THE UNIVERSITY
of EDINBURGH

THE UNIVERSITY *of* EDINBURGH
School of Biological Sciences

Automation-aided High-throughput Technologies for Synthetic Biology

Paulina Julita Kanigowska

Thesis submitted for the degree of

Doctor of Philosophy

Edinburgh, 2018

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise, the work presented is entirely my own.

12th Aug 2019

Paulina Julita Kanigowska

Abstract

Synthetic biology is a research discipline which harnesses technological progress in *de novo* DNA synthesis as well as combining expertise of biological sciences and engineering research fields to facilitate construction of novel artificial biological systems. Since the past two decades, application of its methodologies has led to significant advances in metabolic engineering, providing alternative biochemical routes for the production of therapeutic products, cosmetics and biofuels. However, several challenges remain to be addressed to support development of synthetic biology applications, notably the demand for faster, cheaper and more reliable DNA manufacturing as well as efficient methods for genome-scale engineering of living organisms. This doctoral thesis proposes new interdisciplinary approaches to these problems, taking advantage of the latest laboratory automation technologies to improve efficiency of modern DNA assembly and genome editing methods. The first results chapter proposes application of a robotic platform for an acoustic liquid transfer for miniaturisation of DNA fabrication. This research, published in 2016, demonstrates the possibility to cost-efficiently assemble DNA in sub-microlitre assembly reactions. The second results chapter presents efforts to develop a method for genome-scale engineering of a model eukaryote, the budding yeast. This work capitalises on the recent progress in on-chip DNA synthesis and the next-generation sequencing (NGS) technology. Finally, the last results chapter demonstrates computational studies to predict and accelerate turnaround times of a commercial DNA supply chain using probabilistic simulations. The developed software is used to estimate sequence-specific DNA manufacturing turnaround times in order to help plan DNA manufacturing and guide decisions regarding further automation of different experimental procedures.

Lay Summary

All living organisms, from microbes to mammals, are equipped with thousands of genes which provide them with a range of capabilities needed to survive in their environment. To better understand how genes work and to re-purpose them for medicinal or industrial applications, researchers often create novel organisms, containing modified genes. Better methods for creating new DNA would therefore facilitate and accelerate research. This doctoral thesis shows how the latest progress in robotics can be used to make genetic engineering faster and cheaper. The work presented contributes to an emerging science discipline, called *synthetic biology*, which applies engineering methods to rationalise biology. First, this thesis demonstrates that it is possible to assemble DNA in volumes orders of magnitude smaller than previously, using technology originally developed for inkjet printers. Second, it presents application of a state of the art DNA modification method, called CRISPR, to delete any gene of baker's yeast, a model organism which is easy to work with. Third, it shows results of computer simulation experiments, the objective of which was to help plan DNA manufacturing at one of the leading synthetic DNA production facilities. The work presented therefore further emphasises that a combination of cross-disciplinary methodologies can be used to drive progress in genetic engineering.

Acknowledgements

This project would have not been possible without the support I received.

First, I would like to thank my supervisors, Prof. Christopher French and Prof. Patrick Yizhi Cai, for introducing me to the fields of synthetic biology and laboratory automation, for the opportunity to be a part of so many interesting projects and for the scientific advice they have given me throughout my thesis.

I would also like to say thank you to Dr. Axel Trefzer and Dr. Michael Liss for giving me a chance to learn more about the synthetic DNA manufacturing industry, during my research placement at GeneArt (Thermo Fisher Scientific, Germany), as well as all staff members who supported me throughout my project and made the three months that I have spent in Germany an unforgettable experience.

To Prof. Michael Tyers (The University of Montreal, Canada) and members of the Tyers Lab, thank you All for sharing your expertise in yeast genetics, getting me started with CRISPR and allowing me to be a part of an exciting collaboration.

To Dr. Hille Tekotte and the Edinburgh Genome Foundry team (The University of Edinburgh, UK) for providing me with access to state-of-the-art robotic equipment in one of the most impressive lab automation platforms. Thank you for your kindness and assistance in setting up various synthetic biology protocols.

I would also like to thank all people at University of Edinburgh who mentored me and helped me learn and grow during these years, in particular, Dr. Dariusz Abramczyk for all his support at the bench, Dr. Tomasz Turowski for sharing his knowledge of next-generation sequencing and Dr. Valentin Zulkower for encouraging me to learn computer programming.

Finally, I would like to thank my parents, Julita and Artur, whose support pushed me through ups and downs and who are a constant source of inspiration.

Thank you,
Paulina

List of Abbreviations

ADE – acoustic droplet ejection

ANSI – American National Standards Institute

bp – base pair

CAD – computer-aided design

CAM – computer-aided manufacturing

cDNA – complementary deoxyribonucleic acid

CFU – colony-forming unit

CRISPR – clustered regularly interspaced short palindromic repeats

DBTL – Design, Build, Test and Learn

DNA – deoxyribonucleic acid

DSB – double-stranded break

dsDNA – double-stranded deoxyribonucleic acid

Gbp – Giga base pair

HDR – homology-directed repair

kb – kilobase

LB – lysogeny broth

LCA – ligase chain assembly

LIMS – Laboratory Information Management System

lncRNA – long non-coding RNA

Mbp – Mega base pair

MC – Monte Carlo

MES – Manufacturing Execution System

mRNA – messenger ribonucleic acid

NGS – next-generation sequencing

NHEJ – non-homologous end joining

nt – nucleotide

ODU – optical density unit

ORF – open reading frame

OFAT – one factor at the time (analysis)

PAM – protospacer-adjacent motif

PCA – polymerase chain assembly

PCR – polymerase chain reaction

PP – polypropylene

qPCR – quantitative polymerase chain reaction

RNA - ribonucleic acid

SC – synthetic complete

SCRaMbLE - Synthetic Chromosome Rearrangement and Modification by LoxPsym-mediated Evolution

SGA – synthetic genetic arrays

sgRNA – single-guide ribonucleic acid

SOC – Super Optimal Catabolite repression

SOP – standard operating procedure

ssDNA – single-stranded deoxyribonucleic acid

SVC – support vector classification

SVM – support vector machine

TALE – transcription activator-like effector

tRNA – transfer ribonucleic acid

TSS – transcription start site

TU – transcription unit

YPD - Yeast extract Peptone Dextrose

ZFN – zinc finger nuclease

Table of Contents

Table of Figures

Table of Tables

Chapter 1
Introduction

1.1 Synthetic Biology – Current State

1.1.1 Synthetic Biology Supports Innovation in a Range of Research Fields

Synthetic biology is an interdisciplinary research field which harnesses engineering approaches as well as the latest *de novo* DNA fabrication technologies to address limitations of the classical genetic engineering methods (Cheng and Lu, 2012). Synthetic biologists thus do not rely on natural nucleic acid sequences, but instead choose to rationally design DNA *in silico* to then build it from "scratch". This approach to genetic engineering helps circumventing issues which tend to hamper scientific progress (e.g., DNA material recovery from inaccessible environmental sources or its labor-intensive mutagenesis) and therefore is becoming increasingly popular among researchers (Gardner, 2013).

Synthetic biology methods have been applied in a broad range of scientific projects, with several leading to significant research achievements (Cameron, Bashor and Collins, 2014). For instance, in order to support rational engineering of custom biological functions, various small synthetic biological circuits (Boolean logic gates, bistable switches, oscillators etc.) performing different logical operations have been built and led to construction of more complex functional biological devices (Elowitz and Leibler, 2000; Gardner, Cantor and Collins, 2000; Kobayashi *et al*., 2004; Stricker *et al*., 2008; Friedland *et al*., 2009; Purnick and Weiss, 2009). One notable example are whole-cell and cell-free biosensors, which demonstrate potential to address such challenges as the arsenic groundwater contamination in the regions of South and Southeast Asia or the need for efficient Ebola virus diagnostics (de Mora *et al*., 2011; Pardee *et al*., 2014). Metabolic engineering is yet another research discipline which has been applying synthetic biology methods. In the past years, this research collaboration has contributed to building different metabolically re-wired organisms able to produce diverse biochemical compounds (Smanski *et al*., 2016). For example, baker's yeast, *Saccharomyces cerevisae*, was engineered to cost-efficiently synthesise opioid drugs (thebaine and hydrocodone) and an antimalarial drug precursor (artemisinic acid), compounds which conventional manufacturing methods rely on expensive plant extraction and can be thus difficult to scale (Ro *et*

al., 2006; Galanie *et al*., 2015). Furthermore, a number of environment-friendly gasoline, diesel and jet fuel substitutes, belonging to diverse compound classes (e.g., alcohols, isoprenoids and fatty acid alkyl esters), have been thus far produced by genetically refactored microbes, with some of their biosyntheses utilizing an abundant and sustainable carbon source – plant biomass (Mendez-Perez *et al*., 2017; Peris *et al*., 2017; Ibrahim *et al*., 2018; Katre *et al*., 2018; Singh *et al*., 2018). Current biosynthesis methods are yet not limited to unicellular life nor living organisms. For instance, multicellular structures such as biofilms were also used to manufacture biofuels, while cell-free systems were applied in production of more green alternatives to such synthetic polymers as plastic (Wang and Chen, 2009; Kelwick *et al*., 2017; Tao *et al*., 2017). Synthetic biology approaches to bio-engineering are not limited to microorganism-based systems either. For example, plant and mammalian synthetic biology are two emerging research fields which hold promise for, e.g., crop improvement and development of precision medicine, respectively (Aubel and Fussenegger, 2010; Lienert *et al*., 2014; Liu and Stewart, 2015; Martella *et al*., 2016).

Synthetic biology is therefore supporting various applied research domains, with some of its applications being tangible in everyday life. One such example is Biofene®, a commercial isoprenoid compound produced by genetically engineered budding yeast, *Saccharomyces cerevisiae*, which at present is used as a "drop-in" diesel and jet fuel, powering public transportation in Brazil and jet engines of commercial airplanes, as well as a chemical additive in cosmetic products (Liebsch, 2014; Beller, Lee and Katz, 2015). The scope of potential synthetic biology applications is broadening too and involves such unconventional ideas as "growing your own clothes", using Kambucha tea microbial cultures, or manufacturing fragrances in bioreactors, even those which come from extinct flowers (Rutkin, 2015; Florea *et al*., 2016). Such scientific creativity is possible thanks to synthetic biology's efforts to simplify genetic engineering and as a consequence make it accessible to the youngest scientists, e.g., through such initiatives as the annual international Genetically Engineered Machine, iGEM, competition (Kelwick *et al*., 2015).

1.1.2 DNA "Reading" and "Writing" Technologies

In 1976, one of the smallest (3,569 nucleotides-long) known genomes was sequenced – the viral MS2 bacteriophage genome (Fiers *et al*., 1976). In 1995, the first (1.8 Mbp-long) prokaryotic genome was accomplished and belonged to a bacterial pathogen, *Haemophilus influenzae* (Fleischmann *et al*., 1995). Shortly after, in 1996, the first eukaryotic, budding yeast genome (12.1 Mbp-long), was *read*, and in 2001 the first draft of the human genetic code (3.2 Gbp-long) was published (Goffeau *et al*., 1996; Lander *et al*., 2001). More recently, scientists have been embarking on more challenging projects, e.g., sequencing ~20 Gbp-long gymnosperm plant genomes or genetic material of unculturable microorganisms (i.e., metagenomics studies) (Zimin *et al*., 2014).

Now, synthetic biologists are *writing* DNA to better understand the link between the genotype and phenotype as well as to discover the minimal set of genes required for life. Current state of the DNA manufacturing technology is one of the key factors which make this scientific approach feasible, with the improving *de novo* DNA fabrication methods promising an unconstrained DNA sequence customisation in the future (Clore, 2018; L. Wang *et al*., 2018; Palluk *et al*., 2018; Veneziano *et al*., 2018). Back in 1970s, the first gene was made *de novo* (Agarwal *et al*., 1976). It was a 77 nucleotides-long yeast alanine tRNA gene DNA sequence enzymatically assembled from 17 DNA oligonucleotides, chemical synthesis of which took 5 years. Today, such oligonucleotides are available for a next-day delivery, while larger DNA fragments are likewise commercially available, affordable and ready within reasonable timescales (Kosuri and Church, 2014). Owing to this progress, synthetic biology projects are thus now entering genomic scales.

Just like the DNA sequencing landmarks, the first genomes to be made *de novo* were the most compact viral and bacterial ones. In 2002, Cello *et al*. synthesised the first viral genome DNA - the poliovirus cDNA, which totaled ~7.5 kbp. In 2008, this work served as a stepping stone to construction of an attenuated version of this virus (a vaccine candidate) by re-coding its genetic material with underrepresented codon

variants (Cello, Paul and Wimmer, 2002; Coleman *et al*., 2008). The same year however yet another genome synthesis milestone was accomplished. The J. Craig Venter Institute announced re-fabrication of the first bacterial (*Mycoplasma genitalium*) genome, 583 kb-long (Gibson *et al*., 2008). Two years later, continuation of this work led to a successful re-synthesis of a nearly two times larger 1.08 Mbp *Mycoplasma mycoides* genome and its transplantation into a *Mycoplasma capricolum* recipient cell, which was "rebooted" with this foreign synthetic chromosome (Gibson *et al*., 2010). More recently, in 2016, the same research team has also managed to halve the re-synthesised *M. mycoides* genome (to 531 kbp), using an iterative "design, build, test and learn" synthetic biology methodology, giving rise to the smallest autonomously replicating cell (Hutchison *et al*., 2016).

Eukaryotic chromosomes are a new research focus. In 2011, the International Synthetic Yeast Genome ("Sc2.0") Project was launched (which our research team is a part of) and set itself a goal of custom *S. cerevisiae* yeast genome re-synthesis, incorporating a set of "designer" features into its native chromosomes (Dymond *et al*., 2011). For instance, a synthetic evolution system was added in the form of LoxPsym DNA recombinase recognition sites, inserted downstream of the non-essential gene reading frames by design, which upon heterologous expression of the recombinase enzyme would serve as genomic rearrangement (i.e., deletion, duplication, inversion and translocation) foci, facilitating accelerated adaptation to a given environmental condition. A number of other custom chromosome modifications were also announced and included (1) deletion and relocation of DNA sequences, which are known to elicit genomic instability, i.e., of the Ty transposable elements and tRNA genes; (2) deletion of introns; (3) replacement of the native telomere sequences with much shorter synthetic alternatives; (4) replacement of all TAG stop codons with their TAA counterparts, to allow future expansion of the yeast genetic code with unnatural amino acids; and (5) incorporation of synthetic DNA "watermark" sequences, to discriminate between the synthetic and wild-type DNA regions.

To date, six synthetic yeast chromosomes have been published (synII, III, V, VI, X and XII) as a result of this international collaboration, totalling approximately 3.5 Mbp of synthetic DNA, which constitutes more than one quarter of the native 12.1 Mbp-long yeast genome (Annaluru *et al*., 2014; Mitchell *et al*., 2017; Shen *et al*., 2017; Wu *et al*., 2017; Xie *et al*., 2017; Zhang *et al*., 2017). The rate at which consecutive yeast chromosomes are accomplished can be partially attributed to the DNA construction strategy adopted by the Sc2.0 Project consortium based on allocation of individual chromosome syntheses to different Sc2.0 Project research teams distributed worldwide. As a result however, each synthetic chromosome is built in a separate yeast strain. Therefore, the upcoming challenge is to combine all of them in a single yeast cell. To date, two viable yeast strains containing more than one synthetic chromosome have been reported (Mitchell *et al*., 2017). Future work on combining increasing amounts of synthetic DNA can be viewed both as a scientific challenge and an opportunity to discover unknown genetic interactions, which can give more insight into eukaryotic biology. Larger genome synthesis projects could benefit from this expertise. Recently, the international Human Genome Project-write consortium announced re-fabrication of human chromosomes. Synthetic human chromosomes are currently anticipated to assist in building novel cell lines of therapeutic importance and are a much larger scale undertaking, with chromosomal DNA sizes ranging from 48 to 249 Mbp (Boeke *et al*., 2016).

Due to the fast-increasing pace of DNA *reading* and *writing*, DNA sequencing and fabrication technologies have been both benchmarked against Moore's law, describing an exponential technological growth tendency previously observed in the computer industry. Both technologies exhibited such growth. Yet about a decade ago the DNA sequencing technology has readily surpassed it (Moore, 2006; Stein, 2010). The somewhat parallel technological progress in the *reading* and *writing* research domains facilitates efforts to engineer novel biological systems, which take advantage of the expanding knowledge of natural DNA sequence diversity.

1.1.3 Synthetic DNA as a Functional Biopolymer

Application of nucleic acid manufacturing technologies is not limited to developing custom biological systems. As DNA is becoming a commodity product it is used by some researchers simply as a cheap biopolymer. For instance, scientists are now considering using DNA to store data, owing to its dense nanoscale four-bit (i.e., A, T, G, C) information encoding system (e.g., 1 cubic millimeter of DNA can encode a quintillion, or 10^{18}, bytes of data) as well as its longevity (i.e., a half-life of 521 years at 13.1°C) (Allentoft *et al.*, 2012; Church *et al.*, 2012). As the amount of data generated daily continues to grow, DNA molecules can help to address the data storage demand (Church *et al.*, 2012). To date, several different data objects have been stored in strings of DNA, including poems, entire books, images and GIFs (Church, Gao and Kosuri, 2012; Goldman *et al.*, 2013; Shipman *et al.*, 2017). Larger DNA data storage projects have been announced as well. For instance, in 2016 Microsoft Corporation announced its ongoing research in this area and approximately 200 megabytes of data stored in its DNA data storage system (Extance, 2016).

Aside from its capability to encode data, DNA can also spontaneously form various 3D structures due to its biophysical properties. *In vivo* these properties allow the DNA strands to control access to genetic information. *Ex vivo* however, they can be harnessed to construct customised molecular structures, i.e., so-called "DNA origami", which can perform various tasks. Design and assembly of DNA origami nanostructures can be considered as good examples of rational DNA engineering. First, dedicated pieces of computer-aided design software (e.g., the CADnano 2.0 software) are used to determine the necessary set of DNA fragments required to build a certain nanostructure. Two types of single-stranded DNA are then used in the nanostructure assembly process, i.e., a longer, "scaffold", DNA molecule and numerous shorter, "staple", DNA oligonucleotides, which guide the nucleic acid scaffold towards the desired 3D shape during a thermal annealing step (Wang *et al.*, 2017). Many DNA origami constructs have been published, with drug delivery being one of their potential applications (Douglas *et al.*, 2009; Rajendran *et al.*, 2011;

Wickham *et al*., 2012). For example, in 2012 Jiang *et al*. demonstrated successful delivery of an anti-cancer drug, doxorubicin, into drug-sensitive and drug-resistant cancer cells and its significant cytotoxicity in both cell lines (Jiang *et al*., 2012). Following this work, further developments employing virus-like coating strategies have now also been applied and can help alleviating immune activation (Mikkilä *et al*., 2014).

1.2 The Synthetic Biology Design, Build, Test and Learn Cycle

Research projects and areas outlined above demonstrate efforts to forward engineer biological systems, which is one of the main ambitions of synthetic biology. Despite the scientific progress made so far, this goal however still proves vastly challenging, mainly due to our limited understanding of the way biology works (Kwok, 2010). Therefore, in order to tackle the complexities of forward engineering synthetic biologists try to systematise engineering workflows to avoid any ad hoc experimental work. These are thus typically organised into so-called "design, build, test and learn" (DBTL) optimisation cycles which aim at rationally *designing* and *building* a set of DNA constructs, which will provide sufficient experimental results to efficiently *test* a given experimental hypothesis and *learn* from the output data. Consecutive synthetic biology cycles capitalise on the knowledge gained throughout the preceding iterations and hence continuous progress towards the initially specified objective is maximised through this methodology (Carbonell *et al*., 2018).

