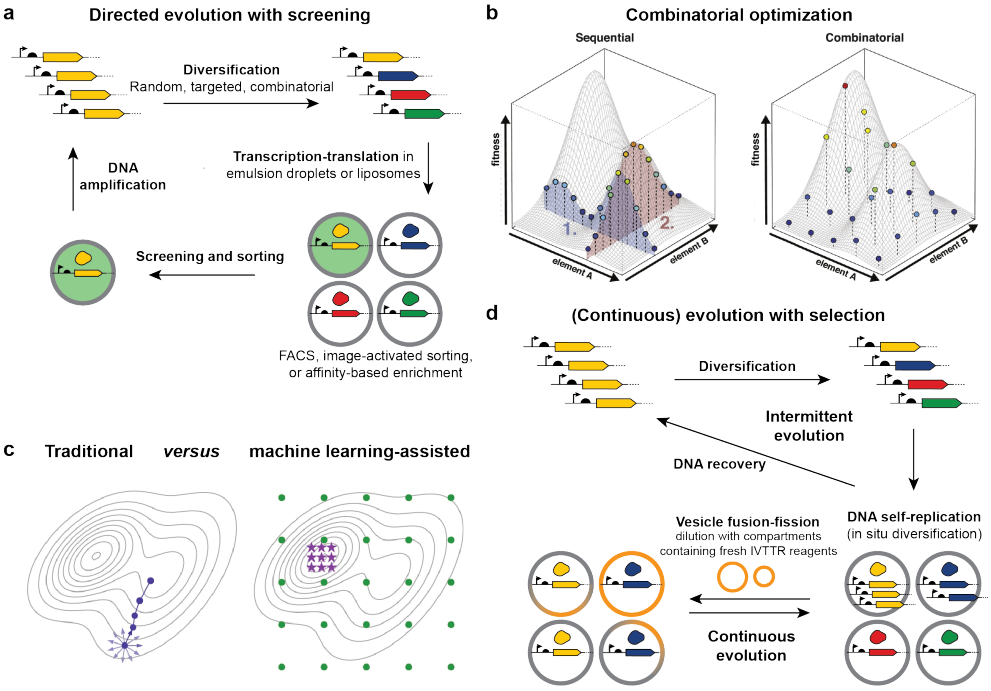


Fig. 3 Methodologies for *in vitro* compartmentalized directed evolution. a Directed evolution workflow with screening for functional DNA variants using a fluorescent reporter as a measure of variant fitness. Mutations are generally introduced by random or targeted mutagenesis prior to DNA encapsulation. With combinatorial diversification, multiple nucleotides or regions on the DNA template are mutagenized simultaneously. The vesicles with the highest fluorescence intensity are physically sorted from the population, and their DNA is amplified and sequenced and/or re-diversified for a next screening round. **b** Sequential versus combinatorial optimization. With sequential optimization, one element is optimized at a time. An optimized element is fixed at the highest fitness and used as starting point for the optimization of the next element. With combinatorial optimization, multiple elements are optimized simultaneously. Because nonlinear interactions can occur between elements, combinatorial optimization can uncover global fitness maxima that are not accessible through sequential optimization. Figure adapted from Jeschek *et al. Curr. Opin. Biotechnol.* 47, 142-151 (2017). **c** Fitness landscapes for traditional versus machine learning-assisted directed evolution. In the left scheme (traditional directed evolution), local searches around a starting point within the fitness landscape lead to improvements in fitness until the search gets stuck in a local maximum. In the right scheme (machine learning-assisted directed evolution), first a global search is performed that covers a larger part of the search space, resulting in a higher chance of converging towards the global fitness optimum. Figure adopted from Yang *et al. Nat. Methods* 16, 687–694 (2019). **d** Intermittent or continuous evolution employing selection of variants through DNA self-replication.

(b) Optimize molecular compositions

The initial molecular composition of synthetic cells is essential to boot up gene expression, support protein folding, and for the functionality of the expressed protein modules. It is important to find the optimal conditions for the best performance of the system as a whole. To illustrate this, Seo and Ichihashi studied the compatibility between transcription, translation, and DNA replication in PURE system[70]. They found that dNTP, rNTPs, tRNAs and potassium glutamate have opposing effects on the three reactions (improving or inhibiting), which could be partly alleviated by adjusting the magnesium concentration.

Similar as for optimizing DNA sequences, for molecular optimizations it is also important to screen the sample space combinatorially, rather than optimizing each component sequentially (**Fig. 3b**). Matsuura *et al*. quantified the epistatic interactions among 69 components in PURE system[71]. The amount of protein produced when changing the concentrations of two PURE system components was on average 16% higher or lower than what would be expected based on the effects of the individual changes[71]. Epistasis could arise from direct physical interactions between e.g., enzymes, but also from indirect physical interactions, e.g., through competition between enzymes for substrates or cofactors when these are limiting[71]. Matsuura *et al*. showed that taking into account interactions between two components of PURE system, and neglecting larger than 2-body interactions, provided a good approximation of epistatic effects in PURE system[71]. Nevertheless, the search space is still incredibly complex, since a wide range of concentrations could in principle be tested for each component. Due to the large combinatorial search space, optimization of cell-free expression systems has been assisted by machine learning, which is further explained in the next paragraph.

3. Active learning-assisted directed evolution

In traditional directed evolution screenings, the fitness information from unimproved variants is discarded, and an improved variant is used as starting point for a next local search in the fitness landscape. In contrast, in machine learning-assisted screening, the *entire* function landscape is modeled based on prior experimental data or computationally designed variants, whereafter a small set of variants with high predicted fitness is selected for experimental validation[72]. The predictive models help to efficiently navigate the search space and escape local minima (**Fig. 3c**). Traditional and machine learning-assisted directed evolution can be considered "***greedy searches***". In contrast, **active learning** allows for deeper ***exploration*** before converging to the fittest variants by iteratively collecting data and improving the machine learning model[73]. Active learning has been used to efficiently scan large search spaces to improve gene expression, including bacterial lysate compositions[74–77] and combinatorial DNA libraries of spacer or regulatory sequences[74,75].

4. (Continuous) evolution with selection

Instead of physically sorting subpopulations of variants (i.e., variant screening followed by sorting), *selection* can be exploited to enrich high-fitness variants. With selection, a desired function promotes the self-replication or the survival of DNA or cells. Hereby, variants with desired functions are enriched in the population while other variants are depleted or eliminated (**Fig. 3d**). This method requires less intervention by the researcher and is not reliant on a reporter signal. In contrast to directed evolution using *screening*, the exploratory power of the *selection* method is thus not limited by a screening rate (e.g., the FACS screening rate) but rather by the number of gene-expressing compartments that can be produced.

Using the selection strategy, adaptive evolution of DNA replicators encoding the Phi29 DNA polymerase has been achieved in water-in-oil droplets and liposomes. This was successful for rolling-circle replication in the presence of a recombinase[36,37,78] as well as for linear DNA replication by co-expressing the DNA polymerase with the Phi29 terminal protein that acts as replication primer[79]. In these studies, mutations were introduced randomly, both during PCR when preparing the DNA template prior to encapsulation, and during replication by the Phi29 DNA polymerase.

Selection can also be applied to optimize functions other than those directly involved in DNA replication. To do so, the function of interest should be *coupled* to replication. Di Meo *et al.* made DNA self-replication conditional to the activity of a deoxyribonucleoside kinase (which produces nucleoside triphosphates) by supplying dNMPs as precursors of dNTPs. Instead of relying on random mutations, they constructed a DNA library prior to encapsulation, targeting the mutations to the region encoding the kinase and its RBS[80]. Sierra *et al.* showed that DNA replication can also be made conditional to functions that are not related to the Central Dogma (transcription, translation, DNA replication)[81]. They exemplified this by enriching for β-galactosidase activity converting lactose into allolactose, which induces the transcription of the Phi29 replication proteins.

Both screening (e.g., by FACS) and selection through DNA self-replication work with an *intermittent* evolution scheme, where the DNA is extracted from the compartments after each round, and re-encapsulated for a subsequent round. However, selection also allows for a more *continuous* evolution scheme, where the DNA is not repeatedly recovered from the compartments (**Fig. 3d**). Instead, the compartments can be diluted with extra compartments containing fresh cell-free expression system. Upon fusion and fission of the compartments with replicated DNA and the fresh compartments, the replicated DNA is redistributed for a next round of evolution[37,79]. With continuous evolution, DNA diversification relies on mutations introduced by the Phi29 DNA polymerase and is thus more difficult for the researcher to control.

The main challenges using directed evolution for building a synthetic cell

In summary, as no defined blueprint exists for building a synthetic cell, directed evolution would be a valuable complement to rational design due to its high exploratory power. There are various ways to generate diversity (random, targeted, combinatorial) and enrich fit DNA variants after expression in synthetic compartments. Considering evolution as a key property of life may suggest that the *selection* approach, where – as in *natural* selection – traits in populations of organisms that promote reproduction are enriched, could help in building a synthetic cell[13]. However, directed evolution through screening is indispensable to engineer functionalities that are difficult to couple to DNA or cell replication[82].

While directed evolution has been well established for optimizing single proteins – both *in vivo* and *in vitro*[57,58], it has barely been applied to the simultaneous engineering of multiple genes *in vitro*. One of the biggest challenges when applying directed evolution to the optimization of synthetic genomes, is the **extremely large search space**. The search space is not only vast due to the long DNA sequence (150-200 kb) and the many variable molecular components, but also because of the large combinatorial space. A large number of genes could increase the number of interactions in the system. These can be both direct interactions between DNA, RNA or proteins, or indirect interactions via other molecular components. Such interactions can make the effects of variant combinations difficult to predict. Consequently, at least some degree of combinatorial testing is required, leading to an extremely large design space that cannot be exhaustively explored experimentally.

A semi-rational approach that limits mutations to specific regions of the DNA sequence could help in efficiently navigating the large sequence space. The key challenge here lies in choosing those regions on the DNA template that are **most likely to improve** the functions of interest.

Lastly, the lack of established techniques for high-throughput **spatiotemporal screening and sorting of vesicles** challenges the assessment of functions with time-dependent morphological or localization features, such as membrane growth, constriction and fission and the distribution of DNA over daughter cells. While such multidimensional phenotypes are commonly characterized by optical microscopy, an image-based screening and sorting method for vesicles is lacking.

Dissertation outline

In this thesis, we develop methodologies for the laboratory evolution of minimal synthetic cells. Synthetic cell engineering involves both the genome and the molecular hardware (i.e., cytoplasm and membrane). In Chapters 2 to 4, we focus on DNA optimization, while in Chapters 5 and 6, we also address the optimization of molecular compositions.

Chapter 2 introduces a new protocol for the construction of **targeted combinatorial plasmid libraries** named MOSAIC. This fast method exploits *E. coli* as the library generation machinery and introduces desired mutations at multiple target sites at once.

Using DNA libraries constructed by MOSAIC, **Chapter 3** studies the **impact of tuning protein translation initiation rates** on the activity of modules composed of multiple genes. Combinatorial DNA libraries are constructed by mutating RBSs or bases in the first codons of genes involved in DNA replication or phospholipid synthesis. After enriching the top performing variants, either through FACS-based screening or selection through DNA self-replication, experimental fitness landscapes are reconstructed that map the DNA sequences to the system's level activities.

In **Chapter 4**, we expand the FACS-based screening and sorting method from Chapter 3 with a microscopy-based pre-screening step to enable **image-based phenotypic sorting** of synthetic cells. The pre-screening step translates rich spatiotemporal features into a simplified one-dimensional readout that can subsequently be measured by FACS. By automating optical microscopy and machine learning-assisted image analysis, we demonstrate the sorting of vesicles based on fluorescence intensity, protein localization, membrane morphology and dynamic behaviors.

In **Chapter 5**, we shift our focus from solely genome optimization to include optimization of the molecular hardware. This chapter describes the design of a workflow for the **co-optimization of genetic and molecular compositions** and demonstrates its feasibility by proof-of-concept experiments.

Finally, **Chapter 6** describes the **future perspectives** on laboratory evolution of synthetic cells by identifying three potential methodological improvements and discussing how these could be implemented in a specific application involving a minimal DNA-encoded cell cycle.

References

1. Jeffery, C. J. Moonlighting proteins. *Trends in Biochemical Sciences* **24**, 8–11 (1999).
2. Glass, J. I., Merryman, C., Wise, K. S., Hutchison, C. A. & Smith, H. O. Minimal cells—real and imagined. *Cold Spring Harbor Perspect. Biol.* **9**, a023861 (2017).
3. Regenmortel, M. H. V. V. Reductionism and complexity in molecular biology. *EMBO Rep.* **5**, 1016–1020 (2004).
4. Glass, J. I. *et al.* Essential genes of a minimal bacterium. *PNAS* **103**, 425–430 (2006).
5. UniProt Mycoplasma genitalium Proteome UP000000807. *UniProt* https://www.uniprot.org/proteomes/UP000000807.
6. Hutchison, C. A. *et al.* Design and synthesis of a minimal bacterial genome. *Science* **351**, aad6253 (2016).

7. Breuer, M. *et al.* Essential metabolism for a minimal cell. *eLife* **8**, e36842 (2019).
8. Schwille, P. & Frohn, B. P. Hidden protein functions and what they may teach us. *Trends Cell Biol.* **32**, 102–109 (2022).
9. Elowitz, M. & Lim, W. A. Build life to understand it. *Nature* **468**, 889–890 (2010).
10. Noor, E., Eden, E., Milo, R. & Alon, U. Central carbon metabolism as a minimal biochemical walk between precursors for biomass and energy. *Mol. Cell* **39**, 809–820 (2010).
11. Luisi, P. L., Oberholzer, T. & Lazcano, A. The notion of a DNA minimal cell: a general discourse and some guidelines for an experimental approach. *Helv. Chim. Acta* **85**, 1759–1777 (2002).
12. Caschera, F. & Noireaux, V. Integration of biological parts toward the synthesis of a minimal cell. *Curr. Opin. Chem. Biol.* **22**, 85–91 (2014).
13. Abil, Z., Giaveri, S., Erb, T. J. & Rothschild, L. J. Integrating metabolism and evolution towards the realization of synthetic life. *Nat. Rev. Bioeng.* **3**, 9–10 (2025).
14. Ornes, S. How nonequilibrium thermodynamics speaks to the mystery of life. *PNAS* **114**, 423–424 (2017).
15. Cleij, C. Building the genome of a minimal synthetic cell. (TU Delft, 2025).
16. Forster, A. C. & Church, G. M. Towards synthesis of a minimal cell. *Mol. Syst. Biol.* **2**, 45 (2006).
17. Olivi, L. *et al.* Towards a synthetic cell cycle. *Nat. Commun.* **12**, 4531 (2021).
18. Kohyama, S., Frohn, B. P., Babl, L. & Schwille, P. Machine learning-aided design and screening of an emergent protein function in synthetic cells. *Nat. Commun.* **15**, 2010 (2024).
19. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).
20. Frohn, B. P., Kohyama, S. & Schwille, P. Protein design and optimization for synthetic cells. *Nat. Rev. Bioeng.* **3**, 645–659 (2025).
21. Hunt, A. C. *et al.* Cell-free gene expression: methods and applications. *Chem. Rev.* **125**, 91–149 (2025).
22. Shimizu, Y. *et al.* Cell-free translation reconstituted with purified components. *Nat. Biotechnol.* **19**, 751–755 (2001).
23. Doerr, A., Foschepoth, D., Forster, A. C. & Danelon, C. In vitro synthesis of 32 translation-factor proteins from a single template reveals impaired ribosomal processivity. *Sci. Rep.* **11**, 1898 (2021).
24. Cho, E. & Lu, Y. Compartmentalizing cell-free systems: toward creating life-like artificial cells and beyond. *ACS Synth. Biol.* **9**, 2881–2901 (2020).
25. Nomura, S. M. *et al.* Gene expression within cell-sized lipid vesicles. *ChemBioChem* **4**, 1172–1175 (2003).
26. Noireaux, V. & Libchaber, A. A vesicle bioreactor as a step toward an artificial cell assembly. *PNAS* **101**, 17669–17674 (2004).
27. Van de Cauter, L. *et al.* Optimized cDICE for efficient reconstitution of biological systems in giant unilamellar vesicles. *ACS Synth. Biol.* **10**, 1690–1702 (2021).
28. Gonzales, D. T., Yandrapalli, N., Robinson, T., Zechner, C. & Tang, T.-Y. D. Cell-free gene expression dynamics in synthetic cell populations. *ACS Synth. Biol.* **11**, 205–215 (2022).
29. Nourian, Z., Roelofsen, W. & Danelon, C. Triggered gene expression in fed-vesicle microreactors with a multifunctional membrane. *Angew. Chem. Int. Ed.* **51**, 3114–3118 (2012).
30. Blanken, D., van Nies, P. & Danelon, C. Quantitative imaging of gene-expressing liposomes reveals rare favorable phenotypes. *Phys. Biol.* **16**, 045002 (2019).
31. Scott, A. *et al.* Cell-free phospholipid biosynthesis by gene-encoded enzymes reconstituted in liposomes. *PLOS ONE* **11**, e0163058 (2016).
32. Blanken, D., Foschepoth, D., Serrão, A. C. & Danelon, C. Genetically controlled membrane synthesis in liposomes. *Nat. Commun.* **11**, 4317 (2020).
33. Castle, S. D., Stock, M. & Gorochowski, T. E. Engineering is evolution: a perspective on design processes to engineer biology. *Nat. Commun.* **15**, 3640 (2024).
34. Huitzil, S. & Huepe, C. Life's building blocks: the modular path to multiscale complexity. *Front. Syst. Biol.* **4**, (2024).
35. van Nies, P. *et al.* Self-replication of DNA by its encoded proteins in liposome-based synthetic cells. *Nat. Commun.* **9**, 1583 (2018).

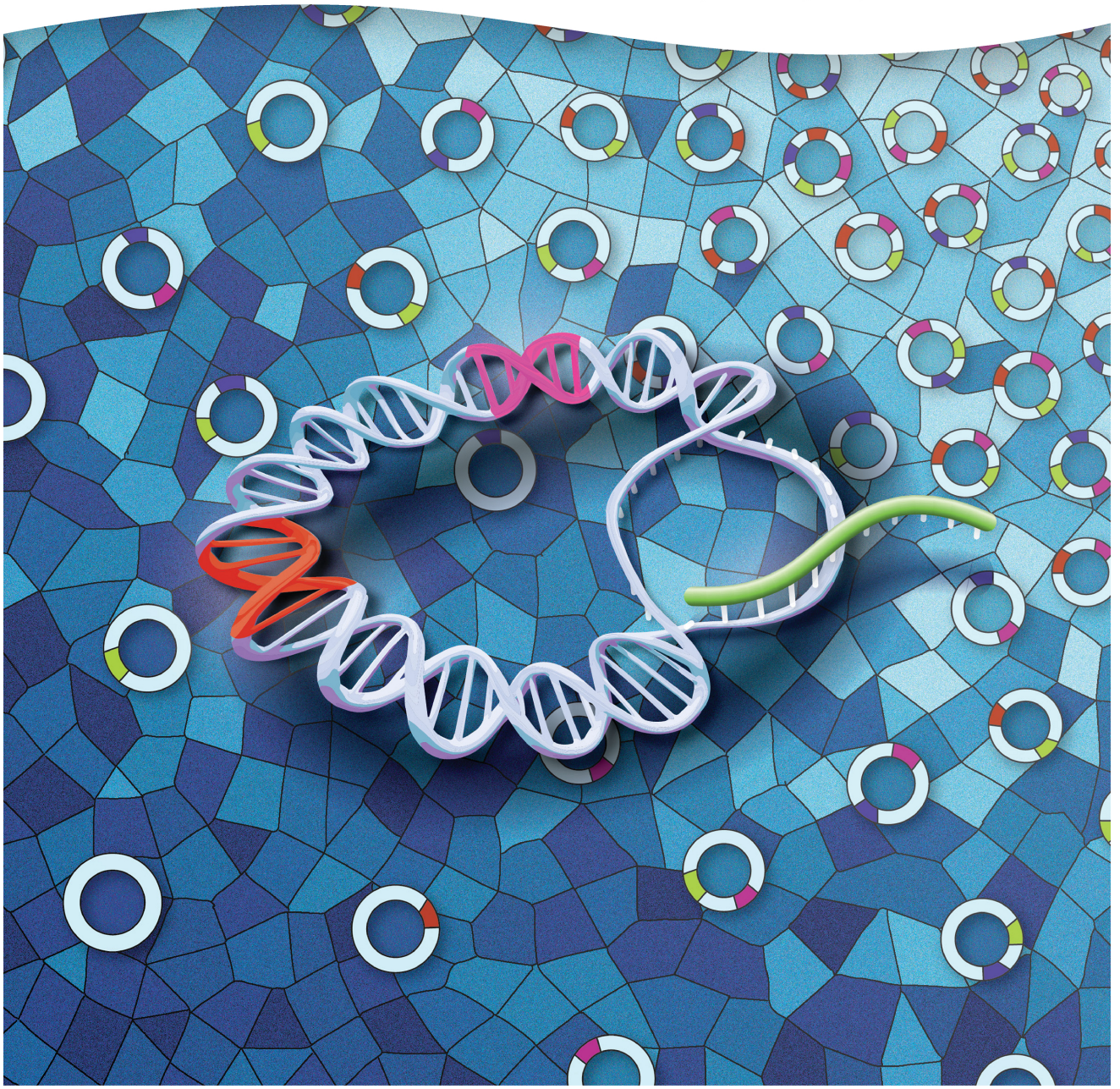
36. Sakatani, Y., Yomo, T. & Ichihashi, N. Self-replication of circular DNA by a self-encoded DNA polymerase through rolling-circle replication and recombination. *Sci. Rep.* **8**, 13089 (2018).
37. Okauchi, H. & Ichihashi, N. Continuous cell-free replication and evolution of artificial genomic DNA in a compartmentalized gene expression system. *ACS Synth. Biol.* **10**, 3507–3517 (2021).
38. Eto, S. *et al.* Phospholipid synthesis inside phospholipid membrane vesicles. *Commun. Biol.* **5**, 1–11 (2022).
39. Bhattacharya, A., Brea, R. J., Niederholtmeyer, H. & Devaraj, N. K. A minimal biochemical route towards de novo formation of synthetic phospholipid membranes. *Nat. Commun.* **10**, 300 (2019).
40. Furusato, T. *et al.* De novo synthesis of basal bacterial cell division proteins FtsZ, FtsA, and ZipA inside giant vesicles. *ACS Synth. Biol.* **7**, 953–961 (2018).
41. Godino, E. *et al.* De novo synthesized Min proteins drive oscillatory liposome deformation and regulate FtsA-FtsZ cytoskeletal patterns. *Nat. Commun.* **10**, 4969 (2019).
42. Godino, E. *et al.* Cell-free biogenesis of bacterial division proto-rings that can constrict liposomes. *Commun. Biol.* **3**, 539 (2020).
43. Kattan, J., Doerr, A., Dogterom, M. & Danelon, C. Shaping liposomes by cell-free expressed bacterial microtubules. *ACS Synth. Biol.* **10**, 2447–2455 (2021).
44. Kohyama, S., Merino-Salomón, A. & Schwille, P. In vitro assembly, positioning and contraction of a division ring in minimal cells. *Nat. Commun.* **13**, 6098 (2022).
45. Hibi, K. *et al.* Reconstituted cell-free protein synthesis using in vitro transcribed tRNAs. *Commun. Biol.* **3**, 350 (2020).
46. Miyachi, R., Shimizu, Y. & Ichihashi, N. Transfer RNA synthesis-coupled translation and DNA replication in a reconstituted transcription/translation system. *ACS Synth. Biol.* **11**, 2791–2799 (2022).
47. Hagino, K. & Ichihashi, N. In vitro transcription/translation-coupled DNA replication through partial regeneration of 20 aminoacyl-tRNA synthetases. *ACS Synth. Biol.* **12**, 1252–1263 (2023).
48. De Capitani, J. & Mutschler, H. The long road to a synthetic self-replicating central dogma. *Biochemistry* **62**, 1221–1232 (2023).
49. Godino, E. & Danelon, C. Gene-directed FtsZ ring assembly generates constricted liposomes with stable membrane necks. *Adv. Biol.* **7**, 2200172 (2023).
50. Panevska, A., Šakanović, A., Paccione, G., Rivas, G. & Schwille, P. Optimizing spatial organization of FtsZ rings for large-scale constriction in synthetic cells. *Nat. Commun.* **17**, 2320 (2026).
51. De Franceschi, N., Barth, R., Meindlhumer, S., Fragasso, A. & Dekker, C. Dynamin A as a one-component division machinery for synthetic cells. *Nat. Nanotechnol.* **19**, 70–76 (2024).
52. Sierra, A. M. R., Gomez, F. R., Tongeren, M. van, Heras, L. S. & Danelon, C. A synthetic cell with integrated DNA self-replication and membrane biosynthesis. 2025.01.14.632951 Preprint at https://doi.org/10.1101/2025.01.14.632951 (2025).
53. Doerr, A. *et al.* Modelling cell-free RNA and protein synthesis with minimal systems. *Phys. Biol.* **16**, 025001 (2019).
54. Chizzolini, F., Forlin, M., Cecchi, D. & Mansy, S. S. Gene position more strongly influences cell-free protein expression from operons than T7 transcriptional promoter strength. *ACS Synth. Biol.* **3**, 363–371 (2014).
55. Chizzolini, F. *et al.* Cell-free translation is more variable than transcription. *ACS Synth. Biol.* **6**, 638–647 (2017).
56. 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).
57. Packer, M. S. & Liu, D. R. Methods for the directed evolution of proteins. *Nat. Rev. Genet.* **16**, 379–394 (2015).
58. Seo, K., Hagino, K. & Ichihashi, N. Progresses in cell-free in vitro evolution. in *Cell-free Production: System Development* (eds. Lu, Y. & Jewett, M. C.) 121–140 (Springer Nature Switzerland, Cham, 2023). doi:10.1007/10_2023_219.
59. Bernath, K. *et al.* In vitro compartmentalization by double emulsions: sorting and gene enrichment by fluorescence activated cell sorting. *Anal. Biochem.* **325**, 151–157 (2004).

60. Nishikawa, T., Sunami, T., Matsuura, T., Ichihashi, N. & Yomo, T. Construction of a gene screening system using giant unilamellar liposomes and a fluorescence-activated cell sorter. *Anal. Chem.* **84**, 5017–5024 (2012).
61. Baret, J.-C. *et al.* Fluorescence-activated droplet sorting (FADS): efficient microfluidic cell sorting based on enzymatic activity. *Lab Chip* **9**, 1850–1858 (2009).
62. Uyeda, A. *et al.* Liposome-based in vitro evolution of aminoacyl-tRNA synthetase for enhanced pyrrolysine derivative incorporation. *ChemBioChem* **16**, 1797–1802 (2015).
63. Uyeda, A., Nakayama, S., Kato, Y., Watanabe, H. & Matsuura, T. Construction of an in vitro gene screening system of the E. coli EmrE transporter using liposome display. *Anal. Chem.* **88**, 12028–12035 (2016).
64. Fujii, S., Matsuura, T., Sunami, T., Kazuta, Y. & Yomo, T. In vitro evolution of α-hemolysin using a liposome display. *PNAS* **110**, 16796–16801 (2013).
65. Murase, Y., Nakanishi, H., Tsuji, G., Sunami, T. & Ichihashi, N. In vitro evolution of unmodified 16S rRNA for simple ribosome reconstitution. *ACS Synth. Biol.* **7**, 576–583 (2018).
66. Fujii, S. *et al.* Liposome display for in vitro selection and evolution of membrane proteins. *Nat. Protoc.* **9**, 1578–1591 (2014).
67. Gantz, M., Neun, S., Medcalf, E. J., van Vliet, L. D. & Hollfelder, F. Ultrahigh-throughput enzyme engineering and discovery in in vitro compartments. *Chem. Rev.* **123**, 5571–5611 (2023).
68. Gan, R. *et al.* High-throughput regulatory part prototyping and analysis by cell-free protein synthesis and droplet microfluidics. *ACS Synth. Biol.* **11**, 2108–2120 (2022).
69. Scheele, R. A. *et al.* Droplet-based screening of phosphate transfer catalysis reveals how epistasis shapes MAP kinase interactions with substrates. *Nat. Commun.* **13**, 844 (2022).
70. Seo, K. & Ichihashi, N. Investigation of compatibility between DNA replication, transcription, and translation for in vitro Central Dogma. *ACS Synth. Biol.* **12**, 1813–1822 (2023).
71. Matsuura, T., Kazuta, Y., Aita, T., Adachi, J. & Yomo, T. Quantifying epistatic interactions among the components constituting the protein translation system. *Mol. Syst. Biol.* **5**, MSB200950 (2009).
72. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687–694 (2019).
73. Yang, J. *et al.* Active learning-assisted directed evolution. *Nat. Commun.* **16**, 714 (2025).
74. Caschera, F. *et al.* Coping with complexity: Machine learning optimization of cell-free protein synthesis. *Biotechnol. Bioeng.* **108**, 2218–2228 (2011).
75. Pandi, A. *et al.* A versatile active learning workflow for optimization of genetic and metabolic networks. *Nat. Commun.* **13**, 3876 (2022).
76. Borkowski, O. *et al.* Large scale active-learning-guided exploration for in vitro protein production optimization. *Nat. Commun.* **11**, 1872 (2020).
77. Zhu, J. *et al.* AI-driven high-throughput droplet screening of cell-free gene expression. *Nat. Commun.* **16**, 2720 (2025).
78. Sakatani, Y., Mizuuchi, R. & Ichihashi, N. In vitro evolution of phi29 DNA polymerases through compartmentalized gene expression and rolling-circle replication. *Protein Eng., Des. Sel.* **32**, 481–487 (2019).
79. Abil, Z. *et al.* Darwinian evolution of self-replicating DNA in a synthetic protocell. *Nat. Commun.* **15**, 9091 (2024).
80. Di Meo, T. *et al.* Functional selection in a population of synthetic cells with a minimal metabolism. 2025.08.25.672132 Preprint at https://doi.org/10.1101/2025.08.25.672132 (2025).
81. Danelon, C. & Heras, L. S. Autocatalytic selection of gene functions in vitro. 2025.08.25.672052 Preprint at https://doi.org/10.1101/2025.08.25.672052 (2025).
82. Abil, Z. & Danelon, C. Roadmap to building a cell: an evolutionary approach. *Front. Bioeng. Biotechnol.* **8**, (2020).

October 2025 • Volume 14, Issue 10

ACS Synthetic Biology

pubs.acs.org/acssynthbio



ACS Publications
Most Trusted. Most Cited. Most Read.

www.acs.org

2

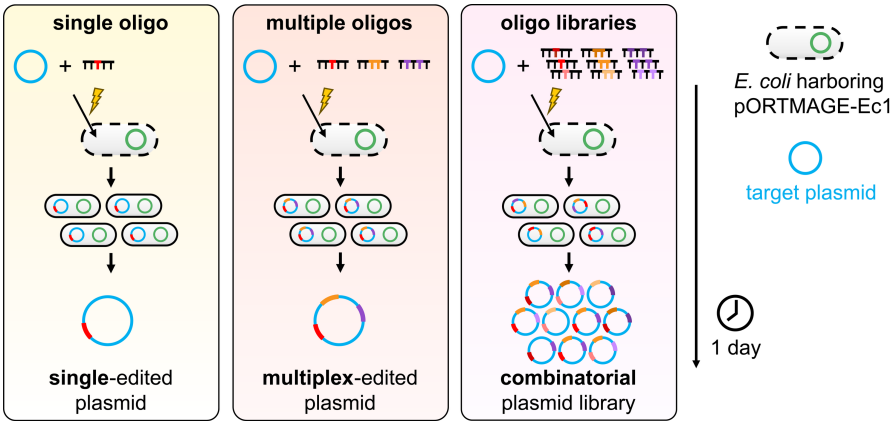
MOSAIC: A highly efficient, one-step recombineering approach to plasmid editing and diversification

The editing of plasmids and construction of plasmid libraries is paramount to the engineering of desired functionalities in synthetic biology. Typically, plasmids with targeted mutations are produced through time- and resource-consuming DNA amplification and/or cloning steps. In this study, we establish MOSAIC, a highly efficient protocol for the editing of plasmids and generation of combinatorial plasmid libraries. This quick protocol employs the efficient single-stranded DNA annealing protein (SSAP) CspRecT to incorporate (libraries of) DNA oligos harboring the desired mutations into a target plasmid in *Escherichia coli*. In addition to up to 90% single-target plasmid editing efficiency, we demonstrate that MOSAIC enables the generation of a combinatorial plasmid library spanning four different target regions on a plasmid, in a single transformation. Lastly, we integrated a user-friendly validation pipeline using Nanopore sequencing reads, requiring minimal computational experience. We anticipate that MOSAIC will provide researchers with a simple, rapid and resource-effective method to edit plasmids or generate large, diverse plasmid libraries for a wide range of *in vivo* or *in vitro* applications in molecular and synthetic biology.

Manuscript published as: van den Brink, M.[1], Althuis, T. Y.[1], Danelon, C., Claassens, N. J. *ACS Synthetic Biology*, 14, 3880-3889 (2025). doi:10.1021/acssynbio.4c00657. [1] shared co-authorship

Illustration on previous page: Artistic impression of MOSAIC, featured on the journal cover. I thank Maarten Wilming for his valuable input in the design and refinement of the illustration.

Graphical abstract



Introduction

The development of tools that enable the construction of DNA parts and variants thereof are driving the field of synthetic biology. With the continuing advances in modelling and machine-learning, our predictive capabilities and *a priori* design of functional genetic systems and proteins are rapidly improving[1–5]. Still, *in silico* designed genetic parts and proteins often do not behave as expected in the complex genetic and molecular contexts of cells or cell-free expression systems. As a result, efforts in pathway, genetic circuit and protein engineering often benefit from the exploration of wide solution spaces provided by (semi-)rational design and/or computational tools[6,7].

Plasmid libraries offer an efficient means to test a range of designs *in vivo* or *in vitro*. Generally, mutant plasmids are constructed from DNA fragments amplified by polymerase chain reaction (PCR) with degenerate primers to introduce the desired variation at specific locations. Subsequently, the produced DNA fragments are (re)assembled into a plasmid *in vitro* or *in vivo* by enzymatic assembly[8,9]. For example, site-directed mutagenesis methods based on high-fidelity polymerases and mutagenic primers are widely used (e.g., QuikChange or Q5 site-directed mutagenesis). However, these methods can only diversify one region at a time. Alternative approaches, including restriction enzyme-based and homology-based assembly methods (e.g., Golden Gate, Gibson Assembly, In-Fusion Cloning or Ligase Cycling Reaction) can create plasmids from multiple DNA parts harboring mutations, but are limited in their efficiency and flexibility for the generation of large combinatorial libraries[10–13]. Specifically, the number of correctly assembled clones decreases as the number of assembly parts increases. Therefore, researchers need to upscale their experimental efforts or rely on laboratory automation to obtain sufficient numbers of clones to generate larger libraries. These challenges are exacerbated for combinatorial libraries with multiple target sites (multiplexing), large plasmids resulting in low transformation efficiencies, or plasmids containing repetitive

DNA. Altogether, there is a need for efficient and flexible strategies to generate large, multiplex plasmid libraries.

Recombineering is a widely used tool to introduce targeted and scarless modifications in bacterial genomes[14]. This approach relies on single-stranded DNA (ssDNA) or double-stranded DNA (dsDNA) fragments containing desired mutations, which are introduced into the organism, usually by electroporation. These DNA fragments are then incorporated into replicating chromosomes using phage-derived ssDNA-annealing proteins (SSAPs). This system is extensively employed for genome engineering, especially in *Escherichia coli*, to make large insertions and deletions using dsDNA recombineering and small edits using ssDNA[15,16]. Multiplex automated genome engineering (MAGE) builds on the latter by introducing mutations to many genomic loci at the same time using iterative, automated or manual, editing cycles[17].

Whilst underutilized, ssDNA-mediated recombineering has also been employed to modify plasmids in *E. coli*[18–24]. Initially, plasmid recombineering efficiencies of 5-10% were observed for single point mutations with the phage λ-derived SSAP Recβ and two sequential transformations of *E. coli,* first with the target plasmid and then with the mutagenic ssDNA oligos[19]. Later, co-electroporation of an optimized ratio of mutagenic ssDNA and the target plasmid yielded editing efficiencies of 20-30%[20]. This was further improved to 60% when combined with a co-selection strategy, in which a restriction site on the plasmid is simultaneously mutated, whereafter unmodified variants are eliminated by restriction digestion. Higher efficiencies have also been obtained by combining recombineering with counterselection of non-mutated variants by a CRISPR-Cas nuclease[22,23]. However, this approach requires additional, time-consuming cloning steps and complicates the experimental design and setup. To the best of our knowledge, only a few studies have applied recombineering-based approaches to produce diversified and multiplex plasmid libraries[20,21,24]. Presumably, the low efficiency of plasmid recombineering and laborious methods relying on co- or counterselection explain the limited application range thus far.

Recently, a systematic screen of phage SSAPs in *E. coli* identified CspRecT, which has a two-fold higher genomic recombineering efficiency than the commonly used Recβ[25]. This prompted us to develop MOSAIC: a multiplex one-step SSAP-mediated plasmid diversification protocol. Its name is derived from *mosaicism*, a phenomenon where mutations give rise to distinct genetic compositions within an organism or a cell population. In this study, we show that MOSAIC's high plasmid editing efficiency enables the generation of large combinatorial plasmid libraries in a single transformation. Furthermore, MOSAIC employs a validation methodology based on Nanopore long-read sequencing, which quantifies the frequency of (multiplex) library variants directly from the plasmid library sample. We believe that the easy experimental and sequence validation protocols of MOSAIC will facilitate plasmid diversification and expand its range of applications throughout many laboratories.

Results and discussion

CspRecT-mediated recombineering achieves ~85% single-locus plasmid editing efficiency

We investigated the efficiency of plasmid recombineering using SSAP CspRecT expressed from plasmid pORTMAGE-Ec1 (**Figure 1A**)[25]. The plasmid recombineering efficiency was first tested with the high-copy number plasmid pUC19. Two types of ssDNA oligos were designed to introduce a single-nucleotide deletion or a three-nucleotide mismatch in the *lacZ* gene on pUC19. The oligos contained two phosphorothioate bonds at the 5′ ends to protect against degradation by exonucleases *in vivo*[17]. Because pUC19 replicates unidirectionally in a DNA sequence-controlled manner[26–28], we designed the oligo sequences such that they target the lagging strand during plasmid replication, as this is believed to lead to the highest efficiency during recombineering[14,20,29]. The ssDNA oligos (2.5 μM) were co-electroporated with 1 ng of pUC19 plasmid into electrocompetent *E. coli* MG1655 expressing CspRecT and the dominant negative *E. coli* MutL mutant (EcMutL^{E32K}) for temporal repression of mismatch repair. Usually, deleterious mutations in *lacZ* can be quantified using blue-white screening on an LB agar plate with X-gal. However, as plasmid recombineering leads to mixed plasmid populations in single colonies, we determined the editing efficiencies by DNA sequencing. Hence, after overnight growth, the plasmids were isolated, and the editing efficiency was quantified from Nanopore sequencing reads. The Nanopore sequencing analysis pipeline is shown in **Figure S1**. Only reads with a quality threshold ≥50 in the target region were used to reduce the chance of incorrect detection of mutations to < 1% (**Figure S2**).

Remarkably, the editing efficiency after a single round of MOSAIC was 88% and 83% for the single-nucleotide deletion and three-nucleotide mismatch, respectively (**Figure 1B**). As expected, the use of complementary oligos that bind the leading strand in the replication fork resulted in lower editing efficiencies (74% and 54% for the single-nucleotide deletion and three-nucleotide mismatch, respectively) (**Figure 1B**). The addition of both the leading and lagging strand-targeting oligos did not further improve the editing efficiency. The observed editing efficiencies were more than double those previously reported for plasmid editing with the Redβ SSAP[20]. This coincides with the previously observed two-fold increase in genomic editing efficiency with CspRecT versus Redβ[25], highlighting the large impact of SSAP on the editing efficiency for both genomic and plasmid recombineering.

Importantly, when the oligos were electroporated into bacteria already harboring the target plasmid (i.e., "2-step electroporation"), the fraction of edited plasmids was very low (<7%) (**Figure 1B**). So, co-electroporation of the target plasmid and ssDNA is key to reach high plasmid editing efficiencies. This is in agreement with a previous study using Redβ for plasmid recombineering[20]. The large effect of co-electroporation is likely explained by the fact that recombineering is most effective during plasmid replication. When a single plasmid or a low

number of plasmids enters the cell, the plasmid(s) will likely be rapidly replicated many times to reach the copy number at which the plasmid is maintained in the cells.

Overall, plasmid recombineering resulted in much higher efficiencies than recombineering on the *E. coli* genome, whose highest reported efficiency is ~50% but in our hands reached only 14% based on a blue-white screening (**Figure S3A**)[25]. We also tested the MOSAIC protocol in *E. coli* DH5α, which is routinely used for transformation and cloning purposes. However, the observed plasmid editing efficiency in this strain was lower than in *E. coli* MG1655 (**Figure 1B**). Hence, unless stated otherwise, *E. coli* MG1655 was used for recombineering in the remainder of this study.

Higher-copy plasmids are edited more efficiently than low-copy plasmids

As we hypothesized that the high editing efficiency was coupled to a high plasmid replication rate, we anticipated that higher-copy plasmids with comparatively higher replication rates after electroporation would be edited more efficiently than lower-copy plasmids[20]. To test this, we applied MOSAIC to a series of pSEVAb vectors that differed only in their origins of replication and, consequently, the copy number at which they are maintained in *E. coli*[30]. The selected origins of replications were RK2 (low-copy number), pBBR1, p15A, and pBR322/ROP (medium-copy number), and pRO1600/ColE1 and pUC (high-copy number)[31,32] (**Figure S4**). We designed oligos to incorporate a deletion or a three-nucleotide mismatch in the gene encoding sfGFP present in all plasmids. We identified the plasmid leading and lagging strands based on the known class B theta replication mechanism of ColE1 and ColE1-like origins (pRO1600/ColE1, pUC, pBR322 and p15A)[26–28,33,34], and similarly for the class A theta replication mechanism of the RK2-plasmid origin[28,35,36]. Based on this, we tested the oligos targeting the lagging strand assuming this would lead to the highest recombineering efficiency. To the best of our knowledge, the precise replication mechanism of pBBR1-derived plasmids is still unknown. Therefore, we tested both (reverse complementary) oligos for this plasmid, which performed equally well (**Figure S3B**). As anticipated, the vector with the low-copy RK2 origin of replication was edited with the lowest efficiency, on average 30%, followed by 58% for the pBBR1 origin of replication (**Figure 1C**). Surprisingly, the four other plasmids that we evaluated were all modified with 70-90% efficiency. As such, it appears that beyond a certain copy number, additional replication events no longer increase the efficiency at which a single oligo is incorporated. Such a threshold might be due to a saturation effect, possibly caused by exceeding a time window during which the recombineering machinery and/or oligos are sufficiently active, and may represent an upper boundary for MOSAIC.

Large insertions and deletions are incorporated with high efficiency

To probe the potential broad applicability of MOSAIC, we investigated if a high editing efficiency could still be obtained with a substantially larger number of mutations per oligo. If so, larger regions, such as regulatory sequences (e.g., promoters, RBSs and operator sites), could readily be inserted, deleted, replaced or diversified. Hence, we designed three ssDNA

oligos that incorporate an 18-nt wide insertion, deletion or mismatch into the *lacZ* gene on pUC19. The efficiency for the insertion and deletion was 65-70% (**Figure 1D**). The efficiency for substituting 18 nucleotides was lower (30%), but still sufficient for many of the aforementioned applications. Additionally, the large substitution was 5-fold more efficient than the earlier reported substitution of identical length in the *E. coli* genome[25]. Altogether, these results demonstrate that MOSAIC is a powerful method to edit and diversify both small and larger regions of plasmid DNA.

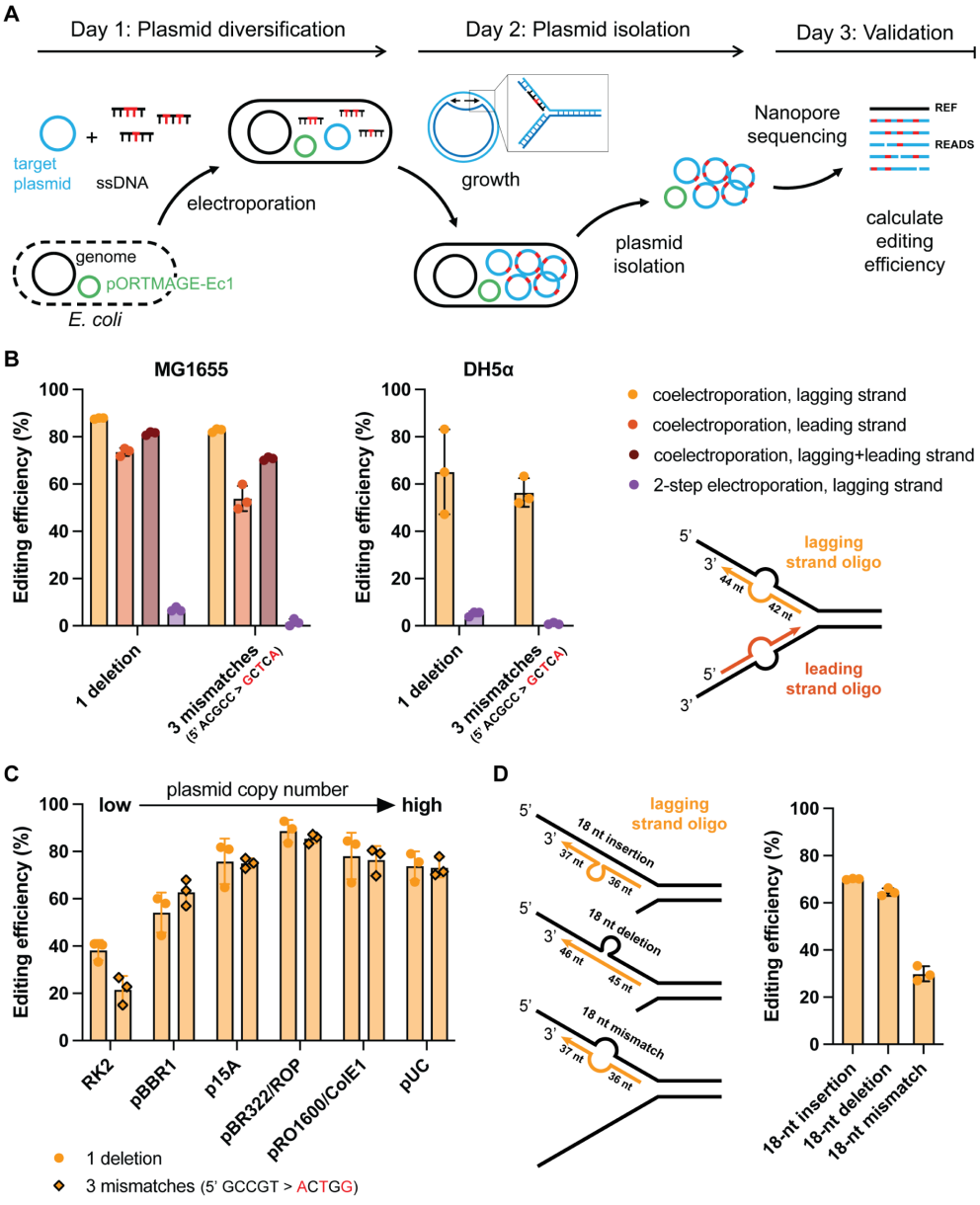


Figure 1. Plasmid editing efficiency of MOSAIC. A) Schematic representation of MOSAIC's protocol for plasmid editing using ssDNA recombineering. The expression of SSAP CspRecT and the dominant negative mutant EcMutL^{E32K} is induced in *E. coli* cells harboring pORTMAGE-Ec1. The cells are made electrocompetent and transformed with the target plasmid and mutagenic ssDNA oligos. During plasmid replication, the oligos anneal to one of the two DNA strands in the replication fork with the help of CspRecT, introducing mutations into the plasmid sequence. Plasmids are then isolated from the cells, and the editing efficiency is calculated from Nanopore sequencing reads. **B)** Plasmid editing efficiencies for the incorporation of a single-nucleotide deletion or a three-nucleotide mismatch in the *lacZ* gene in high-copy number plasmid pUC19. We investigated the effects of lagging strand versus leading strand oligos, the use of co- or 2-step electroporation and the use of *E. coli* strains MG1655 or DH5α. **C)** Plasmid editing efficiency for the incorporation of a single-nucleotide deletion or a three-nucleotide mismatch in the gene encoding sfGFP on the pSEVAb plasmids with different origins of replication. **D)** DNA editing efficiency for the incorporation of an 18-nt long insertion, deletion or mismatch in the *lacZ* gene on the high-copy number plasmid pUC19. For panels C and D, lagging strand oligos were used, and they were co-electroporated into *E. coli* MG1655 together with the plasmid DNA.

One round of MOSAIC yields a large multiplex plasmid library

Next, we investigated if we could apply MOSAIC to create a combinatorial plasmid library of ~10^4 variants in a single electroporation step. As a proof-of-principle, we diversified the ribosome binding sites (RBSs) of four genes encoding a phospholipid synthesis pathway on the pUC19-derived plasmid G555 (**Figure 2A**)[37]. Mutations in RBS sequences are expected to change the absolute and relative abundances of the four encoded proteins, hence phospholipid production. This enzymatic cascade from *E. coli* has been reconstituted in cell-free systems[37] and is amenable to phenotypic characterization of RBS modifications both *in vivo* and *in vitro*. To modulate translation, RBS Calculator was used to design RBS variants with a wide range of predicted translation initiation rates (1). The resulting variants contained up to 7 mismatches per RBS relative to the wild-type DNA and yielded a total library of 13 x 11 x 9 x 9 = 11,583 theoretical DNA variants (**Figure 2B**). Because two of the four target sites were identical, three degenerate oligo libraries targeting four different sites on the plasmid were sufficient to produce the combinatorial library. In a single electroporation reaction, the G555 plasmid and a 1:1:1 mix of the three degenerate oligo libraries were transformed into *E. coli* MG1655 cells expressing CspRecT and EcMutL^{E32K}. G555 contains multiple highly similar sequences (e.g., transcriptional promoters, RBSs and terminators) and is prone to recombination in *E. coli* MG1655. As such, the cells were directly plated on selective agar plates to prevent that some cells, harboring incomplete plasmids with a lower expression burden, outcompete the cells with full-length plasmids. Such competition would be a potential risk if the libraries were cultivated in liquid media. Following incubation overnight, the plasmid libraries from four plasmid recombineering transformations were isolated and sequenced.

Through Nanopore sequencing, the target loci are (physically) linked in a single read enabling the identification of the full genotype of each plasmid variant. Using the analysis pipeline outlined in **Figure S1**, the RBS variants in the Nanopore reads mapping to the phospholipid synthesis pathway genes were compared to the library variants designed by RBS Calculator

and to the wild-type RBSs, and their frequencies were counted. Based on 150-300 Nanopore reads, we calculated the per-locus editing efficiencies as 48% ± 3% (*plsB* RBS), 51% ± 5% (*plsC* RBS), 53% ± 5% (*cdsA* RBS) and 52% ± 6% (*pssA* RBS) (mean ± standard deviation, n=4) (**Figure 2C**, left graph). Thus, the editing frequency was consistent across the target sites and across replicates. With an average editing frequency of 51%, we expected that approximately 7% (i.e., 0.51^4) of the DNA molecules would have mutations in all four RBS sequences. However, we observed that 21 ± 3% (n=4) of the sequenced library variants had all four RBS sequences mutated (**Figure 2D**, left graph). This represents an unexpectedly high multiplexing efficiency of 21% after a single round of plasmid engineering. This suggests that there is a subpopulation of cells or plasmids with a higher-than-average editing efficiency and, thus, a higher chance that all loci are edited simultaneously. In addition, 79% of the library variants had at least one target site mutated, 57% at least two, and 39% had at least three target sites mutated. In contrast, previous work seeking to modify multiple sites on a plasmid with the Redβ recombinase observed >1 mutation in only 25% of their library[20]. Additionally, Higgins *et al.* observed 2 mutated target sites in 25% of their population after 5 rounds of plasmid recombineering[21].

To estimate the library coverage, we sequenced one of the four plasmid libraries in more depth. In the previous, shallow sequencing, we obtained only ~50 high-quality reads that spanned all target loci in a single read. To obtain as many reads spanning the full mutagenized region of interest on G555 as possible, the library was purified from pORTMAGE-Ec1 by retransformation and linearized at a unique restriction site outside the region of interest. We obtained ~300,000 high-quality reads spanning all four target sites. Of these reads, ~60,000 or 22% had all four RBS sequences mutated, echoing the multiplexing efficiency we observed in our shallow sequencing dataset (**Figure 2D**). Furthermore, all designed RBS variants were present with similar frequencies (**Figure S5**), and 99% of the library variants were present in the same order of magnitude (**Figure S6A**), suggesting unbiased RBS diversification. Altogether, we detected 2,839 of the 11,583 designed library variants, representing a library coverage of at least 25%. We believe this value represents a minimum, as 38% of the library variants were represented by one or two reads, suggesting that deeper sequencing may be required for full coverage of the DNA library. This was also indicated by the rarefaction curve (**Figure 6SB**), which was still increasing at 60,000 reads, suggesting that the sequencing depth is limiting the number of variants detected and the library size is likely larger than 2,839 variants.

Notably, only ~1-3% of RBS sequences were neither the wild-type sequence nor a designed library variant ("Other", **Figure 2C**). These unintended mutations are likely sequencing errors that were not excluded by our analysis pipeline (**Figure S2B**), mutations incorporated by oligos intended to bind at other target sites, and/or spontaneous mutations retained due to the suppression of the mismatch repair system during recombineering. The low fraction of unintended mutations is particularly noteworthy given that the oligos share an almost identical

40-nt left arm. It appears that having only one unique arm in the oligos is sufficient for the specific introduction of mutations. This is especially useful for mutating, for example, multiple 5'UTR or terminator regions with high sequence similarity.

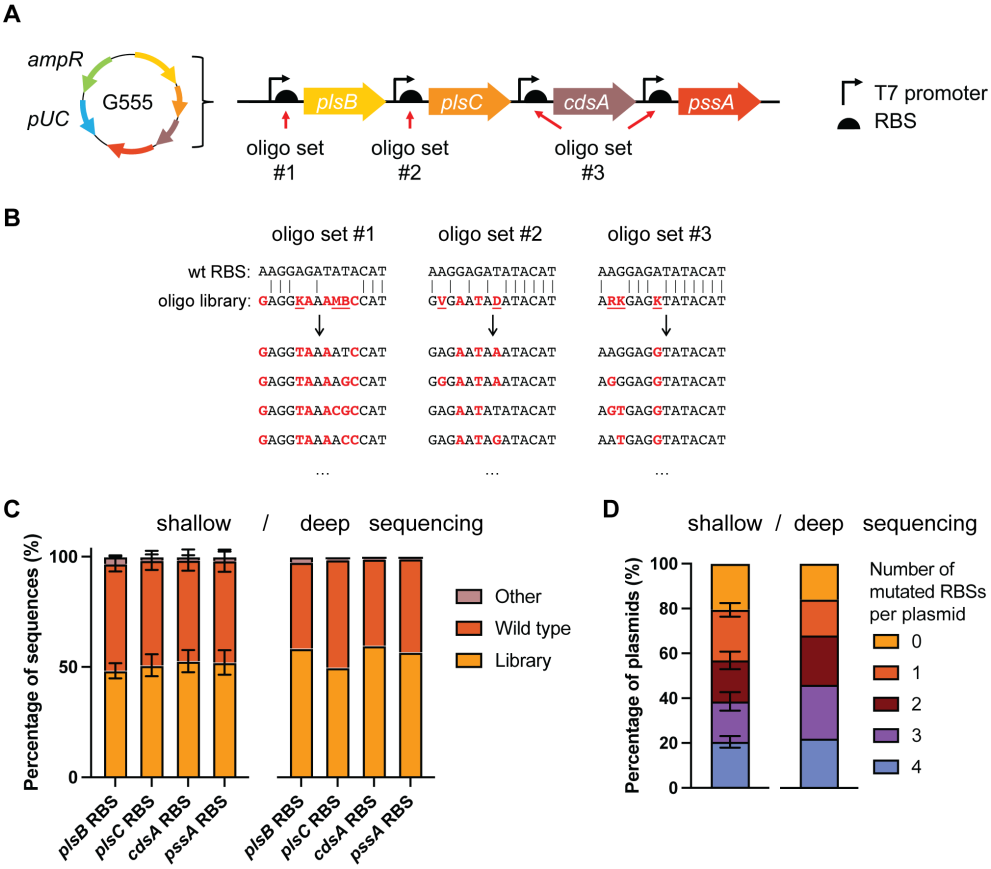


Figure 2. Construction of a multiplex plasmid library with four diversified RBSs using MOSAIC. A) pUC-derived target plasmid G555 containing four genes from the Kennedy phospholipid biosynthesis pathway (*plsB*, *plsC*, *cdsA* and *pssA*). Three ssDNA oligo sets of degenerate sequences (9-13 variants per set) targeted four regions on plasmid G555. **B)** The three sets of ssDNA oligos were designed using RBS Calculator [1] to target the RBSs of four genes. Letters in red indicate changes relative to the wild-type DNA. Underlined letters indicate degenerate nucleotides. **C)** Percentage of wild-type and library sequences per target locus, quantified from shallow Nanopore sequencing (left graph, n=4 with each 150-300 reads per target locus) or a deep sequencing dataset (right graph, n=1 with 279,454 reads per target locus). **D)** Number of mutated RBSs per plasmid, quantified from shallow Nanopore sequencing (left graph, n=4 with each 20-50 full-length reads) or the deep sequencing dataset (right graph, n=1 with 279,454 full-length reads).

Lastly, we tested whether performing successive rounds of plasmid recombineering leads to the accumulation of mutations at the four target regions in plasmid G555. As G555 is prone to recombine in MG1655, we performed these experiments in DH5α harboring pORTMAGE-Ec1. Following each round of plasmid recombineering, the target plasmid was selected for in liquid media overnight, the plasmids were isolated, and 30 ng was used as the starting point for the subsequent round. The editing efficiencies were quantified from Nanopore sequencing reads for three successive rounds. Despite the slightly lower plasmid editing efficiency observed for DH5α (**Figure 1B**), we observed an increase in the number of mutated RBSs per plasmid with each subsequent round of plasmid recombineering (**Figure S7A**). Interestingly, the *plsB* locus was edited roughly >2 fold more often than the other loci (**Figure S7B**). Altogether, these results demonstrate that multiple rounds of retransformation and recombineering can be used to increase the frequency of mutations across multiple target loci on a plasmid.

To conclude, these results demonstrate that MOSAIC enables the generation of large, diversified plasmid libraries in a single transformation. Moreover, key parameters for library characterization can be accurately determined by commercial, fast and low-cost Nanopore sequencing.

Attaining clonality: purifying plasmid mixtures after recombineering

In assembly-based cloning methods, cells or colonies typically harbor a single plasmid variant after transformation. In contrast, plasmid recombineering yields colonies with multiple DNA variants, as mutations are incorporated in the plasmids after uptake by the cells. Single-locus plasmid edits yield colonies harboring both the mutant and wild-type plasmids. To determine the number of plasmid variants present in a single colony following multiplex plasmid recombineering, we miniprepped and sequenced plasmid mixtures from six single colonies following our G555 RBS diversification experiment. On average, we found 6 ± 4 (n=6 colonies) different DNA variants per colony (**Table S2**). To ensure that the modified plasmids are suitable for an application of interest, the following steps can be taken to attain clonality and/or remove pORTMAGE-Ec1 from the plasmid pool (**Figure 3A**).

To obtain clonality and remove the pORTMAGE-Ec1 helper plasmid, plasmids can be isolated from the MOSAIC-derived population and used for retransformation of *E. coli* or other hosts, while only selecting for target plasmids. We demonstrate this by retransforming *E. coli* with the plasmid mixture from an earlier, single-locus pUC19 editing experiment. After transformation and growth of *E. coli* DH5α on plates selective for pUC19, plasmids from several colonies were individually extracted, analyzed on agarose gel and sequenced. Following retransformation, the band in the gel corresponding to the size of pORTMAGE-Ec1 was no longer present, suggesting the successful removal of pORTMAGE-Ec1 (**Figure S8**). This was confirmed by Nanopore sequencing as none of the colonies yielded reads that mapped to pORTMAGE-Ec1. More importantly, the sequencing data showed that each individual colony

was associated with a single genotype; there was a strict separation of mutated and wild-type plasmids (**Figure 3B**). Four of the five (replicate 1) and five out of five (replicate 2) colonies harbored the mutated variant, which is in line with the high single-locus editing efficiency of plasmids with this origin of replication (~83%). With these results, we verify clonality at the sequence level and demonstrate that modified and unmodified plasmids can be untangled after plasmid recombineering by a simple retransformation.

Due to the attained clonality, we presume that libraries generated by plasmid recombineering can be applied directly *in vivo* by retransformation into a strain of interest, as long as a sufficient number of transformants are obtained to ensure adequate library coverage. Alternatively, libraries can be applied *in vitro* following pooled plasmid isolation. To demonstrate recovery of the mutagenized G555 library whilst purifying the mixture of pORTMAGE-Ec1, we retransformed 155 ng of the mutagenized RBS plasmid library into highly electrocompetent NEB 10-beta cells yielding around 500,000 colonies, which is sufficient to cover the full library. Subsequent pooled plasmid isolation, digestion and Nanopore sequencing confirmed complete removal of pORTMAGE-Ec1, as no sequencing reads could be mapped to pORTMAGE-Ec1. Alternatively, the target DNA could be amplified by PCR directly from the mixture of target plasmids and pORTMAGE-Ec1.

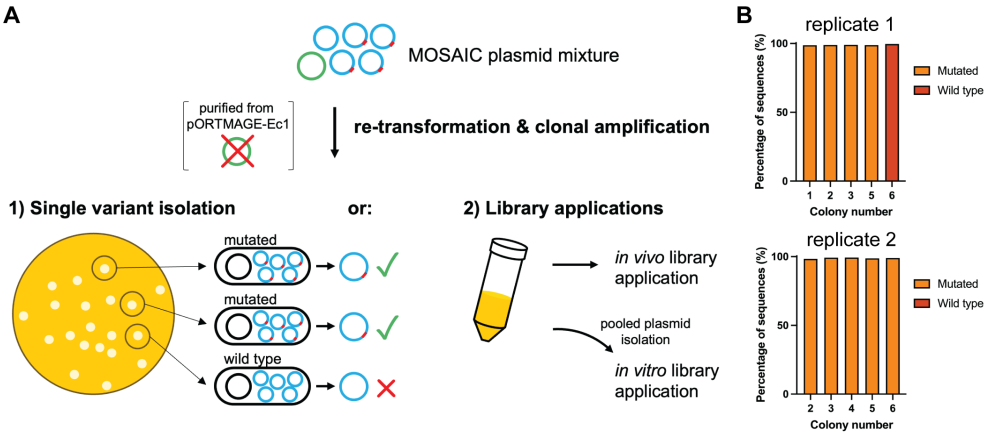


Figure 3. Retransformation of plasmid mixtures generated by plasmid recombineering is sufficient to remove pORTMAGE-Ec1 and establish clonality. A) Pure, mutated plasmids can be isolated by single-colony picking after retransformation. Plasmid libraries can be purified from pORTMAGE-Ec1 by retransformation, using either a cloning strain (e.g., DH5α or NEB 10-beta electrocompetent cells) or presumably the strain of interest for an *in vivo* application. For *in vitro* applications, libraries can be isolated from the cloning strain after retransformation. **B)** Clonality of re-transformants from single-edited plasmids.

Conclusions

In contrast to state-of-the-art plasmid editing methods, such as Golden Gate and Gibson assembly, MOSAIC does not require PCR or plasmid assembly from fragments. The only requirements for MOSAIC are fast-to-order mutagenic oligos and a publicly available *E. coli* strain harboring pORTMAGE-Ec1. A simple co-electroporation of the target plasmid and oligos is sufficient to perform the desired mutagenesis. This study has built on earlier work by increasing the complexity of plasmid engineering through the new, more efficient pORTMAGE-Ec1 system. The remarkably high editing efficiency and user-friendly library sequence characterization protocols, made accessible to scientists with minimal computational experience, now enable the use of plasmid recombineering and library sequence analysis on a regular basis.

The simplicity of MOSAIC's protocol lends itself to a plethora of *in vivo* and cell-free synthetic biology applications ranging from protein engineering to the optimization of natural or synthetic metabolic pathways[13,38–40]. More specifically, MOSAIC could enable the rapid prototyping of pathway or enzyme variants when coupled to high-throughput phenotypic screening or growth-coupled selection approaches to isolate well-performing variants. These isolated plasmid variants could in principle be rapidly subjected to further rounds of diversification using MOSAIC during subsequent design-build-test-learn cycles.

To synthesize large plasmid libraries, a high transformation efficiency is required. In contrast to cloning-based library synthesis methods, plasmid recombineering benefits from the transformation of preassembled plasmids. In other words, transformation efficiency is not limited by incomplete assemblies. However, the presence of full and partial wild-type plasmid variants alongside the mutant DNA variants in combinatorial MOSAIC libraries is currently unavoidable. As such, increasing the number of target sites or the variability at each site may hinder library coverage. The library coverage could be increased by performing multiple reactions in parallel or improving the transformation efficiency. Alternatively, the fraction of fully mutated library variants could be improved by employing iterative rounds of MOSAIC, co-selection using restriction enzymes, or CRISPR-Cas-based counterselection[20,22]. Another promising strategy is to restrain the library size by computational design. Excitingly, recent efforts to preselect library variants from a larger pool using machine learning prior to wet lab characterization showed promising results in generating small but smart libraries to accelerate the evolutionary optimization[41]. All in all, we anticipate that MOSAIC, combined with the continuing advances in computationally aided design of genetic and protein libraries, will enable the rapid exploration of biological solution spaces throughout many labs and research projects.

Material and methods

Reagents and equipment

Chemicals were purchased from Sigma-Aldrich, unless stated otherwise. *m*-toluic acid was dissolved in ethanol at a concentration of 1 M and stored at -20 °C. Plasmids were isolated from bacterial cells using PureYield Plasmid Miniprep System (Promega) or QIAprep Spin Miniprep Kit (Qiagen). Linear DNA was purified using the QIAquick PCR & Gel Cleanup Kit (Qiagen). DNA concentrations were measured using NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific), DS-11 FX spectrophotometer (DeNovix) or Qubit 4 Fluorometer (Invitrogen). Electroporation was performed with 1-mm gap Gene Pulser/MicroPulser electroporation cuvettes (Bio-Rad Laboratories). Different electroporators were used, which all performed robustly for the MOSAIC protocol: the Eppendorf Eporator electroporator (Eppendorf) (1.8 kV) and the ECM 630B electroporator (BTX) (1.8 kV, 200 Ω, 25 µF).

Strains, cultivation and plasmid construction

Bacterial strains for transformation experiments included *E. coli* K-12 MG1655 (Leibniz Institute DSMZ, Germany), *E. coli* K-12 DH5α and NEB 10-beta *E. coli* (NEB, C3020K). The bacteria were grown in Lysogeny Broth (LB) medium or on LB agar plates containing antibiotics (kanamycin, ampicillin, apramycin) at a concentration of 50 µg/mL unless indicated otherwise. The plasmids used in this study are listed in **Table 1**. pORTMAGE-Ec1 was a gift from the George Church lab (Addgene plasmid #138474; http://n2t.net/addgene:138474; RRID:Addgene_138474). pUC19 was acquired from New England Biolabs. pSEVAb plasmids were cloned according to the method reported earlier[30]. Plasmid G555 was constructed by subcloning of the construct containing genes *plsB*, *plsC*, *cdsA* and *pssA* (amplified by primers 1285 ChD and 1286 ChD from plasmid G363) into the backbone of plasmid G340 (amplified by primers 1287 ChD and 1288 ChD) via restriction enzyme digestion (NcoI/XhoI) and ligation. Primers 1285 ChD-1288 ChD are listed in **Table S1**. G340 was constructed as described elsewhere[42]. G363 was assembled using a stepwise Golden Gate ligation of six PCR fragments containing independent transcriptional cassettes. First, *plsB*, *plsC* (fragment 1) and *cdsA*, *pssA* (fragment 2) and *tp*, *dnap*, Phi29 origins (fragment 3) were ligated. Then, these three fragments and the pTU1 backbone (Addgene #72934) were ligated to form G363.

Table 1. List of plasmids used in this study.

Plasmid name	Addgene number	Origin of replication	Antibiotic resistance marker	Target region in plasmid for MOSAIC
pORTMAGE-Ec1	#138474	RSF1010	kanamycin	-
pUC19	#50005	pUC	ampicillin	*lacZ*
pSEVAb827	#217500	RK2	apramycin	*sfGFP*
pSEVAb837	#217501	pBBR1	apramycin	*sfGFP*
pSEVAb847	#217502	pRO1600/ColE1	apramycin	*sfGFP*
pSEVAb867	#217503	p15A	apramycin	*sfGFP*
pSEVAb887	#217504	pUC	apramycin	*sfGFP*
pSEVAb897	#217505	pBR322/ROP	apramycin	*sfGFP*
G555	#216483	pUC	ampicillin	RBSs of *plsB*, *plsC*, *cdsA* and *pssA*

Recombineering oligos and library design

Mutagenic ssDNA oligos of 89-91 nucleotides were designed to anneal with at least 30 nucleotides at both ends to the target-DNA. The ssDNA oligos are listed in **Table S1**. The oligos were modified with two phosphorothioate bonds at the 5′ end. The ssDNA oligos were synthesized and purified by desalting by Sigma-Aldrich (oligos BG31272 and BG31273) or synthesized and purified by HPLC by ELLA Biotech GmbH (Germany) (all other oligos). The oligos were diluted in Milli-Q water to a concentration of 100 µM and stored at -20 °C.

RBS variants were designed using the RBS Library Calculator in the "Optimize Expression Levels" mode (https://salislab.net/software/design_rbs_library_calculator) with the following input parameters: the host organism was *Escherichia coli*; target minimum and maximum translation initiation rates were 1 and 1,000,000, respectively; the genomic RBS sequence was the mRNA sequence from the 5′ end until the start codon[1].