1.2.1 Design

Each optimisation cycle starts with accurately specifying the intended nucleic acid/organism structure or function, e.g., by mathematically describing the desired biological system's behavior. Once a formal specification is ready, suitable DNA constructs and edits have to be designed to satisfy the researcher's requirements. Two kinds of DNA design methodologies are employed by synthetic biologists: rational and combinatorial. Rational methodologies are those which decide on a certain DNA design based on its predicted performance, as judged by an appropriate mathematical model. Combinatorial methods on the other hand seek to find a set of

design variants which will likely encompass the optimal one. In order to find the optimal design, these variants have to be screened or undergo selection under a suitable environmental pressure. Such search tactics are often labor-intensive and combinatorial approaches are typically used when there is a strong scientific belief that valuable DNA designs might not be considered by the rational approach due to a limited understanding of a particular biological system. Some common applications of the combinatorial DNA design methodology are identification of optimal metabolic pathway variants with, e.g., promoter, terminator, ribosome-binding site and enzyme libraries; or construction of genome-wide editing libraries to find the most favorable genotypes under a certain, e.g., chemical, condition (Smanski *et al*., 2014; Kim, Moore and Yoon, 2015; Wong *et al*., 2015; Wong, Choi and Lu, 2016).

Repositories of accurately characterised and standardised parts as well as various computer-aided design (CAD) software tools are two key resources necessary throughout the design process (Endy, 2005; Xia *et al*., 2011; Ham *et al*., 2012; Madsen *et al*., 2016; Appleton *et al*., 2017). Synthetic biologists thus aim at developing their own suite of CAD tools as well as implementing unified modular biological part standards (e.g., the BioBricks and PhytoBricks standards) and policies for *in silico* design representation (e.g., SBOL – Synthetic Biology Open Language), part characterisation and laboratory protocol documentation (i.e., SOPs – Standard Operating Procedures) (Knight, 2003; Shetty, Endy and Knight, 2008; Shetty *et al*., 2011; Galdzicki *et al*., 2014; Patron *et al*., 2015). These efforts seek to simplify the design process and facilitate knowledge transfer in the synthetic biology community through centralised data repositories (e.g., Registry of Standard Biological Parts, SBOL Stack, ICE – Inventory of Composable Elements, or Clotho) (Endy, 2005; Xia *et al*., 2011; Ham *et al*., 2012; Madsen *et al*., 2016).

The number of synthetic biology CAD tools is increasing and comprises computational solutions to different design challenges (Appleton *et al*., 2017). For instance, various CAD tools are available to rationally design DNA parts with the desired output. For example, the PromoterCAD software facilitates design of synthetic promoter sequences with tailored properties, whereas the online RBS

Calculator tool helps researchers build ribosome-binding sites with customised strengths (Salis, Mirsky and Voigt, 2009; Nishikata *et al*., 2014). Other pieces of software are dedicated to full DNA construct design and model-driven simulation of costruct's behavior, with examples including the GenoCAD, BioJADE and iBioSim CAD tools, which are all equipped with access to public or their own DNA parts repositories (Goler, 2004; Czar, Cai and Peccoud, 2009; Watanabe *et al*., 2018). Synthetic biologists can hence use these design frameworks to conveniently evaluate performance of their designs and to then rationally choose the optimal DNA constructs. Even more convenient "hands-off" CAD solutions are now emerging and are there to completely automate DNA design. For example, Cello is a CAD tool which generates optimal DNA designs for synthetic biological circuits based on a formal specification of the circuit function. This tool applies a hardware description language, Verilog, which is used to describe the desired function as well as to specify various circuit components (i.e., sensors and actuators) and constraints, including the host organism context (Nielsen *et al*., 2016). Such rational and automated DNA design frameworks are no yet available for many other synthetic biology projects and so combinatorial approaches have to be employed instead. Standardised, modular and "one-pot" DNA assembly techniques allow easy construction of DNA sequence variants by swapping various DNA parts in their designated DNA assembly slots. However, swapping DNA parts at multiple such positions leads to a rapid expansion of the combinatorial space, which can prove challenging to screen (Wong, Choi and Lu, 2016). Synthetic biologists therefore take advantage of mathematical methods such as the design of experiments (DOE) method to rationally manage the scale of combinatorial experiments, so that these can efficiently elucidate the relationship between the DNA parts and a given function as well as between the DNA parts themselves. Such analyses are complex to tackle manually and therefore a number of CAD tools are available to perform these more easily (e.g., the Double Dutch and RedLibs CAD software) (Jeschek, Gerngross and Panke, 2016; Roehner *et al*., 2016).

Determining a suitable DNA sequence, or a set of them, to elicit the desired function marks a near completion of the *Design* phase. DNA sequences now require further attention to plan the synthesis and assembly process. This final *Design* stage lies

right at the border with the *Build* phase and aims at establishing the most economical and straightforward DNA fabrication strategy. Therefore, it concerns itself with maximizing DNA part reuse as well as "polishing" nucleic acid sequences, so that they conform to certain DNA assembly standards and are easy to manufacture (i.e., have reasonable GC contents and lack stable secondary structures, as well as repeat and homopolymer stretches) (Villalobos *et al.*, 2006; Appleton *et al.*, 2014; Blakes *et al.*, 2014; Oberortner *et al.*, 2016). Such multi-objective optimisation is a computationally complex task and is difficult to accomplish using manual approaches. Various CAD solutions are hence being developed to automatically solve user-defined constraints, e.g., the DNALD and Raven DNA fabrication planning tools (Appleton *et al.*, 2014; Blakes *et al.*, 2014). Processed DNA sequences can be viewed and further edited (if necessary) with different DNA sequence editors (e.g., SnapGene, VectorEditor and Benchling), which offer user-friendly interfaces and an increasing array of DNA view and edit features (e.g., visualisation of DNA restriction sites, cloning simulation and DNA editing sites design). Their use is therefore becoming increasingly widespread among laboratory teams, which commit to one or a number of such tools and use them in general DNA file management and sharing (Appleton *et al.*, 2017).

1.2.2 Build

Complete DNA designs, with their nucleic acid sequences optimised (to facilitate the construction process) and the manufacturing strategies outlined can now be *built*. DNA fabrication is a hierarchical process featuring various chemical and enzymatic DNA construction methods, which are each suitable for different DNA fabrication scales and strategies (**Figure 1.1 A**). The process of building nucleic acid sequences thus starts with chemically synthesizing short (up to ~200 nucleotides-long) oligonucleotides, which are subsequently enzymatically assembled into longer (~1,000 nucleotides-long) DNA fragments, during a process termed "gene synthesis". The longer DNA fragments can be next combined using various other *in vitro* and *in vivo* enzymatic methods, depending on the application, to form even longer DNA pieces (up to chromosome sizes) (Ellis, Adie and Baldwin, 2011; L. Wang *et al.*, 2018).

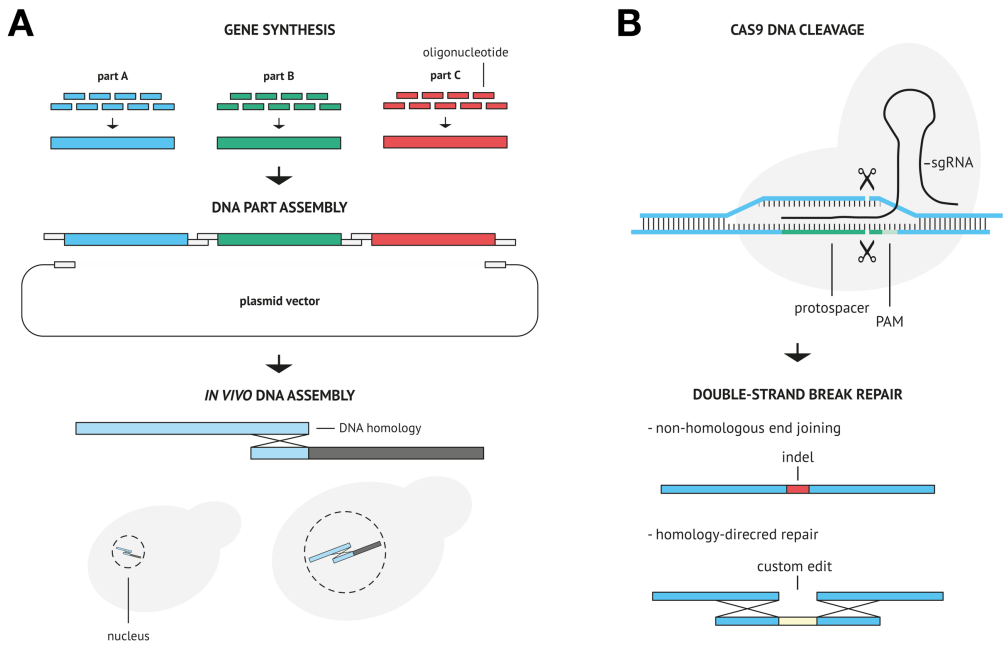


Figure 1.1 DNA Assembly and Editing

(**A**) DNA assembly; Different DNA fabrication methods are used to assemble DNA, depending on its length. First, multiple short ssDNA oligonucleotides (~60 bp) are annealed and fused to form larger dsDNA fragments, up to ~1 kb, in a process referred to as gene synthesis. Larger DNA fragments are next combined with sequence homology-based (e.g., Gibson assembly) or restriction-ligation-based methods (e.g., Golden Gate DNA assembly). Assembly of DNA pieces > 10 kb is often done *in vivo*, e.g., using the budding yeast homologous recombination machinery. (**B**) DNA editing; DNA can also be edited to obtain the desired nucleic acid sequence. *S. pyogenes* CRISPR-Cas9 machinery is currently one of the most popular DNA editing tools. Cas9 endonuclease forms a ribonucleoprotein complex with the single-guide RNA moiety (a crRNA and tracrRNA hybrid), which recognises 20 bp protospacer DNA sequences, followed by an NGG protospacer adjacent motif (PAM). Two Cas9 endonuclease domains, RuvC and HNH, then cleave both strands of the target DNA locus, resulting in a blunt-end cut 3 bp upstream of a given PAM sequence. Double-stranded DNA breaks (DSBs) are lethal for cells and have to be repaired. Two main mechanisms of DSB repair are used by cells to re-ligate broken DNA strands, i.e., non-homologous end joining (NHEJ) and homology-directed repair (HDR). NHEJ is considered error-prone, as it does not use any reference DNA sequence to repair DNA, and thus often leads to insertion-deletion DNA mutations (indels). HDR uses a homologous nucleic acid template to guide repair of DSBs. In haploid and polyploid cells, sister chromatids and additional chromosomal DNA sets, respectively, are used to precisely (without any errors) repair DNA. However, if cells are provided with a sufficient amount of external homologous DNA, custom DNA alterations can be introduced.

1.2.2.1 Oligonucleotide DNA Synthesis

The first synthetic oligonucleotides were made in the 1950s (Michelson and Todd, 1955). In 1980s, oligonucleotide synthesis started utilizing solid-phase phosphoramidite chemistry methods developed by Martin Caruthers and his colleagues and became commercialised (Caruthers *et al*., 1987). Since then, these methods have prevailed as the standard oligonucleotide synthesis methods and have become fully automated, with batches of 96 or 384 oligonucleotides being now processed simultaneously and synthesis scales of ~10-100 nmol (Kosuri and Church, 2014). Standard oligonucleotide synthesis proceeds through cycles of: (1) deprotection, (2) coupling, (3) capping and (4) oxidation steps, with each cycle adding one base to the growing oligonucleotide strand (**Fig. 1.2 A**). First, a protecting dimethoxytrityl (DMT) group is removed from the 5' end of the last nucleoside phosphoramidite, using trichloroacetic acid. Next, another DMT-protected nucleoside phosphoramidite is added and reacts with the unprotected 5' hydroxide group, through activation with a tetrazole or an imidazole activator. Unreacted 5' ends can be then capped, using an acetate group, and thus excluded from the subsequent synthesis cycles to prevent any deletion mutations. The last step of the cycle is an iodine oxidatation of phosphite groups, which results in a cyanoethyl-protected phosphate backbone. At the last synthesis cycle, this chemical reaction is followed by a final deprotection procedure, including deprotection of the 5' end and the backbone using e.g., gaseous ammonia. Deprotected oligonucleotides are next eluted off the solid support (typically, a controlled pore glass (CPG) column) by treating them with a denaturing base (Roy and Caruthers, 2013; Kosuri and Church, 2014).

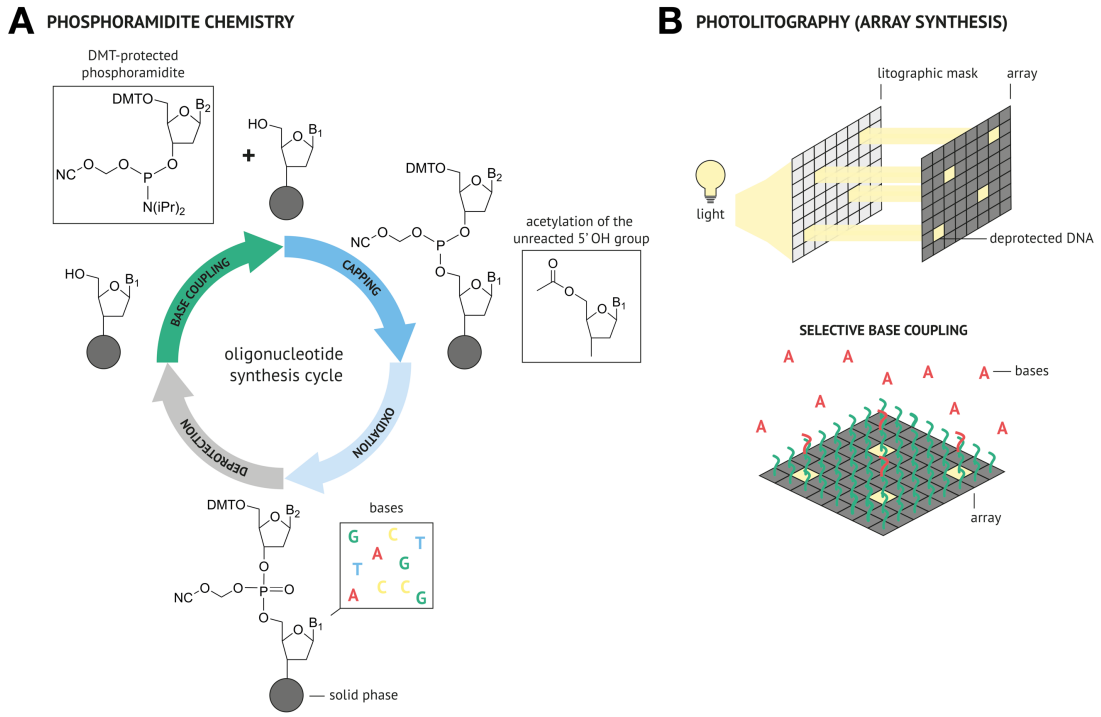


Figure 1.2 Oligonucleotide DNA Synthesis

(**A**) Oligonucleotide synthesis using solid-phase phorsphoramidite chemistry: A growing oligonucleotide DNA strand is attached to a solid support (typically, a controlled pore glass column). DNA synthesis proceeds through cycles of deprotection, base coupling, capping and oxidation. First, a dimethoxytrityl (DMT) group is removed from the 5' end of the last nucleoside phosphoramidite, using an acid-catalysed reaction. A second DMT-protected nucleoside phosphoramidite is then coupled with the deprotected phosphoramidite moiety, using, e.g., a tetrazole activator. Some 5' OH groups are left unreacted and have to be capped in order to prevent deletion mutations. The hydroxyl groups are thus capped with an acetate group. The backbone phosphite groups are next oxidised through iodine oxidation and resulting a cyanoethyl-protected phosphate nucleic acid backbone. Once the desired DNA sequence is synthesised, via multiple such synthesis cycles, all chemical protection groups are removed using, e.g., gaseous ammonia and the oligonucleotide DNA is eluted off the solid support. (**B**) Arrayed oligonucleotide synthesis with photolithography: Photolithography is one of the on-chip oligonucleotide synthesis methodologies, using light and photomasking for selective deprotection and coupling of photo-labile phosphoramidites.

Depending on the target oligonucleotide length, an appropriate number of synthesis cycles is necessary. However, this number is limited, with DNA syntheses reaching oligonucleotide lengths of up to ~200 bases (Clore, 2018). There are a couple of reasons behind this limit. First, the correct sequence yield has to be high at each synthesis cycle. Yet a seemingly high 99% yield only leads to a final yield of 55% for a 60 nucleotides-long oligo and decreases exponentially with every base added. Second, depurination can occur during the acid-catalysed deprotection step and pose problems for longer oligonucleotides, which tend to break at the resulting abasic sites during the final deprotection procedure (Kosuri and Church, 2014). Novel chemistries as well as enzymatic DNA synthesis technologies (e.g., the commercial DNA Script technology) are therefore emerging to address these issues (Palluk *et al.*, 2018; Veneziano *et al.*, 2018).

Solid-phase phosphoramidite chemistry does not solely suffer from its chemical limitations. Its other drawback is high running costs (Eroshenko *et al.*, 2012). To address this issue, alternative technologies have been therefore arising, allowing parallel synthesis of thousands to millions of oligonucleotides on a single DNA chip, and contributed to at least a two order of magnitude drop in the synthesis costs (Eroshenko *et al.*, 2012). Affymetrix was the first company to develop an on-chip synthesis method, in 1990s, and used a light-activated chemistry to selectively deprotect photolabile nucleoside phosphoramidites (**Fig. 1.2 B**) (Fodor *et al.*, 1993; Pease *et al.*, 1994). Today, a number of other commercial on-chip synthesis methods exist (Eroshenko *et al.*, 2012). For example, Agilent developed an ink jet DNA chip printing technology, while CustomArray (recently acquired by GenScript) developed a semiconductor-based electrochemical method to selectively produce acid and thus selectively deprotect nucleoside phosphoramidites (Eric LeProust *et al.*, 2000; Gao *et al.*, 2001; Cleary *et al.*, 2004; Egeland and Southern, 2005). Nonetheless, these more economical DNA synthesis approaches did not readily reduce costs and increased lengths of the downstream oligonucleotide assemblies (Eroshenko *et al.*, 2012). There are a number of technological limitations which hinder further progress. First, the substantial throughput of on-chip synthesis can be problematic. Namely, it makes it difficult to select and co-localise oligonucleotides which will be further assembled

into larger DNA constructs. Second, typically, on-chip-synthesised oligonucleotides exhibit higher error rates as compared to the tradition column-based methods. Finally, lower oligonucleotide yields are achieved using the on-chip synthesis and fall below the nmol range (Eroshenko *et al.*, 2012). Several solutions to these issues have nevertheless been demonstrated. For instance, in 2010 Kosuri *et al.* designed primer-binding site barcodes, uniquely identifying sets of oligonucleotides needed for a single assembly reaction, and used PCR amplification to enrich for a specific oligonucleotide set. Barcode sequences were subsequently enzymatically cleaved off and the oligonucleotide DNA was assembled into larger DNA fragments (Kosuri *et al.*, 2010). Quan *et al.*, on the other hand, developed a technology to synthesise such oligonucleotide sets in physically separated micro-wells and thus perform one assembly reaction per such micro-compartment (Quan *et al.*, 2011). This technology later paved the way for, e.g., such commercial DNA products as the ones offered by Gen9 (part of Gingko Bioworks) (Goldberg, 2013). This year, Plesa *et al.* reported yet another method, DropSynth, to concentrate oligonucleotide assembly sets. DropSynth uses barcoded beads, which harvest oligonucleotide assembly sets. Oligonucleotide DNA is later processed and assembled in picolitre emulsion droplets (Plesa *et al.*, 2018). Lastly, methods to alleviate oligonucleotide error rates were also reported and mostly relied on proteins capable of detecting DNA mis-hybridisations upon DNA strands denaturation and re-annealing. For example, MutS binds to DNA error-containing heteroduplexes, which can be later detected and filtered out by an electrophoretic mobility shift assay (EMSA) or a MutS-binding column, while other proteins can additionally cleave these off (e.g., the T7 endonuclease I) and so they can also be used to repair DNA errors (Carr *et al.*, 2004; Matzas *et al.*, 2010; Sequeira *et al.*, 2016).

1.2.2.2 Oligonucleotide DNA Assembly

DNA oligonucleotides exhibiting a sufficient DNA sequence quality can be assembled into larger nucleic acid fragments using one of the two available enzymatic methods – a DNA ligation-based method (LCA – ligase cycling assembly) or a PCR-based method (PCA – polymerase cycling assembly). The DNA ligation method requires overlapping oligonucleotides, which fully cover both strands of the

target DNA sequence, and utilises a thermostable DNA ligase enzyme to ligate the oligonucleotide ends (Chandran, 2017). PCA, on the other hand, does not require full coverage of the target DNA sequence since gaps between the overlapping oligonucleotides are filled by a high-fidelity DNA polymerase (TerMaat *et al.*, 2009). Both LCA and PCA involve analogous thermal cycling protocols of: DNA denaturation, annealing and ligation (LCA)/extension (PCA), to progressively enrich for full-length DNA fragments, and a final DNA amplification step using a standard PCR reaction. Despite the protocol similarities, both methods however have their inherent limitations. For instance, LCA requires an extra 5' end oligonucleotide phosphorylation reaction prior to the thermal cycling protocol, which increases overall assembly costs. Moreover, its full oligonucleotide coverage requirement further increases the expenses, as compared to the PCA method which requires less oligonucleotide DNA. Nonetheless, PCA is more prone to nucleic acid mis-hybridisation due to its shorter DNA strand overlap regions, and therefore it might not be suitable for some DNA sequences, e.g., those with repeats.

1.2.2.3 Higher Level DNA Assemblies

Typical gene synthesis reaches DNA lengths of up to 1 kb and suitable enzymatic methods are used to assemble larger DNA pieces. Notably, these methods do not require PCR amplification of the full-length DNA constructs, which is increasingly prone to errors as the DNA constructs grow larger, and do not directly rely on the generally error-prone oligonucleotides. Following gene synthesis, or any other DNA assembly stage, DNA products are typically further examined using restriction digestion, colony PCR and Sanger sequencing to assess correctness of their DNA sequences. Once a given DNA fragment passes such a quality control check, it can be then used as a substrate in another DNA assembly round. Sequence-verified gene synthesis products can be therefore joined to obtain larger DNA pieces. A variety of enzymatic methods exist to perform these higher-level DNA assemblies and these can be generally split into two major classes – the DNA overlap-directed class and the restriction digestion-based class (**Fig. 1.3**).

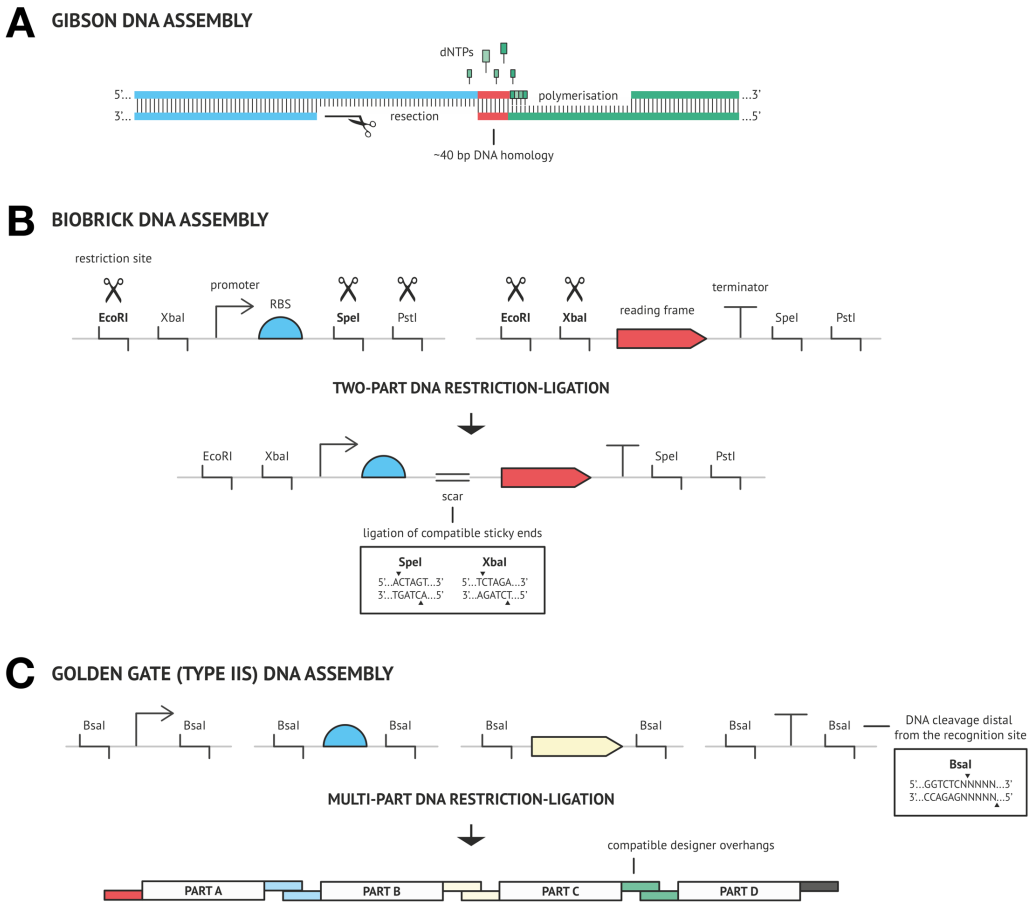


Figure 1.3 Sequence Homology- and Restriction-Digestion-based Assembly

(**A**) Gibson assemblies are a type of DNA homology-based assemblies, which use a combined action of three enzymes to join DNA fragments sharing terminal DNA sequence homologies (~40 bp). First, the T5 exonuclease 5' end resection exposes homologous nucleic acid strands. Complementary DNA strands then anneal (thanks to a favorable 50°C reaction temperature). A DNA polymerase re-synthesises DNA degraded by the T5 exonuclease in a 3' to 5' direction. DNA ligase is next used to ligate the nicked dsDNA. (**B**) The BioBrick method is based on DNA restriction digestion and ligation mechanism and allows assembling two DNA parts at the time. The DNA parts feature standard prefix and suffix sequences, which encode EcoRI and XbaI, and SpeI and PstI restriction sites, respectively. Through appropriate restriction digestion of the two DNA fragments (one with the EcoRI and SpeI pair of endonucleases and the other one with the EcoRI and XbaI restriction enzyme combination), SpeI and XbaI sticky ends anneal, are ligated by a DNA ligase and form a 6 bp scar sequence, which cannot be re-cut. (**C**) The Golden Gate method is based on the restriction-ligation mechanism as well. However, it utilises a distinct class of Type IIS restriction enzymes, which cleave DNA outside of their recognition sites. BsaI is an example Type IIS endonuclease, which recognition and restriction sites are illustrated by the figure. BsaI digestion leads to formation of staggered DNA ends, which sequences can be defined by the researcher. Design of uniquely compatible DNA overhangs allows multi-part DNA assembly.

The first class relies on ~40 bp, or longer, overlaps between the DNA fragments, which become exposed upon a 5 to 3' nucleic acid resection and undergo thermal annealing to yield larger DNA products. Such overlaps are typically already encoded by design by the DNA fragments to be assembled. However, if they are not, they can also be easily added by PCR amplification with DNA primers encoding these. Methods belonging to the DNA overlap-directed class can be further split into *in vitro* and *in vivo* methods. *In vitro* methods rely on isothermal (e.g., Gibson Assembly, SLIC – sequence- and ligase-independent cloning and SLiCE – seamless ligation cloning extract) or thermal cycling protocols (e.g., CPEC – circular polymerase extension cloning) to carry out the necessary enzymatic reactions. Here, the enzymatic DNA fabrication can involve a single enzyme – commonly, a DNA polymerase (e.g., the CPEC and SLIC DNA assemblies); a mixture of enzymes, e.g., Gibson assembly, which utilises T5 DNA exonuclease, Phusion DNA polymerase and Taq DNA ligase (**Fig. 1.3 A**); or a crude cellular enzymatic extract, which circumvents the need of purchasing various (at times, costly) enzymatic components (e.g., the SLiCE method) (Gibson *et al.*, 2009; Quan and Tian, 2009; Li and Elledge, 2012; Zhang, Werling and Edelmann, 2014). *In vivo* methods, on the other hand, rely on DNA transfection protocols, which deliver the DNA assembly fragments inside a eukaryotic cell (**Fig. 1.1 A**). These methods assemble DNA *in vivo*, harnessing the host DNA homologous recombination machinery, and typically use budding yeast or bacterial, *Bacillus subtilis*, cells thanks to their proficiency in DNA recombination as well as abilities to handle DNA assemblies of large DNA pieces (e.g., the TAR, transformation-associated recombination, and "domino" methods) (Itaya *et al.*, 2008; Kouprina and Larionov, 2016).

Restriction digestion-based methods constitute the second major DNA assembly class. These methods rely on DNA restriction enzymes, which generate standard sticky end DNA sequences allowing compatible DNA parts to be ligated with each other. Such standardised DNA assembly approaches are therefore suitable for establishing modular and hierarchical DNA assembly systems, which are particularly useful in building combinatorial DNA libraries (Wong, Choi and Lu, 2016). For instance, a scientist could be interested in swapping and analysing the function of

various components of a biochemical pathway (level 2 and 3 assemblies), which consists of several transcription units (level 1 assembly), which in turn consist of, e.g., DNA promoter, open reading frame (ORF) and terminator parts (level 0 assembly). The first restriction digestion-based DNA assembly method to be established was the BioBrick DNA assembly method, utilizing 4 restriction enzymes (i.e., EcoRI, XbaI, SpeI and PstI) which have their restriction sites encoded in standardised "prefix" and "suffix" sequences, flanking the DNA assembly parts (i.e., "BioBricks"). Two BioBricks can be combined through digestion with two separate sets of enzymes (I – EcoRI and SpeI, II – XbaI and PstI) and a subsequent DNA ligation. The resulting DNA assembly product is a new larger BioBrick. This product however contains a "scar", which can prove problematic at some DNA part junctions. Another limitation of this method is its throughput. Namely, a labour-intensive sequential ligation of DNA parts (two DNA parts per one DNA assembly) is required (**Fig. 1.3 B**) (Knight, 2003; Shetty, Endy and Knight, 2008; Shetty *et al*., 2011). More recent restriction digestion-based methods address these issues by utilizing Type IIS DNA restriction enzymes (e.g., BsaI, BsmBI, BbsI, AarI or SapI), which make staggered DNA cuts outside of their recognition sites, therefore allowing customizing sticky end sequences, which can be designed to minimise or completely eliminate DNA scars (e.g., a DNA scar can encode a stop codon). User-defined DNA overhang sequences also open the possibility of assembling all DNA parts in a single, "one-pot", reaction, since they allow the user to set the order of the DNA fragments. To date, the most popular Type IIS method is the Golden Gate method, which has been harnessed by various DNA fabrication frameworks (e.g., the MoClo and Golden Braid systems) featuring their unique overhang sequences, DNA part slots as well as hierarchical DNA construction techniques (e.g., utilizing different Type IIS DNA endonucleases at different levels of DNA assembly) (**Fig. 1.3 C and Table 1.1**) (Engler, Kandzia and Marillonnet, 2008). Nevertheless, despite the increasing adoption of the Golden Gate method, this approach still does not address one remaining issue, namely the occasional presence of a relevant Type IIS site inside a DNA part, which can lead to decreased efficiencies of DNA assemblies or spurious assembly products. Parts to be assembled with the Golden Gate method thus similarly to BioBricks require domestication, i.e., removal of the forbidden

restriction sites. More recent frameworks (e.g., Mobious Assembly) attempt to overcome this issue by using rare Type IIS cutters such as the AarI DNA endonuclease (Andreou and Nakayama, 2018).

Table 1.1 Standard Frameworks Using Golden Gate Assembly

framework name	domestication requirements	assembly levels	number of plasmid vectors per each assembly level (L)	cloning capacity (number of TUs)	organism preference	additional comments	references
MoClo	BsaI and BpiI site removal	0, 1, 2, M and P	L0 – 5, L1 – 7 vectors and 21 linkers, L2 – 7 vectors, L M/P – 7 vectors and 7 linkers each	up to 6 TUs in L2 and M; alternating between M and P levels allows custom TU capacity expansion	eukaryotic cells	-	(Weber *et al*., 2011; Werner *et al*., 2012)
MoClo-YTK (MoClo derivative)	BsaI and BsmBI site removal	0, 1, 2	L0 – 8 part plasmids, L1 - custom cassette plasmids can be created with L0 plasmids, L2 – custom multigene plasmids can be created with 5 assembly connectors and custom cassette plasmids	up to 6 transcription units	yeast	features a library of characterised DNA parts	(Lee *et al*., 2015)
mMoClo (MoClo derivative)	BsaI and BpiI site removal	0, 1, 2	L0 – 6 part positioning plasmids, L1 – 9 TU positioning and 9 linker plasmids, L2 – 1 destination vector	up to 9 transcription units	mammalian cells	-	(Duportet *et al*., 2014)
CIDAR MoClo (MoClo derivative)	BsaI and BbsI (BpiI) site removal	0, 1, 2	L0 – 16 basic part vectors (DVA), L1 – 6 TU vectors (DVK), L2 – 16 device plasmids (DVA)	custom TU expansion by alternating between DVA and DVK vector assemblies	*E. coli*	reaction cost and duration reductions (as compared to the MoClo framework); features a library of characterised DNA parts	(Iverson *et al*., 2016)

framework name	domestication requirements	assembly levels	number of plasmid vectors per each assembly level (L)	cloning capacity (number of TUs)	organism preference	additional comments	references
EcoFlex (MoClo derivative)	BsaI, BsmBI and BpiI site removal	0, 1, 2, 3	L0 – 2 BioPart vectors, L1 – 6 TU vectors, L2 – 6 pathway plasmids, L3 – 3 pathway plasmids, and additional "secondary modules"	up to 20 transcription units	*E. coli*	secondary modules allow optimising expression of a TU subset; features a library of characterised DNA parts	(Moore *et al.*, 2016)
Golden Braid 2.0	BsaI, BsmBI and BtgZI site removal	0, α and Ω	L0 – 1 part domestication pUPD vector, L alpha and omega - 16 pDGB vectors	custom pairwise TU expansion, by alternating between alpha and omega vector assemblies	plants	-	(Sarrion-Perdigones *et al.*, 2013; Vazquez-Vilar *et al.*, 2017)
Mobious Assembly	BsaI and AarI site removal	0, 1, 2	L0 – 1 part domestication mUAV vector, L1 and 2 – 4 acceptor vectors and 7 auxiliary plasmids each	custom geometric series-like transcription unit expansion	*E. coli*	usage of a rare cutter (AarI) simplifies the domestication process	(Andreou and Nakayama, 2018)
YeastFab	BsaI and BsmBI site removal	0, 1, 2	L0 – 3 plasmids (HcKan_P, O and T), L1 – 11 POT plasmids, L2 – 2 plasmids (low- and high-copy)	up to 6 transcription units (by assembling appropriate POT units into L2 plasmids)	budding yeast	features an expanding genome-wide library of characterised TU parts	(Guo *et al.*, 2015)

framework name	domestication requirements	assembly levels	number of plasmid vectors per each assembly level (L)	cloning capacity (number of TUs)	organism preference	additional comments	references
EMMA	BsaI and BsmBI site removal	0, 1, 2	L0 - 25 part domestication plasmids, L1 and 2 - 1 receiver plasmid each	up to 4 transcription units	mammalian cells	enables using a wide range of DNA part types (25 part slots) per each L1 plasmid	(Martella *et al*., 2017)

1.2.2.4 DNA Editing

Some synthetic biology projects additionally require various DNA alterations in a given host genome, introduction of which can be challenging, depending on the host organism (Gaj *et al*., 2016). Only a handful of characterised living organisms are capable of efficiently recombining homologous DNA strands (e.g., *Saccharomyces cerevisiae* or *Bacillus subtilis*) (Juhas and Ajioka, 2016). Therefore, DNA cassettes bearing the desired DNA alterations have to be often flanked by lengthy DNA homology regions in order to succeed at altering a given genomic locus. Moreover, such DNA cassettes need to include DNA markers as well, to select for the successfully modified clones, which means that even more extra DNA is required and a suitable DNA screening protocol has to be implemented. Furthermore, one might even run at risk of interfering with organism's phenotype in an unwanted manner (Sauer, 1994; Jessop-Fabre *et al*., 2016; Leng and Song, 2016).

DNA editing methods try to address these issues by harnessing organisms' native DNA break repair pathways, with the two major DNA break repair routes being the NHEJ (non-homologous end joining) and HDR (homology-directed repair) pathways (Ceccaldi, Rondinelli and D'Andrea, 2016). Unrepaired double-strand DNA breaks (DSBs) are lethal to any living cell and therefore cells are under strong selective pressure to repair them. These are repaired either by polishing and ligating the exposed ends (NHEJ), which can lead to indel mutations, or by finding a homologous piece of DNA (e.g., coming from a sister chromatid). Therefore, intentional triggering of DSBs offers an opportunity to randomly mutagenise the corresponding DNA loci or to rationally modify them by providing appropriate DNA homology cassettes, without the need of using any selective DNA markers.

Currently, there are three major DNA editing technologies which allow targeted introduction of DSBs. The first one established uses custom proteins, called Zinc Finger Nucleases (ZFNs). ZFNs are transcription factors in which two "fingers" recognise ~3 to 6 DNA nucleotide triplets each. Therefore, appropriate mixing and matching of the finger modules allows recognition of almost any DNA sequence. ZFNs use the FokI endonuclease domain to cleave a given DNA locus. This domain

is active only as a dimer and therefore a good DNA targeting accuracy is typically achieved, since two DNA-binding events are required (Urnov *et al*., 2010). Transcription activator-like effector nucleases (TALENs) are another editing technology which similarly relies on dimeric transcription factor nucleases. TALEN DNA-binding domains however each recognise a single nucleotide. Moreover, TALENs' DNA-protein interactions are less complex, as compared to ZFNs, and hence TALENs are generally more straightforward to design, and so are currently more frequently used (Joung and Sander, 2012).

One common limitation of ZFNs and TALENs however is that a custom protein has to be designed and manufactured every time a researcher wants to introduce a new DSB. CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) DNA editing technology circumvents this issue by utilizing a single type of DNA endonuclease (a Cas protein), which is guided to the target locus by a short crRNA sequence. CRISPR loci are naturally present in bacteria and archaea and were first characterised by Francisco Mojica. In 2005, further work led this researcher to an observation that these repeat DNA sequences include fragments of bacteriophage genomes, which in turn led to a (now confirmed) hypothesis that CRISPR loci play a role in adaptive immunity (Mojica, Juez and Rodríguez-Valera, 1993; Mojica *et al*., 2005).

Upon viral infection of a prokaryotic cell, Cas proteins cleave foreign DNA into small DNA pieces (~20 bp) and incorporate them into the CRISPR locus, in-between the DNA repeat sequences (these DNA fragments are referred to as "protospacer" sequences). This DNA information is later used to combat any other analogous viral infection. Here, a separate set of Cas proteins expresses the CRISPR locus and processes the resulting pre-crRNA molecules into mature crRNAs. An additional RNA species is however usually necessary (in Type II systems) to process pre-crRNAs and trigger formation of the targeting Cas ribonucleoprotein complex – a trans-activating crRNA (tracrRNA). TracrRNA binds to pre-crRNA, through its homology to the DNA repeat sequences, forming an RNA duplex which triggers recruitment of an RNase III and an RNA-guided Cas endonuclease, which then work together to form separate defensive crRNA:tracrRNA:Cas ribonucleoprotein

complexes. To date, many distinct CRISPR/Cas systems have been identified and classified into different classes, yet the one derived from *Streptococcus pyogenes* (belonging to the Type II class) is by far the most popular. The system utilises an RNA-guided Cas9 protein, which requires an NGG protospacer-adjacent motif (PAM) downstream of the targeting sequence. Once bound to the target DNA, Cas9 uses its two endonuclease domains (RuvC and HNH) to cleave both DNA strands, 3 nucleotides upstream of the PAM sequence (**Figure 1.1 B**) (Haurwitz *et al.*, 2010).