Plasmid recombineering with the MOSAIC protocol

E. coli cells harboring pORTMAGE-Ec1 were streaked from glycerol stocks on a kanamycin-supplemented LB agar plate and grown overnight at 37 °C. The day before the MOSAIC experiment, an individual colony was picked and grown overnight in LB medium supplemented with kanamycin in a shaking incubator at 37 °C and 180-250 rpm. The following day, the overnight culture was diluted 1:100 in LB supplemented with kanamycin in a 50-mL falcon tube and incubated at 37 °C and 180-250 rpm. At an OD_{600} of 0.2-0.3, expression of the pORTMAGE-Ec1 machinery was induced by adding *m*-toluic acid to the culture to a final concentration of 1 mM. Following induction, the cells were incubated for an additional 45 minutes before being placed on ice for 1 hour. To make the cells electrocompetent, the culture was pelleted by centrifugation at 3200 rcf and 4 °C for 10 minutes. Next, the supernatant was

carefully decanted before the cells were resuspended in 1 mL of ice-cold Milli-Q water containing 10% glycerol (v/v) and transferred to a 1.5-mL Eppendorf tube. The cells were washed another two to three times. Following the last wash step, the cells were resuspended in 250 µL of ice-cold Milli-Q water per 10 mL of initial culture. Next, 40 µL of cell suspension, 1 ng of target plasmid and 1 µL of 100 µM ssDNA oligos were combined in a 1.5-mL Eppendorf tube. For multi-target MOSAIC reactions, the oligos of interest were premixed at equimolar concentrations and added to the cells to a final concentration of 2.5 µM per oligo or degenerate set of oligos. For the RBS library, degenerate oligos were mixed with one or two additional single oligos per target locus as additional library variants (**Table S1**). Next, 40 µL of the cell-DNA mixture were transferred to a 1-mm gap electroporation cuvette and electroporated. Immediately after electroporation, 960 µL of prewarmed LB were added to the cell suspension and the cells were allowed to recover for 1 hour at 37 °C and 180-250 rpm. Following recovery, single-target transformants were transferred to a 50-mL falcon tube, supplemented with 4 mL of LB containing the appropriate antibiotic and incubated overnight at 37 °C and 180-250 rpm. The next day, the plasmids were isolated from the cells. For multi-target MOSAIC transformations, the recovered cells were plated on large (15 cm diameter) selective agar plates and incubated at 37 °C overnight. The following day, the colonies were counted by hand, whereafter the colonies were scraped off the plate for plasmid isolation. All plasmids were eluted from the plasmid purification columns using Milli-Q water. The DNA purity and concentration were validated by spectrophotometry and fluorometry.

To quantify the number of DNA variants present in single colonies for multi-target MOSAIC with degenerate oligos, six single colonies were picked and grown overnight in ampicillin-supplemented LB (100 µg/mL ampicillin) for plasmid isolation and subsequent Nanopore sequencing.

Genomic recombineering control experiment

When recombineering was performed on the *E. coli* genome, electroporation of the cells was followed by 1 hour of incubation in 1 mL of LB and, subsequently, 2 hours of incubation in 6 mL of kanamycin-supplemented medium, whereafter the cells were plated on LB agar plates containing kanamycin, 100 µM isopropyl-β-D-thiogalactopyranoside (IPTG) and 100 µg/mL 5-bromo-4-chloro-3-indolyl-beta-D-galacto-pyranoside (X-gal) (Thermo Fisher Scientific). After incubation overnight at 37 °C, the fraction of white colonies relative to the total number of colonies was counted and used as a measure for the genomic recombineering efficiency.

Retransformation of MOSAIC plasmid mixtures

To separate mutated pUC19 from wild-type pUC19 and pORTMAGE-Ec1 after recombineering, we retransformed competent *E. coli* cells with the resulting plasmid mixtures and screened single colonies for the desired mutations. More specifically, we transformed chemically competent DH5α with 2 ng of the plasmid mixture by heat shock at 42 °C for 45 seconds. After 1 hour recovery in 1 mL of prewarmed LB, 50 µL of cells were plated on prewarmed ampicillin-

supplemented (50 µg/mL) agar plates and grown overnight at 37 °C. Single colonies were picked and grown in 3 mL of ampicillin-supplemented (50 µg/mL) LB overnight. The plasmids were isolated from the cells, and the isolation of pure and correctly mutated plasmid was verified by spectrophotometry, gel electrophoresis and Nanopore sequencing.

To separate the RBS plasmid library from pORTMAGE-Ec1, 40 µL of NEB 10-beta electrocompetent *E. coli* cells (C3020K) were transformed with 2.5 µL (155 ng) of plasmid mixture by electroporation, following the instructions from NEB (Electroporation Protocol C3020). All cells were plated on five large (15 cm diameter) ampicillin-supplemented (100 µg/mL) agar plates and grown overnight at 37 °C. The colonies were scraped off the plates for plasmid isolation. The plasmids were then linearized at a unique restriction site outside of the mutagenized region of interest using StuI (0.8 U/µL) in rCutSmart buffer to ensure that the obtained Nanopore sequencing reads spanned the full region of interest. The linearized DNA was purified by excision of the expected band from a 0.7% agarose gel and further purification using the QIAquick PCR & Gel Cleanup Kit (Qiagen). The linear DNA was diluted in Milli-Q water to a concentration of 99 ng/µL and sequenced by Nanopore sequencing.

Nanopore sequencing and analysis

Nanopore sequencing was performed with the mixture containing target plasmid and pORTMAGE-Ec1 unless indicated otherwise. Samples were prepared by diluting the DNA in Milli-Q water to a concentration of 30-40 ng/µL as quantified by Qubit. Nanopore sequencing was performed by Plasmidsaurus (Oregon, US). To extract the sequencing reads that map to the target plasmid, the reads were filtered based on size (the target plasmid size plus and minus 100 bp) and mapped to the wild-type DNA sequence of the target plasmid using the *Filter FASTQ reads by quality score and length*[43] and *Map with minimap2*[44] tools, respectively (accessed in the Galaxy web platform (https://usegalaxy.org))[45]. To minimize the numbers of insertions/deletions rather than mismatches during mapping, the minimap2 alignment parameters gap open penalty for deletions and insertions were increased from 4 (default) to 16 and from 24 (default) to 48, respectively, for the MOSAIC experiments incorporating 18 nt-insertions and deletions in pUC19 and diversifying the four RBSs in plasmid G555. For the 18-nt mismatch samples, these were increased to 32 and 72, respectively. For the 18-nt insertion samples, the reads were mapped to the designed modified DNA sequence instead of the wild-type sequence. If the reads from multiple Nanopore sequencing runs of the same sample were used for analysis, the FASTQ datasets were first merged using *Concatenate datasets tail-to-head* tool in the Galaxy web platform.

In-house developed R scripts were run in Rstudio (Version 1.1.456) to determine the editing efficiency (available at https://dx.doi.org/10.4121/4464ab86-9214-49b3-a808-10ca655385a6). The sequential steps in the analysis pipeline are illustrated in **Figure S1**. In short, the DNA sequences from the target loci were extracted from the mapped reads and filtered based on the per-base quality scores recorded in the FASTQ files; the target sequences that contained

at least one base with a score lower than 50 were excluded from the analysis. Then, the target sequences were identified as the wild type or as successfully mutated based on 100% similarity. The fraction of mutated sequences relative to the total number of target sequences was used to determine the editing efficiency. If the plasmid was modified in multiple loci, the number of mutated target loci was also counted per plasmid. To that end, an additional filtering step was applied to remove all reads that did not span the full sequence from the first to the last target region.

Statistics

The editing efficiencies described in the main text were averaged over at least three biological replicates. The mean and standard deviation are given where indicated.

Data availability

The data and code underlying this article are available in the 4TU.ResearchData repository at https://dx.doi.org/10.4121/4464ab86-9214-49b3-a808-10ca655385a6.

Supporting information

Additional data for Nanopore sequencing, bacterial plating, agarose gel electrophoresis, as well as a list of DNA oligo sequences used in this study.

Acknowledgements

We thank Ana Restrepo Sierra for her contribution to the conceptualization and design of the RBS library. We also thank Suzan Yilmaz for assistance in setting up the recombineering experiments. We are thankful to Akos Nyerges for providing advice on recombineering and the use of pORTMAGE. Lastly, we thank Haroun Bensaadi for performing preliminary recombineering experiments. MvdB, CD and NJC acknowledge financial support from The Netherlands Organization of Scientific Research (NWO/OCW) Gravitation program Building A Synthetic Cell (BaSyC) (024.003.019). TYA and NJC acknowledge funding from an NWO-XL grant (OCENW.XL21.XL21.007). CD acknowledges funding from ANR (ANR-22-CPJ2-0091-01). NJC also acknowledges support from an NWO-Veni Fellowship (VI.Veni.192.156).

Author contributions

MvdB and TYA contributed equally to this study. Conceptualization and Methodology: MdvB, TYA, CD and NJC. Investigation and formal analysis: MvdB and TYA. Software and Data Curation (analysis pipeline development): MvdB. Visualization and Writing – Original draft: MvdB and TYA. Writing – Review & Editing: MdvB, TYA, CD and NJC. Supervision & Funding acquisition: CD and NJC.

References

1. 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).
2. LaFleur, T. L., Hossain, A. & Salis, H. M. Automated model-predictive design of synthetic promoters to control transcriptional profiles in bacteria. *Nat. Commun.* **13**, 5159 (2022).
3. Leman, J. K. *et al.* Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat. Methods* **17**, 665–680 (2020).
4. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
5. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
6. Yilmaz, S., Nyerges, A., van der Oost, J., Church, G. M. & Claassens, N. J. Towards next-generation cell factories by rational genome-scale engineering. *Nat. Catal.* **5**, 751–765 (2022).
7. Simon, A. J., d'Oelsnitz, S. & Ellington, A. D. Synthetic evolution. *Nat. Biotechnol.* **37**, 730–743 (2019).
8. Currin, A. *et al.* The evolving art of creating genetic diversity: From directed evolution to synthetic biology. *Biotechnol. Adv.* **50**, 107762 (2021).
9. Alejaldre, L., Pelletier, J. N. & Quaglia, D. Methods for enzyme library creation: Which one will you choose? *BioEssays* **43**, 2100052 (2021).
10. Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. *PLOS ONE* **3**, e3647 (2008).
11. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
12. Kok, S. de *et al.* Rapid and reliable DNA assembly via ligase cycling reaction. *ACS Synth. Biol.* **3**, 97–106 (2014).
13. Naseri, G. & Koffas, M. A. G. Application of combinatorial optimization strategies in synthetic biology. *Nat. Commun.* **11**, 2446 (2020).
14. Wannier, T. M. *et al.* Recombineering and MAGE. *Nat. Rev. Methods Primers* **1**, 1–24 (2021).
15. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *PNAS* **97**, 6640–6645 (2000).
16. Wang, X. *et al.* Improved dsDNA recombineering enables versatile multiplex genome engineering of kilobase-scale sequences in diverse bacteria. *Nucleic Acids Res.* **50**, e15–e15 (2022).
17. Wang, H. H. *et al.* Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460**, 894–898 (2009).
18. Swaminathan, S. *et al.* Rapid engineering of bacterial artificial chromosomes using oligonucleotides. *Genesis* **29**, 14–21 (2001).
19. Thomason, L. C., Costantino, N., Shaw, D. V. & Court, D. L. Multicopy plasmid modification with phage λ Red recombineering. *Plasmid* **58**, 148–158 (2007).
20. Li, Y. *et al.* Multiplex iterative plasmid engineering for combinatorial optimization of metabolic pathways and diversification of protein coding sequences. *ACS Synth. Biol.* **2**, 651–661 (2013).
21. Higgins, S. A., Ouonkap, S. V. Y. & Savage, D. F. Rapid and programmable protein mutagenesis using plasmid recombineering. *ACS Synth. Biol.* **6**, 1825–1833 (2017).
22. Geng, Y. *et al.* A highly efficient in vivo plasmid editing tool based on CRISPR-Cas12a and phage λ Red recombineering. *J. Genet. Genomics* **46**, 455–458 (2019).
23. Zhang, J. J. & Moore, B. S. Site-directed mutagenesis of large biosynthetic gene clusters via oligonucleotide recombineering and CRISPR/Cas9 targeting. *ACS Synth. Biol.* **9**, 1917–1922 (2020).
24. Hueso-Gil, A., Nyerges, Á., Pál, C., Calles, B. & de Lorenzo, V. Multiple-site diversification of regulatory sequences enables interspecies operability of genetic devices. *ACS Synth. Biol.* **9**, 104–114 (2020).
25. Wannier, T. M. *et al.* Improved bacterial recombineering by parallelized protein discovery. *PNAS* **117**, 13689–13698 (2020).
26. Tomizawa, J.-I., Sakakibara, Y. & Kakefuda, T. Replication of colicin E1 plasmid DNA in cell extracts. Origin and direction of replication. *PNAS* **71**, 2260–2264 (1974).

27. Selzer, G., Som, T., Itoh, T. & Tomizawa, J. The origin of replication of plasmid p15A and comparative studies on the nucleotide sequences around the origin of related plasmids. *Cell* **32**, 119–129 (1983).
28. Lilly,J. & Camps,M. Mechanisms of theta plasmid replication. *Microbiol. Spectr.* **3**, PLAS-0029-2014 (2015).
29. 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).
30. Damalas, S. G., Batianis, C., Martin-Pascual, M., de Lorenzo, V. & Martins dos Santos, V. A. P. SEVA 3.1: enabling interoperability of DNA assembly among the SEVA, BioBricks and Type IIS restriction enzyme standards. *Microb. Biotechnol.* **13**, 1793–1806 (2020).
31. Jahn, M., Vorpahl, C., Hübschmann, T., Harms, H. & Müller, S. Copy number variability of expression plasmids determined by cell sorting and Droplet Digital PCR. *Microb. Cell Fact.* **15**, 211 (2016).
32. Silva-Rocha, R. *et al.* The Standard European Vector Architecture (SEVA): a coherent platform for the analysis and deployment of complex prokaryotic phenotypes. *Nucleic Acids Res.* **41**, D666-675 (2013).
33. Kuzminov, A., Schabtach, E. & Stahl, F. W. Study of plasmid replication in Escherichia coli with a combination of 2D gel electrophoresis and electron microscopy. *J. Mol. Biol.* **268**, 1–7 (1997).
34. Doran, K. S., Konieczny, I. & Helinski, D. R. Replication origin of the broad host range plasmid RK2: positioning of various motifs is critical for initiation of replication. *J. Biol. Chem.* **273**, 8447–8453 (1998).
35. Meyer, R. J. & Helinski, D. R. Unidirectional replication of the P-group plasmid RK2. *Biochim. Biophys. Acta, Nucleic Acids Protein Synth.* **478**, 109–113 (1977).
36. Stalker, D. M., Thomas, C. M. & Helinski, D. R. Nucleotide sequence of the region of the origin of replication of the broad host range plasmid RK2. *Mol. Gen. Genet.* **181**, 8–12 (1981).
37. Blanken, D., Foschepoth, D., Serrão, A. C. & Danelon, C. Genetically controlled membrane synthesis in liposomes. *Nat. Commun.* **11**, 4317 (2020).
38. Gargiulo, S. & Soumillion, P. Directed evolution for enzyme development in biocatalysis. *Curr. Opin. Chem. Biol.* **61**, 107–113 (2021).
39. Li, Z., Deng, Y. & Yang, G.-Y. Growth-coupled high throughput selection for directed enzyme evolution. *Biotechnol. Adv.* **68**, 108238 (2023).
40. Orsi, E., Claassens, N. J., Nikel, P. I. & Lindner, S. N. Growth-coupled selection of synthetic modules to accelerate cell factory development. *Nat. Commun.* **12**, 5295 (2021).
41. Marchal, D. G. *et al.* Machine learning-supported enzyme engineering toward improved CO2-fixation of glycolyl-CoA carboxylase. *ACS Synth. Biol.* **12**, 3521–3530 (2023).
42. Abil, Z. *et al.* Darwinian evolution of self-replicating DNA in a synthetic protocell. *Nat. Commun.* **15**, 9091 (2024).
43. Blankenberg, D. *et al.* Manipulation of FASTQ data with Galaxy. *Bioinformatics* **26**, 1783–1785 (2010).
44. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
45. The Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res.* **50**, W345–W351 (2022).

Supporting Information

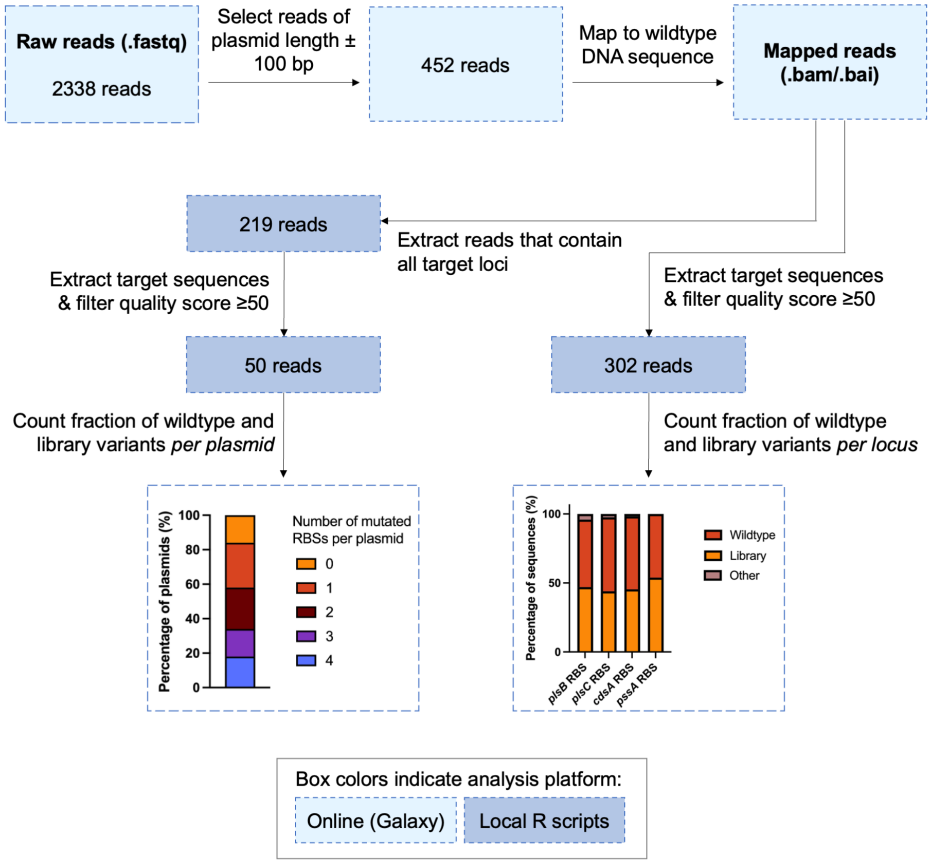


Figure S1. Nanopore sequencing analysis pipeline used in this study. As an example, the graphs and numbers of reads are given for a single MOSAIC sample where four RBSs were diversified in four different loci on plasmid G555.

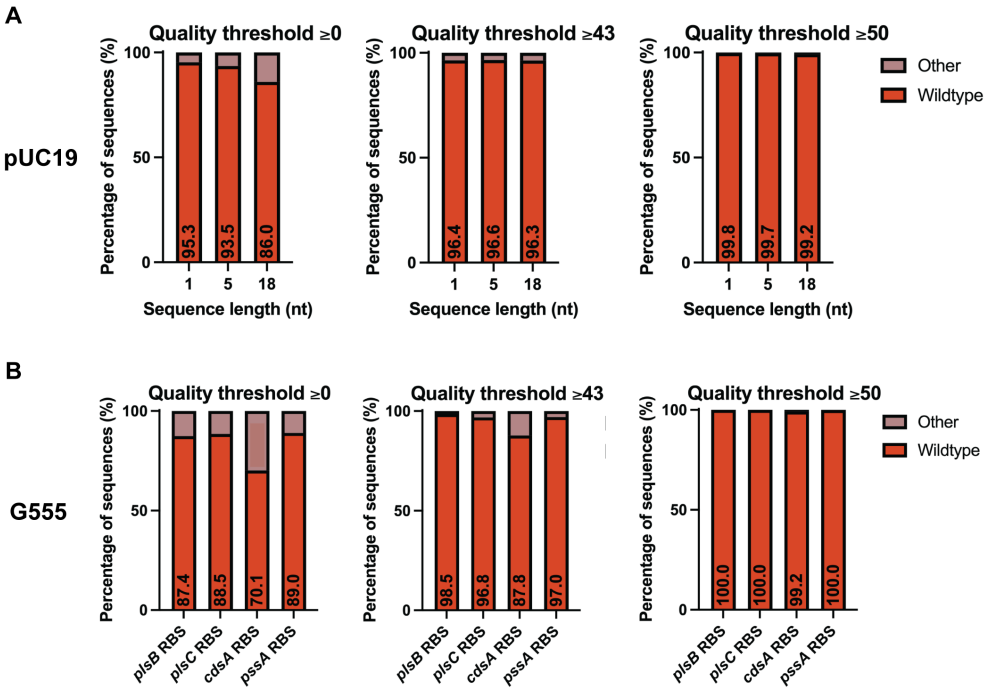


Figure S2. Analysis of Nanopore sequencing reads for wild-type pUC19 and wild-type G555 using different quality thresholds. A) Different quality thresholds applied to specific loci in wild-type pUC19 sequencing reads (1, 5 or 18 nt long) that correspond to the target loci in the experiments of Figure 1B and Figure 1D. **B)** Different quality thresholds applied to the specific loci in wild-type plasmid G555 sequencing reads (20 nt long) that correspond to the target loci in the experiments of Figure 2 and Figure S5. While for short target regions of 1-5 nucleotides the fraction of sequencing errors was relatively small without quality control (4-6%), for larger target regions (18-20 nt) the fraction of sequencing errors became large (11-30%). The value 50 was applied to the experiments in this study as the quality score threshold to minimize sequencing errors based on the sequencing data of the wild-type plasmids to <1%, while maintaining as many sequences for analysis as possible.

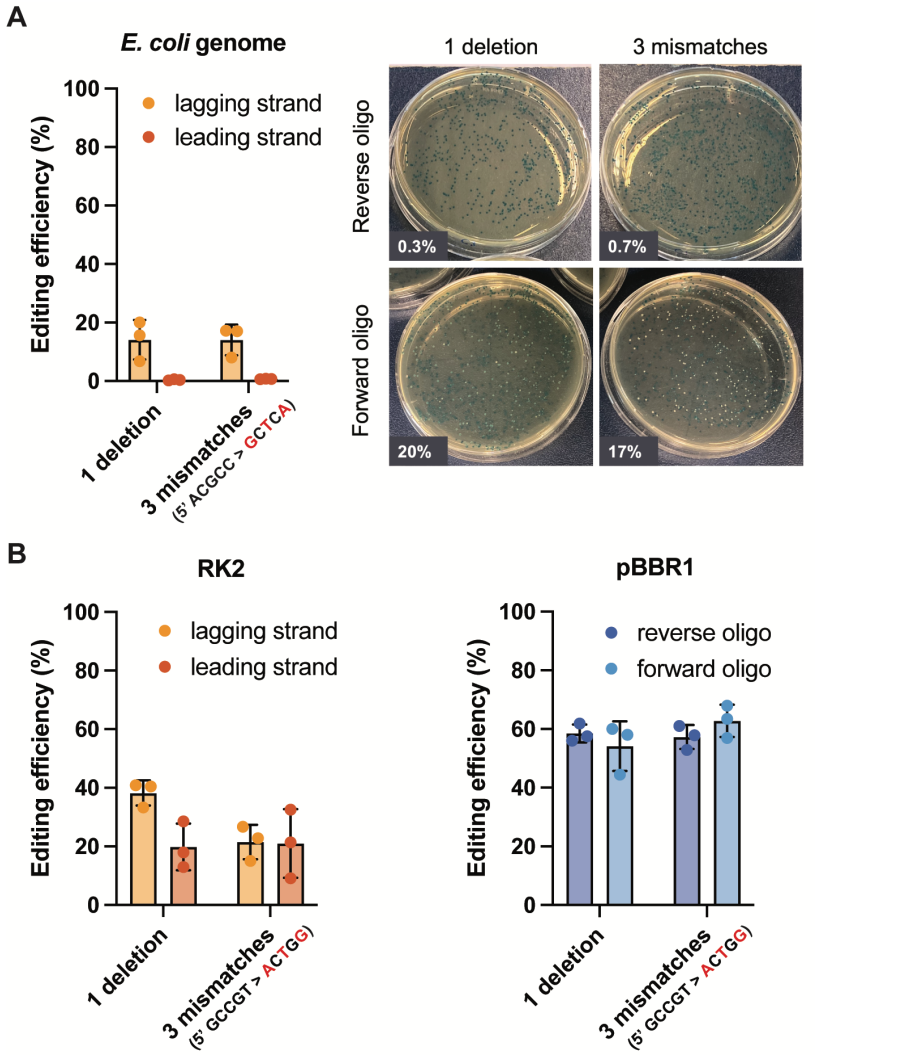


Figure S3. Editing efficiency targeting the *E. coli* genome or pSEVA plasmids using lagging or leading strand oligos. A) DNA editing efficiency for the incorporation of one deletion or three nucleotide mismatches in the *lacZ* gene in the genome of *E. coli* MG1655 harboring pORTMAGE-Ec1. The editing efficiencies were quantified by blue-white screening. The deletion in the *lacZ* gene caused a frameshift resulting in white colonies (deficient β-galactosidase), and the three substitutions led to a premature stop codon, resulting in light-blue colonies, while the fully functional variants appeared as dark blue colonies. **B)** DNA editing efficiency for the incorporation of one deletion or three nucleotide mismatches in the *sfGFP* gene on RK2-derived plasmid pSEVAb827 (left) or pBBR1-derived plasmid pSEVAb837 (right). As it is unknown which of the strands are the leading and lagging strands for pBBR1, we indicate the reverse complementary oligos with "forward" and "reverse" oligo, which anneal to the noncoding and coding strand of the *sfGFP* gene, respectively. The oligos were co-electroporated with the target plasmid into the *E. coli* MG1655 cells harboring pORTMAGE-Ec1.

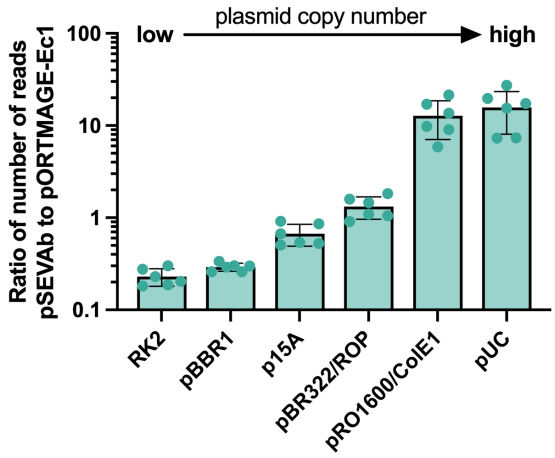


Figure S4. Ratio of the number of Nanopore sequencing reads for the pSEVAb plasmids to those for pORTMAGE-Ec1. *E. coli* MG1655 cells harboring pORTMAGE-Ec1 were electroporated with one of the pSEVAb plasmids and a mutagenic oligo. After plasmid isolation and Nanopore sequencing, the number of sequencing reads corresponding to the pSEVAb or pORTMAGE-Ec1 plasmids were determined by counting the reads with a size corresponding to that of the plasmids (± 100 bp). Then, the ratio of pSEVAb reads to pORTMAGE-Ec1 reads was calculated. pSEVAb plasmids with higher copy number origins of replication exhibited correspondingly elevated pSEVAb-to-pORTMAGE-Ec1 read ratios, thus affirming the consistency of the copy numbers reported in literature with the experimental (relative) copy numbers in our study[31,32].

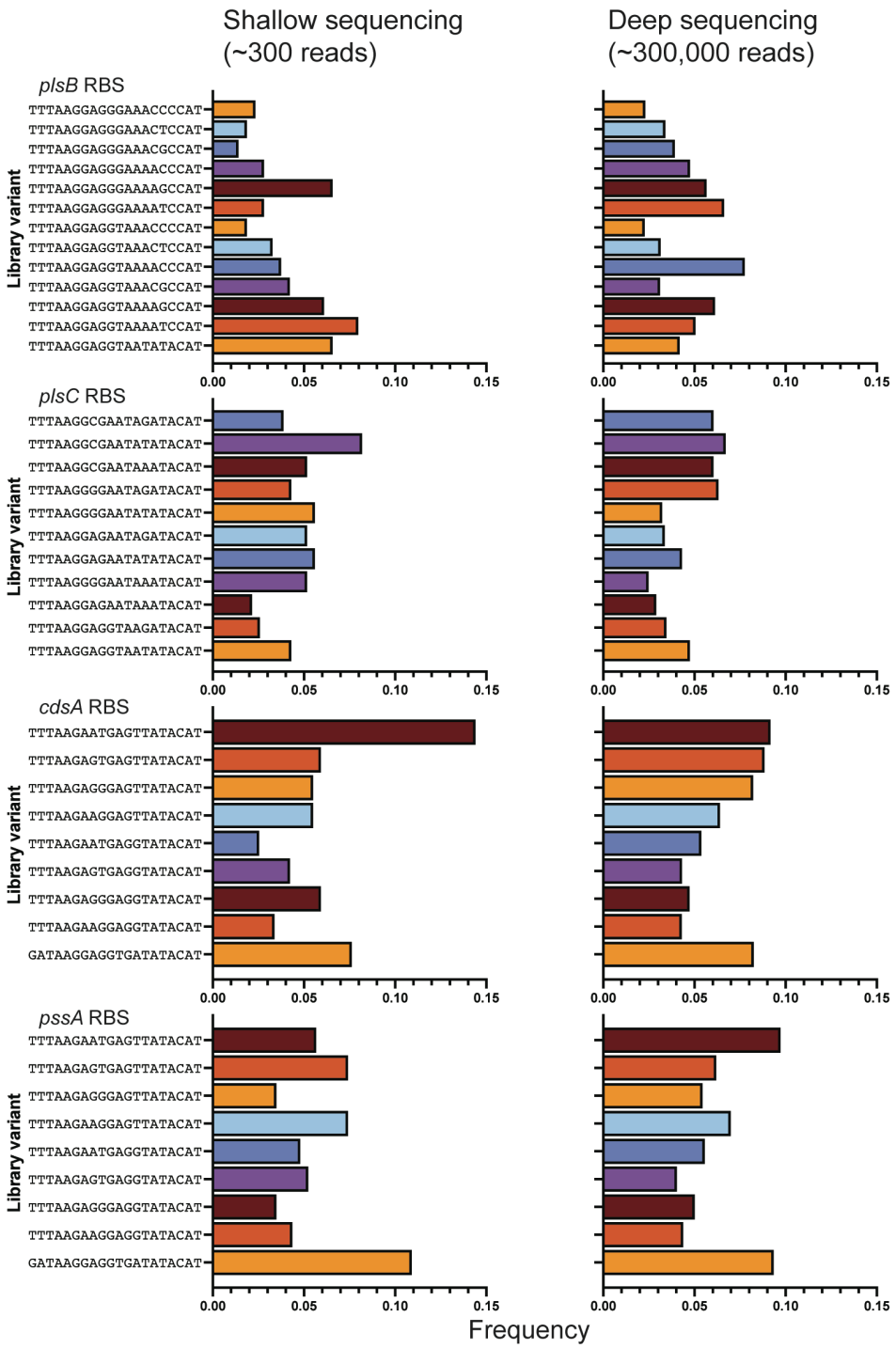


Figure S5. Frequency of each of the RBS variants per target locus. The graphs display the data for one of the four replicates from Figure 2C. The variant frequencies were quantified from shallow Nanopore sequencing (n=4, 150-300 reads) or deep sequencing (n=1, 279,454 reads).

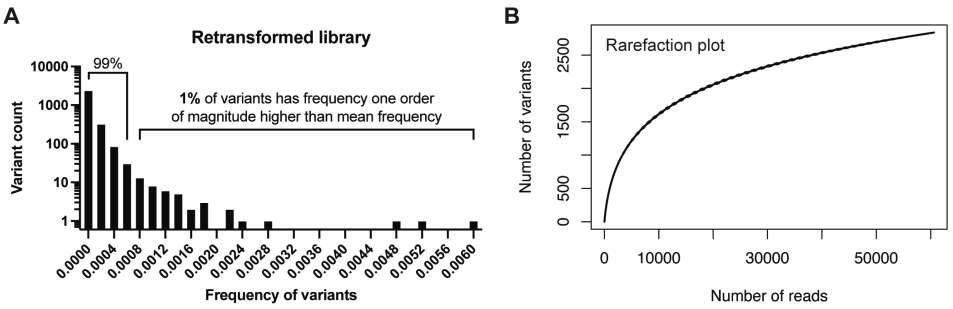


Figure S6. Sequencing depth for the retransformed RBS library. A) Frequency distribution of designed library variants in the deep sequencing dataset (279,454 reads). Sequencing reads containing at least one unmodified RBS were not considered. The average frequency was 0.0001 ± 0.0002 (mean ± standard deviation). **B)** Rarefaction plot of the sequenced DNA variants with all four RBSs mutated.

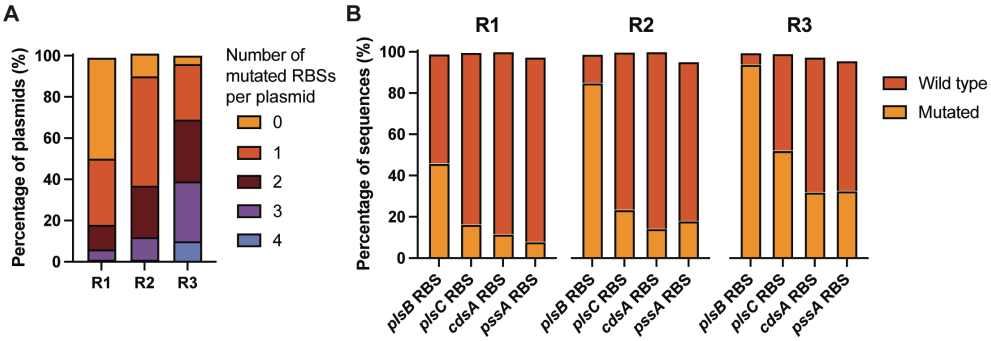


Figure S7. Three successive rounds of MOSAIC increase the RBS library diversity of plasmid G555. Strain used: DH5α. Oligos used: *1544 ChD*, *1545 ChD* and *1549 ChD* (Table S1). **A)** Number of mutated RBSs per plasmid. **B)** Percentage of wild-type and library sequences per target locus. The stacked bars do not match up to exactly 100%, because a small fraction of the sequences was neither the wild-type sequence nor a designed library variant. These unintended mutations were likely sequencing errors that were not excluded by our analysis pipeline (Figure S2B), mutations incorporated by oligos intended to bind at other target sites, and/or spontaneous mutations retained due to the suppression of the mismatch repair system during recombineering.

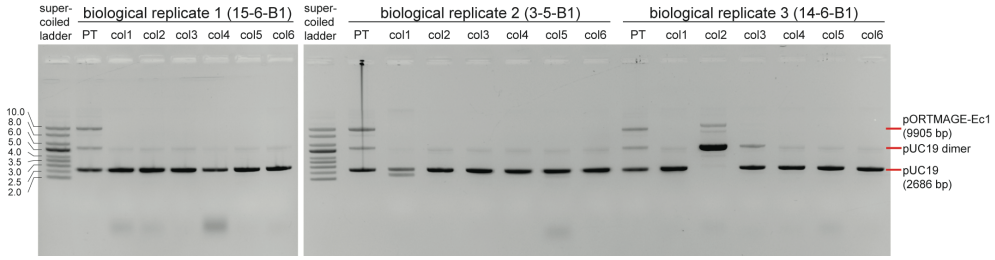


Figure S8. Isolation of mutated pUC19 by retransformation. pUC19 was mutated in the *lacZ* gene as shown in Figure 1B (3 mismatches, co-electroporation, MG1655). The resulting plasmid mixtures (n=3 biological replicates), containing mutated pUC19, wild-type pUC19 and pORTMAGE-Ec1, were used for chemical transformation of *E. coli* DH5α cells. Per sample, 6 colonies were picked from selective agar plates, grown in liquid LB medium, and plasmids were isolated from each cell culture. 100 ng of plasmids were run on a 0.7% agarose gel. Plasmid sizes were determined using the supercoiled DNA ladder (NEB) with plasmids ranging from 2.0-10.0 kb. PT=pre-transformation plasmid mixture containing pUC19 and pORTMAGE-Ec1. Col[x]=isolated plasmids from colony [x] after transformation. Nanopore sequencing results of some of the isolated plasmids are presented in Figure 3B. A small fraction of pUC19 dimer was formed, as indicated in the figure. The formation of multimers during plasmid recombineering has been reported before[19,20]. However, in our experiment, it remained unclear whether the dimers were formed due to the expressed recombineering machinery, or due to an alternative mechanism in *E. coli* MG1655 and/or in DH5α. Due to the small fraction of dimers formed as compared to the monomer, this did not cause a problem during the isolation of mutated pUC19 monomer.

Table S1. DNA oligos used in this study. Phosphorothioate bonds are indicated by an asterisk (*). Degenerate oligos (DEG) targeting G555 were mixed with single oligos (SINGLE) as additional library variants before mixing of oligos, plasmid and cell suspension for electroporation.

Oligo name	Sequence	Target DNA	Description of use in this study
1285 ChD / 1286 ChD	AACCATGGCTAGAACGTCTCAATCTCT ATTAATAC / TTCTCGAGTTACAGGATGCGGCTAATT AATCGGTC	Plasmid G363	Forward and reverse PCR primers for amplification of fragment containing four genes of phospholipid synthesis pathway for the assembly of plasmid G555
1287 ChD / 1288 ChD	AACTCGAGTAGCATAACCCCTTGGGG CCTCTAAAC / AACCATGGACATGCGACACAGACGAA GCGCTAAAAC	Plasmid G340	Forward and reverse PCR primers for amplification of backbone for the assembly of plasmid G555
1495 ChD	C*T*GGCGAAAGGGGGATGTGCTGCA AGGCGATTAAGTTGGGTAACGCAGGG TTTTCCCAGTCACGACGTTGTAAAACG ACGGCCAGTGAAT	pUC19/ *E. coli* genome	Single deletion in *lacZ* gene (introducing frame shift), lagging strand oligo for pUC19, leading strand oligo for *E. coli* genome
1496 ChD	C*T*GGCGAAAGGGGGATGTGCTGCA AGGCGATTAAGTTGGGTAGCTCAAGG GTTTTCCCAGTCACGACGTTGTAAAAC GACGGCCAGTGAAT	pUC19/ *E. coli* genome	3-nt mismatch in *lacZ* gene (introducing stop codon), lagging strand oligo for pUC19, leading strand oligo for *E. coli* genome
1489 ChD	A*T*TCACTGGCCGTCGTTTTACAACGT CGTGACTGGGAAAACCCTGCGTTACC CAACTTAATCGCCTTGCAGCACATCCC CCTTTCGCCAG	pUC19/ *E. coli* genome	Single deletion in *lacZ* gene (introducing frame shift), leading strand oligo for pUC19, lagging strand oligo for *E. coli* genome
1490 ChD	A*T*TCACTGGCCGTCGTTTTACAACGT CGTGACTGGGAAAACCCTTGAGCTAC CCAACTTAATCGCCTTGCAGCACATCC CCCTTTCGCCAG	pUC19/ *E. coli* genome	3-nt mismatch in *lacZ* gene (introducing stop codon), leading strand oligo for pUC19, lagging strand oligo for *E. coli* genome

Continuation of Table S1

1505 ChD	C*T*GGCGAAAGGGGGATGTGCTGCA AGGCGATTAAGTGTCCGCTGAAGGAA TCGATCCCAGTCACGACGTTGTAAAA CGACGGCCAGTGAAT	pUC19	18-nt mismatch in *lacZ* gene, lagging strand oligo for pUC19
1506 ChD	T*T*ACGCCAGCTGGCGAAAGGGGGA TGTGCTGCAAGGCGATTAAGTTCCCA GTCACGACGTTGTAAAACGACGGCCA GTGAATTCGAGCTCG	pUC19	18-nt deletion in *lacZ* gene, lagging strand oligo for pUC19
1507 ChD	A*G*GGGGATGTGCTGCAAGGCGATT AAGTTGGGTAACAAAGTATCAGGACG GTATGCCAGGGTTTTCCCAGTCACGAC GTTGTAAAACGACG	pUC19	18-nt insertion in *lacZ* gene, lagging strand oligo for pUC19
BG312 70	G*T*CATAAGTTTAGCGTTCGCGGTGA AGGTGAGGGCGACGCGACCACGGCA AACTGACCCTGAAGTTCATCTGCACCA CCGGTAAACTGC	pSEVAb plasmids	Single deletion in *sfGFP* gene, leading strand oligo
BG312 71	G*T*CATAAGTTTAGCGTTCGCGGTGA AGGTGAGGGCGACGCGACCACCAGT AAACTGACCCTGAAGTTCATCTGCACC ACCGGTAAACTGC	pSEVAb plasmids	3-nt mismatch in *sfGFP* gene, leading strand oligo
BG312 72	G*C*AGTTTACCGGTGGTGCAGATGAA CTTCAGGGTCAGTTTGCCGTGGTCGC GTCGCCCTCACCTTCACCGCGAACGC TAAACTTATGAC	pSEVAb plasmids	Single deletion in *sfGFP* gene, lagging strand oligo
BG312 73	G*C*AGTTTACCGGTGGTGCAGATGAA CTTCAGGGTCAGTTTACTGGTGGTCGC GTCGCCCTCACCTTCACCGCGAACGC TAAACTTATGAC	pSEVAb plasmids	3-nt mismatch in *sfGFP* gene, lagging strand oligo
PlsB DEG	C*A*ACGGTTTCCCTCTAGAAATAATTT TGTTTAACTTTAAGGAGGKAAAMBCC ATATGACTTTCTGCTATCCTTGCCGCG CATTTGCATTA	Plasmid G555	12-variant library targeting RBS of *plsB* gene, lagging strand oligo
PlsC DEG	C*A*ACGGTTTCCCTCTAGAAATAATTT TGTTTAACTTTAAGGVGAATADATACA TATGCTATATATCTTTCGTCTTATTATT ACCGTGATT	Plasmid G555	9-variant library targeting RBS of *plsC* gene, lagging strand oligo
PssA_CdsA DEG	A*T*AACAATTCCCCTCTAGAAATAATT TTGTTTAACTTTAAGARKGAGKTATAC ATATGGCTAGCATGACTGGTGGACAG CAAATGGGTCG	Plasmid G555	8-variant library targeting RBSs of *cdsA* and *pssA* genes, lagging strand oligo
PlsB SINGLE	C*A*ACGGTTTCCCTCTAGAAATAATTT TGTTTAACTTTAAGGAGGTAATATACA TATGACTTTCTGCTATCCTTGCCGCGC ATTTGCATTA	Plasmid G555	Single library variant targeting RBS of *plsB* gene, mixed with degenerate oligo set before recombineering, lagging strand oligo
PlsC_1 SINGLE	C*A*ACGGTTTCCCTCTAGAAATAATTT TGTTTAACTTTAAGGAGGTAATATACA TATGCTATATATCTTTCGTCTTATTATT ACCGTGATT	Plasmid G555	Single library variant targeting RBS of *plsC* gene, mixed with degenerate oligo set before recombineering, lagging strand oligo
PlsC_2 SINGLE	C*A*ACGGTTTCCCTCTAGAAATAATTT TGTTTAACTTTAAGGAGGTAAGATACA TATGCTATATATCTTTCGTCTTATTATT ACCGTGATT	Plasmid G555	Single library variant targeting RBS of *plsC* gene, mixed with degenerate oligo set before recombineering, lagging strand oligo
PssA_CdsA SINGLE	A*G*CGGATAACAATTCCCCTCTAGAA ATAATTTTGTTTAACGATAAGGAGGTG ATATACATATGGCTAGCATGACTGGTG GACAGCAAATG	Plasmid G555	Single library variant targeting RBSs of *cdsA* and *pssA* genes, mixed with degenerate oligo set before recombineering, lagging strand oligo

Continuation of Table S1

1544 ChD	C*A*ACGGTTTCCCTCTAGAAATAATTT TGTTTAACTTTAAGGAGGGAAAACCC ATATGACTTTCTGCTATCCTTGCCGCG CATTTGCATTA	Plasmid G555	Single library variant targeting RBS of *plsB* gene, lagging strand oligo *Not used to generate RBS library, used only for Figure S7*
1545 ChD	C*A*ACGGTTTCCCTCTAGAAATAATTT TGTTTAACTTTAAGGAGAATATATACA TATGCTATATATCTTTCGTCTTATTATT ACCGTGATT	Plasmid G555	Single library variant targeting RBS of *plsC* gene, lagging strand oligo *Not used to generate RBS library, used only for Figure S7*
1549 ChD	A*T*AACAATTCCCCTCTAGAAATAATT TTGTTTAACTTTAAGAATGAGTTATAC ATATGGCTAGCATGACTGGTGGACAG CAAATGGGTCG	Plasmid G555	Single library variant RBSs of *cdsA* and *pssA* genes, lagging strand oligo *Not used to generate RBS library, used only for Figure S7*

Table S2. Number of sequenced DNA variants per colony. From the Nanopore reads that span all four target loci, the different DNA variants were counted. The average number of DNA variants per colony was 6 ± 4 (n=6 colonies).

Colony #	**DNA variants (shown are the four target sequences: PlsB RBS … PlsC RBS … CdsA RBS … PssA RBS)**	**Count**
1 (total 60 reads, 5 different DNA variants)	TTTAAGGAGGTAAACCCCAT…TTTAAGAAGGAGATATACAT… TTTAAGAAGGAGATATACAT…TTTAAGAGTGAGGTATACAT	43
	TTTAAGGAGGAGATATACAT…TTTAAGGAGGAGATATACAT… TTTAAGAAGGAGTTATACAT…TTTAAGAAGGAGATATACAT	8
	TTTAAGAAGGAGATATACAT…TTTAAGAAGGAGATATACAT… TTTAAGAAGGAGATATACAT…TTTAAGAAGGAGATATACAT	7
	TTTAAGGAGGTAAAATCCAT…TTTAAGGAGGAGATATACAT… TTTAAGAAGGAGTTATACAT…TTTAAGAAGGAGATATACAT	1
	TTTAAGGAGGTAAACCCCAT…TTTAAGAAGGAGATATACAT… TTTAAGAAGGAGATATACAT…TTTAAGAAGGAGATATACAT	1
2 (total 22 reads, 11 different DNA variants)	TTTAAGAAGGAGATATACAT…TTTAAGGGGAATAGATACAT… TTTAAGAAGGAGTTATACAT…TTTAAGAAGGAGATATACAT	7
	TTTAAGAAGGAGATATACAT…TTTAAGGGGAATAGATACAT… TTTAAGAAGGAGATATACAT…TTTAAGAAGGAGATATACAT	1
	TTTAAGAAGGAGATATACAT…TTTAAGAAGGAGATATACAT… TTTAAGAAGGAGTTATACAT…GATAAGGAGGTGATATACAT	1
	TTTAAGAAGGAGATATACAT…TTTAAGAAGGAGATATACAT… TTTAAGAAGGAGATATACAT…GATAAGGAGGTGATATACAT	1
	TTTAAGAAGGAGATATACAT…TTTAAGGGGAATAGATACAT… TTTAAGAAGGAGTTATACAT…GATAAGGAGGTGATATACAT	1
	TTTAAGAAGGAGATATACAT…TTTAAGAAGGAGATATACAT… TTTAAGAAGGAGATATACAT…TTTAAGAAGGAGATATACAT	4
	TTTAAGAAGGAGATATACAT…TTTAAGAAGGAGATATACAT… TTTAAGAAGGAGTTATACAT…TTTAAGAAGGAGATATACAT	1
	TTTAAGAAGGAGATATACAT…TTTAAGAAGGAGATATACAT… TTTAAGAAGGAGGTATACAT…TTTAAGAAGGAGATATACAT	2
	TTTAAGAAGGAGATATACAT…TTTAAGAAGGAGATATACAT… TTTAAGAAGGAGGTATACAT…GATAAGGAGGTGATATACAT	1
	TTTAAGAAGGAGATATACAT…TTTAAGGAGAATATATACAT… TTTAAGAAGGAGTTATACAT…GATAAGGAGGTGATATACAT	2
	TTTAAGAAGGAGATATACAT…TTTAAGGAGAATATATACAT… TTTAAGAAGGAGGTATACAT…TTTAAGAAGGAGATATACAT	1

Continuation of Table S2

3 (total 30 reads, 9 different DNA variants)	TTTAAGAAGGAGATATACAT...TTTAAGGCGAATATATACAT... TTTAAGAGGGAGGTATACAT...TTTAAGAAGGAGGTATACAT	2
	TTTAAGGAGGTAAACGCCAT...TTTAAGAAGGAGATATACAT... TTTAAGAAGGAGATATACAT...TTTAAGAAGGAGATATACAT	1
	TTTAAGGAGGTAAAATCCAT...TTTAAGAAGGAGATATACAT... TTTAAGAAGGAGATATACAT...TTTAAGAAGGAGATATACAT	4
	TTTAAGAAGGAGATATACAT...TTTAAGGAGGTAAGATACAT... TTTAAGAAGGAGGTATACAT...TTTAAGAAGGAGATATACAT	3
	TTTAAGAAGGAGATATACAT...TTTAAGAAGGAGATATACAT... TTTAAGAGGGAGGTATACAT...TTTAAGAAGGAGATATACAT	11
	TTTAAGAAGGAGATATACAT...TTTAAGAAGGAGATATACAT... TTTAAGAAGGAGATATACAT...TTTAAGAAGGAGATATACAT	3
	TTTAAGAAGGAGATATACAT...TTTAAGAAGGAGATATACAT... TTTAAGAGGGAGGTATACAT...GATAAGGAGGTGATATACAT	1
	TTTAAGGAGGTAAAACCCAT...TTTAAGAAGGAGATATACAT... GATAAGGAGGTGATATACAT...TTTAAGAAGGAGATATACAT	2
	TTTAAGAAGGAGATATACAT...TTTAAGGAGAATATATACAT... TTTAAGAAGGAGATATACAT...TTTAAGAAGGAGATATACAT	3
4 (total 29 reads, 4 different DNA variants)	TTTAAGAAGGAGATATACAT...TTTAAGAAGGAGATATACAT... TTTAAGAGTGAGGTATACAT...TTTAAGAAGGAGATATACAT	10
	TTTAAGGAGGGAAACCCCAT...TTTAAGAAGGAGATATACAT... TTTAAGAAGGAGATATACAT...TTTAAGAAGGAGATATACAT	5
	TTTAAGGAGGGAAAATCCAT...TTTAAGAAGGAGATATACAT... TTTAAGAATGAGTTATACAT...TTTAAGAGTGAGTTATACAT	13
	TTTAAGAAGGAGATATACAT...TTTAAGAAGGAGATATACAT... TTTAAGATGAAGGTATACAT...TTTAAGAAGGAGATATACAT	1
5 (total 27 reads, 3 different DNA variants)	TTTAAGAAGGAGATATACAT...TTTAAGAAGGAGATATACAT... TTTAAGAAGGAGATATACAT...TTTAAGAAGGAGATATACAT	23
	TTTAAGAAGGAGATATACAT...TTTAAGAAGGAGATATACAT... TTTAAGAGGGAGGTATACAT...TTTAAGAAGGAGATATACAT	1
	TTTAAGAAGGAGATATACAT...TTTAAGAAGGAGATATACAT... TTTAAGAATGAGGTATACAT...TTTAAGAAGGAGATATACAT	3
6 (total 29 reads, 2 different DNA variants)	TTTAAGAAGGAGATATACAT...TTTAAGAAGGAGATATACAT... TTTAAGAAGGAGATATACAT...TTTAAGAAGGAGATATACAT	28
	TTTAAGAAGGAGATATACAT...TTTAAGGAGGTAAGATACAT... TTTAAGAAGGAGATATACAT...TTTAAGAAGGAGATATACAT	1

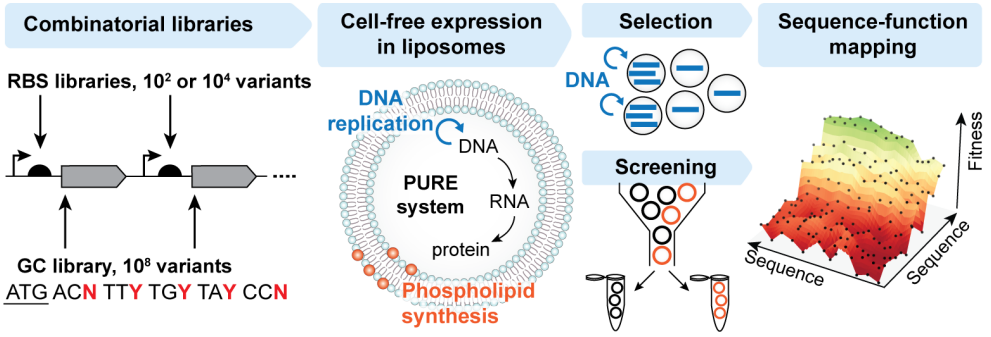
3

Combinatorial optimization of protein systems in synthetic cells

In vitro reconstitution of protein systems – e.g., metabolic pathways, genetic circuits or biosensors – often requires optimization to enhance their activity. Combinatorial DNA libraries that simultaneously target multiple genes allow for a holistic optimization strategy by studying the interplay between the systems' components, which may reveal DNA variants that would be hidden when testing each element in isolation. Here, we screen large populations of synthetic vesicles that express combinatorial DNA variants of a DNA self-replicator or a phospholipid synthesis pathway. We simultaneously vary the strengths of multiple RBSs or synonymously mutate the first codons of multiple genes to explore the effects of the protein translation rates directly on the functionality of the two core synthetic cell modules. We isolated high performers through DNA self-selection or functional screening by fluorescence-activated cell sorting. Long-read sequencing of the fittest variants informed on the optimal RBS strengths and base substitutions in the first codons and indicated which genes were most impactful in regulating the functionality of the protein systems. Single-mutation data were used to predict the fitness of combinatorial variants, which was compared with the experimental fitness observed. The theoretical fitness of combinatorial variants was extremely predictive for the two-gene library of the DNA replicator but less for the larger pathway library. Altogether, our approach exemplifies how combinatorial testing can be expanded from single proteins to multiprotein systems, which can in the future be extended to the evolutionary engineering of even larger genetic and metabolic networks, and eventually an entire artificial cell.

Manuscript published as: van den Brink, M., Claassens, N. J., Danelon, C. *ACS Synthetic Biology* (2026). doi: 10.1021/acssynbio.6c00166.

Graphical abstract



Introduction

Cell-free gene expression offers an ideal playground to explore biological designs. It finds applications in prototyping, biomanufacturing, and has emerged as a platform to build a synthetic cell[1,2]. The reconstitution of metabolic pathways, genetic circuits or biosensors often requires optimization steps to improve functional performance. For example, the engineering of regulatory parts, such as promoters, terminators and ribosome binding sites (RBSs) has become a routine method to control gene expression – and thereby protein concentrations – *in vivo* and *in vitro*, thus enhancing a certain objective function, for instance maximizing the yield of a desired enzymatic product[1,3].

Large libraries of DNA variants can be assayed by high-throughput screening of femto- to picoliter cell-free expression reactions encapsulated in lipid vesicles or droplets[4–13]. While many examples of such *in vitro* evolution screens have been reported for the optimization of single genes, multivariate optimization involving multiple proteins is more challenging due to the vast combinatorial space and unpredictable epistatic effects. Examples of multi-protein system optimization targets include (i) the production of proteins at stoichiometric amounts within a macromolecular complex, (ii) the balanced production of enzymes involved in a pathway to enhance metabolic fluxes and avoid accumulation of (toxic) intermediate products, (iii) fine tuning gene expression profiles for efficient allocation of resources, and (iv) the regulation of genetic networks to promote the integration of and coordinate different biological functions. Living organisms navigate such complexity through natural evolution across extremely long timespans. Accelerating the optimization of multigene functions through directed evolution would require approaches that reduce the problem of combinatorial explosion, for instance by generating 'small but smart' libraries, and implement selection procedures that retain the best performing designs[14–16].

In this study, we performed *in vitro* evolution of two genetic modules, each encoding multiple proteins involved in DNA replication or phospholipid synthesis. The genomes were expressed with PURE system (Protein synthesis Using Recombinant Elements)[17] in liposome-based synthetic cells. We here define 'synthetic cell' as an umbrella term for the compartmentalized expression of genes encoding a cellular function. We investigated the effects of modified RBS sequences or the GC content of the first codons on the functionality of the two genetic systems. It is known from previous studies that translation initiation is often the rate-limiting step in bacterial translation[3,18], and that the initiation rate also controls protein yield in PURE system[19,20]. However, it remains unclear how changing the translation rates of multiple genes impacts the overall expression pattern in PURE system, possibly distorted by epistatic interactions between genes competing for translation resources (e.g., tRNAs, tRNA synthetases, amino acids). It is also unknown whether control over protein amounts is effective in regulating the functionality of the protein systems in synthetic cells.

To address these questions, we expressed semi-rational, combinatorial DNA libraries in synthetic cells. The fittest variants were isolated either through selection by DNA self-replication or through a FACS-based liposome screening method with a reporter to detect the output molecule of the phospholipid synthesis pathway. Long-read sequencing of the fittest variants informed on the optimal RBS strengths and base substitutions in the first codons. The data also suggested which genes were most impactful in regulating the functionality of the modules, and whether the effects of multiple mutations were predictable or unexpected. Altogether, these results lay the groundwork for future evolutionary engineering efforts by demonstrating how increased DNA template complexity can be harnessed in synthetic cell systems.

Results

Design of a combinatorial library of RBS variants in a DNA replicator

Protein-primed DNA replication was established in liposomes by the *in vitro* expression of two proteins from the *Bacillus subtilis* bacteriophage Phi29[21] (**Fig. 1a**). The Phi29 DNA polymerase (DNAP) and terminal protein (TP) were produced from a linear template flanked with the origins of replication from the Phi29 genome. DNAP and TP form a 1:1 heterodimer and recognize the terminal origins of replication, where the TP forms a covalent bond with the first deoxyadenosine monophosphate (dAMP) of the new 5′ DNA strand[22]. The polymerase then elongates the TP-primed DNA strand with strand-displacement of the original DNA. The process is further facilitated by the addition of purified double-stranded binding protein (DSB) and single-stranded binding protein (SSB) to unwind the DNA at the origins and stabilize displaced single-stranded DNA, respectively.

The effects of changing the absolute and stoichiometric amounts of DNAP, TP and DNA on replication efficacy has been studied using purified proteins in simple buffer conditions[23–25]. However, it is unknown how the production rates of DNAP and TP impact DNA replication in this reconstituted autocatalytic network, where protein concentrations vary over time and DNA is also the substrate of RNA polymerases which may collide with DNAP[26]. To test a range of protein production levels, we mutated the RBSs for each of the two coding sequences on the template. We targeted the RBSs, because (i) protein production level is better correlated with translation rate than transcription rate in PURE system[19,20], and (ii) translation initiation is often a major rate-limiting step in protein translation in *Escherichia coli*[18,27] thus possibly also in PURE system[28]. Because the strength of an RBS is influenced by the ~35 nucleotides upstream and downstream the start codon[18,29–31], we designed a specific set of RBS sequences for each of the protein coding sequences using RBS Calculator[18]. The library was designed to maximize the range of predicted translation initiation rates (TIRs) per gene, while keeping the number of mutations between one and five base pair substitutions relative to each wild-type RBS (**Table S1**). The combinatorial library consisted of 12 RBS variants for DNAP and 13 RBS variants for TP, thus in total 156 combinatorial variants. We constructed the library by our recently published protocol for multiplex one-step single-stranded DNA annealing protein (SSAP)-mediated plasmid diversification (MOSAIC)[32]. In short, this method relies on plasmid recombineering in an *E. coli* strain equipped with recombination machinery and requires just one co-electroporation of the target plasmid and degenerate ssDNA oligos encoding the mutations of interest. After one round of MOSAIC, 33% of the DNA library variants harbored two mutated RBSs (**Fig. S1**). The remaining fraction of the library consisted of variants with only one modified RBS (46%) and the wild type (22%).

Self-selection of RBS variants through DNA replication

To establish a genotype-phenotype link, the library of DNA replicators was expressed by PURE system in liposomes at an average DNA copy number of one per liposome (λ = 1, Methods section). After 16 h of incubation at 30 °C, quantitative PCR (qPCR) analysis confirmed 10 to 100 fold DNA amplification (**Fig. S2a-b**). The resulting population of full-length DNA was amplified from the liposome sample by PCR and re-encapsulated in liposomes for a second round of *in vitro* transcription-translation-replication (IVTTR) (**Fig. 1a**, **Fig. S2c**). Nanopore sequencing of the initial library and the recovered DNA after round one and two revealed changes in combinatorial variant frequencies (**Fig. 1b**). Certain combinations of RBS variants were enriched (log2 fold change (log2FC) > 0), while others decreased in frequency (log2FC < 0) or disappeared from the library (**Fig. 1b**, **Fig. S3**). Overall, the sequence landscape showed a gradual increase of fitness, defined as log2FC, with increasing predicted TIR. The most enriched variants harbored strong RBSs for both TP and DNAP. Interestingly, the fitness was mostly dependent on the predicted TIR of TP and less of DNAP (**Fig. S4**). This could mean that either TP is required in higher amounts and/or the TP RBS library covers a lower expression range than predicted. Specifically, for both DNAP and TP, the log2FC of the RBS variants increased with the predicted TIR up to a certain TIR, whereafter the log2FC plateaued (TP) or

slightly decreased (DNAP) (**Fig. S4**). Some RBS variants led to unpredictable results, indicated by a rugged pattern, most pronounced for TP RBSs s8-s11, possibly due to inaccurate predictions of the TIRs for some RBSs (**Fig. S3, S4**). Three biological replicates resulted in similar frequency changes, indicating high reproducibility of the self-selection system behavior (**Fig. S3**). Furthermore, while mutations could be introduced during in-liposome replication by the Phi29 DNA polymerase or during DNA recovery by PCR, no mutations outside the targeted RBS regions were detected in the DNA population with frequency ≥1%, demonstrating that the observed variations in DNA replicator fitness exclusively come from the mutations in the RBS sequences.

Reverse engineering of clonal self-replicator variants reveals high performance

We then sought to validate that the observed changes in variant frequency were proportional to the ability to self-replicate (i.e., the fitness of each DNA variant). Therefore, we constructed a few clonal variants from the sequence landscape, and we assessed their self-replication ability in liposomes using qPCR (**Table S2, Fig. 1c-d**). Two high-frequency variants, V1 and V2, showed high replication folds. In contrast, V4 and V5, two variants in the "valley" of the sequence landscape, did not replicate at all. We tested another DNA variant, V3, which had the same TP RBS as V1 but a DNAP RBS that was not part of the theoretical library. This DNAP RBS variant was most likely generated as by-product during library generation with MOSAIC or linearization of the library with PCR, as it was already present before encapsulation in liposomes. V3 started at a 50-fold lower frequency than V1 in the initial library (R0), but V3 was enriched 3.2 ± 0.5 fold more than V1 in R2 (mean ± standard deviation, n=3). Indeed, clonal V3 also replicated well (**Fig. 1c-d**).

In addition to the population-level qPCR analysis, we then investigated the performance of each replicator variant at the single-liposome level using confocal fluorescence microscopy (**Fig. 1e**). We detected DNA-replicating liposomes based on the intensity of the fluorescent dye dsGreen, which binds specifically to double-stranded DNA. We plotted the average dsGreen intensity per liposome (**Fig. 1f, Fig. S5**) and found similar results as the qPCR data, corroborating that the variant frequency change (log2FC) is indeed an accurate measure for its fitness.

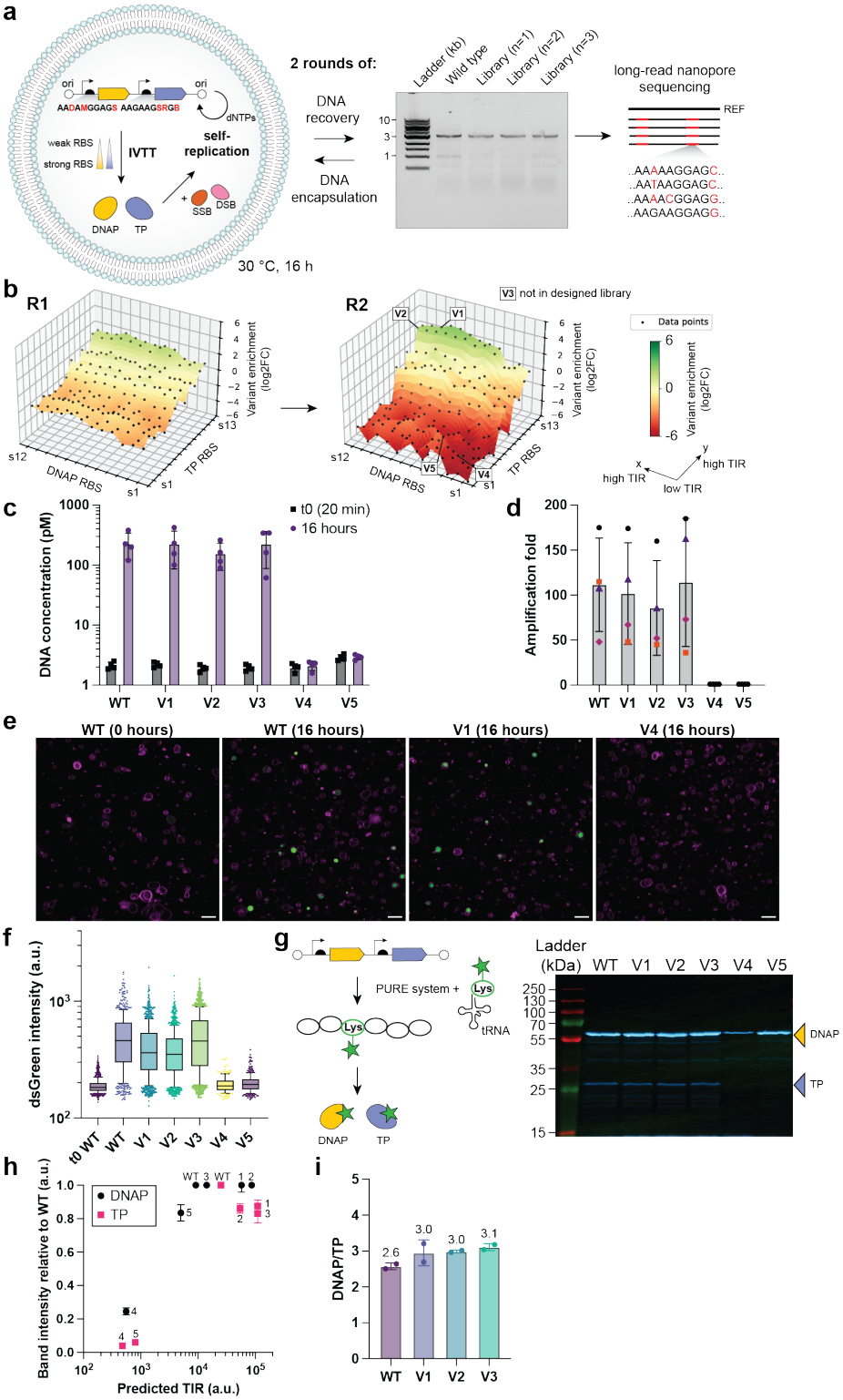

a
ori
AADAMGGAGS AAGAAGSRGB
dNTPs
weak RBS
strong RBS
IVTT
self-replication
DNAP
TP
SSB
DSB
30 °C, 16 h
2 rounds of:
DNA recovery
DNA encapsulation
Ladder (kb)
Wild type
Library (n=1)
Library (n=2)
Library (n=3)
10
3
1
long-read nanopore sequencing
REF
..AAAAAGGAGC..
..AATAAGGAGC..
..AAAACGGAGG..
..AAGAAGGAGG..
b
R1
R2
V3 not in designed library
Data points
Variant enrichment (log2FC)
DNAP RBS
TP RBS
s1
s12
s13
V1
V2
V4
V5
high TIR
low TIR
c
DNA concentration (pM)
t0 (20 min)
16 hours
WT
V1
V2
V3
V4
V5
d
Amplification fold
e
WT (0 hours)
WT (16 hours)
V1 (16 hours)
V4 (16 hours)
f
dsGreen intensity (a.u.)
t0 WT
g
PURE system +
Lys
tRNA
DNAP
TP
Ladder (kDa)
250
130
100
70
55
35
25
15
h
Band intensity relative to WT (a.u.)
Predicted TIR (a.u.)
i
DNAP/TP
2.6
3.0
3.0
3.1

Fig. 1 Selection of RBS variants through DNA self-replication. a Schematic illustration of *in vitro* transcription-translation (IVTT) and DNA self-replication of the RBS library. DNAP and TP were expressed in liposomes by PURE system. SSB and DSB were supplied as purified proteins. DNA from the liposome samples was recovered by PCR, verified by gel electrophoresis and analyzed by nanopore sequencing. **b** 3D landscapes displaying the log2FC in read frequency of combinatorial variants. The RBS variants on the x- and y-axes are ordered from low-to-high predicted TIR. The landscapes show the log2-transformed average FC of three biological replicates. The number of reads covering the 156 variants are 28,978 in R0 (pre-selection), 10,140 to 16,998 per replicate in R1, and 3,900 to 6,060 per replicate in R2. Data from the individual repeats are shown in Fig. S3. **c** Quantification of DNA concentration by qPCR for clonal DNA variants V1-5 before replication (t0, incubation 20 min) and after replication (16 h) at 30 °C (n=4 biological replicates). **d** Amplification fold of the data in panel c, calculated as the DNA concentration at 16 h divided by the concentration at 20 min. Symbols represent the different biological replicates (n=4). **e** Confocal fluorescence images of liposomes with different clonal DNA variants. Magenta, Cy5-conjugated lipids; green, dsGreen; Scale bars: 10 µm. **f** Quantification of the dsGreen intensity of dsGreen-positive liposomes (>129 a.u.) from confocal microscopy images. Number of dsGreen-positive liposomes: 599 (t0 WT), 487 (WT), 1,056 (V1), 1,040 (V2), 1,255 (V3), 170 (V4), 332 (V5). **g** SDS-PAGE of bulk (no liposomes) PURE reaction samples from clonal DNA variants with GreenLys protein labeling reagent. The fainter bands for TP compared to DNAP could be the result of the lower number of lysine residues of TP (27) compared to DNAP (64). **h** Band intensities from the SDS-PAGE gel in panel g relative to the wild type versus the corresponding predicted TIRs (n=2). The DNA variant is indicated for each data point. **i** Ratio of corrected band intensities for DNAP over TP. Corrected band intensities were calculated by dividing the adjusted band intensity (intensity with background subtracted) by the number of lysines per protein (64 for DNAP, 27 for TP).