Thanks to their straightforward RNA-guided mechanism, CRISPR/Cas systems have quickly become popular DNA editing tools, with one of their biggest advantages being the ease of design and construction of genome-wide DNA editing libraries, which are powerful phenotypic screening tools (Smith *et al.*, 2016; Kurata *et al.*, 2017; Metzakopian *et al.*, 2017; Bao *et al.*, 2018; Guo *et al.*, 2018; Roy *et al.*, 2018; T. Wang *et al.*, 2018). Moreover, Cas proteins have not only been exploited as DNA endonucleases, but also as proteins capable of silencing or activating selected DNA loci – applications, which however first require mutagenizing the Cas DNA cleavage domains and re-engineering the resulting "dead" Cas, dCas, proteins to bear suitable silencing/activating modules (e.g., the Krüppel-associated box, KRAB, domain is frequently used to repress transcription of a target gene) (Qi *et al.*, 2013; La Russa and Qi, 2015; Zheng *et al.*, 2018). Nonetheless, CRISPR/Cas systems also have their limitations. Notably, they have been reported to exhibit off-target DNA-binding activities (Pattanayak *et al.*, 2013; Zhang *et al.*, 2015). Various scoring metrics have therefore been developed to estimate performance of individual crRNAs as well as partially dead "nickase" Cas protein variants, which similarly to ZFNs and TALENs have to dimerise on the target DNA to catalyse DSBs (Ran *et al.*, 2013; Doench *et al.*, 2014; Haeussler *et al.*, 2016; Abadi *et al.*, 2017). More recently, chemical modifications of single-guide RNAs have also been shown to increase CRISPR system specificity (Ryan *et al.*, 2018).

1.2.3 Test

Once the products of DNA fabrication are ready, they have to be screened and tested for the desired function. Typically, this is accomplished using a range of standard and well-established laboratory methods, e.g., spectrometry, chromatography, cytometry or microscopy (Petzold *et al.*, 2015). However considering our growing ability to rapidly build DNA and generate large combinatorial nucleic acid variant libraries, the commonly used screening and testing approaches are increasingly incapable of rapidly evaluating numerous DNA designs and therefore are often becoming synthetic biology's cycle bottleneck (Rogers, Taylor and Church, 2016).

For instance, metabolic engineers are currently applying synthetic biology methodologies to rapidly construct numerous biosynthetic pathway variants, which require further screening and testing to identify the most efficient biomolecule producers. Therefore, scientific progress of the metabolic engineering field is largely dependent on surpassing the DNA construct evaluation bottleneck (Rogers, Taylor and Church, 2016). Liquid chromatography and mass spectrometry are gold standard methods of metabolite measurement. Nevertheless, they are commonly limited to approximately a thousand measurements per instrument daily (Rogers, Taylor and Church, 2016). Recent high-throughput screening methods are trying to address these technological limitations, including mass spectrometry techniques directly analyzing microbial colonies or harnessing acoustic dispensing methods (Fang and Dorrestein, 2014; Sinclair *et al.*, 2016). Nonetheless, rational synthetic biology approaches are still desirable to efficiently and more economically navigate through combinatorial spaces. At present, biosensor-based methods are one of the main synthetic biology solutions to the metabolic pathway evaluation issue. Namely, genetic biosensor modules can be engineered in production hosts, *chassis*, to sense the biomolecules being synthesised and report on their yields. Different biosensor outputs can be used to support either screening or selection efforts. For example, fluorescent protein signal strength can indicate production levels and flow cytometry methods can be later used to screen and sort cells according to their metabolite yields. Alternatively, expression of an antibiotic resistance marker can be linked a with metabolite's

biosynthesis, so that only top producers will be able to withstand a certain level of selection pressure (Rogers, Taylor and Church, 2016). To date, several studies have reported successful application of such biosensor devices, including published work on 1-butanol, mevalonate and vanillin biosensors (Pfleger *et al*., 2007; de los Santos *et al*., 2016; Shi *et al*., 2017). Some of the more recent examples feature much more extended functionalities, e.g., can alter mutagenesis rates, proportionally to a biomolecule's yield; and monitor biosynthesis in real-time (Chou and Keasling, 2013; Rogers and Church, 2016).

1.2.4 Learn

In an attempt to build DNA constructs and organisms capable of performing certain functions, one iteration of the synthetic biology cycle might not be enough to meet a previously specified objective. Despite the best efforts of synthetic biologists to forward and rationally engineer biology, researchers continue to stumble across various unexpected obstacles, due to our still limited knowledge of how biological systems function (Kwok, 2010). Such discrepancies between expectations and observations provide an opportunity to *learn* new biology revise our prior beliefs and update any established molecular interaction models. Furthermore, the increasing amounts of manufacturing and biological data we are currently generating provide yet another occasion to discover any previously unknown patterns and dependencies through data mining (McCulloch, 2013). Therefore, the field of synthetic biology is now also intersecting with such research fields as the systems biology and machine learning fields (Nesbeth *et al*., 2016).

Systems biology is a research discipline which concerns itself with computational and mathematical modelling of complicated biological systems to better understand their function. Machine learning, on the other hand, is a computer science research field which uses statistical methods to allow computer systems to learn patterns and dependencies in data without being thoroughly programmed to do so. Thus, systems biology and machine learning expertise can both be used to allow new insights into biological data.

During the *Test* stage, different kinds of experimental data can be collected and their subsequent analysis reveals whether performance of a rationally designed biological system is consistent with the specifications stated beforehand. If not, revision and updating of various model parameters and assumptions is required. Here, a synthetic biologist could, e.g., go back to one of the CAD modelling frameworks used and adjust the relevant model parameters. Otherwise, different modelling languages could provide more flexible means of studying the unexpected results. For instance, rule-based modelling languages, such as the Kappa modelling language, allow the user to indirectly construct a mathematical model by specifying a set of rules. These modelling frameworks are therefore straightforward to use and were, e.g., applied by undergraduate students participating in the iGEM competition, to model a complex mechanism of light-based communication in *Escherichia coli* (Stewart and Wilson-Kanamori, 2011; Wilson-Kanamori *et al*., 2015).

Modelling analyzes might however be insufficient to account for the observed experimental results. Information needed to elucidate a certain biological phenomenon might be often inaccessible to scientists, which impedes efficient model updating and refining. Therefore, machine learning methods can be used to mine data in search of its hidden characteristics. Thus far, machine learning approaches have helped in gaining of several new insights into different biological systems. For example, data mining studies allowed mapping of the relationship between promoter sequence and promoter strength, which led to a further development of custom synthetic promoters (Meng *et al*., 2013). Work exploring messenger RNA (mRNA) secondary structure, on the other hand, made it possible to develop a prediction tool capable of predicting translation rate given an mRNA sequence; while analysis of the CRISPR system target nucleic acid sequences and the corresponding Cas9 cutting efficiencies allowed construction of a machine learning model predicting the cleavage performance (Huang *et al*., 2011; Abadi *et al*., 2017). More widespread application of the machine learning algorithms however requires a continuing effort to make these methods more accessible to researchers who do not have the relevant background knowledge. This is now achieved by developing more high-level

machine learning software and computer programming libraries (e.g., the e1071 and scikit-learn R and Python programming language packages, respectively) (Piccinini *et al*., 2017).

1.3 High-throughput Technologies Help Accelerating the DBTL Cycle

Increasingly widespread application of the synthetic biology approaches coincides with the growing implementation of high-throughput technologies in the field of life sciences and creates an even bigger demand for high-throughput laboratory methods (Chao *et al*., 2017). This trend reflects synthetic biology's main goal of making forward engineering of living organisms easier.

1.3.1 A Brief History of Laboratory Automation

In the past few decades, laboratory automation has been focusing on reduction of protocol durations (e.g., through simultaneous processing of large sample batches) and progressive elimination of the error-prone human labour. One of the first reports of an automated laboratory device was in 1875 and described an automated filtration device (Stevens, 1875). In 1894, Greiner published his work on developing an automated pipette, which was designed to help scientists determining milk fat content using the Babcock test (Greiner, 1894). At the beginning of the new century, electrical equipment for conductivity measurements facilitated development of the first commercially-available automated gas detection devices for both laboratory and field use (Taylor and Hugh, 1922). In the 1950s, laboratory automation was becoming even more mainstream. For instance, for the first time a high-throughput device was connected with a digital computer, namely the Atlantic Refining Company released a mass spectrometer capable of performing a biochemical analysis of a complex organic mixture in about 10 minutes and automatically typing out its results (Olsen, 2012). Moreover, microplates were starting to become a popular laboratory consumable (their design specifications became later standardised by the American National Standards Institute (ANSI)) (Olsen, 2012; Astle, 2016b, 2016a). Three decades later, in 1984, the first computer-controlled and programmable robotic arm was developed and was able to pick up and transfer labware (Olsen, 2012). This

achievement paved the way for integrating multiple laboratory automation devices to conduct larger scale processes.

1.3.2 Modern Lab Automation for Synthetic Biology

1.3.2.1 Accelerating DNA Fabrication

Today, a life sciences laboratory has a variety of commercial automated laboratory devices to choose from and synthetic biology workflows can benefit from this choice (Chao *et al*., 2017). A typical synthetic biology DNA construction workflow involves: (1) preparation of DNA assembly mixes; (2) incubation or thermal cycling of the DNA assembly reactions; (3) a heat shock *E. coli* bacterial transformation of the assembly products; (4) bacterial colony picking, to identify clones propagating the right DNA assembly construct; (5) DNA extraction from the candidate clones; and (6) DNA restriction digestion and/or sequencing verification of the purified DNA samples. The verified DNA constructs can be subsequently characterised in a chosen host organism by, e.g., analyzing its growth and responses to various environmental cues. Currently, all these steps can be facilitated by dedicated automation systems (**Fig. 1.4**).

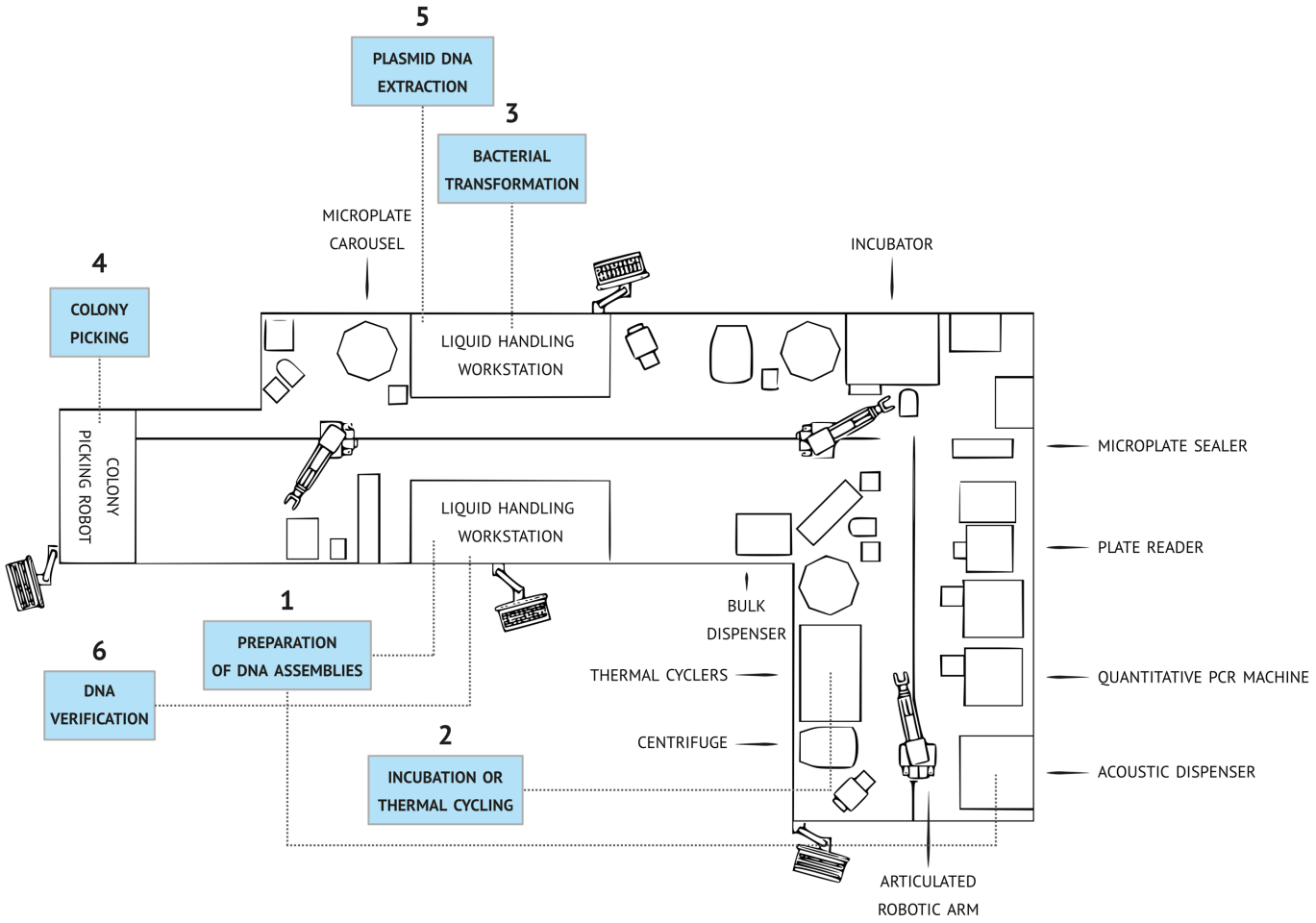


Figure 1.4 Integrated Automation of DNA Fabrication

At present, most of the DNA fabrication operations can be automated. Example DNA manufacturing steps and the corresponding robotic devices are indicated. Integrated automation of different DNA manufacturing processes requires use of articulated robot arms, which transport microplates (bearing biological samples) between different pieces of automation equipment, as well as specialised software, to enable communication between robots. Robotic setup outline - courtesy of the Edinburgh Genome Foundry.

First, several liquid handling technologies exist, e.g., using tips, pins, valves or sound waves, to transfer liquids, which can be used to automatically prepare the DNA assembly mixes (Kong *et al*., 2012). Liquid handling devices can prove especially useful when dealing with large 96, 384 or 1,536 sample batches, which are tedious to handle and prone to human error. Moreover, they can also accurately transfer even nanolitre liquid amounts, which is something that cannot be done by hand using a standard single-channel pipette. Liquid handling devices have now also become larger workstations (e.g., the Tecan, Hamilton Robotics and Beckman Coulter workstations) performing various additional tasks, e.g., purifying nucleic acid samples or shaking and incubating microplates. Therefore, they can be used to run isothermal or thermal cycling DNA assembly protocols, set up DNA transformations or extract DNA from microbial cultures. Pin tools are yet another type of high-throughput laboratory device, able to pick and re-array microbial colonies as well as plate various microorganisms onto solid media in a sterile manner (Cleveland and Koutz, 2005; Kong *et al*., 2012). Microbial clones can be picked, based on their colony size or colour, which facilitates the downstream verification efforts. DNA restriction digestion is one of the typical primary DNA verification methods and involves digestion of a DNA assembly product with a suitable restriction endonuclease (or a mixture of these),to yield a specific pattern of DNA fragments, which is used by researchers to validate DNA assembly. The fragments first have to be separated and visualised by gel electrophoresis before they can be analysed. At present, a number of high-throughput solutions exist to automate gel electrophoresis protocols. For instance, liquid handling workstations can prepare and incubate digestion reactions and load their DNA products into wells of suitable pre-cast gels (e.g., the E-Gel agarose gels, from Thermo Fisher Scientific) (Wu *et al*., 2016). Furthermore, high-throughput capillary gel electrophoresis instruments can perform electrophoretic separations with greater sensitivity, if necessary; are able to quantify DNA more accurately, as compared with regular agarose gels and gel imaging systems; and can perform automated gel image analyses, outputs of which are computer-readable (Li *et al*., 2018). PCR is yet another primary DNA verification method, which uses DNA primers designed to differentiate between the right and wrong DNA assembly products. Similarly to the DNA restriction digestion methods,

PCR reactions, including their downstream electrophoretic analyses, can be run by suitable liquid handling workstations (i.e., including an on-deck thermal cycler and a gel electrophoresis unit) (Chao *et al*., 2017). High-throughput quantitative PCR (qPCR) methods can also be used, as alternatives to the traditional end-point PCR approaches. These methods do not require any downstream gel electrophoresis analyses and, furthermore, provide digital PCR reaction results right at the end of a thermal cycling run. A recent publication by Mitchell *et al*. reports such a method, developed to genotype an entire synthetic yeast chromosome in a single qPCR run, which involved 1,536 reaction wells (Mitchell *et al*., 2015).

1.3.2.2 NGS for High-throughput DNA Sequence Analysis

Besides DNA restriction digestion and PCR methods, more stringent approaches are commonly required to confirm correctness of the DNA assembly products (notably, when PCR-based DNA assembly protocols are used). DNA sequencing can detect small nucleic acid sequence polymorphisms, which might be overlooked by the methods mentioned above. Here, Sanger sequencing methods are a common choice. However, their limited throughput is currently encouraging scientists to harness some of the next-generation sequencing (NGS) technologies instead (Green, Rubin and Olson, 2017; Shendure *et al*., 2017). In contrast to the Sanger sequencing methods, NGS allows sequencing of heterogenous DNA samples and combining tens of such samples in a single sequencing run. A single NGS run generates millions of sequencing reads (as compared to a single Sanger sequencing read). Therefore, bioinformatics analyses are carried out to map these against the reference nucleic acid regions. Each of these regions is usually covered by multiple read sequences, which allows an accurate insight into unknown DNA variations (Magi *et al*., 2010). To date, NGS technologies have made it possible to, e.g., completely sequence 4,000 plasmid DNA assemblies, in a single NGS run, with a coverage of 15 reads per plasmid region and a cost of less than $3 per plasmid, which led to an overall 20-fold cost reduction as compared to the Sanger sequencing process (Shapland *et al*., 2015). Moreover, NGS workflows were also used to screen sequence errors in chip-synthesised oligonucleotides and track off-target CRISPR genome editing events

(e.g., the GUIDE-Seq method) (Matzas *et al*., 2010; Tsai *et al*., 2014). Lastly, NGS is increasingly used in large-scale phenotypic screening projects, as it allows researchers to analyze multiple screening experiments in parallel (i.e., with the so-called "Bar-Seq" methods). Here, short, pre-designed barcode sequences are used to measure growth of individual cells in complex pools by quantifying their read depths. Enrichment, depletion or dropout of clones can be thus studied under certain environmental conditions (Robinson *et al*., 2014). Clones of interest can be later further analysed, if necessary. For instance, high-throughput platforms such as microplate readers or micro-bioreactors can more accurately monitor cellular growth as well as dynamically control some of the culture parameters (e.g., the dissolved oxygen concentrations and pH) (Mühlmann *et al*., 2017; Paytubi *et al*., 2017).

1.3.3 Integrated Platforms for DNA Manufacturing

1.3.3.1 DNA Foundries

While high-throughput devices are already assisting synthetic biologists in many tedious tasks, the ultimate goal of automation technologies is to completely free scientists from manual sample handling. This can be achieved by connecting the necessary high-throughput platforms with each other, so that they can work together to accomplish an entire process (e.g., to build, verify and test an entire library of standard promoter DNA parts). Currently, integration of robotic equipment is not only limited to the life sciences industry, but is also becoming increasing employed by academic facilities. For instance, so-called "DNA foundries" are emerging academic DNA manufacturing facilities, which feature semi- or fully-integrated laboratory automation systems using robotic arms to shuttle biological samples between various high-throughput devices. These facilities serve various academic customers and help them advance their scientific projects faster by providing various high-throughput DNA construction and cellular screening services (Chao *et al*., 2017). For example, some of the work presented in this thesis benefited from access to the Edinburgh Genome Foundry (United Kingdom). To date, a number of publications have reported successful applications of the DNA foundry high-throughput platforms. For example, Liang *et al*. demonstrated a high-throughput

process capable of assembling hundreds of DNA constructs encoding transcription activator-like effector (TALE) proteins within one day and with a cost of $5 per TALE (Liang *et al*., 2013). Si *et al*., on the other hand, reported a high-throughput process capable of generating and screening complex genome-wide *S. cerevisiae* overexpression and knockdown libraries and using these to optimise various phenotypes, e.g., isobutanol production or acetic acid tolerance (Si *et al*., 2017). These positive case studies can be encouraging for many academic organisations and laboratories to build their own integrated laboratory automation platforms. This task is now becoming easier as various companies offer custom construction of such systems and assistance in setting up various automated workflows (e.g., HighRes Biosolutions or Thermo Fisher Scientific); thereby minimizing the amount of expert electrical and mechanical engineering knowledge as well as hardware programming skills required.

Nevertheless, efficient operation of such systems requires a suitable software infrastructure. For instance, laboratory information management systems (LIMS) are typically needed to, e.g., track information on numerous biological samples (DNA parts, plasmids, primers, microbial strains etc.) or manage inventories (e.g., of lab consumables and reagents). Manufacturing execution systems (MES), on the other hand, track high-throughput workflows, i.e., transformation of raw biological materials into finished DNA assembly (or, e.g., microbial strain) products (Chao *et al*., 2017; Craig *et al*., 2017). Therefore, synthetic biology labs and facilities have to develop their own software suites, or "stacks", to plan, schedule, execute, manage, track, control and analyze high-throughput workflows. A growing number of computer applications supporting automated synthetic biology protocols is now conveniently emerging. These are often open-source and web-based, and so readily available to the synthetic biology community, which can benefit from these without the need for individual research teams to develop their own software (Appleton *et al*., 2017). For instance, the CIDAR (Cross-disciplinary Integration of Design Automation Research) Lab, based at Boston University, offers versatile software automation tools, which include both CAD and CAM (computer-aided manufacturing) tools (Xia *et al*., 2011; Appleton *et al*., 2014; Nielsen *et al*., 2016).

Puppeteer is one of these, and combines utilities of LIMS and MES systems. Puppeteer allows users to define laboratory protocols, which are later interpreted and transformed into a set of robot- and human-readable manufacturing instructions, as well as assists in tracking and controlling laboratory tasks, reagents and equipment (Vasilev *et al*., 2011). PaR-PaR, on the other hand, is a robot programming language developed by JBEI (Joint BioEnergy Insitute). This robot communication tool features a biologist-friendly syntax and is designed to simplify the task of writing complex biological protocols for robotic platforms (Linshiz *et al*., 2013). In the future, widespread adoption of a single robot programming standard could facilitate sharing of high-throughput protocols between different labs worldwide.

1.3.3.2 Cloud Labs

Despite the increasing engineering and software support, costs of setting up and operating high-throughput systems are however still considerably high (i.e., oscillating around millions of British pounds) (Chambers, Kitney and Freemont, 2016). Therefore, since not every academic organisation can afford such an expense, and consequently not every laboratory can have access to a DNA foundry, other high-throughput services are emerging. For instance, "cloud labs" such as the Emerald and Transcriptic Cloud Laboratories, offer remote laboratory automation services, which involve customers specifying their high-throughput experiments and shipping the starting biological materials. Results data can be then accessed and downloaded online (Check Hayden, 2014). The straightforwardness with which high-throughput data can be obtained through such remote services encourages reflection on the future of high-throughput synthetic biology. Outsourcing high-throughput experiments might at some point become a routine practice and perhaps results data will be commonly used to automatically trigger other experiments necessary to reach a pre-specified objective. Recent work on "robot scientists" able to autonomously formulate and test biological hypotheses suggests that such developments can be anticipated (King *et al*., 2009).

1.4 Remaining Challenges Facing Synthetic Biology

Despite recent advances in the field of synthetic biology, including the increasing application of high-throughput methodologies, new technological challenges are nevertheless arising. Therefore, this thesis will focus on three particular challenges: (1) further boosting throughput of nucleic acid manufacturing to meet the growing demand for synthetic DNA, which requires more cost-effective DNA fabrication methods; (2) development of more efficient genome-scale DNA engineering methodologies, to facilitate mapping of complex genotype-phenotype relationships; and, (3) finally, harnessing the growing amounts of data generated by high-throughput nucleic acid manufacturing pipelines to allow more efficient predictive analytics and therefore more robust DNA manufacturing process optimisation (McCulloch, 2013; Esvelt and Wang, 2014; Katz *et al*., 2018).

1.4.1 1st Challenge: Lowering DNA Assembly Costs

DNA assembly is one of the main enabling technologies for synthetic biologists and as synthetic biology workflows are becoming increasingly high-throughput there is a growing need for making it cheaper (Katz *et al*., 2018). Miniaturisation of DNA assembly reactions is one of the possible approaches towards achieving a more cost-efficient DNA fabrication.

Traditional, tip-based high-throughput liquid handling technologies are usually incapable of accurately transferring liquids below the microlitre threshold (Ellson *et al*., 2016). Nevertheless, a couple of alternative methodologies now exist, which are adept at moving sub-microlitre volumes. For instance, pin tools are able to transfer liquid amounts as low as ~2 nL, and have proven useful for, e.g., high-throughput screening of compound libraries, usually dissolved in dimethyl sulfoxide (DMSO) (**Fig. 1.5 A**) (Cleveland and Koutz, 2005; Kong *et al*., 2012). High concentrations of DMSO in the screening assay reactions are undesirable due to toxicity of this solvent and therefore minimal compound library transfers are desirable, to e.g., avoid any intermediate dilution steps. Furthermore, other common applications of this technology include replication and re-arraying of large biological libraries.

In order to ensure accurate liquid transfer, pin tools however require control of various factors. For example, pin diameter, surface tension of the liquid material being transferred, pin retraction speed and immersion depth (both in the source and destination liquid), pin dwell period in an empty destination well and volume of the pin slot/groove are some of the key factors controlling the amount of fluid material being moved. Moreover, every pin tool liquid transfer requires a thorough wash cycle to avoid contamination issues. Different concerted approaches, combining use of chemical solutions (e.g., bleach), sonication, UV light sterilisation and mechanical cleaning methods (e.g., using brushes) as well as lint-free blotting strategies, are therefore applied. Additional pin coatings are also exploited, to prevent non-specific binding of proteins and lipids onto the pin surface as well as liquid dislocation throughout the transfer process. Lastly, pin tools are rather inflexible when it comes to moving different liquid volumes in a single liquid handling workflow and can necessitate mounting of several suitable pin arrays to complete a given liquid transfer operation (Cleveland and Koutz, 2005).

Other liquid transfer technologies are now emerging to address these limitations, which originate from the contact-based approach. For example, acoustic dispensing technologies are based on a contactless acoustic droplet ejection (ADE) phenomenon and allow flexible transfer of variable liquid amounts through ejection of nanolitre droplets from specialised source plates into various destination microplates, using appropriate sound wave parameters (**Fig. 1.5 B**). The physical ADE phenomenon was first described in 1920s and since then it has been applied in, e.g., ink jet printing (in the 1970s). In the 2000s, this "drop-on-demand" technology was adopted by the field of life sciences and is at present pioneered by such companies as Labcyte Inc. and EDC Biosystems (Ellson *et al*., 2016).

Microfluidic technologies are alternatives to robotic liquid handlers and are becoming increasingly accessible to researchers as high-resolution 3D printing services and benchtop 3D printers are turning into commodities (Gach *et al*., 2017). To date, microfluidic devices helped downscaling various biological protocols,

including DNA fabrication methods such as Gibson and Golden Gate DNA assemblies as well as PCR amplification protocols (Patrick *et al*., 2015; Khilko *et al*., 2018). However, successful application of microfluidic technologies to DNA fabrication and verification (via PCR) requires relevant expertise and has been also shown to necessitate additional supplementation of, e.g., molecular crowding agents and surfactants as well as excess amounts of enzyme, to optimise nanoscale reaction systems, which might not be desirable and cost-efficient (Khilko *et al*., 2018).

In the first results chapter, I therefore present published work on further miniaturisation of the Gibson and Golden Gate methods, using the acoustic dispensing technology, which thus far has not been used to assemble DNA. Miniaturisation efforts led to scaling down the reaction sizes by up to two orders of magnitude. Nanolitre DNA assemblies did not require supplementation of any additional reaction components and led to up to ~200-fold reagent cost reductions. Aside from the DNA fabrication methods, end-point PCR reactions, which are often used to validate synthetic DNA constructs, were also downscaled and were functional in volumes as low as 250 nL.

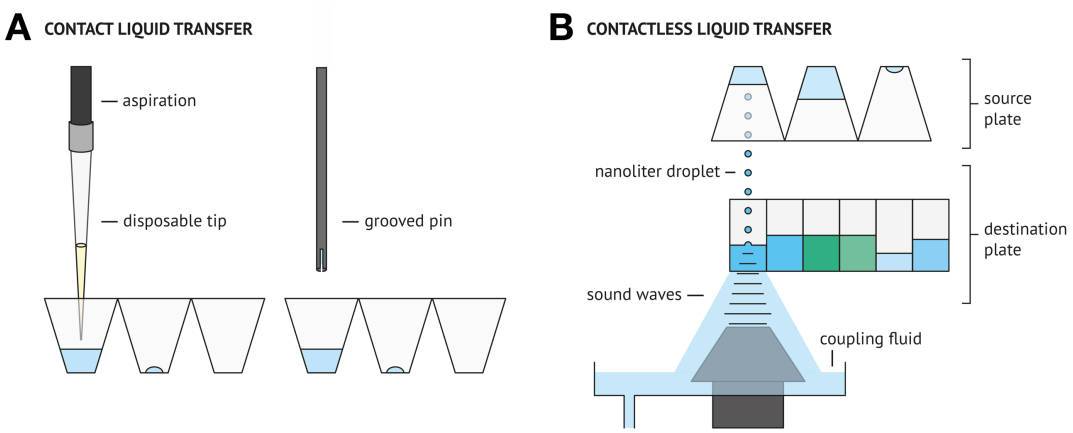


Figure 1.5 Contact and Contactless Liquid Handling Technologies

(**A**) Contact liquid transfer, Contact liquid transfer methods mainly use tips (disposable or fixed) and pins to transfer different types of liquids. Tip-based methods use air or liquid displacement mechanisms to aspirate and dispense liquids. Pin tools, on the other hand, transfer fixed liquid amounts, depending on the physical properties of a given pin and the fluid being transferred (e.g., the liquid's surface tension). Grooved pins are one type of pin tool used, which harness properties of capillaries to move various liquids. (**B**) Contactless liquid transfer, Acoustic dispensing is a contactless liquid transfer method, which uses sound waves to shoot nanolitre droplets from standard source plates to inverted destination plates (empty or pre-filled). Coupling fluid (water) is used to ensure efficient and fast propagation of sound, which travels faster in liquid than in air.

1.4.2 2nd Challenge: Building Tools for Mapping Genotype-Phenotype Relationships

By making DNA fabrication more economical and straightforward, synthetic biologists hope to better elucidate genotype to phenotype relationships. In eukaryotic cells these relationships are primarily governed by complex genetic interaction networks, which ensure cellular robustness against various environmental perturbations (Sanchez *et al*., 1999). Baker's yeast, *Saccharomyces cerevisiae*, is a unicellular eukaryotic organism which for decades has been serving as an accessible model biological system to investigate more complicated molecular mechanisms, e.g., control of human cell function, and consequently has become one of the key eukaryotes to study this intricate cellular wiring (Duina, Miller and Keeney, 2014; Costanzo *et al*., 2016; Kuzmin *et al*., 2018). *S. cerevisiae* cells can be easily genetically modified (thanks to their short doubling times, efficient transfection protocols and robust homologous recombination capabilities), are able to rapidly proliferate on relatively inexpensive carbon sources, are resistant to many growth inhibitors and feature a metabolism rich in various useful biosynthesis precursors. These characteristics have been therefore encouraging application of *S. cerevisiae* strains in both molecular biology and applied research fields (Duina, Miller and Keeney, 2014). The genetic "interactome" studies are valuable for both of these research domains, by increasing our understanding of compound phenotypes, including industrially relevant ones (Si *et al*., 2015, 2017; Costanzo *et al*., 2016; Kuzmin *et al*., 2018).

Synthetic genetic array (SGA) technology was the first high-throughput approach for studying *S. cerevisiae* genetic interactions genome-wide (**Fig. 1.6**) (Kuzmin *et al*., 2016). SGA harnesses standard yeast deletion collections, which comprise S288c genetic background strains individually harbouring complete single open reading frame (ORF) deletions. The ORF deletion mutations target all ~6,000 *S. cerevisiae* genes, which are replaced by a *KanMX4* antibiotic resistance selectable marker and uniquely tagged with short nucleic acid sequences. Automated pin tool devices are later used to mate a given query gene knockout strain with every member of an appropriate deletion library, i.e., of an opposite mating type; and distinct

auxotrophies and antibiotic resistances are used to select for the resulting digenic knockout strains (Giaever and Nislow, 2014). Double gene mutants can feature unexpected phenotypes, considering the underlying single gene deletion effects, which indicate a genetic interaction. Such a genetic interaction can be either positive or negative. Double mutants which exhibit higher than expected viabilities indicate positive gene-gene interactions, while those which become unexpectedly sick or die (i.e., exhibit "synthetic sickness" or "synthetic lethality", respectively) imply negative ones (Kuzmin *et al*., 2016). To date, more than 23 million digenic knockouts have been investigated in *S. cerevisiae* yeast using the SGA method and revealed ~550,000 negative and ~350,000 positive gene-gene interactions. Statistical correlation analysis of these allowed computation of interaction strengths, which were later used to construct a genetic interaction network diagram. This visual representation of the cellular wiring led to a number of observations. For instance, negative gene-gene interactions tend to connect functionally related genes, whereas positive relationships tend to indicate regulatory connections (Costanzo *et al*., 2016). More recently, the first attempts at mapping trigenic interactions (~200,000) in *S. cerevisiae* using SGA were also reported. Kuzmin *et al*., revealed that these link more distantly functionally-related bioprocesses, as compared with the digenic interaction networks (Kuzmin *et al*., 2018).

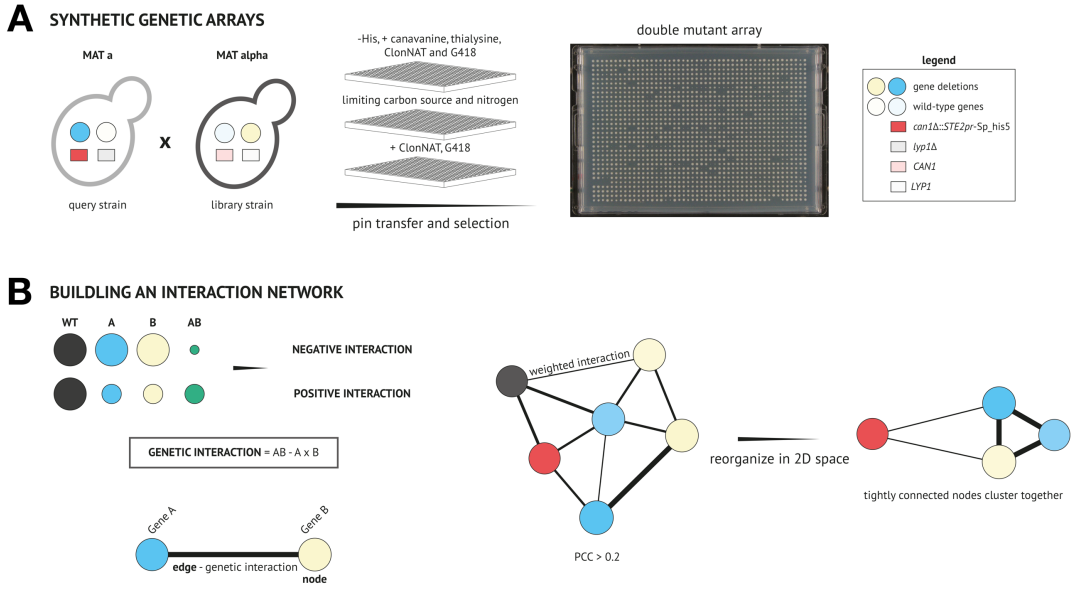


Figure 1.6 SGA Genetic Interaction Studies

(**A**) Synthetic genetic arrays, A query strain harbouring a gene deletion of interest is mated with the Yeast Knockout Collection. Dense mating arrays are generated by high-throughput pin replicators. Diploid yeast strains are selected using a double antibiotic selection, corresponding to different antibiotic resistance cassettes integrated into the query and library strains (kanMX4 and natMX6). Replication of diploids onto a growth medium with limiting carbon source and nitrogen amounts leads to sporulation. Double haploid mutants are selected using a combined antibiotic (G418 and ClonNat), auxotrophic (histidine) and chemical (canavanine and thialysine) selection. Reduced growth or lack of growth on the resulting double mutant array indicates synthetic sick and lethal genetic interactions, respectively. Image taken by Young and Loewen (Young and Loewen, 2013) (**B**) Building a genetic interaction network, Deviation from the expected double mutant phenotype (A x B) indicates either a positive (yeast cells are less sick than expected) or a negative genetic interaction (more sick than expected). Positive genetic interactions indicate genetic suppression, while negative genetic interactions indicate genes which impinge on the same biological process. Pearson correlation coefficient (PCC) is used to measure strength of genetic interactions and interactions with PCC > 0.2 are used in construction of an interaction map. Statistical measures of interaction strengths provide means for re-organizing the interaction diagram in 2D space, leading to clustering of strong gene-gene interactions.

Nevertheless, despite these valuable insights, throughput of classical SGA methods has its physical scalability limits. Namely, classical SGA methods rely on yeast colony size measurements of densely populated colony arrays as proxies for strain fitness and so the number of double mutants that can be investigated, e.g., per day, depends on the number of colonies which can be fitted onto a single growth medium plate (e.g., too densely populated arrays could cause cross-contamination issues and complicate image analyses) (Kuzmin *et al*., 2016). Therefore, alternative approaches have been emerging to address this bottleneck. These, e.g., study digenic interactions in heterogenous double mutant pools proliferating in liquid growth media, using the unique yeast deletion collection "barcode" sequences in microarray hybridisation (e.g., the dSLAM method) or NGS studies (e.g., the Bar-Seq experiments). Intensity of each of the resulting barcode readouts is later used to infer individual strain fitness and is a readily quantitative measure, as compared with the image analysis approaches (Pan *et al*., 2007; Robinson *et al*., 2014). Moreover, modern NGS analyses are capable of analyzing tens of such genetic interaction screens in parallel, while ensuring statistically significant read depths. Other methods concerned themselves with eliminating SGA's strain S288c genetic background, auxotrophy and selectable drug resistance marker requirements, which can be impractical for some applications. For instance, studies of industrially-relevant phenotypes are often performed in more suitable laboratory strains (e.g., the CEN.PK strains with an enriched set of maltose metabolism genes), rather than the S288c derivatives. Moreover, such studies often aim not only at dissecting the genotypic basis of a given phenotype, but also at evolving an optimal one, which is frequently accomplished by investigating effects of genomic knockouts (or, alternatively, knockdowns, overexpression, or a combination of those) in the presence of, e.g., a certain synthetic metabolic pathway or chemical stress associated with the industrial bioprocess of interest (e.g., elevated concentrations of vanillin or acetic acid, which can lead to a growth defect, during lignocellulosic biofuel production) (Si *et al*., 2015, 2017). Hence, genome editing protocols which heavily dependent on auxotrophic and drug resistance selection are not favored by such applications, as they are limited by the number of available selectable markers as well as being often constrained by sub-optimal and costly growth media formulations, which are

required to maintain a suitable selective amino acid and/or antibiotic supplementation (Sauer, 1994; Leng and Song, 2016).

Novel marker-free genome editing approaches can tackle these drawbacks. For example, CRISPR genome editing approaches take advantage of lethal double-strand DNA breaks as a means of selecting correctly edited cells and are readily accessible in various *S. cerevisiae* strains, given an appropriate short crRNA sequence design. Consequently, a growing number of different CRISPR engineering methods for this budding yeast is being developed. These methods rely on *S. cerevisiae*'s preferred homology-directed repair (HDR) DSB repair pathway, and include DNA editing approaches which are suitable for constructing genome-wide CRISPR libraries for this model eukaryote, as well as allowing introduction of multiple, "multiplex", genome edits in a single cell (e.g., the HI-CRISPR method or approaches targeting repetitive retrotransposon delta sequences) (Stovicek, Holkenbrink and Borodina, 2017). HDR CRISPR methods have already proved effective in genome-wide targeting of bacterial loci (i.e., the CREATE method) and are now being exploited in *S. cerevisiae* yeast (Garst *et al*., 2016). Recently, three such methods have been reported for this eukaryotic organism (Bao *et al*., 2018; Guo *et al*., 2018; Roy *et al*., 2018).