Lastly, we directly measured the amounts of synthesized proteins from the different RBS variants to examine the correlation with fitness. The proteins were expressed in bulk reactions in the presence of tRNA preloaded with fluorescently labelled lysine, and we quantified the fluorescence intensity of the bands after SDS-PAGE (**Fig. 1g-h**, **Fig. S6**). V1-3 showed clear bands for DNAP and TP, with intensities similar to the wild type, despite having higher predicted TIR values for both DNAP (14,117-86,579 a.u.) and TP (54,294-110,313) (comparison to wild type (WT): DNAP: 9,123; TP: 25,228). This observation coincides with the plateauing fitness (log2FC) at high predicted TIRs for DNAP and TP in the library dataset (**Fig. S4**), and with the similar DNA amplification folds for the high-TIR clonal variants (**Fig. 1d-f**). The estimated ratio of molecules DNAP over TP, derived by normalizing the band intensity to the number of lysine residues per protein, increased slightly from 2.6 for the wild type to 3.0-3.1 for the library variants V1-3 (**Fig. 1i**). In contrast, V4 and V5, which have lower TIRs for both DNAP and TP, showed reduced DNAP and TP expression levels, most likely impairing their inability to replicate. Notably, we here quantified the end-point expression level, which may differ from the protein yield and stoichiometry at earlier time points in the reaction.

Design of a combinatorial RBS library of a 4-gene phospholipid synthesis pathway

Next, we performed multivariate optimization of a four-gene phospholipid synthesis pathway, generating a cascade of reactions catalyzed by membrane-associated or transmembrane proteins (**Fig. 2a**). We selected the four upstream enzymes of the *E. coli* Kennedy phospholipid pathway PlsB, PlsC, CdsA and PssA. These enzymes stepwise convert oleoyl-coenzyme A

(oleoyl-CoA) and glycerol-3-phosphate (G3P) to phosphatidylserine (PS) in the presence of the co-substrates cytidine triphosphate (CTP) and L-serine. The amount of final product PS is positively correlated with the fluorescence intensity of the PS-specific reporter LactC2-mCherry, enabling the quantitative assessment of PS production and membrane incorporation at the single-liposome level[33].

While *in vivo*, the phospholipid synthesis pathway is mainly regulated post-translationally by an (unknown) allosteric regulator of PlsB activity[34] and metabolic regulation of phosphatidylglycerol (PG) and phosphatidylethanolamine (PE)[35], we reasoned that a minimal synthetic approach based on translational control could modulate phospholipid production *in vitro*. We hypothesized that the yield of PS production depends on the stoichiometry of the expression of the four enzymes, as observed in other enzymatic cascades[14,36,37]. Previously, it was found that phosphatidic acid (PA), the product of PlsC, significantly accumulates in the Kennedy pathway reconstituted in liposomes, suggesting potential for further optimization[33]. However, the yield of PS has a nontrivial dependency on the four RBS sequences, as the rate-limiting step is unknown and unexpected effects of mutating RBSs could arise from protein folding, membrane insertion/binding, allosteric regulation, inhibitory effects, or resource competition.

To tackle this multivariate problem, we generated a 4-site combinatorial RBS library by MOSAIC consisting of 13 (PlsB) × 11 (PlsC) × 9 (CdsA) × 9 (PssA) = 11,583 variants (**Table S3**). With a MOSAIC editing efficiency of ~50% per target locus, we obtained a library that consisted for 11% of variants with all RBSs mutated, and the remainder of the library consisted of library fractions where zero, one, two or three RBSs were mutated (**Fig. S7a, c**). Using nanopore sequencing, we detected all designed variants that had one or two mutated RBSs, and 84% and 35% of all designed variants with three and four mutated RBSs, respectively, but more of the designed variants are likely covered in the generated library as suggested by rarefaction plots (**Fig. S8**). Hereafter, we name the fractions of the library with *N* mutated target loci (RBSs) "*N*-site fraction" (e.g., 1-site fraction).

Enrichment of library variants with high PS yield through screening by FACS

We encapsulated the linearized library in liposomes at an average DNA copy number per liposome $\lambda = 0.2$ (Methods). Functionality screening of the liposomes was performed by measuring the fluorescence intensity of the PS probe LactC2-mCherry with flow cytometry. We typically screened ~3-6 million liposomes of which 2-13% showed PS production (**Fig. S9-S10**). We calculated the screening coverage of the DNA library as shown in **Table S4**, taking into account that 13 ± 8% of a clonal liposome population is active (i.e., the performance of the wild type, **Fig. S9a**) (mean ± standard deviation, n=5 biological replicates). The screening coverage was approximately $3x10^3$-fold for the 1-site fraction (42 variants), $3x10^2$-fold for the 2-site fraction (656 variants), $3x10^1$-fold for the 3-site fraction (4,518 variants) and 6-fold of the 4-site combinatorial library fraction (11,583 DNA variants). Using FACS, we sorted the

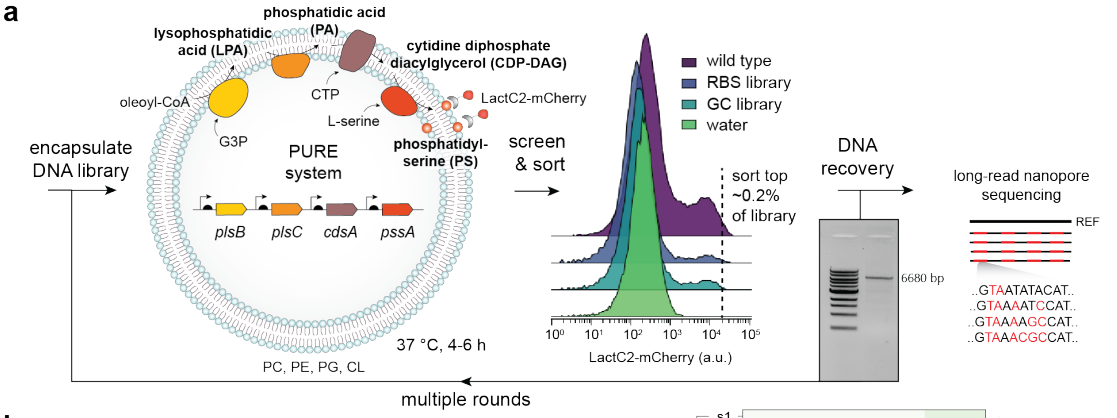
liposomes with the highest LactC2-mCherry intensity, less than 1% of all liposomes, corresponding to 10,000-30,000 liposomes. We amplified the DNA from the sorted liposome sample by PCR and analyzed the DNA sequences by nanopore sequencing. We performed four subsequent rounds of sorting, DNA amplification and re-encapsulation in liposomes.

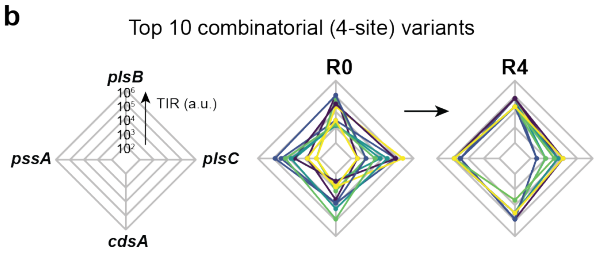
The combinatorial (4-site) genotypes of the top ten most frequent DNA variants are presented in **Fig. 2b** (and **Fig. S12**). Due to the relatively low sequencing coverage of the 3-site and 4-site library fractions in the presort library (R0) (on average 4 ± 4 (3-site) and 2 ± 2 (4-site) reads per variant), we used the read frequency as a measure of fitness instead of calculating the fold change in frequency relative to R0. Comparing the initial library to that after four rounds of sorting, we see a strong shift towards high protein expression levels for PlsB, CdsA and PssA, while a range of RBS strengths were present for PlsC. To obtain a more holistic view on the full dataset, we then plotted the average frequency of each RBS for each fraction of the library (**Fig. 2c, S12**). These results support the observation that high-fitness variants contain *plsB*, *cdsA* and *pssA* RBSs with high predicted TIRs (*plsB*: 21,409-112,986 a.u.; *cdsA*/*pssA*: 31,879-78,824 a.u.) but a more even distribution of frequencies is observed for *plsC* RBSs. An exception is s5 from cdsA, which has a relatively high frequency considering its TIR of 4,501.

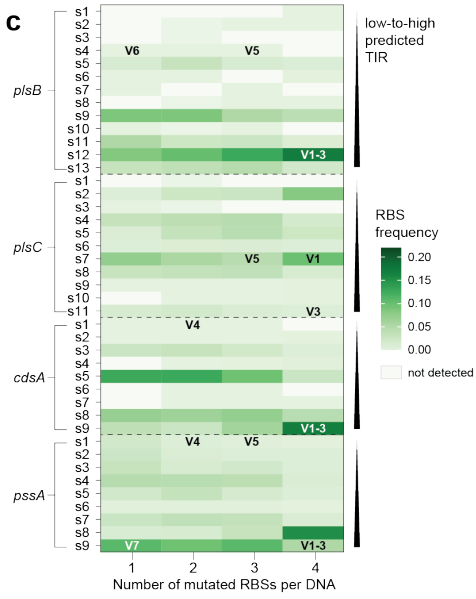
The wider range of RBS strengths for PlsC among the top performers can be interpreted in the light of our previous study reporting that among the intermediate products of the Kennedy pathway only PA accumulated in significant amounts[33]. The conversion of PA by CdsA limits the flux rendering PS production poorly sensitive to the PA production rate and thus also to the amount of PlsC.

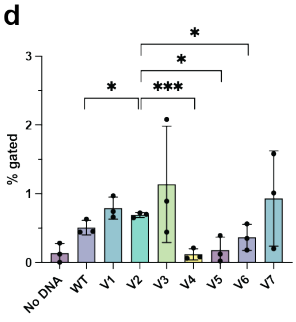
Reverse engineering of clonal variants of the phospholipid synthesis pathway

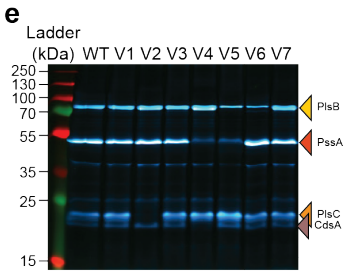
Due to the repeated PCR steps involved, some mutations accumulated with a frequency >0.04 at off-target locations. To validate the fitness of DNA variants as predicted by the sequencing data, we constructed three clonal variants of expected high fitness with strong RBSs for *plsB*, *cdsA* and *pssA* and varying strengths of the *plsC* RBS (V1-3) (**Fig. 2g (top row), Table S5**), and two clonal variants of expected low fitness, one with weak *cdsA* and *pssA* RBSs (V4) and one with weak *plsB* and *pssA* RBSs and stronger *plsC* RBS (V5). All DNA variants led to PS production in liposomes (**Fig. S13, S14**). However, the population size of the top-producing liposomes differed among the variants. Specifically, we quantified the percentage of liposomes in the strict LactC2-mCherry gate, similar to the sorting gates in the library enrichment phase (i.e., the selection pressure) (**Fig. 2d**). As expected, we found that V1-3 performed well, with 1.4- to 2.2-fold larger gated population sizes (mean values across three replicates) compared to that of the wild type (unpaired *t*-test for WT versus V2 was significant (P = 0.05), but those for WT versus V1 or V3 were not). In contrast, V4 and V5 performed poorly. We also found suboptimal performance of a variant with a predicted weak RBS for *plsB* (V6) and good performance of a variant with a predicted strong RBS for *pssA* (V7), as expected from their frequencies in the sorted library as shown in **Fig. 2c**. We also compared the V3, V5 and WT designs on the single-

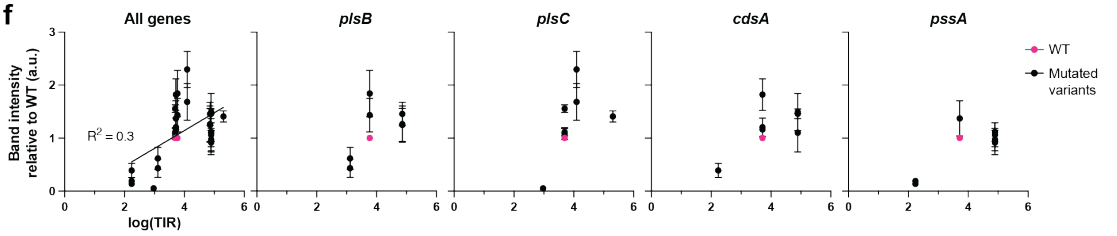

a
phosphatidic acid (PA)
lysophosphatidic acid (LPA)
cytidine diphosphate diacylglycerol (CDP-DAG)
oleoyl-CoA
CTP
L-serine
LactC2-mCherry
G3P
PURE system
phosphatidyl-serine (PS)
plsB plsC cdsA pssA
encapsulate DNA library
screen & sort
wild type
RBS library
GC library
water
sort top ~0.2% of library
LactC2-mCherry (a.u.)
DNA recovery
6680 bp
long-read nanopore sequencing
REF
..GTAATATACAT..
..GTAAAATCCAT..
..GTAAAAGCCAT..
..GTAAACGCCAT..
37 °C, 4-6 h
PC, PE, PG, CL
multiple rounds


b
Top 10 combinatorial (4-site) variants
plsB
TIR (a.u.)
pssA
plsC
cdsA
R0
R4


c
plsB
plsC
cdsA
pssA
low-to-high predicted TIR
RBS frequency
0.20
0.15
0.10
0.05
0.00
not detected
Number of mutated RBSs per DNA


d
% gated
No DNA
WT
V1
V2
V3
V4
V5
V6
V7


e
Ladder
(kDa) WT V1 V2 V3 V4 V5 V6 V7
PlsB
PssA
PlsC
CdsA


f
All genes
plsB
plsC
cdsA
pssA
Band intensity relative to WT (a.u.)
R2 = 0.3
log(TIR)
WT
Mutated variants

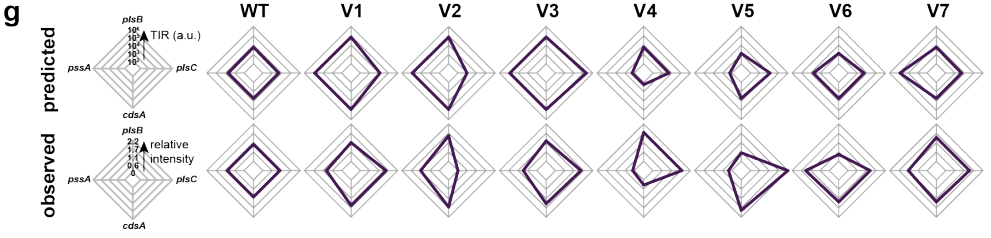

g
WT
V1
V2
V3
V4
V5
V6
V7
predicted
observed
TIR (a.u.)
relative intensity

Fig. 2 Activity-based selection of DNA designs for PS production from a library of four-site combinatorial variants. a Schematic illustration of the workflow. Liposomes express four enzymes from a DNA template with diversified RBSs. The enzymes form a pathway to produce PS, measured by the LactC2-mCherry probe. Four subsequent rounds of sorting the liposomes with highest LactC2-mCherry signal and re-encapsulation of the recovered DNA were performed prior to long-read sequencing. **b** Radar charts displaying the top 10 combinatorial variants with the highest frequency in the initial library (R0) and after four rounds of sorting. The variants are listed in Table S6. **c** Heatmap displaying the frequency of each RBS variant. The frequency was calculated relative to the library fractions with *N* number of mutated RBSs per DNA. Number of reads per library fraction (each column in the heatmap): 9,123 (1-site), 8,510 (2-site), 3,119 (3-site), 697 (4-site). **d** Analysis of clonal DNA variants from the RBS library by flow cytometry. The percentage of gated liposomes was calculated using a similar gate as the one used for sorting the DNA library. Unpaired *t*-test: * $P \leq 0.05$; *** $P \leq 0.001$. n = 3 biological replicates. **e** SDS-PAGE with bulk PURE system reactions (no liposomes) containing the GreenLys reagent for co-translational labeling of proteins expressed from clonal DNA variants. **f** Band intensities were measured from the SDS-PAGE gel in panel g, normalized to the wild type, and plotted against the corresponding predicted TIRs (n=2). The appended curve is the linear regression for all genes combined with $R^2 = 0.3$, $P < 0.0001$. **g** Radar charts displaying the predicted TIRs and observed relative protein expression levels.

liposome level using confocal fluorescence microscopy. The data suggest improved PS synthesis of V3 over the WT, and decreased activity for V5 (**Fig. S14**). Altogether, the experimentally determined performances of the clonal variants are in agreement with our expectations based on the sequenced libraries.

Next, we quantified the relative protein expression levels for V1-7 using a GreenLys SDS-PAGE assay (**Fig. 2e-g, Fig. S15**) and interpreted the efficiency of the lipid biosynthesis pathway considering these enzyme amounts. This assay confirmed that PS is still produced even with very low PlsC expression (V2), while a reduction in the production of PlsB (V6), or CdsA and PssA (V4), or PlsB and PssA in combination with an increase in PlsC production (V5), reduced PS production significantly (unpaired *t*-tests relative to V2, $P \leq 0.05$, **Fig. 2d**). High expression of PlsC (V1) did not significantly change PS production compared to low expression (V2). Note that CdsA migrates faster than expected based on its molecular weight (31 kDa), as described earlier[38].

We found a weak positive linear correlation between the predicted log TIR and the observed protein expression level as quantified from GreenLys SDS-PAGE ($R^2 = 0.3$) (**Fig. 2f**). We observed a few instances suggesting that a change in production level of a given protein could influence the production levels of other proteins whose RBSs remained the same (**Fig. 2f**, plots for separate genes). Hence, the same TIR may give different band intensities for different genetic backgrounds. For instance, in V4, upon reducing CdsA and PssA amounts by introducing weak RBSs, the expression levels of PlsB and PlsC increased despite their wild-type RBSs were unmodified (**Fig. S15c**). Furthermore, a *plsC* RBS with TIR 12,360 a.u. led to a higher PlsC production in combination with the downregulation of PlsB and PssA (V5) than in combination with strong RBSs for PlsB, CdsA and PssA (V1) (**Fig. S15c**). On the contrary, a strong reduction in PlsC production in V2 compared to V1 and V3 did not result in a higher

expression of the other three proteins. These findings suggest that mutating RBSs of multiple genes leads to unpredictable behavior of the overall expression pattern in PURE system.

Mutations within the first six codons modulate phospholipid synthesis

We applied a second strategy to modulate the translation initiation rates of the four enzymes of the phospholipid synthesis pathway. The GC content of the sequence directly downstream of the start codon affects the translation rate *in E. coli*, most likely through the tendency of GC-rich sequences to form stronger secondary mRNA structures, which hinder translation initiation[31]. This relationship has recently been observed in PURE system as well[19,39,40]. Inspired by these studies, we varied the GC content of the first six codons between 28-56% for *plsB*, 22-50% for *plsC* and 33-61% for *cdsA* and *pssA*. Only the third base of the codons were mutated such that the mutations were silent, except for the fourth and sixth codons of *cdsA* and *pssA*. A methionine was mutated to isoleucine at the fourth codon of both genes to study a potential positive or negative effect of introducing a second start codon close to the canonical one. At the sixth codon, glycine was mutated to valine to expand the range of GC content, while accepting a potential effect of amino acid substitution.

The theoretical DNA library consisted of 127 (PlsB) × 191 (PlsC) × 127 (CdsA) × 127 (PssA) = 391,241,153 combinatorial (4-site) DNA variants, which is over 33,000-fold larger than the RBS library. Based on the performance of the wild type (**Fig. S9a**), we obtained approximately 3×10^2-fold screening coverage for the 1-site fraction (572 variants), while only ≤ 1-fold coverage for the fractions with multiple modified target loci (calculation explained in **Table S7**). However, since many DNA variants have the same GC content, we screen the combinatorial GC variants many times (**Table S7**). Therefore, the screening was not an exhaustive search of all combinatorial DNA variants, but of all GC contents.

After four rounds of sorting, we found that some DNA variants were enriched and many variants were depleted from the library (**Fig. S16**). Generally, most DNA variants were enriched in all library fractions, where each library fraction has a different number (*N*) of mutated loci per DNA (*N* = 1, 2, 3 or 4). This observation suggests that these variants were beneficial in different genetic contexts. However, there was no trend visible towards high or low GC contents for the different genes. The most enriched GC variants and their predicted TIRs are listed in **Table S8**.

We also investigated the base-specific effects of each randomized position on PS synthesis (**Fig. 3a**). Site-specific preferences were identified for the tested bases, although no overall preference for A or T over G or C was visible. Interestingly, the effect of the base G at the third position of the second codon was found to be very negative for all genes. This negative effect of the G at this specific position was also discovered in a recent systematic *in vivo* study involving an ultradeep characterization of a large range of randomizations at the 5′-UTR and the start of the coding sequence[31]. Thus, this shows an analogy between *in vivo* and *in vitro*

conditions. A few other bases were found to have a severing effect: G in sixth codon of *plsB*, G in fifth codon of *cdsA* and A in second codon of *pssA*. Lastly, changing the second methionine to isoleucine at the fourth codon seems beneficial for both *cdsA* and *pssA*.

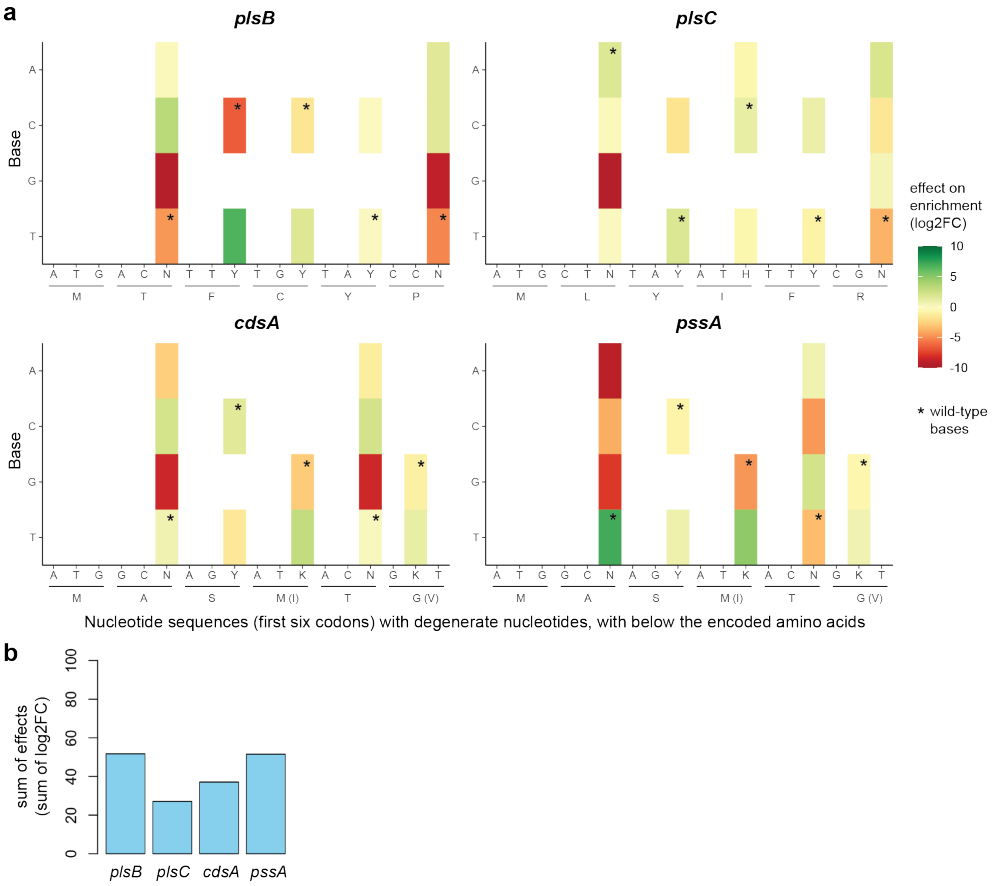


Fig. 3 Base-specific effects at specific nucleotide positions in the first six codons of the phospholipid synthesis pathway genes. Data are presented for the 1-site library fraction in R4, but the results across the four library fractions are similar (plots available in the accompanying dataset in the 4TU.ResearchData repository). Number of sequencing reads: 5,827 (*plsB*), 10,254 (*plsC*), 14,594 (*cdsA*), 3,350 (*pssA*). **a** The effect on enrichment (log2FC) was calculated by the fold change of the mean frequency of variants with a certain base at the specific nucleotide position over the mean frequency of variants with the other tested bases. The WT bases are indicated with an asterisk. The encoded amino acids in brackets match with the non-WT bases. **b** Sum of absolute per-base effects on enrichment for each gene.

To determine the total impact of the base substitutions for each gene on PS synthesis, we summed the effects of each tested base. The sum was the highest for *plsB* and *pssA*, suggesting that the mutations in these genes had the largest effect on PS synthesis. In contrast, mutations in the *plsC* gene had the smallest effect on PS production of all genes (**Fig. 3b**). This is in

agreement with the RBS library screening which showed that both low and high amounts of PlsC promoted PS synthesis (**Fig. 2d**).

We further investigated the effect of the mutations for the most impactful genes *plsB* and *pssA* by constructing clonal variants with bases yielding a more positive effect for PlsB (V8-11), or a more positive (V10) or negative (V11) effect for PssA (**Table S9**). SDS-PAGE analysis revealed that the PssA expression level was lower with V11, as expected (**Fig. S15c**). However, no significant increase in PlsB or PssA quantity was observed. Nevertheless, a deletion upstream the *plsB* gene of V11 and a deletion upstream the *pssA* gene of V9, which were likely introduced during the cloning of these variants, may have affected the TIR of the proteins.

Predicting fitness of combinatorial variants based on single variants

Lastly, we examined how predictable the combinatorial variants from the three tested libraries are. Understanding whether combinations of mutations lead to expected or unexpected effects can shed light on the presence of epistatic interactions and clarify whether testing variants in a combinatorial manner is necessary compared to evaluating them individually. Epistasis occurs when the observed fitness (here, variant frequency) deviates from the expected fitness when multiple mutations are combined. For example, the effect of a combination of two RBSs may be larger or smaller than the effect of the first RBS multiplied by the effect of the second RBS (multiplicative model), such that the fitness of the double mutant cannot directly be predicted from analyzing each position in isolation. Epistatic effects may for example appear when multiple genes compete for translation resources. For example, changing the TIR of some genes may indirectly influence the expression level of other genes (improving or severing) by modifying the utilization of tRNAs and tRNA synthetases, a phenomenon that is still poorly understood in cell-free protein synthesis systems.

We first studied the presence of epistatic interactions for the DNA replication RBS library by comparing the observed fitness of the combinatorial variants in the 2-site library fraction with the expected fitness derived from the single variants in the 1-site library fraction. We calculated the expected FC of the combinatorial variants without any epistatic effects by multiplying the FC of each individual RBS (**Fig. 4a**). We then compared the expected and observed sequence landscapes (log2FC, **Fig. 4b**). Both are very similar, with an R^2 of 0.94, suggesting that no large epistatic interactions govern DNA replication (**Fig. 4c**). Alternative calculations using the 2-site library fraction only (taking the average fold change) or using frequency instead of FC led to similar conclusions (**Fig. S17**). Therefore, the combinatorial variants of this DNA replication library were extremely predictable and, if there was any competition for translation resources, it had minimal effect on DNA self-replication.

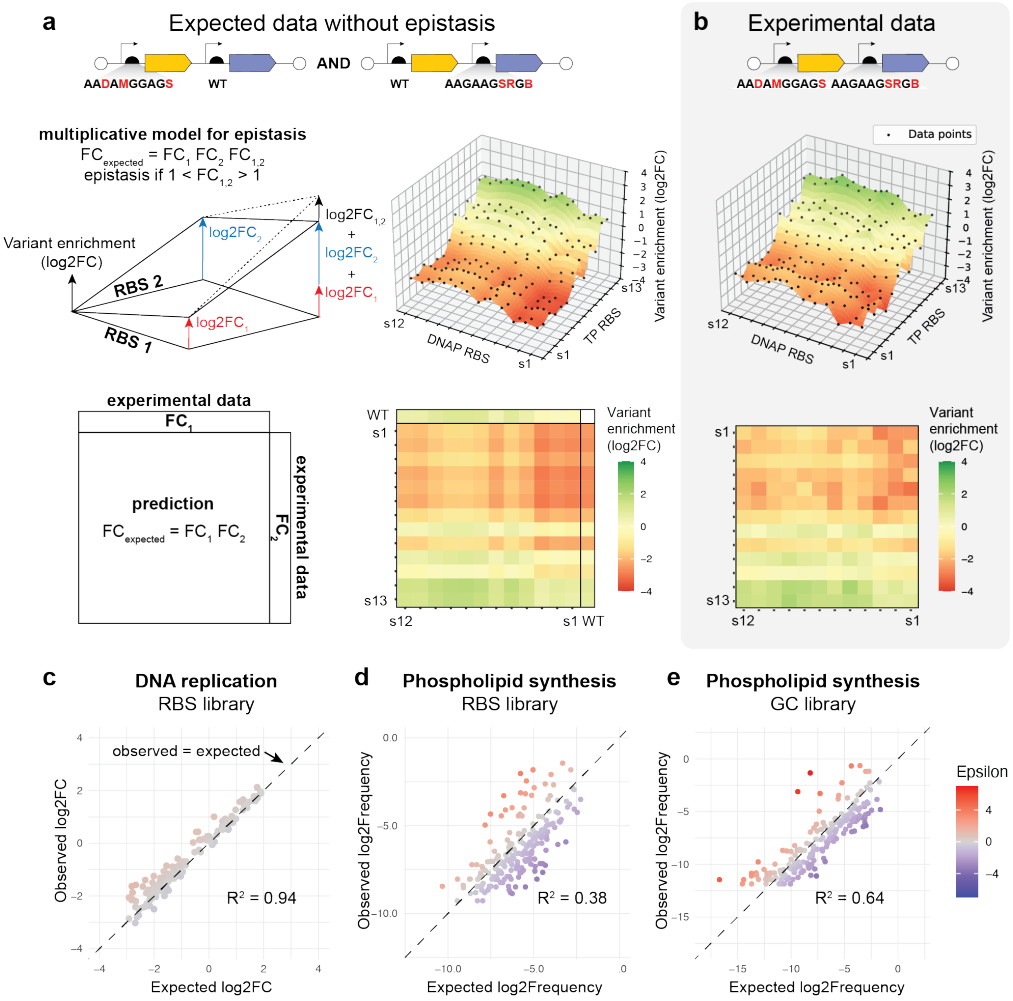


Fig. 4 Predicting the fitness of combinatorial variants based on single-mutated variants. a Expected combinatorial variant enrichment (log2FC) for the DNA replication RBS library. Prediction was based on a multiplicative model using the enrichment data of DNA variants with one WT and one modified RBS (1-site library fraction). The average of three biological replicates is plotted, where R0 contains 39,928 reads, and R1 33,834 to 56,684 reads per replicate. **b** Experimental data of variant enrichment for the combinatorial library of RBSs of the DNA replicator (average of 3 replicates). **c** Observed versus expected log2FC for the DNA replication RBS library (average of 3 replicates) for all 156 combinatorial variants. **a-c** Data for selection round 1 (R1) were used, because, unlike R2, all DNA variants were detected by Nanopore sequencing, offering a better resolution for low-fitness variants. **d-e** Observed versus expected log2Frequency for the phospholipid synthesis RBS library (d) and GC library (e). Data are from R4 (one biological replicate). Prediction is based on 1-site library fraction. Data points: 205 of the 656 (d) and 189 of the 121,158 (e) pairwise combinatorial variants. **c-e** Dashed line indicates $y = x$. Epsilon is the difference between the observed and expected fitness, i.e., Epsilon = $\log2FC_{observed} - \log2FC_{expected}$ (c), or Epsilon = $\log2Frequency_{observed} - \log2Frequency_{expected}$ (d-e).

Similarly, we predicted the fitness of the pairwise combinatorial DNA variants of the phospholipid synthesis pathway libraries. Here, we calculated the expected frequency instead of the expected log2FC because of the lower sequencing depth of the pre-sort, 2-site library fraction (24 ± 13 reads per variant for the RBS library) compared to the DNA replication library (186 ± 100 reads per variant) (mean ± standard deviation). The correlation between the expected versus observed fitness for the two PS synthesis libraries was weaker than that for the DNA replication library (R^2=0.38 for RBS library, 0.64 for GC library) (**Fig. 4d-e**). Alternative calculations using the 2-site library fraction only or using FC instead of frequency led to similar conclusions (**Fig. S18**). This means that the fitness of these combinatorial variants is less predictable than for the DNA replicators, possibly due to epistatic interactions and/or the more complex sorting workflow, which may increase the level of noise in the sequencing dataset. Nevertheless, it is reasonable that more complex protein systems like the PS synthesis pathway, which involve more genes than the DNA replication system, are more subjected to epistasis affecting gene expression levels as we have also observed from the GreenLys protein gels in **Fig. S15**: RBSs lead to different protein production levels for different genetic backgrounds.

Discussion

Targeted combinatorial randomization strategies have extensively been applied to optimize expression regulatory sequences for multiprotein systems *in vivo*, for instance genetic circuits[41] and metabolic pathways[14], but not in cell-free systems, where the screening of combinatorial libraries is only starting to be explored and so far only targeted single proteins[13,42,43]. Here, we used combinatorial library designs targeting the RBSs or first six codons of coding sequences to tune cell-free expression of multiple genes from multi-gene constructs. Through self-selection or functionality screening, followed by next-generation sequencing (NGS) retrieval of the high-performing variants, we identified variants regulating the functionality of its genetic module at the system's level. The proposed holistic engineering concept will aid in integrating biological functions in synthetic cells[44,45].

Several factors may bias the abundance of variants after selection, independent of the mutational effects: the large inherent heterogeneity in functionality within clonal populations of liposomes[46], differences between variant frequencies in the input library (**Fig. S3a**, **Fig. S12**), and the FACS sorting accuracy of only 80% as determined earlier[47,48]. We assumed that each DNA variant produces an output signal distribution of identical shape (e.g., width of the distribution), leading to a single output quantity that represents its fitness (i.e., amplification fold, or LactC2-mCherry intensity). Increasing the sampling resolution of the variant distributions through implementation of multiple sorting gates could improve the inference of the fitness value and provide insights into the variant's mechanism of action[49]. In addition,

information about low-fitness variants is valuable, especially when combined with active learning-assisted directed evolution. In active learning, machine learning models are used to efficiently navigate large combinatorial sequence spaces, but their predictive accuracy could be improved by including training data on low-fit variants[42,50–52]. Another improvement could be the implementation of unique molecular identifiers (UMIs) during the first 1-2 cycles of PCR to correct for changes in variant frequencies during DNA recovery[53–55]. Despite potential sources of 'noise', we demonstrated the functional selection of high-performing variants in one selection round of DNA replication, and within four rounds of sorting for the phospholipid synthesis pathway.

The simultaneous optimization of multiple genes led to mutations that were kilobases apart. Therefore, we used long-read nanopore sequencing for sequence-function mapping. Despite the relatively low single-read per-base sequencing accuracy (99.0-99.7%)[56], quality control applied to the targeted mutagenesis loci allowed for the detection and counting of DNA variants with minimal mutational distance (i.e., down to one nucleotide difference), as validated earlier[32]. Thus, we linked the mutations within single full-length reads without additional DNA processing steps, such as adding barcodes to identify the DNA variants[43] or UMIs for deriving consensus sequences from multiple reads[57].

Our experiments revealed that different target loci have distinct levels of control over the system's performance. In future optimization campaigns, mutagenesis should preferentially target these most impactful regions to navigate the fitness landscape efficiently, while the expression of other genes could be fixed at a lower rate to save translation resources. We also found that co-expressing multiple proteins with RBSs having theoretical strengths (much) higher than the wild-type RBS, did not lead to large increases in protein production compared to the wild type (**Fig. 1h**, **2f**). Instead, the most significant increases in the amounts of expressed proteins (maximum 1.8 to 2.3-fold relative to the wild type) were obtained in combination with downregulating the co-expression of another protein (**Fig. S15, V4 and V5**), which appears as a more powerful way to balance protein systems and metabolic fluxes than the sole implementation of strong RBSs. More extensive quantitative proteomics analysis would be needed to determine the precise genetic context effects on gene expression and to potentially generate certain design rules that could help optimize expression patterns of multigene constructs in PURE system.

Lastly, pairwise epistasis calculations showed high predictability for the RBS library for self-selection of a DNA replicator and lower predictability for the phospholipid synthesis pathway libraries (**Fig. 4**). These findings underscore the importance of combinatorial scanning over testing each mutation in isolation for different protein systems and/or selection procedures, especially for larger multi-gene systems. However, combinatorial testing comes at the expense of larger library sizes. As mentioned above, active learning strategies can accelerate the search of good designs in a large sequence space.

Other types of libraries could be explored to improve the expression and function of protein systems in synthetic cells. Possible targets include promoters[20,58], terminators, operator sites, gene order and orientation, promoter insulators (i.e., spacers upstream a promoter sequence)[59], and operon (multicistronic) designs[20,60]. For example, enhanced transcription can improve protein production[20], but excess mRNA molecules can sequester initiation factors of the translation machinery, thereby reducing protein synthesis rate[2,61]. Furthermore, protein sequence regions known to govern protein activity, such as active sites or binding sites, could be mutated. Lastly, mutating the type of amino acids in the N- and C-terminal regions, or nucleotide patches that might induce ribosome frameshifting (X/XXY/YYZ) or stalling, might increase the efficiency of protein folding and full-length synthesis[2,61].

In the future, the system's level optimization workflow presented here could benefit from the sorting of liposomes based on microscopy images instead of FACS[48], followed by multi-omics characterization[62–64]. An alternative strategy to variant screening is to couple gene-encoded protein systems to DNA replication, such that a higher fitness of the desired system promotes DNA replication[43,45,65]. In pursuing the development of a synthetic cell, establishing a holistic approach that harnesses combinatorial optimization of multi-protein systems will surely be needed to manage the increasing number of possible genetic designs.

Methods

Reagents and equipment

Chemicals were purchased from Sigma-Aldrich unless stated otherwise. MilliQ water used in buffers and solutions had a resistivity of 18.2 MΩ. DNA concentrations were measured using Qubit dsDNA HS (High Sensitivity) Assay Kit and the Qubit 4 Fluorometer (Invitrogen). Eppendorf Eporator electroporator was used at 1.8 kV. Plasmid isolation was performed with PureYield Plasmid Miniprep System (Promega). PCR-amplified DNA was purified using the QIAquick PCR & Gel Cleanup Kit (Qiagen).

Biological resources

Bacterial strains *E. coli* K-12 MG1655 and *E. coli* K-12 DH5α were used. pORTMAGE-Ec1 was a gift from the George Church lab (Addgene plasmid #138474; http://n2t.net/addgene:138474; RRID:Addgene_138474). Plasmid G435 was constructed by introducing a point mutation (to remove a PmeI recognition site) in G340, whose synthesis was described earlier[66]. Plasmid G555 (Addgene plasmid #216483; http://n2t.net/addgene:216483; RRID:Addgene_216483) was constructed as described earlier[32]. Plasmids and DNA oligos are listed in **Table S10 and S11**, respectively. LactC2-mCherry was recombinantly expressed and purified as previously described[47]. SSB and DSB were expressed and purified as described in Soengas *et al.*, 1995[67]

and Mencía *et al*., 2011[25], respectively, and stored at 10 mg/mL in 50 mM Tris, pH 7.5, 1 mM EDTA, 7 mM β-mercaptoethanol, 50% glycerol and 60 mM (SSB) or 100 mM (DSB) ammonium sulfate. Purified proteins were aliquoted and stored at −80 °C.

Design and synthesis of DNA libraries

The RBS libraries for plasmids G435 (encoding DNAP and TP) and G555 (encoding PlsB, PlsC, CdsA and PssA) were designed using the RBS calculator[18] in the "Optimize Expression Levels" mode (https://salislab.net/software/design_rbs_library_calculator) with parameters: host organism, *E. coli*; target minimum/maximum translation initiation rates, 1/1,000,000; genomic RBS sequence, mRNA sequence from transcription start until (and including) stop codon or terminator. The G555 GC library was designed manually by changing primarily the third bases of the codons to vary GC content of the first six codons of each gene between 22 to 61% (depending on the target locus), while maintaining the originally encoded amino acids (with two exceptions as indicated). The libraries were generated by MOSAIC, a multiplex one-step plasmid diversification protocol, which we recently developed[32]. In this method, DNA oligo libraries are co-electroporated with the target plasmid into *E. coli* cells expressing the single-stranded DNA annealing protein CspRecT from the pORTMAGE-Ec1 plasmid. During plasmid replication in the host cells, the oligos anneal to their target sites and integrate into the plasmid sequence creating a library of plasmids. The (degenerate) DNA oligos used in this study are listed in **Table S11**. For the G435 RBS library, a single electroporation yielded approximately 1,600 colonies, from which the plasmids were isolated. For the G555 RBS library, four electroporation reactions were performed in parallel, yielding in total 3,700 colonies. For the G555 GC library, two parallel electroporation reactions yielded in total 1,800 colonies. The plasmids were isolated from each parallel reaction separately.

PCR was performed on the purified plasmids to linearize the DNA and simultaneously purify the target DNA from the pORTMAGE-Ec1 plasmid. The PCR reaction mix contained 1× Phusion HF buffer, 200 µM dNTPs, 500 nM of each primer (491 ChD and 492 ChD), 2 U Phusion High-Fidelity DNA polymerase and 200 ng plasmid at 100 µL reaction volume. The mixture was split into 2×50 µL, and the thermocycling reaction was performed as follows: 98 °C for 30 s, 20-25 cycles of (98 °C for 5 s and 72 °C for 3 min (G435) or 6 min 20 s (G555)), 72 °C for 5 min. Ten units of DpnI were added to each 50 µL PCR mix and incubated for 1 h at 37 °C to digest the plasmids. The expected PCR products orip2p3 (3,214 bp, derived from G435) and oriPL (6,680 bp, derived from G555) were excised from a 0.7% agarose gel and further purified using QIAquick PCR & Gel Cleanup Kit (Qiagen). The DNA concentration was measured using the Qubit fluorometer. The four oriPL RBS library PCR products were then combined at equimolar amounts to increase library coverage. Similarly, the two oriPL GC library PCR products were combined.

Synthesis of clonal DNA variants

Clonal plasmid variants were also generated using MOSAIC with a few modifications. Instead of one electroporation, three subsequent rounds of electroporation and plasmid isolation were performed. The *E. coli* strain DH5α harboring pORTMAGE-Ec1 was used instead of MG1655 pORTMAGE-Ec1 to prevent the emergence of smaller plasmids due to recombination during repeated transformation steps. Electroporation was performed with 10-50 ng of plasmid mixture and 2.5 µM of each DNA oligo variant added to the reaction. The plasmids were isolated from colonies grown on ampicillin-supplemented agar plates (50 µg/mL) or from liquid cell culture when the number of colonies on the plates was lower than 50. After three subsequent electroporation steps for mutagenesis, DH5α cells were chemically transformed by heat shock using 3 ng of purified plasmid mixture containing the target plasmids and pORTMAGE-Ec1 to obtain clonally pure plasmids. The transformed cells were plated on ampicillin-supplemented plates to pick single colonies for subsequent plasmid isolation and sequence validation by nanopore sequencing.

In-liposome gene expression for DNA self-replication

For the expression of DNA libraries in liposomes, swelling solutions were prepared at a volume of 10 µL with 5 µL of PURE*frex* 2.0 (GeneFrontier) Solution I, 0.5 µL of Solution II, 1 µL of Solution III, 0.25 U/µL SUPERase·In RNase inhibitor (Invitrogen), 300 µM of dNTP Mix (Thermo Fisher), 20 mM of ammonium sulfate, 750 µg/mL of purified SSB, 210 µg/mL of purified DSB, and 50 pM of linear orip2p3 DNA. During this study, a change in the commercial PURE*frex* 2.0 solution resulted in impaired DNA replication. Therefore, swelling solutions with clonal DNA variants consisted of an adjusted PURE composition, which yielded better replication compared to the standard kit: 2.5 µL of custom PURE*frex* Solution I lacking tRNAs and NTPs (PFC-Z1608-EX) (i.e., half of the recommended volume), 0.5 µL of Solution II, 1 µL of Solution III, 0.25 U/µL SUPERase·In, 300 µM of dNTP Mix (Thermo Fisher), 20 mM of ammonium sulfate, 750 µg/mL of purified SSB, 210 µg/mL of purified DSB, 50 pM of linear orip2p3 DNA and 1/3rd of the recommended concentrations of NTPs and tRNAs.

The preparation of lipid-coated beads was performed as described earlier[46] using 50 mol % DOPC, 36 mol % DOPE, 12 mol % DOPG, 2 mol % 18:1 cardiolipin, 0.01 mol % DOPE-Cy5 and 1 mass % DSPE-PEG(2000)-biotin. Five to six milligrams of lipid-coated beads were added to 10 µL of the swelling solution and the tubes were tumbled at 4 °C for 1 h. After four freeze-thaw cycles by dipping the tubes in liquid nitrogen and thawing on ice, 3 µL of liposomes were transferred to PCR tubes containing 0.25 µL of DNase I (2 U/µL) using a cut pipette tip. The liposomes were incubated in a thermocycler at 30 °C for 16 h to enable gene expression, DNA replication, and degradation of non-encapsulated DNA by DNase I. Negative control samples were incubated at 30 °C for 20 min to activate DNase I. The liposomes were then incubated at 75 °C for 15 min to heat-inactive DNase I before amplifying the intraliposomal DNA by (quantitative) PCR.

Verification of in-liposome DNA replication using quantitative PCR

Quantitative PCR reaction mix was set up at 10 µL volume and contained 5 µL PowerUp SYBR Green Master Mix (Applied Biosystems), 400 nM of each primer (976 ChD and 977 ChD) and 1 µL of heat-inactivated liposome sample diluted 100-fold in MilliQ water. The thermocycling was performed on a Quantstudio 5 Real-Time PCR instrument (Thermo Fisher) as follows: 50 °C for 2 min, 94 °C for 5 min, 45 cycles of (94 °C for 15 s, 56 °C for 15 s and 68 °C for 30 s), 68 °C for 5 min, followed by a melting curve from 65 °C to 95 °C. The DNA concentration was determined using a standard curve of the wild-type DNA with 10-fold dilutions ranging from 0.0001 to 100 pM and the Quantstudio Design and Analysis software v2.6.0 (Thermo Fisher).

DNA recovery from liposomes after DNA replication

The DNA was recovered from the liposomes by PCR amplification. The PCR reaction mix (100 µL) contained 1× Phusion HF buffer, 200 µM dNTPs, 500 nM of each primer (491 ChD and 492 ChD), 2 U Phusion High-Fidelity DNA polymerase, 5% DMSO and 10 µL of liposome sample diluted 100-fold in MilliQ water. The mixture was split into 2×50 µL, and the thermocycling reaction was performed as follows: 98 °C for 30 s, 30 cycles of (98 °C for 5 s and 72 °C for 3 min), 72 °C for 5 min. The expected PCR products (3214 bp) were excised from a 0.7% agarose gel and further purified using QIAquick PCR & Gel Cleanup Kit (Qiagen). The DNA concentration was measured using a Qubit fluorometer.

In-liposome gene expression for phospholipid synthesis

The lipid film swelling solutions were prepared at a volume of 20 µL with 10 µL of PURE*frex* 2.0 Solution I, 1 µL of Solution II, 2 µL of Solution III, 0.5 U/µL SUPERase·In, 1 mM CTP, 0.5 mM G3P, 0.5 mM L-serine, 5 mM β-mercaptoethanol, and 10 pM linear oriPL DNA. Lipid-coated beads were prepared and added to the swelling solution as described above, except that the mass of beads was doubled, and the tubes were tumbled at 4 °C for 1 h. After four freeze-thaw cycles, 15 µL of liposomes and 0.5 µL of DNase I (2 U/µL) were transferred to PCR tubes containing a dried film of oleoyl-CoA (Avanti Polar Lipids) to reach a final concentration of 176 µM oleoyl-CoA. The liposomes were incubated in a thermocycler at 37 °C for 6 h (R1), 5 h (R2) or 4 h (R3-5 and clonal variants).

Liposome screening and sorting with FACS

For each library (RBS or GC library), two identically prepared 15-µL liposome samples were screened and sorted using the BD FACSMelody cell sorter (BD Biosciences). 15 µL of liposomes were diluted in 1,300 µL buffer containing 20 mM HEPES-KOH, pH 7.6, 180 mM potassium glutamate, 14 mM magnesium acetate, and 300 nM LactC2-mCherry. The diluted liposomes were filtered using the 35 µm nylon mesh of a cell strainer cap of 5-mL round bottom polystyrene test tubes (Falcon). LactC2-mCherry fluorescence was measured using a 561-nm laser line and emission filter 610/20 nm. The photon multiplier tube voltages were set at 342 V for FSC, 405 V for SSC (with SSC threshold 359 V), and 461 V for LactC2-mCherry. First, a gate was set on FSC and SSC to select the main liposome population (~60-80% of all events)

(**Fig. S19**). For each sample, 5,000-30,000 liposomes with the highest LactC2-mCherry signal were sorted in purity mode, becoming more stringent in the later sorting rounds. The sorted liposomes from two identical samples were combined. The recorded data was further processed using Cytobank (https://community.cytobank.org/). Here, in addition to the gating based on FSC and SSC, events with LactC2-mCherry signal lower than 2 a.u. were discarded (**Fig. S19**).

DNA recovery from liposomes sorted by FACS

To recover the DNA from the sorted liposomes, the sorted fractions were spun down at 12,000 g for 3 min, whereafter the supernatant was removed. The remaining ~10 µL was incubated at 75 °C for 15 min to heat-inactivate DNase I. PCR reactions were prepared at 100-µL reaction volume by mixing 1× Xtreme buffer, 400 µM dNTPs, 300 nM of each primer (1459 ChD and 1460 ChD), 2 units of KOD Xtreme Hotstart DNA polymerase, and 10 µL of heat-inactivated liposome solution. The mixture was split into 2×50 µL, and the thermocycling reaction was performed as follows: 98 °C for 2 min, 35 cycles of (98 °C for 10 s and 68 °C for 6 min 40 s). The expected PCR products (6,616 bp) were excised from a 0.7% agarose gel and further purified using QIAquick PCR & Gel Cleanup Kit (Qiagen). A second PCR reaction was performed with ~60 ng of the purified DNA using primers 1301 ChD and 1518 ChD to restore the full-length DNA template (6,680 bp), and the amplicon was purified as described above. The DNA concentration was measured using a Qubit fluorometer.

Nanopore sequencing and analysis

DNA samples were diluted to a concentration of 34 ng/µL (orip2p3) or 71 ng/µL (oriPL). Nanopore sequencing was performed by Plasmidsaurus (Oregon, US). First, analysis of the raw fastq sequencing reads was carried out using the Galaxy web platform[68] (https://usegalaxy.org). The reads were filtered on length (3,020-3,300 nt for orip2p3, 6,400-7,000 nt for oriPL) using the *Filter FASTQ reads by quality score and length* tool[69], and mapped to the wild-type sequences using *Map with minimap2*[70]. *Minimap2* was executed using default parameters, except *retain at most INT secondary alignments* that was set to 0. For oriPL data, the *gap open penalty deletion* and *gap open penalty insertion* were set to 16 and 48, respectively.

The mapped reads were further processed using custom R scripts (R v4.4.1, RStudio v2023.09.1+494). A quality control was applied to the mutated regions of interest using a per-base quality cutoff of 50; this quality control method was validated as described earlier[32]. The variant frequency was calculated by dividing the number of reads of a given variant by the total number of reads within the corresponding *N*-site library fraction (i.e., the library fraction containing all sequences with *N* mutated target loci). The variant enrichment (defined as fold change, FC) was calculated by dividing the frequency of a variant by the frequency of that variant in the initial library before sorting (R0). The variant enrichment was plotted as the log2FC (log2 Fold Change). The sum of frequencies per RBS within each *N*-site library fraction,

as plotted in **Fig. 2c, S12, S16**, was divided by the number (*N*) of mutated target loci per DNA to normalize the total frequency within each *N*-site library fraction to 1.

Positional base-specific effects

The positional base-specific effects on variant enrichment were calculated as previously described[31]. The mean frequency of all DNA variants containing a specific base on a specific position was divided by the mean frequency of all DNA variants containing another base on that position. The effect on enrichment was plotted as the log2FC.

Fitness predictions of combinatorial variants without epistasis

The expected fitness of combinatorial variants was calculated by multiplying the fitness of each individual variant. As the frequency can be seen as the expected probability to sort and sequence a DNA variant, the Bayes' law about conditional probability can be used to support the choice for the multiplicative model, as suggested by a previous *in vitro* screening study[13]. DNA variants with less than three reads in the 2-site library fraction were excluded from the analysis.

For the DNA replicator library, the expected fitness and the difference between the observed and expected fitness (Epsilon) were calculated as follows:

$$FC_{expected} = FC_1 \times FC_2, \quad \text{(Eq. 1)}$$

where FC_i is the FC of the single variant in the 1-site library fraction.

$$Epsilon = log2FC_{observed} - log2FC_{expected}, \quad \text{(Eq. 2)}$$

where $FC_{observed}$ is the FC of the combinatorial variant in the 2-site library fraction.

For the PS synthesis libraries, we used frequency instead of FC due to the lower sequencing depth of the pre-sort libraries:

$$Freq_{expected} = Frequency_1 \times Frequency_2, \quad \text{(Eq. 3)}$$

where $Frequency_i$ is the frequency of the single variant relative to all variants with that target site mutated in the 1-site library fraction. The difference between the observed and expected fitness is thus defined as:

$$Epsilon = log2Frequency_{observed} - log2Frequency_{expected}, \quad \text{(Eq. 4)}$$

where $Frequency_{observed}$ is the frequency of the combinatorial variant relative to all variants with those two target sites mutated in the 2-site library fraction.

We repeated the calculation of the expected fitness using the 2-site library fraction instead of the 1-site library fraction, and we reached identical conclusions (**Fig. S17-S18**). The expected fitness and Epsilon were calculated for the DNA replicator and PS synthesis libraries as follows:

$$FC_{expected,2\text{-}site} = FC_1 \times FC_2, \quad \text{(Eq. 5)}$$

where FC_i is the *average* FC of all variants in the 2-site library fraction with mutation *i*.

$$Epsilon = log2FC_{observed} - log2FC_{expected,2\text{-}site}, \quad \text{(Eq. 6)}$$

where $FC_{observed}$ is the FC of the combinatorial variant in the 2-site library fraction.

$$Freq_{expected,2\text{-}site} = Frequency_1 \times Frequency_2, \quad \text{(Eq. 7)}$$

where $Frequency_i$ is the *sum* of frequencies of all variants in the 2-site library fraction with mutation *i*.

$$Epsilon = log2Frequency_{observed} - log2Frequency_{expected,2\text{-}site}, \quad \text{(Eq. 8)}$$

where $Frequency_{observed}$ is the frequency of the combinatorial variant relative to all variants in the 2-site library fraction.

Off-target mutations

The presence of off-target mutations (outside RBS regions) was analyzed by comparing the mapped reads of the DNA recovered from liposomes with the initial library in the Integrative Genomics Viewer (IGV) v2.16.0 software using a coverage allele-fraction threshold of 0.01 (DNA replication) or 0.04 (PS synthesis libraries).

GreenLys SDS-PAGE

Relative quantification of protein production was performed by running bulk PURE system reactions. Reactions were set up at 10 µL volume, containing 5 µL of PURE*frex* 2.0 Solution I, 0.5 µL of Solution II, 1 µL of Solution III, 0.25 U/µL SUPERase·In, 0.5 µL of 5-fold diluted FluoroTect $Green_{Lys}$ in vitro Translation Labeling System (Promega) and 1 nM of linear DNA template. The samples were incubated at 30 °C (DNA replication proteins) or 37 °C (phospholipid synthesis enzymes) for 2 h, unless indicated otherwise. Each sample was supplemented with 0.25 µL of RNase A Solution (4 mg/µL) (Promega) and incubated for 30 min at 37 °C. 5 µL of sample was mixed with 1.9 µL of solution containing 1× Laemmli SDS sample buffer, non-reducing (Thermo Fisher), and 10 mM dithiothreitol. The samples were incubated for 5 min at 70 °C to denature the proteins and run on an SDS-PAGE gel consisting of a 12% polyacrylamide resolving gel and a 4% stacking gel. Fluorescently labelled proteins were visualized with a Typhoon fluorescence gel imager (Amersham Biosciences) using a 488-nm laser and a 520-nm band-pass emission filter. Band intensities (adjusted band volume with background subtracted) were quantified using Image Lab v6.1.0 (Bio-Rad Laboratories) relative to the wild-type DNA sample.

Confocal microscopy and image analysis

Samples for confocal fluorescence microscopy were prepared by mixing 1-2 µL of liposomes with 8 µL of buffer containing 20 mM HEPES-KOH, pH 7.6, 180 mM potassium glutamate, 14 mM magnesium acetate and 5× dsGreen (Lumiprobe). Using a cut pipet tip, the 10-µL sample was transferred to a BSA-functionalized custom-made glass imaging chamber. The sample was incubated at 4 °C for 1-1.5 h before imaging for liposome sinking. Image acquisition was performed with the A1R Laser scanning confocal microscope (Nikon) with an SR Apo TIRF 100× oil immersion objective (NA 1.49) and NIS-Elements v5.30.07 software (Nikon). Imaging was performed using laser line 488 nm with emission filter 525/50 nm (dsGreen), laser line 561 nm with emission filter 595/50 nm (LactC2-mCherry), and laser line 640 nm with emission filter 700/75 nm (DOPE-Cy5). Acquisition settings and adjustments of brightness and contrast were

identical for multiple images when comparing fluorescence intensities. Using Fiji (ImageJ2 v2.9.0)[71], the dsGreen signal inside liposomes was quantified by detecting particles on the binarized dsGreen channel followed by measuring the mean pixel intensity for each particle. The quantification of LactC2-mCherry intensity was performed using SMELDit (https://github.com/DanelonLab/SMELDit).

Statistics for the number of DNA molecules per liposome

Following Poissonian partitioning, the probability of encapsulating k DNA molecules per liposome is:

$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad \text{(Eq. 9)}$$

where λ is the expected average number of DNA molecules per liposome, which is calculated as:

$$\lambda = \frac{\pi N_A C d^3}{6} \quad \text{(Eq. 10)}$$

where N_A is the Avogadro constant, C is the concentration of DNA and d is the average liposome diameter.

With 10 pM DNA and a liposome diameter of 4 μm (λ = 0.2), the probability of encapsulating one DNA molecule per liposome is 0.16, and the probability of encapsulating more than one molecule is 0.02. Experimental validation that encapsulation is mostly clonal at 10 pM and 50 pM was performed previously[47]. A more accurate approximation of λ values and encapsulation probabilities can be derived by accounting for the experimentally measured distribution of the liposome diameters (https://github.com/DanelonLab/ACS), but the main conclusions remain valid.

Visualizations

Radar charts, heatmaps and rarefaction plots were generated with R v4.4.1 in RStudio v2023.09.1+494. 3D sequence landscapes were generated with Python in Google Colaboratory. To generate the 3D landscapes, the variants with zero sequencing reads were assigned a log2 fold change identical to the lowest finite number in that dataset. Other graphs were made using Prism v10.0.3 and Illustrator v28.5.

Statistics

Bar height in bar plots is the mean value across replicates (the individual data points are often appended), and error bars indicate the standard deviation (SD). Box-whisker plots in **Fig. 1f, S5b and S9b** display the median (middle line), the 75/25 percentiles at the boxes and the 90/10 at the whiskers. Statistical tests (unpaired *t*-test, **Fig. 2d, S14**) were performed using GraphPad Prism v10.0.3 and were significant if P value (two-tailed) ≤ 0.05. R^2 is the coefficient of determination. The P value in linear regression (**Fig. 2f**) indicates whether the slope is significantly different from zero. The number of biological replicates n (i.e., separate liposome samples) is stated in the figure captions.

Data availability statement

The data are available in the main manuscript and Supporting Information. Raw data, processed data and the code for data processing underlying this article are available in the 4TU.ResearchData repository with DOI: 10.4121/1d798f6e-7071-4b7d-9d79-7425e4b9d247.v1.

Supporting information

Supplementary figures presenting data of nanopore sequencing, agarose gel electrophoresis, SDS-PAGE, fluorescence microscopy, FACS and qPCR. Supplementary tables listing plasmids, DNA oligos, DNA variant sequences and variant read counts.

Acknowledgements

We thank Ana Restrepo Sierra for her contribution to the conceptualization and design of the RBS libraries. We are grateful to Ilja Westerlaken for purifying the LactC2-mCherry proteins, Federico Ramirez Gomez for acquiring and analyzing microscopy images of the phospholipid-producing liposomes and Laura Sierra Heras for the adjusted PURE system composition. We also thank Liedewij Laan and Stefan Loonen for fruitful discussions, GeneFrontier for sponsoring our research, and Miguel de Vega and Alicia del Prado from the Centro de Biología Molecular Severo Ochoa (Madrid) for providing purified SSB and DSB proteins. MvdB, CD and NJC acknowledge financial support from the "BaSyC – Building a Synthetic Cell" Gravitation grant (024.003.019) of the Netherlands Ministry of Education, Culture and Science (OCW) and the Netherlands Organisation for Scientific Research (NWO). CD also acknowledges funding from Agence Nationale de la Recherche (ANR-22-CPJ2-0091-01). NJC also acknowledges funding by the NWO-SUMMIT grant "Evolving life from non-life (EVOLF)" (SUMMIT.1.004).

Author contributions

MvdB and CD conceived the project; CD and NJC supervised the project and acquired funding; MvdB and CD designed the experiments; MvdB performed experiments and data analysis and created visualizations; MvdB and CD wrote the paper with input from NJC.

References

1. Hunt, A. C. *et al.* Cell-free gene expression: methods and applications. *Chem. Rev.* **125**, 91–149 (2025).
2. De Capitani, J. & Mutschler, H. The long road to a synthetic self-replicating central dogma. *Biochemistry* **62**, 1221–1232 (2023).
3. Nieuwkoop, T., Finger-Bou, M., Oost, J. van der & Claassens, N. J. The ongoing quest to crack the genetic code for protein production. *Mol. Cell* **80**, 193–209 (2020).
4. Bernath, K. *et al.* In vitro compartmentalization by double emulsions: sorting and gene enrichment by fluorescence activated cell sorting. *Anal. Biochem.* **325**, 151–157 (2004).
5. Nishikawa, T., Sunami, T., Matsuura, T., Ichihashi, N. & Yomo, T. Construction of a gene screening system using giant unilamellar liposomes and a fluorescence-activated cell sorter. *Anal. Chem.* **84**, 5017–5024 (2012).

6. Fujii, S., Matsuura, T., Sunami, T., Kazuta, Y. & Yomo, T. In vitro evolution of α-hemolysin using a liposome display. *PNAS* **110**, 16796–16801 (2013).
7. Fujii, S. *et al.* Liposome display for in vitro selection and evolution of membrane proteins. *Nat. Protoc.* **9**, 1578–1591 (2014).
8. Uyeda, A. *et al.* Liposome-based in vitro evolution of aminoacyl-tRNA synthetase for enhanced pyrrolysine derivative incorporation. *ChemBioChem* **16**, 1797–1802 (2015).
9. Uyeda, A., Nakayama, S., Kato, Y., Watanabe, H. & Matsuura, T. Construction of an in vitro gene screening system of the E. coli EmrE transporter using liposome display. *Anal. Chem.* **88**, 12028–12035 (2016).
10. Murase, Y., Nakanishi, H., Tsuji, G., Sunami, T. & Ichihashi, N. In vitro evolution of unmodified 16S rRNA for simple ribosome reconstitution. *ACS Synth. Biol.* **7**, 576–583 (2018).
11. Seo, K., Hagino, K. & Ichihashi, N. Progresses in cell-free in vitro evolution. in *Cell-free Production: System Development* (eds. Lu, Y. & Jewett, M. C.) 121–140 (Springer Nature Switzerland, Cham, 2023). doi:10.1007/10_2023_219.
12. Gan, R. *et al.* High-throughput regulatory part prototyping and analysis by cell-free protein synthesis and droplet microfluidics. *ACS Synth. Biol.* **11**, 2108–2120 (2022).
13. Scheele, R. A. *et al.* Droplet-based screening of phosphate transfer catalysis reveals how epistasis shapes MAP kinase interactions with substrates. *Nat. Commun.* **13**, 844 (2022).
14. Jeschek, M., Gerngross, D. & Panke, S. Combinatorial pathway optimization for streamlined metabolic engineering. *Curr. Opin. Biotechnol.* **47**, 142–151 (2017).
15. Höllerer, S., Desczyk, C., Muro, R. F. & Jeschek, M. From sequence to function and back – High-throughput sequence-function mapping in synthetic biology. *Curr. Opin. Syst. Biol.* **37**, 100499 (2024).
16. Packer, M. S. & Liu, D. R. Methods for the directed evolution of proteins. *Nat. Rev. Genet.* **16**, 379–394 (2015).
17. Shimizu, Y. *et al.* Cell-free translation reconstituted with purified components. *Nat. Biotechnol.* **19**, 751–755 (2001).
18. 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).
19. Chizzolini, F. *et al.* Cell-free translation is more variable than transcription. *ACS Synth. Biol.* **6**, 638–647 (2017).
20. Chizzolini, F., Forlin, M., Cecchi, D. & Mansy, S. S. Gene position more strongly influences cell-free protein expression from operons than T7 transcriptional promoter strength. *ACS Synth. Biol.* **3**, 363–371 (2014).
21. van Nies, P. *et al.* Self-replication of DNA by its encoded proteins in liposome-based synthetic cells. *Nat. Commun.* **9**, 1583 (2018).
22. Blanco, L. *et al.* Effect of NH4+ ions on φ29 DNA-protein p3 replication: formation of a complex between the terminal protein and the DNA polymerase. *J. Virol.* **61**, 3983–3991 (1987).
23. Blanco, L., Lázaro, J. M., de Vega, M., Bonnin, A. & Salas, M. Terminal protein-primed DNA amplification. *PNAS* **91**, 12198–12202 (1994).
24. Esteban, J. A., Blanco, L., Villar, L. & Salas, M. In vitro evolution of terminal protein-containing genomes. *PNAS* **94**, 2921–2926 (1997).
25. Mencía, M., Gella, P., Camacho, A., de Vega, M. & Salas, M. Terminal protein-primed amplification of heterologous DNA with a minimal replication system based on phage φ29. *PNAS* **108**, 18655–18660 (2011).
26. Elías-Arnanz, M. & Salas, M. Bacteriophage φ29 DNA replication arrest caused by codirectional collisions with the transcription machinery. *The EMBO Journal* **16**, 5775–5783 (1997).
27. Mutalik, V. K. *et al.* Quantitative estimation of activity and quality for collections of functional genetic elements. *Nat. Methods* **10**, 347–353 (2013).
28. Doerr, A. *et al.* Modelling cell-free RNA and protein synthesis with minimal systems. *Phys. Biol.* **16**, 025001 (2019).
29. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in Escherichia coli. *Science* **324**, 255–258 (2009).

30. Nieuwkoop, T. *et al.* Revealing determinants of translation efficiency via whole-gene codon randomization and machine learning. *Nucleic Acids Res.* **51**, 2363–2376 (2023).
31. Höllerer, S. & Jeschek, M. Ultradeep characterisation of translational sequence determinants refutes rare-codon hypothesis and unveils quadruplet base pairing of initiator tRNA and transcript. *Nucleic Acids Res.* **51**, 2377–2396 (2023).
32. van den Brink, M., Althuis, T. Y., Danelon, C. & Claassens, N. J. MOSAIC: A highly efficient, one-step recombineering approach to plasmid editing and diversification. *ACS Synth. Biol.* **14**, 3880–3889 (2025).
33. Blanken, D., Foschepoth, D., Serrão, A. C. & Danelon, C. Genetically controlled membrane synthesis in liposomes. *Nat. Commun.* **11**, 4317 (2020).
34. Noga, M. J. *et al.* Posttranslational control of PlsB is sufficient to coordinate membrane synthesis with growth in Escherichia coli. *mBio* **11**, 10.1128/mbio.02703-19 (2020).
35. Kumar Saha, S., Satomi, N., Matsuzaki, H., Shibuya, I. & Matsumoto, K. A regulatory mechanism for the balanced synthesis of membrane phospholipid species in Escherichia coli. *Biosci. Biotechnol. Biochem.* **60**, 111–116 (1996).
36. Jeschek, M., Gerngross, D. & Panke, S. Rationally reduced libraries for combinatorial pathway optimization minimizing experimental effort. *Nat. Commun.* **7**, 11163 (2016).
37. Zelcbuch, L. *et al.* Spanning high-dimensional expression space using ribosome-binding site combinatorics. *Nucleic Acids Res.* **41**, e98 (2013).
38. Scott, A. *et al.* Cell-free phospholipid biosynthesis by gene-encoded enzymes reconstituted in liposomes. *PLOS ONE* **11**, e0163058 (2016).
39. Hagino, K. & Ichihashi, N. In vitro transcription/translation-coupled DNA replication through partial regeneration of 20 aminoacyl-tRNA synthetases. *ACS Synth. Biol.* **12**, 1252–1263 (2023).
40. Influence of 5'UTR and N-terminal ORF region on protein synthesis efficiency in PURExfrex® | Posters | Resources | PURExfrex® | GeneFrontier Corporation. *PURExfrex®* https://purefrex.genefrontier.com/resources/posters/P040_CFSC2023.html (2023).
41. Schaerli, Y. & Isalan, M. Building synthetic gene circuits from combinatorial libraries: screening and selection strategies. *Mol. BioSyst.* **9**, 1559–1567 (2013).
42. Pandi, A. *et al.* A versatile active learning workflow for optimization of genetic and metabolic networks. *Nat. Commun.* **13**, 3876 (2022).
43. Di Meo, T. *et al.* Functional selection in a population of synthetic cells with a minimal metabolism. 2025.08.25.672132 Preprint at https://doi.org/10.1101/2025.08.25.672132 (2025).
44. Abil, Z. & Danelon, C. Roadmap to building a cell: an evolutionary approach. *Front. Bioeng. Biotechnol.* **8**, (2020).
45. Abil, Z., Giaveri, S., Erb, T. J. & Rothschild, L. J. Integrating metabolism and evolution towards the realization of synthetic life. *Nat. Rev. Bioeng.* **3**, 9–10 (2025).
46. Blanken, D., van Nies, P. & Danelon, C. Quantitative imaging of gene-expressing liposomes reveals rare favorable phenotypes. *Phys. Biol.* **16**, 045002 (2019).
47. Abil, Z., Restrepo Sierra, A. M. & Danelon, C. Clonal amplification-enhanced gene expression in synthetic vesicles. *ACS Synth. Biol.* **12**, 1187–1203 (2023).
48. van den Brink, M., Stam, M., Claassens, N. J. & Danelon, C. Image-based phenotypic sorting of synthetic cells. *bioRxiv* 2025.09.25.678515 (2025) doi:10.1101/2025.09.25.678515.
49. Peterman, N. & Levine, E. Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations. *BMC Genomics* **17**, 206 (2016).
50. Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *PNAS* **116**, 8852–8858 (2019).
51. Yang, J. *et al.* Active learning-assisted directed evolution. *Nat. Commun.* **16**, 714 (2025).
52. Vornholt, T. *et al.* Enhanced sequence-activity mapping and evolution of artificial metalloenzymes by active learning. *ACS Cent. Sci.* **10**, 1357–1370 (2024).
53. Fu, G. K., Hu, J., Wang, P.-H. & Fodor, S. P. A. Counting individual DNA molecules by the stochastic attachment of diverse labels. *PNAS* **108**, 9026–9031 (2011).
54. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2012).