In the second results chapter, I therefore present an approach to *S. cerevisiae* DNA engineering on a genomic scale using an HDR CRISPR mechanism. This approach harnesses the HDR DSB repair pathway to insert unique DNA barcode cassettes at Cas9 cleavage sites, which were designed to knock out every gene in the *S. cerevisiae* genome via generation of pre-mature stop codons. Barcode cassettes allowed tracking individual single-gene mutants in heterogenous populations of *S. cerevisiae* yeast using next-generation sequencing and served as direct (one-to-one) proxies for mutant abundance tracking. Moreover, the constructed library is compatible with the synthetic yeast chromosomes designed by the Sc2.0 consortium (i.e., it complies with all of their design features, e.g., the relocation of tRNA genes and presence of the synthetic DNA watermark sequences). This genetic tool is therefore meant to ultimately facilitate the process of mapping novel genetic

interactions in the Sc2.0 project strains as well as "debugging" their synthetic genomes.

1.4.3 3rd Challenge: Using Manufacturing Data to Optimise DNA Fabrication

As automation solutions for synthetic biology emerge, DNA fabrication workflows are increasingly resembling industrial production lines. Therefore, there is a growing need for statistical tools for rational setup and examination of DNA manufacturing processes (Chao *et al*., 2017). At present, researchers still tend to make ad hoc decisions and draw intuitive conclusions regarding high-throughput bioprocesses, while these should be supported by logical statistical studies (Jahangirian *et al*., 2010; Cameron, Bashor and Collins, 2014; Mourtzis, Doukas and Bernidaki, 2014). Therefore, quantitative answers are required to questions such as: How beneficial is it to run a given DNA fabrication step overnight? How significant is a particular process parameter, e.g., PCR mean duration or failure rate? How do technician working hours impact total duration of a DNA manufacturing process?

Analysis of DNA manufacturing processes has one more caveat, i.e., every DNA sequence has its unique degree of "manufacturability". Too high or too low GC content, nucleic acid secondary structure, DNA repeats and homopolymers are some of the most problematic DNA features which synthetic biologists have to cope with when synthesizing, assembling and validating different DNA species. Hence, sometimes two distinct DNA constructs, although originating from the same production workflow, can have a distinct impact on the manufacturing costs and turnaround times (TATs) (Oberortner *et al*., 2016). DNA sequence features should therefore ideally also be considered when examining high-throughput DNA fabrication workflows. Such DNA sequence-specific studies could lead to better predictions of the expected process costs and TATs, which could in turn help companies and other DNA manufacturing services (e.g., DNA foundries) to provide better order turnaround time estimates to their customers or to better constrain features of the incoming DNA sequences.

Since the number of high-throughput DNA manufacturing processes is growing, both in industry and in academia, there is an increasing amount of data available concerning durations and failure rates of individual process steps as well as nucleic acid sequence-specific failures (Chao *et al*., 2017). This data is often stored in computer databases supporting function of lab automation processes and can be readily used to develop predictive tools which could predict expected manufacturing costs and TATs, given an input nucleic acid sequence (Craig *et al*., 2017). Such tools require two components. First, means of predicting the probability of failure of a given DNA sequence in a given process step is needed, to incorporate DNA-sequence specific information into the analysis pipeline. This could be accomplished with a statistical classification model, which could be trained using a machine learning algorithm and the DNA construct failure data to infer failure probabilities of other nucleic acid sequences, e.g., based on their chosen features (i.e., supervised learning). Such model could be continuously updated, using the new incoming DNA data, so that its predictive capabilities could increase over time. Second, means of analysing the underlying experimental workflows is needed, given their unique experimental step progressions, step failure "rescue" paths etc. Since these can be complex, and therefore hard to analyze using analytical methods, Monte Carlo (MC) methods can be used instead. MC simulations are computer experiments sampling random numbers from probability distributions pre-defined by a modeler and best describing a given process random variable (e.g., a PCR reaction duration) (Kroese *et al*., 2014).

In the third results chapter, I therefore propose a simple MC simulation framework, usable by non-specialists, which allows the user to define a manufacturing process model (including its experimental steps, their order, failure rates, rescue paths, step duration probability distributions, as well as the work schedules of the employees involved), using a simple Microsoft Excel spreadsheet, and to simulate it. Apart from the turnaround time estimations, this simulation framework is also able to identify manufacturing steps which have the most significant impact on the production turnaround times, using a one-factor-at-a-time sensitivity analysis, and therefore to guide the user towards the most promising optimisation targets. If desired, DNA

sequence-specific failure estimates can be incorporated into MC simulation studies, thanks to a statistical model which was constructed using industrial DNA manufacturing data from GeneArt (Thermo Fisher Scientific).

Chapter 2
Materials and Methods

2.1 General Materials and Methods

2.1.1 Microbial Strains, Media and Chemicals

2.1.1.1 Microbial Strains

One Shot TOP10 (F⁻ *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) ϕ80*lac*ZΔM15 Δ*lac*X74 *rec*A1 *ara*D139 Δ(*ara- leu*)7697 *gal*U *gal*K *rps*L (Str^R) *end*A1 *nup*G) and MAX Efficiency DH5α Competent Cells (F⁻ Φ80*lac*ZΔM15 Δ(*lac*ZYA-*arg*F) U169 *rec*A1 *end*A1 *hsd*R17 (rk⁻, mk⁺) *pho*A *sup*E44 λ⁻ *thi*-1 *gyr*A96 *rel*A1) from Thermo Fisher Scientific were used to perform chemical transformations of *E. coli* cells. SURE Electroporation-Competent Cells (*e*14⁻(McrA⁻) Δ(*mcrCB-hsdSMR-mrr*)*171 endA1 gyrA96 thi-1 supE44 relA1 lac recB recJ sbcC umuC*::Tn5 (Kan^r) *uvrC* [F ´ *proAB lacI*^{q}*ZΔM15* Tn*10* (Tet^r)]) from Agilent Technologies were used to electroporate DNA into *E. coli*. All experimental budding yeast work used an S288C-derivative BY4741 laboratory strain (***MATa*** *his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*).

2.1.1.2 Chemicals

Chemicals from Sigma-Aldrich and Formedium were used to prepare microbial growth media. Chemicals used to make various buffer solutions were purchased from Sigma-Aldrich only, unless otherwise stated.

2.1.1.3 Bacterial Media

LB (Lysogeny Broth) and SOC (Super Optimal Catabolite repression) media were used in bacterial cell experiments. The LB medium contained 1% w/v tryptone and NaCl, and 0.5% w/v yeast extract. Its pH was adjusted to 7.0 and the resulting liquid medium was then sterilised. LB plates were made by combining 2-times concentrated liquid LB medium with sterile 4% w/v agar. LB media harbouring antibiotic selection were used in this thesis project. Therefore, depending on the plasmid antibiotic resistance marker, sterile carbenicillin and kanamycin at final

concentrations of 50 μg/mL were added. The SOC medium was used to recover cells after transformation. The medium was sterilised and contained 0.5% w/v yeast extract, 2% w/v tryptone, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$ and 20 mM dextrose.

2.1.1.4 Yeast Media

All yeast growth media were first prepared as 2-times concentrated (2X) solutions and filter-sterilised. Depending on whether a liquid or a solid medium was needed, these were 2-times diluted with sterile ddH_2O or sterile 4% w/v agar. The rich YPD (Yeast extract Peptone Dextrose) medium contained 1% w/v yeast extract, 2% w/v peptone and 2% w/v dextrose. The minimal SC (Synthetic Complete) medium contained 6.9 g/L of yeast nitrogen base (Formedium) without amino acids and with ammonium sulfate, 2% w/v dextrose carbon source and an appropriate amount of the Kaiser drop-out mixture (as indicated by Formedium). For the purpose of *URA3* marker plasmid counter-selection experiments, SC media (containing all amino acid supplements) were in addition supplemented with 5-fluoroorotic acid (5-FOA), at a final concentration of 1 g/L. Ingredients of all media formulations were always combined in their powder forms and dissolved in sterile ddH_2O, with a sterile magnetic stirrer and a heat block (set at ~50°C) with a stirring option.

2.1.2 Molecular Cloning

2.1.2.1 Commercial DNA Purchasing

Primer DNA (DNA Oligonucleotides and Ultramer DNA Oligonucleotides) and larger DNA constructs (gBlocks Gene Fragments) were purchased from Integrated DNA Technologies. The oligonucleotide pool used in experiments presented in Chapter 4 was bought from Agilent Technologies.

2.1.2.2 DNA Assembly

Gibson and Golden Gate DNA assembly was performed in this thesis project. 15 μL Gibson assembly master mixes were used to assemble DNA pieces with terminal ~40 bp homologies. Each master mix contained 4 μL of a 5-times concentrated ISO buffer, 0.16 μL of T5 Exonuclease (cat. no: T5E4111K; from Epicentre), 0.25 μL of Phusion High-Fidelity DNA Polymerase (cat. no: F-530L; from NEB), 2 μL of Taq DNA Ligase (cat. no: M0208L; from NEB) and 8.59 μL of distilled water. The ISO buffer consisted of 0.5 M Tris-HCl pH 7.5, 50 mM $MgCl_2$, 1 mM dATP, 1 mM dTTP, 1 mM dCTP, 1 mM dGTP (dNTPs from NEB; cat. no: N0446S), 50 mM DTT, 25% w/v PEG 8000 and 5 mM NAD^+. Gibson assembly was performed by combining 15 μL of the master mix with 5 μL of an equimolar DNA fragment mixture (~0.02 – 0.5 pmoles of DNA fragments) and incubating the resulting reaction mix at 50°C for 1 hour. Golden Gate DNA assembly was used to assemble DNA fragments with compatible 4 nt ssDNA overhangs generated by cleaving flanking BsaI restriction sites. A 7.5 μL Golden Gate reaction mix contained 0.7 μL of concentrated T4 DNA Ligase (2,000 units/μL; from NEB), 0.7 μL of BsaI-HF restriction endonuclease (NEB), 1 μL of T4 DNA Ligase Reaction Buffer (NEB), 0.5 μL of bovine serum albumin (cat. no: B9000S; NEB) and 4.6 μL of an equimolar DNA fragment mixture. Once prepared, the mix was subjected to a thermal cycling protocol of 15 cycles of 5 min at 37°C and 10 min at 16°C, 5 min at 50°C, 10 min at 80°C and final storage at 4°C.

2.1.2.3 Bacterial DNA Transformation

E. coli cells were transformed with plasmid DNA using either a chemical heat shock method or electroporation. The chemical protocol started with a brief thawing of chemically-competent bacterial cells on ice. Plasmid DNA was added and mixed with the cells, with the DNA volume always constituting less than 10% of the competent cell volume. The DNA-cell mixture was next incubated on ice for 20 min, heat shocked at 42°C for 45 sec and, again, incubated on ice for 5 min. 950 µL of pre-warmed (to 37°C) SOC medium was then pipetted into the mixture and the

transformant cells were allowed to recover at 37°C for 1 hour, with 200 rpm shaking. Following this step, bacterial cells were concentrated by brief centrifugation at 17,000 x g and/or diluted, and plated onto solid LB medium plates harbouring a relevant antibiotic for plasmid selection. The plates were incubated overnight at 37°C. A MicroPulser Electroporator (Bio-Rad) was used to conduct DNA electroporation, and protocols of the electroporation device and the competent cells manufacturers were followed when electroporating *E. coli* cells.

2.1.2.3 Yeast DNA Transformation

Mid-logarithmic growth phase cells were used as competent cells in the yeast transformation procedures. Yeast competent cells were always prepared fresh, prior to the transformation protocol. A 10 mL single-colony culture was first prepared by overnight incubation at 30°C in YPD medium with 200 rpm shaking. The next day, a pre-warmed (30°C) 2X YPD medium was inoculated to a starting concentration of 0.5 ODU/mL using the overnight culture. Pre-warmed (30°C) flasks were always used and were filled with the inoculated medium up to a 20% of their total volume to ensure sufficient culture aeration. The cells were then allowed to divide twice and reach a concentration of 2.0 ODU/mL. This growth takes about 4 hours. Each transformation reaction typically requires 10 ODU. Therefore, in order to carry out, e.g., 10 of them, a 50 mL mid-logarithmic phase culture is needed. Once the mid-logarithmic phase was reached, the cells were collected by centrifugation at 3,000 x g for 10 min. Two wash cycles with 25 mL of sterile ddH_2O were next conducted, using the same centrifugation settings. The washed cell pellets were then transferred, using the remaining liquid, to 1.5 mL polypropylene (PP) tubes (100 ODU per tube). Again, the yeast cells were spun down by centrifugation (17,000 x g for 30 sec) and were next re-suspended, with sterile ddH_2O to a volume of 1 mL. 100 µL portions of cells (10 ODU) were distributed across all transformation reaction 1.5 mL PP tubes. Cells were pelleted by 30 seconds centrifugation with the same speed settings. A transformation reaction mix was next transferred to each of the tubes and mixed with the cells. Each transformation reaction consisted of 273 µL of 44% w/v PEG 4000, 36 µL of 1 M LiOAc, 20 µL of 10 mg/mL herring sperm DNA (pre-boiled at 100°C

for 10 min and briefly cooled on ice; Promega), up to 5 µg of plasmid DNA and sterile ddH_2O up to a volume of 360 µL. The transformation mixtures were incubated at 30°C for 30 min and were next subjected to heat shock at 42°C for 15 min. Following this step, the transformant cells were collected by a brief 30 sec centrifugation at 17,000 x g and re-suspended in sterile ddH_2O (typically > 400 µL). The cells were next plated onto pre-warmed (30°C) solid SC medium plates with suitable auxotrophic plasmid selection and incubated for 4 to 7 days at 30°C. 200 and 400 µL liquid volumes were typically plated onto 90 and 150 mm diameter media plates, respectively. Appropriate dilutions were prepared and plated, if necessary. This transformation protocol yielded a CFU range of ~10^5 – 10^6 CFU/µg of plasmid DNA, using the BY4741 yeast strain.

2.1.2.4 Bacterial DNA Extraction

When working with *E. coli* cells, plasmid DNA extraction was performed using a commercial kit from Qiagen (QIAprep Spin Miniprep Kit), unless stated otherwise. The extracted DNA was quantified with a NanoDrop spectrophotometer (Thermo Fisher Scientific), unless a more accurate method was required, e.g., fluorometry.

2.1.2.5 Yeast DNA Extraction

Total DNA was extracted from yeast cells using a phenol-chloroform method. A 50 ODU yeast cell pellet was first re-suspended in 400 µL of lysis buffer (10 mM Tris-HCl pH 8.0, 0.1 M NaCl and 1% w/v Triton X-100). A ~200 µL volume of acid-washed glass beads (Sigma-Aldrich; cat. no: G8772) was combined with the cell suspension. 400 µL of a 25:24:1 v/v phenol:chloroform:isoamyl alcohol mixture (pH 8.0) was then added. The resulting mixture was vigorously vortexed for 10 min and then centrifuged at 17,000 x g. All experimental work involving phenol and chloroform was performed in a suitable fume hood. About 300 µL of the upper aqueous phase was collected and transferred to a clean 1.5 mL PP tube. The DNA was then precipitated with absolute isopropanol (1:1 sample to alcohol volume ratio), by addition of the alcohol followed by 30 min incubation on ice. The precipitated

DNA was collected by centrifugation at 17,000 x g for 10 min and washed once with 80% v/v ethanol. The washed nucleic acid pellet was then briefly dried in a Vacufuge Concentrator (Eppendorf) for 5 min and gently re-suspended in 100 µL of elution buffer (10 mM Tris-Cl pH 8.5). 1 µL of 10 mg/mL RNase A (Thermo Fisher Scientific; cat. no: EN0531) was added and the solution was incubated at 37°C for 30 min to degrade residual RNA. 200 µL of TE buffer (10 mM Tris-HCl, pH 7.5, 1 mM EDTA, pH 8.0) was next added, followed by 300 µL of 24:1 v/v chloroform:isoamyl alcohol mixture. The resulting solution was vigorously vortexed for 5 min and centrifuged at 17,000 x g for 10 min. The upper aqueous phase (about 200 µL) was collected and transferred to a clean 1.5 mL PP tube. DNA was then precipitated using absolute isopropanol, dried with the Vacufuge Concentrator and the resulting DNA pellet was gently re-suspended in 35 µL of nuclease-free water. DNA concentration was quantified with both spectrometry (i.e., with the NanoDrop instrument) and fluorometry (with a Qubit fluorometer and the dsDNA HS Assay Kit from Thermo Fisher Scientific), to account for the presence of any impurities and accurately quantify dsDNA, respectively.

2.1.2.6 DNA Precipitation

Purified plasmid and genomic DNA were precipitated either with absolute ethanol (1:2.5 sample to alcohol volume ratio) or with absolute isopropanol (1:1 sample to alcohol volume ratio). Both alcohol precipitations followed the same experimental procedure. First, 0.1 volume 3 M sodium acetate was added to a given DNA sample. Visibility of precipitated DNA was increased by subsequent addition of GlycoBlue Coprecipitant (Thermo Fisher Scientific) at a final concentration of 50-150 µg/mL. Ice-cold ethanol or room temperature isopropanol was next combined with the resulting DNA mixture. Following this step, ethanol precipitations were incubated at -80°C for ~2 hours or overnight, while isopropanol precipitations were kept on ice for 30 min. The precipitated nucleic acid material was collected by centrifugation at 17,000 x g for 10 min, washed once with 80% v/v ethanol and dried for 5 min, using the Vacufuge Concentrator.

2.1.2.7 Plasmid DNA Restriction Digestion

Restriction endonucleases from New England Biolabs (NEB) were used to cleave plasmid DNA to assess its correctness or to obtain particular plasmid DNA fragments. The manufacturer's instructions were followed when setting up the digestion reactions.

2.1.2.8 Thermal Cycling Equipment

ProFlex PCR System thermal cycling instruments (Thermo Fisher Scientific) were used to conduct endpoint PCR thermal cycling protocols, DNA assemblies and scaled-down bacterial DNA transformations. The LightCycler 96 System (Roche Life Science) was used to perform quantitative PCR experiments.

2.1.2.9 Gel Electrophoresis Assay

Gel electrophoresis was used to separate digested DNA fragments and to judge appropriate restriction digestion DNA products, purity of DNA and its quantity. 1% w/v agarose gels, pre-stained with SYBR Safe intercalating dye (Thermo Fisher Scientific), were used, unless otherwise stated. TAE buffer (40 mM Tris-acetate and 1 mM EDTA; at a pH of ~8.3) was used to prepare agarose gels and run electrophoresis. Bio-Rad equipment was used to run gel electrophoreses at ~100 V for ~40 min as well as image the resulting nucleic acid gels with the Gel Doc XR+ Gel Documentation System. NEB gel loading dyes and DNA ladders were used to track DNA migration through the agarose gel matrix and judge size of the separated DNA, respectively.

2.1.2.10 DNA Extraction from Agarose Gels

MiniElute Gel Extraction Kits (Qiagen) were used to extract DNA from agarose gels and ensured high concentration of the final eluted DNA solutions. The manufacturer's instructions were followed. The Bio-Rad gel imaging system was used to visualise DNA bands targeted for excision.

2.1.2.11 Sanger DNA Sequencing

Plasmid and PCR amplicon DNA was sequenced at the University of Edinburgh DNA sequencing facility (Edinburgh Genomics). Prior to DNA sample submission, chain terminating PCR reactions were performed, using the BigDye Terminator v3.1 Cycle Sequencing Kit (Thermo Fisher Scientific). Edinburgh Genomics PCR protocol specifications were followed (URL: http://genomics.ed.ac.uk/services/sanger-sequencing).

2.1.3 Computational Work

Programming scripts were used to plot graphs illustrating experimental data, analyze output of next-generation sequencing experiments, mine DNA sequence data, run machine learning algorithms and develop computational tools for probabilistic simulation studies. Python programming language (version 3.6.5) was used to write programming scripts. Atom source code editor and Jupyter Notebook were used to document, edit and store computer code. Several Python programming language libraries were used to accomplish different computational tasks. Pandas, NumPy and SciPy libraries were used to conduct general statistical data analyses. Matplotlib and Seaborn libraries were used for data visualisation. Nucleotide sequence alignment analyses were performed with the blastn tool from the NCBI BLAST+ software suite (via Biopython programming tools).