55. Schmitt, M. W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *PNAS* **109**, 14508–14513 (2012).
56. Nanopore sequencing accuracy. *Oxford Nanopore Technologies* https://nanoporetech.com/platform/accuracy.
57. Zurek, P. J., Knyphausen, P., Neufeld, K., Pushpanath, A. & Hollfelder, F. UMI-linked consensus sequencing enables phylogenetic analysis of directed evolution. *Nat. Commun.* **11**, 6023 (2020).
58. Komura, R., Aoki, W., Motone, K., Satomura, A. & Ueda, M. High-throughput evaluation of T7 promoter variants using biased randomization and DNA barcoding. *PLOS ONE* **13**, e0196905 (2018).
59. Carr, S. B., Beal, J. & Densmore, D. M. Reducing DNA context dependence in bacterial promoters. *PLOS ONE* **12**, e0176013 (2017).
60. Yoshida, A., Kohyama, S., Fujiwara, K., Nishikawa, S. & Doi, N. Regulation of spatiotemporal patterning in artificial cells by a defined protein expression system. *Chem. Sci.* **10**, 11064–11072 (2019).
61. Li, J. *et al.* Dissecting limiting factors of the Protein synthesis Using Recombinant Elements (PURE) system. *Translation* **5**, e1327006 (2017).
62. Kosuri, S. *et al.* Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *PNAS* **110**, 14024–14029 (2013).
63. Su, P.-R. *et al.* Microscopy-based single-cell proteomic profiling reveals heterogeneity in DNA damage response dynamics. *Cell Rep. Methods* **2**, 100237 (2022).
64. You, L. *et al.* Linking the genotypes and phenotypes of cancer cells in heterogenous populations via real-time optical tagging and image analysis. *Nat. Biomed. Eng.* **6**, 667–675 (2022).
65. Danelon, C. & Heras, L. S. Autocatalytic selection of gene functions in vitro. 2025.08.25.672052 Preprint at https://doi.org/10.1101/2025.08.25.672052 (2025).
66. Abil, Z. *et al.* Darwinian evolution of self-replicating DNA in a synthetic protocell. *Nat. Commun.* **15**, 9091 (2024).
67. Soengas, M. S., Gutiérrez, C. & Salas, M. Helix-destabilizing activity of φ29 single-stranded DNA binding protein: effect on the elongation rate during strand displacement DNA replication. *J. Mol. Biol.* **253**, 517–529 (1995).
68. The Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res.* **50**, W345–W351 (2022).
69. Blankenberg, D. *et al.* Manipulation of FASTQ data with Galaxy. *Bioinformatics* **26**, 1783–1785 (2010).
70. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
71. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).

Supporting Information

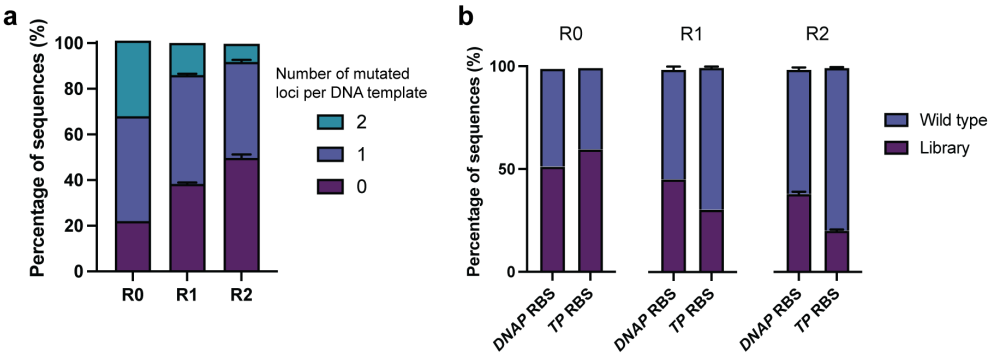


Fig. S1 Characterization of the orip2p3 RBS library before (R0) and after one (R1) or two (R2) rounds of IVTTR in liposomes. a Percentage of sequences with zero, one or two modified target loci. Percentage relative to the complete sequenced library after excluding reads with low-quality bases in the target sites. Number of biological replicates: n=1 (R0); n=3 (R1-2). Number of reads: 89,066 (R0); 72,884 (R1, replicate 1); 122,112 (R1, replicate 2); 72,501 (R1, replicate 3); 57,041 (R2, replicate 1); 76,140 (R2, replicate 2); 47,249 (R2, replicate 3). **b** Percentage of wild-type and modified sequences per target locus. The sums of percentages per bar for wild-type and intended library sequences add up to > 98%.

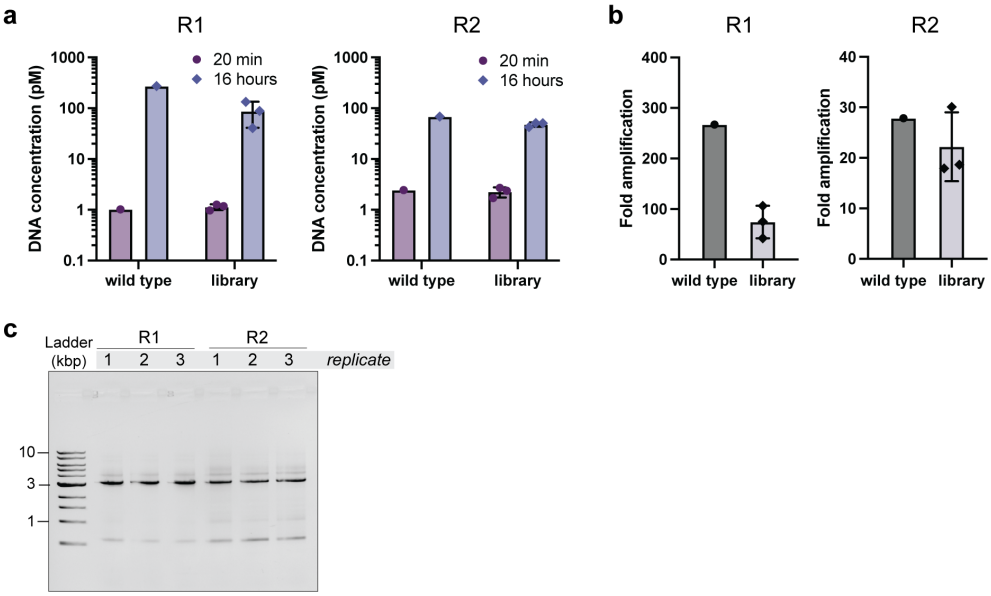


Fig. S2 Quantification and recovery of wild-type and library orip2p3 templates replicated in liposomes. a Absolute DNA quantification from in-liposome IVTTR reactions incubated for 16 hours or 20 min (negative control). **b** Replication fold from IVTTR reactions displayed in panel a. Replication fold was determined by the ratio of DNA concentration for reactions incubated for 16 h and 20 min. The wild type sample in round 2 (R2) was the wild-type template recovered from round 1 (R1). **c** PCR-amplified DNA templates (expected size: 3.2 kb) for library IVTTR after 16 h visualized by agarose gel electrophoresis. Purified PCR products were sequenced by nanopore sequencing.

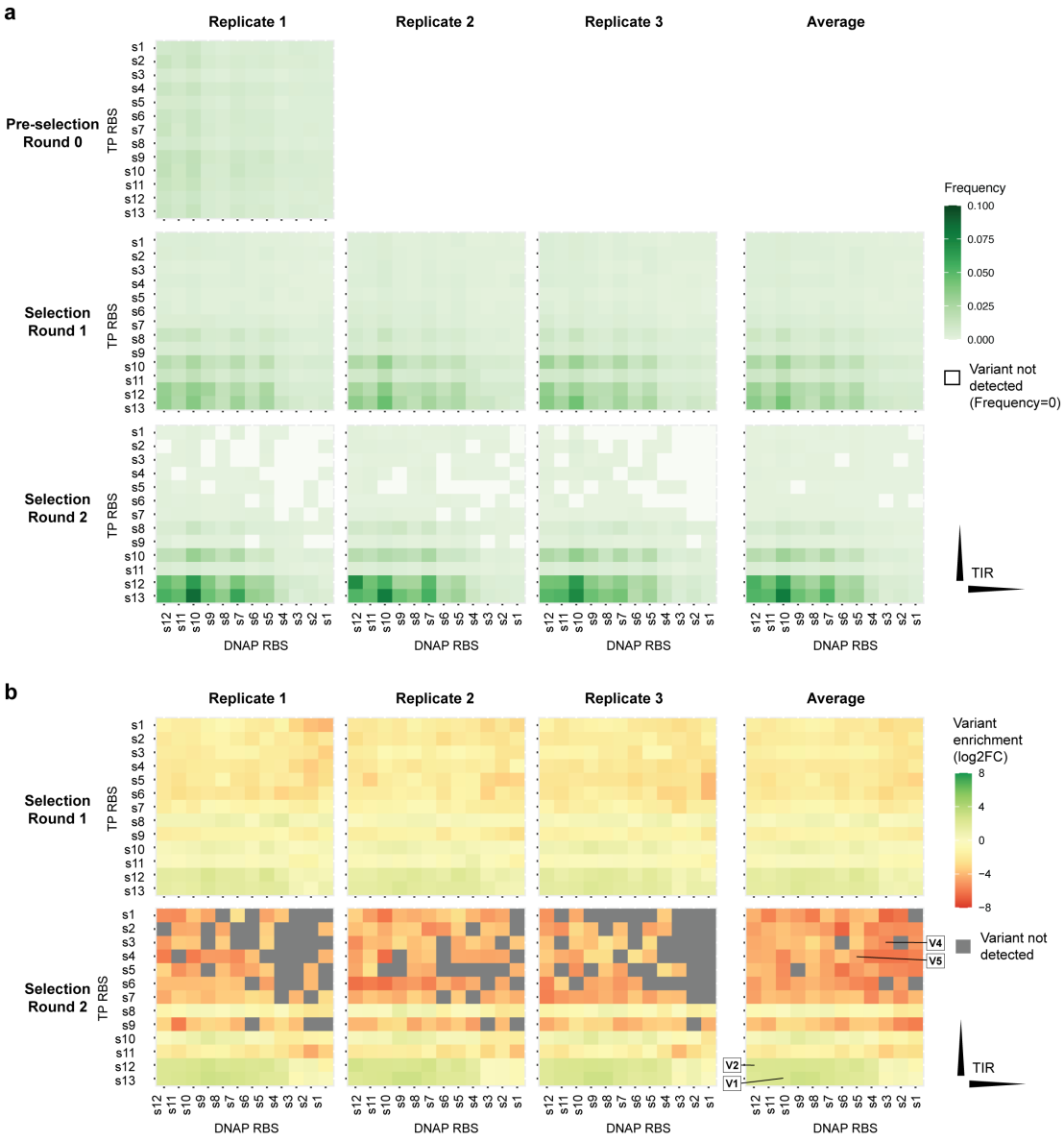


Fig. S3 Heatmaps displaying the frequency (a) or log2 fold change (b) of variants for each combination of *DNAP* and *TP* RBSs. a The frequency was calculated by dividing the number of reads by the total number of reads for the fraction of the library containing DNA variants with both RBSs mutated. Number of reads: 28,978 (R0); 10,436 (R1, replicate 1); 16,998 (R1, replicate 2); 10,140 (R1, replicate 3); 4,512 (R2, replicate 1); 6,060 (R2, replicate 2); 3,900 (R2, replicate 3). **b** To calculate the average log2FC, the average frequency was calculated, whereafter this value was divided by the initial frequency of the variant (round 0), and log2 transformed. Clonal DNA variants V1, V2, V4 and V5 are indicated in the right bottom heatmap. In panels a and b the variants are ranked from high to low predicted TIR.

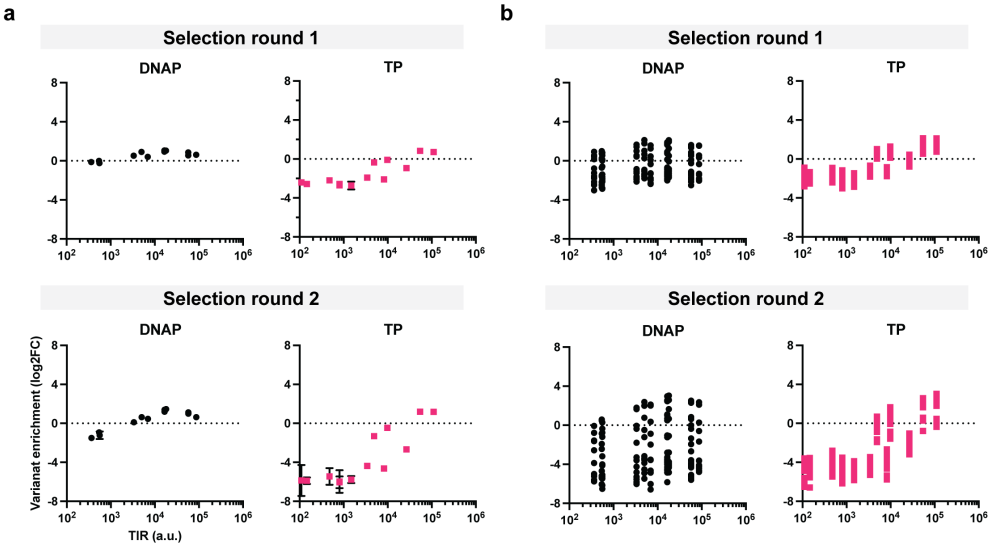


Fig. S4. Variant enrichment (log2FC) versus predicted TIR for rounds 1 and 2. a Individual data points for DNA variants that have one wild-type RBS and one mutated RBS. Error bars indicate standard deviation (n=3). **b** Data for the DNA variants that have two mutated RBSs (n=3). For each TIR value, there are multiple data points corresponding to the RBS variants of the other gene on the DNA template.

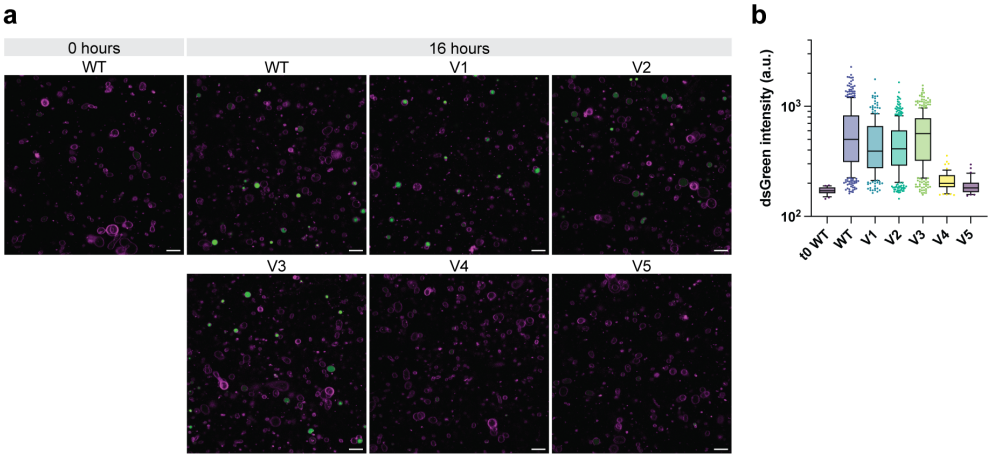


Fig. S5 Microscopy images of liposomes with replication of clonal DNA variants. a Additional microscopy images (partly identical to Fig. 1e). Magenta, Cy5-conjugated lipids; green, dsGreen; Scale bars: 10 µm. **b** Second biological replicate for dsGreen quantification in addition to replicate 1 shown in Fig. 1f. Number of dsGreen-positive liposomes: 23 (t0 WT), 364 (WT), 222 (V1), 356 (V2), 340 (V3), 55 (V4), 30 (V5).

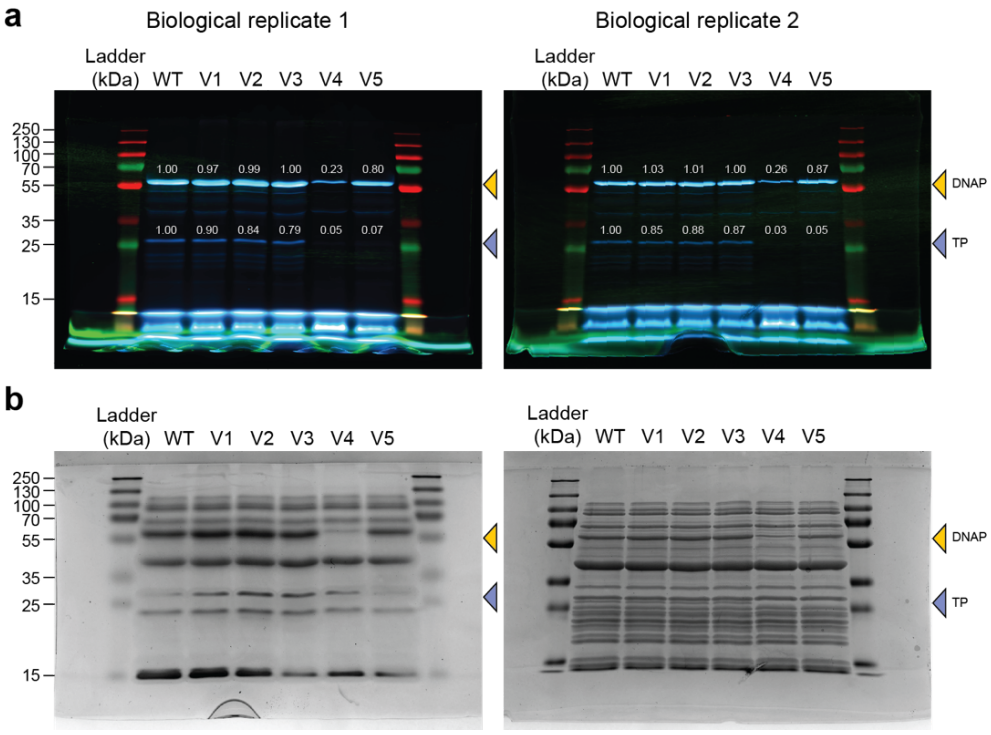


Fig. S6 Bulk PURE reactions with GreenLys protein labeling for orip2p3 DNA templates with different DNAP and TP RBSs. Biological replicate 1 was incubated for 8 hours at 30 °C. Biological replicate 2 was incubated for 2 hours at 30 °C. **a** Fluorescence imaging of SDS-PAGE gels for visualization of the bands corresponding to the proteins expressed by PURE system. Band intensities relative to the wild type (WT) are appended in the image. The gel for replicate 2 is also displayed in Fig. 1g. **b** Coomassie staining of the same gels shown in panel a to visualize the bands of all proteins. The staining of the gel for biological replicate 1 was most likely less sensitive due to the use of older Coomassie staining solution.

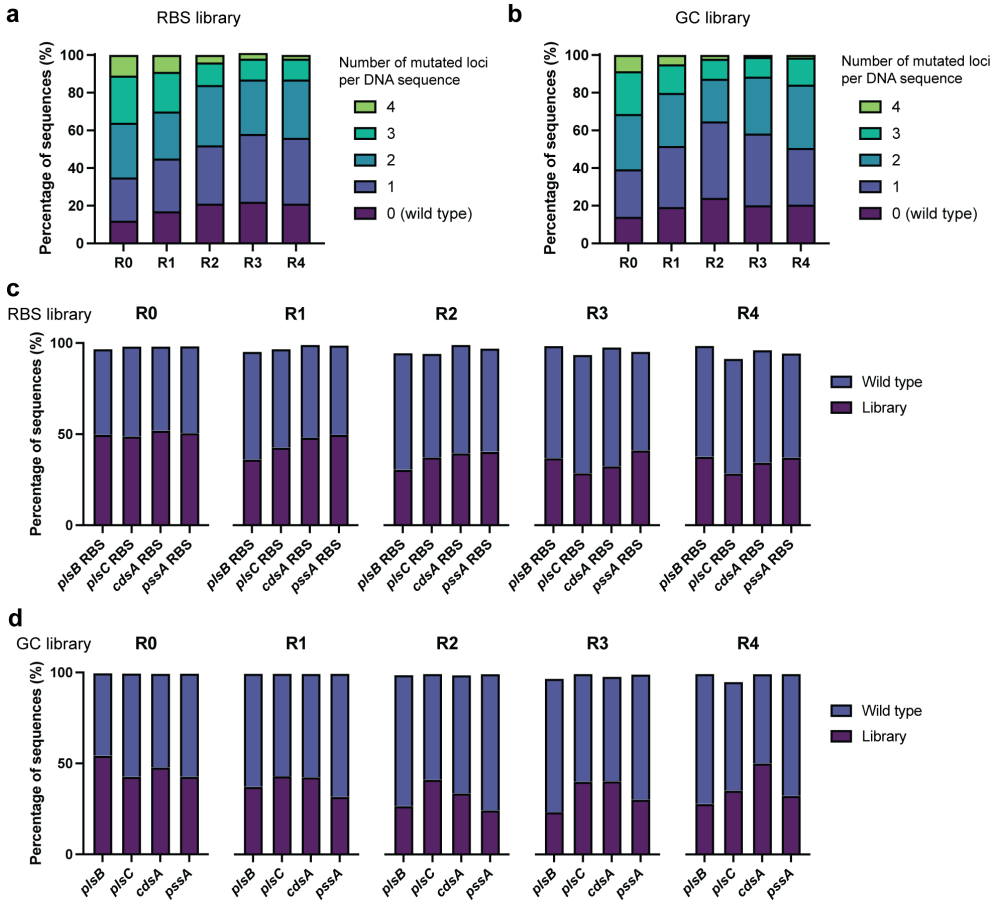


Fig. S7 Characterization of the phospholipid pathway RBS and GC libraries before (R0) and after one to four (R1-4) rounds of sorting of the liposomes with highest PS production by FACS. a-b Percentage of sequences with zero, one, two, three and four modified target loci for the RBS library (a) and GC library (b). Percentage relative to the complete sequenced library after excluding reads with low-quality bases in the target sites. Number of reads RBS library (a): 58,988 (R0); 77,727 (R1); 46,248 (R2); 31,776 (R3); 31,777 (R4). Number of reads GC library (b): 41,913 (R0); 116,996 (R1); 142,201 (R2); 118,839 (R3); 120,292 (R4). **c-d** Percentage of wild-type and modified sequences per target locus for the RBS library (c) and GC library (d).

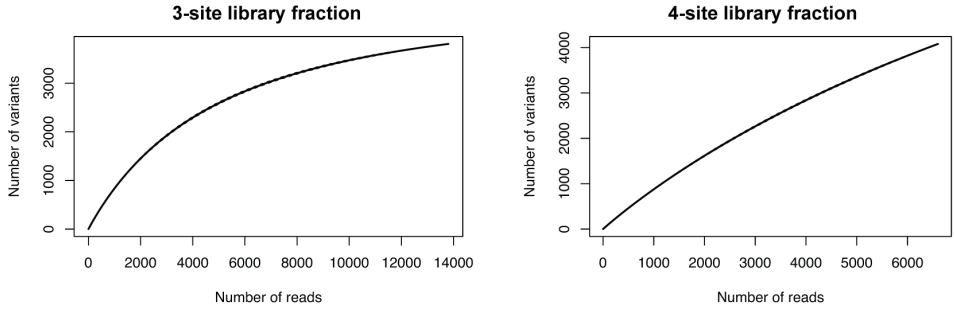


Fig. S8 Rarefaction plots for DNA variants with three or four RBSs mutated. Samples were an equimolar mix of four PCR-linearized plasmid library replicates. The graphs do not show a plateau, indicating that the library coverage is most likely larger than the number of variants detected in the nanopore sequencing dataset. For the DNA variants with three mutated RBSs, we detected 3,811 of the 4,518 variants (84%), based on 13,835 reads. For the DNA variants with four mutated RBSs, we detected 4,108 of the 11,583 variants (35%), based on 6,666 reads.

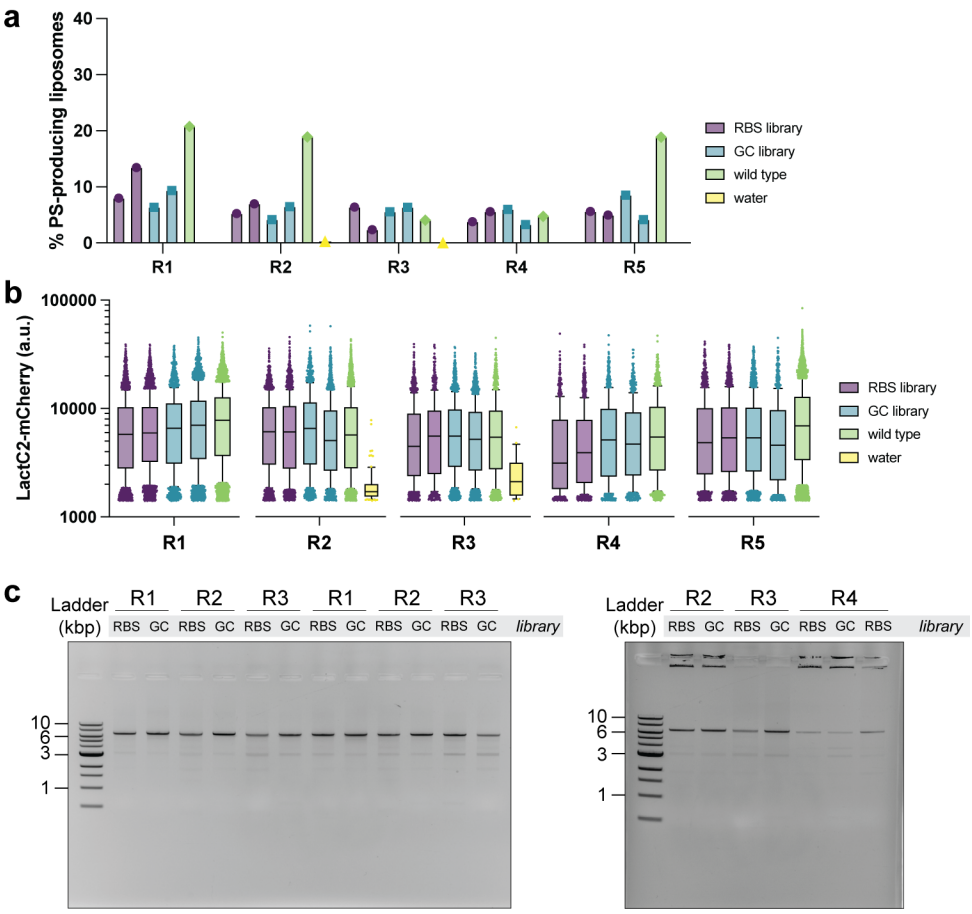


Fig. S9 FACS sorting and DNA recovery of liposomes producing PS. a Fraction of PS-producing liposomes measured by FACS. **b** LactC2-mCherry signal of the PS-producing liposomes measured by FACS. **a-b** For the RBS and GC libraries, two biological replicates were analyzed, indicated by identical bar colors. **c** Recovery of DNA from sorted liposomes by PCR (expected template length: 6.7 kb), visualized by agarose gel electrophoresis.

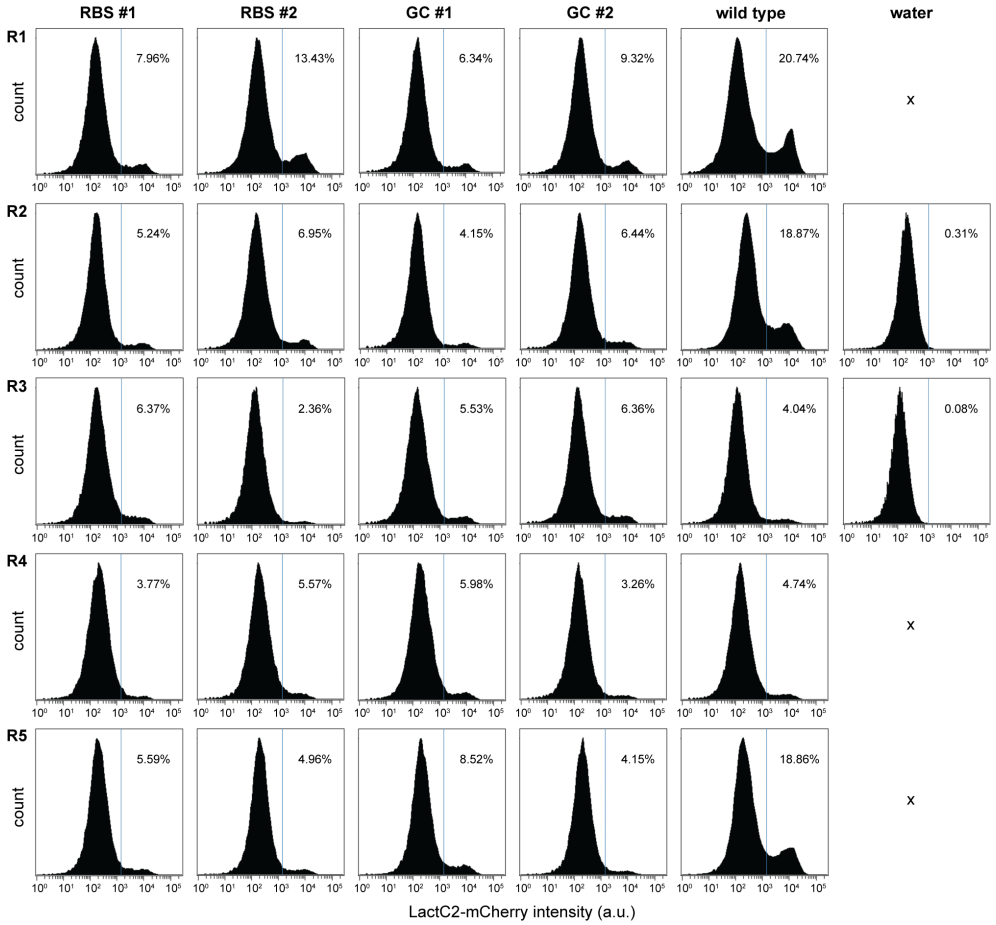


Fig. S10 Histograms displaying the distribution of LactC2-mCherry intensity of total liposome population. Liposomes exhibiting an intensity above the threshold, indicated with blue vertical line, are considered PS-producing liposomes. For the RBS and GC libraries, two biological replicates were analyzed, indicated by #1 and #2.

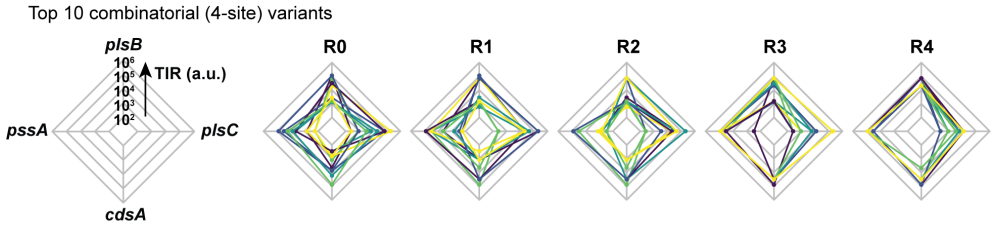


Fig. S11 Radar charts displaying the top 10 most frequent DNA variants for the phospholipid pathway RBS library for sorting round zero to four. The top 10 was derived from the sequencing reads containing a library variant for each RBS (i.e., 4-site library fraction).

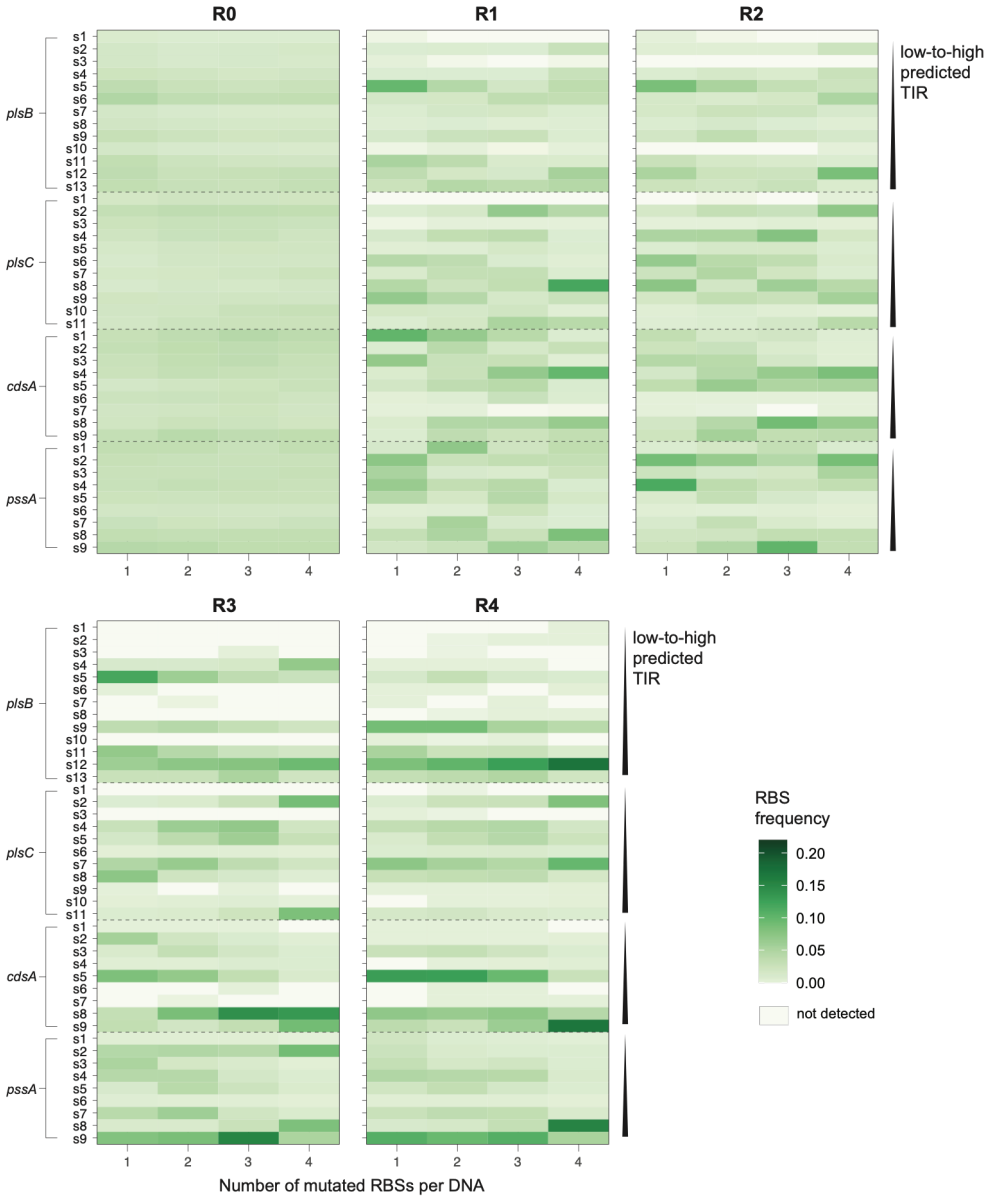


Fig. S12 Heatmaps showing the frequency of each variant of the PS synthesis RBS library for each round. The frequency was calculated relative to the library fractions with n number of mutated RBSs per DNA. The heatmap for R4 is also shown in **Fig. 2c**. Number of reads used to calculate the frequency are noted in **Table S12**.

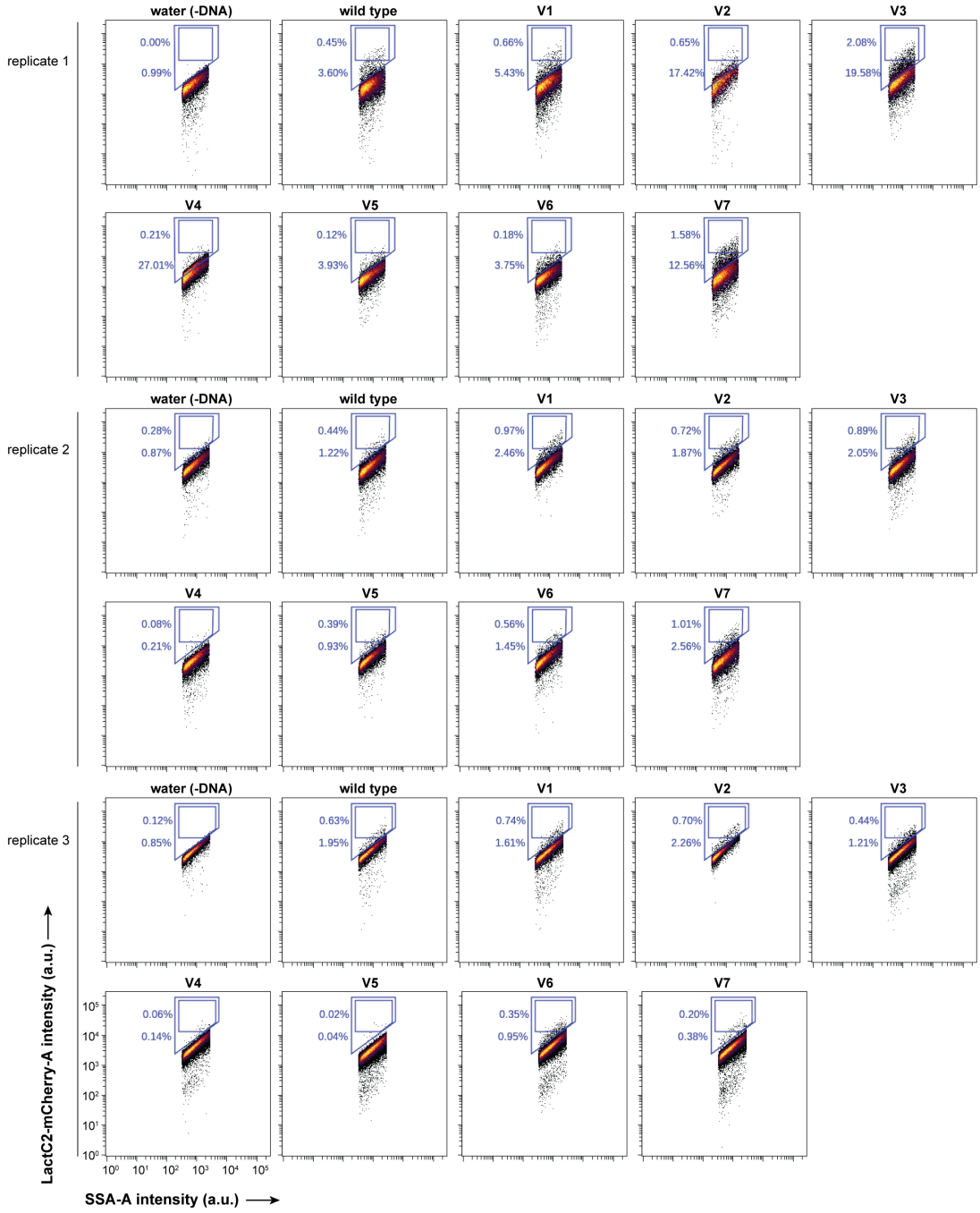


Fig. S13 Flow cytometry measurements of clonal DNA variants V1-7 for phospholipid synthesis. 20,000 events per sample (except plot for V2 of replicate 1 which shows 10,000 events).

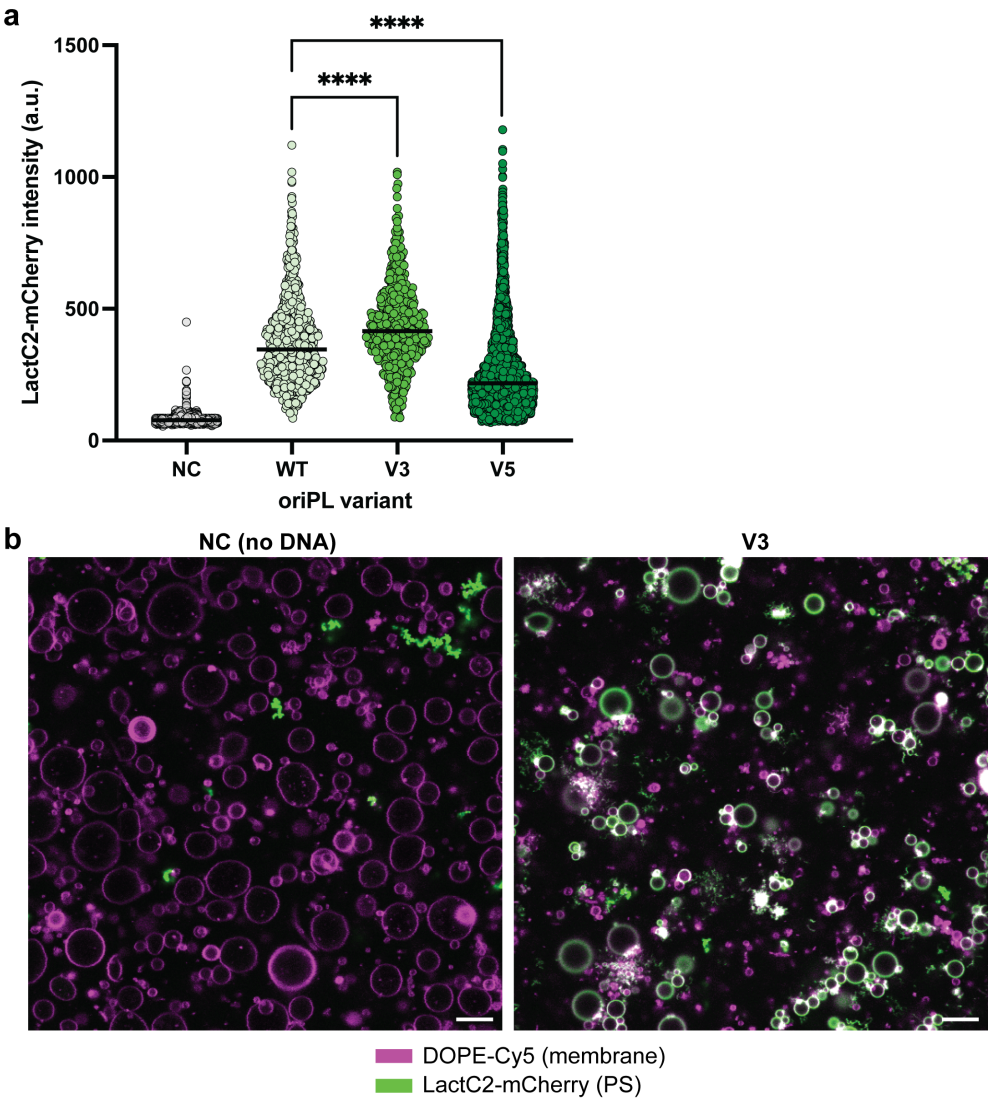


Fig. S14 Confocal image analysis of LactC2-mCherry fluorescence intensity with different DNA variants. 1 nM linear DNA was encapsulated in liposomes together with PURE system. The liposomes were incubated for 12 hours at 37 °C. The negative control (NC) contained MilliQ water instead of DNA. **a** Fluorescence intensity profiles from individual liposomes for different samples. Horizontal line indicates the median. Unpaired *t*-test: **** $P < 0.0001$. Analyzed liposomes per sample: 4,188 (NC), 859 (WT), 476 (V3) and 3,566 (V5). **b** Representative confocal fluorescence microscopy images for the negative control (NC) and V3. Scale bars: 10 µm.

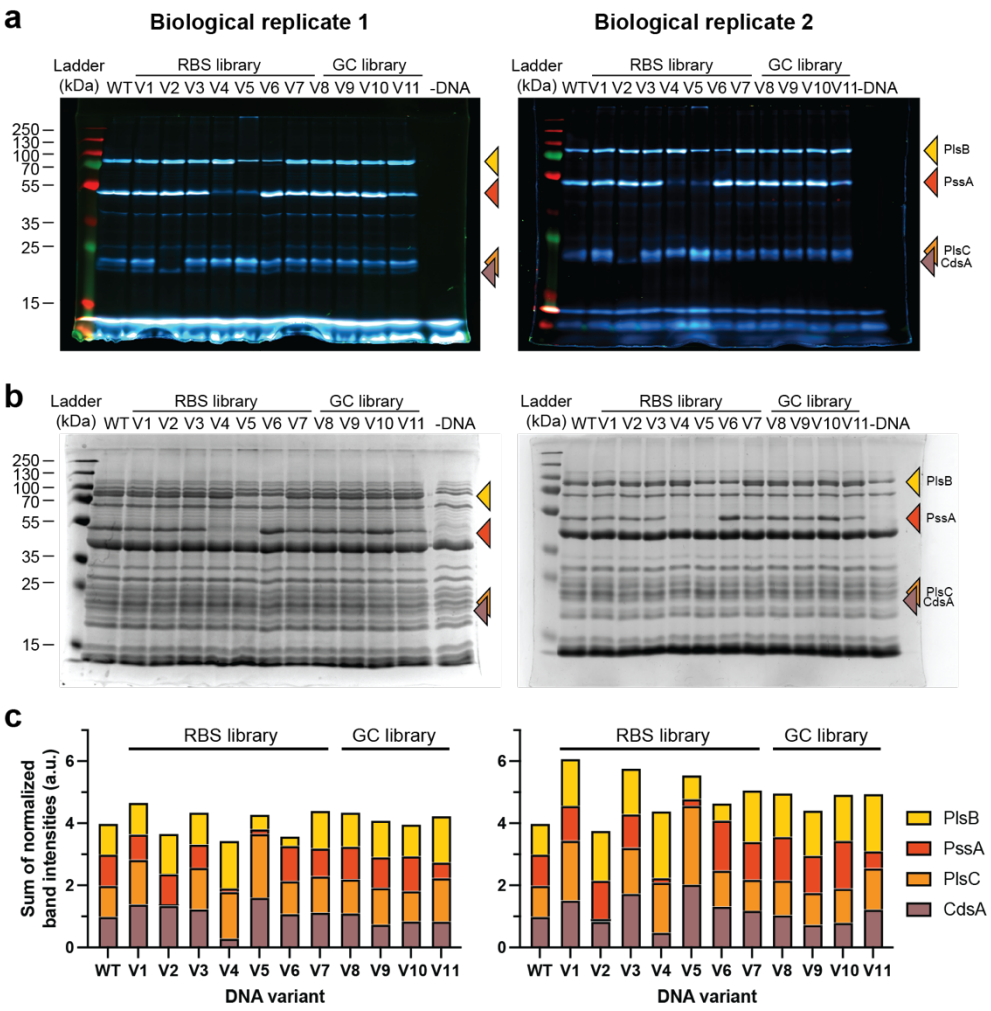


Fig. S15 Bulk PURE reactions with GreenLys protein labeling for the four phospholipid synthesis enzymes with different RBSs (clonal variants V1-7) or GC content at the start of the coding region (clonal variants V8-11). The variant sequences and predicted TIRs are listed in Tables S5 and S9. **a** Fluorescence imaging of SDS-PAGE gels for visualization of the bands corresponding to the proteins expressed by PURE system. The gel for replicate 1 is also displayed in Fig. 2e. **b** Coomassie staining of the same gels shown in panel a to visualize the bands of all proteins. **c** Quantification of the band intensities relative to the wild type (WT). Note that V9 and V11 have 4 (unintended) deletions upstream the *pssA* gene, and 1 deletion upstream the *plsB* gene, respectively, which may bias the interpretation of the effect of the targeted GC mutations for these variants.

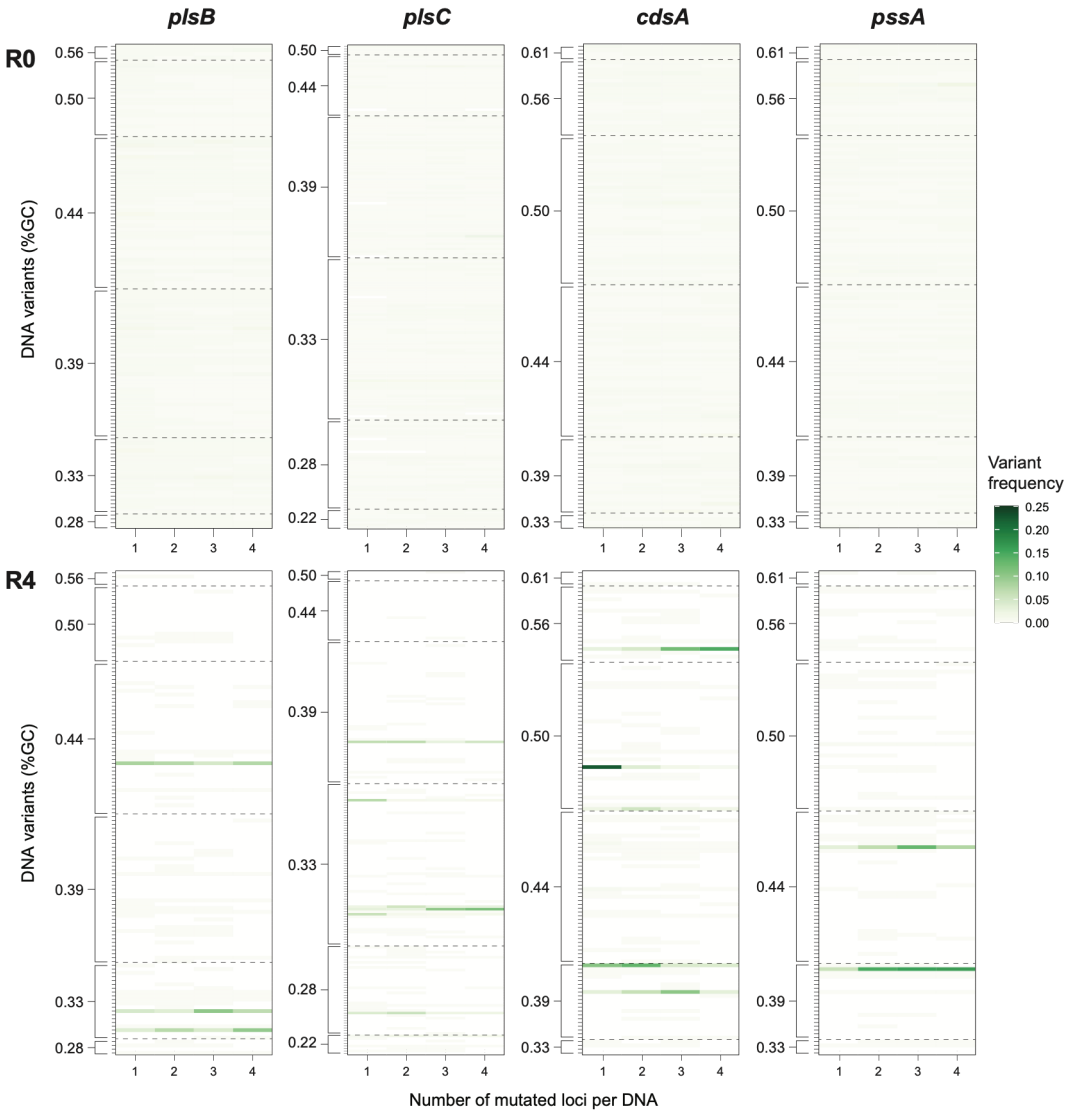


Fig. S16 Heatmaps showing the frequency of each DNA variant in the PS synthesis GC library for R0 (top row) and R4 (bottom row). The frequency was calculated relative to the library fractions with n number of mutated RBSs per DNA. Each tick mark on the y-axis corresponds to one DNA variant (127 DNA variants for *plsB*, *cdsA* and *pssA*, and 191 variants for *plsC*). The DNA variants are grouped based on the GC content of the first six codons of the genes, which ranged from 28-56% for *plsB*, 22-50% for *plsC* and 33-61% for *cdsA* and *pssA*. The GC content in the first six codons of wild-type *plsB*, *plsC*, *cdsA*, *pssA* are 39%, 28%, 50% and 50%, respectively. The number of reads used to calculate the frequency are noted in **Table S13**.

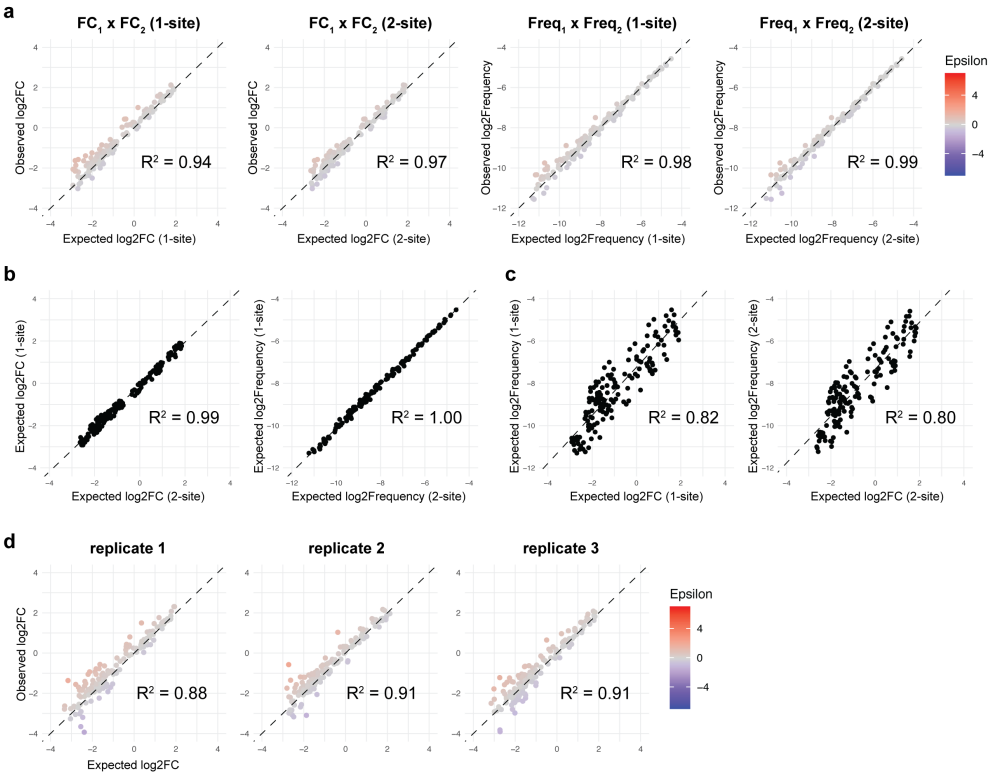


Fig. S17 Different methods for calculating the expected fitness of combinatorial DNA variants for the DNA replicator RBS library. Data of selection round 1 (R1). Dashed line is y = x (panels **a** and **d**) or the result of linear regression (panels **b-c**). All plots contain 156 data points. **a** Observed versus expected fitness plots, where fitness is log2FC or log2Frequency, and the expected fitness was calculated using the 1-site or 2-site library fraction data. Epsilon is the log2 observed fitness – log2 expected fitness. **b** The use of the 1-site or 2-site library fraction to calculate the expected fitness has similar outcomes. **c** Correlation between frequency and fold change for calculating the expected variant fitness. **a-c** Average of three biological replicates. **d** Observed versus expected log2FC of the three individual biological replicates. The expected log2FC was calculated using the 1-site library fraction data.

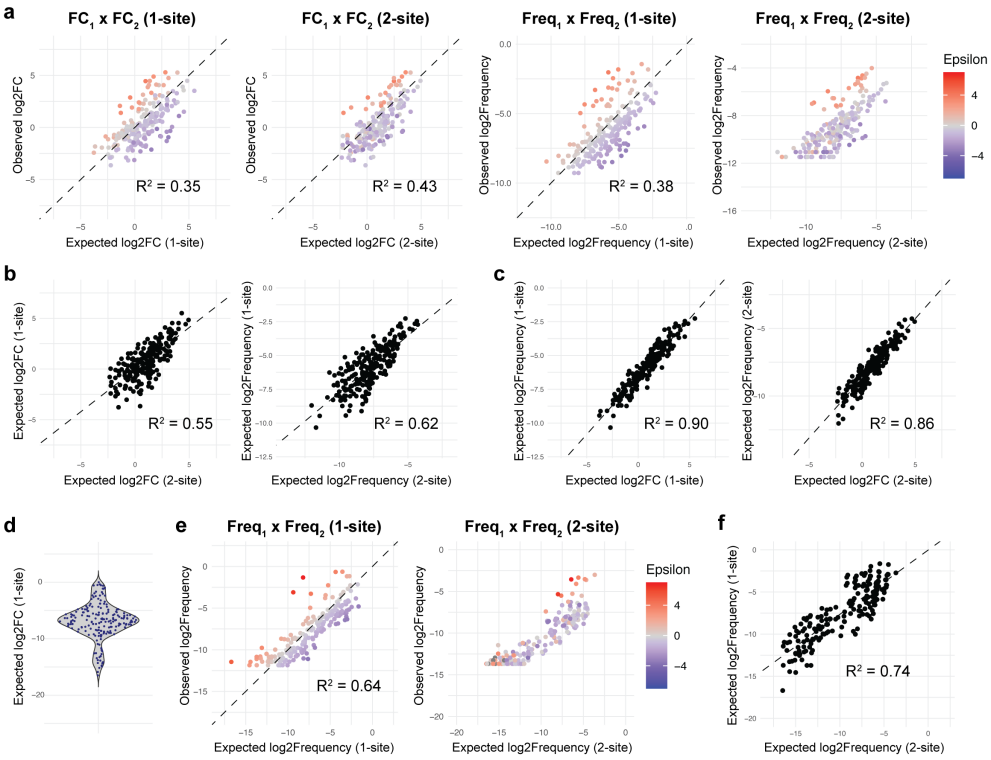


Fig. S18 Different methods for calculating the expected fitness of pairwise combinatorial DNA variants for the PS synthesis RBS and GC libraries. Data of sorting round 4 (R4). Dashed line is y = x (panels **a**, **e**) or the result of linear regression (panels **b-c**, **f**). **a-d** RBS library. **e-f** GC library. **a** Observed versus expected fitness plots for the RBS library, where fitness is log2FC or log2Frequency, and the expected fitness was calculated using the 1-site or 2-site library fraction data. Note that expected log2Frequency (2-site) does not scale with observed log2Frequency in the rightmost plot, because the full combinatorial library (with 4 mutagenic sites) is larger than the pairwise variants. The plotted Epsilon is identical for all plots to show that the high-epistasis variants (in red) are detected by all calculation methods (Epsilon is the observed log2Frequency – expected log2Frequency (1-site) for all plots). **b** Correlation between the expected fitness calculated using the 1-site versus 2-site fraction of the RBS library. **c** The use of FC or Frequency as a measure of the variant fitness leads to similar results. **a-c** All plots show 205 of the 656 pairwise combinatorial variants (DNA variants with less than three reads in the 2-site library fraction were excluded from the analysis). **d** Expected log2FC (calculated using the 1-site library fraction) for combinatorial DNA variants with less than three reads in the 2-site library fraction. The expected log2 fold change was low (mostly negative enrichment factors) compared to the sequenced combinatorial variants in panel a (left plot), explaining why the combinatorial DNA variants were not sequenced. 200 variants were plotted. **e** Observed versus expected fitness plots for the GC library. Note that expected log2Frequency (2-site) does not scale with observed log2Frequency in the right plot, because the full combinatorial library (with 4 mutagenic sites) is larger than the pairwise variants. The plotted Epsilon is identical for the two plots to show that the high-epistasis variants (in red) are detected by the two calculation methods (Epsilon is the observed log2Frequency –expected log2Frequency (1-site)). The left plot is identical to Fig. 4e. **f** Correlation between the expected fitness calculated using the 1-site versus 2-site fraction of the GC library. **e-f** 189 of the 121,158 pairwise combinatorial variants are plotted.

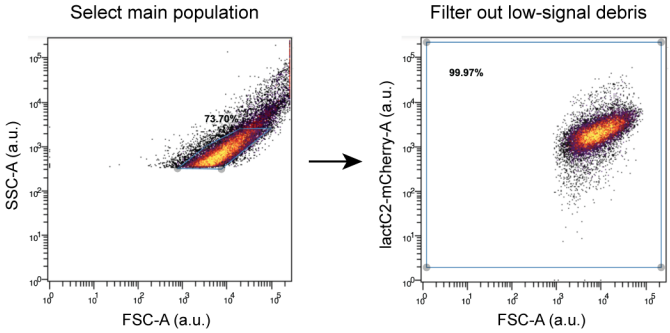


Fig. S19 Example of gating strategy for flow cytometry data. Only the first step (gating based on forward and side scatter light) was applied during sorting. The two filtering steps were applied during post-measurement analysis.

Table S1 List of RBS sequences for DNAP and TP in orip2p3 DNA template. Red color indicates degenerate nucleotides.

	DNAP RBS		TP RBS	
	Sequence	**Predicted TIR**	**Sequence**	**Predicted TIR**
Wild type	TTTAAGAAGGAGATATACAT	9,123	TTTAAGAAGGAGATATACTA	25,228
Library degenerate sequence	TTTAADAMGGAGSTATACAT	-	TTTAAGAAGSRGBTTGACTA	-
s1	TTTAAAACGGAGCTATACAT	362	TTTAAGAAGCGGCTTGACTA	110
s2	TTTAAGACGGAGCTATACAT	543	TTTAAGAAGCGGTTTGACTA	146
s3	TTTAATACGGAGCTATACAT	555	TTTAAGAAGCGGGTTGACTA	477
s4	TTTAAAAAGGAGCTATACAT	3,329	TTTAAGAAGCAGTTTGACTA	801
s5	TTTAAGAAGGAGCTATACAT	4,992	TTTAAGAAGCAGCTTGACTA	801
s6	TTTAATAAGGAGCTATACAT	6,933	TTTAAGAAGCAGGTTGACTA	1,476
s7	TTTAAGACGGAGGTATACAT	16,525	TTTAAGAAGGGGCTTGACTA	3,425
s8	TTTAAAACGGAGGTATACAT	16,525	TTTAAGAAGGAGCTTGACTA	4,909
s9	TTTAATACGGAGGTATACAT	17,920	TTTAAGAAGGGGTTTGACTA	8,224
s10	TTTAAGAAGGAGGTATACAT	57,485	TTTAAGAAGGAGTTTGACTA	9,818
s11	TTTAAAAAGGAGGTATACAT	57,485	TTTAAGAAGGGGGTTGACTA	26,545
s12	TTTAATAAGGAGGTATACAT	86,579	TTTAAGAAGGAGGTTGACTA	54,294
s13	-	-	TTTAAGAAGGAGGTTTACTA	110,313

Table S2 Clonal DNA variants from orip2p3 RBS library. The *plsC* RBS of V3 (with predicted TIR: 938 a.u.) was not a variant from the theoretical library, but likely originated from a mutation during MOSAIC or PCR.

	DNAP RBS		**TP RBS**	
	Sequence	**Predicted TIR**	**Sequence**	**Predicted TIR**
WT	WT	9,123	WT	25,228
V1	s10	57,485	s13	110,313
V2	s12	86,579	s12	54,294
V3	TTTAATAAGGAGATATACAT	14,117	s13	110,313
V4	s3	555	s3	477
V5	s5	4,992	s4	801

Table S3 List of RBS sequences for *plsB*, *plsC*, *cdsA* and *pssA* in oriPL DNA template. Red color indicates degenerate nucleotides.

	***plsB* RBS**		***plsC* RBS**	
	Sequence	**Predicted TIR**	**Sequence**	**Predicted TIR**
Wild type	TTTAAGAAGGAGATATACAT	5,896	TTTAAGAAGGAGATATACAT	4,947
Library degenerate sequence	TTTAAGGAGGKAAAMBCCAT	-	TTTAAGGVGAATADATACAT	-
s1	TTTAAGGAGGGAAACCCCAT	246	TTTAAGGCGAATAGATACAT	144
s2	TTTAAGGAGGGAAACTCCAT	608	TTTAAGGCGAATATATACAT	249
s3	TTTAAGGAGGGAAACGCCAT	981	TTTAAGGCGAATAAATACAT	878
s4	TTTAAGGAGGGAAAACCCAT	1,309	TTTAAGGGGAATAGATACAT	3,327
s5	TTTAAGGAGGGAAAAGCCAT	1,561	TTTAAGGGGAATATATACAT	5,725
s6	TTTAAGGAGGGAAAATCCAT	2,803	TTTAAGGAGAATAGATACAT	7,182
s7	TTTAAGGAGGTAAACCCCAT	6,346	TTTAAGGAGAATATATACAT	12,360
s8	TTTAAGGAGGTAAACTCCAT	15,661	TTTAAGGGGAATAAATACAT	20,220
s9	TTTAAGGAGGTAAAACCCAT	21,409	TTTAAGGAGAATAAATACAT	43,653
s10	TTTAAGGAGGTAAACGCCAT	25,289	TTTAAGGAGGTAAGATACAT	68,987
s11	TTTAAGGAGGTAAAAGCCAT	31,839	TTTAAGGAGGTAATATACAT	201,909
s12	TTTAAGGAGGTAAAATCCAT	72,247		
s13	TTTAAGGAGGTAATATACAT	112,986		

	***cdsA* and *pssA* RBSs**	
	Sequence	**Predicted TIR**
Wild type	TTTAAGAAGGAGATATACAT	5,198
Library degenerate sequence	TTTAAGARKGAGKTATACAT	-
s1	TTTAAGAATGAGTTATACAT	174
s2	TTTAAGAGTGAGTTATACAT	289
s3	TTTAAGAGGGAGTTATACAT	659
s4	TTTAAGAAGGAGTTATACAT	1,231
s5	TTTAAGAATGAGGTATACAT	4,501
s6	TTTAAGAGTGAGGTATACAT	7,484
s7	TTTAAGAGGGAGGTATACAT	17,054
s8	TTTAAGAAGGAGGTATACAT	31,879
s9	GATAAGGAGGTGATATACAT	78,824

Table S4 Calculation of FACS screening coverage for RBS library. Calculation based on 4,500,000 screened events and 13% PS-producing liposomes (for wild type). Calculation example (row 2): Per-variant screening coverage = 4,500,000 (number of events) * 0.23 (frequency of this library fraction of total library) * 0.13 (% PS-producing liposomes) / 42 (number of variants) = 3,204.

Library fraction	Number of mutated RBSs in this library fraction	Number of variants	Fraction of total library	Per-variant screening coverage
1	0 (WT)	1	0.12	70,200
2	1	42	0.23	3,204
3	2	656	0.29	259
4	3	4,518	0.25	32
5	4	11,583	0.11	6

Table S5 Clonal DNA variants from oriPL RBS library.

	***plsB* RBS**		***plsC* RBS**		***cdsA* RBS**		***pssA* RBS**	
	Sequence	**Predicted TIR**	**Sequence**	**Predicted TIR**	**Sequence**	**Predicted TIR**	**Sequence**	**Predicted TIR**
WT	WT	5,896	WT	4,947	WT	5,198	WT	5,198
V1	s12	72,247	s7	12,360	s9	78,824	s9	78,824
V2	s12	72,247	TTTAAGGGGAATAGATATAT	938	s9	78,824	s9	78,824
V3	s12	72,247	s11	201,909	s9	78,824	s9	78,824
V4	WT	5,896	WT	4,947	s1	174	s1	174
V5	s4	1,309	s7	12,360	WT	5,198	s1	174
V6	s4	1,309	WT	4,947	WT	5,198	WT	5,198
V7	WT	5,896	WT	4,947	WT	5,198	s9	78,824

Table S6 Top 10 enriched variants in the 4-site fraction of the RBS oriPL library after round 4.

***plsB* RBS sequence**	***plsC* RBS sequence**	***cdsA* RBS sequence**	***pssA* RBS sequence**	***plsB* RBS TIR**	***plsC* RBS TIR**	***cdsA* RBS TIR**	***pssA* RBS TIR**	**Number of reads**
TTTAAGGAG GTAAAATCC AT	TTTAAGGAG AATATATAC AT	GATAAGGA GGTGATATA CAT	TTTAAGAAG GAGGTATAC AT	72,247	12,360	78,824	31,879	181
TTTAAGGAG GTAAAATCC AT	TTTAAGGCG AATATATAC AT	GATAAGGA GGTGATATA CAT	TTTAAGAAG GAGGTATAC AT	72,247	249	78,824	31,879	172
TTTAAGGAG GTAAAATCC AT	TTTAAGGGG AATAGATAT AT	TTTAAGAAG GAGGTATAC AT	GATAAGGA GGTGATATA CAT	72,247	938	31,879	78,824	162
TTTAAGGAG GTAAAACCC AT	TTTAAGGGG AATAGATAT AT	TTTAAGAAG GAGGTATAC AT	GATAAGGA GGTGATATA CAT	21,409	938	31,879	78,824	18
TTTAAGGAG GTAAAATCC AT	TTTAAGGGG AATAGATAC AT	TTTGAGAAG GAGATATAC AT	GATAAGGA GGTGATATA CAT	72,247	3,327	4,794	78,824	17
TTTAAGGAG GTAAAATCC AT	TTTAAGGGG AATATATAC AT	TTTAAGAAG GAGGTATAC AT	GATAAGGA GGTGATATA CAT	72,247	5,725	31,879	78,824	16
TTTAAGGAG GTAAAACCC AT	TTTAAGGCG AATATATAC AT	GATAAGGA GGTGATATA CAT	TTTAAGAAG GAGGTATAC AT	21,409	249	78,824	31,879	15
TTTAAGGAG GTAAAACCC AT	TTTAAGGGG AATATATAC AT	TTTAAGAAG GAGGTATAC AT	GATAAGGA GGTGATATA CAT	21,409	5,725	31,879	78,824	14
TTTAAGGAG GTAAAACCC AT	TTTAAGGGG AATAGATAC AT	TTTGAGAAG GAGATATAC AT	GATAAGGA GGTGATATA CAT	21,409	3,327	4,794	78,824	13
TTTAAGGAG GTAAAACCC AT	TTTAAGGAG AATATATAC AT	TTTAAGAAG GAGGTATAC AT	GATAAGGA GGTGATATA CAT	21,409	12,360	31,879	78,824	11

Table S7 Calculation of FACS screening coverage for the theoretical GC library. Calculation based on 4,500,000 screened events and 13% PS-producing liposomes (for wild type). With a MOSAIC editing efficiency of ~50% per target locus, we obtained a library that consisted for 9% of variants with all target loci mutated, and the remainder of the library consisted of library fractions where zero, one, two or three loci were mutated (**Fig. S11a**). Using nanopore sequencing, we detected 565, 9,236, 7,785 and 2,922 library variants with one, two, three and four mutated target sites, respectively, in the initial library. This corresponded to 99%, 8%, 0.1% and 0.001% of all theoretical library variants with one, two, three and four mutated RBSs, respectively. Potentially, more of the theoretical library is covered but not detected with the current sequencing depth.

Screening coverage for DNA variants:				
Library fraction	**Number of mutated loci**	**Number of DNA variants**	**Fraction of total library**	**Per-variant screening coverage**
1	0 (WT)	1	0.14	81,900
2	1	572	0.25	256
3	2	121,158	0.29	1
4	3	11,290,300	0.23	0.01
5	4	391,241,153	0.09	0.0001
Screening coverage for %GC variants:				
Library fraction	**Number of mutated loci**	**Number of %GC variants**	**Fraction of total library**	**Per-variant screening coverage**
1	0 (WT)	1	0.14	81,900
2	1	6	0.25	24,375
3	2	36	0.29	4,713
4	3	216	0.23	623
5	4	1,296	0.09	41

Table S8 Most enriched variants (variant frequency > 0.02) of the GC oriPL library after round 4.

Gene	**DNA sequence**	**Predicted TIR (a.u.)**
plsB	CTTTTGTTATCCA	5,870
plsB	ATTTTGCTATCCA	20,695
plsB	CTTTTGTTACCCC	5,870
plsC	ATATATCTTTCGC	8,605
plsC	TTATATATTCCGG	34,414
plsC	CTACATTTTCCGA	21,844
plsC	ATATATCTTCCGA	2,975
plsC	ATATATCTTTCGA	4,859
plsC	ATATATCTTTCGG	6,809
cdsA	CAGCATTACCGT	13,195
cdsA	CAGCATGACCGT	10,073
cdsA	TAGTATTACTGG	13,618
cdsA	CAGCATTACCGT	13,195
cdsA	CAGCATGACCGT	10,073
pssA	TAGTATTACGGT	5,611
pssA	TAGCATTACAGG	14,700

Table S9 Clonal DNA variants from the oriPL GC library.

	***plsB* RBS**			***plsC* RBS**		
	Sequence	**%GC**	**Predicted TIR**	**Sequence**	**%GC**	**Predicted TIR**
WT	WT (TTTCTGCTATCCT)	0.38	5,896	WT (ATATATCTTTCGT)	0.23	4,947
V8	ATTTTGCTATCCA	0.31	20,695	ATATATCTTTCGC	0.31	8,605
V9	ATTTTGCTATCCA	0.31	20,695	ATATATCTTTCGC	0.31	8,605
V10	ATTTTGCTATCCA	0.31	20,695	ATATATCTTTCGC	0.31	8,605
V11[$]	ATTTTGCTATCCA	0.31	12,389[$]	ATATATCTTTCGC	0.31	8,605

	***cdsA* RBS**			***pssA* RBS**		
	Sequence	**%GC**	**Predicted TIR**	**Sequence**	**%GC**	**Predicted TIR**
WT	WT (TAGCATGACTGG)	0.50	5,198	WT (TAGCATGACTGG)	0.50	5,198
V8	WT	0.50	5,198	WT	0.50	5,198
V9	CAGCATGACCGT	0.58	10,073	WT*	0.50	9,289*
V10	TAGTATTACTGT	0.25	13,618	TAGTATTACTGT	0.25	13,618
V11	CAGCATGACCGT	0.58	10,073	CAGCATGACCGT	0.58	10,073

[$] V11 has a deletion at the 22[nd] nucleotide upstream the start codon of the *plsB* gene.
* V9 has 4 deletions 17-20 nucleotides upstream the start codon of the *pssA* gene.

Table S10 List of plasmids.

Plasmid name	**Description**
G555	Expression of PlsB, PlsC, CdsA and PssA under the control of T7 promoters. The 4-gene insert is flanked by Phi29 replication origins. Addgene #216483.
G435	Expression of DNAP and TP under the control of T7 promoters. The 2-gene insert is flanked by Phi29 replication origins.
pORTMAGE-Ec1	Enables high-efficiency plasmid and genomic recombineering in *E. coli*. Expresses single-stranded DNA annealing protein (SSAP) CspRecT and the dominant negative *E. coli* MutL mutant (EcMutL^{E32K}), controlled by XylS-Pm expression system. Addgene #138474.

Table S11 List of DNA oligos. '*' indicates phosphorothioate bond. Red color indicates degenerate nucleotides.

Primer name	**Sequence (5' to 3')**	**Description**
DNAP DEG	C*A*CAACGGTTTCCCTCTAGAAA TAATTTTGTTTAACTTTAADAMGG AGSTATACATATGCCGCGTAAAA TGTACAGCTGCGATTTTGAAAC	G435 RBS library Degenerate MOSAIC oligo for mutating the *DNAP* RBS in plasmid G435.
TP DEG	G*A*CACCAGAGGGTTTACATGTT TATTTGTTTAACTTTAAGAAGSRG BTTGACTAATGGCACGCAGCCCG CGCATCCGCATCAAAGATAACG	G435 RBS library Degenerate MOSAIC oligo for mutating the *TP* RBS in plasmid G435.

Table S11 continued

TP_1 custom	G*A*CACCAGAGGGTTTACATGTT TATTTGTTTAACTTTAAGAAGGAG GTTTACTAATGGCACGCAGCCCG CGCATCCGCATCAAAGATAACG	G435 RBS library Additional MOSAIC oligo for mutating the *TP* RBS in plasmid G435. This oligo was mixed with the degenerate *TP* oligo set before addition to the electroporation sample.
1585 ChD	C*A*CAACGGTTTCCCTCTAGAAA TAATTTTGTTTAACTTTAAGAAGG AGGTATACATATGCCGCGTAAAA TGTACAGCTGCGATTTTGAAAC	G435 RBS variant (s10) targeting the *DNAP* RBS (MOSAIC oligo).
1586 ChD	C*A*CAACGGTTTCCCTCTAGAAA TAATTTTGTTTAACTTTAATAAGG AGGTATACATATGCCGCGTAAAA TGTACAGCTGCGATTTTGAAAC	G435 RBS variant (s12) targeting the *DNAP* RBS (MOSAIC oligo).
1587 ChD	C*A*CAACGGTTTCCCTCTAGAAA TAATTTTGTTTAACTTTAATACGG AGCTATACATATGCCGCGTAAAA TGTACAGCTGCGATTTTGAAAC	G435 RBS variant (s3) targeting the *DNAP* RBS (MOSAIC oligo).
1588 ChD	C*A*CAACGGTTTCCCTCTAGAAA TAATTTTGTTTAACTTTAAGAAGG AGCTATACATATGCCGCGTAAAA TGTACAGCTGCGATTTTGAAAC	G435 RBS variant (s5) targeting the *DNAP* RBS (MOSAIC oligo).
1589 ChD	C*A*CAACGGTTTCCCTCTAGAAA TAATTTTGTTTAACTTTAATAAGG AGATATACATATGCCGCGTAAAA TGTACAGCTGCGATTTTGAAAC	G435 RBS variant targeting the *DNAP* RBS (MOSAIC oligo).
1590 ChD	G*A*CACCAGAGGGTTTACATGTT TATTTGTTTAACTTTAAGAAGGAG GTTGACTAATGGCACGCAGCCCG CGCATCCGCATCAAAGATAACG	G435 RBS variant (s12) targeting the *TP* RBS (MOSAIC oligo).
1591 ChD	G*A*CACCAGAGGGTTTACATGTT TATTTGTTTAACTTTAAGAAGCGG GTTGACTAATGGCACGCAGCCCG CGCATCCGCATCAAAGATAACG	G435 RBS variant (s3) targeting the *TP* RBS.
1592 ChD	G*A*CACCAGAGGGTTTACATGTT TATTTGTTTAACTTTAAGAAGCAG TTTGACTAATGGCACGCAGCCCG CGCATCCGCATCAAAGATAACG	G435 RBS variant (s4) targeting the *TP* RBS (MOSAIC oligo).
PlsB DEG	C*A*ACGGTTTCCCTCTAGAAATA ATTTTGTTTAACTTTAAGGAGGKA AAMBCCATATGACTTTCTGCTAT CCTTGCCGCGCATTTGCATTA	G555 RBS library Degenerate MOSAIC oligo for mutating the *plsB* RBS in plasmid G555.
PlsC DEG	C*A*ACGGTTTCCCTCTAGAAATA ATTTTGTTTAACTTTAAGGVGAAT ADATACATATGCTATATATCTTTC GTCTTATTATTACCGTGATT	G555 RBS library Degenerate MOSAIC oligo for mutating the *plsC* RBS in plasmid G555.
PssA_CdsA DEG	A*T*AACAATTCCCCTCTAGAAAT AATTTTGTTTAACTTTAAGARKGA GKTATACATATGGCTAGCATGAC TGGTGGACAGCAAATGGGTCG	G555 RBS library Degenerate MOSAIC oligo for mutating the *cdsA and pssA* RBSs in plasmid G555.

Table S11 continued

PlsB custom	C*A*ACGGTTTCCCTCTAGAAATA ATTTTGTTTAACTTTAAGGAGGTA ATATACATATGACTTTCTGCTATC CTTGCCGCGCATTTGCATTA	G555 RBS library Additional MOSAIC oligo for mutating the *plsB* RBS in plasmid G555. This oligo was mixed with the degenerate *plsB* oligo set before addition to the electroporation sample.
PlsC_1 custom	C*A*ACGGTTTCCCTCTAGAAATA ATTTTGTTTAACTTTAAGGAGGTA ATATACATATGCTATATATCTTTC GTCTTATTATTACCGTGATT	G555 RBS library Additional MOSAIC oligo for mutating the *plsC* RBS in plasmid G555. This oligo was mixed with the degenerate *plsC* oligo set before addition to the electroporation sample.
PlsC_2 custom	C*A*ACGGTTTCCCTCTAGAAATA ATTTTGTTTAACTTTAAGGAGGTA AGATACATATGCTATATATCTTTC GTCTTATTATTACCGTGATT	G555 RBS library Additional MOSAIC oligo for mutating the *plsC* RBS in plasmid G555. This oligo was mixed with the degenerate *plsC* oligo set before addition to the electroporation sample.
PssA_CdsA custom	A*G*CGGATAACAATTCCCCTCTA GAAATAATTTTGTTTAACGATAAG GAGGTGATATACATATGGCTAGC ATGACTGGTGGACAGCAAATG	G555 RBS library Additional MOSAIC oligo for mutating the *cdsA and pssA* RBSs in plasmid G555. This oligo was mixed with the degenerate *cdsA/pssA* oligo set before addition to the electroporation sample.
1543 ChD	C*A*ACGGTTTCCCTCTAGAAATA ATTTTGTTTAACTTTAAGGAGGTA AAATCCATATGACTTTCTGCTATC CTTGCCGCGCATTTGCATTA	G555 RBS variant (s12) targeting the *plsB* RBS (MOSAIC oligo).
1544 ChD	C*A*ACGGTTTCCCTCTAGAAATA ATTTTGTTTAACTTTAAGGAGGGA AAACCCATATGACTTTCTGCTATC CTTGCCGCGCATTTGCATTA	G555 RBS variant (s4) targeting the *plsB* RBS (MOSAIC oligo).
1545 ChD	C*A*ACGGTTTCCCTCTAGAAATA ATTTTGTTTAACTTTAAGGAGAAT ATATACATATGCTATATATCTTTC GTCTTATTATTACCGTGATT	G555 RBS variant (s7) targeting the *plsC* RBS (MOSAIC oligo).
1547 ChD	C*A*ACGGTTTCCCTCTAGAAATA ATTTTGTTTAACTTTAAGGGGAAT AGATATATATGCTATATATCTTTC GTCTTATTATTACCGTGATT	G555 RBS variant targeting the *plsC* RBS (MOSAIC oligo).
1549 ChD	A*T*AACAATTCCCCTCTAGAAAT AATTTTGTTTAACTTTAAGAATGA GTTATACATATGGCTAGCATGAC TGGTGGACAGCAAATGGGTCG	G555 RBS variant (s1) targeting the *cdsA/pssA* RBSs (MOSAIC oligo).
1481 ChD	A*T*AATTTTGTTTAACTTTAAGAA GGAGATATACATATGACNTTYTG YTAYCCNTGCCGCGCATTTGCATT ATTAACCAGAGGCTTTACAT	G555 GC library Degenerate MOSAIC oligo for modification of the GC content in the first five codons after the start codon of the *plsB* gene in plasmid G555.
1482 ChD	A*T*AATTTTGTTTAACTTTAAGAA GGAGATATACATATGCTNTAYAT HTTYCGNCTTATTATTACCGTGAT TTACAGCATCTTAGTCTGTG	G555 GC library Degenerate MOSAIC oligo for modification of the GC content in the first five codons after the start codon of the *plsC* gene in plasmid G555.

Table S11 continued

1483 ChD	T*C*TAGAAATAATTTTGTTTAACT TTAAGAAGGAGATATACATATGG CNAGYATKACNGKTGGACAGCA AATGGGTCGCGGATCCGGCTGC	G555 GC library Degenerate MOSAIC oligo for modification of the GC content in the first five codons after the start codon of the T7 tag preceding the *cdsA* and *pssA* genes in plasmid G555.
1551 ChD	A*T*AATTTTGTTTAACTTTAAGAA GGAGATATACATATGACATTTTG CTATCCATGCCGCGCATTTGCATT ATTAACCAGAGGCTTTACAT	G555 GC variant targeting *plsB* (MOSAIC oligo).
1554 ChD	A*T*AATTTTGTTTAACTTTAAGAA GGAGATATACATATGCTATATAT CTTTCGCCTTATTATTACCGTGAT TTACAGCATCTTAGTCTGTG	G555 GC variant targeting *plsC* (MOSAIC oligo).
1557 ChD	T*C*TAGAAATAATTTTGTTTAACT TTAAGAAGGAGATATACATATGG CCAGCATGACCGTTGGACAGCAA ATGGGTCGCGGATCCGGCTGC	G555 GC variant targeting *cdsA/pssA* (MOSAIC oligo).
1560 ChD	T*C*TAGAAATAATTTTGTTTAACT TTAAGAAGGAGATATACATATGG CTAGTATTACTGTTGGACAGCAA ATGGGTCGCGGATCCGGCTGC	G555 GC variant targeting *cdsA/pssA* (MOSAIC oligo).
491 ChD	5′Phos-AAAGTAAGCCCCCACCCTCACATG	PCR primer binding to oriL191 in G435 and G555.
492 ChD	5′Phos-AAAGTAGGGTACAGCGACAACATACAC	PCR primer binding to oriR194 in G435 and G555.
1523 ChD	5′Phos-AAAGTAAGCCCCCACCCTCACATGATACC	Alternative PCR primer for 491 ChD.
1524 ChD	5′Phos-AAAGTAGGGTACAGCGACAACATACACCATTTCC	Alternative PCR primer for 492 ChD.
1459 ChD	CTCCTAATATCGACATAATCCGTCGATCCTCG	PCR primer binding to oriL191 used for recovery of oriPL from liposomes after FACS.
1460 ChD	CCCCATTGACCGACTATCTTCGACAAG	PCR primer binding to oriR194 used for recovery of oriPL from liposomes after FACS.
1301 ChD	5′Phos-AAAGTAAGCCCCCACCCTCACATGATACCATTCTCCTAATATCGACATAATCCGTCGATCC	PCR primer binding to oriL191 used for recovery of oriPL from liposomes after FACS.
1518 ChD	5'Phos-AAAGTAGGGTACAGCGACAACATACACCATTTCCCCATTGACCGACTATCTTCG	PCR primer binding to oriR194 used for recovery of oriPL from liposomes after FACS.
976 ChD	GGATGAAGACTACCCGCTGC	qPCR primer binding in *DNAP*.
977 ChD	ACAGGTCTGCGATTTCACCG	qPCR primer binding in *DNAP*.

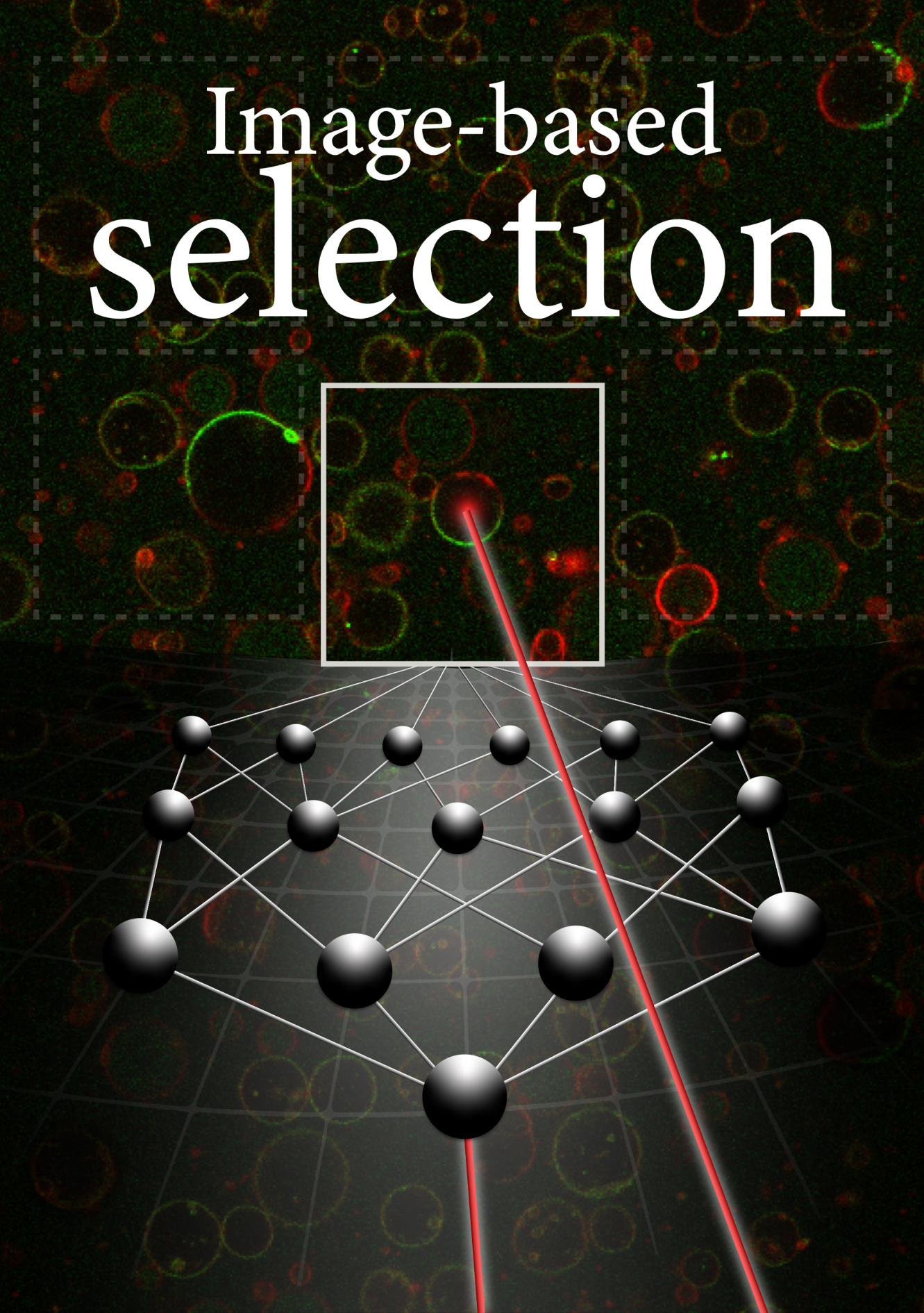
Table S12 Total number of reads of the different library fractions (n-site fractions) for the RBS library of the phospholipid synthesis pathway, supplementing the data in **Fig. S12**:

Round	1-site	2-site	3-site	4-site
R0	11,977	15,586	13,835	6,666
R1	19,404	17,740	14,456	7,134
R2	11,897	13,071	5,189	2,030
R3	9,528	7,790	3,268	970
R4	9,123	8,510	3,119	697

Table S13 Total number of reads for the different library fractions of the GC library of the phospholipid synthesis pathway, supplementing the data in **Fig. S16**.

Round	1-site	2-site	3-site	4-site
R0	10,314	12,067	9,410	3,607
R1	37,127	32,459	17,614	5,850
R2	55,764	31,071	14,882	2,988
R3	41,912	34,681	12,152	1,309
R4	34,025	39,306	17,120	1,639

Image-based selection



4

Image-based phenotypic sorting of synthetic cells

Understanding the relationships between genotype and phenotype is key to many areas of biological research and to the development of synthetic cells. We describe an image-based screening and sorting workflow that explores the phenotypes of gene-expressing vesicles within nonclonal populations and selects the desired variants. Using automated confocal microscopy and real-time, neural-network-assisted image analysis, we demonstrate that liposomes can be selected for fluorescence intensity, protein localization, membrane morphology, and dynamic behaviors, and their phenotype can be linked to genetic content. This approach could substantially accelerate the evolution of cellular functions in a minimal synthetic context.

Manuscript published as: van den Brink, M., Stam, M., Claassens, N. J., Danelon, C. *Science Advances* 12, eaed0860 (2026) doi:10.1126/sciadv.aed0860.

Illustration on previous page: Artistic impression of the method described in this chapter.

Introduction

Modular reconstitution of basic cellular functions in liposomes has laid the groundwork for building up an entire synthetic cell. For instance, DNA replication[1], transcription-translation[2–4], phospholipid synthesis[5–8], and membrane remodeling[9–14] have already been reconstituted in liposomes. The functional states of the engineered liposomes are commonly assessed by fluorescence microscopy techniques, including imaging flow cytometry[15]. Image-based methods allow for the exploration of a large observation space, defined as the set of features making up a liposome's phenotype. Therefore, vesicles exhibiting complex phenotypes, e.g., dynamic behaviors, spatial organizations and membrane morphologies, can be identified, expanding the range of target functions toward which liposomes can be engineered, ultimately accessing a high degree of "aliveness".

When the liposome's phenotype is directed by a gene expression program, methods of in vitro evolution can be applied to accelerate the search for a desired function[16,17]. Here, diversity is created by producing genetic variants directly inside liposomes[18] or externally prior to random encapsulation[19], leading to heterogenous—nonclonal—populations of liposomes exhibiting various, potentially distinct phenotypes. A major challenge resides in linking the specific phenotypes to their respective genotypes for sequence-function mapping. Pooled genetic screening of library variants in liposomes and isolation of the best-performing ones can be achieved by fluorescence-activated cell sorting (FACS)[19–25]. However, phenotypic interrogation here relies on unidimensional parameters, which restricts selection to simple properties.

Technological advances in image-based screening and sorting of cells are revolutionizing our understanding of genotype-to-phenotype relationships with great implications in single-cell omics and clinical research[26]. High-throughput image-activated cell sorters make sort decisions within micro- to milliseconds by real-time image analysis using predefined image parameters[27,28] and/or more powerful machine learning strategies[29] to classify complex phenotypes. However, drawbacks of these sorters are the low equipment accessibility, the limited spatial resolution, the limited number of fluorescence imaging channels, the inability to acquire time-lapse images or images in 3D (at multiple z planes), or the possible impact of the flow on some phenotypic traits[30]. Improved spatial resolution is obtained with imaging flow cytometry (IFC), but this technology lacks the ability to sort[31]. An alternative method that offers improved spatial and temporal resolution is the two-step combination of microscopy and FACS. Microscopy images are analyzed in real-time, whereafter the cells of interest are optically tagged for sorting by FACS. While this technique has been established for the analysis of mammalian and yeast cells[32–38], its potential in the research field of synthetic cells remains unexplored.

In this study, we describe a confocal microscopy-based method that assesses the phenotypes of gene-expressing liposomes and marks variants that exhibit the desired properties through

selective photoactivation. Photoactivated liposomes are then isolated by FACS, and the enrichment of the functional variants is determined by quantitative polymerase chain reaction (qPCR) or DNA sequencing. High-resolution image acquisition, analysis, and photoactivation are fully automated, employing trained neural networks and pre-defined parameters (e.g., intensity, size, circularity) for liposome classification. This approach allowed us to isolate liposomes with specific spatial and temporal features—including protein localization, membrane morphology and dynamic behaviors—from large populations of hundreds of thousands of liposomes per experiment. Our work thereby establishes microscopy-based phenotypic selection of gene-expressing liposomes as a core technology for the construction of synthetic cells using a system's level evolutionary approach.

Results

Establishing image-based liposome sorting

We aim to sort liposomes based on the visual phenotypes that arise from the expression of a library of different DNA variants. Synthesis of proteins that direct the phenotypic states is established directly in liposomes by the co-encapsulation of DNA and a cell-free transcription-translation system (PURE system)[39] (**Fig. 1**). We used a laser-scanning confocal fluorescence microscope with a high numerical aperture objective to screen 50-150 liposomes per field-of-view (FOV) enabling precise examination of vesicle morphologies and subcellular features. The liposomes also encapsulate purified PAmCherry2, a photoactivatable mCherry, which marks individual liposomes for sorting when its fluorescence is activated by point-stimulation with 405-nm light. Automated image acquisition, multimodal image analysis, and stimulation of targeted liposomes iterate over each FOV, allowing for the screening of hundreds of thousands of liposomes per experiment of ~8 hours. Next, the liposomes with photoactivated PAmCherry2 are isolated using FACS and the sorted DNA variants are analyzed by qPCR or PCR-amplified for DNA sequencing. The recovered variants could also be subjected to another round of diversification, re-encapsulation, and sorting.

First, we verified that co-encapsulation of 20 μM purified PAmCherry2 was compatible with the expression of a DNA template encoding YFP (**Fig. 2A**). Point-stimulation (405-nm light at 84 μW laser power for 300 ms) of single liposomes located in the center xy coordinate of the FOV led to an increase in PAmCherry2 intensity in the entire lumen. Imaging at different z planes confirmed that, while a variety of liposome sizes was present (from 2 up to 20 μm in diameter) most of the vesicles were captured in a single z plane ~2-3 μm above the glass surface (**Fig. 2A**). This is especially important for preventing off-target photoactivation in liposomes located within the cone-shaped illumination volume above and below the focal point. After photoactivating 29 liposomes with expressed YFP, all liposomes showed an increase in PAmCherry2 fluorescence intensity (**Fig. 2A**).

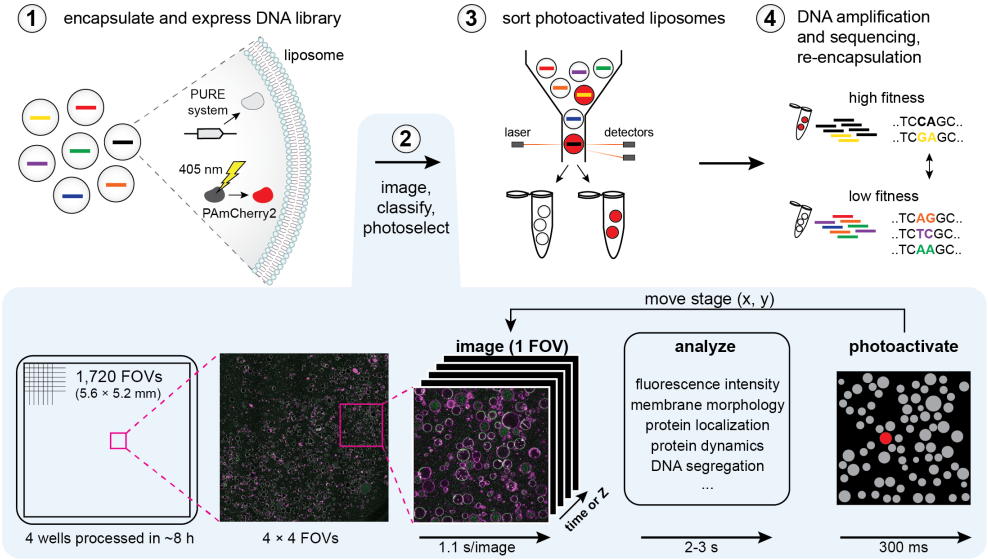


Fig. 1. Schematic workflow of image-based liposome sorting. Liposomes encapsulate a DNA library alongside PURE system and purified PAmCherry2. Automated imaging, real-time analysis, and point-stimulation by 405-nm light to photoactivate PAmCherry2 in liposomes with a desired visual phenotype, is iterated for each FOV. The duration of each step is indicated. The photoactivated liposomes (in red on the right-most picture) are then sorted by FACS and DNA is recovered for analysis or re-encapsulation. The microscopy images were acquired in an experiment described in Fig. 4C and reused here as representative examples.

Next, we examined the accuracy of PAmCherry2 photoactivation for different laser powers (34, 84 or 167 µW). All target liposomes were successfully tagged for the three photoactivation conditions with no significant differences in the median intensity values (**Fig. S1A**). We then measured the lumen intensities of all other liposomes within 15-µm xy distance to the point of stimulation and found activation precisions of 96% at 34 µW, 89% at 84 µW, and 86% at 167 µW (**Fig. S1A**). Thus, 34 µW laser power is the preferred choice, enabling photoactivation of all target liposomes, while limiting off-target activation. We also measured the movement of liposomes over a period of 10 seconds, a time sufficient for imaging, real-time analysis and photoactivation. The displacement of the liposome center coordinate was 0.5 ± 0.3 µm (n=15 liposomes), thus too small to cause off-target photoactivation.

Interestingly, we observed a positive correlation between the fluorescence intensities of YFP and activated PAmCherry2 (**Fig. S1B**). Specifically, upon photostimulation of liposomes that did not show YFP expression, these liposomes did also not exhibit post-activation PAmCherry2 signal. This result suggests that the encapsulation efficiency of different macromolecules in the swelling solution may be correlated. Furthermore, the YFP signal did not bleach upon photostimulation (**Fig. S1C**), and the activated PAmCherry2 was stable for at least 21 hours, giving enough time to sort the tagged liposomes by FACS (**Fig. S1D-E**).

Image-based enrichment of YFP-expressing vesicles

To assess the performance of the imaged-based selection pipeline, we quantified the precision and sensitivity for sorting YFP-expressing liposomes starting from a two-gene mock library. A mixed population of liposomes containing PAmCherry2, PURE system, and either a *YFP* gene or a control gene (1:39 volume ratio) was processed (**Fig. 2B-D**). The YFP-expressing liposomes were identified in real-time using basic analysis parameters including intensity thresholding, object size, and circularity, and were photoactivated. Per replicate experiment, we screened liposomes in 2 to 5 thousand FOVs (31 to 77 mm^2), and photoactivated ~800 to 7,200 liposomes (**Table S1**). Stimulated and non-stimulated samples were analyzed by flow cytometry (**Fig. 2E**, **Fig. S2**, **Table S2**). Approximately 1 to 2% of the liposomes were expressing YFP. This percentage is in agreement with the ratio at which the liposome populations were mixed (1:39) and the percentage of YFP-expressing liposomes in the initial population prior mixing with the control population (40 to 80%). While the non-stimulated sample only showed liposomes with a low PAmCherry2 signal, the stimulated sample yielded 0.3% ± 0.1% of the liposomes with a high PAmCherry2 signal (n=3 biological repeats). By processing the events in the four quadrants as indicated in **Fig. 2C**, we found that the YFP-expressing liposomes were detected and photoactivated with a precision of 100% (using laser power 34 µW) (**Fig. 2E**). Two replicate experiments with laser power 84 µW and where stimulation was not restricted to center of FOV resulted in a precision of 47% ± 8% (**Fig. S2**). Additionally, 15.3% ± 0.2% (n=3) of the YFP-expressing liposomes were photoactivated. This value defines the sensitivity and is specific to this experiment. It could be improved by increasing the fraction of the sample that was screened (here, only 50-80% of the chamber bottom-surface was imaged) and making the detection less stringent on liposome size and circularity. Ten to 20% of the photoactivated liposomes were detected by FACS and sorted, corresponding to 150 to 900 liposomes (**Table S1**). Lastly, the enrichment of the *YFP* gene over the control gene was quantified by qPCR (**Fig. 2F**, **Table S3**). The percentage of *YFP* amount increased from 3% ± 1% to 78% ± 17%, corresponding to a 39 ± 28 fold enrichment (n=3).

We further investigated the possible contribution of off-target stimulation to the presence of the control gene in the sorted fraction by directly sorting YFP-expressing liposomes from the mixed population by FACS, i.e., without imaging. Here, we obtained a percentage of *YFP* in the DNA pool of 81% ± 6% (n=4) after sorting (**Fig. S3**, **Tables S4-S6**) (to be compared to 78% with the image-based workflow), indicating that photostimulation did not significantly affect the sorting precision.

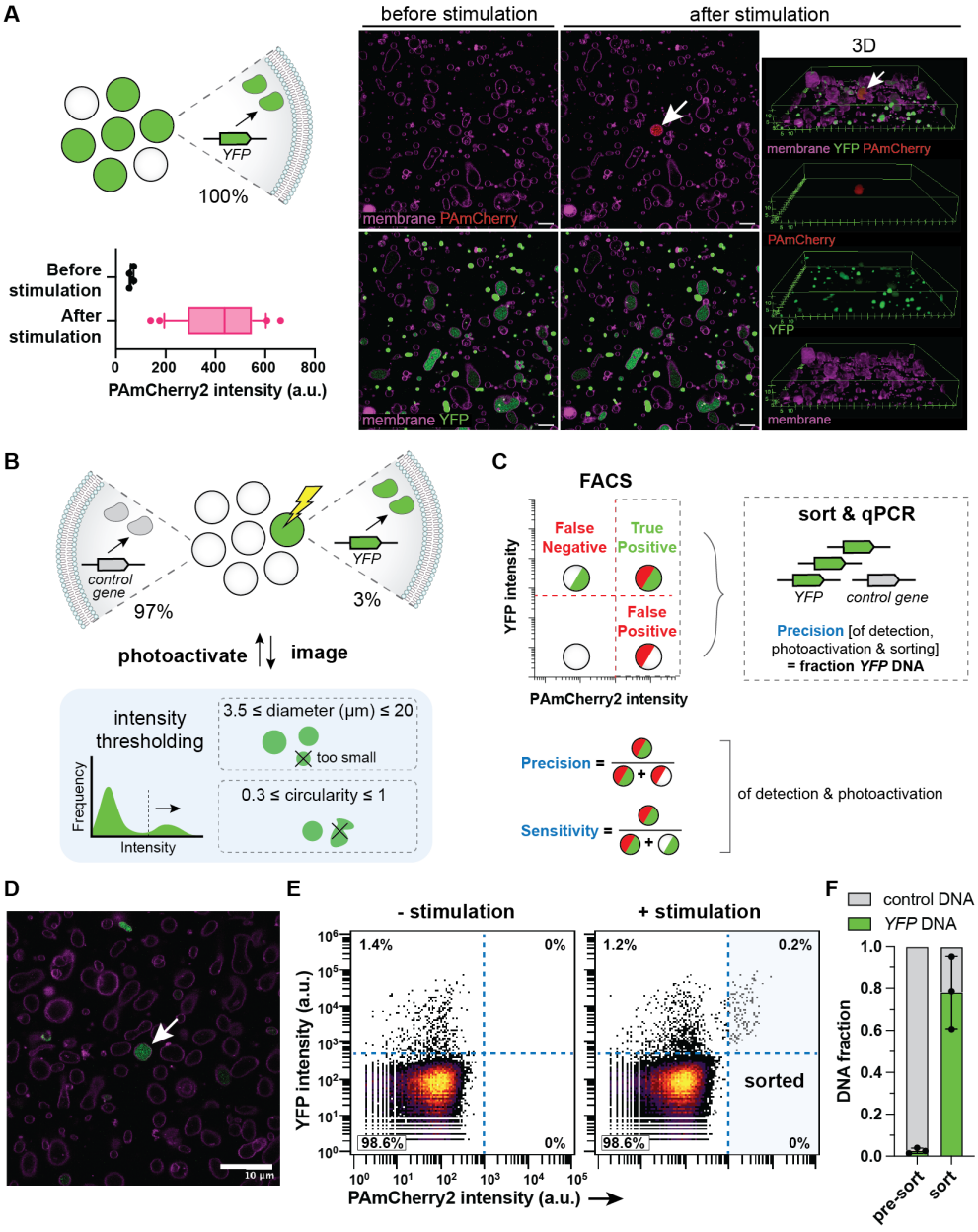


Fig. 2. Image-based sorting of YFP-expressing liposomes. (**A**) Photoactivation of PAmCherry2 in YFP-expressing liposomes. Liposomes contained 5 nM of *YFP* DNA template and 20 μM of purified PAmCherry2. Stimulation was performed in the center of the FOV using 84 μW laser power for 300 ms. The PAmCherry2 fluorescence intensity was measured 10 seconds after stimulation. Graph displays n=29 liposomes. Scale bars: 10 μm. (**B**) Image-based sorting of YFP-expressing liposomes from a two-gene mock library prepared by mixing two populations of liposomes at a 1:39 volume ratio, each containing an equal amount of linear DNA template, either *YFP* or *btubB* (control gene). Automated image analysis was performed using intensity thresholding on the YFP signal. (**C**) The precision and sensitivity of the method were assessed using FACS and qPCR analysis of the *YFP* and control DNA in the sorted liposomes. (**D**)

Example microscopy image of the mock library. Green, YFP; magenta, Cy5-conjugated lipids. Scale bar: 10 µm. Arrow indicates liposome selected for photostimulation (300 ms at 34 µW (replicates 1 and 2) or 84 µW (replicate 3) laser power). (**E**) Flow cytometry measurements of liposome populations without (28,898 events) or with (54,490 events) stimulation. (**F**) Fractions of *YFP* and control DNA in the sorted and pre-sort liposome samples. n=3 biological replicates (distinct mixed liposome populations).

Real-time, neural network-assisted detection of liposomes and spatial protein organization

Deep learning models have become state-of-the-art for image analysis thanks to their high prediction speed, robustness to biological variability, noisy data and varying imaging conditions, and reduced bias from manual parameter tuning[40,41]. Therefore, the single-stage object detector You Only Look Once (YOLO)[42–44] was chosen and used to identify liposomes in real-time based on their membrane morphology and the spatial organization of fluorescent proteins. As a proof-of-principle, we applied YOLO to detect proteins of the Min system, a regulator of bacterial cytokinesis and a candidate for the reconstitution of division processes in synthetic cells[45,10]. This protein system displays spatiotemporal organization when encapsulated in liposomes, providing an ideal study case for selection based on dynamic behaviors. We expressed the membrane binding protein MinD from a *minD* DNA template and added the purified eGFP-MinC fusion protein as a reporter of the presence of active MinD (**Fig. 3A**). The localization of eGFP-MinC in the lumen would indicate the absence of MinD, while its localization at the membrane would reveal the functional expression of MinD.

As training data, we used 700 images of liposomes containing both the lipid membrane and eGFP-MinC channels (**Fig. 3B**). The dataset originated from multiple liposome samples (biological replicates) to increase the model's robustness to variation in sample preparation and imaging conditions. Each image displayed many liposomes with various phenotypes: no eGFP-MinC signal, eGFP-MinC on the membrane or/and in the lumen, unilamellar or multilamellar vesicles, circular liposomes or structures with other shapes, and a wide distribution of sizes. We annotated 688 objects in one class, aiming for unilamellar, circular liposomes of 3 to 15 µm in diameter with eGFP-MinC on the membrane. Using 80% of the labelling dataset, we trained YOLOv3-tiny[43] with pretrained weights, and both YOLOv3-tiny and YOLOv7-tiny[44] with and without pretrained weights (**Fig. S4A**). We tested the two models using the remaining 20% images and found that the best-performing model was YOLOv7-tiny with a precision of 82% and recall of 82% at a confidence threshold of 0.8 (**Fig. 3C**). Communication between the image acquisition software and the trained YOLOv7-tiny model was established for real-time image analysis and photoactivation.

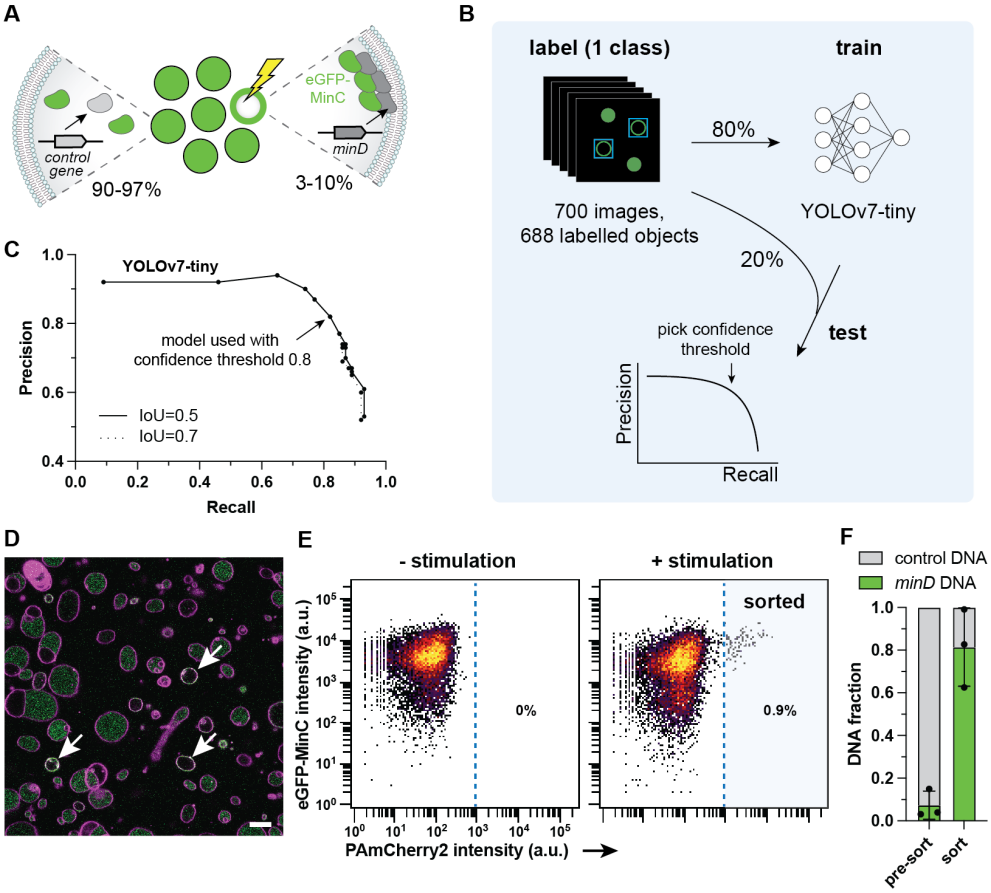


Fig. 3. Image-assisted sorting of membrane-localized MinD using deep learning. (**A**) A mock DNA library was prepared by mixing two populations of liposomes (1:9 or 1:39 volume ratio) expressing either *minD* or *btubB* (control gene) and containing 500 nM of purified eGFP-MinC. When MinD is expressed, it binds to the membrane and recruits eGFP-MinC. (**B**) Labelling, training, and testing of YOLOv7-tiny using images displaying the membrane reporter dye Cy5 and GFP signals. (**C**) Performance of the trained YOLOv7-tiny model. IoU is Intersection over Union. (**D**) Example microscopy image of the mock library. Green, eGFP-MinC; magenta, Cy5-conjugated lipids. Scale bar: 10 µm. Automated image analysis was performed by the trained YOLOv7-tiny model. Arrows indicate liposomes classified for photostimulation during 300 ms at 34 µW or 84 µW laser power (Table S7). (**E**) Flow cytometry measurements of liposome populations without (10,365 events) or with (10,262 events) stimulation. Between 170 and 900 liposomes were sorted per replicate. (**F**) Fractions of *minD* and control DNA in the sorted and pre-sort liposome samples. n=3 biological replicates.

Two populations of liposomes containing PAmCherry2, PURE system, purified eGFP-MinC, and either the *minD* gene or a control gene were mixed in a 1:9 (n=1) or 1:39 (n=2) ratio (*minD* : control gene) (**Fig. 3A**). Liposomes exhibiting eGFP-MinC signal at the membrane were photostimulated to select vesicles containing *minD* (**Fig. 3D**, **Table S7**). Photoactivated liposomes representing 0.03 to 0.94% of the total population were sorted by FACS (**Fig. 3E, Fig. S5, Table S8**), and the fractions of gene variants were quantified by qPCR in the pre-sort

and sorted liposome samples. The percentage of *minD* increased from 7% ± 7% in the pre-sort sample to 81% ± 18% after sorting, corresponding to a 19 ± 12 fold enrichment (n=3 biological repeats) (**Fig. 3F, Table S9**). Because neural network analysis can be applied to identify images with rich spatial and morphological information, these results show the potential of our method for the selection of synthetic cells with complex phenotypes from heterogeneous samples.

Video-based photoactivation of liposomes displaying dynamic protein oscillations

As flow-based image-activated sorters can only detect static features, we sought to exploit our method to photoactivate liposomes based on dynamic processes. In *Escherichia coli*, the Min system undergoes oscillatory behaviors that aid in positioning the division site[46]. In liposomes, co-expression of MinD and MinE leads to different types of dynamic behavior, including pulsing, circling, and pole-to-pole oscillations with periods typically ranging from 15 to 45 seconds (**Fig. 4A**)[10]. Liposomes can also exhibit an uncategorized or halted dynamic mode, and modes can switch from one to another or change directionality, leading to a large variability in dynamic and static phenotypes within one liposome sample. Consequently, time-lapse microscopy is required to assess whether a liposome exhibits a dynamic or static organization of Min proteins. To detect the temporal behaviors of the Min proteins in real time, the liposomes were imaged with 12 second intervals over a period of 48 seconds. We trained YOLOv7-tiny to detect two classes: class 1 consists of liposomes with eGFP-MinC not located on the membrane, while class 2 consists of liposomes with eGFP-MinC (partly) located on the membrane (**Fig. 4B**). The model was trained on 80% of the images and tested using the separate 20% of the dataset, yielding a precision of 83% and recall of 86% at a confidence threshold of 0.9 (**Fig. S4B**). Hereafter, we retrained a model on the complete labelled dataset (700 to 900 objects per class).

Detected liposomes were tracked across multiple time frames using a nearest-neighbor search (**Fig. 4B**). While most liposomes remained stationary during the measurements, a few moved substantially in xy direction, increasing the likelihood of being misidentified as a different liposome. Therefore, we applied a 5-µm maximum displacement threshold relative to the position in the first frame. Then, each class for a given liposome was compared over the consecutive time frames. The change from one class to the other reveals a dynamic event and these liposomes were photoactivated. In 8 hours, we screened approximately 40,000 liposomes and stimulated one liposome in every 2 FOVs (~0.5% of all liposomes) (**Fig. 4C**, **Table S10**). Manual labelling of liposomes showing both the presence and absence of membrane-localized eGFP-MinC across different time frames resulted in a precision of 76 ± 10% and a sensitivity of 44 ± 5% (n=3 biological repeats), validating video-based classification of liposomes (**Fig. 4D**).

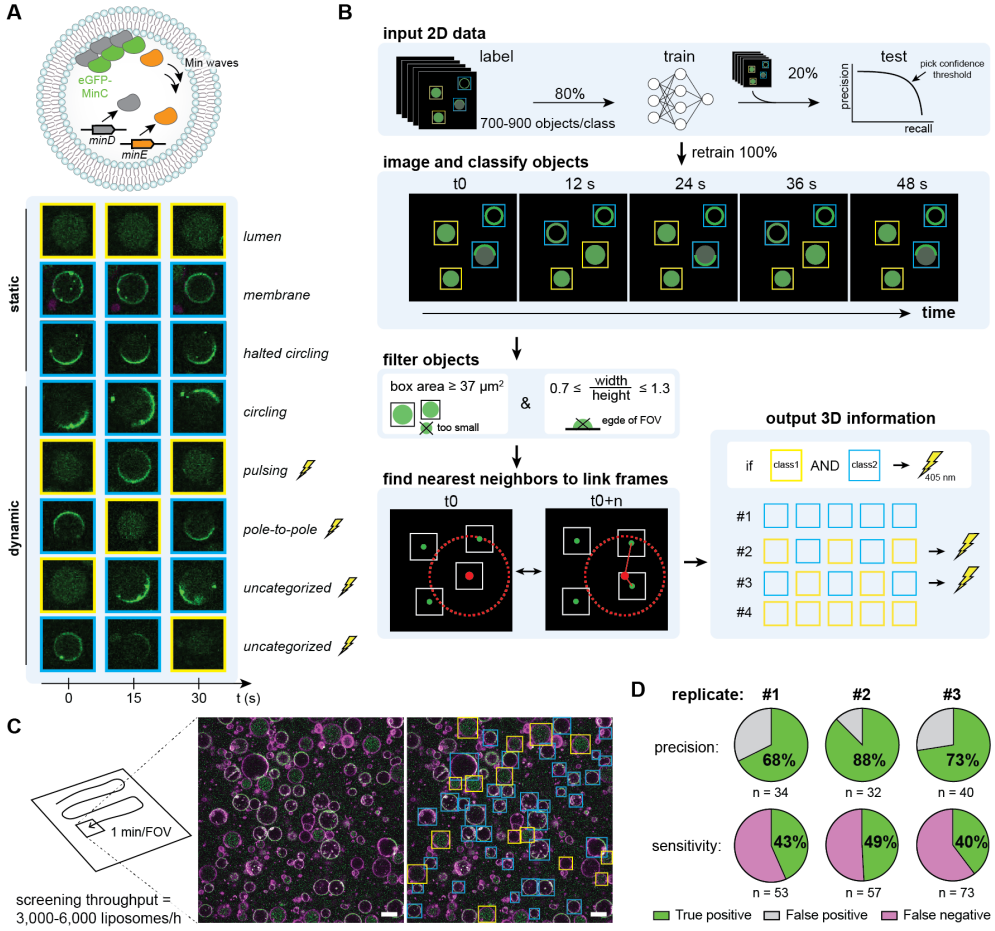


Fig. 4. Real-time detection and tagging of liposomes with dynamic Min oscillations. (**A**) Min oscillations driven by the co-expression of MinD and MinE were visualized using purified eGFP-MinC. Representative images of eight liposomes exhibiting static or dynamic behaviors. Only liposomes with dynamic patterns were photoactivated. (**B**) Automated image analysis pipeline. Liposomes were detected using a YOLOv7-tiny model, trained on 2D data using two classes (indicated with blue and yellow bounding boxes). To assure accurate liposome tracking across time frames, small liposomes (diameter < 6 µm), which tend to move in and out of focus more than larger liposomes, and liposomes with aspect ratio below 0.7 and above 1.3, were excluded. Linking multiple time frames through a nearest neighbor search enabled the discrimination of static vs dynamic behaviors. (**C**) Automated acquisition of liposome images and real-time identification of static vs dynamic self-organization of Min proteins. One example microscopy image is presented twice: without (left) and with (right) annotation of the detected liposomes. Blue boxes indicate detected liposomes where Min proteins localize at the membrane, and yellow boxes indicate liposomes where the fluorescent proteins did not localize to the membrane. Scale bars: 10 µm. (**D**) Performance of the classification pipeline. Precision and sensitivity were calculated relative to the ground truth, which entailed manual classification of the liposomes in the videos. n=3 biological replicates. Stimulation time was 300 ms, and laser power was 34 µW. The YOLOv7-tiny model from replicate 1 was retrained with 9 additional images to strengthen the model (Fig. S4b). For replicates 1 and 2, a population of liposomes expressing MinD and MinE was mixed 1:1 with liposomes expressing MinD and a control protein (BtubB). For replicate 3, only liposomes expressing MinD and MinE were used.

Selection of best-performing membrane-binders from a library of *minD* variants

At last, we showed that image-based screening and sorting can be successfully coupled to the recovery and sequencing of DNA from the sorted liposomes. We sought to find the best-performing variants within a *minD* library harboring mutations in the hydrophobic patch of the membrane targeting sequence (MTS)[47], most likely conferring different abilities to bind to the liposome membrane (**Fig. 5A**). Nine DNA variants were separately encapsulated and expressed in liposomes, whereafter the clonal liposome populations were mixed and subjected to the image-based selection pipeline to sort the liposomes with membrane-localized eGFP-MinC (**Fig. 5B**, **Fig. S6**, **Tables S11-S12**). The DNA was recovered by PCR from 88 up to 704 sorted liposomes per biological replicate (**Fig. 5C**, **Fig. S7C**), and the fraction of each variant before and after sorting was quantified by Nanopore sequencing (**Fig. 5D**). From the nine variants present in the input library, only two were found in the sorted pool: Var1 (the wild type) and Var7 (amino acids: FL>LH). Encapsulation and expression of the individual variants confirmed that Var1 and Var7 encode a membrane-binding MinD, while the other mutants could not (**Fig. 5E, Fig. S7A-B**). These results are in agreement with a previous study that showed that substitution of hydrophobic amino acids to charged amino acids in the MTS caused defective binding of MinD to the membrane in *E. coli*[47]. Altogether, these findings demonstrate the feasibility to isolate and genotype active gene variants from a library using image-based liposome phenotyping.

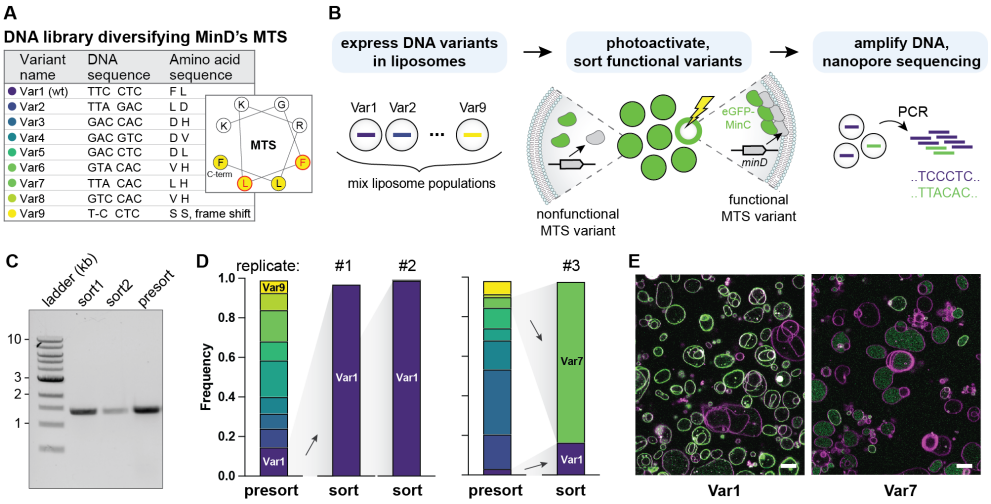


Fig. 5. Enrichment of active *minD* variants from a library. (**A**) Design of the library of MTSs in the gene *minD*. Mutations were introduced in two of the codons in the hydrophobic patch of the highly conserved MTS motif of MinD. (**B**) Image-based sorting of liposomes expressing MinD variants with different affinities to the membrane. Populations of clonal liposomes, each containing one library variant, were pooled and subjected to image-based selection. Mutations lead to differential localization of eGFP-MinC to the membrane (functional MTS) or to the lumen (nonfunctional MTS). DNA was amplified from the sorted liposomes by PCR and sequenced. Stimulation time was 300 ms and laser power was 34 µW. (**C**) Gel electrophoresis with PCR-amplified *minD* (1,398 bp) recovered from liposomes sorted by FACS. (**D**) Frequency of the DNA variants was quantified using Nanopore sequencing before and after liposome

sorting. Two starting ratios of the nine DNA variants were prepared, from which in total three sorting replicates were performed. (**E**) Confocal microscopy images of the two enriched variants (Var1 and Var7) separately encapsulated and expressed in liposomes. Green, eGFP-MinC; magenta, Cy5-conjugated lipids. Scale bar: 10 µm.

Discussion

We describe a confocal microscopy-based method that explores the phenotypes of gene-expressing liposomes within large populations and sorts the desired variants using photo-tagging followed by FACS. Our results offer a new promising way to accelerate the development of synthetic cells through in vitro evolution whereby the system's fitness is assessed by visual screening. With this objective in view, we here discuss the current assets of the proposed technology compared to existing ones (see **Fig. S8** and **Supplementary Data 1** for benchmarking analysis), as well as possible extensions that could enable the discovery of phenotypically complex synthetic cells from a large library of genetic and molecular variants.

Unlike traditional methods that enrich variants based on one-dimensional readouts in time and space, such as FACS[19–25], affinity-based selection[17,48], and self-selection through DNA replication[18], visual phenotyping allowed us to classify liposomes based on morphology and spatiotemporal protein organization. While photoselection-based methods similar to ours have been developed for biological cells[32–38], working with liposomal synthetic cells imposes several challenges: the small size (diameter < 10 µm), polydispersity and low phase-contrast of the vesicles, stochastic encapsulation of the photoactivatable probe, and positioning of the liposomes on a 2D surface. Because of these constraints, we chose confocal-fluorescence imaging using a high numerical aperture (NA 1.49, 100× oil-immersion) objective to measure with a high spatial resolution. Despite the small size of the liposomes, photoactivation of single vesicles was accurate and sensitive; all liposomes that displayed gene expression also exhibited higher PAmCherry2 fluorescence after stimulation (**Fig. 2, Fig. S1**).

In this work, we explored a combined set of liposome features including fluorescence intensity, protein co-localization to the membrane or lumen, and liposome morphology. We also introduced a temporal dimension by linking these phenotypes across time frames. By capturing 50 to 150 liposomes per FOV (128 µm × 128 µm), we screened up to 700,000 liposomes per experiment of ~8 hours (including photoactivation). We demonstrated successful recovery and sequencing of the DNA underlying the selected phenotypes for typical sorting fractions of 150 to 1,000 liposomes, and we could even recover the DNA from as low as 88 sorted liposomes. We showcased the enrichment of the most active DNA variants from a library of nine variants (**Fig. 5**). The next step in image-based sorting of synthetic cells is enlarging the library size to hundreds or thousands of variants to make better use of the available screening capacity. This will require establishing a method for multiplexing liposome production, where many liposome samples can be prepared in parallel, for example using lab

automation. Alternatively, the DNA library could be encapsulated in a single liposome sample at a low DNA concentration (~1 copy per liposome), but this strategy demands an increase in the fraction of 'active' liposomes to enable screening of large libraries, for instance by coupling gene expression with clonal amplification[49]. Such approaches should go hand-in-hand with smart library design and active learning-assisted directed evolution to enable efficient screening of large genetic and molecular spaces by focusing experimental testing to a smaller subset containing the most informative variants[50,51].

In clonal liposome samples, only 40-80% of the liposomes show gene expression above the detected threshold (**Fig. 2a**, **Fig. S2**). Nongenetic factors that may affect protein production include liposome size, membrane permeability to low-molecular weight compounds (e.g. NTPs, amino acids, energy substrates), and encapsulation efficiency of PURE system components[52]. Because such a heterogeneity similarly manifests for each DNA variant in the library, phenotypic selection still hinges on vesicles encapsulating the best performing genetic variants. However, these effects raise the number of liposomes to be screened for efficient library coverage.

In this work, we typically photoactivated 1,000-10,000 liposomes of which we sorted 100-1,000 by FACS (**Tables S1, S7, S11**). The 10% recovery rate of photoactivated liposomes could limit the sorting efficiency when the phenotypes of interest are extremely rare. Given that not all liposomes within a clonal population are active (say 10% as a lower bound), our method can assess 200,000 DNA variants when screening 2 million liposomes, which takes ~24 h microscope time. Considering a minimum of 1,000 photoactivated liposomes, our method is currently most suited for DNA libraries with at least 1 in 200 beneficial DNA variants. The 10% recovery rate of photoactivated liposomes by FACS is somewhat lower than those reported for cells (26-54%)[33,38]. The disruption of liposomes, incomplete harvesting during transfer from the imaging well to the FACS instrument, or the low scattered light intensity of the liposomes may explain this observation. Potential improvements include using the PAmCherry2 signal instead of the scattered light intensity to trigger event recording during FACS and designing microscopy chambers that would facilitate harvesting of the liposomes. Furthermore, the PAmCherry2 concentration could be increased, or an alternative photoactivatable protein or dye could be used.

While the screening throughput depends on the number of stimulated liposomes per FOV, it is on par with that of existing methods applied to biological cells (**Supplementary Data 1**). Yet it is low compared to flow-based imaging sorters that can typically analyze samples at a speed of 10^2 to 10^3 events per second[27,29]. Strategies to increase the throughput include (combinations of) shortening (i) the acquisition time by using a spinning-disk confocal microscope, (ii) the analysis time by processing multiple FOVs simultaneously or speeding up the communication between acquisition and analysis, and (iii) the stimulation time while increasing the 405-nm laser power. Noteworthy, ultrawide-FOV microscopes could

dramatically increase the throughput by imaging FOVs of ~5 to 42 mm^2 at millisecond temporal resolution[36,37,53]. However, a better tradeoff will have to be found for improving the spatial resolution (NA objective 0.25 or 0.5)[36,37,53] as it is currently insufficient for the characterization of internal liposome structures. The same limitation of the spatial resolution also applies to the flow-based imaging sorters IACS2.0[29] and BD FACSDiscover S8[27,28].

High-content imaging, which captures multiple features of gene-expressing liposomes, enables multi-objective optimization that reflects the integrated activities of several biological processes. As the number of reconstituted biological modules increases, the liposome phenotypes become more complex, and the observation space becomes larger. Therefore, sorting multiple subpopulations would be very useful to obtain a more holistic view of the effects of genetic and molecular variations on the phenotype—beyond the best-performing variants only. This allows better identification of neutral and deleterious mutations and may provide deeper insights into mutational effects across different genetic or molecular backgrounds[54]. Sorting multiple subpopulations might be achieved by varying the activation time or laser power and sort multiple phenotypes by FACS based on PAmCherry2 intensity thresholding[32,34]. However, the overlap of PAmCherry2 intensity distributions at laser powers spanning 34 to 167 µW currently challenges multi-way sorting of photoactivated liposomes. Adjusting both the laser power and stimulation time might generate distinct PAmCherry2 intensity profiles.

While in this work, liposomes were prepared via gentle hydration of lipid films dried on glass beads[55], the screening-sorting workflow could also be applied to gene-expressing giant liposomes formed via alternative protocols, such as water-in-oil droplet transfer across an oil-water interface[3,56] or droplet-based microfluidics[57]. Furthermore, in addition to DNA sequencing, the sorted liposome fractions could be subjected to other analytical investigations to reveal a rich array of cellular properties and functions, which can be coupled to the fitness level and genotype of each subpopulation. Recent advances in single-cell 'omics' could be applied to gene-expressing liposomes for mapping genotype and detailed molecular information, e.g., proteins, RNA transcripts and metabolites[32,36,58,59].

Lastly, the current classification strategy relies on pre-defined metrics of physical parameters or on supervised machine learning with liposome images annotated by the user. This approach introduces potential human bias that narrows the exploration of phenotypic states. Unsupervised methods based on the structure of the data itself and lacking an explicit optimization target may increase robustness and expand the range of possible discoveries by revealing difficult-to-identify or hidden properties of the liposomes[60]. Implementing unsupervised learning to our imaging-based pipeline would lay the foundation for open-ended evolution of synthetic cells[16].

Materials and Methods

Reagents and plasmid DNA

Solutions were prepared using Milli-Q water with 18.2 MΩ resistivity (Merck Millipore). Reagents were purchased from Sigma-Aldrich unless indicated otherwise. Primers and plasmids are listed in **Tables S13** and **S14**, respectively. Plasmids G365 (pUC-*YFP*) and G437 (pUC-*minD*) were constructed as described earlier[49]. G439 (pUC-*btubB*) was constructed as described earlier[12]. G376 (pUC-*minE_opt*) was ordered from GenScript, as described earlier[10]. The expression vector G595 (pET15b-*PAmCherry2*) was constructed by cloning of the *PAmCherry2.1* gene[61] (amplified with primers 1429 ChD and 1430 ChD) into the pET15b backbone (amplified with primers 1431 ChD and 1432 ChD) using Gibson Assembly. The expression vector G577 (pET15b-mEos2) was constructed by cloning of the *mEos2* gene (amplified from G571 with primers 1355 ChD and 1356 ChD) into the pET15b backbone (amplified with 1353 ChD and 1354 ChD) using Gibson Assembly.

Protein purification

eGFP-MinC was purified as described in published protocols[62]. To synthesize PAmCherry2 and mEos2, 10 mL of overnight culture of *E. coli* BL21 (DE3) (New England Biolabs) carrying G595 (PAmCherry2) or G577 (mEos2) was diluted 1:100 in Lysogeny Broth (LB) medium containing 50 µg mL^{-1} ampicillin. The diluted culture was grown in a shaking incubator at 37 °C until the OD_{600} reached 0.4-0.6. Gene expression, under control of the T7-LacO promotor and in frame with an N-terminal His-tag, was induced by addition of 0.25 mM (PAmCherry2) or 1 mM (mEos2) isopropyl-β-D-thiogalactopyranoside (IPTG). The cells were incubated overnight at 16 °C in a shaking incubator (PAmCherry2) or at 26 °C for 4 h (mEos2). The cells from 1 L of culture were pelleted by centrifugation and resuspended in 10 mL of lysis buffer (50 mM HEPES-KOH pH 7.5, 500 mM NaCl and 10% glycerol). The cells were sonicated in an ice bath six times 20-30 seconds with an interval of 1 min and an amplitude of 40%. After centrifugation at 16,000 g for 15 min at 4 °C, the cell-free extract (supernatant) was harvested. 10 mM imidazole and SetIII protease inhibitor (EDTA-free, 1:1000 dilution, Calbiochem) were added. The cell-free extract was mixed with a HisPure Ni-NTA resin (Thermo Scientific), priorly equilibrated with lysis buffer, and incubated at 4 °C for 2 hours while tumbling. The unbound fraction was removed by spinning the resin at 750 g for 2 min. The resin with bound protein was washed twice with 20 equivalents (volume) of wash buffer (50 mM HEPES-KOH pH 7.5, 500 mM NaCl, 50 mM (PAmCherry2) or 40 mM (mEos2) imidazole, 10% glycerol). The protein was eluted with 6 mL of elution buffer (50 mM HEPES-KOH pH 7.5, 500 mM NaCl, 500 mM imidazole, 10% glycerol). The fluorescent fractions (for mEos2) were pooled together and buffer-exchanged with storage buffer (50 mM HEPES-KOH pH 7.5, 50 mM (PAmCherry2) or 100 mM (mEos2) NaCl, 10% glycerol) using a 10 kDa MWCO Amicon Ultra-15 centrifugal filter unit (Merck). Protein concentration was determined with a Bradford assay.

Preparation of linear DNA templates for encapsulation in liposomes

Linear DNA templates for in-liposome expression were amplified by PCR from plasmids G365 (*YFP*), G439 (*btubB*), G437 (*minD*), G376 (*minE*) with primers 757 ChD and 709 ChD annealing to the T7 promoter and T7 terminator, respectively. The PCR reactions were set up at 50-µL volume and contained 200 µM of dNTPs, 500 nM of each primer, 200 pg μL^{-1} of plasmid DNA and 1 unit of Phusion High-Fidelity DNA Polymerase (Thermo Scientific) in 1× Phusion HF Buffer. The thermocycling reaction was performed as follows: 98 °C for 30 s, (98 °C for 10 s, 56 °C for 20 s, 72 °C for 30 s) × 30, and 72 °C for 10 min. The amplicons were purified directly or by gel extraction using the QIAquick PCR & Gel Cleanup Kit (Qiagen). The DNA concentration was measured with the Qubit dsDNA High Sensitivity Assay Kit and the Qubit 4 Fluorometer (Invitrogen).

Synthesis of *minD* plasmid library and isolation of single variants

A 32-variant *minD* plasmid library (G598) with mutations in the MTS of MinD was synthesized by full-plasmid amplification of plasmid G437 (pUC19-*minD*) with a mutagenic primer set and subsequent re-circularization using the protocol adapted from Galka *et al.*[63]. PCR amplification was performed with primers 1445 ChD and 1448 ChD, of which the latter contains degeneracy at the MTS. The PCR reaction was set up at 20-µL volume and contained 400 µM of dNTPs, 300 nM of each primer, 350 pg μL^{-1} plasmid DNA and 0.4 units of KOD Xtreme Hot Start DNA Polymerase in 1× KOD Xtreme PCR Buffer and 5% DMSO. A touchdown thermocycling reaction was performed as follows: 94 °C for 2 min, (98 °C for 10 s, 74 °C for 4.2 min) × 4, (98 °C for 10 s, 72 °C for 4.2 min) × 4, (98 °C for 10 s, 70 °C for 4.2 min) × 4, (98 °C for 10 s, 60 °C for 20 s, 70 °C for 4.2 min) × 12, and 70 °C for 5 min.

Because the two primers are (partly) complementary, the two ends of the amplicon were identical, enabling re-circularization by mixing 18 µL of the PCR reaction with 54 µL of a Gibson reaction mix containing 1.3× ISO Buffer (0.5 M of Tris-HCl pH 7.5, 50 mM of $MgCl_2$, 4 mM of dNTP mix, 50 mM of DTT, 0.25 g mL^{-1} of PEG-8000, 5 mM NAD), 0.053 U μL^{-1} of T5 exonuclease (Epicentre), 0.0033 U μL^{-1} of Phusion High-Fidelity DNA polymerase (Thermo Scientific), 5.3 U μL^{-1} of Taq DNA ligase (NEB) and 50 U mL^{-1} of DpnI endonuclease (NEB). The reaction was incubated in a thermocycler at 50 °C for 2 hours. The full reaction was loaded onto a 0.7% agarose gel, and the band corresponding to the circularized amplicon was purified by gel excision and further purified using QIAquick PCR & Gel Cleanup Kit (Qiagen).

Next, 50 µL of competent *E. coli* K-12 DH5α cells were transformed with 3 ng of purified plasmid using heat shock. Single colonies were picked from ampicillin-supplemented LB agar plates and grown overnight in ampicillin-supplemented (50 µg mL^{-1}) LB medium. The plasmids were isolated with the PureYield Plasmid Miniprep System (Promega), and their sequences were analyzed by Sanger sequencing (Macrogen Europe, Netherlands) using primer 709 ChD.

Liposome preparation

The swelling solutions (10 µL) were prepared on ice by mixing 5 µL of PURE*frex* 2.0 Solution I, 0.5 µL of Solution II, 1 µL of Solution III, 0.75 U μL^{-1} SUPERase·In RNase Inhibitor (Ambion), 20 µM purified PAmCherry2 and 5 nM linear DNA template (unless indicated otherwise) in a 1.5-mL Eppendorf tube. When expressing proteins from the Min system, the solution was supplemented with 0.5 µM of purified eGFP-MinC, 0.5 µL DnaK Mix (GeneFrontier) and 2.5 mM of ATP. The DNA concentration was adjusted to 6 nM *minD* and 3.8 nM *minE* when both proteins were co-expressed. Next, 5-6 mg of lipid-coated glass beads were added to the swelling solutions. The lipid-coated beads were prepared as described earlier[52] with minor modifications. The glass beads (0.6 g) were coated with 50 mol % DOPC, 36 mol % DOPE, 12 mol % DOPG, 2 mol % 18:1 cardiolipin, 0.05 mass % DOPE-Cy5 and 1 mass % DSPE-PEG(2000)-biotin with a total mass of lipids of 2 mg. The mixture containing PURE system and lipids was gently rotated at 4 °C for 1 hour to allow for liposome swelling. Four freeze-thaw cycles were applied by dipping the samples in liquid nitrogen for 10 seconds followed by thawing on ice.