2.2 Chapter 3 Materials and Methods

2.2.1 Acoustic Dispensing Equipment and Consumables

The Echo 550 acoustic dispenser (Labcyte Inc.) was used to set up nanolitre reaction systems. Echo Plate Reformat software was used to define and run liquid handling protocols. Echo Qualified Source Plates were used to store and transfer individual reaction components into inverted destination microplates. In particular, 384-Well LDV COC Source Plates were used to store DNA, while 384-Well Polypropylene

Source Plates were used to store nuclease-free water and reaction master mixes. MicroAmp Endura Optical 96-Well Reaction Plates (Thermo Fisher Scientific) were used as the destination plates.

2.2.2 Nanolitre DNA Assembly

2.2.2.1 Gibson Assembly

Two ~3 kb DNA fragments were amplified from the pPC025 plasmid vector using Phusion High-Fidelity DNA Polymerase (NEB) and following the manufacturer's guidelines. Two pairs of primers were used to amplify the fragments (YCp2391-YCp2392 and YCp2393-YCp2394; all primer DNA sequences are listed in the **Supplementary Table 1.1.1 in Appendix 1.1**). The two resulting DNA amplicons harboured terminal 40 bp homologies and spanned the complete pPC025 sequence. DNA assembly junctions were located within an ampicillin resistance gene (*bla*) and a red fluorescent protein reading frame, therefore providing means for antibiotic selection and colorimetric confirmation of successful DNA assemblies. Successful PCR amplification was confirmed by gel electrophoresis and the PCR products were extracted from the agarose gel and purified. Triplicate 50 nL, 250 nL, 500 nL, 1 µL and 20 µL (positive control) Gibson assembly reactions were set up with the Echo acoustic dispenser in a pre-chilled destination microplate (except for the positive control experiment, which was set up manually) and incubated at 50°C for 1 hour. As the Echo 550 dispenser transfers liquids with 2.5 nL droplets, accurate setup of the smallest reaction systems was challenging (a 50 nL volume constitutes of only 20 droplets). Therefore, volumes of some reaction components had to be rounded up or down to obtain liquid amounts corresponding to multiples of 2.5 nL (**Supp. Table 1.1.2 in Appendix 1.1**)

2.2.2.2 Golden Gate Assembly

Golden Gate DNA assembly experiments aimed at cloning a 500 bp yeast promoter (*pMBP1*) into a ~3 kb plasmid acceptor vector (HcKan_P). The acceptor plasmid encoded a red fluorescent protein ORF flanked by outward facing BsaI endonuclease

recognition sites. The *pMBP1* fragment was amplified from purified BY4741 yeast genomic DNA with primers YCp2395 and YCp2396, using a similar PCR protocol to the nanolitre Gibson assembly experiments. DNA amplification flanked the promoter part with inward facing BsaI endonuclease recognition sites. Subsequent restriction digestion of the plasmid vector and amplicon DNA with BsaI led to generation of compatible 4 nt ssDNA overhang sequences allowing assembly of the promoter part into HcKan_P and replacing the red fluorescent protein reading frame. The amplified promoter DNA was purified using the PureLink PCR Purification Kit (Thermo Fisher Scientific) and used to set up the nanolitre DNA assemblies. Triplicate 50 nL, 250 nL, 500 nL, 1 µL and 7.5 µL (positive control) Golden Gate assemblies were set up similarly, using the acoustic dispenser, and subjected to the assembly thermal cycling protocol (**Supp. Table 1.1.3 in Appendix 1.1**).

2.2.2.3 Bacterial Transformation of the Assembled DNA

Assembled DNA was next transformed into MAX Efficiency DH5α Competent Cells using the chemical DNA transformation method. However, there were several differences in the transformation protocol. Specifically, 20 µL of thawed competent cells were added to microplate wells containing the nanolitre reactions, as it was not feasible to pipette these out of the microplate. Following the heat shock procedure, conducted in a thermal cycling machine, 200 µL of SOC medium were added to each DNA transformation well. Microplates were then incubated at 37°C for 1 hour with 200 rpm shaking. 100 µL of the recovered bacterial cells were plated onto selective LB media (Gibson assembly – w/ carbenicillin, Golden Gate DNA assembly – w/ kanamycin). The resulting inoculated solid media were incubated at 37°C overnight. Only red colonies were expected to grow following the Gibson assembly experiments, and indicated correct DNA assemblies. Both red and white colonies were anticipated from the Golden Gate assemblies where white colonies indicated successful assembly.

2.2.2.4 Validation of the Assembled DNA

1-2 bacterial clones (red – Gibson method, white - Golden Gate method) were verified for each successful miniaturised assembly. Assembled plasmid DNA was extracted from these and purified. The purified plasmids had both of their assembly junctions sequenced by Sanger sequencing to validate successful and correct (without mutations) DNA assembly.

2.2.3 Nanolitre PCR Reactions

The HcKan_P vector (120 ng/μL) was used as a template in the nanolitre PCR experiments, and a pair of primers (YCp2214 and YCp2215) was designed to amplify its red fluorescent protein reading frame region (1,378 bp). GoTaq Green Master Mix (Promega) was used to amplify the target nucleic acid fragment. 50 nL, 250 nL, 500 nL, 750 nL, 1 μL and 10 μL (positive control) PCR reactions were set up in quadruplicate, using the Echo acoustic dispenser and a pre-chilled destination microplate (details of the reaction setups are in the **Supp. Table 1.1.4 in Appendix 1.1**). The reactions were loaded into a pre-heated thermal cycler and the following thermal cycling protocol was run: 2 min at 95°C, 32 cycles of 10 sec at 95 °C, 30 sec at 50°C and 2 min at 72°C, 7 min at 72°C, and final storage at 4°C. Once finished, 5 μL of nuclease-free water was added to each reaction to facilitate their recovery from the microplate. The resulting ~5-6 μL solutions, together with the positive control, were analysed by agarose gel electrophoresis.

2.3 Chapter 4 Materials and Methods

2.3.1 Single-Gene CRISPR Knockout Experiments

2.3.1.1 CRISPR Plasmid Construction

The *ADE2, CAN1* and *SEN34* gene knockout cassettes were ordered as Ultramer DNA oligonucleotides (details of the corresponding crRNA sequences are in the **Supp. Table 3.1.1 in Appendix 3.1**). For the HDR and NHEJ tests, ssDNA gene knockout cassettes were amplified and flanked with 50 bp CRISPR library vector

homologies via PCR. PKp001 and PKp002 primers and Phusion High-Fidelity DNA Polymerase (NEB) were used. The manufacturer's protocol was followed. Amplification products were run on a 2% w/v agarose gel. DNA from correct size bands was extracted and purified. The resulting cassettes were next cloned into SapI restriction enzyme-linearised pGZ110 (with *LEU2* marker) CRISPR system plasmid using the Gibson method. The assembly products were transformed into chemically-competent One Shot TOP10 *E. coli*. Transformant cells were plated onto LB w/ carbenicillin plates and incubated overnight at 37°C. Several clones were later picked per construct and their plasmid DNA was subjected to Sanger sequencing verification of the insert region. Correct plasmid isolates were used in the subsequent yeast knockout experiments. For the HDR positive control experiments, double-stranded donor DNA was prepared by annealing equimolar amounts of forward and reverse Ultramer DNA strands in a thermal cycler (program used: 95°C for 5 min, cool down at a ramp rate of 0.1°C/sec to 25°C and hold at 25°C for 5 min).

2.3.1.2 Yeast Transformation with CRISPR Plasmids

BY4741 yeast strain was transformed with the constructed DNA. Each transformation was set up in duplicate and contained 200 ng of plasmid DNA and, if applicable, 2 μg of the 90 bp double-stranded oligo DNA donor. Dilution series of transformant cells were plated onto SC-Leu medium plates.

2.3.1.3 CRISPR Knockout Efficiency Assessment

All SC-Leu plates were incubated for 3 days at 30°C. Following this incubation period, the total number of colonies recovered after each transformation was recorded for all CRISPR deletion tests as well as the negative control (empty pGZ110 plasmid). For the *ADE2* knockout tests, the plates were in addition refrigerated at 4°C for up to a day to facilitate red pigment development and the ratio of red (mutant) to white clones (a knockout efficiency measure) was quantified. For the the *CAN1* knockout tests the plates were replica plated onto SC-Arg-Leu plates with canavanine (final conc. = 50 mg/L) and the resulting replica plates were further

incubated at 30°C for 2 days. Ratio of colony counts on plates with and without canavanine was computed as the knockout efficiency measure.

2.3.1.4 Further DNA Sequencing Validation of the Gene Mutants

The *ADE2* gene mutant colonies (plasmid HDR donor and barcoded plasmid HDR donor tests) were subsequently subjected to an extra DSB locus Sanger sequencing validation step to confirm 8 bp deletions and barcode insertions, respectively. In order to investigate the corresponding genomic DNA regions using the Sanger sequencing method they had to be first amplified. Mutant yeast colonies were used as colony PCR templates. Each colony was suspended in 50 µL of 20 mM NaOH and incubated at 95°C for 5 min. 10 µL of the resulting lysate was next used in a 50 µL PCR reaction following the GoTaq Green Master Mix (Promega) protocol with an extended 1.5 min annealing step and using primers PKp003 and PKp004. Correct amplification was assessed by gel electrophoresis. The remaining amplicon DNA was purified using the PureLink PCR Purification Kit (Thermo Fisher Scientific) and used as a template in the Sanger sequencing verification.

2.3.2 Genome-wide CRISPR Knockout Experiments

2.3.2.1 CRISPR Deletion Library Cloning

2.3.2.1.1 Oligonucleotide Pool Amplification and Purification

A lyophilised DNA library was re-suspended in ddH_2O to a final concentration of 50 nM as storage stock. A 5 nM working stock was prepared through a 10-fold dilution of the storage stock. The library DNA was amplified and flanked with 50 bp cloning vector (pGZ110 with *URA3* marker) homologies via PCR, using 20 µL of 5X Herculase II Reaction Buffer (Agilent Technologies), 4 µL of the working stock (0.02 pmol), 1 µL of dNTP mix (final conc. of 0.25 mM/dNTP; NEB), 2.5 µL of 10 µM forward (PKp001) and reverse (PKp002) primers, 1 µL of Herculase II Fusion DNA Polymerase (Agilent Technologies) and 63 µL of nuclease-free ddH_2O for a total reaction volume of 100 µL. A 15 cycle thermal cycling protocol was followed:

3 cycles of 95°C for 20 sec, 50°C for 20 sec and 72°C for 30 sec, 12 cycles of 95°C for 20 sec, 56°C for 20 sec and 72°C for 30 sec, followed by a final storage at 4°C to avoid entering the PCR saturation phase and therefore prevent DNA amplification bias (Aird *et al*., 2011). Amplicon DNA was run on a high-resolution 2% w/v MetaPhor agarose gel (Lonza) to confirm successful PCR amplification. A correctly sized DNA band was extracted from the gel and purified.

2.3.2.1.2 Cloning of the Amplified Pool and Purification of the Plasmid Library

The amplified library cassettes were cloned into SapI-linearised pGZ110 CRISPR vector, using the Gibson method. A 3:1 molar ratio of insert to vector DNA was followed. The assembled DNA was electroporated into SURE Electroporation-Competent *E. coli* Cells. The number of electroporations performed was based on efficiency of the test DNA assemblies and aimed at sufficient library representation (~20-fold). The transformed cells were pooled in a pre-chilled tube (kept on ice) and vortexed briefly. 400 µL aliquots of the resulting cell suspension were each plated onto a single 150 x 15 mm LB w/ carbenicillin medium plate. Plates were incubated at 37°C for 18 hours. The colonies were well-separated, to avoid excessive competition for growth resources and therefore growth bias. The colonies were scraped off the agar plates with 2 x 5 mL of cold LB medium and the resulting cell suspensions were transferred to a common pre-chilled bottle and kept on ice. The cell pool was mixed by swirling. Aliquots were made and *E. coli* cells were pelleted by centrifugation. The pellets were stored at -20°C until plasmid DNA extraction. 300 mg wet cell weight pellet portions (approx. 3×10^{11} cells) were used per single plasmid DNA extraction. Qiagen Plasmid Maxi Kit was used to purify the plasmid library. DNA yield and quality were assessed by the NanoDrop spectrophotometer (Thermo Fisher Scientific) and gel electrophoresis.

2.3.2.2 CRISPR Library Yeast Transformation

BY4741 was transformed with the purified plasmid pool, so that ~50-100 colonies were obtained per plasmid. The yeast transformation protocol was appropriately

scaled up to accomplish this coverage, and 5 μg of plasmid DNA were transformed per each ten ODUs used in the transformation procedure. Transformant cells were plated onto the large SC-Ura media plates and the plates were incubated at 30°C for 4 days. A small portion of cells was also plated onto two small SC-Ura media plates to assess coverage of the CRISPR library.

2.3.2.3 First CRISPR Library Screen in Yeast

Following the plate outgrowth period, yeast cells were harvested from the large Petri dishes by scraping them with 2 × 5 mL of sterile cold 15% w/v glycerol solution using sterile L-shaped spreaders. The scraped cells were stored in a sterile bottle and kept on ice until all plates were processed. The resulting cell mixture was mixed by swirling and aliquoted into cryovials, which were later flash-frozen on dry ice and stored at -80°C as glycerol stocks. Serial dilutions of a small sample were also performed to assess cell concentration with spectrometry. One glycerol stock was later thawed and re-suspended in pre-warmed YPD medium and the re-suspended cells were transferred to a pre-warmed 1 L flask, containing 200 mL of the rich YPD medium (pre-warmed glassware was always used, filled up to 20% of its volume, to ensure good aeration). Cell concentration of this starting culture was measured with a spectrophotometer and did not exceed a 1 ODU/mL concentration (2×10^5 library coverage). Further sub-culturing steps aimed at a starting cell concentration of 0.1 ODUs/mL. The YPD culture was incubated at 30°C with 200 rpm shaking until mid-logarithmic cell growth was observed. A portion of cells was washed with sterile water and sub-cultured into 200 mL of pre-warmed plasmid counter-selection medium (SC complete medium w/ 5-Fluoroorotic acid). A portion of yeast cells (50 ODUs) was also washed with nuclease-free water, pelleted by centrifugation and stored at -20°C until DNA extraction. Again, cells were allowed to grow in the minimal counter-selection medium until mid-logarithmic phase and were sub-cultured into the same medium until the counter-selection period reached 24 hours. Once the plasmid counter-selection was completed, again a 50 ODUs portion of cells was washed, pelleted and stored at -20°C. Three more portions were washed with sterile water. One of them was sub-cultured into pre-warmed YPD medium (200 mL)

and serial dilutions of two others were plated onto both SC complete and SC-Ura plates to confirm successful plasmid loss. Plates were incubated at 30°C for 3 days, while the liquid yeast culture was similarly grown until mid-logarithmic phase. Following the rich medium outgrowth, once again, 50 ODU of cells were collected, washed and stored at -20°C. Once the three days plate outgrowth was completed, on the other hand, growth on both auxotrophic and non-auxotrophic minimal media was assessed and the ratio between the SC-Ura and the SC complete media yeast colony counts was computed to assess the magnitude of the CRISPR plasmid loss. Once confirmed, total DNA was extracted from all of the frozen 50 ODU cell pellets using the phenol-chloroform extraction method.

2.3.2.4 Yeast Library Screens with an Additional Plasmid-selective Outgrowth

To assess whether more time was needed for the plasmid-encoded CRISPR system to introduce the designed gene knockouts, a second library glycerol stock was recovered (as described above). The same sub-culturing and sample collection procedure was followed, however this time the library was grown in plasmid-selective SC-Ura liquid medium. Cell samples were collected at two timepoints, at 48 and 96 hours. Total DNA (including plasmid and genomic DNA) was extracted using the phenol-chloroform method.

2.3.2.5 NGS Sequencing Library Preparation, Purification and Quantification

2.3.2.5.1 Illumina Sequencing Library Preparation

Illumina next-generation sequencing libraries were constructed by initial amplification of the barcode sequences from the plasmid or yeast (total or genomic) DNA. PAGE-purified forward P5 and reverse P7 w/ BC1-10 primers were used to amplify barcode DNA, and encoded read 1 or read 2 Illumina sequences as well as the P5 and P7 flow cell adapters. Reverse primers harboured additional 6 bp Illumina index sequences. Different index sequences were used in amplification of samples coming from different growth condition tests. Amplified samples could therefore be mixed and sequenced together in a single NGS run. Q5 High-Fidelity DNA

Polymerase (NEB) was used to perform PCR amplification according to the manufacturer's instructions. For the plasmid DNA templates, a 22-cycle thermal protocol was used. For the yeast DNA samples, the number of cycles was increased to 28. Efforts were made to minimize the number of DNA amplification cycles to avoid reaching the PCR plateau phase, which can lead to barcode abundance bias (Aird *et al.*, 2011). Decreasing the number of PCR cycles can however be impeded by, e.g., poor DNA template purity, in which case the cycle number might have to be increased to achieve a high enough yield for the downstream protocol steps. Increasing DNA template concentrations, reaction volumes or setting up replicate PCR reactions can also help remedying any yield issues.

2.3.2.5.2 Illumina Sequencing Library Purification

A small portion of each PCR reaction sample was next run on a high-resolution 2% w/v MetaPhor agarose gel (Lonza) to confirm efficient amplification of 266-271 bp NGS library cassettes. The remainder PCR reaction volumes were purified using AMPure XP magnetic beads (Beckman Coulter). A 1:1.3 DNA sample to beads ratio was followed and the protocol was carried out using the DynaMag-2 Magnet (Thermo Fisher Scientific), according to the manufacturer's instructions.

2.3.2.5.3 Illumina Sequencing Library Quantification

The eluted amplicon DNA was next quantified with the Qubit fluorometer (Thermo Fisher Scientific), using the double-stranded DNA HS Assay Kit (Thermo Fisher Scientific). An additional qPCR DNA quantification was performed as well, using an Illumina-compatible qPCR master mix kit, from Kapa Biosystems (cat. no: KK4854). The manufacturer's protocol was followed. This extra measurement step allowed assessment and quantification of specific DNA amplification, as opposed to the previous total dsDNA assessment. Results of both quantification protocols should however be consistent, as any DNA misquantification errors can lead to over- or under-clustering of sequencing clusters.

2.3.2.6 NGS Runs and Raw Sequencing Data Processing

2.3.2.6.1 Illumina DNA Sequencing

The Illumina MiniSeq sequencer was used to perform paired-end sequencing of the CRISPR barcode cassettes. Illumina MiniSeq High Output Reagent Kits (75 cycles) were used. These kits achieve a maximum output of 25 million reads per run. Therefore, to achieve an approximate 100X read depth per barcode sequence, up to 10 DNA samples were typically processed by a single NGS procedure. Prior to NGS, the barcode amplicon samples, tagged with the different 6 bp indices, were mixed to yield an equimolar 10 nM sequencing library. This library was later further diluted, denatured, neutralised and combined with Hybridization Buffer, according to the Illumina Denature and Dilute Libraries Guide (URL: https://support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miniseq/miniseq-denature-dilute-libraries-guide-1000000002697-00.pdf). The prepared mixed sequencing library was loaded into a previously thawed and mixed Illumina cartridge, which was then inserted together with a clean flow cell into a primed sequencing machine. Optionally, the NGS sample can also be supplemented with a suitable amount of a PhiX control from Illumina (cat. no: FC-110-3001), as per the above guide instructions. Such a DNA "spike-in" can be harnessed as an NGS clustering reaction or a sequencing error control. Alternatively, it can also introduce sequence diversity to DNA samples lacking it.

2.3.2.6.2 Raw Sequencing Data Processing and Mining

Raw FASTQ format sequencing files were analysed and used to align read 1 and read 2 sequences to the reference 27,000 oligo pool list. A Python script was written to convert this list to a FASTA file. Bowtie 2 software was then used to create a reference alignment library, which was later used to perform a paired-end read alignment of the FASTQ read files. SAM format alignment files were obtained and analysed using a Python script, to compute number of sequencing reads per barcode in the CRISPR library. Barcode read depths were next used as proxies for

investigating individual gene deletion mutant viability in different environmental conditions. Python scripts were used to further analyse the barcode read count datasets.