In-liposome gene expression

For liposome samples containing either *YFP*, *btubB*, or *minD* DNA, 0.25 µL of DNase I (NEB, 2 U μL^{-1}) was transferred into an empty PCR tube. Using a cut pipette tip, 7 µL of liposomes were transferred to the PCR tubes and mixed with the DNase I by pipetting up and down three times. Liposomes were incubated at 37 °C for 3-16 hours in a thermocycler. The 5.7 × 6.1 mm chambers (length × width) of a µ-Slide 18 Well Glass Bottom (Ibidi) were pre-incubated with 50 µL of 1 mg mL^{-1} BSA in Milli-Q water for 10 minutes and subsequently washed three times with Milli-Q water and once with PB buffer (20 mM HEPES-KOH pH 7.6, 180 mM potassium glutamate, and 14 mM magnesium acetate). Liposomes were diluted 50-fold in PB buffer of which 50 µL was transferred to each chamber. The imaging slide was kept overnight in the dark at 4 °C to allow for the liposomes to sediment.

For experiments with MinD and MinE, the liposome suspension obtained after swelling was diluted 2-fold using a mixture of 1:1 PURE*frex* 2.0 Solution I and Milli-Q water supplemented with 2.5 mM ATP and 0.13 U μL^{-1} DNase I. Next, 10 µL of the diluted liposome sample was transferred to an in-house made, BSA-functionalized round imaging chamber (diameter ~3 mm). The sample slide was incubated in a laboratory incubator at 37 °C for 2-3 hours before imaging at room temperature.

Confocal microscopy

Liposomes were imaged using the A1R Laser scanning confocal microscope (Nikon) equipped with an SR Apo TIRF 100× oil immersion objective (NA 1.49). Imaging was performed using laser line 488 nm with emission filter 525/50 nm (gallium arsenide phosphide (GaAsP) detector; YFP and eGFP-MinC), laser line 561 nm with emission filter 595/50 nm (GaAsP detector; PAmCherry2) and laser line 640 nm with emission filter 700/75 nm (photomultiplier tube detector; DOPE-Cy5). For image-based screening, channels 1 (YFP, eGFP-MinC) and 3 (DOPE-

Cy5) were imaged simultaneously. Imaging was performed at a fixed dwell time (2.2 µs) and image size (512 × 512 pixels, 0.25 µm pixel^{-1}) to optimize for high imaging speed. The pinhole size was 42.1 µm (1 A.U. for 405 nm). Point stimulation was performed using the 405 nm laser for 300 ms at 10-50% laser power (34-167 µW). Image acquisition and stimulation were operated by NIS-Elements version 5.30.07 (Nikon) at room temperature with the focus drift correction module Perfect Focus System (Nikon) enabled.

For optimal stimulation accuracy, point stimulation was performed in the center of the field-of-view (FOV) using the Galvano scanner mirrors. The point stimulation accuracy was regularly checked by the green-to-red photoconvertion of mEos2 using a 3-µL droplet of 50 µM protein in PB buffer sandwiched between a glass slide and coverslip and 1% laser power (5 µW) for 300 ms (**Fig. S9**). When a misalignment of a few pixels was observed, the stage position was adjusted accordingly. The absolute laser powers (in µW) for point stimulation were measured by a Microscope Slide Photodiode Power Sensor with 18 mm × 18 mm Active Area (S170C, THORLABS Germany) placed on top of the 100× objective (without oil) and connected to a Digital Optical Power and Energy Meter (PM100D, THORLABS Germany). The absolute power was measured for 1%, 10%, 25% and 50% relative laser powers while point-stimulating in the center of the FOV. The power value was averaged over 10 s.

Deviations from the above-mentioned protocols for some experiments are described in **Tables S1**, **S7**, **S10** and **S11**.

Quantification of the lumen intensities of target and off-target liposomes

To quantify the change of PAmCherry2 fluorescence intensity upon stimulation, liposomes expressing YFP were imaged before and 10 seconds after stimulation. Using Fiji (ImageJ2 v2.9.0)[64], the PAmCherry2 and YFP intensities were measured for the target liposome and all off-target liposomes within a distance of 15 µm. For each liposome, an area was defined manually that covered most of the lumen, of which the average intensity was calculated.

Training and testing of YOLO

Train and test sets for YOLOv3-tiny[43] and YOLOv7-tiny[44] algorithms were generated by manually drawing bounding boxes using the open-source software Label Studio[65] (https://github.com/heartexlabs/label-studio). While labelling, the individual imaging channels were also inspected using Fiji (ImageJ2 v2.9.0)[64]. The YOLO models were trained and tested on GPU using a Google Colaboratory notebook (available at https://github.com/DanelonLab/Image-based-Phenotypic-Selection/tree/main/Training-YOLO), which was adapted from https://github.com/theAIGuysCode/YOLOv3-Cloud-Tutorial. Unless indicated otherwise, the training of YOLO algorithms was performed from scratch (i.e., without using pre-trained weights) using jpg images of 512 × 512 pixels. During training, saturation and exposure (brightness) augmentations were enabled, but color and rotation augmentations were disabled. The 'random' parameter was disabled to prevent up- and

downscaling of images during training. For YOLO algorithms trained on 1 class (MinD localized on the membrane), a dataset was used consisting of 700 images containing 688 labelled instances. For YOLO algorithms trained on 2 classes, 40 images were used containing 1,645 instances, 915 without membrane-localized eGFP-MinC (class 1) and 730 instances with membrane-localized eGFP-MinC (class 2). For the 2-class YOLO model (used for detection of dynamical behaviors), special care was taken to exclude (i) small liposomes (approximately <6 μm), which tend to move in and out of focus more frequently than larger liposomes, (ii) liposomes with a blurry, undefined membrane signal, and (iii) liposomes at the edges of the image. The YOLO models were trained and tested using 80% and 20% of the images, respectively. After validation of the 2-class model, a new model was trained on 100% of the data and used for the experiments.

Real-time image analysis

Real-time image analysis was performed on a HP Z4 workstation with Intel(R) Xeon(R) W-2145 CPU, 64 GB ram, NVIDIA Quadro RTX 4000 GPU and Windows 10 Enterprise operating system. To run YOLO, open-source software CUDA v10.2, cuDNN v8.0.2, OpenCV v4.2.0 and Darknet were installed. Darknet was downloaded from https://github.com/AlexeyAB/darknet and compiled using Visual Studio 2019 as described on https://github.com/AlexeyAB/darknet#requirements-for-windows-linux-and-macos.

Automated image acquisition, image analysis and laser stimulation were controlled by the JOBS module of the NIS-Elements image acquisition software (Nikon). Real-time image analysis by intensity thresholding and simple morphology constraints was performed by the General Analysis module in NIS-Elements (JOB available at https://github.com/DanelonLab/Image-based-Phenotypic-Selection/tree/main/Detection-By-Intensity-Thresholding). This module applied general convolution (size: 3), intensity thresholding, and size (3.5 to 20 μm) and circularity (0.3 to 1) constraints to the YFP channel. For object detection by YOLO, the OutProc function within JOBS exported the image to a local computer folder and initiated a Windows batch script which started YOLO and redirected the resulting coordinates back to NIS-Elements for stimulation (JOB and code available at https://github.com/DanelonLab/Image-based-Phenotypic-Selection/tree/main/Detection-By-YOLO-Single-Time-Frame).

For video-based classification, the liposomes detected in multiple time frames were identified as the same object based on the nearest neighbor with a maximum distance of 5 μm. Bounding boxes with an x-to-y ratio smaller than 0.7 or larger than 1.3 (these were typically liposomes at the edge of the FOV) and/or an area smaller than 37 μm^2 were excluded from the analysis. Liposomes whose classes changed across the different time frames were classified as dynamic and their coordinates were imported by NIS-Elements for stimulation. The JOB and code for video-based classification are available at https://github.com/DanelonLab/Image-based-

Phenotypic-Selection/tree/main/Detection-By-YOLO-Using-Videos. ChatGPT was used to assist in generating parts of the Windows batch scripts.

Flow cytometry

Liposomes were transferred from the imaging chamber to an Eppendorf tube using a cut pipette tip. The imaging chambers were washed two times with 50 µL of PB buffer, and this volume was added to the same Eppendorf tube. The sample was filtered using the 35 µm nylon mesh of a cell strainer cap of 5-mL round bottom polystyrene test tubes (Falcon). Liposomes were screened and sorted with the BD FACSMelody cell sorter (BD Biosciences) using a 100-µm nozzle, laser line 488 nm with emission filter 527/32 nm (YFP, eGFP-MinC) and laser line 561 nm with emission filter 582/15 nm (PAmCherry2). The photomultiplier tube voltages were set at 342 V for FSC, 405 V for SSC (with SSC threshold 359 V), 646 V for YFP and eGFP-MinC, and 660 V for PAmCherry2. Data were processed with BD FACSChorus version 3 (while sorting) or Cytobank (https://community.cytobank.org/) as shown in **Fig. S10**. The liposomes sorted by FACS were heated at 75 °C for 15 min to inactivate DNase I.

Quantitative PCR

Quantitative PCR reactions contained 1× PowerUP SYBR Green Master Mix (Applied Biosystems), 400 nM of each primer (1121 ChD and 1122 ChD for *YFP*, 1208 ChD and 1209 ChD for *minD*, 1414 ChD and 1415 ChD for *btubB*) and 1 µL of heat-inactivated liposome sample in a total volume of 10 µL. Quantitative PCR reactions were performed using the Quantstudio 5 Real-Time PCR system (Thermo Fisher). The samples were incubated sequentially for 2 min at 50 °C, 5 min at 94 °C, 45 cycles of 15 s at 94 °C, 15 s at 56 °C, 30 s at 68 °C, followed by 5 min at 68 °C and a melting curve from 65 to 95 °C. The DNA concentration was calculated using calibration curves ranging from 0.00001 pM to 100 pM in the Quantstudio Design and Analysis software v1.4.3 (Thermo Fisher).

DNA recovery from liposomes and Nanopore sequencing

DNA (*minD*) was recovered from liposomes by PCR. The reactions were set up at 50-µL volume and contained 400 µM of dNTPs, 300 nM of each primer (1525/1526 ChD), 0.4 units of KOD Xtreme Hot Start DNA Polymerase and 2 µL of heat-inactivated liposome sample in 1× KOD Xtreme PCR Buffer. The thermocycling reaction was performed as follows: 94 °C for 2 min, (98 °C for 10 s, 56 °C for 30 s, 68 °C for 84 s) × 39. The amplicons were purified with the QIAquick PCR Cleanup Kit (Qiagen). Recovery of the DNA (1,398 bp) from liposomes was verified using gel electrophoresis and the DNA concentration was measured by the Nanodrop 2000c spectrophotometer (Isogen Life Science). Nanopore sequencing was performed by Plasmidsaurus. Using the Galaxy web platform[66] (https://usegalaxy.org), the raw sequencing reads were filtered based on size (1,100-1,400 bp) and mapped to the reference sequence using the *Filter FASTQ reads by quality score and length*[67] and *Map with minimap2*[68] tools, respectively, using default parameters. The mapped reads were further processed with in-house written R scripts run in Rstudio (R v4.4.1)

(https://github.com/DanelonLab/MOSAIC/tree/main/Application-MinD-MTS-Library). The DNA sequences encoding the MTS of MinD were extracted from the mapped reads and subjected to a quality control, whereafter the different DNA variants amongst those sequences were counted.

Statistical analysis

Error bars in bar plots indicate the standard deviation (SD). The box-whisker plots in **Fig. 2A and Fig. S1** display the median (middle line), the 75/25 percentiles at the boxes and the 90/10 at the whiskers. Statistical tests in **Fig. S1** (Mann-Whitney tests, correlation analysis, linear regression, unpaired t tests) were performed using GraphPad Prism v10.0.3 and were significant if *P* value (two-tailed) < 0.05. Pearson's *r* is the Pearson correlation coefficient. R^2 is the coefficient of determination. The number of biological replicates (i.e., distinct liposome samples) and technical replicates (repeated measurements) are indicated in the figure legends. The brightness and contrast of microscopy images were adjusted identically for multiple images when comparing fluorescence intensities using Fiji (ImageJ2 v2.9.0)[64].

Data, code and materials availability

All data are available in the main text, the Supplementary Information or Supplementary Data 1 (Benchmarking overview). Supplementary Information includes confocal fluorescence images, FACS plots, qPCR data, performance of object detection models, agarose gel electrophoresis, lists of primers and plasmids, tables with experimental details, and DNA sequences. Raw and processed data underlying this article are available in the 4TU.ResearchData repository with DOI: https://doi.org/10.4121/60cdce27-5bd5-44ec-9d8f-53f2ef6b9d58. The codes used in this article are available in the 4TU.ResearchData repository with DOI: https://doi.org/10.4121/60cdce27-5bd5-44ec-9d8f-53f2ef6b9d58, as well as on GitHub at https://github.com/DanelonLab under the MIT license, allowing for free use, modification, and distribution. Correspondence and requests for materials should be addressed to Christophe Danelon.

Acknowledgements

PAmCherry2.1 gene was a kind gift from the Van der Oost group (Wageningen University, the Netherlands). G571 (pRSET-mEos2) was a kind gift from Kristin Grüßmayer (TU Delft, the Netherlands). We thank Ilja Westerlaken for the purification of PAmCherry2 and mEos2. We thank Rosalie Knot for preliminary experimental work, and Mats van Tongeren for performing preliminary tests with YOLO. We are also thankful to Michal Shemesh, Martin Holub, Timo Kuijt (Nikon Instruments) and the technical support team from Laboratory Imaging s.r.o. for their help with the microscopy automation, Thierry Langlois and Damien Loras from BD Life Sciences – Biosciences for fruitful discussions and testing the BD FACSDiscover S8, and Miao-Ping Chien and the Chien lab for testing their ultrawide-FOV microscope with our liposome samples. MvdB, CD, MS and NJC acknowledge financial support from the "BaSyC – Building a

Synthetic Cell" Gravitation grant (024.003.019) of the Netherlands Ministry of Education, Culture and Science (OCW) and the Netherlands Organisation for Scientific Research (NWO). CD acknowledges funding from Agence Nationale de la Recherche (ANR-22-CPJ2-0091-01). NJC acknowledges funding by the NWO-SUMMIT grant "Evolving life from non-life (EVOLF)" (SUMMIT.1.004).

Author contributions

Conceptualization: C.D. and M.v.d.B. Methodology: C.D., M.v.d.B. and M.S. Investigation: M.v.d.B. and M.S. Formal analysis: M.v.d.B. Validation: C.D. and M.v.d.B. Software: M.v.d.B. Visualization: M.v.d.B. Resources: C.D. Data curation: M.v.d.B and C.D. Supervision: C.D. and N.J.C. Writing–original draft: M.v.d.B. and C.D. Writing–review and editing: C.D., N.J.C. and M.v.d.B. Project administration: C.D. Funding acquisition: C.D. and N.J.C.

References

1. van Nies, P. *et al.* Self-replication of DNA by its encoded proteins in liposome-based synthetic cells. *Nat. Commun.* **9**, 1583 (2018).
2. Nomura, S. M. *et al.* Gene expression within cell-sized lipid vesicles. *ChemBioChem* **4**, 1172–1175 (2003).
3. Noireaux, V. & Libchaber, A. A vesicle bioreactor as a step toward an artificial cell assembly. *PNAS* **101**, 17669–17674 (2004).
4. Saito, H. *et al.* Time-resolved tracking of a minimum gene expression system reconstituted in giant liposomes. *ChemBioChem* **10**, 1640–1643 (2009).
5. Scott, A. *et al.* Cell-free phospholipid biosynthesis by gene-encoded enzymes reconstituted in liposomes. *PLOS ONE* **11**, e0163058 (2016).
6. Blanken, D., Foschepoth, D., Serrão, A. C. & Danelon, C. Genetically controlled membrane synthesis in liposomes. *Nat. Commun.* **11**, 4317 (2020).
7. Eto, S. *et al.* Phospholipid synthesis inside phospholipid membrane vesicles. *Commun. Biol.* **5**, 1–11 (2022).
8. Bhattacharya, A., Brea, R. J., Niederholtmeyer, H. & Devaraj, N. K. A minimal biochemical route towards de novo formation of synthetic phospholipid membranes. *Nat. Commun.* **10**, 300 (2019).
9. Furusato, T. *et al.* De novo synthesis of basal bacterial cell division proteins FtsZ, FtsA, and ZipA inside giant vesicles. *ACS Synth. Biol.* **7**, 953–961 (2018).
10. Godino, E. *et al.* De novo synthesized Min proteins drive oscillatory liposome deformation and regulate FtsA-FtsZ cytoskeletal patterns. *Nat. Commun.* **10**, 4969 (2019).
11. Godino, E. *et al.* Cell-free biogenesis of bacterial division proto-rings that can constrict liposomes. *Commun. Biol.* **3**, 539 (2020).
12. Kattan, J., Doerr, A., Dogterom, M. & Danelon, C. Shaping liposomes by cell-free expressed bacterial microtubules. *ACS Synth. Biol.* **10**, 2447–2455 (2021).
13. Kohyama, S., Merino-Salomón, A. & Schwille, P. In vitro assembly, positioning and contraction of a division ring in minimal cells. *Nat. Commun.* **13**, 6098 (2022).
14. De Franceschi, N., Barth, R., Meindlhumer, S., Fragasso, A. & Dekker, C. Dynamin A as a one-component division machinery for synthetic cells. *Nat. Nanotechnol.* **19**, 70–76 (2024).
15. Godino, E., Restrepo Sierra, A. M. & Danelon, C. Imaging flow cytometry for high-throughput phenotyping of synthetic cells. *ACS Synth. Biol.* **12**, 2015–2028 (2023).
16. Abil, Z. & Danelon, C. Roadmap to building a cell: an evolutionary approach. *Front. Bioeng. Biotechnol.* **8**, (2020).
17. Packer, M. S. & Liu, D. R. Methods for the directed evolution of proteins. *Nat. Rev. Genet.* **16**, 379–394 (2015).

18. Abil, Z. *et al.* Darwinian evolution of self-replicating DNA in a synthetic protocell. *Nat. Commun.* **15**, 9091 (2024).
19. Fujii, S. *et al.* Liposome display for in vitro selection and evolution of membrane proteins. *Nat. Protoc.* **9**, 1578–1591 (2014).
20. Sunami, T. *et al.* Femtoliter compartment in liposomes for in vitro selection of proteins. *Anal. Biochem.* **357**, 128–136 (2006).
21. Nishikawa, T., Sunami, T., Matsuura, T., Ichihashi, N. & Yomo, T. Construction of a gene screening system using giant unilamellar liposomes and a fluorescence-activated cell sorter. *Anal. Chem.* **84**, 5017–5024 (2012).
22. Fujii, S., Matsuura, T., Sunami, T., Kazuta, Y. & Yomo, T. In vitro evolution of α-hemolysin using a liposome display. *PNAS* **110**, 16796–16801 (2013).
23. Uyeda, A. *et al.* Liposome-based in vitro evolution of aminoacyl-tRNA synthetase for enhanced pyrrolysine derivative incorporation. *ChemBioChem* **16**, 1797–1802 (2015).
24. Uyeda, A., Nakayama, S., Kato, Y., Watanabe, H. & Matsuura, T. Construction of an in vitro gene screening system of the E. coli EmrE transporter using liposome display. *Anal. Chem.* **88**, 12028–12035 (2016).
25. Murase, Y., Nakanishi, H., Tsuji, G., Sunami, T. & Ichihashi, N. In vitro evolution of unmodified 16S rRNA for simple ribosome reconstitution. *ACS Synth. Biol.* **7**, 576–583 (2018).
26. Walton, R. T., Singh, A. & Blainey, P. C. Pooled genetic screens with image-based profiling. *Mol. Syst. Biol.* **18**, e10768 (2022).
27. Schraivogel, D. *et al.* High-speed fluorescence image–enabled cell sorting. *Science* **375**, 315–320 (2022).
28. Paulsen, M. S. Image-enabled cell sorting using the BD CellView Technology. in *Flow Cytometry Protocols* (eds. Hawley, T. S. & Hawley, R. G.) 145–158 (Springer US, New York, NY, 2024). doi:10.1007/978-1-0716-3738-8_8.
29. Isozaki, A. *et al.* Intelligent image-activated cell sorting 2.0. *Lab Chip* **20**, 2263–2273 (2020).
30. Kuhn, T. M., Paulsen, M. & Cuylen-Haering, S. Accessible high-speed image-activated cell sorting. *Trends Cell Biol.* **34**, 657–670 (2024).
31. Barteneva, N. S., Fasler-Kan, E. & Vorobjev, I. A. Imaging flow cytometry: coping with heterogeneity in biological systems. *J. Histochem. Cytochem.* **60**, 723–733 (2012).
32. Hasle, N. *et al.* High-throughput, microscope-based sorting to dissect cellular heterogeneity. *Mol. Syst. Biol.* **16**, e9442 (2020).
33. Lee, J. *et al.* Versatile phenotype-activated cell sorting. *Sci. Adv.* **6**, eabb7438 (2020).
34. Yan, X. *et al.* High-content imaging-based pooled CRISPR screens in mammalian cells. *J. Cell Biol.* **220**, e202008158 (2021).
35. Kanfer, G. *et al.* Image-based pooled whole-genome CRISPRi screening for subcellular phenotypes. *J. Cell Biol.* **220**, e202006180 (2021).
36. You, L. *et al.* Linking the genotypes and phenotypes of cancer cells in heterogenous populations via real-time optical tagging and image analysis. *Nat. Biomed. Eng.* **6**, 667–675 (2022).
37. Tian, H. *et al.* Video-based pooled screening yields improved far-red genetically encoded voltage indicators. *Nat. Methods* **20**, 1082–1094 (2023).
38. Lin, C., Liu, L. & Zou, P. Functional imaging-guided cell selection for evolving genetically encoded fluorescent indicators. *Cell Rep. Methods* **3**, 100544 (2023).
39. Shimizu, Y. *et al.* Cell-free translation reconstituted with purified components. *Nat. Biotechnol.* **19**, 751–755 (2001).
40. Moen, E. *et al.* Deep learning for cellular image analysis. *Nat. Methods* **16**, 1233–1246 (2019).
41. Chai, B., Efstathiou, C., Yue, H. & Draviam, V. M. Opportunities and challenges for deep learning in cell dynamics research. *Trends Cell Biol.* **34**, 955–967 (2024).
42. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. Preprint at http://arxiv.org/abs/1506.02640 (2016).
43. Redmon, J. & Farhadi, A. YOLOv3: an incremental improvement. Preprint at https://doi.org/10.48550/arXiv.1804.02767 (2018).

44. Wang, C.-Y., Bochkovskiy, A. & Liao, H.-Y. M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. Preprint at https://doi.org/10.48550/arXiv.2207.02696 (2022).
45. Litschel, T., Ramm, B., Maas, R., Heymann, M. & Schwille, P. Beating vesicles: encapsulated protein oscillations cause dynamic membrane deformations. *Angew. Chem. Int. Ed.* **57**, 16286–16290 (2018).
46. Rowlett, V. W. & Margolin, W. The Min system and other nucleoid-independent regulators of Z ring positioning. *Front. Microbiol.* **6**, (2015).
47. Szeto, T. H., Rowland, S. L., Rothfield, L. I. & King, G. F. Membrane localization of MinD is mediated by a C-terminal motif that is conserved across eubacteria, archaea, and chloroplasts. *PNAS* **99**, 15693–15698 (2002).
48. Seo, K., Hagino, K. & Ichihashi, N. Progresses in cell-free in vitro evolution. in *Cell-free Production: System Development* (eds. Lu, Y. & Jewett, M. C.) 121–140 (Springer Nature Switzerland, Cham, 2023). doi:10.1007/10_2023_219.
49. Abil, Z., Restrepo Sierra, A. M. & Danelon, C. Clonal amplification-enhanced gene expression in synthetic vesicles. *ACS Synth. Biol.* **12**, 1187–1203 (2023).
50. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687–694 (2019).
51. Pandi, A. *et al.* A versatile active learning workflow for optimization of genetic and metabolic networks. *Nat. Commun.* **13**, 3876 (2022).
52. Blanken, D., van Nies, P. & Danelon, C. Quantitative imaging of gene-expressing liposomes reveals rare favorable phenotypes. *Phys. Biol.* **16**, 045002 (2019).
53. Werley, C. A., Chien, M.-P. & Cohen, A. E. Ultrawidefield microscope for high-speed fluorescence imaging and targeted optogenetic stimulation. *Biomed. Opt. Express* **8**, 5794–5813 (2017).
54. Peterman, N. & Levine, E. Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations. *BMC Genomics* **17**, 206 (2016).
55. Nourian, Z., Roelofsen, W. & Danelon, C. Triggered gene expression in fed-vesicle microreactors with a multifunctional membrane. *Angew. Chem. Int. Ed.* **51**, 3114–3118 (2012).
56. Van de Cauter, L. *et al.* Optimized cDICE for efficient reconstitution of biological systems in giant unilamellar vesicles. *ACS Synth. Biol.* **10**, 1690–1702 (2021).
57. Gonzales, D. T., Yandrapalli, N., Robinson, T., Zechner, C. & Tang, T.-Y. D. Cell-free gene expression dynamics in synthetic cell populations. *ACS Synth. Biol.* **11**, 205–215 (2022).
58. Su, P.-R. *et al.* Microscopy-based single-cell proteomic profiling reveals heterogeneity in DNA damage response dynamics. *Cell Rep. Methods* **2**, 100237 (2022).
59. Bennett, H. M., Stephenson, W., Rose, C. M. & Darmanis, S. Single-cell proteomics enabled by next-generation sequencing or mass spectrometry. *Nat. Methods* **20**, 363–374 (2023).
60. Pratapa, A., Doron, M. & Caicedo, J. C. Image-based cell phenotyping with deep learning. *Curr. Opin. Chem. Biol.* **65**, 9–17 (2021).
61. Olivi, L. *et al.* Live-cell imaging reveals the trade-off between target search flexibility and efficiency for Cas9 and Cas12a. *Nucleic Acids Res.* **52**, 5241–5256 (2024).
62. Loose, M., Fischer-Friedrich, E., Herold, C., Kruse, K. & Schwille, P. Min protein patterns emerge from rapid rebinding and membrane interaction of MinE. *Nat. Struct. Mol. Biol.* **18**, 577–583 (2011).
63. Galka, P., Jamez, E., Joachim, G. & Soumillion, P. QuickLib, a method for building fully synthetic plasmid libraries by seamless cloning of degenerate oligonucleotides. *PLOS ONE* **12**, e0175146 (2017).
64. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
65. Tkachenko, M., Malyuk, M., Holmanyuk, A. & Liubimov, N. Label Studio: Data labeling software. (2020).
66. The Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res.* **50**, W345–W351 (2022).
67. Blankenberg, D. *et al.* Manipulation of FASTQ data with Galaxy. *Bioinformatics* **26**, 1783–1785 (2010).
68. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

Supplementary Materials

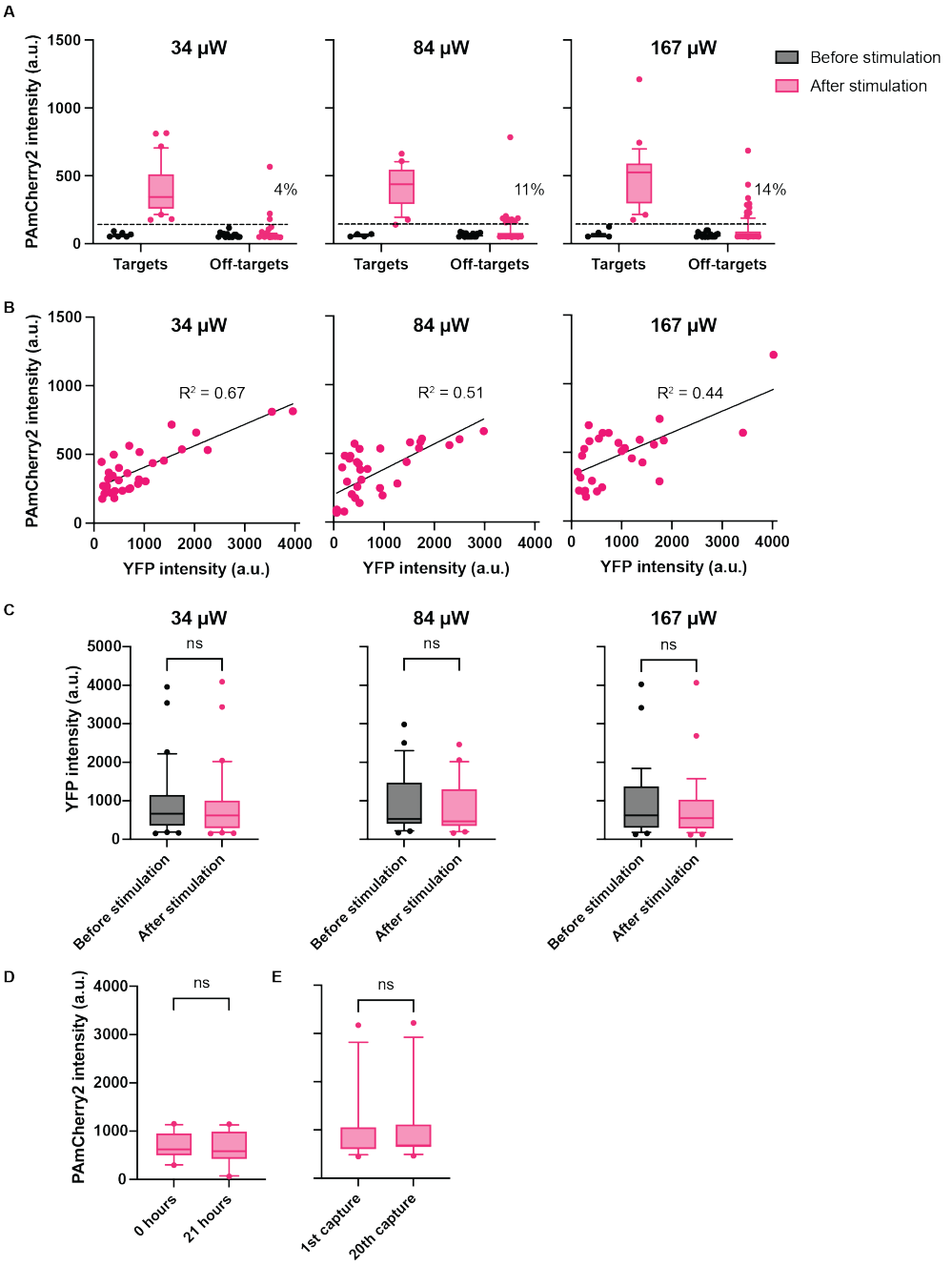


Fig. S1. PAmCherry2 and YFP fluorescence intensities before and after liposome stimulation. YFP was expressed in liposomes containing 5 nM *YFP* DNA template and 20 µM purified PAmCherry2. Liposomes were stimulated in the center of the FOV for 300 ms using different laser powers (34, 84 and 167 µW). The fluorescence intensity after stimulation was measured 10 s after stimulation. (**A**) PAmCherry2

fluorescence intensity in target and off-target liposomes before and after stimulation. All target liposomes were successfully tagged for the three photoactivation conditions with no significant differences in the median intensity values (Mann-Whitney test, $P = 0.38$ for 34 µW vs 84 µW; $P = 0.11$ for 34 µW vs 167 µW). All liposomes within 15 µm of the target liposome were considered as off-targets. Per condition, 29-31 target liposomes and 78-94 off-target liposomes were measured. The percentages of off-target liposomes with increased PAmCherry2 intensity (above a certain threshold indicated with the dashed line) relative to all off-target liposomes are appended in the graphs. The data for the target liposomes of the 84 µW condition are also shown in Fig. 2A. (**B**) Positive correlation between the YFP intensity (before stimulation) and the PAmCherry2 intensity after stimulation (Pearson's $r = 0.82$ for 34 µW; $r = 0.72$ for 84 µW; $r = 0.67$ for 167 µW; $P < 0.0001$ for all). A linear regression curve is shown in the graphs with the corresponding R^2 value. Per condition, 29-31 target liposomes were measured. For the 84 µW condition, ten additional liposomes were measured with low YFP intensity (< 65 a.u.) and low PAmCherry2 intensity (< 95 a.u.) values. (**C**) YFP intensities before and after stimulation were not significantly different (unpaired t test, $P = 0.88$, $t = 0.1540$, degrees of freedom = 60 for 34 µW; $P = 0.50$, $t = 0.6778$, degrees of freedom = 56 for 84 µW; $P = 0.52$, $t = 0.6528$, degrees of freedom = 56 for 167 µW), thus no significant YFP bleaching occurred. Per condition, 29-31 target liposomes were measured. (**D**) PAmCherry2 fluorescence values are constant up to 21 hours in liposomes. Twelve YFP-expressing liposomes were activated (34 µW laser power, 300 ms) and tracked over time by imaging every hour. PAmCherry2 intensities at t0 and t21 h were not significantly different (unpaired t test, $P = 0.71$, $t = 0.3782$, degrees of freedom = 22). (**E**) Control experiment for panel d. PAmCherry2 was not bleached due to repeated imaging (20 captures within 30 min) (unpaired t test, $P = 0.84$, $t = 0.2027$, degrees of freedom = 22). Twelve target liposomes were measured.

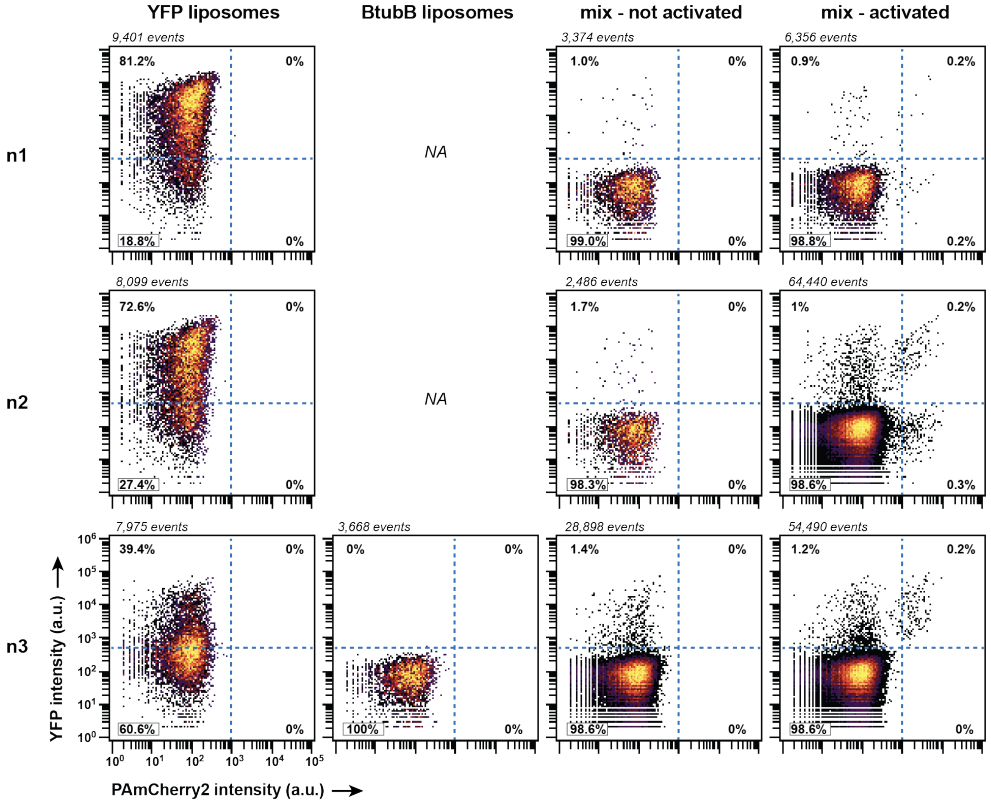


Fig. S2. Flow cytometry measurements of (a mix of) YFP- and BtubB-expressing liposomes, without or with activation of PAmcherry2. This figure supplements Fig. 2E with three biological replicates. The two right-most plots from replicate 3 (n3) are also shown in Fig. 2E. The number of events varies per plot as indicated above the graphs. The percentage of liposomes in each gate is also indicated. Experiment 3 (n3) was performed with laser power 34 µW and stimulation in the center of the FOV, while experiments 1 and 2 (n1-2) were performed with laser power 84 µW and stimulation was not restricted to the center of the FOV, likely explaining the lower stimulation precision for experiments 1-2.

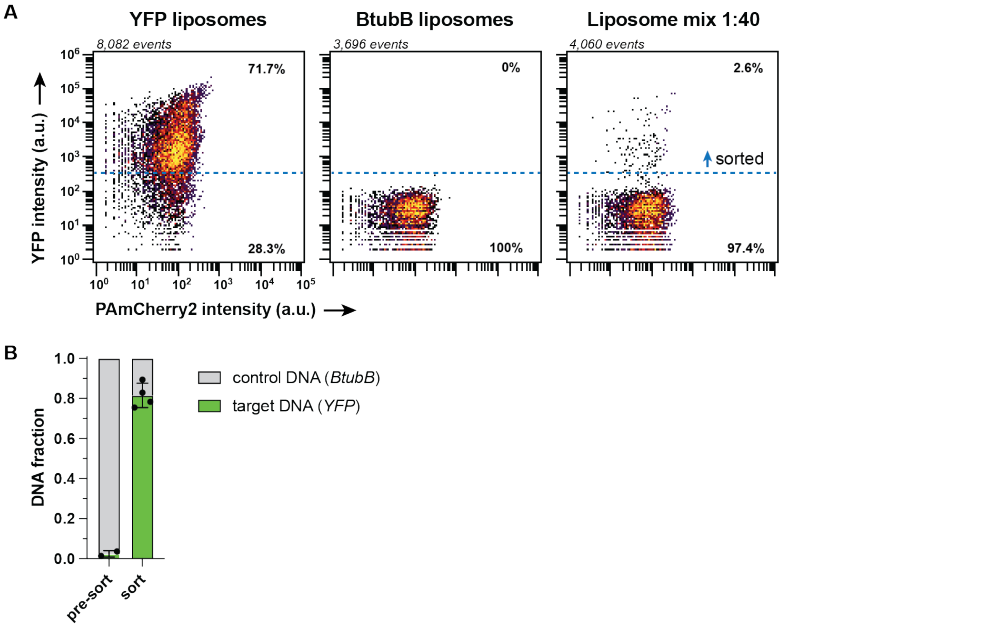


Fig. S3. FACS sorting of YFP-expressing liposomes from a mixed population of YFP- and BtubB-expressing liposomes based on YFP fluorescence intensity. (**A**) FACS plots showing YFP intensity vs PAmCherry2 intensity for one of two biological replicates. The second replicate corresponds to the left three plots of n3 in Fig. S2. Number of events are indicated above the plots. The percentage of liposomes in each gate is also indicated. (**B**) qPCR analysis of the sorted liposomes showing enrichment of YFP DNA upon sorting. Two biological replicates (mixed liposome samples) with two technical replicates (sorts) each.

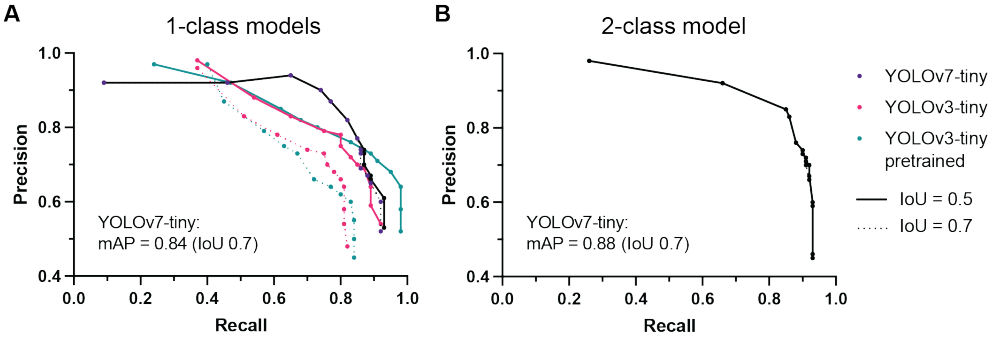


Fig. S4. Performance of trained YOLO-tiny models. (**A**) YOLO-tiny models trained on one class to detect liposomes with eGFP-MinC at the membrane. mAP is mean Average Precision, and IoU is Intersection over Union. For YOLOv7-tiny (trained from scratch), YOLOv3-tiny (trained from scratch) and YOLOv3-tiny (pretrained), the mAP is 0.85, 0.85 and 0.87 for IoU 0.5, and 0.84, 0.73 and 0.72 for IoU 0.7, respectively. The mAP remains very similar for YOLOv7-tiny when increasing the IoU from 0.5 to 0.7, while for YOLOv3-tiny the mAP decreases. Therefore, the YOLOv7-tiny model has better localization performance of the bounding boxes than the YOLOv3-tiny models. The trained YOLOv7-tiny model was used in liposome experiments at a confidence threshold of 0.8, with both a precision and recall of 82% (IoU 0.5 and 0.7). The YOLOv7-tiny data are also shown in Fig. 3C. (**B**) YOLOv7-tiny model trained on two classes to detect liposomes with eGFP-MinC at the membrane, and liposomes with eGFP-MinC in the lumen. mAP is 0.88 for both IoU 0.5 and 0.7. The model was used at a threshold of 0.9 (precision 83%, recall 86%).

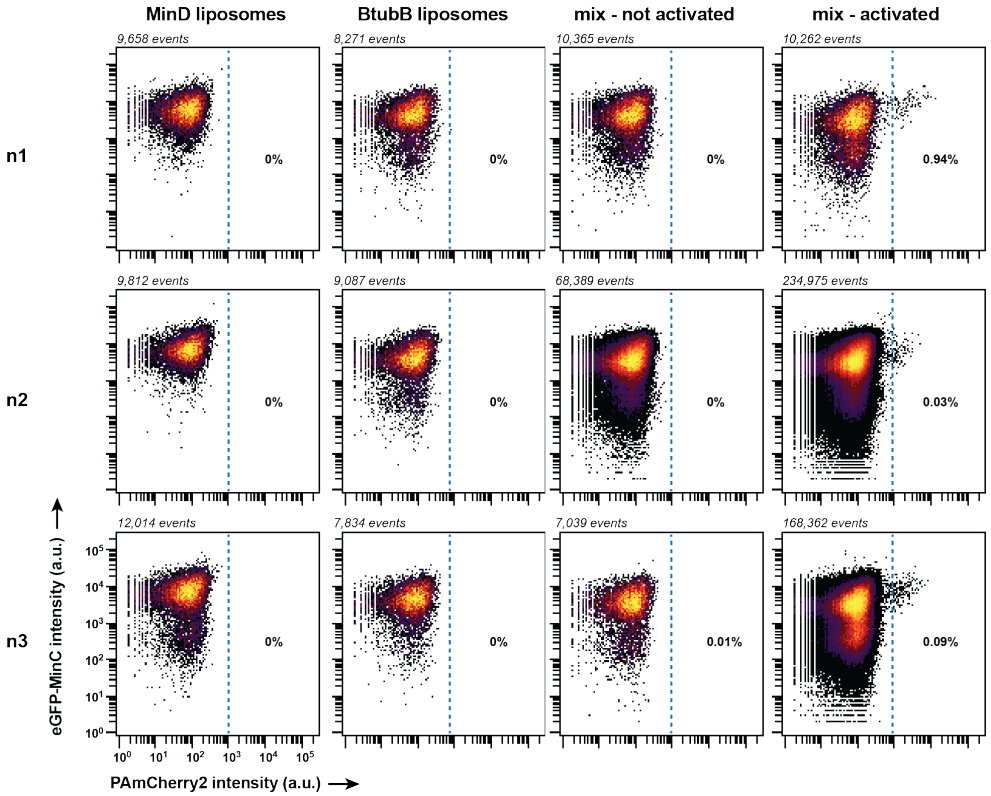


Fig. S5. Flow cytometry measurements of (a mix of) MinD- and BtubB-expressing liposomes, without or with activation of PAmcherry2. This figure supplements Fig. 3E with three biological replicates. The two right-most plots from replicate 1 (n1) are also shown in Fig. 3E. The number of events varies per plot as indicated above the graphs. The percentage of liposomes in each gate is also indicated.

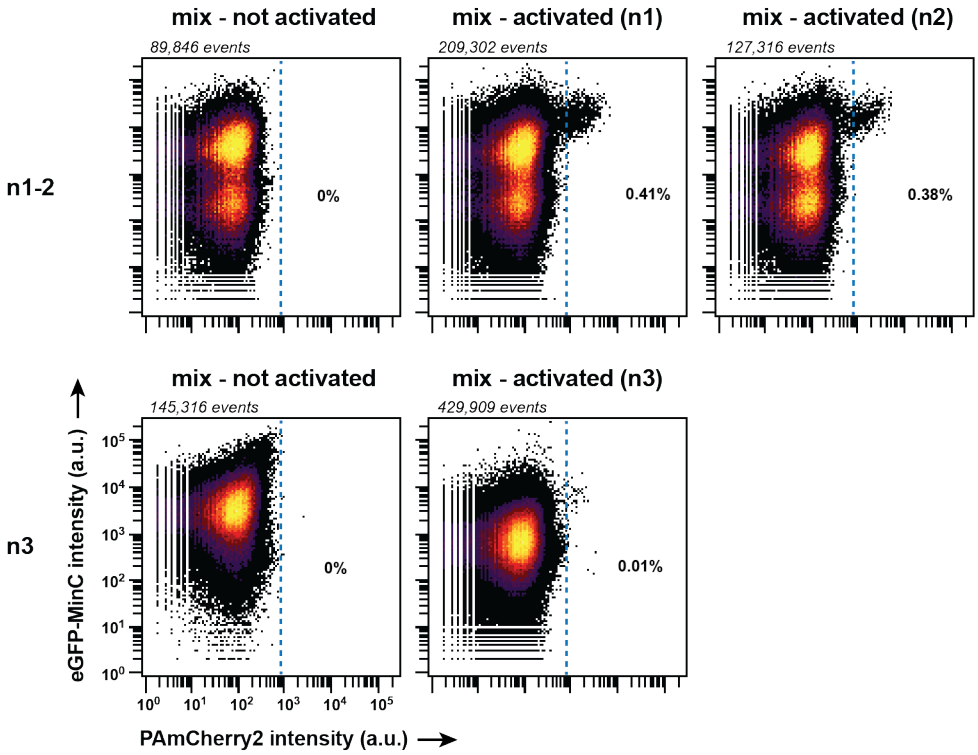


Fig. S6. Flow cytometry measurements of liposomes expressing a DNA library encoding MinD with mutations in the MTS, without or with activation of PAmcherry2. This figure supplements Fig. 5. Two biological replicates, for which the first replicate entailed two independent sorts (n1 and n2) and the second replicate entailed one sort (n3). The number of events varies per plot as indicated above the graphs. The percentage of liposomes in each gate is also indicated.

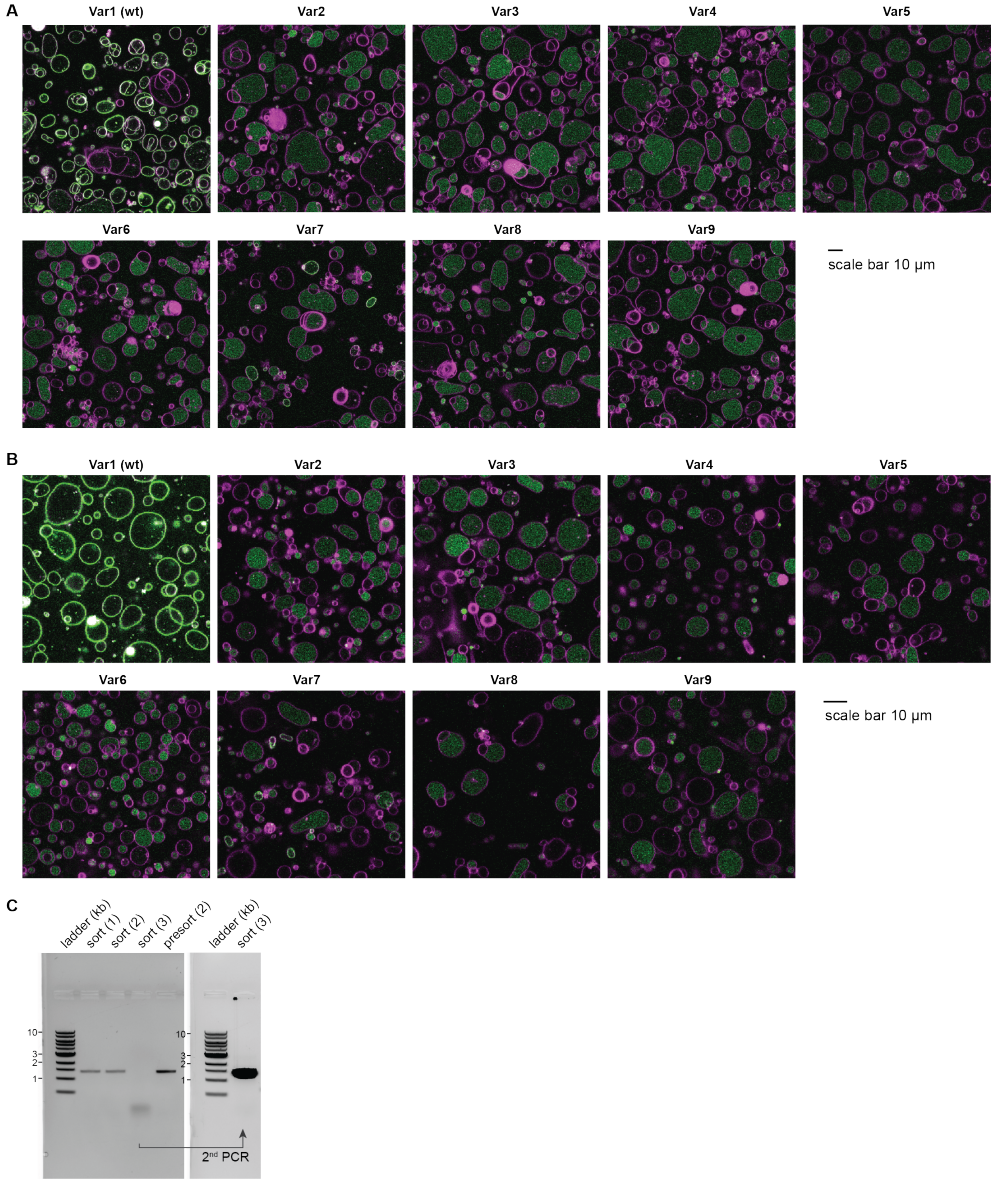


Fig. S7. Additional data for the library of MinD MTSs. (**A-B**) Fluorescence confocal microscopy images of liposomes expressing clonal DNA variants with mutations in the MinD MTS. Green, eGFP-MinC; magenta, Cy5-conjugated lipids. Scale bars: 10 µm. In replicate 1 (**A**), the liposomes were diluted 15-fold before imaging using PB buffer. In replicate 2 (**B**), the liposomes were diluted 15-fold using a 1:1 mixture of PURE*frex* 2.0 Solution I and MilliQ. The images for Var1 (wt) and Var 7 in replicate 1 (A) are also presented in Fig. 5E. (**C**) Gel electrophoresis with PCR-amplified *minD* (1,398 bp) recovered from sorted liposomes. For sort 3, the full-length DNA was recovered using two subsequent, identical PCRs, where the second PCR contained 1 µL of the first PCR as the source of DNA template instead of the sorted liposomes.

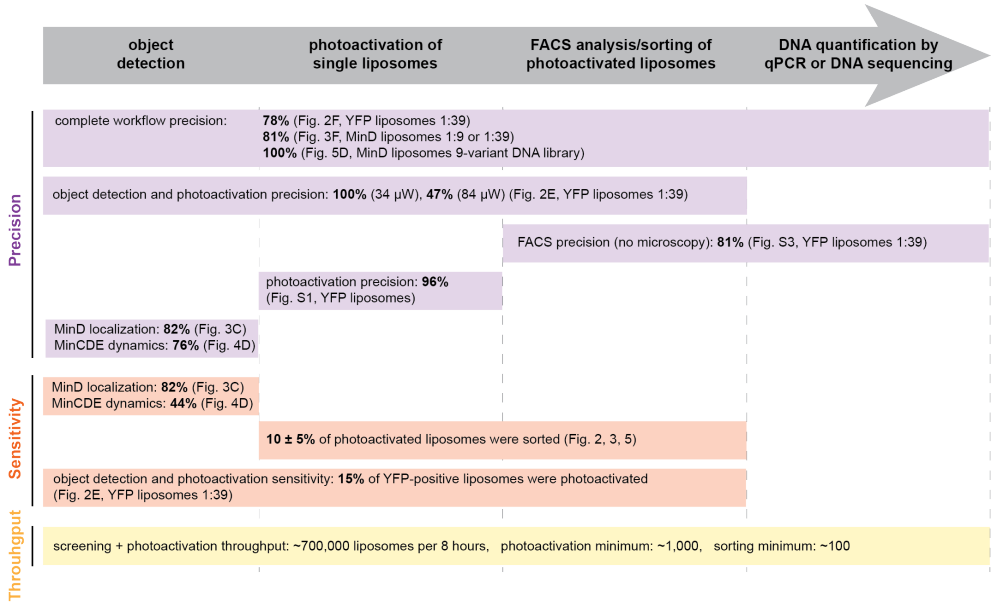


Fig. S8. Precision, sensitivity and throughput parameters of the method presented in this work. The precision, sensitivity and throughput were extensively quantified for the different stages in the workflow. A comparison of the metrics to those of other studies is provided as separate supplementary material: Supplementary Data 1.

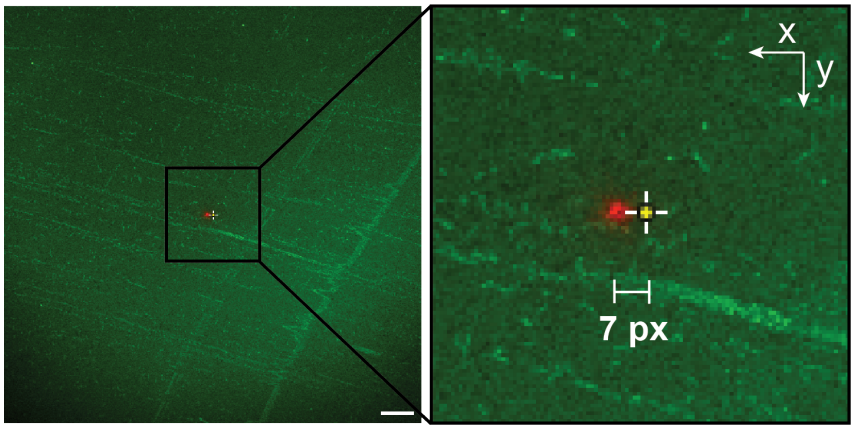


Fig. S9. Difference between the target and observed location of point-stimulation. mEos2 protein was imaged at the glass surface of a cover slip and point-stimulated in the center pixel of the field of view (405 nm, 300 ms, 5 µW laser power). The difference in xy location between the target stimulation point (yellow dot) and the observed green-to-red photoconversion and bleaching of mEos2 was quantified (here: 7 pixels in the x-direction, i.e., 1.75 µm). The differences in x and y locations were corrected for in the JOBS protocol (Nis-Elements software). Scale bar: 10 µm.

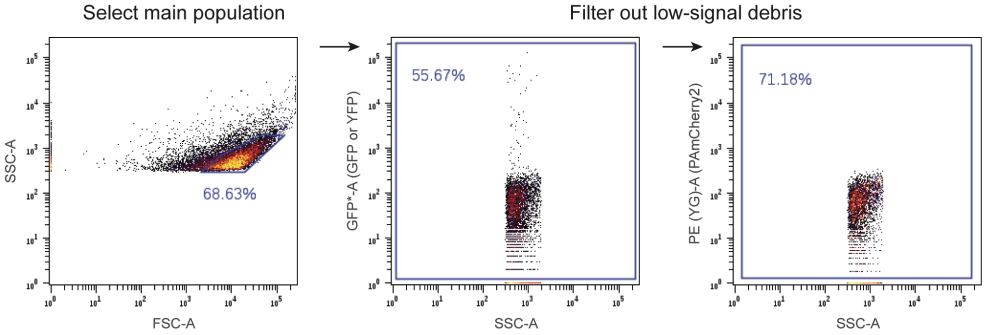


Fig. S10. Example of gating strategy applied to FACS data. While sorting liposomes, only the first step (gating based on forward and side scatter light) was applied. Post-measurement analysis involved all three filtering steps.

Table S1. Experimental details for photoselection and sorting of liposomes from *YFP-btubB* mock library. The number of stimulated liposomes was estimated by counting the stimulated liposomes for a fraction of all images.

Exp	Exp ID	Stim at center FOV	Laser power (µW)	Stim time (ms)	Correction xy point stimulation	Number of FOVs screened	Number of stimulated liposomes	FACS sorting threshold	Number of sorted liposomes	Initial ratio *YFP*: *btubB*
1	20230713	no	84	300	no	1,914	870	1,000	193	1:40
2	20230721	no	84	300	no	4,687	7,218	1,800	856	1:40
3	20240429	yes	34	300	yes, 7 px in x direction	3,440	1,668	910	154	1:50

Table S2. Percentages of gated liposomes (extended data for Fig. S2). Gating thresholds for PAmCherry2 signal and YFP were 988 a.u. and 542 a.u., respectively. ROI 1-4 are the top left, top right, bottom left and bottom right gates, respectively.

	YFP liposomes				BtubB liposomes			
Exp	**ROI 1**	**ROI 2**	**ROI 3**	**ROI 4**	**ROI 1**	**ROI 2**	**ROI 3**	**ROI 4**
1	81.16%	0.01%	18.83%	0.00%	*NA*	*NA*	*NA*	*NA*
2	72.63%	0.00%	27.37%	0.00%	*NA*	*NA*	*NA*	*NA*
3	39.39%	0.00%	60.61%	0.00%	0.00%	0.00%	100.00%	0.00%

Continuation Table S2

	Mix - not activated				Mix - activated			
Exp	**ROI 1**	**ROI 2**	**ROI 3**	**ROI 4**	**ROI 1**	**ROI 2**	**ROI 3**	**ROI 4**
1	0.98%	0.00%	99.02%	0.00%	0.93%	0.17%	98.75%	0.15%
2	1.73%	0.00%	98.27%	0.00%	1.02%	0.18%	98.55%	0.25%
3	1.38%	00.0%	98.62	0.00%	1.21%	0.22%	98.57%	0.00%

Table S3. DNA concentrations in the sorted fractions of the *YFP-btubB* mock library as measured by qPCR.

	[*btubB*] (pM)		[*YFP*] (pM)		*YFP* fraction	
Exp	**Pre-sort**	**Post-sort**	**Pre-sort**	**Post-sort**	**Pre-sort**	**Post-sort**
1	0.030018	0.000287	0.000756	0.001045	0.0245	0.7846
2	0.030467	0.000754	0.001257	0.001159	0.0396	0.6060
3	0.101736	0.000003[$]	0.001403	0.000051	0.0136	0.83507* -0.9525[$]

* when taking the lowest calibration curve data point (0.00001 pM)
[$] if calibration curve was forced to extend below the lowest calibration data point (0.00001 pM)

Table S4. Experimental details for sorting liposomes from *YFP-btubB* mock library based on the YFP signal (liposomes were not stimulated in microscope). Data consist of 2 biological replicates (mock library liposome mixes), each containing 2 technical sorting replicates.

Exp	**Exp ID**	**FACS sorting threshold**	**Number of sorted liposomes**	**Initial ratio *YFP*:*btubB***
1	20230912	350	7,021	1:40
		350	5,006	
2	20240429	418	513	1:50
		418	1,338	

Table S5. Percentages of gated liposomes (extended data for Fig. S3). Gating thresholds for PAmCherry2 signal and YFP were 988 a.u. and 542 a.u., respectively. ROI 1-4 are the top left, top right, bottom left and bottom right gates, respectively.

	YFP liposomes				**BtubB liposomes**				**Mix - not activated**			
Exp	**ROI 1**	**ROI 2**	**ROI 3**	**ROI 4**	**ROI 1**	**ROI 2**	**ROI 3**	**ROI 4**	**ROI 1**	**ROI 2**	**ROI 3**	**ROI 4**
1	71.74%	0.00%	28.26%	0.00%	0.00%	0.00%	100.00%	0.00%	2.62%	0.00%	97.38%	0.00%
2	39.39%	0.00%	60.61%	0.00%	0.00%	0.00%	100.00%	0.00%	1.38%	00.0%	98.62%	0.00%

Table S6. DNA concentrations in the sorted fractions of the YFP-BtubB mock library as measured by qPCR. Liposomes were not stimulated in microscope but sorted based on the YFP signal.

	[*btubB*] (pM)		[*YFP*] (pM)		*YFP* fraction	
Exp	**Pre-sort**	**Post-sort**	**Pre-sort**	**Post-sort**	**Pre-sort**	**Post-sort**
1	0.190905	0.002459	0.007178	0.008871	0.0362	0.7829
		0.001243		0.010397		0.8932
2	0.101736	0.000012	0.001403	0.000037	0.0136	0.7542
		0.000019		0.000090		0.8289

Table S7. Experimental details for photoselection and sorting of liposomes from *minD-btubB* mock library. The number of stimulated liposomes for experiment 1 was estimated by counting the stimulated liposomes for a fraction of all images. The exact numbers are presented for experiments 2 and 3.

Exp	Exp ID	Stim at center FOV	Laser power (µW)	Stim time (ms)	Correction xy point stimulation	Number of FOVs screened	Number of stimulated liposomes	FACS sorting threshold	Number of sorted liposomes	Initial ratio *minD*:*btubB*
1	2023 1206	yes	84	300	Yes, 7 px in x direction	2,880	8,640	1,000	905	1:9
2	2024 0503	yes	34	300	Yes, 7 px in x direction	5,160	1,797	700	171	1:40
3	2024 0508	yes	34	300	Yes, 7 px in x direction	6,880	3,893	700	189	1:40

Table S8. Percentages of gated liposomes (extended data for Fig. S5). Gate > 988 a.u (PAmCherry2 signal). Percentages based on dataset with most events.

Exp	MinD liposomes	BtubB liposomes	Mix - not activated	Mix - activated
1	0.00%	0.00%	0.00%	0.94%
2	0.00%	0.00%	0.00%	0.03%
3	0.00%	0.00%	0.01%	0.09%

Table S9. DNA concentrations in the sorted fractions of the *minD-btubB* mock library as measured by qPCR.

	[*btubB*] (pM)		[*minD*] (pM)		*minD* fraction	
Exp	Pre-sort	Post-sort	Pre-sort	Post-sort	Pre-sort	Post-sort
1	0.031877	0.000262	0.005609	0.000437	0.1496	0.6253
2	0.244735	0.000030	0.007848	0.000144	0.0311	0.8274
3	0.399657	0.000005	0.016422	0.000535	0.0395	0.9817*

* when taking the lowest calibration curve data point (0.00001 pM)

Table S10. Experimental details for photoselection of liposomes displaying Min oscillations.

Exp	Exp ID	Stim at center FOV	Laser power (µW)	Stim time (ms)	Correction xy point stimulation	Number of FOVs screened	Number of stimulated liposomes	Time of experiment (h)
1	2024 1009	yes	34	300	Yes, 11 px in x direction	170	89	3.3
2	2024 1113	yes	34	300	Yes, 11 px in x direction	401	190	7.8
3	2024 1203	yes	34	300	Yes, 11 px in x direction	337	146	7.9

Table S11. Experimental details for photoselection of liposomes from the MinD library with mutations in the MTS.

Exp	Exp ID	Stim at center FOV	Laser power (µW)	Stim time (ms)	Correction xy point stimulation	Number of FOVs screened	Number of stimulated liposomes	FACS sorting threshold	Number of sorted liposomes
1	20240522	yes	34	300	Yes, 7 px in x direction	1,720	6,155	700	704
2	20240522	yes	34	300	Yes, 7 px in x direction	1,720	5,024	700	420
3	20240621	yes	34	300	Yes, 7 px in x direction	5,160	2,471	700	88

Table S12. Percentages of gated liposomes (extended data for Fig. S6). Gate > 988 a.u (PAmCherry2 signal). Percentages based on dataset with most events.

Exp	Mix - not activated	Mix - activated
1	0.00%	0.41%
2		0.38%
3	0.00%	0.01%

Table S13. List of primers. Phosphorothioate bonds are indicated with an asterisk (*).

Oligo name	Purpose	DNA sequence (5′ – 3′)
709 ChD	Gene amplification from T7 promoter to T7 terminator; also used for Sanger sequencing of *minD* variants from G598 containing mutations in the MTS of MinD	CAAAAAACCCCTCAAGACCCGTTTAGAGG
757 ChD	Gene amplification from T7 promoter to T7 terminator	TAATACGACTCACTATAGGG
1429 ChD	*PAmCherry2* gene amplification for the assembly of G595	AGCGGCCTGGTGCCGCGCGGCAGCCATAT GGTGTCTAAAGGGGAAGAGGACAAT
1430 ChD	*PAmCherry2* gene amplification for the assembly of G595	TTTCGGGCTTTGTTAGCAGCCGGATCCTTAC TTATAAAGTTCGTCCATACCCCCCGTTG
1431 ChD	pET15b vector amplification for the assembly of G595	TAAGGATCCGGCTGCTAACAAAGCC
1432 ChD	pET15b vector amplification for the assembly of G595	CATATGGCTGCCGCGCGGCAC
1355 ChD	*mEos2* gene amplification from G571 for the assembly of G577	AGCAGCGGCCTGGTGCCGCGCGGCAGCCA TATGAGTGCGATTAAGCCAGACATG
1356 ChD	*mEos2* gene amplification from G571 for the assembly of G577	TCCTTTCGGGCTTTGTTAGCAGCCGGATCCT TATCGTCTGGCATTGTCAGGCAATC
1353 ChD	pET15b vector amplification for the assembly of G577	GGATCCGGCTGCTAACAAAGCCCGAAAG
1354 ChD	pET15b vector amplification for the assembly of G577	ATGGCTGCCGCGCGGCACCAGGCCGCTG
1121 ChD	DNA quantification (qPCR) *YFP* gene	TGCAACTGGCTGACCACTAC
1122 ChD	DNA quantification (qPCR) *YFP* gene	AATGATTGTCCGGCAGCAGA
1208 ChD	DNA quantification (qPCR) *minD* gene	CGCGACTCTGACCGTATTT
1209 ChD	DNA quantification (qPCR) *minD* gene	AGCATGTCACCTCTGCTTAC
1414 ChD	DNA quantification (qPCR) *btubB* gene	CTGACCCTGCAACGTATTCT
1415 ChD	DNA quantification (qPCR) *btubB* gene	TACGGTTCAGTTTCGCCTTC
1445 ChD	Full-plasmid amplification of G437 to introduce mutations in the MTS of MinD	GCCTTTCTTCTCTTCTTCAATGAAGCGGAA
1448 ChD	Full-plasmid amplification of G437 to introduce mutations in the MTS of MinD	T*T*CCGCTTCATTGAAGAAGAGAAGAAAGG CKWMSWCAAACGCTTGTTCGGAGGATAAG GATCCGGCTGCTAACAAAGCCC
1525 ChD	Forward PCR primer for amplification of *minD* gene	AAAGTAAGCCCCCACCCTCACATGATACC
1526 ChD	Reverse PCR primer for amplification of *minD* gene	AAAGTAGGGTACAGCGACAACATACACCAT TTCC

Table S14. List of plasmids.

Plasmid name	Description
G365	pUC19-*YFP*
G376	pUC57-*minE_opt*
G437	pUC19-*minD*
G439	pUC19-*btubB*
G595	pET15b-PAmCherry2
G571	pRSET-mEos2
G577	pET15b-mEos2
G598	pUC19-*minD* 32-variant MTS library

DNA sequences

DNA sequence of PAmCherry2

ATGGTGTCTAAAGGGGAAGAGGACAATCTTGCAATCATCAAGGAGTTCATGCGCTTTAAGGTTCATCTGGAAGGCTCTG
TTAATGGACATGAATTTGAAATCGAAGGAGAAGGGGAGGGTCGTCCTTATGAGGGAACTCAAACTGCTAAATTGAAGG
TAACAAAGGGAGGACCACTTCCGTTCGCGTGGGATATTCTGTCGCCCCAGTTCATGTACGGAAGCAATGCGTACGTCAA
GCATCCCGCTGACATTCCTGACTACTTTAAGCTTTCCTTCCCCGAAGGTTTCAAATGGGAACGTGTTATGAACTTCGAAG
ACGGCGGGGTGGTAACTGTGACACAAGATTCGTCGTTGCAAGACGGTGAGTTTATCTACAAGGTAAAATTGCGCGGGA
CAAACTTTCCCTCTGACGGGCCCGTCATGCAAAAAAAGACCATGGGGTGGGAAACCTTATCAGAGCGTATGTACCCCG
AGGATGGTGCGCTTAAAGGCGAGTTGAAAGCTCGTACTAAGTTAAAGGATGGCGGACACTATGATACCGAAGTTAAAA
CGACTTATAAAGCGAAGAAACCCGTGCAATTACCAGGAGCATACAATGTTAATCGTAAGTTAGACATTACTAGTCACAA
CGAGGATTATACAATTGTCGAGCAATACGAGCGTGCTGAGGGTCTTCATTCAACGGGGGGTATGGACGAACTTTATAA
GTAA

Degenerate DNA sequence of 32-variant *minD* library with mutations in the MTS

ATGGCACGCATTATTGTTGTTACTTCGGGCAAAGGGGGTGTTGGTAAGACAACCTCCAGCGCGGCCATCGCCACTGGTT
TGGCCCAGAAGGGAAAGAAAACTGTCGTGATAGATTTTGATATCGGCCTGCGTAATCTCGACCTGATTATGGGTTGTGA
ACGCCGGGTCGTTTACGATTTCGTCAACGTCATTCAGGGCGATGCAACGCTAAATCAGGCGTTAATTAAAGATAAGCGT
ACTGAAAATCTCTATATTCTGCCGGCATCGCAAACACGCGATAAAGATGCCCTCACCCGTGAAGGGGTCGCCAAAGTTC
TTGATGATCTGAAAGCGATGGATTTTGAATTTATCGTTTGTGACTCCCCGGCAGGGATTGAAACCGGTGCGTTAATGGC
ACTCTATTTTGCAGACGAAGCCATTATTACCACCAACCCGGAAGTCTCCTCAGTACGCGACTCTGACCGTATTTTAGGCA
TTCTGGCGTCGAAATCACGCCGCGCAGAAAATGGCGAAGAGCCTATTAAAGAGCACCTGCTGTTAACGCGCTATAACC
CAGGCCGCGTAAGCAGAGGTGACATGCTGAGCATGGAAGATGTGCTGGAGATCCTGCGCATCAAACTCGTCGGCGTG
ATCCCAGAGGATCAATCAGTATTGCGCGCCTCTAACCAGGGTGAACCGGTCATTCTCGACATTAACGCCGATGCGGGT
AAAGCCTACGCAGATACCGTAGAACGTCTGTTGGGAGAAGAACGTCCTTTCCGCTTCATTGAAGAAGAGAAGAAAGGC
KWMSWCAAACGCTTGTTCGGAGGATAA

DNA sequence of *minE_opt*

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAG
ATATACATATGGCGCTGCTGGATTTCTTTCTGAGCCGTAAGAAAAACACCGCGAACATCGCGAAAGAGCGTCTGCAAAT
CATTGTTGCGGAGCGTCGTCGTAGCGATGCGGAACCGCACTACCTGCCGCAGCTGCGTAAAGATATCCTGGAAGTGAT
TTGCAAGTATGTTCAAATTGACCCGGAGATGGTGACCGTTCAGCTGGAACAAAAGGACGGTGATATCAGCATTCTGGA
GCTGAACGTTACCCTGCCGGAAGCGGAGGAACTGAAGTAAGGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAG
TTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTG

Supplementary Data 1 (Benchmarking overview) is provided as a separate file.

5

Co-optimization of the genetic and molecular compositions of a synthetic cell

Marijn van den Brink, Enrica Mammana, Yannick Bernard-Lapeyre and Christophe Danelon

Large-scale optimization of a synthetic cell genome and its molecular composition, constituting the cytoplasm and membrane, has largely been approached in isolation of each other. However, interactions between the genetic and molecular compositions ask for a combined optimization approach. Here, we develop a method for the co-optimization of the DNA sequence and the swelling solutions for gene-expressing liposomes using a barcoding strategy. Proof-of-concept experiments with a wild-type DNA template validated that the barcodes did not bias variant enrichment and allowed the retrieval of well-performing molecular compositions.

Introduction

The engineering of synthetic cells can be viewed as two-sided: we write the genetic code, and we control the initial molecular composition constituting the "cytoplasm" and the membrane. The DNA sequence can be refined by, for instance, mutating regulatory sequences (Chapter 3)[1], codon optimization, changing gene order and orientation, implementing operon design[2], as well as directed mutagenesis of protein active sites or screening for improved or new protein functionalities by machine-learning approaches[3]. The molecular composition can be optimized by adjusting the concentrations of cell-free expression system components and other substrates, cofactors and crowding agents required for reconstituting biological functions, as well as the pH, temperature, and lipid composition of the membrane. The search for the right initial conditions that would boot up gene expression, protein activity and the emergence of higher-order functions is a grand challenge.

Relatively small experimental screening datasets (considering the vast design space) have iteratively guided the optimization of DNA sequences or molecular compositions by exploiting active learning[4–8]. While the two engineering targets have mostly been approached independently, a few studies have highlighted their interdependence. Bernard-Lapeyre *et al*. found that the optimal composition of PURE system[9] is specific to the gene that is expressed and the DNA concentration[8]. Moreover, Caschera *et al*. found synergistic effects in *Escherichia coli* lysate between DNA variants (with different spacer sequences between regulatory sites) and molecular components, including the concentration of amino acids and the time at which different concentrations of amino acids, cofactors and energy source molecules were (re-)supplied to the system[4]. The highest fitness was only obtained when multiple components were changed simultaneously. Lastly, Nishikawa *et al*. studied how the translation of proteins containing consecutive proline residues, which can cause ribosome stalling, could be improved in PURE system. They found synergistic effects between encoding a SKIK peptide tag in the DNA and the addition of translation elongation factor P (EF-P)[10]. These examples underscore that the combined optimization of DNA sequence and molecular composition is likely key to the engineering success of synthetic cells.

In PURE system, when gene expression is driven by a fixed DNA sequence, up to two-body epistatic interactions between the system's components were found to increase or reduce protein synthesis[11]. However, it remains unclear how many components are involved per interaction when the DNA sequence is also variable, which components within the large genetic and molecular design space exhibit the strongest interactions, and how large the effects of those interactions are. Deeper insight into these factors would inform experimental design choices affecting the library design and size. Here, we devise a co-optimization strategy that enables the screening of DNA libraries in a background of ~35 different molecular compositions in PURE system. By using **barcoded DNA as a unique identifier of the PURE composition, all variants can be pooled** into one sample for screening and sorting,

minimizing the experimental efforts. In this chapter, we present proof-of-principle experiments that demonstrate the feasibility of this idea.

Results

Design of the co-optimization method

The Danelon Lab has reconstituted several biological functions (hereafter named *modules*) in liposomes by expressing genes from bacteria or bacteriophages using PURE system. These modules include DNA replication[12], lipid synthesis[13,14] and membrane remodeling and constriction[15–18]. The DNA replication module, derived from the Phi29 bacteriophage replication system, copies linear DNA templates that are flanked by two origins of replication derived from the Phi29 bacteriophage genome. We designed a method to co-optimize DNA sequence and molecular composition that is compatible with (i) the synthetic cell chassis based on gene-expressing liposomes encapsulating DNA and PURE system, and (ii) the linear Phi29 DNA replication scheme. The workflow can be applied to the optimization of one module or, more excitingly, to the integration of multiple modules.

The workflow starts by constructing plasmid libraries for the modules of interest, for example targeting regulatory sites by MOSAIC as demonstrated in Chapters 2-3 (**Fig. 1**). The DNA libraries are linearized, and multiple parts are assembled into one linear construct via Golden Gate Assembly while introducing a DNA barcode between the parts. Note that the barcode cannot be appended at the end of the DNA, as the final template should contain an origin of replication at each end to ensure it is recognized by the DNA replication machinery. Importantly, the assembly is performed in parallel reactions using different barcodes. Each barcode will be an identifier of a different molecular composition. For instance, the concentrations of ribosome, RNA polymerase, amino acids, or additives, may vary between the conditions.

Next, lipid films are dried in tubes or a multi-well plate. The lipid films can be identical or vary in lipid composition. Different molecular compositions are prepared and added to the dried lipid films together with a uniquely barcoded DNA library. The lipid films will swell to form liposomes, encapsulating the PURE system and DNA with at most one DNA copy per liposome. Lastly, all liposome samples are pooled, screened and those with highest performance are sorted. Pooled screening is preferred over parallel screening, because sorting and recovering the DNA from approximately a hundred separate liposome samples is time- and resource-consuming. Finally, the DNA recovered from the sorted liposomes is sequenced. The molecular composition, derived from the barcode sequence, and the DNA library variant can be linked to the variant's fitness (i.e., the variant frequency). This information is fed into an active learning model to predict a new set of conditions for a subsequent screening round to effectively navigate the multidimensional search space.

Barcoding approach

To validate the feasibility of the barcoding strategy, a set of 35 barcodes was generated (**Table S1**). Each barcode has a length of 10 nucleotides, a GC content between 0.25 and 0.75 and a maximum homopolymer length of 3. These guidelines were applied to promote high structural stability and accurate DNA sequencing. Moreover, the barcodes had a minimum Hamming distance of 3 nucleotides between each other, allowing the retrieval of the barcode identity even if one or two nucleotides are sequenced incorrectly or mutated during PCR or in-liposome DNA replication.

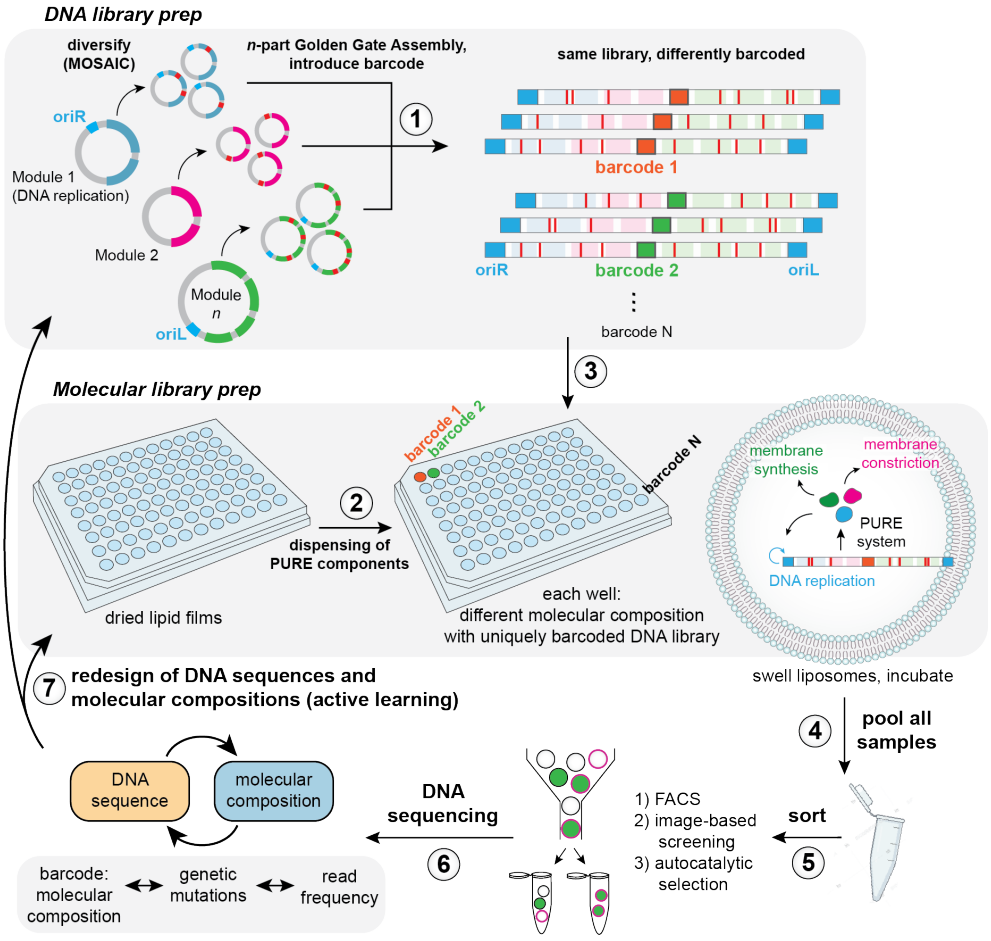


Fig. 1 Workflow for the co-optimization of genotype and molecular composition of synthetic cells. This chapter of the thesis (Chapter 5) describes the general concept. Chapter 6 provides a more in-depth discussion of the different steps in the pipeline and possible applications.

As proof-of-principle, the barcodes were incorporated into a DNA template encoding the two modules described in Chapter 3: DNA replication and phosphatidylserine (PS) synthesis. This six-gene system could benefit from further optimization, since the fraction of liposomes showing the combined phenotypes is small, and the activity of phospholipid synthesis too low for significant membrane growth[19].

The addition of the barcode at the end of the DNA template by overhang-PCR would be straightforward; however, this is not feasible because the DNA should be flanked by the origins of replication. An alternative barcoding method would be overhang-PCR with a long primer containing the barcode and a minimized replication origin (68 bp). However, shortening the replication origins reduces replication efficiency[20], and this method would depend on very long primers. Therefore, we here propose to insert the barcode in a central position of the DNA, which allows the use of the full 191- and 194-nucleotide long replication origins and minimizes the potential effects of the barcodes on initiation of DNA replication (**Fig. 2a**). For that, we optimized a Golden Gate Assembly protocol for the *in vitro* assembly of two linear DNA templates. Two templates were amplified separately by PCR, each encoding the genes required for one of the modules. To the template encoding the phospholipid synthesis genes, a DNA barcode was appended through primer overhang. Both resulting amplicons contained restriction sites, which after restriction generate compatible overhangs for the ligation of the two templates. The correctly assembled DNA was amplified using the crude restriction-ligation reaction as template, and its size (9.6 kb) was verified by gel electrophoresis (**Fig. 2b**). The correct sequence was verified by long-read nanopore sequencing. The assembly reaction could easily be scaled up to the production of the 35 differently barcoded templates.

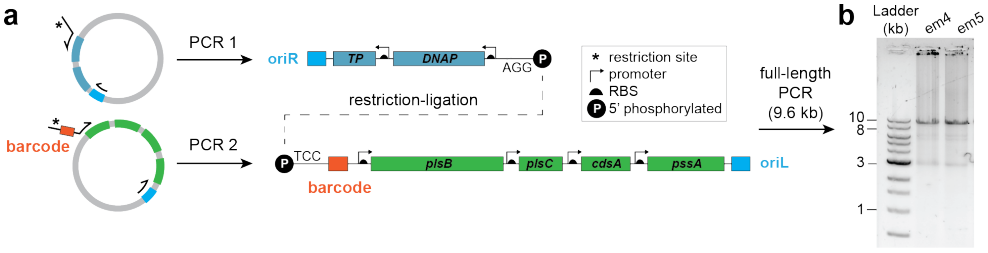


Fig. 2 Barcoding method for a two-module DNA template flanked by origins of replication. a The genes of interest are amplified from two plasmids. Plasmid 1: two genes for DNA replication. Plasmid 2: four genes for phospholipid synthesis. Each amplicon contains one origin of replication (oriR or oriL). Linear amplicons are treated with the type IIS restriction enzyme BspQI-HF to generate compatible overhangs on each template for subsequent ligation. **b** PCR amplification of the two-module template from oriR to oriL (9.6 kb) was confirmed by agarose gel electrophoresis. The two samples in the gel contain barcode em4 or em5.

Barcodes do not introduce an enrichment bias

Next, we validated that the barcodes did not bias the enrichment of variants, for instance by affecting the module activities, DNA amplification from sorted liposomes or sequencing. We prepared a mix of 11 barcoded templates, which we encapsulated and expressed in liposomes at 200 pM (λ = 4 DNA molecules per liposome, on average) (**Fig. 3a**). A lower DNA concentration (λ = 1) yielded too few liposomes with the combined phenotype for sorting. The calculation of the expected average number of DNA molecules per liposome (λ) is presented in the Methods section *Statistics for the number of DNA molecules per liposome* in Chapter 3.

The liposome quality and module activities were verified by confocal fluorescence microscopy (**Fig. 3b**) and fluorescence-activated cells sorting (FACS) (**Fig. 3c**). DNA replication was reported by the fluorescent dye dsGreen[21], which intercalates specifically with double-stranded DNA. Phospholipid synthesis was reported by the PS-specific probe LactC2-mCherry[14]. We sorted the liposomes with combined phenotypes by FACS, recovered the full-length DNA by PCR (9.6 kb), and quantified the abundance of each barcode before and after sorting by nanopore sequencing (**Fig. 3d**). All barcodes were retrieved at similar frequencies. On average, the barcode frequency was 9 ± 4% (mean ± standard deviation, n=11 barcodes, both before and after sorting, confirming that the barcode sequences did not bias the sorting, recovery and sequencing of the DNA.

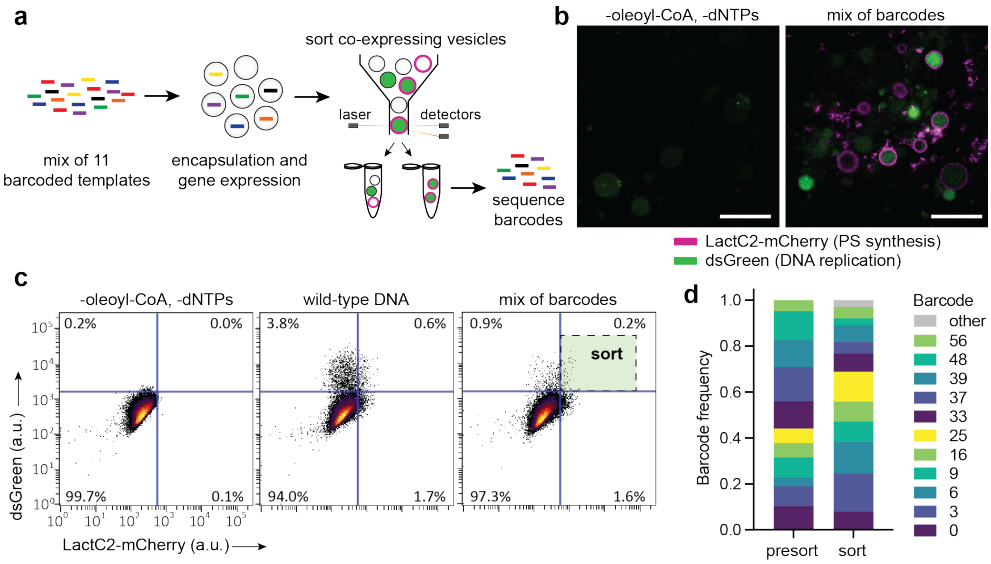


Fig. 3 Barcode sequences do not introduce a bias in DNA enrichment. a Eleven uniquely barcoded DNA were mixed before encapsulation in liposomes. The liposomes were screened by FACS for DNA replication and PS synthesis activity. The liposomes with the joint phenotype were sorted, their DNA was recovered by PCR and analyzed by long-read nanopore sequencing. **b** Confocal fluorescence images of liposomes expressing the mix of DNA templates with different barcodes. The negative control (left image) lacked oleoyl-CoA and dNTPs. Scale bars: 10 µm. **c** FACS plots of the negative control (barcode mix lacking

oleoyl-CoA and dNTPs), positive control (wild-type DNA without barcode), and the mix of barcodes. **d** Frequency of barcodes before and after sorting (n=1 biological replicate). Only reads with length >7000 nucleotides were used for analysis. Presort: 127 reads. Sort: 138 reads. For the sorted sample, 3% of the reads contained a sequence other than the barcodes, possibly caused by mutations during in-liposome DNA replication, PCR or sequencing errors.

Well-performing molecular compositions were enriched and identified by barcodes

Next, we sought to enrich molecular variants that promote DNA replication. We prepared 11 swelling solutions, varying in the presence and absence of substrates/cofactors required for DNA replication and/or phospholipid synthesis (**Tables S2-4**). The wild-type variant was included in duplo (samples 1 and 11). A molecular variant that was designed for low DNA replication activity (sample 2) lacked dNTPs and the accessory single- and double-stranded binding proteins (SSB and DSB, respectively). Molecular variants designed for reduced PS synthesis activity (samples 3 and 4) lacked the fatty acid chain substrate (oleoyl-CoA) and/or the head-modifying substrates (cytidine triphosphate (CTP) and L-serine), glycerol-3-phosphate and reducing agent. Negative controls for both modules (samples 5, 6, 7) lacked combinations of the previously mentioned components, lacked ribosomes, or were incubated at 34 °C for 20 minutes instead of 16 hours. Solutions were based on a PURE system formulation previously optimized for *in vitro* transcription, translation and DNA replication[22]. Three other PURE system formulations were tested with no clear expectations on the outcomes, including the commercial PURE*frex* 2.0 (sample 10), and two other PURE compositions optimized in our group for gene expression in bulk reactions: Top Day 5 (sample 8) and Top Day 8 (sample 9)[8]. To each swelling solution, a uniquely barcoded DNA template was added at 500 pM concentration (λ = 10 DNA molecules per liposome, on average) (**Fig. 4a**). After encapsulating the swelling solutions in liposomes and activating gene expression by incubation at 34 °C, all liposome samples were combined into one sample. Active DNA replication and PS synthesis were confirmed for the mixed sample by fluorescence microscopy (**Fig. 4b**). The liposomes with active DNA replication were sorted by FACS (**Fig. 4c**). The 1.3-kb region around the barcode was recovered by PCR and analyzed by nanopore sequencing. We did not sort liposomes with the combined phenotypes due to the inability to detect LactC2-mCherry signal by FACS, for a still unknown reason. We compared the expected distributions of the molecular variants based on the substrates and cofactors that we depleted (**Fig. 4d**) with the sequencing results (**Fig. 4e**). We concluded that the samples expected to promote DNA replication were indeed identified by sequencing, and no barcodes from the samples uncapable of DNA replication were detected. PURE composition "Top Day 5" was not detected, while "Top Day 8" and the commercial PURE*frex* 2.0 were detected, indicating that they promote DNA replication. Additional biological replicates or larger-scale experiments with 20-35 different compositions would further demonstrate the applicability of the method.

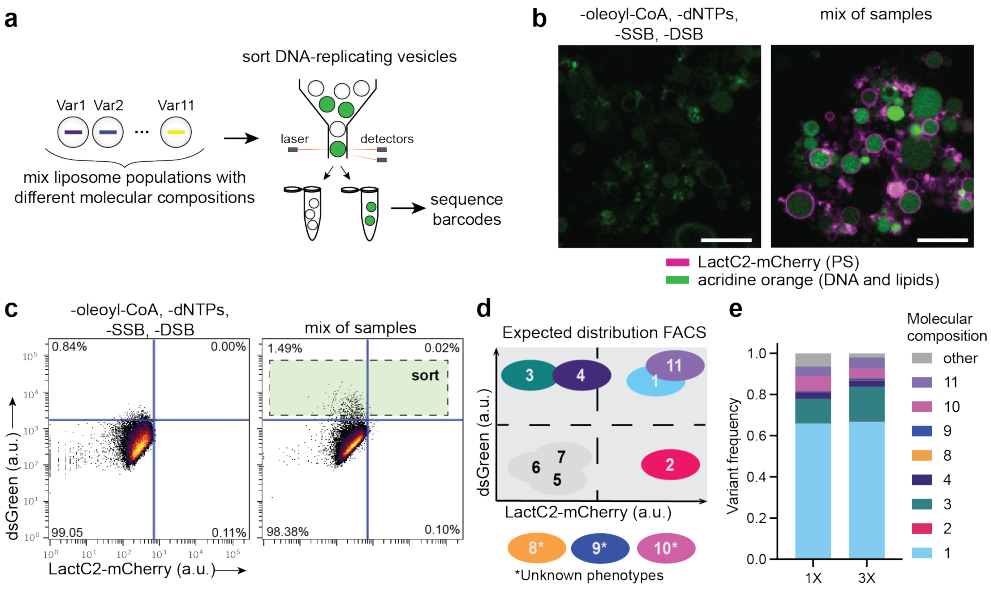


Fig. 4 Enrichment of molecular variants that promote DNA replication. a Eleven swelling solutions were prepared varying in substrates/cofactors for DNA replication and/or PS synthesis. Each swelling solution contained a uniquely barcoded DNA template (Table S2). The liposomes were screened by FACS for DNA replication activity, those with dsGreen signal higher than the negative control were sorted, and their DNA was amplified and sequenced. **b** Confocal fluorescence image of the mixed populations of liposomes prepared from different swelling solutions (right image). One of the variants (sample 5, lacking oleoyl-CoA, dNTPs, SSB and DSB) was imaged in isolation as negative control (left image). Scale bars: 10 µm. **c** FACS plots for one of the samples serving as negative control (sample 5), and the mix of all 11 samples. **d** Schematic FACS plot displaying the expected distribution of the molecular variants. **e** Frequency of barcodes, indicating the frequency of the molecular variants after sorting (n=1 biological replicate, n=2 technical replicates). One technical replicate (1X PCR) is a single PCR. The second technical replicate is a mix of three identical PCRs, to compensate for possible random bias in template amplification. 1X PCR: 3159 reads. 3X PCR: 1704 reads. A small fraction of the reads contained a sequence other than the barcodes, possibly caused by mutations during in-liposome DNA replication, PCR or sequencing errors.

Upscaling the preparation of molecular compositions and liposomes

Preparing multiple liposome swelling solutions with varying concentrations of PURE system components and module substrates can quickly become highly labor-intensive. Automated liquid dispensing can be a solution to that. For example, with the Echo acoustic droplet ejection technology, PURE system reagents can be dispensed in 384-well plates in an automated fashion at ~10- to 20-µL final reaction volumes (**Fig. 5a**)[8].

To test if solutions prepared with automated liquid handling are compatible with liposome swelling and lead to active DNA replication and PS synthesis, the M1 PURE system components and all module substrates were dispensed using the Echo liquid handler. Only proteins SSB and DSB were added manually to conserve these reagents (the dead volume in the source

plate for droplet transfer is about 20 µL). A total of 27 different components were distributed at one location of our lab (Toulouse, France). The reaction mixtures were snap-frozen, shipped to the other location (Delft, the Netherlands), and encapsulated in liposomes via natural swelling of lipid films dried on glass beads[23] in Eppendorf tubes. After incubation at 37 °C, the liposomes were imaged (**Fig. 5b**). Many lipid aggregates were observed, which could be due to sample-to-sample variability, or possibly due to the new way the swelling solutions were prepared. Nevertheless, co-occurrence of DNA replication and phospholipid synthesis in single liposomes was confirmed. Scaling up liquid dispensing of swelling solutions to generate 100 samples with varying molecular compositions is possible[8]. The next step is to also scale up the swelling of lipid films in multi-well plates to produce 100 liposome samples. A sample number of this order has proved sufficient for the optimization of cell-free expression systems by active learning in the past[4–6], while maintaining the experimental effort manageable.

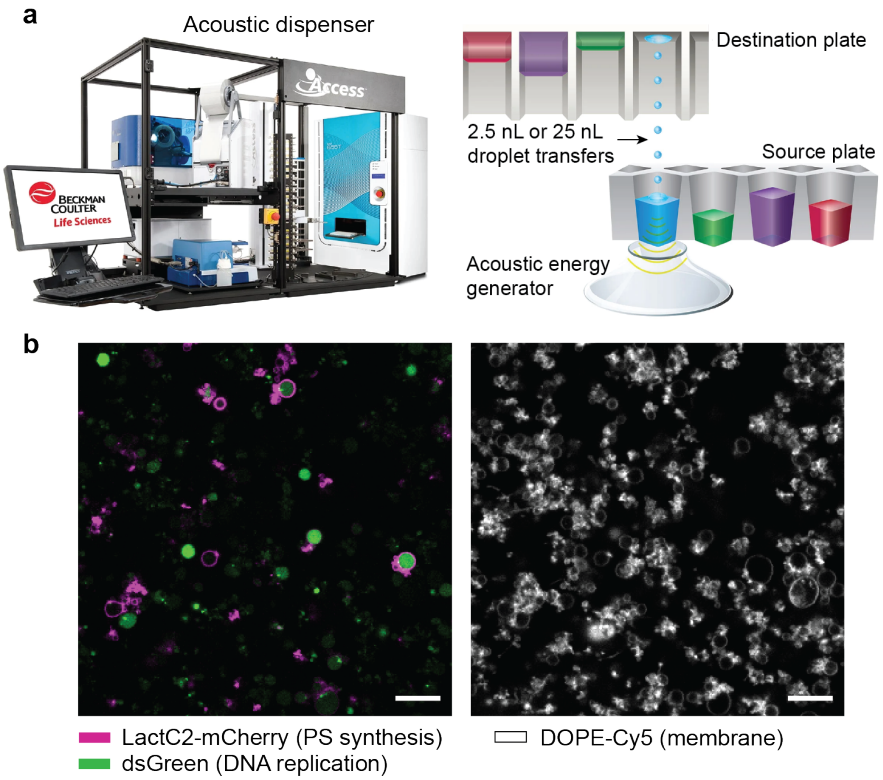


Fig. 5 Liposomes prepared with swelling solutions dispensed by the Echo liquid handler are capable of DNA replication and PS synthesis. a The Echo acoustic droplet ejection technology from Beckman Coulter. Nanoliter droplets are transferred from a source plate to a destination plate by an acoustic energy generator. Figures adapted from: mybeckman.nl/liquid-handlers/echo-acoustic/. **b** Confocal fluorescence microscopy images of liposomes with active DNA replication and PS synthesis. Liposomes prepared in Eppendorf tubes using the Echo-dispensed M1 PURE system containing 500 pM DNA. The liposomes were incubated at 37 °C for 16 hours before imaging. Scale bars: 10 µm.

Discussion

The combined testing of different molecular compositions (constituting the synthetic cell "cytoplasm" and the membrane) and DNA sequences is difficult due to the large sampling space. This is particularly the case when the template DNA is long and contains multiple genes, as each additional gene increases the sequence space and the number of interactions in the system. Each additional gene also intensifies the competition for transcription-translation resources, making the system's level optimization increasingly important. Our barcode-assisted screening method enables a shift from testing one DNA variant per molecular variant, to evaluating many DNA variants per molecular variant. By pooling all samples together before screening, the screening, sorting and DNA recovery steps become effort- and cost-efficient.

The barcodes were integrated during the assembly of the construct, making the method flexible and efficient by avoiding additional cloning. We incorporated the barcodes in the DNA templates via a fully *in vitro* assembly and amplification method; due to the multiple repeats in the DNA template (promoters, RBSs, terminators), DNA recombination could occur when the construct is amplified *in vivo*. Nevertheless, for larger constructs used in the future (> 20 kb), PCR may no longer be a feasible option for amplification. The Golden Gate Assembly product (when designed as circular template) could also be transformed into *E. coli* for amplification, whereafter all colonies are harvested together to maintain the library. The circular amplified product could then be linearized using blunt-end restriction enzymes. Alternatively, assembly itself could be performed *in vivo* in *Saccharomyces cerevisiae*. The choice of assembly/amplification method will depend on the construct size, the number of fragments, and whether the construct will contain sequence repeats and DNA heterologous or homologous to the assembly host cell[24]. Specifically, *S. cerevisiae* is useful for large assemblies (up to Mb size) and robust to the presence of heterologous DNA, but yields relatively low DNA quantity and purity[24,25].

A downside of our approach is that the encapsulation of the DNA library should ideally happen at maximum one molecule per liposome ($\lambda < 1$) to maintain a genotype-phenotype link. This generates a larger fraction of non-active (empty) liposomes compared to encapsulation at higher DNA concentrations. Therefore, the method relies on high-throughput screening, such as FACS. The screening capacity for liposomes using FACS is 10^6-10^7 variants per day for λ = 0.1-0.2 DNA molecules per liposome on average (Chapter 3 and [26]). Thus, when testing in the order of 30-100 molecular variants, we expect that every DNA variant in every molecular environment can be screened at least once using a maximum DNA library size of 10^4-10^5 variants. For screening methods with lower throughputs, such as the image-based screening and sorting method described in Chapter 4, the DNA library size could be downscaled accordingly. Alternatively, a first high-throughput screen could be performed with FACS to establish a functional gene expression level for most genes, whereafter the sorted variants proceed to a subsequent lower-throughput screen for visual inspection using microscopy

images to engineer morphological, localization or dynamic features. This will be further elaborated on in Chapter 6.

Methods

DNA oligos and plasmid

DNA oligos (**Tables S1 and S5**) were ordered from Ella Biotech. Plasmid G616 (encoding the DNA replication and phospholipid synthesis modules) was assembled *in vivo* in *Saccharomyces cerevisiae* following the procedure described by Restrepo Sierra *et al.*[19]. Plasmid G616 encodes (i) the genes for the DNA replication module (*p2* and *p3*), (ii) the genes for the phospholipid synthesis module (*plsB*, *plsC*, *cdsA* and *pssA*), (iii) the Phi29 origins of replication, (iv) the centromeric origin of replication *CEN6/ARS4* and the auxotrophic marker *URA3* for maintenance and selection in *S. cerevisiae*, and (v) the *OriE1* origin of replication for amplification and long-term storage in *E. coli*. After assembly in *S. cerevisiae*, *E. coli* DH5α cells were transformed with 1 ng of the plasmid via heat shock. After 1 h recovery in LB medium, the transformed cells were plated on agar plates containing 50 µg/mL ampicillin. Single colonies were grown overnight at 37 °C in 6 mL of LB medium containing ampicillin (50 µg/mL). The plasmids were isolated from the cells using the PureYieldTM Plasmid Miniprep System (Promega) according to the manufacturer's instructions. The DNA sequence was verified using nanopore sequencing (Plasmidsaurus).

Protein purification

Single-stranded binding protein (SSB) and double-stranded binding protein (DSB) were produced and stored as described earlier[12].

Barcode design

A custom Python script (Creating_barcodes.py, available in the 4TU.ResearchData repository with DOI: https://doi.org/10.4121/692a826b-35c7-491d-a8a8-ddddc69e1dd1) was used to create a set of DNA sequences based on user-defined parameters: set size of 60 sequences, sequence length of 10 nucleotides, GC ratio between 0.25-0.75, maximum 3 consecutive identical nucleotides in the barcode region, and a minimum Hamming distance of 3 between any two barcodes. From a set of 60 generated barcodes, a manual selection was performed to exclude nucleotide repeats when the barcodes are integrated into the DNA template of interest. The scripts were implemented in Python 3 with Visual Studio Code (v1.105.1) using the random, os and pandas libraries.

PCR amplification of (barcoded) templates

The wild-type linear DNA template (no barcode) was amplified from plasmid G616 by PCR with primers 1523 ChD and 1524 ChD. The fragment with two genes for DNA replication was amplified from plasmid G616 by PCR using primer 1526 ChD and the overhang primer 1605

ChD carrying the restriction site for BspQI-HF. The fragment with genes for PS synthesis (*plsB*, *plsC*, *cdsA* and *pssA*) was amplified from plasmid G616 by PCR using primer 1525 ChD and an overhang primer carrying a barcode and the restriction site for BspQI-HF (1604 ChD, 1617-1626 ChD or 1631-1654 ChD) (barcoded primers are listed in **Table S1**). PCR was performed with RepliQa Hi-Fi ToughMix (QuantaBio) following the manufacturer's instructions using thermocycling conditions: 98 °C for 30 s, 30 cycles of (98 °C for 10 s, 68 °C for 5-10 s/kb). To digest the parent plasmids, 0.5 µL of restriction endonuclease DpnI (20,000 U/mL, NEB) was added per 50 µL PCR mix after thermocycling, and the solution was incubated for 1 h at 37 °C. DNA was purified from the PCR reagents using the QIAquick PCR Cleanup Kit (Qiagen). The size of the amplicons was verified by agarose gel electrophoresis (0.7% agarose in TAE buffer) using the 1 kb DNA Ladder (NEB). DNA concentrations were measured with the Qubit dsDNA High Sensitivity Assay Kit and the Qubit 4 Fluorometer (Invitrogen).

Golden Gate Assembly

For Golden Gate Assembly of the DNA replication and phospholipid synthesis modules, 20 ng of each DNA fragment were mixed with 1 µL of BspQI-HF (10,000 U/mL, NEB), 0.5 µL of T4 DNA ligase (400,000 U/mL, NEB), 2 µL of 10× T4 ligase reaction buffer (NEB), 0.5 µL of DpnI (20,000 U/mL, NEB) and Milli-Q water up to 20 µL. The reaction mix was incubated for 1 h at 37 °C (restriction) followed by 15 min at 16 °C (ligation).

Amplification of the assembled DNA template

The full-length DNA (9.6 kb) was amplified from the assembly reaction by PCR with primers 1523 ChD and 1524 ChD using 2.5 µL of GGA product per 50 µL of PCR reaction volume. PCR was performed with RepliQa Hi-Fi ToughMix (QuantaBio) following the manufacturer's instructions using thermocycling conditions: 98 °C for 30 s, 30 cycles of (98 °C for 10 s, 68 °C for 5-10 s/kb). The amplicons were purified directly or by gel extraction using the QIAquick PCR & Gel Cleanup Kit (Qiagen). Amplicon size (9.6 kb) and concentration were verified by agarose gel electrophoresis and the Qubit dsDNA High Sensitivity Assay Kit, respectively.

Liposome preparation

Liposome swelling solutions (20 µL) were prepared on ice. The wild-type molecular composition was based on a PURE system formulation previously optimized for *in vitro* transcription, translation and DNA replication in our group[22], M1 PURE system, which was produced by GeneFrontier. The wild-type molecular composition consisted of: 7.2 µL M1 Solution I, 4.2 µL M1 Solution II, 1.3 µL M1 Solution III, 0.5 U μL^{-1} SUPERase·In RNase Inhibitor (Ambion), 1 mM CTP, 0.5 mM glycerol-3-phosphate, 0.5 mM L-serine, 5 mM beta-mecaptoethanol, 0.3 mM dNTP Mix (Thermo Fisher), 20 mM ammonium sulphate, 0.75 mg/mL SSB, 0.21 mg/mL DSB, DNA template and MilliQ-water. Other molecular compositions based on M1 PURE system or other PURE system formulations are presented in **Tables S3** and **S4**, respectively.

Next, 10-12 mg of lipid-coated glass beads were added to 20 µL of swelling solution. The lipid-coated beads were prepared as described earlier[27] with minor modifications. The glass beads (0.6 g) were coated with 50 mol % DOPC, 36 mol % DOPE, 12 mol % DOPG, 2 mol % 18:1 cardiolipin and 1 mass % DSPE-PEG(2000)-biotin with a total mass of lipids of 2 mg. 0.05 mass % DOPE-Cy5 was also added when indicated. The liposomes were swelled at 4 °C for 1 h, whereafter four freeze-thaw cycles were applied by dipping the samples in liquid nitrogen for 10 s followed by thawing on ice.

Using a cut pipette tip, 12 µL of liposomes and 0.5 µL of DNase I (2 U/µL) were transferred to PCR tubes containing a dried film of oleoyl-CoA (Avanti Polar Lipids) to reach a final concentration of 176 µM oleoyl-CoA. The liposomes were incubated in a thermocycler at 34 °C for 16 h. Time 0 controls were incubated for 20 min at 34 °C to degrade the DNA not encapsulated in liposomes by DNase I, and were then kept at 4 °C.

Confocal fluorescence microscopy

Using a cut pipette tip, 2 µL of liposomes were mixed with 13 µL of buffer containing 20 mM HEPES-KOH, pH 7.6, 180 mM potassium glutamate, 14 mM magnesium acetate, 500 nM LactC2-mCherry and 10X dsGreen (Lumiprobe) (or 62.5 µM acridine orange (Sigma-Aldrich) instead of dsGreen). Imaging wells were constructed by resizing a glass coverslip using a diamond engraving pen to fit the microscope holder. The necks of pipette tips were cut to form short cylindrical rims. The flat edge of each cut tip was lightly dipped into a thin layer of 2-component epoxy glue (Bison), and the resulting sticky rim was immediately placed on coverslip to create the imaging well. After washing the wells three times with 10 µL of Milli-Q water, 10 µL of BSA (1 mg/mL) were added. BSA was removed after 10 min, and the wells were washed twice with Milli-Q water and once with buffer containing 20 mM HEPES-KOH, pH 7.6, 180 mM potassium glutamate, 14 mM magnesium acetate. The well was filled with 15 µL of liposome sample and covered with 10 µL of mineral oil. The samples were imaged with the A1R Laser scanning confocal microscope (Nikon) with an SR Apo TIRF 100× oil immersion objective (NA 1.49). Imaging was performed using laser line 488 nm with emission filter 525/50 (dsGreen), 561 nm with emission filter 595/50 (LactC2-mCherry), and 640 nm with emission filter 700/75 (Cy5). Images were processed in Fiji (ImageJ2 v2.9.0)[28].

Liposome screening and sorting by FACS

1-3 µL of liposomes were diluted in 200 µL buffer containing 20 mM HEPES-KOH, pH 7.6, 180 mM potassium glutamate, 14 mM magnesium acetate, 300 nM LactC2-mCherry and 10X dsGreen (Lumiprobe). The diluted liposomes were filtered using the 35 µm nylon mesh of a cell strainer cap of 5-mL round bottom polystyrene test tubes (Falcon). The filtered samples were analysed and sorted with BD FACSMelody Cell Sorter (BD Biosciences) with a 100-µm nozzle and operated with BD FACSChorus application (v1.4.3.0). LactC2-mCherry was measured with laser line 561 nm with emission filter 610/20 nm, and dsGreen with laser line 488 nm with emission filter 527/32 nm. The side scatter (SSC) threshold for triggering was set

at 359 V. For sorting, first a gate was set on FSC and SSC to select the main liposome population (~60-80% of all events) (example shown in **Fig. S9 in Chapter 4**). For this population, additional gates based on the LactC2-mCherry and dsGreen signals were applied. Sorting was performed in purity mode. The recorded data was further processed using Cytobank (https://community.cytobank.org/). Here, in addition to the gating based on FSC and SSC, events with LactC2-mCherry and/or dsGreen signals lower than 2 a.u. were discarded.

DNA recovery from sorted populations

Liposomes sorted by FACS were centrifuged at 4 °C and 12,000 rcf for 3 min, whereafter the supernatant was discarded. The remaining 12-20 µL were incubated at 75 °C for 15 min to heat-inactivate DNase I and stored at -20 °C. 1-3 µL were added to a 50-µL PCR reaction. Primer pairs used were 1567 ChD and 1627 ChD (full-length DNA template, experiment of Fig. 3) or 1138 ChD and 756 ChD (smaller fragment, experiment of Fig. 4). PCR was performed with RepliQa Hi-Fi ToughMix (QuantaBio) following the manufacturer's instructions using thermocycling conditions: 98 °C for 30 s, 30 cycles of (98 °C for 10 s, 68 °C for 5-10 s/kb). The amplified DNA was purified and sequenced by Plasmidsaurus.

Nanopore sequencing and analysis

Sequencing of the barcoded DNA constructs with the Oxford Nanopore Technology was performed by Plasmidsaurus using the Amplicon PCR Premium option. The reads were filtered based on a minimum size of 7 kb (experiment of Fig. 3) or 1 kb (experiment of Fig. 4) using the *Filter FASTQ reads by quality score and length* tool[29] from the Galaxy web platform (https://usegalaxy.org)[30]. Alignments to a reference sequence were performed with the *Map with minimap2* tool[31] from the same platform, using default parameters except the gap open penalties for deletions and insertions were set to 16 and 48, respectively, to minimize the number of insertions/deletions rather than mismatches during mapping, and the *retain at most INT secondary alignments* parameter was set to 0. For samples containing a mix of barcodes, the reads were aligned to a reference sequence containing a 10-nucleotide repeat of adenines instead of the barcode to minimize misalignments due to the heterogeneity of the reads. The mapped reads were further analysed with an R script developed previously[32], with slight modifications to detect barcode variants (adapted script available in the 4TU.ResearchData repository with DOI: https://doi.org/10.4121/692a826b-35c7-491d-a8a8-ddddc69e1dd1).

While we here identified barcodes using 100% identity between nanopore sequencing reads and the reference barcodes, future analysis could map the reads to the reference barcodes while allowing one or two incorrectly sequenced bases (typical single-read per-base accuracy of nanopore sequencing is 99-99.7%[33], rendering one in 10-30 sequenced barcodes unusable), because any two barcodes differ with at least three nucleotides from each other.

Acoustic liquid dispensing by the Echo Liquid Handler

Liquid dispensing by the Echo 525 Acoustic Liquid Handler (Beckman Coulter) was performed as described earlier[8] in a 384-wells plate at a final volume of 22 µL per well. 27 components were dispensed, excluding purified proteins DSB and SSB, which were added manually afterwards (**Table S6**). One of the components was the wild-type DNA, which was PCR-amplified from plasmid G616 as described earlier and added at 500 pM final concentration. Directly after dispensing, the plate with PURE solutions was kept on ice until it was snap-frozen using liquid nitrogen for shipping. After shipping, the plate was thawed on ice. Liposomes were prepared in Eppendorf tubes as described in earlier sections, with the following adaptations: a final oleoyl-CoA concentration of 100 µM, and the liposomes were incubated at 37 °C for 16 h after swelling. A negative control sample was incubated at 37 °C for 20 min. The samples were analyzed by fluorescence microscopy as described in an earlier section. The samples were kept at room temperature during imaging.

Data and code availability

The code, raw data and processed data from this chapter, including DNA sequencing data, gel electrophoresis, microscopy images and FACS data are available in the 4TU.ResearchData repository at https://doi.org/10.4121/692a826b-35c7-491d-a8a8-ddddc69e1dd1.

References

1. Chizzolini, F. *et al.* Cell-free translation is more variable than transcription. *ACS Synth. Biol.* **6**, 638–647 (2017).
2. Chizzolini, F., Forlin, M., Cecchi, D. & Mansy, S. S. Gene position more strongly influences cell-free protein expression from operons than T7 transcriptional promoter strength. *ACS Synth. Biol.* **3**, 363–371 (2014).
3. Kohyama, S., Frohn, B. P., Babl, L. & Schwille, P. Machine learning-aided design and screening of an emergent protein function in synthetic cells. *Nat. Commun.* **15**, 2010 (2024).
4. Caschera, F. *et al.* Coping with complexity: Machine learning optimization of cell-free protein synthesis. *Biotechnol. Bioeng.* **108**, 2218–2228 (2011).
5. Pandi, A. *et al.* A versatile active learning workflow for optimization of genetic and metabolic networks. *Nat. Commun.* **13**, 3876 (2022).
6. Borkowski, O. *et al.* Large scale active-learning-guided exploration for in vitro protein production optimization. *Nat. Commun.* **11**, 1872 (2020).
7. Zhu, J. *et al.* AI-driven high-throughput droplet screening of cell-free gene expression. *Nat. Commun.* **16**, 2720 (2025).
8. Bernard-Lapeyre, Y., Cleij, C., Sakai, A., Huguet, M.-J. & Danelon, C. Optimization of PURE system composition using automation and active learning. Preprint at https://doi.org/10.64898/2026.03.23.713685 (2026).
9. Shimizu, Y. *et al.* Cell-free translation reconstituted with purified components. *Nat. Biotechnol.* **19**, 751–755 (2001).
10. Nishikawa, Y., Fujikawa, R., Nakano, H., Kanamori, T. & Ojima-Kato, T. Effect of translation-enhancing nascent SKIK peptide on the arrest peptides containing consecutive proline. *ACS Synth. Biol.* **13**, 3908–3916 (2024).

11. Matsuura, T., Kazuta, Y., Aita, T., Adachi, J. & Yomo, T. Quantifying epistatic interactions among the components constituting the protein translation system. *Mol. Syst. Biol.* **5**, MSB200950 (2009).
12. van Nies, P. *et al.* Self-replication of DNA by its encoded proteins in liposome-based synthetic cells. *Nat. Commun.* **9**, 1583 (2018).
13. Scott, A. *et al.* Cell-free phospholipid biosynthesis by gene-encoded enzymes reconstituted in liposomes. *PLOS ONE* **11**, e0163058 (2016).
14. Blanken, D., Foschepoth, D., Serrão, A. C. & Danelon, C. Genetically controlled membrane synthesis in liposomes. *Nat. Commun.* **11**, 4317 (2020).
15. Godino, E. *et al.* De novo synthesized Min proteins drive oscillatory liposome deformation and regulate FtsA-FtsZ cytoskeletal patterns. *Nat. Commun.* **10**, 4969 (2019).
16. Godino, E. *et al.* Cell-free biogenesis of bacterial division proto-rings that can constrict liposomes. *Commun. Biol.* **3**, 539 (2020).
17. Godino, E., Doerr, A. & Danelon, C. Min waves without MinC can pattern FtsA-anchored FtsZ filaments on model membranes. *Commun. Biol.* **5**, 675 (2022).
18. Godino, E. & Danelon, C. Gene-directed FtsZ ring assembly generates constricted liposomes with stable membrane necks. *Adv. Biol.* **7**, 2200172 (2023).
19. Sierra, A. M. R., Gomez, F. R., Tongeren, M. van, Heras, L. S. & Danelon, C. A synthetic cell with integrated DNA self-replication and membrane biosynthesis. 2025.01.14.632951 Preprint at https://doi.org/10.1101/2025.01.14.632951 (2025).
20. Mencía, M., Gella, P., Camacho, A., de Vega, M. & Salas, M. Terminal protein-primed amplification of heterologous DNA with a minimal replication system based on phage φ29. *PNAS* **108**, 18655–18660 (2011).
21. Abil, Z., Restrepo Sierra, A. M. & Danelon, C. Clonal amplification-enhanced gene expression in synthetic vesicles. *ACS Synth. Biol.* **12**, 1187–1203 (2023).
22. Danelon, C. & Heras, L. S. Autocatalytic selection of gene functions in vitro. 2025.08.25.672052 Preprint at https://doi.org/10.1101/2025.08.25.672052 (2025).
23. Nourian, Z., Roelofsen, W. & Danelon, C. Triggered gene expression in fed-vesicle microreactors with a multifunctional membrane. *Angew. Chem. Int. Ed.* **51**, 3114–3118 (2012).
24. Koster, C. C., Postma, E. D., Knibbe, E., Cleij, C. & Daran-Lapujade, P. Synthetic genomics from a yeast perspective. *Front. Bioeng. Biotechnol.* **10**, (2022).
25. Cleij, C. *et al.* Assembly and cell-free expression of a partial genome for the synthetic cell. 2025.10.16.682769 Preprint at https://doi.org/10.1101/2025.10.16.682769 (2025).
26. Nishikawa, T., Sunami, T., Matsuura, T., Ichihashi, N. & Yomo, T. Construction of a gene screening system using giant unilamellar liposomes and a fluorescence-activated cell sorter. *Anal. Chem.* **84**, 5017–5024 (2012).
27. Blanken, D., van Nies, P. & Danelon, C. Quantitative imaging of gene-expressing liposomes reveals rare favorable phenotypes. *Phys. Biol.* **16**, 045002 (2019).
28. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
29. Blankenberg, D. *et al.* Manipulation of FASTQ data with Galaxy. *Bioinformatics* **26**, 1783–1785 (2010).
30. The Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res.* **50**, W345–W351 (2022).
31. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
32. van den Brink, M., Althuis, T. Y., Danelon, C. & Claassens, N. J. MOSAIC: A highly efficient, one-step recombineering approach to plasmid editing and diversification. *ACS Synth. Biol.* **14**, 3880–3889 (2025).
33. Nanopore sequencing accuracy. *Oxford Nanopore Technologies* https://nanoporetech.com/platform/accuracy.

Supporting Information

Table S1 List of 35 barcodes and corresponding DNA oligos for PCR amplification of the phospholipid synthesis genes.

Barcode name	Barcode sequence	Oligo name	Oligo sequence (5' to 3')
em0	GGATGGACCT	1617 ChD	GCTCAGATCAGCTCTTCATCCGGATGGACCTCGG CCCTTCTAGAACGTCTCAATCTC
em1	AGCGGCCAGT	1631 ChD	GCTCAGATCAGCTCTTCATCCAGCGGCCAGTCGG CCCTTCTAGAACGTCTCAATCTC
em3	GGAAACTAGA	1618 ChD	GCTCAGATCAGCTCTTCATCCGGAAACTAGACGG CCCTTCTAGAACGTCTCAATCTC
em4	TAGTCTCTAC	1633 ChD	GCTCAGATCAGCTCTTCATCCTAGTCTCTACCGGC CCTTCTAGAACGTCTCAATCTC
em5	GTCGGTTATT	1634 ChD	GCTCAGATCAGCTCTTCATCCGTCGGTTATTCGGC CCTTCTAGAACGTCTCAATCTC
em6	AGATATTTCC	1619 ChD	GCTCAGATCAGCTCTTCATCCAGATATTTCCCGGC CCTTCTAGAACGTCTCAATCTC
em7	TAACCGTAAC	1635 ChD	GCTCAGATCAGCTCTTCATCCTAACCGTAACCGGC CCTTCTAGAACGTCTCAATCTC
em8	TGGTCCATAC	1636 ChD	GCTCAGATCAGCTCTTCATCCTGGTCCATACCGGC CCTTCTAGAACGTCTCAATCTC
em9	CATTCGCGTT	1620 ChD	GCTCAGATCAGCTCTTCATCCCATTCGCGTTCGGC CCTTCTAGAACGTCTCAATCTC
em11	AAACGTGCAG	1637 ChD	GCTCAGATCAGCTCTTCATCCAAACGTGCAGCGG CCCTTCTAGAACGTCTCAATCTC
em12	GTCAGATCCT	1638 ChD	GCTCAGATCAGCTCTTCATCCGTCAGATCCTCGGC CCTTCTAGAACGTCTCAATCTC
em13	GCGCTTTACA	1639 ChD	GCTCAGATCAGCTCTTCATCCGCGCTTTACACGGC CCTTCTAGAACGTCTCAATCTC
em14	GCCTCTACTC	1640 ChD	GCTCAGATCAGCTCTTCATCCGCCTCTACTCCGGC CCTTCTAGAACGTCTCAATCTC
em15	TACACCCTCG	1641 ChD	GCTCAGATCAGCTCTTCATCCTACACCCTCGCGGC CCTTCTAGAACGTCTCAATCTC
em16	AGCTGGATCT	1621 ChD	GCTCAGATCAGCTCTTCATCCAGCTGGATCTCGGC CCTTCTAGAACGTCTCAATCTC
em18	AGGGCCATTC	1642 ChD	GCTCAGATCAGCTCTTCATCCAGGGCCATTCCGGC CCTTCTAGAACGTCTCAATCTC
em19	ACCTGTTTCG	1643 ChD	GCTCAGATCAGCTCTTCATCCACCTGTTTCGCGGC CCTTCTAGAACGTCTCAATCTC
em20	AGCCTCGTCG	1644 ChD	GCTCAGATCAGCTCTTCATCCAGCCTCGTCGCGGC CCTTCTAGAACGTCTCAATCTC
em21	GTGATTCCGT	1645 ChD	GCTCAGATCAGCTCTTCATCCGTGATTCCGTCGGC CCTTCTAGAACGTCTCAATCTC
em22	GATGAGGATA	1646 ChD	GCTCAGATCAGCTCTTCATCCGATGAGGATACGG CCCTTCTAGAACGTCTCAATCTC
em23	TCTCGAGATC	1647 ChD	GCTCAGATCAGCTCTTCATCCTCTCGAGATCCGGC CCTTCTAGAACGTCTCAATCTC
em24	ACACGCAAGT	1648 ChD	GCTCAGATCAGCTCTTCATCCACACGCAAGTCGG CCCTTCTAGAACGTCTCAATCTC

Continuation of Table S1

em25	CCCGCGTTAG	1622 ChD	GCTCAGATCAGCTCTTCATCCCCCGCGTTAGCGGC CCTTCTAGAACGTCTCAATCTC
em26	AACGTCCGCA	1649 ChD	GCTCAGATCAGCTCTTCATCCAACGTCCGCACGGC CCTTCTAGAACGTCTCAATCTC
em27	ACTTGTTGCT	1650 ChD	GCTCAGATCAGCTCTTCATCCACTTGTTGCTCGGC CCTTCTAGAACGTCTCAATCTC
em28	GGTCCTGACC	1651 ChD	GCTCAGATCAGCTCTTCATCCGGTCCTGACCCGGC CCTTCTAGAACGTCTCAATCTC
em30	TCATAGCGTC	1653 ChD	GCTCAGATCAGCTCTTCATCCTCATAGCGTCCGGC CCTTCTAGAACGTCTCAATCTC
em31	GTGCTAGGCT	1654 ChD	GCTCAGATCAGCTCTTCATCCGTGCTAGGCTCGGC CCTTCTAGAACGTCTCAATCTC
em33	GCAAAGGTAA	1604 ChD	GCTCAGATCAGCTCTTCATCCGCAAAGGTAACGG CCCTTCTAGAACGTCTCAATCTC
em34	ACCGTGGTAT	1632 ChD	GCTCAGATCAGCTCTTCATCCACCGTGGTATCGGC CCTTCTAGAACGTCTCAATCTC
em36	ATAGCTATAG	1652 ChD	GCTCAGATCAGCTCTTCATCCATAGCTATAGCGGC CCTTCTAGAACGTCTCAATCTC
em37	CAATCTCATA	1623 ChD	GCTCAGATCAGCTCTTCATCCCAATCTCATACGGC CCTTCTAGAACGTCTCAATCTC
em39	TATGTGAATC	1624 ChD	GCTCAGATCAGCTCTTCATCCTATGTGAATCCGGC CCTTCTAGAACGTCTCAATCTC
em48	GCGACTCCCT	1625 ChD	GCTCAGATCAGCTCTTCATCCGCGACTCCCTCGGC CCTTCTAGAACGTCTCAATCTC
em56	GCTTTCTGGA	1626 ChD	GCTCAGATCAGCTCTTCATCCGCTTTCTGGACGGC CCTTCTAGAACGTCTCAATCTC

Table S2 Molecular compositions and DNA barcodes used for the experiment of Fig. 4. G3P = glycerol-3-phosphate; L-ser = L-serine; bme = beta-mercaptoethanol.

Sample #	**PURE system variant**	**Additives ("-" means lacking)**	**Expected phenotype DNA replication**	**Expected phenotype PS synthesis**	**Barcode**
1	M1	All	+	+	em0
2	M1	-dNTPs -SSB -DSB	-	+	em3
3	M1	-oleoyl-CoA	+	-	em6
4	M1	-CTP -G3P -L-ser -bme	+	less	em9
5	M1	-dNTPs -SSB -DSB -oleoyl-CoA	-	-	em16
6	M1	-ribosomes -SSB -DSB	-	-	em25
7	M1	All, minimal incubation (20 min)	-	-	em33
8	Top day 5	All	unknown	unknown	em37
9	Top day 8	All	unknown	unknown	em39
10	PURE*frex* 2.0	All	unknown	unknown	em48
11	M1	All	+	+	em56

Table S3 Swelling solutions used for the experiment of Fig. 4, with M1 PURE system.

Sample # >		#1 standard, M1	#2 no DNArep	#3 no Plsyn	#4 less Plsyn	#5 no both	#6 no both	#7 no both	#11 standard, M1
	Final conc (standard)		no dNTP, SSB, DSB	no oCoA	no CTP, G3P, L-ser, b-mercapt	no dNTP, oCoA, SSB, DSB	no ribosomes, SSB, DSB	time0	standard M1
PURE Sol. I (B) new batch (tRNAs, substrates, AAs, NTPs)		7.24	7.24	7.24	7.24	7.24	7.24	7.24	7.24
PURE Sol. II (E) new batch (proteins)		4.2	4.2	4.2	4.2	4.2	4.2	4.2	4.2
PURE Sol. III (R) new batch (ribosomes)		1.33	1.33	1.33	1.33	1.33		1.33	1.33
SUPERase (20 U/µL)	0.5 U/µL	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Mix of CTP (20 mM), G3P (10 mM) and L-serine (10 mM)	1 mM CTP, 0.5 mM G3P, 0.5 mM L-ser	1	1	1		1	1	1	1
300 mM beta-mercaptoethanol (195.8 uL mQ + 4.2 uL bME) *	5 mM	0.3334	0.3334	0.3334		0.3334	0.3334	0.3334	0.3334
dNTP mix (10 mM)	300 µM	0.6		0.6	0.6		0.6	0.6	0.6
Ammonium sulphate (800 mM)	20 mM	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
SSB (p5, 10 mg/mL)	0.75 mg/mL	1.5		1.5	1.5			1.5	1.5
DSB (p6, 7 mg/mL)	0.21 mg/mL	0.6		0.6	0.6			0.6	0.6
DNA	500 pM								
MQ + DNA stock of barcoded to final conc 200		2.1966	4.8966	2.1966	3.53	4.8966	5.6266	2.1966	2.1966
		17.8034	15.1034	17.8034	16.47	15.1034	14.3734	17.8034	17.8034
Total		20 µL	20 µL	20 µL	20 µL	20 µL	20 µL	20 µL	20 µL

Table S4 Swelling solutions used for the experiment of Fig. 4, with PURE system variants other than M1: PURE variant "Top Day 5" (sample 8), "Top Day8" (sample 9) and PURE*frex* 2.0 (sample 10).

Sample # >	#8 Top Day5	#9 Top Day8	#10 PUREfrex
PURE Sol. I (B)	7.37	10.97	10
PURE Sol. II (E)	0.95	0.47	1
PURE Sol. III (R)	0.666	1.332	2
SUPERase (20 U/µL)	0.5	0.5	0.5
Mix of CTP (20 mM), G3P (10 mM) and L-serine (10 mM)	1	1	1
300 mM beta-mercaptoethanol (195.8 uL mQ + 4.2 uL bME) *	0.3334	0.3334	0.3334
dNTP mix (10 mM)	0.6	0.6	0.6
Ammonium sulphate (800 mM)	0.5	0.5	0.5
SSB (p5, 10 mg/mL)	1.5	1.5	1.5
DSB (p6, 7 mg/mL)	0.6	0.6	0.6
DNA			
MQ + DNA stock of barcoded to final conc 200	5.9806	2.1946	1.9666
	14.0194	17.8054	18.0334
Total	20 µL	20 µL	20 µL

Table S5 DNA oligos. "(5'P)" indicates 5' phosphorylation.

Oligo name	Oligo sequence (5' to 3')
756 ChD	ACTTCGCCTTTTTCACGCCC
1138 ChD	GCTCAGGGTCGGGAACACTTTC
1523 ChD	(5'P)-AAAGTAAGCCCCCACCCTCACATGATACC
1524 ChD	(5'P)-AAAGTAGGGTACAGCGACAACATACACCATTTCC
1525 ChD	AAAGTAAGCCCCCACCCTCACATGATACC
1526 ChD	AAAGTAGGGTACAGCGACAACATACACCATTTCC
1567 ChD	AAAGTAGGGTACAGCGACAACATACACCATTTCCCCATTGACCG
1605 ChD	GCATCAGTCAGCTCTTCAGGAGCCGCGAATTCTGTCGCATGTG
1627 ChD	(5'P)-AAAGTAAGCCCCCACCCTCACATGATACCATTCTCCTAATATCGACATAATCCG
1628 ChD	GCATCAGTCAGCTCTTCATGCCGCCGCGAATTCTGTCGCATGTG

Table S6 Molecular composition of swelling solutions dispensed by the Echo Liquid Handler (Beckman Coulter). SSB and DSB were not dispensed by the machine but added manually afterwards. The final concentrations of the PURE system components cannot be disclosed for confidentiality reasons with GeneFrontier.

Components	**Final concentration**	**1X standard reaction (µL)**
HEPES-KOH		1.55
Potassium glutamate		0.5
Magnesium acetate		0.35
Dithiolthreitol		0.425
Spermidine		0.3
FTHF		0.625
Creatine phosphate		0.35
Amino acid group 1		1.25
Amino acid group 2		0.825
Amino acid group 3		0.425
ATP, GTP		0.4
CTP, UTP		0.275
tRNA mix		0.25
Aminoacyl-tRNA synthetases, MTF		0.625
Initiation factors (IFs)		0.625
Elongation factors (EFs)		0.625
Release factors (RFs)		0.6
T7 RNA polymerase		0.6
NDK-MK-CK-PPiase		0.6
Factors		0.6
Ribosomes		1.375
Mixture of CTP (20 mM), G3P (10 mM), and L-serine (10 mM)	1 mM CTP, 0.5 mM G3P, 0.5 mM L-ser	1.1
300 mM beta-mercaptoethanol	5 mM	0.375
dNTP mix (10 mM)	300 µM	0.65
Ammonium sulphate (800 mM)	20 mM	0.5
SSB (p5, 10 mg/mL)	0.75 mg/mL	0.775
DSB (p6, 7 mg/mL)	0.21 mg/mL	0.225
Linear DNA template (5 nM)	500 pM	2.2
MilliQ-water		2.975
Total		**22 µL**

6

Perspective on laboratory evolution of synthetic cells

In this chapter, we discuss our perspective on the laboratory evolution of synthetic cells. The chapter consists of three parts. First, we shortly reflect on the key findings in this thesis and highlight remaining challenges. We define three major methodological improvements for future research, which promote a more holistic variant characterization and analysis within the laboratory evolution workflow to strengthen the knowledge gained from each experiment. Second, we describe how these improvements could be implemented in the context of a specific case study: the integration of DNA replication, phospholipid synthesis and a membrane contractile ring. Third, we outline the challenges and potential extensions of the methods discussed thus far that would be required when the synthetic genome size further increases to over 150 genes, the estimated minimum number of genes required for a minimal, synthetic cell.

Part 1: Where do we stand and what lies ahead?

The ultimate engineering approach to understanding the smallest unit of life – the cell – is the building of a minimal, synthetic version of a cell. Because no defined blueprint exists, we postulate that rational design must be complemented by a randomness-leveraging research method with a high exploratory power: laboratory evolution. In this thesis, we developed tools for the laboratory evolution of synthetic cells as a means to ultimately build a minimal cell that can self-replicate.

The large search space available for tuning the cell's genome and/or its molecular hardware calls for a targeted diversification approach. Here, *hardware* refers to the compositions of the cytoplasm (mainly the cell-free expression system) and the membrane. Moreover, we foresee epistatic interactions between (i) mutations in the synthetic genome (**Chapter 3**), (ii) between components of PURE system[1], and (iii) between the genome sequence and molecular composition (**Chapter 5**) that influence the fitness of the synthetic cell in unpredictable ways. Therefore, combinatorial testing will be crucial to effectively screen the vast design space for high-fitness variants.

Chapter 2 introduced a new protocol for the construction of targeted combinatorial plasmid libraries named MOSAIC. Using this protocol, DNA libraries were constructed for a minimal DNA replicator and phospholipid synthesis pathway to vary the expression levels of the involved genes. By expressing these libraries in liposomes in **Chapter 3**, we found that tuning protein translation initiation rates is effective in modulating the activity of the DNA-encoded protein systems at the system's level. After enriching the top performing variants with highest activities through *selection* or fluorescence-activated cell sorting (FACS)-based *screening*, we reconstructed experimental fitness landscapes mapping DNA sequences to the module activities. These landscapes provide a basis for the integration of these modules with each other and with additional protein systems.

The combinatorial DNA library approach allowed for studying the effects of combinations of mutations. While general insights into the occurrence of epistatic effects were gained, the generation of quantitative sequence-fitness mapping datasets with wider dynamic ranges for fitness (as discussed later) might in the future enable the identification of specific epistatic interaction partners and the quantification of their interaction strengths.

In **Chapter 4**, we expanded the FACS-based screening and sorting method with a microscopy-based pre-screening step, which enabled the sorting of vesicles based their spatiotemporal properties. The biggest challenge for this method is the screening throughput, which decreases as the information used to define the sorting function is richer. Generally, 200,000 variants can be screened in 24 hours of microscope time (i.e., 2 million liposomes of which 10% are active). This is relatively low in comparison to FACS (10^6 to 10^7 variants per day). Thus,

the method's unique strength of handling complex phenotypes involves a tradeoff with throughput.

In the future, we would like to couple the image-based sorting workflow to the laboratory evolution strategy designed and validated in **Chapter 5**, which enables the co-optimization of the DNA sequence and the molecular hardware of synthetic cells. Using this strategy, we estimate that up to 100 molecular variants can be screened in combination with 10^3 (image-based screening) or 10^4 to 10^5 DNA variants (FACS-based screening). While improving the throughput of the microscopy-based screening technology may enable the screening of larger libraries, smart library design choices will remain essential to avoid exhaustive scanning of too many variants. In addition, active learning is an effective strategy to bridge the gap between the large search space and the relatively low experimental capacity[2–6].

What are the upcoming challenges?

While sorting and genotyping only top performing variants can be useful to direct systems towards a target property, it may fail to explain *why* certain variants outperform others. In my view, moving away from a rational engineering approach – as we do with a randomness-leveraging research method such as directed evolution – can only be justified if the mechanisms underlying the improvements can be discovered; we should not only use laboratory evolution to *obtain* synthetic cells that perform desired functions, but also to *increase our understanding* of the processes of interest, one of the primary reasons for constructing a minimal synthetic cell after all.

Beside this more fundamental reason, increasing the knowledge gained from each design-build-test cycle may also accelerate building a synthetic cell. Library design choices could be more effectively informed by previous testing rounds using quantitative models describing the sequence-fitness landscape[7–9]. Such models can propose variants to test that would either navigate to high-fitness optima (i.e., exploitation) or focus on deeper exploration of the landscape by targeting regions of high model uncertainty. The predictive models are more accurate if the datasets contain accurate quantitative fitness data on a diverse sequence space[7–9]. Including quantitative fitness measures for medium- and low-fitness variants, and not only the highest-fitness variants, can be important, as they help to define the topology of the fitness landscape and allow for calculating epistatic interactions, which is needed to efficiently navigate sequence-fitness landscapes[9,10]. Thus, collecting better informative datasets in each screening round could impact our next hypotheses and avoid wasted experimental efforts, thereby reducing the number of engineering rounds required to reach the final engineering goal, increasing the success rate of building a synthetic cell.

We can increase the knowledge gained from each test cycle by aiming for a holistic and system-level analysis of the screened library. We propose three improvements:

1. **Sort multiple library subfractions** to capture the full population – not only the best-performing ones – to reconstruct more accurately the variants' fitness distributions[9] and gain insights also into lower-fitness variants.
2. Deepen synthetic cell characterization by quantitative analysis of its full molecular content using **proteomics, transcriptomics, metabolomics and lipidomics** – maybe even at different points in time.
3. **Optimize the DNA recovery method** to enable more accurate quantification of variant frequencies.

In Part 2 of this chapter, we discuss how these improvements can be implemented in the light of a specific application.

Part 2: Methodological improvements applied to a case study: a DNA-encoded cell cycle

Integration of three modules: DNA replication, phospholipid synthesis and membrane constriction

We envision the next optimization target to be a three-module template encoding minimal protein systems for DNA replication[11], phospholipid synthesis[12], and membrane constriction[13] (**Fig. 1a**). We aim to integrate and optimize these modules in concert, even before each individual module is fully established in isolation, as the processes are likely to influence each other once combined[14]. The integrated modules will form the basis of an encoded minimal cell-cycle. Previously, DNA replication and phosphatidylserine (PS) synthesis were integrated using a six-gene DNA template[15]. However, when expressing the genes with PURE system in liposomes, the fraction of liposomes exhibiting the combined phenotype was small, with most liposomes showing only one active module. The fraction of liposomes with PS production was reduced by 50% upon co-expression of the genetic module for DNA replication. Making PS synthesis more robust to the co-expression of other genes could be one of the optimization targets for the three-module template.

Recently, Ramirez Gomez *et al.* (Danelon and Koenderink labs, manuscript in preparation) integrated phospholipid synthesis with membrane constriction through co-expression of the PS synthesis pathway with the protein FtsA and fusion protein FtsZ-mVenus. Upon GTP hydrolysis, FtsZ polymerizes into protofilaments, which are anchored to the liposome membrane by FtsA and form proto-rings that constrict liposome membranes generating budding vesicles with elongated membrane necks[16,17]. However, this system too yielded only a small fraction of liposomes with the combined phenotype. An FtsZ concentration above its critical concentration of 1.5-2.5 μM is required to establish polymerization *in vitro*[18,19]. This

concentration is high considering that the total protein yield of PURE system is 3 to 5 μM in bulk reactions[20,21] and varies between 1 and 12 μM in a clonal population of liposomes[22]. For comparison, the total protein concentration in *Escherichia coli* is approximately 4 mM[23]. Thus, the expression of many other genes besides *ftsZ-mVenus* at *functional* levels in the same PURE reaction may be challenging. Furthermore, tuning the FtsZ-to-FtsA protein ratio to approximate the *in vivo* protein ratio of 3:1 to 5:1[24] may be key for successful membrane constriction.

In summary, overcoming the impairing genetic-context effects and optimizing protein concentrations and ratios would be the first optimization targets for the three-module system. In the next sections, we discuss what DNA and molecular libraries could be designed for this template, and how the libraries can be screened, sorted and sequenced.

DNA template

The wild-type three-module DNA template with eight genes has already been constructed via Golden Gate Assembly of three parts (**Fig. 1b**). During assembly, it was barcoded to enable the (co-)optimization of its sequence and the molecular composition of synthetic cells (this method was described in Chapter 5). The correctly assembled template (13.1 kb) was amplified from the assembly reaction (**Fig. 1c**) and its sequence was verified by long-read nanopore sequencing. The genes from the template were expressed in bulk PURE reactions in the presence of tRNA preloaded with fluorescently labelled lysine. The sizes of the proteins produced by PURE system were analyzed by SDS-PAGE, which confirmed that all genes from the DNA template were expressed (**Fig. 1d**).

Because the three modules were derived from three different plasmids (**Fig. 1b**), a DNA library can be prepared for each module prior to combinatorial assembly, for instance to modulate the expression of the eight genes. The knowledge gained in Chapter 3, where we screened ribosomal binding site (RBS) libraries for the DNA replication and PS synthesis modules, can help us design small but smart DNA libraries. Chapter 3 showed that varying translation initiation rates effectively modulated the modules' functions at the system's level. For the DNA replication module, changing the RBS of TP has a wider dynamic range in DNA replication fold than DNAP, and therefore the TP RBS could be a better target to modulate DNA replication than the DNAP RBS. Because FtsZ-mVenus should be produced in relatively high quantities, the library could be designed to upregulate FtsZ-mVenus, while downregulating other genes. Especially, *plsC* expression could be downregulated relative to the other phospholipid synthesis enzymes, since a low amount of produced PlsC did not impair PS synthesis (Chapter 3). Beside RBSs, other possible targets include promoters[25,26], promoter insulators (i.e., spacers upstream a promoter sequence)[27], operator sites, terminators, gene order and orientation, and operon (multicistronic) design[25,28].

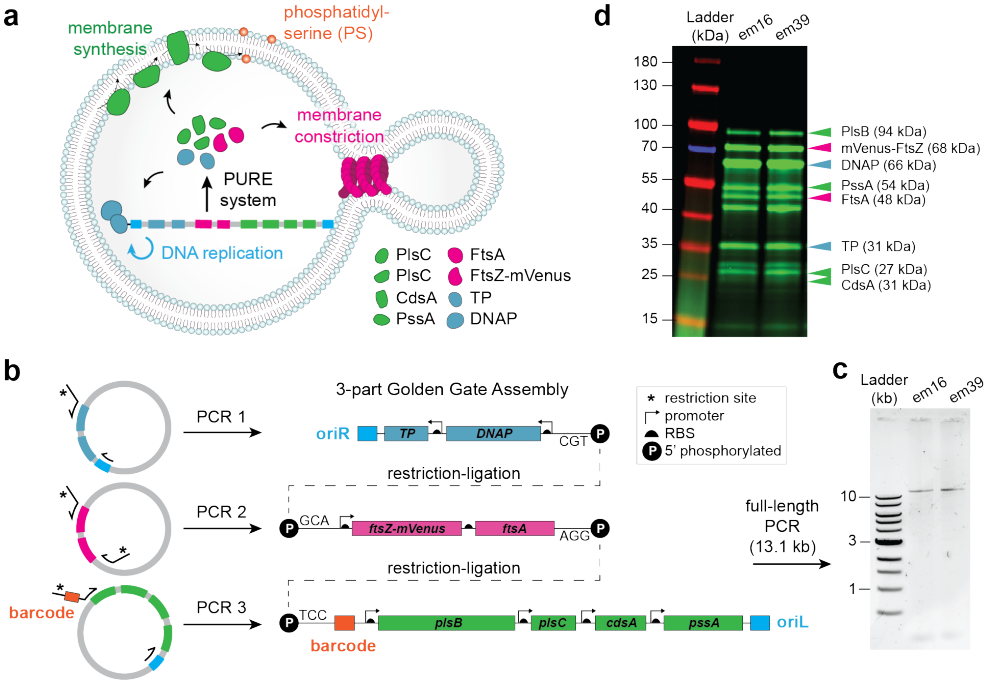


Fig. 1 Assembly and barcoding method for a three-module DNA template flanked by origins of replication. a Three-module DNA construct is expressed in liposomes by PURE system. The produced proteins act in DNA replication, lipid synthesis or membrane constriction. **b** The genes of interest are amplified by PCR from three plasmids. Plasmid 1: two genes for DNA replication. Plasmid 2: two genes for membrane constriction. Plasmid 3: four genes for phospholipid synthesis. Plasmid 1 and 3 each contain one origin of replication (oriR or oriL). Linear amplicons are treated with the type IIS restriction enzyme BspQI-HF to generate compatible overhangs on each template for subsequent ligation. Note that *ftsZ-mVenus* and *ftsA* form an operon under the control of one T7 promoter. **c** PCR amplification of the three-module template from oriR to oriL (13.1 kb) was confirmed by agarose gel electrophoresis. **d** Fluorescence imaging of an SDS-PAGE loaded with bulk PURE reactions (no liposomes) from the assembled DNA template and GreenLys protein labeling reagent. Because the GreenLys reagent contains tRNAs charged with fluorescently labelled lysine, the proteins produced by PURE system were fluorescently labelled. One reaction contained the DNA template with barcode em16 and the second reaction the template with em39 (sequences are reported in Table S1 in Chapter 5).

Such targeted libraries can be constructed using PCR and/or cloning. For example, in Chapter 4, we mutated a single site on a plasmid by PCR-amplification of the full plasmid sequence using a mutagenic primer set, followed by re-circularization. When mutations should be incorporated at multiple target sites, MOSAIC could be a more efficient alternative. With MOSAIC, 1- to 18- nucleotide substitutions, insertions and deletions can be performed[29], as demonstrated in Chapter 2, and possibly also larger modifications (this remains to be validated). If sequences of 70 nucleotides long containing a transcription terminator of one gene and a promoter of the downstream gene can be efficiently deleted using MOSAIC, operons could be created via this mutation technique. Lastly, the gene order and orientation

could be changed through combinatorial Golden Gate Assembly of fragments containing multiple pairs of restriction sites[30].

An important consideration when selecting a library construction method is the editing efficiency. When using MOSAIC in Chapter 3, the 50% per-site editing efficiency resulted in libraries that contained plasmids with all possible combinations of mutated and wild-type target sites. For example, the DNA replicator RBS library comprised roughly one-third molecules with two edited RBSs, one-third with one edited and one wild-type RBS, and one-third fully wild-type. While this feature facilitates epistatic effect calculations based on separate experimental fractions, it also increases the total library size, which could hinder exhaustive screening.

Molecular library design

The solution encapsulated in the liposomes consists of many components whose initial concentration could be varied. PURE system contains ribosomes and 36 proteins, including 20 aminoacyl tRNA synthetases, ten initiation, elongation and release factors, four enzymes for energy regeneration, methionyl-tRNA-formyltransferase and T7 RNA polymerase[31]. The PURE system buffer consists of HEPES-KOH, potassium glutamate, magnesium acetate, 20 amino acids, tRNA mix, creatine phosphate, spermidine, reducing agent, four rNTPs and 10-formyl-5,6,7,8-tetrahydrofolic acid. Additional components can be supplemented to PURE system, such as chaperones to promote protein folding, and elongation factor P (EF-P) to alleviate ribosome stalling caused by consecutive proline residues[32]. The variable space also includes all essential substrates and cofactors required for the activities of the produced proteins, including the fatty acid chain substrate, glyceraldehyde-3-phosphate and L-serine (for phospholipid synthesis), and dNTPs, ammonium sulphate and single- and double-stranded DNA binding proteins (for DNA replication). Lastly, the initial phospholipid composition of the liposomes could be varied as this may influence membrane-associated processes. For instance, the fraction of PG in large unilamellar vesicles can influence the association-dissociation of PssA to the membrane and thereby modulate PssA activity[12].

The total set thus comprises at least 76 components. The initial set of molecular compositions could be generated by starting from a reference composition encompassing the regular PURE system composition and varying the concentration of components. If for each component, three concentrations were to be tested, this would lead to an extremely large combinatorial space of 3^{76} variants. One way to reduce the combinatorial space is to restrict the combinations to 2-body interactions, which was found earlier to provide a good approximation of epistatic effects in PURE system[1]. This reduces the library size to 3×10^4 variants (Eq. 1). Even when some components are grouped to reduce the number of components as done in an earlier work[6], for instance grouping the amino acids, tRNA synthetases and enzymes involved in energy regeneration, the resulting combinatorial set on the order of 10^3 variants is currently still prohibitive to exhaustive experimental testing in a single screen. However, repeated screening

of a fraction of this variable space, where the design of each test set is informed by the results from the previous rounds, is promising to effectively screen the variable space[2–6].

For 76 components, the number of 2-body interactions × number of variants per interaction =

$$\left(\frac{76\times75}{2}\right) \times (3 \times 3) = 3 \times 10^4\ variants \quad \text{(Eq. 1)}$$

Fluorescent reporters of module functions

Liposomes expressing the three genetic modules could be characterized based on four fluorescent readouts: DOPE-Cy5 for membrane morphology, LactC2-mCherry to quantify the amount of produced PS, fusion protein FtsZ-mVenus for its expression and recruitment to the membrane, and a double-stranded DNA-intercalating dye to quantify DNA replication (**Fig. 2a**). While we previously used dsGreen to stain double-stranded DNA[15] (Chapters 3 and 5), its excitation/emission spectra overlap with those of mVenus. The use of alternative dyes should be tested, for instance DAPI or DRAQ5[33]. In the next sections, we discuss how these fluorescent readouts can be used for library screening and sorting. Given the tradeoff between throughput and the ability to assess spatiotemporal features, we think an effective strategy is to begin with high-throughput screening by FACS, and hereafter apply image-based screening to optimize for specific spatiotemporal properties.

Screening and sorting by FACS

First, high-throughput screening and sorting could be performed using FACS. The goal here is to obtain as many liposomes as possible with the three phenotypes: positive signals for LactC2-mCherry (PS synthesis), DAPI/DRAQ5 (DNA replication) and FtsZ-mVenus. Three-module activity would suggest that *ftsZ-mVenus* is expressed and that the genes involved in PS synthesis and DNA replication are expressed at *functional* levels enabling module activity. Eventually, we would be satisfied with a fraction of 1-10% of the liposome population having the combined phenotype, as such fraction sizes were obtained previously for YFP expression[34] and the phospholipid synthesis module at 10 pM DNA concentration (λ = 0.2 DNA molecules per liposome, on average) (Chapter 3).

Generally, with FACS screening based on a single readout, enrichment can be achieved using a single sorting gate (usually liposomes with highest fluorescence are sorted), or multiple gates[9,35] (**Fig. 2b**). Sorting one population (as performed in Chapter 3) is the simplest and most practical approach. This will allow for the identification of best performing variants, with the assumption that the variant distributions have similar shapes and dispersions (e.g., similar standard deviation)[9]. If the variability differs among variants, this method could limit accurate quantitative sequence-function mapping. Also, as can be seen from **Fig. 2b**, sorting only the population with the highest fluorescence will not inform on the variants that fall outside this gate. Higher-resolution quantitative data could be obtained by sorting multiple populations and, after DNA sequencing, statistically inferring a quantitative fitness value for each variant based on the variant's relative per-bin read counts[9,35]. This multiple-gate method could also

enable the quantification of the variability of each variant in the library[9]. While a larger number of sorted populations increases the efforts and costs for DNA recovery and sequencing, sequencing costs can be minimized by adding a gate-specific barcode at the end of the DNA template during DNA recovery by PCR to identify the sorted fraction, whereafter all samples are pooled and sequenced at once.

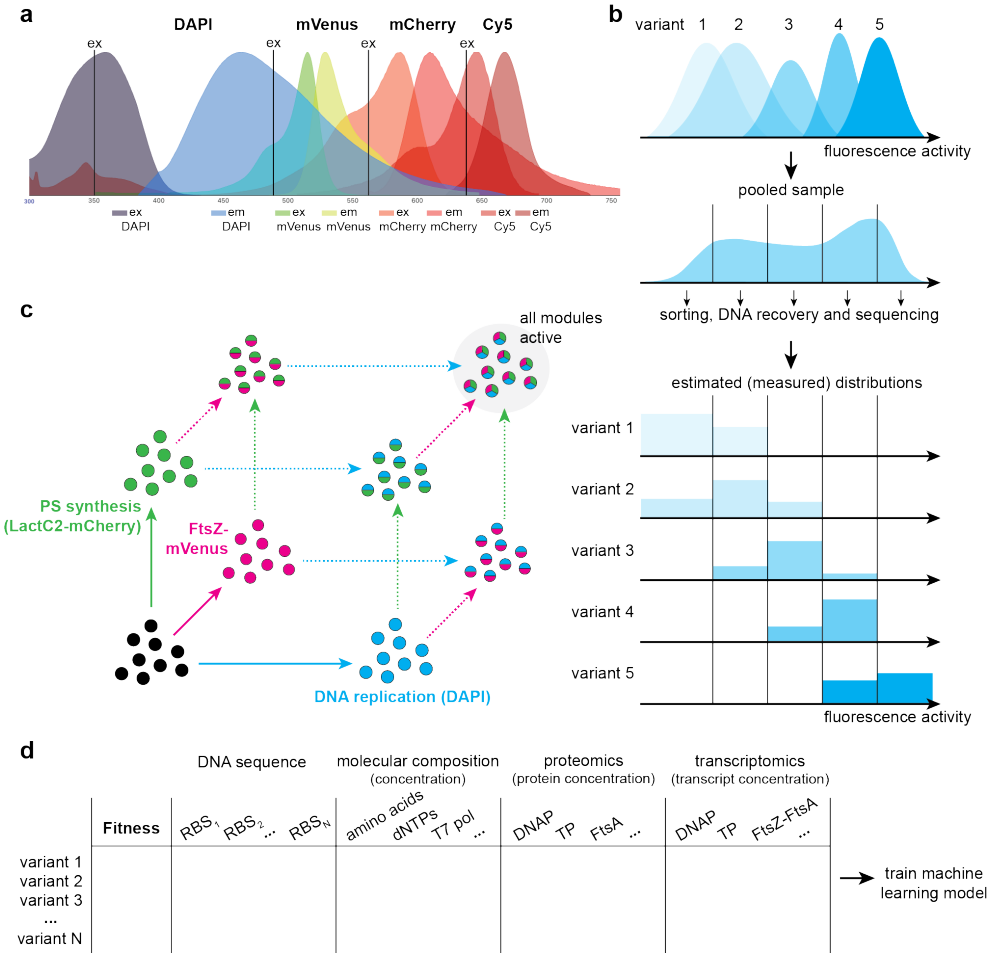


Fig. 2 Screening and sorting strategies for the three-module system from Fig. 1. a Excitation and emission spectra for DAPI, mVenus, mCherry and Cy5, obtained from fpbase.org/spectra/. Suggested lasers for excitation, indicated with 'ex', are examples from the spectral BD FACSDiscover S8 flow cytometer. **b** Sorting scheme for a one-dimensional readout. If only the population with the highest fluorescence activity were sorted (variant 5), the information about all other variants 1-4 would be lost. Figure inspired from Peterman and Levine *BMC Genomics* 17:206 (2016). **c** Schematic of a three-dimensional FACS plot. **d** Example of a dataset generated by screening, followed by genomics, proteomics and transcriptomics. In this example, RBSs are mutated in the DNA sequence.

Our DNA template of interest will not have one but three fluorescent readouts. Analogous to the gating strategies described above, the liposomes positive for all three readouts could be sorted, or additional populations could be sorted where only two, one or zero modules are active (**Fig. 2c**). After DNA sequencing of each sorted subpopulation, the genetic and molecular compositions leading to each phenotype will be revealed. Comparing subpopulations will give more insight into which factors are important for the (dis)ability of each module, informing us on promising variants for the next screening round. The sorting of multiple subfractions could be especially important when liposomes with the three combined phenotypes are initially absent. The fractions of liposomes with one or two phenotypes could be sorted first, and these variants could be diversified further aiming to establish the combined phenotype.

Quantitative variant-fitness datasets could be used to train machine learning models to predict the fitness of unscreened variants within the variable space[7]. Using an active learning algorithm called Bayesian optimization, predictive machine learning models can be used to select the most informative variants to test in a new screening round to efficiently probe the variable space[3,8]. These models could be trained on data retrieved after library screening and sorting (**Fig. 2d**). By DNA sequencing, the inferred variant fitness, the variant's genetic mutations and molecular composition – identified by a DNA barcode as described in Chapter 5 – could be linked. In addition to DNA sequencing, performing proteomics, transcriptomics and lipidomics on the sorted liposome populations could give insight into various factors underlying the high- and low-fitness phenotypes, for instance gene expression patterns, including (relative) amounts of RNA and protein and whether there is a correlation between transcription and translation, production of truncated proteins, and bottlenecks in metabolic networks/pathways, e.g., the phospholipid synthesis pathway. By adding the multi-omics data to the machine learning training dataset, the features that contributed most to the model's predictions (feature importance in machine learning) could potentially give mechanistic insight into the variants' phenotypes. This could potentially increase our understanding of the system and accelerate building the synthetic cell.

Ultimately, *single-liposome* sorting followed by multi-omics analysis would provide a direct link between the genotype and the other cell omics profiles and would provide more detailed information on the sorted liposome subpopulations (e.g., shape and dispersion of the population distribution). However, the low amounts of protein, DNA, RNA and lipids in single liposomes could make this challenging.

Lastly, it is important to understand that, on top of sorting those variants selected by the researcher, in-liposome DNA replication by the Phi29 system will contribute to the enrichment of DNA molecules that promote DNA replication and not necessarily the activity of the other two modules. This feature could potentially interfere with quantitative sequence-function mapping. In that case, screening and sorting by FACS could be performed in absence of DNA

replication, for instance by not providing dNTPs. Such screening steps could be alternated with steps of self-selection of DNA replication, as performed in Chapter 3, which would serve to optimize the DNA replication module.

Image-based screening and sorting

Once the expression of seven out of the eight genes (not *ftsA*) has been established by FACS-based screening, and DNA replication and PS synthesis are functional, we could shift our goal to the recruitment of FtsA and FtsZ-mVenus to the membrane and membrane constriction. For this, the microscopy-based screening and sorting method described in Chapter 4 could be used. As discussed previously, the throughput of image-based screening is low at $\lambda < 1$ DNA molecules per liposome: 10^4 to 10^5 variants per day can be screened using image-based sorting, compared 10^6 to 10^7 variants per day using FACS. Therefore, for image-based screening, small but smart libraries should be designed with specific optimization targets. For example, the ratio of FtsZ-mVenus to FtsA would be a suitable target for this optimization stage, as a proper ratio of FtsZ to FtsA has been found to be critical for cell division *in vivo*[36].

A high-throughput sorting instrument that would be very useful at this stage would be the BD FACSDiscover S8[37–39], a spectral flow cytometer capable of making sort decisions in *realtime* based on images. During a trial experiment, we successfully sorted YFP-expressing liposomes from non-YFP-expressing liposomes using an imaging parameter of this system (**Fig. 3**). However, the imaging resolution of this system is currently not sufficient to visually inspect liposome morphologies and intravesicular protein localizations. Additionally, with only one image per vesicle, it is limiting in assessing dynamic behaviors. In summary, the lack of compatibility between the imaging resolution of this system and our sample type currently limits the application of this technology for high-throughput image-based sorting.

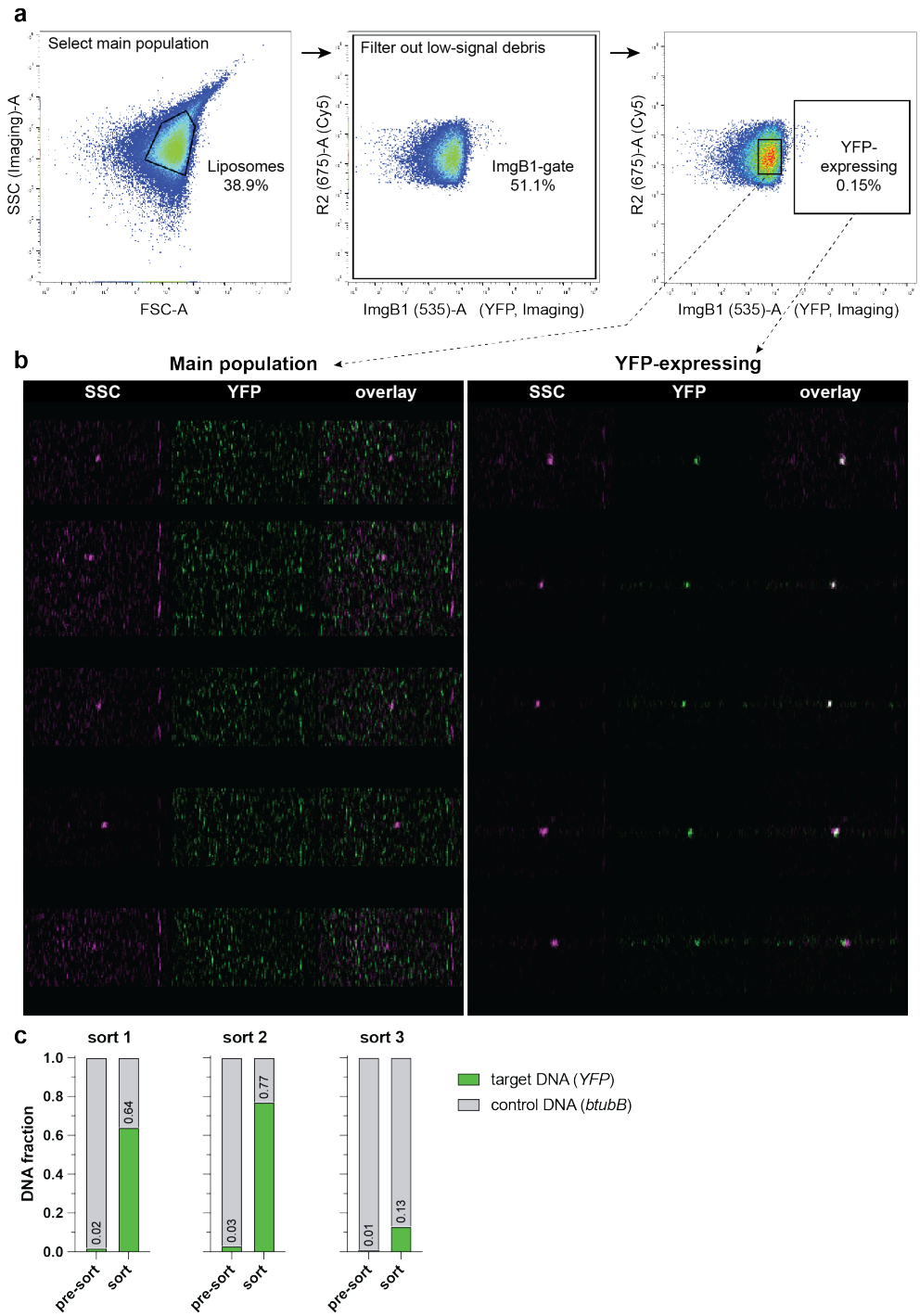


Fig. 3 Sorting YFP-expressing liposomes from a mock library using the imaging spectral flow cytometer BD FACSDiscover S8 Cell Sorter. Two liposome populations were mixed at a 1:100 ratio. The rare population consisted of liposomes expressing the fluorescent YFP protein, while the population in

excess consisted of liposomes expressing a nonfluorescent control protein (BtubB). The liposome membrane dye (DOPE-Cy5) was used to trigger the recording of events. Using the imaging parameter ImgB1 (535)-A, we sorted ~2,000 liposomes aiming for high YFP expression. **a** FACS plots showing the gating strategy applied to the BD FACSDiscover S8 data. **b** Example images from FACSDiscover S8 from the main population (left) and the YFP-expressing population (right). The imaging data acquired in *realtime* confirmed that the gated liposomes were expressing YFP in contrast to the main population. **c** Quantitative PCR analysis of sorted fractions showing an enrichment of *YFP* DNA upon sorting (n=3 sorts). Specifically, the fraction of *YFP* DNA increased from 0.02 ± 0.01 to 0.5 ± 0.3, an enrichment of 24 ± 10 fold (mean ± standard deviation, n=3 sorts).

DNA recovery

After liposome sorting, the DNA should be recovered from the liposomes. The full-length DNA (13 kb) could be amplified from the sorted liposome populations by PCR and sequenced by long-read nanopore sequencing as performed in Chapter 3. Amplification of long templates is often challenged by a low yield and a bias towards amplification of smaller fragments. Nevertheless, there are high-fidelity polymerases, such as the Quantabio RepliQa Hifi Toughmix, that can amplify up to ~20 kb fragments with minimal sequence bias for instance due to differences in GC content[40].

The implementation of unique molecular identifiers (UMIs) could further reduce bias in variant frequencies due to PCR amplification and sequencing[41–43]. Accurate variant counts are important for quantitative sequence-fitness mapping. The UMIs can be 10-20 nucleotides long and introduced at the end of the DNA sequence during PCR by using an overhang primer containing a degenerate sequence, e.g., *NNNN*… (20×*N*) (**Fig. 4a**). It is important that the number of unique primer overhangs exceeds the number of template molecules in the PCR reaction, ensuring that every DNA molecule receives a unique identifier. By grouping all sequencing reads with the same UMI, and considering each group as one variant count, the frequency of each variant in the original DNA population can be determined. By generating a consensus sequence from all grouped reads with the same UMI, this method can also be used to improve the sequencing accuracy of DNA library variants[44,45] (**Fig. 4b**). In Chapters 2-4, RBS variants were accurately identified by discarding reads containing low-quality bases in the target regions. However, this quality control method may not be feasible when dealing with many or large target regions (with in the extreme case random mutagenesis), where it would leave to few usable reads. This limitation could be overcome by the generation of consensus sequences using UMIs. Optionally, the UMIs could be built from homotrimeric nucleotide blocks to enable the identification of UMIs that contain a PCR or sequencing error (**Fig. 4c**)[46]. Lastly, the primer and UMI structure could be refined to minimize non-specific PCR products. For example, undesirable primer interactions have been reduced by designing stem-loop structures in the primer overhang containing the UMI[47].

If full-length DNA recovery of long templates (>10 kb) results in too low amplification yield or too few full-length sequencing reads, a "dictionary" could be made prior to DNA encapsulation

that stores the identity of the variants[48] (**Fig. 4d**). Prior to encapsulation in liposomes, the DNA templates would be labelled with UMIs (by PCR or ligation) and sequenced to link the molecular identifier to the genotype[48]. Then, after liposome sorting, only the UMI region would need to be amplified from the liposomes and sequenced. Here, short-read Illumina sequencing (~150-nucleotide reads) could be used instead of long-read nanopore sequencing, generally yielding a lower error rate and larger sequencing depth. Downsides of this method are: (i) mutations incorporated during in-liposome DNA replication cannot be detected, and (ii) the recovered DNA cannot be re-encapsulated in liposomes for additional screening rounds.

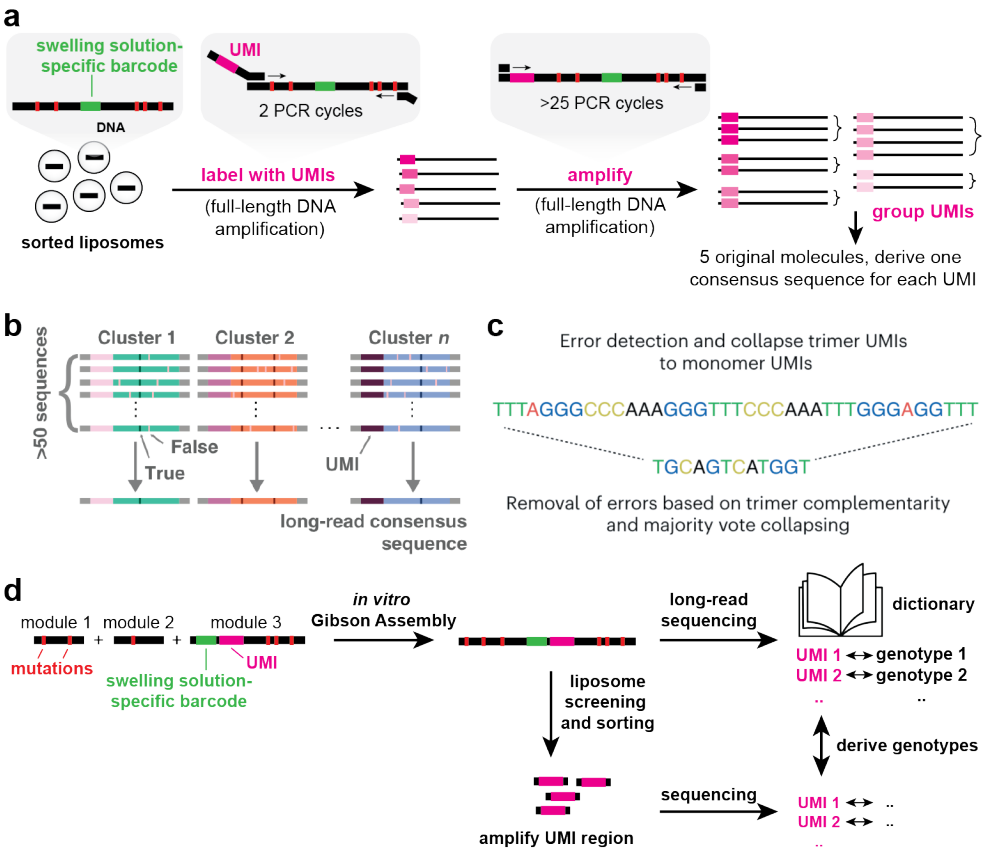


Fig. 4 Proposed strategies for DNA recovery from sorted liposomes and DNA sequencing using unique molecular identifiers (UMIs). a Recovery of the full-length DNA after liposome sorting using a primer containing an UMI in its overhang. Grouping of sequencing reads with identical UMIs improves variant frequency counts and sequencing accuracy. **b** Consensus sequences derived from sequencing reads with identical UMIs. Figure adapted from Zurek *et al. Nat. Comm.* 11:6023 (2020). **c** UMIs built up from homotrimeric nucleotide blocks (AAA, TTT, GGG and CCC) instead of single nucleotides improve the accuracy of identification of the UMI sequences. Figure adapted from Sun *et al. Nat. Methods* 21, 401–405 (2024). **d** Incorporation of the UMI before liposome screening and sorting allows for recovery of only the UMI region after sorting.

Part 3: Laboratory evolution of large (>150 genes) synthetic genomes

Ultimately, the three modules described in the Part 2 could be further extended by the addition of genes to expand the phospholipid synthesis pathway to produce phospholipids PE, PG and PC[12], or the addition of genes encoding single- and double-stranded DNA binding proteins, which are currently added as purified proteins due to the high amount required[11]. Furthermore, genes could be added to improve control over membrane constriction for generating equal-sized daughter cells (e.g., the Min system[17]), or to enable abscission[49]. Coordination of processes in time could be implemented, where DNA replication and phospholipid synthesis precede constriction and abscission. This could be implemented, for example, through feedback loops that trigger the expression of specific modules, or by timing processes via the accumulation of proteins to threshold levels that activate their functions[50]. Lastly, the genome should be expanded with RNA- and protein-coding genes required to regenerate PURE system, such as genes for RNA polymerase, translation factors, tRNA synthetases, tRNAs, energy regeneration enzymes and ribosomes[21,51–54]. As described in the Introduction of this thesis, prospective synthetic genomes of minimal cells will eventually contain at least 150 RNA- and protein-coding genes[55–57]. Laboratory evolution with such large genomes requires adaptations to the steps described until now which we shortly outline in this last part.

1. DNA and library synthesis

In vivo DNA assembly in *Saccharomyces cerevisiae* is a powerful method for the assembly of large DNA constructs (up to 1.66 Mb) thanks to *S. cerevisiae*'s robustness to heterologous DNA and its ability to carry large amounts of DNA and assemble many fragments at once (up to 44)[58]. Recently, this method has been used to build a partial synthetic genome of 41 kb which encodes 15 genes for DNA replication, membrane growth and cell division[59]. Initially, expression from this construct in PURE system was most likely impaired by a low DNA yield and purity. However, this was overcome by amplification of the yeast-assembled construct in *E. coli*, and expression of all genes in PURE system was confirmed by proteomics analysis[59].

DNA libraries can be constructed either prior to DNA assembly, as discussed earlier in Part 2 of this chapter, or after DNA assembly. After DNA assembly in yeast, the order and orientation of fragments could be varied by the method SCRaMbLE in yeast or *in vitro*[60]. Alternatively, insertions, deletions or edits could potentially be made in *E. coli* by MOSAIC, although the editing efficiency of MOSAIC to mutate large plasmids (>20 kb) remains to validated. A challenge that may arise using MOSAIC is *in vivo* recombination, especially for DNA constructs containing multiple repeated elements, such as similar regulatory sequences for different genes on the same construct. Diversification of regulatory sequences prior to assembly could potentially promote correct assembly, amplification and editing *in vivo*[55].

2. Screening

Upon expanding the number of genes in the synthetic genome, it is important to keep selecting for all functions in the system to preserve previous evolutionary gains. For large genomes with tens to hundreds of genes, it can be difficult to design a screening method for each newly integrated gene or function. Image-based phenotyping could be particularly helpful to classify liposomes based on many functions, because images can be rich in quantitative phenotypic parameters (Chapter 4). Moreover, aggregating the activities of many components into system-level readouts would also be essential to cope with the increasing number of functions. We exemplified such a system's level approach in this thesis, for example, by detecting the end-product of a phospholipid synthesis pathway, which was the result of successful expression and activity of four enzymes. As the synthetic genome grows, this readout could be conditional to an increasing number of activities in the system. For example, functional levels of expressed phospholipid synthesis enzymes could be dependent on successful regeneration of PURE components, including those required for transcription, translation, aminoacylation, or energy regeneration.

Final conclusion and outlook

The aim of this thesis was to explore directed evolution as research methodology to accelerate the building of a minimal, synthetic cell. This was achieved by building targeted combinatorial DNA libraries, subjecting these to various screening and selection methods, and exploring nanopore sequencing analysis and visualization methods for sequence-function mapping. This thesis exemplified how targeted combinatorial testing can be applied to the optimization of genetic modules by assessing biological functions at the system's level. In the future, synthetic cell research will largely benefit from increasing the knowledge gained from each prototyping round by technical methodological improvements and balancing exploration versus exploitation, impacting the library design, screening and analysis steps.

In Chapter 1, we referred to a minimal, synthetic cell as a unit capable of self-replication, self-maintenance and evolution. In this thesis, we mostly focused on directed evolution as an engineering tool for building up modules establishing autonomous self-replication and self-maintenance (metabolism). However, considering evolution as a key property of life, we could rethink the role of evolution within the synthetic cell framework: what can we learn from evolution as a more autonomous key property of life? As summarized in **Fig. 5**, less intervention by the researcher in evolutionary processes could be achieved by:

1. Instead of encapsulating pre-made DNA libraries, diversification could be performed *in situ* by the DNA polymerase during DNA replication[61–65].
2. Instead of screening and sorting variants based on the sort decision of the researcher, functions can be selected that are coupled to processes of the Central Dogma (transcription, translation and (other) processes that are required for DNA replication)[66].

3. Redistribution of sorted/selected DNA to new cells ("daughter cells") via DNA extraction and re-encapsulation can be replaced by semi-continuous processes by vesicle fusion-fission[61] or eventually full self-replication.

Random diversification may be inefficient for the evolutionary optimization of large DNA constructs compared to targeted diversification. Instead, the strength of a more "autonomous evolution mode" lies in the opportunity to study the evolutionary processes itself. For instance, the evolvability of the synthetic cell could be studied – involving which regions are more susceptible or more robust to mutations, evolutionary constraints, modularity, epistasis, robustness to changes in the environment – and whether the evolvability is evolvable[67]. Thus, while it is important to use directed evolution as an engineering tool to gain knowledge about self-replication and self-maintenance processes in the cell and accelerate its optimization towards the engineering goal, the evolutionary process itself can also be studied by implementing a more "autonomous evolution mode". The dual role of laboratory evolution of synthetic cells, depending on the choices that you make on diversification, selection and DNA progression, makes it a fascinating direction to further explore in the future!

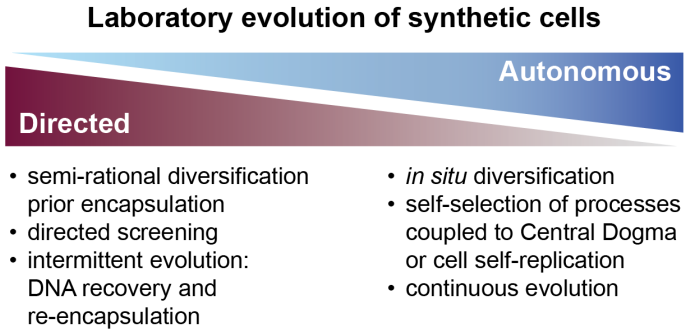


Fig. 5 Different methods for directed versus more autonomous laboratory evolution of synthetic cells.

Methods

DNA oligos and plasmids

DNA oligos (**Table 1**) were ordered from Ella Biotech. The synthesis of plasmid G616 was described in Chapter 5. Plasmid G613 encodes four genes for phosphatidylserine synthesis (*plsB*, *plsC*, *cdsA* and *pssA*) and *ftsZ-mVenus* and *ftsA*. G613 was constructed through *in vivo* assembly in *Saccharomyces cerevisiae* by Laura Sierra Heras (Danelon group, manuscript in preparation).

Table 1 DNA oligos. "(5'P)" indicates 5' phosphorylation.

Oligo name	Oligo sequence (5' to 3')
1523 ChD	(5'P)-AAAGTAAGCCCCCACCCTCACATGATACC
1524 ChD	(5'P)-AAAGTAGGGTACAGCGACAACATACACCATTTCC
1525 ChD	AAAGTAAGCCCCCACCCTCACATGATACC
1526 ChD	AAAGTAGGGTACAGCGACAACATACACCATTTCC
1628 ChD	GCATCAGTCAGCTCTTCATGCCGCCGCGAATTCTGTCGCATGTG
1629 ChD	GACCTGTGACAGCTCTTCGGCATGAATCAGTCCTACGTCAGCACAG
1630 ChD	CATGGTCACAGCTCTTCGGGACTCCCAGTCGGTGTAGCAGCAAT
1621 ChD	GCTCAGATCAGCTCTTCATCCAGCTGGATCTCGGCCCTTCTAGAACGTCTCAATCTC
1624 ChD	GCTCAGATCAGCTCTTCATCCTATGTGAATCCGGCCCTTCTAGAACGTCTCAATCTC

PCR amplification

The genes for DNA replication (*DNAP* and *TP*) were amplified from plasmid G616 by PCR using primer 1526 ChD and the overhang primer 1628 ChD carrying the restriction site for BspQI-HF. The genes for PS synthesis (*plsB*, *plsC*, *cdsA* and *pssA*) were amplified from plasmid G616 by PCR using primer 1525 ChD and an overhang primer carrying a barcode and the restriction site for BspQI-HF (1621 ChD or 1624 ChD). The genes for membrane constriction (*ftsZ-mVenus* and *ftsA*) were amplified from plasmid G613 by PCR using overhang primers 1629 ChD and 1630 ChD both carrying the restriction site for BspQI-HF. PCR was performed with RepliQa Hi-Fi ToughMix (QuantaBio) following the manufacturer's instructions using thermocycling conditions: 98 °C for 30 s, 30 cycles of (98 °C for 10 s, 68 °C for 5-10 s/kb). To digest the parent plasmids, 0.5 µL of restriction endonuclease DpnI (20,000 U/mL, NEB) was added per 50 µL PCR mix after thermocycling, and the solution was incubated for 1 h at 37 °C. DNA was purified from the PCR reagents using the QIAquick PCR Cleanup Kit (Qiagen). The sizes of the amplicons were verified by agarose gel electrophoresis (0.7% agarose in TAE buffer) using the 1 kb DNA Ladder (NEB). DNA concentrations were measured with the Qubit dsDNA High Sensitivity Assay Kit and the Qubit 4 Fluorometer (Invitrogen).

Golden Gate Assembly

For the 3-fragment Golden Gate Assembly, 20 ng of barcoded fragment encoding *plsB*, *plsC*, *cdsA* and *pssA*, 20 ng of the fragment encoding *DNAP* and *TP*, and 30 ng of the fragment encoding *ftsZ-mVenus* and *ftsA* were added to the assembly master mix. The master mix

contained 1 µL of BspQI-HF (10,000 U/mL, NEB), 0.5 µL of T4 DNA ligase (400,000 U/mL, NEB), 2 µL of 10× T4 ligase reaction buffer (NEB), 0.5 µL of DpnI (20,000 U/mL, NEB) and Milli-Q water up to 20 µL. The reaction mix was incubated for 1 h at 37 °C (restriction) followed by 15 min at 16 °C (ligation).

Amplification of the assembled DNA template

The full-length DNA (13.1 kb) was amplified from the assembly reaction by PCR with primers 1523 ChD and 1524 ChD using 2.5 µL of GGA product per 50 µL of PCR reaction volume. PCR was performed with RepliQa Hi-Fi ToughMix (QuantaBio) following the manufacturer's instructions using thermocycling conditions: 98 °C for 30 s, 30 cycles of (98 °C for 10 s, 68 °C for 5-10 s/kb). The amplicons were purified by gel extraction using the QIAquick PCR & Gel Cleanup Kit (Qiagen). Amplicon size (13.1 kb) and concentration were verified by agarose gel electrophoresis and the Qubit dsDNA High Sensitivity Assay Kit, respectively.

SDS-PAGE protein assay

FluoroTect GreenLys in vitro Translation Labeling System (Promega) was diluted 1:4 in Milli-Q water. PURE*frex* 2.0 reactions were prepared on ice following the manufacturer instructions, with the addition of 1 µL of the 1:4 diluted GreenLys labeling system and DNA at 10-20 ng/µL final concentration at a final volume of 20 µL. The samples were incubated at 37 °C for 4–8 h. The reactions were diluted 1:1 in Milli-Q and treated with 1 µL RNase A (4 mg/mL) at 37 °C for 1 h. Samples were prepared for SDS-PAGE by mixing 22 µL of GreenLys sample with 7.5 µL of preheated 4× Laemmli buffer (non-reducing) and 0.56 µL DTT (0.5 M), followed by incubation at 95 °C for 5 min to denature proteins. After brief centrifugation, the wells of a precast stain-free gel (4-15% Mini-PROTEAN TGX, Bio-Rad) were loaded with 5-12 µL of each sample and 4 µL of PageRuler Prestained Protein Ladder. The gels were run at 200 V for 30 min with the electrophoresis tank covered in aluminium foil to prevent photobleaching. The gels were rinsed in Milli-Q water for 30 min and imaged using the Typhoon Imager (Amersham Biosciences) with a 488-nm laser and a 520-nm band-pass emission filter. Next, the gels were stained with InstantBlue Coomassie protein stain overnight, rinsed in Milli-Q and imaged with a ChemiDoc Imaging System (Bio-Rad).

Imaging flow cytometry screening and sorting

Liposomes expressing *YFP* or *btubB* were prepared as described in Chapter 4 using a final DNA concentration of 5 nM. After gene expression, the two liposome populations were mixed at a 1:100 ratio. The rare population consisted of liposomes expressing the fluorescent YFP protein, while the population in excess consisted of liposomes expressing a nonfluorescent control protein (BtubB). The liposomes were measured using the BD FACSDiscover S8 Cell Sorter at an event rate of ~3,000-10,000 events per second and a flow rate of 1. Because the liposomes were covalently labelled with Cy5 (DOPE-Cy5), the R2 (675)-H signal was used for triggering event registration to distinguish liposomes from undesired events due to noise from the buffer

or instrument. We sorted ~2,000 liposomes with YFP expression, indicated by a high ImgB1 (535)-A signal. Quantitative PCR was performed as described in Chapter 4.

Data availability

Raw and processed data underlying this chapter, including nanopore sequencing, agarose gel electrophoresis, SDS-PAGE, imaging flow cytometry and quantitative PCR data are available in the 4TU.ResearchData repository at https://doi.org/10.4121/271054e7-0240-4966-9ac8-044c08a80ecc.

Acknowledgements

I thank Enrica Mammana for performing the Golden Gate Assembly experiments, Laura Sierra Heras for constructing plasmid G613 and Federico Ramirez Gomez and Yannick Bernard-Lapeyre for insightful discussions. We also thank Thierry Langlois and Damien Loras from BD Life Sciences – Biosciences for the possibility to test our liposome samples using BD FACSDiscover S8.

References

1. Matsuura, T., Kazuta, Y., Aita, T., Adachi, J. & Yomo, T. Quantifying epistatic interactions among the components constituting the protein translation system. *Mol. Syst. Biol.* **5**, MSB200950 (2009).
2. Caschera, F. *et al.* Coping with complexity: Machine learning optimization of cell-free protein synthesis. *Biotechnol. Bioeng.* **108**, 2218–2228 (2011).
3. Pandi, A. *et al.* A versatile active learning workflow for optimization of genetic and metabolic networks. *Nat. Commun.* **13**, 3876 (2022).
4. Borkowski, O. *et al.* Large scale active-learning-guided exploration for in vitro protein production optimization. *Nat. Commun.* **11**, 1872 (2020).
5. Zhu, J. *et al.* AI-driven high-throughput droplet screening of cell-free gene expression. *Nat. Commun.* **16**, 2720 (2025).
6. Bernard-Lapeyre, Y., Cleij, C., Sakai, A., Huguet, M.-J. & Danelon, C. Optimization of PURE system composition using automation and active learning. Preprint at https://doi.org/10.64898/2026.03.23.713685 (2026).
7. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687–694 (2019).
8. Yang, J. *et al.* Active learning-assisted directed evolution. *Nat. Commun.* **16**, 714 (2025).
9. Peterman, N. & Levine, E. Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations. *BMC Genomics* **17**, 206 (2016).
10. Johnston, K. E. *et al.* A combinatorially complete epistatic fitness landscape in an enzyme active site. *Proceedings of the National Academy of Sciences* **121**, e2400439121 (2024).
11. van Nies, P. *et al.* Self-replication of DNA by its encoded proteins in liposome-based synthetic cells. *Nat. Commun.* **9**, 1583 (2018).
12. Blanken, D., Foschepoth, D., Serrão, A. C. & Danelon, C. Genetically controlled membrane synthesis in liposomes. *Nat. Commun.* **11**, 4317 (2020).
13. Godino, E. & Danelon, C. Gene-directed FtsZ ring assembly generates constricted liposomes with stable membrane necks. *Adv. Biol.* **7**, 2200172 (2023).
14. Abil, Z. & Danelon, C. Roadmap to building a cell: an evolutionary approach. *Front. Bioeng. Biotechnol.* **8**, (2020).

15. Sierra, A. M. R., Gomez, F. R., Tongeren, M. van, Heras, L. S. & Danelon, C. A synthetic cell with integrated DNA self-replication and membrane biosynthesis. 2025.01.14.632951 Preprint at https://doi.org/10.1101/2025.01.14.632951 (2025).
16. Godino, E. *et al.* Cell-free biogenesis of bacterial division proto-rings that can constrict liposomes. *Commun. Biol.* **3**, 539 (2020).
17. Kohyama, S., Merino-Salomón, A. & Schwille, P. In vitro assembly, positioning and contraction of a division ring in minimal cells. *Nat. Commun.* **13**, 6098 (2022).
18. Mukherjee, A. & Lutkenhaus, J. Dynamic assembly of FtsZ regulated by GTP hydrolysis. *EMBO J* **17**, 462–469 (1998).
19. Mukherjee, A. & Lutkenhaus, J. Analysis of FtsZ Assembly by Light Scattering and Determination of the Role of Divalent Metal Cations. *Journal of Bacteriology* **181**, 823–832 (1999).
20. Doerr, A. *et al.* Modelling cell-free RNA and protein synthesis with minimal systems. *Phys. Biol.* **16**, 025001 (2019).
21. Doerr, A., Foschepoth, D., Forster, A. C. & Danelon, C. In vitro synthesis of 32 translation-factor proteins from a single template reveals impaired ribosomal processivity. *Sci. Rep.* **11**, 1898 (2021).
22. Blanken, D., van Nies, P. & Danelon, C. Quantitative imaging of gene-expressing liposomes reveals rare favorable phenotypes. *Phys. Biol.* **16**, 045002 (2019).
23. Dolgalev, G. V., Safonov, T. A., Arzumanian, V. A., Kiseleva, O. I. & Poverennaya, E. V. Estimating Total Quantitative Protein Content in Escherichia coli, Saccharomyces cerevisiae, and HeLa Cells. *International Journal of Molecular Sciences* **24**, 2081 (2023).
24. Rueda, S., Vicente, M. & Mingorance, J. Concentration and Assembly of the Division Ring Proteins FtsZ, FtsA, and ZipA during the Escherichia coli Cell Cycle. *Journal of Bacteriology* **185**, 3344–3351 (2003).
25. Chizzolini, F., Forlin, M., Cecchi, D. & Mansy, S. S. Gene position more strongly influences cell-free protein expression from operons than T7 transcriptional promoter strength. *ACS Synth. Biol.* **3**, 363–371 (2014).
26. Komura, R., Aoki, W., Motone, K., Satomura, A. & Ueda, M. High-throughput evaluation of T7 promoter variants using biased randomization and DNA barcoding. *PLOS ONE* **13**, e0196905 (2018).
27. Carr, S. B., Beal, J. & Densmore, D. M. Reducing DNA context dependence in bacterial promoters. *PLOS ONE* **12**, e0176013 (2017).
28. Yoshida, A., Kohyama, S., Fujiwara, K., Nishikawa, S. & Doi, N. Regulation of spatiotemporal patterning in artificial cells by a defined protein expression system. *Chem. Sci.* **10**, 11064–11072 (2019).
29. van den Brink, M., Althuis, T. Y., Danelon, C. & Claassens, N. J. MOSAIC: A highly efficient, one-step recombineering approach to plasmid editing and diversification. *ACS Synth. Biol.* **14**, 3880–3889 (2025).
30. Engler, C. & Marillonnet, S. Combinatorial DNA assembly using Golden Gate cloning. in *Synthetic Biology* (eds. Polizzi, K. M. & Kontoravdi, C.) 141–156 (Humana Press, Totowa, NJ, 2013). doi:10.1007/978-1-62703-625-2_12.
31. Kuruma, Y. & Ueda, T. The PURE system for the cell-free synthesis of membrane proteins. *Nat Protoc* **10**, 1328–1344 (2015).
32. Nishikawa, Y., Fujikawa, R., Nakano, H., Kanamori, T. & Ojima-Kato, T. Effect of translation-enhancing nascent SKIK peptide on the arrest peptides containing consecutive proline. *ACS Synth. Biol.* **13**, 3908–3916 (2024).
33. Smith, P. J., Wiltshire, M., Davies, S., Patterson, L. H. & Hoy, T. A novel cell permeant and far red-fluorescing DNA probe, DRAQ5, for blood cell discrimination by flow cytometry. *Journal of Immunological Methods* **229**, 131–139 (1999).
34. Abil, Z., Restrepo Sierra, A. M. & Danelon, C. Clonal amplification-enhanced gene expression in synthetic vesicles. *ACS Synth. Biol.* **12**, 1187–1203 (2023).
35. Höllerer, S., Desczyk, C., Muro, R. F. & Jeschek, M. From sequence to function and back – High-throughput sequence-function mapping in synthetic biology. *Curr. Opin. Syst. Biol.* **37**, 100499 (2024).

36. Dai, K. & Lutkenhaus, J. The proper ratio of FtsZ to FtsA is required for cell division to occur in Escherichia coli. *J. Bacteriol.* **174**, 6145–6151 (1992).
37. Schraivogel, D. *et al.* High-speed fluorescence image–enabled cell sorting. *Science* **375**, 315–320 (2022).
38. Paulsen, M. S. Image-enabled cell sorting using the BD CellView Technology. in *Flow Cytometry Protocols* (eds. Hawley, T. S. & Hawley, R. G.) 145–158 (Springer US, New York, NY, 2024). doi:10.1007/978-1-0716-3738-8_8.
39. Kuhn, T. M., Paulsen, M. & Cuylen-Haering, S. Accessible high-speed image-activated cell sorting. *Trends Cell Biol.* **34**, 657–670 (2024).
40. Quail, M. A., Corton, C., Uphill, J., Keane, J. & Gu, Y. Identifying the best PCR enzyme for library amplification in NGS. *Microbial Genomics* **10**, 001228 (2024).
41. Fu, G. K., Hu, J., Wang, P.-H. & Fodor, S. P. A. Counting individual DNA molecules by the stochastic attachment of diverse labels. *PNAS* **108**, 9026–9031 (2011).
42. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2012).
43. Schmitt, M. W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *PNAS* **109**, 14508–14513 (2012).
44. Karst, S. M. *et al.* High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nat. Methods* **18**, 165–169 (2021).
45. Zurek, P. J., Knyphausen, P., Neufeld, K., Pushpanath, A. & Hollfelder, F. UMI-linked consensus sequencing enables phylogenetic analysis of directed evolution. *Nat. Commun.* **11**, 6023 (2020).
46. Sun, J. *et al.* Correcting PCR amplification errors in unique molecular identifiers to generate accurate numbers of sequencing molecules. *Nat. Methods* **21**, 401–405 (2024).
47. Micallef, P. *et al.* Digital sequencing is improved by using structured unique molecular identifiers. *Genome Biol* **26**, 37 (2025).
48. Gantz, M. *et al.* Microdroplet screening rapidly profiles a biocatalyst to enable its AI-assisted engineering. 2024.04.08.588565 Preprint at https://doi.org/10.1101/2024.04.08.588565 (2025).
49. De Franceschi, N., Barth, R., Meindlhumer, S., Fragasso, A. & Dekker, C. Dynamin A as a one-component division machinery for synthetic cells. *Nat. Nanotechnol.* **19**, 70–76 (2024).
50. Olivi, L. *et al.* Towards a synthetic cell cycle. *Nat. Commun.* **12**, 4531 (2021).
51. Hibi, K. *et al.* Reconstituted cell-free protein synthesis using in vitro transcribed tRNAs. *Commun. Biol.* **3**, 350 (2020).
52. Miyachi, R., Shimizu, Y. & Ichihashi, N. Transfer RNA synthesis-coupled translation and DNA replication in a reconstituted transcription/translation system. *ACS Synth. Biol.* **11**, 2791–2799 (2022).
53. Hagino, K. & Ichihashi, N. In vitro transcription/translation-coupled DNA replication through partial regeneration of 20 aminoacyl-tRNA synthetases. *ACS Synth. Biol.* **12**, 1252–1263 (2023).
54. De Capitani, J. & Mutschler, H. The long road to a synthetic self-replicating central dogma. *Biochemistry* **62**, 1221–1232 (2023).
55. Cleij, C. Building the genome of a minimal synthetic cell. (TU Delft, 2025).
56. Forster, A. C. & Church, G. M. Towards synthesis of a minimal cell. *Mol. Syst. Biol.* **2**, 45 (2006).
57. Luisi, P. L., Oberholzer, T. & Lazcano, A. The notion of a DNA minimal cell: a general discourse and some guidelines for an experimental approach. *Helv. Chim. Acta* **85**, 1759–1777 (2002).
58. Koster, C. C., Postma, E. D., Knibbe, E., Cleij, C. & Daran-Lapujade, P. Synthetic genomics from a yeast perspective. *Front. Bioeng. Biotechnol.* **10**, (2022).
59. Cleij, C. *et al.* Assembly and cell-free expression of a partial genome for the synthetic cell. 2025.10.16.682769 Preprint at https://doi.org/10.1101/2025.10.16.682769 (2025).
60. Jones, S. SCRaMbLE does the yeast genome shuffle. *Nat Biotechnol* **36**, 503–503 (2018).
61. Abil, Z. *et al.* Darwinian evolution of self-replicating DNA in a synthetic protocell. *Nat. Commun.* **15**, 9091 (2024).
62. Sakatani, Y., Yomo, T. & Ichihashi, N. Self-replication of circular DNA by a self-encoded DNA polymerase through rolling-circle replication and recombination. *Sci. Rep.* **8**, 13089 (2018).
63. Okauchi, H. & Ichihashi, N. Continuous cell-free replication and evolution of artificial genomic DNA in a compartmentalized gene expression system. *ACS Synth. Biol.* **10**, 3507–3517 (2021).

64. Sakatani, Y., Mizuuchi, R. & Ichihashi, N. In vitro evolution of phi29 DNA polymerases through compartmentalized gene expression and rolling-circle replication. *Protein Eng., Des. Sel.* **32**, 481–487 (2019).
65. Okauchi, H., Sakatani, Y., Otsuka, K. & Ichihashi, N. Minimization of elements for isothermal DNA replication by an evolutionary approach. *ACS Synth. Biol.* **9**, 1771–1780 (2020).
66. Abil, Z., Giaveri, S., Erb, T. J. & Rothschild, L. J. Integrating metabolism and evolution towards the realization of synthetic life. *Nat. Rev. Bioeng.* **3**, 9–10 (2025).
67. Pigliucci, M. Is evolvability evolvable? *Nat. Rev. Genet.* **9**, 75–82 (2008).

Acknowledgements

I remember learning in school about some of the most famous scientists known to date. I pictured a scientist in a white lab coat, doing an experiment in an old laboratory or sitting behind a microscope, working quite solitarily. The scientist was experimenting for years and years, purely driven by their curiosity, and only once in a while sent a letter to other scientists to communicate some findings. I know now that we cannot do science without many people surrounding us who we can learn from, who inspire us with their ideas and support us personally. Fortunately, my image of a scientist in school was not quite correct and I have been so lucky with all the support I've had! In this section, I would like to thank everyone who contributed to this work and supported me during my PhD time, both with research as well as in any other way, because this thesis would not have existed without them!

Thank you, **Christophe**, for all the time we spent on research, for your advice and encouragement, your kindness, and your trust in me since the beginning. I really admire your calm, open, and thoughtful way of supervising. Sometimes experimenting can feel chaotic, where I sometimes even started doubting the reasons for why or how I was performing the experiments. It makes me think about the chaotic mess of printed papers on your desk ;-) It looks like chaos, but actually you know how these papers are organized on the surface and you can pull out a specific paper without effort (we only learned after shuffling them, sorry). You let me explore the chaos to reassure why and how I did things and how to improve things in the future. But what was most helpful was to know and feel that you are so involved; you are available to discuss anything at the moment it pops up, you think along with the greatest care, are very knowledgeable, and want to know every detail. Thank you for always being available to talk. Thank you for putting such a nice group of people together, shaping projects within the group towards a shared goal so we can work closely together, and highlighting positive things when unexpected things happen, which made my PhD experience one to never forget.

I am also very thankful to **Gijsje**, for your support, kindness and for welcoming me to the Biosoft group. I really appreciated the nice meetings we've had, especially your valuable advice in the last year and your feedback on my thesis chapters which was always spot on. Thank you for creating the exceptional group dynamics, where people support and care about each other, share their diverse research interests, and have fun together within and outside work.

Nico, thank you for all your advice, your enthusiasm and being so nice and approachable to brainstorm with about any scientific ideas or future career steps. Thank you for inviting me to visit to your lab in Wageningen multiple times, so I could learn what the others in your group are working on, present my work there and join the group outings. I also enjoyed organizing

the BaSyC workshop with you and Tim, seeing the results of our work together being shared with others in this fun way.

I'd also like to thank **Miao-Ping Chien**, **Liedewij Laan**, **Yannick Rondelez**, **Sander Tans**, and **Marie-Eve Aubin-Tam**, for taking the time to read my thesis and being part of the doctoral committee. A special thanks to **Miao** and her lab for demonstrating their cool microscopy techniques to us, and **Liedewij** for diving into the details of epistasis calculations with me. I am also very grateful to **Dimphna Meijer** for being part of my go/no-go committee and inviting me to present my project at the inspiring SURF symposium this year, and **Kristin Grußmayer** for being part in MEP assessment committees multiple times.

I want to thank those with whom I worked closely together on projects in this thesis over the years. Working together and sharing excitement for big and small things gave me so much energy. Thank you, **Tim**, for our enjoyable collaboration on the plasmid editing project. I appreciated that we could work on this project together and support each other throughout the process of writing our first manuscript. I am also very grateful to **Haroun**, **Marlena** and **Enrica** who dedicated their time during their studies to work on different projects now part of this thesis. Thank you for your great work and commitment to the projects and your contribution to the *gezelligheid* in the team! **Haroun**, thank you for your curiosity and boosting the MOSAIC project with your lab results. **Marlena**, thank you for setting up and shaping the automated microscopy project together, and for persisting when errors kept popping up for inconsistent reasons. I admire your creativity and have especially enjoyed when you presented your work, showcasing your project with your own personal touch. **Enrica**, you have made the last year of my PhD so enjoyable with your contagious enthusiasm for exploring new things: small things, like sending your first sequencing sample, or big things, like building a synthetic cell ;) Getting to know you better as a friend, starting with our (crazy) holiday in the snow and cold, I know now that your drive to explore new things is not limited to you at work but applies to you as a person, which I really admire about you. I am very happy for you that you will continue your synthetic cell work with Christophe in Toulouse and wish you all the best there!

I also want to thank all other colleagues and friends from the Danelon lab, who created such a welcoming and supportive atmosphere, and from whom I learned a lot through sharing our projects, struggles and findings. **Ana**, you have made me feel so welcome from the beginning. You have such a positive energy! Both in your passion for science and hard work, as in your love for sports, travelling and food. When I started, you introduced me to some of the ideas you had already kickstarted. Being able to discuss those with you, helped me so much allowing a smooth start of my PhD. I have so many great memories together, including our visit to Wageningen to learn about MAGE, the triathlons and our trip to New York with Céline, for which I am so grateful! **Fede**, together we were the last two "orphans".. ;) But without joking, you were my *rots in de branding;* I don't know how, but you seemed to know earlier then myself when I was about to ask you again for advice. Your enormous knowledge and drive for

your work and helping other people are truly admirable. Thank you for being so open in your friendship, and sharing your passions for food, gardening and, soon, hiking! I have many fun memories, just to name one of them: me chasing you in a chicken suit. Thank you for helping me out with experiments and supporting Enrica in the lab when I was in the office the past year. I hope I have been of some support for you too, and I will miss working together as office desk neighbours! **Céline**, thank you for your happy energy, both at work as outside university! You have brought many fun initiatives to the group, both science and non-science related, and I loved our trips to the US and France together. Thank you for always being open to help and that you also do that with great care, which is for instance shown by the extensive list of good tips you always gave on our practice presentations ;) I am very happy that you found a nice position at TU Delft! **Laura**, **Yannick** and **Thibault**, thank you for the great time we had at SynCell, during our visits to Toulouse and your visit to Delft! Thank you for sharing your science updates, tips and tricks, I was always curious to see how things evolved in your projects. **Yannick**, thank you for exchanging code and ideas, and for taking me along with you to demonstrate the Echo machine (let's not talk about what happened with the PURE), and to the bouldering gym during my visits to Toulouse! **Ilja**, **Elisa**, **Andreea**, thank you for your guidance, feedback and support during the first years working in the group. **Rosalie**, thank you for introducing me to experiments and techniques in the lab when I just started. **David**, thank you for introducing me to synthetic biology lab skills, your good supervision during my BEP, and keeping in touch every now and then! **Amélie**, **Kelly**, **Ellen**, **Bastiaan**, **Rebecca** and **Dustin**, thank you for sharing your enthusiasm for your projects and contributing to the nice atmosphere in the group!

I want to thank all the people from the **Danelon lab**, **Biosoft group**, the **BN department**, and the **BaSyC/EVOLF consortia** for contributing to such a nice PhD experience. Without you, my PhD experience would have been very different. I have so much enjoyed the BaSyC/EVOLF meetings and workshops, and the Biophysics and SynCell conferences, where many BN and BaSyC/EVOLF colleagues also went. There, we had the opportunity to share our research and learn a lot from each other's work, but also to get to know each other better in a different environment, and have fun together, for instance going on hikes in the Pyrenees, exploring Minneapolis, and taking a swim in the North Sea at night :) Thank you for those moments! **Jan**, **Wiel**, **Anke**, **Michal**, **Theo**, **Jeffrey** and **Sonam**, thank you for your support and advice in the lab and with the microscopes. **Nikolina**, **Marcel**, **Josh**, **Esengül**, **Justas**, **Friedrich**, and **Gabriela**, thank you for our time supervising the iGEM teams, during which I learned a lot and had fun! **Essie**, thank you for making me excited to join the iGEM supervisor team in the first place. I am grateful for the time with the students in the lab and our trip to Paris together. Thank you for walking into my office every now and then to catch up about iGEM and other things! **Jaïrus**, thank you for introducing me to TA-ing the biochemistry course, for sharing your insights on teaching with me, and for giving the cat tree a second life :) **Lars** and **Joris**, thank you for being such kind and relaxed office mates, who I could ask for advice when I needed it. **Marieke**, you are such a fun and caring person, you've always made me feel very

comfortable sharing things with you. I am looking forward to our planned hiking trip! **Irène**, **Lisa**, **Enrica**, **Charu**, **Beatriz** and **Nikki**, thank you for the fun times doing sports and creative courses at X together! **Agathe**, thank you for testing the plasmid editing method on your big construct, and of course for the delicious ossobuco you made! **Nynke**, **Inês**, **Esther**, **Bram**, **Tijn**, **Christos**, **Gianluca**, **Marcos**, **Irene**, **Hidde**, **Anouk**, **James**, **SaFyre**, **Bert**, **Christine**, **Ivy**, **Alin**, **Nicky**, **Ward**, **Stein**, **Ramon**, **Frank**, **Leander**, **Yoldaş**, **Stefan**, **Pietro**, **Angelica**, **Leanid** and all other past and present people from BN, thank you for the *gezellige* lunch, tea breaks and drinks!

Als laatste wil ik mijn familie en vrienden bedanken, waar ik leuke momenten mee heb kunnen delen en beleven, die de tijd tijdens mijn PhD zo fijn maakten, en die er voor mij waren wanneer ik ergens mee zat. **Fenna**, **Helle**, **Linda**, **Hélène**, **Pauline** en **Eline**, het betekent veel voor mij dat we elkaar na zoveel jaar vriendschap nog steeds geregeld zien en samen leuke dingen doen, en dat het nog zo vertrouwd voelt. **Papa**, ik kan met jou mijn interesse delen voor het onderzoek, het priegelen met code en grafiekjes, maar ik ben het aller-blijst met jouw dikke knuffels die zeggen dat alles goed komt. **Mama**, bedankt dat je er altijd voor mij bent en dit laat weten met lieve woorden en onverwachtse brievenbuscadeautjes waar ik altijd blij van word. **Anne**, ik ben heel blij dat we de laatste jaren nog dichter tot elkaar zijn gekomen, en ik vind het zo leuk om te zien en ook heel knap hoe jij achternagaat wat jou inspireert en energie geeft! **Stefan** en **Marieke**, het is heel fijn om zoveel enthousiasme bij jullie te zien over wat ik doe, wat ik ook doe. **Maarten**, **Nynke**, **Auke**, **Sytze**, ik voel altijd jullie lieve oprechte betrokkenheid en interesse in hoe het met mij gaat, en dat waardeer ik enorm! **Nieko** en **Veronique**, jullie zijn altijd heel lief voor mij en ik voel me heel welkom bij jullie. **Opa** en **oma**, ik hou van jullie, lieve **opa**, je blijft voor altijd bij ons. Dank jullie wel voor alle support en liefde.

Lieve **Jeroen**, ik ben zo gelukkig met jouw liefde en steun. Twee promovendi in één huis is niet altijd makkelijk; we begrijpen elkaar, maar kunnen elkaars gevoelens ook versterken. Soms nam ik jou mee in mijn gedachten en soms niet, zei ik het één en deed ik toch het ander, of veranderde ik van gedachten meteen nadat ik ze had uitgesproken. Toch voel ik jouw onvoorwaardelijke steun en daar heb ik zo veel aan gehad. Zie je die twee poppetjes aan de voet van de berg op de omslag? Ik hou van je, en heb zoveel zin in onze avonturen die gaan komen!

Curriculum vitae

Marijn van den Brink

20-01-1998	Born in Breda, the Netherlands (NL)
2010-2016	Secondary education Stedelijk Gymnasium Breda, Breda, NL
2016-2019	Bachelor Life Science and Technology Leiden University, Leiden, NL Delft University of Technology, Delft, NL
2019	Bachelor thesis • Minimal synthetic cells Bionanoscience department, Delft University of Technology, Delft, NL
2019-2022	Master Life Science and Technology Leiden University, Leiden, NL
2020	Member iGEM team Leiden 2020 • Molecular diagnostics Leiden University, Leiden, NL
2020-2021	Master thesis • Biomaterials, synthetic hydrogels Leiden Institute of Chemistry, Leiden University, Leiden, NL
2022-2026	PhD • Laboratory evolution of synthetic cells Bionanoscience department, Delft University of Technology, Delft, NL

List of publications

van den Brink, M., Claassens, N. J., Danelon, C. Combinatorial optimization of protein systems in synthetic cells. *ACS Synthetic Biology* (2026). https://doi.org/10.1021/acssynbio.6c00166

van den Brink, M., Stam, M., Claassens, N. J., Danelon, C. Image-based phenotypic sorting of synthetic cells. *Science Advances* 12, eaed0860 (2026). https://doi.org/10.1126/sciadv.aed0860

Tong, C.[1], Chen, Y.[1], Janssen, M. L.[1], Sariol, I., Wondergem, J. A. J., **van den Brink, M.**, Özel, M., Nelissen, R. G. H. H., Meulenbelt, I., Heinrich, D., Ramos, Y. F. M., Kieltyka, R. E. Supramolecular engineering of fluid pressure in filamentous hybrid double network hydrogels for 3D chondrocyte culture. *Advanced Healthcare Materials* 15, e05238 (2026). https://doi.org/10.1002/adhm.202505238

van den Brink, M.[1], Althuis, T. Y.[1], Danelon, C., Claassens, N. J. MOSAIC: A highly efficient, one-step recombineering approach to plasmid editing and diversification. *ACS Synthetic Biology*, 14, 3880-3889 (2025). https://doi.org/10.1021/acssynbio.4c00657

Tong, C., Wondergem, J. A. J., **van den Brink, M.**, Kwakernaak, M. C., Chen, Y., Hendrix, M. M. R. M., Voets, I. K., Danen, E. H. J., Le Dévédec, S., Heinrich, D., Kieltyka, R. E. Spatial and temporal modulation of cell instructive cues in a filamentous supramolecular biomaterial. *ACS Applied Materials & Interfaces* 14, 17042-17054 (2022). https://doi.org/10.1021/acsami.1c24114

van den Brink, M.[1], Tandar, S.[1], van den Akker, T., Jovikj, S., Defourt, V., Langelaar, T., Delzenne, T. O., van Strien, K., Schonk, A., Beers, A., Golov, E., Chong, L., Bergman, G., Meijdam, J., Crooijmans, M., Claessen, D., de Winde, J. Rapidemic, a versatile and label-free dnazyme-based platform for visual nucleic acid detection. *bioRxiv* (2020). https://doi.org/10.1101/2020.10.14.337808

[1] shared co-authorship

A PhD journey, navigating through
the fitness landscape of a synthetic cell.