2.3.3 Development of a Novel NGS Library Preparation Procedure

2.3.3.1 Construction of a Yeast Strain Encoding the New Barcode System

In order to construct a yeast strain carrying the proof-of-concept barcode cassette, a CRISPR homology-based DNA insertion method was used. A previously constructed plasmid, targeting the *ADE2* reading frame, was used to introduce a double-strand break in the yeast genome, which was later repaired with a double-stranded oligonucleotide donor harbouring the new barcode cassette. Donor DNA (139 bp) was purchased as a gBlock. 200 ng of plasmid DNA was co-transformed with 2 µg of the donor fragment into the BY4741 strain. A sample of red (putative) mutant colonies was collected and, as previously, subjected to colony PCR amplification of the *ADE2* gene locus region, followed by gel electrophoresis verification, amplicon PCR purification and Sanger sequencing of the amplicon fragment. A single yeast clone containing a correct barcode insertion was used in the downstream new protocol testing experiments.

2.3.3.2 Purification and Quantification of Test DNA Templates

Genomic DNA was extracted from the positive barcode-containing yeast strain using the phenol-chloroform method and served as one of the two DNA templates in the new protocol tests. The second template was obtained by amplification of a 775 bp genomic DNA region, centrally-encoding the barcode cassette. Q5 High-Fidelity DNA Polymerase (NEB) and primers PKp005 and PKp006 were used to amplify the target genomic DNA fragment. The manufacturer's instructions were followed. Correct amplicon size was confirmed by running a small portion of the finished PCR reaction on an agarose gel. Once confirmed, the remaining DNA was purified using the QIAquick PCR Purification Kit (Qiagen). Large quantities of both template samples were used in the downstream tests (1 µg of each per test) and therefore

replicate extraction and amplification reactions were set up. Purified DNA was then accurately quantified using the Qubit fluorometer and the dsDNA HS Assay Kit (Thermo Fisher Scientific).

2.3.3.3 Streptavidin Magnetic Beads-bound Reactions

2.3.3.3.1 Pre-protocol Preparation

Streptavidin Magnetic Beads (NEB) and the DynaMag-2 Magnet (Thermo Fisher Scientific) were used to perform bead-bound enzymatic reactions. Prior to starting the protocol, magnetic beads were brought to room temperature and the required buffers were prepared (Wash/Bind, Low Salt and Elution Buffers) according to the manufacturer's instructions. The Low Salt and Elution Buffers were later pre-chilled and pre-warmed (to 70°C), respectively.

2.3.3.3.2 Biotinylated ssDNA Elongation

The initial biotinylated ssDNA elongation step was performed using Q5 High-Fidelity DNA Polymerase (NEB) and a biotinylated primer (PKp007). A 50 µL reaction was set up and included 10 µL of 5X Q5 Reaction Buffer (NEB), 2.5 µL of 10 µM biotinylated primer DNA, 1 µg of either of the DNA templates (genomic or amplicon DNA), 1 µL of a 10 mM dNTP mix (NEB), 0.5 µL of Q5 High-Fidelity DNA Polymerase, and nuclease-free water up to a 50 µL volume. The elongation reaction was performed in the ProFlex PCR System thermal cycler and consisted of a single thermal cycle: 3 min denaturation at 98°C, rapid cool down to 72°C (6°C/s), slow cool down to 50°C (0.1°C/s) and a 30 sec hold at 50°C (DNA annealing step), rapid ramp to 72°C (6°C/s), with a 15 min incubation (a prolonged extension step), and a rapid ramp (6°C/s) to 98°C, followed by a final 3 min 98°C denaturation step. Once the program was completed, PCR tubes were immediately put on ice to keep the DNA strands denatured.

2.3.3.3.3 Setup of the Bead-bound Reaction System

Bead-containing 1.5 mL PP reaction tubes were next prepared. 50 µL of mixed streptavidin magnetic beads were aliquoted into two tubes. The magnet was then applied to the side of each tube for approx. 30 sec and the bead storage buffer was removed. A bead wash cycle was next performed. 100 µL of Wash/Binding Buffer were added to each tube, the tubes were vortexed to suspend the beads, the magnet was applied for approx. 30 sec and the supernatant was discarded. The elongation DNA products were transferred into their respective tubes and the resulting DNA-bead mixtures were vortexed briefly and incubated at room temperature for 5 min, with occasional agitation by hand. Following the incubation step, the magnet was applied to the side of each tube and the supernatants were carefully removed. Two wash cycles (as above) were performed.

2.3.3.3.4 ssDNA 5' End Phosphorylation

The first bead-bound reaction was the 5' end phosphorylation. A single phosphorylation reaction mix included 5 µL of 10X T4 PNK Reaction Buffer (NEB), 5 µL of 10 mM ATP (Illumina), 1 µL of T4 Polynucleotide Kinase (NEB) and 39 µL of nuclease-free water (up to a total volume of 50 µL). Once prepared, this was added to each tube and the resulting reaction mixtures were mixed by vortexing. The tubes were then incubated at 37°C for 30 min, with occasional agitation by hand. Once the phosphorylation reactions were completed, the magnet was applied and the supernatants were discarded. Two wash cycles (as above) were performed.

2.3.3.3.5 ssDNA Circularisation

Single-stranded DNA circularisation was the next enzymatic step. Two circularisation reaction mixes were set up, containing: 5 µL of 10X CircLigase II Reaction Buffer (Illumina), 2.5 µL of 50 mM $MnCl_2$, 2.5 µL of CircLigase II ssDNA Ligase (Illumina) and 40 µL of nuclease-free water (a 50 µL reaction). Once prepared, they were added to the bead-containing tubes. The resulting reaction mixes were vortexed briefly and incubated at 60°C for 1 hour, with occasional agitation by

hand. Following this incubation period, the magnet was applied, the supernatants discarded and, again, two wash cycles were performed.

2.3.3.3.6 DNA Linearisation

A single-stranded oligonucleotide (PKp008) was subsequently annealed to the circularisation DNA products at a pre-designed BamHI restriction site, to allow DNA linearisation. Each 49 µL annealing mix contained 5 µL of 10X NEB 3.1 Buffer (NEB), 1.5 µL of 10 µM oligonucleotide DNA and 42.5 µL of nuclease-free water. The mixes were dispensed into the PP tubes, and their contents were mixed and incubated at 95°C for 5 min in a heat block. Once the incubations were completed, the heat block was switched off and allowed to passively cool down to room temperature. This process took up to 1 hour and the tubes were thus, during this time, agitated by hand several times. 1 µL of BamHI restriction endonuclease was next added to each tube. The tubes were vortexed briefly and incubated at 37°C for 30 min, with occasional mixing. Following the linearisation reaction, the magnet was applied and the supernatants were discarded. Two wash cycles were performed.

2.3.3.3.7 DNA Elution and Precipitation

The linearised DNA products were eluted off the magnetic beads. First, a DNA wash with cold Low Salt Buffer was performed. 25 µL of pre-warmed (70°C) Elution Buffer were next dispensed into each tube. The tubes were quickly vortexed and incubated at room temperature for 2 min. The magnet was applied to the side of each tube and the supernatants were carefully collected to clean PP tubes. This elution procedure was then repeated one more time. The resulting 50 µL eluents were later precipitated, using the ethanol precipitation method, and re-suspended in 10 µL of nuclease-free water.

2.3.3.4 Test Sequencing Library Amplification, Purification and Validation

2.3.3.4.1 PCR Amplification

The Q5 High-Fidelity DNA 2X Master Mix (NEB) was used to amplify the linearised DNA. PCR reactions contained 10 µL of nuclease-free water, 2.5 µL of each of the standard Illumina read 1 and read 2 primers, containing their respective P5 and P7 flow cell adapters (P5 and P7 w/ BC1), 10 µL of either of the DNA samples and 25 µL of the Q5 High-Fidelity DNA 2X Master Mix (final reaction volume of 50 µL). A 30-cycle PCR program was used :98°C for 30 sec; 30 cycles of 98°C for 10 sec, 68°C for 20 sec and 72°C for 15 min; 72°C for 2 min and a final 4°C incubation.

2.3.3.4.2 Gel Electrophoresis and Nucleic Acid Extraction

PCR products were run on a 3% w/v agarose gel of 100 V for 40 min. Entire reaction volumes were loaded into the gel wells. Following the electrophoretic separation, the resulting DNA smears were extracted from the gel matrix. Caution was taken to omit gel areas below 200 bp as indicated by the 50 bp DNA ladder, as these were likely to contain primer dimer DNA (Illumina read primers share a 13 bp 3' end homology).

2.3.3.4.3 Blunt-end Ligation and Sanger Sequencing of the PCR Products

4 µL of each of the two purified DNA smears (40% of the total amplicon library) were ligated to a plasmid vector using the Zero Blunt TOPO PCR Cloning Kit (Thermo Fisher Scientific). Total 6 µL reaction mixture volumes were next transformed into 100 µL of the commercial One Shot TOP10 *E. coli* cells. The resulting transformant cells were then plated onto LB media plates with 50 µg/mL kanamycin and incubated at 37°C overnight. 12 colonies were next picked per DNA template test (24 colonies in total) and plasmid DNA was extracted. The DNA inserts were then sequenced using the Sanger sequencing method and a standard M13 primer encoded by the plasmid backbone.

2.4 Chapter 5 Materials and Methods

2.4.1 Development of the Monte Carlo Simulation Software

Pandas Python library was used to parse and analyze Microsoft Excel spreadsheets. The SciPy library (scipy.stats.truncnorm.rvs function) was used to sample random manufacturing step durations from the truncated normal distributions. The NumPy library was used to perform other general statistical analyses, e.g., percentile calculations. Datetime Python module was needed to conduct calendar date-aware simulation studies. Matplotlib and Seaborn libraries were later used for plotting simulation results data.

2.4.2 DNA Sequence Data Mining and Classification

Primer3 and ViennaRNA Python programming packages were used to analyze nucleic acid thermodynamics, i.e., calculate primer hairpin and PCR amplicon free energies, respectively. The nearest-neighbour method was used to calculate primer 3' end - PCR template duplex free energies (Watkins *et al*., 2005). Scikit-learn Python package was used to conduct machine learning studies. Prior to application of the machine learning algorithm, GC content and DNA sequence length data was standardised by dividing nucleic acid features by arithmetic means of the corresponding datasets. Sklearn.svm.SVC class was used to perform support vector classification, with kernel, C (regularisation parameter) and gamma (kernel coefficient) parameters chosen by the sklearn.grid_search.GridSearchCV, a hyperparameter assessment method which used a default three-fold cross-validation. To make it possible to estimate probabilities of belonging to a given data class (i.e., failed gene synthesis or successful gene synthesis), SVC classification object's probability parameter was enabled.

Chapter 3
Acoustic Dispensing for DNA Assembly Miniaturisation

3.1 Work Contributions

Work presented in this chapter was published in the Journal of Laboratory Automation (**Appendix 2.4**) (Kanigowska *et al*., 2016).

Together with Yue Shen, I performed the nanolitre DNA fabrication and endpoint PCR experiments, analysed data and wrote the published manuscript. Besides the DNA assembly reagent cost calculations, described in the manuscript, I also performed additional expected reagent cost calculations, taking into account failure rates of the nanolitre DNA assemblies tested. Yijing Zhang helped in setting up the acoustic dispensing instrument. Prof. Yizhi Cai wrote the published manuscript and together with Prof. Susan Rosser allowed access to the Edinburgh Genome Foundry (UK) acoustic dispenser.

3.2 Introduction

High-throughput DNA manufacturing requires more economical approaches to liquid handling. To date, DNA has been assembled mainly in microlitre-scale reactions (Kong *et al*., 2012). However, there is an interest in scaling their volumes down to a nanolitre range, as such miniaturisation would reduce the use of expensive enzymatic components and valuable DNA parts and therefore lead to cost savings, as well as establishment of more scalable DNA fabrication methodologies (Ellson *et al*., 2016; Gach *et al*., 2017).

Today, tip-based methods are a dominant approach towards liquid handling, but are unable to accurately transfer sub-microlitre volumes. Moreover, due to their strong dependence on disposable tips, they are expensive and unsustainable, i.e., generating large amounts of plastic waste (Kong *et al*., 2012; Ellson *et al*., 2016). Fixed tips eliminate the costs associated with disposable tips, however as they still require physical contact to transfer liquids their usage necessitates thorough wash cycles to prevent any contamination issues (Fregeau *et al*., 2007). Pin tools are liquid handling instruments which similarly require contact with biological samples and the reagents used and thus also involve precise cleaning procedures. However, they are able to accurately transfer nanolitre amounts of liquid. Nevertheless, various physical factors have to be appropriately adjusted and controlled to ensure accurate transfer of volumes as small as ~2 nL. For instance, pin diameter, size of the pin slots/grooves (if applicable) which take up liquid material through capillary action, pin dwell time in an empty destination well, retraction speed and immersion depth (both in the source and destination liquids) are some of the parameters which have to be considered. Furthermore, pin tool instruments are not practical when several different liquid volumes have to transferred in a single liquid handling workflow, since this would require exchanging pin arrays (Cleveland and Koutz, 2005). Microfluidic chips are an alternative to large liquid handling workstations and have been used to downscale DNA assembly (Patrick *et al*., 2015; Gach *et al*., 2017; Khilko *et al*., 2018). For instance, Khilko *et al*. demonstrated a sub-microlitre (600 nL) protocol consisting of a 12 DNA parts Gibson Assembly, PCR and enzymatic error correction

of a human influenza virus hemagglutinin gene (339 bp-long), using a commercial digital microfluidics device, while Patrick *et al*. demonstrated a 2 DNA parts nanolitre (490 nL) Golden Gate assembly of a plasmid, using a cheaper 3D printed microfluidics chip (Patrick *et al*., 2015, Khilko *et al*., 2018). Synthetic DNA has also been recently assembled in picolitre volumes. Plesa *et al*. has demonstrated assembly of more than 7,000 genes, using barcoded beads co-localizing individual DNA assembly array-synthesised oligonucleotides in picolitre droplets (Plesa *et al*., 2018). Such microfluidic and water-in-oil emulsion methodologies however require significant expertise and therefore are not yet readily accessible to researchers.

Contactless acoustic liquid handling methodologies are also able to handle nanolitre volumes with high accuracy (Ellson *et al*., 2016). These methods are easy to use and are based on the acoustic droplet ejection phenomenon (ADE), which involves sound waves causing ejection of discrete liquid droplets. ADE phenomenon has been, e.g., exploited by the inkjet printing industry and since the early 2000s has been progressively employed by life sciences, mainly by compound library screening procedures (Dawes *et al*., 2016).

This project therefore applied acoustic liquid transfer technologies to miniaturise DNA assembly. Using an acoustic liquid transfer instrument, dispensing 2.5 nL droplets with high frequency, two popular DNA fabrication methods, the Gibson and the Golden Gate methods, as well as a common DNA validation method, endpoint PCR, were miniaturised down to 250, 50 and 250 nL, respectively. Reaction miniaturisation efforts led to up to a ~200-fold reduction of reagent costs and thus paved the way towards more cost-efficient high-throughput DNA construction. This chapter further discusses how this ultimate goal can be achieved, taking into account failure rates of the nanolitre DNA assemblies, and investigates the “sweet spot” of reaction miniaturisation.

3.3 Results

To test reaction miniaturisation limits of the Gibson and Golden Gate DNA assembly reactions their volumes were downscaled to 50 nL, 250 nL, 500 nL and 1,000 nL, using the Echo 550 acoustic dispenser. This corresponded to 20-400- and 7.5-150-fold miniaturisation of the Gibson and Golden Gate methods, respectively, as compared with the reaction volumes of the positive control nucleic acid assemblies (**Fig. 3.2 and 3.4**).

3.3.1 DNA Assembly Methodology

As miniaturisation of DNA assembly reactions had not been previously demonstrated using the acoustic droplet ejection methodology, a minimal number of the DNA assembly fragments (2 fragments) was first tested. However, the DNA fragment design differed for the Gibson and Golden Gate DNA assemblies. For the Gibson assembly experiments, the two fragment junctions were placed inside a red fluorescent protein (RFP) reading frame and a beta-lactamase open reading frame (*bla*), conferring resistance to ampicillin (**Fig. 3.1**). Upon bacterial transformation, correct DNA assemblies therefore resulted in red *E. coli* colonies, capable of growth on a solid medium which contained ampicillin. Cells propagating incorrect DNA assembly products were prevented from producing red pigmentation and surviving in the presence of ampicillin. The Golden Gate experiments, on the other hand, aimed at cloning a DNA insert encoding a budding yeast promoter into a plasmid vector, likewise encoding a red fluorescent protein. To do that, the promoter and RFP DNA sequences were flanked by inward and outward facing Type IIS BsaI endonuclease recognition sites, respectively, generating compatible sticky ends (**Fig. 3.3**). Correct assemblies hence led to white colonies and incorrect ones to red colonies.

Two types of metrics were computed to evaluate success of the miniaturisation experiments. For the Gibson assembly, the total number of (red) colony forming units (CFUs) was counted. Golden Gate DNA assemblies, in contrast with the Gibson experiments, yielded correct (white) and incorrect (red) colony forming units. The percentage of correct clones was termed assembly efficiency. Low efficiencies

mean that more colonies have to be tested to find correct DNA constructs, in scenarios without an appropriate, e.g., colorimetric, marker. To stay consistent with the Gibson protocol metrics, the number of correct (white) colonies was also counted.

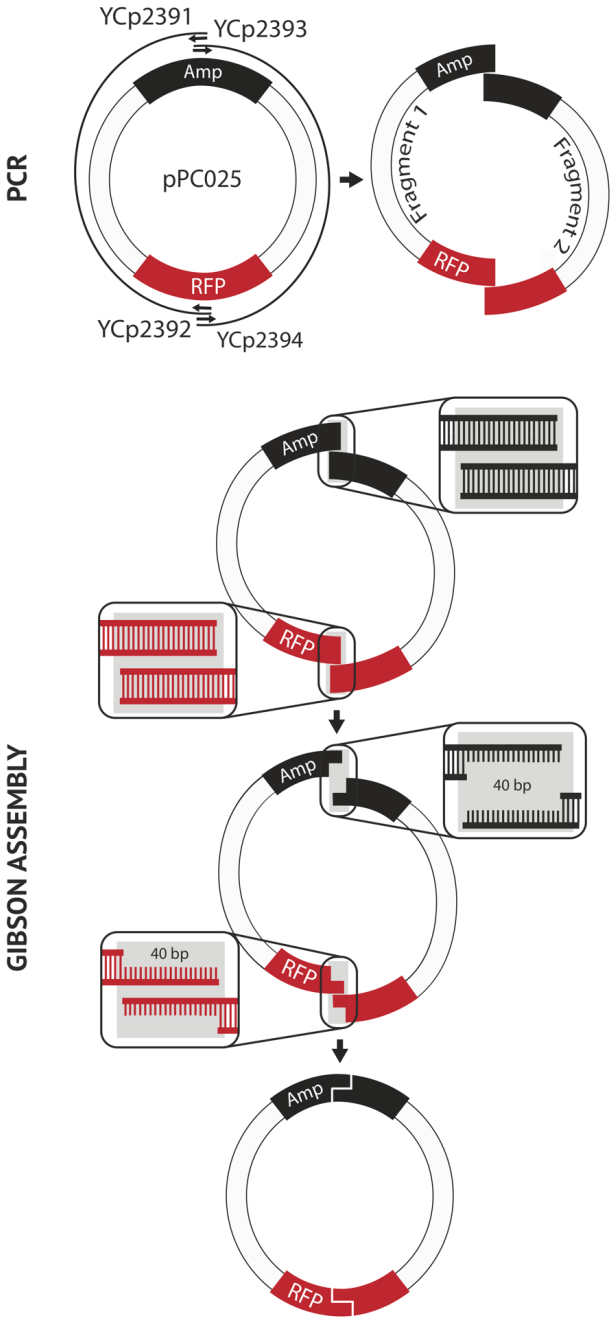


Figure 3.1 Gibson Assembly Methodology

DNA assembly junctions were located inside the RFP and *bla* marker reading frames to facilitate identification of bacteria propagating correct DNA constructs.

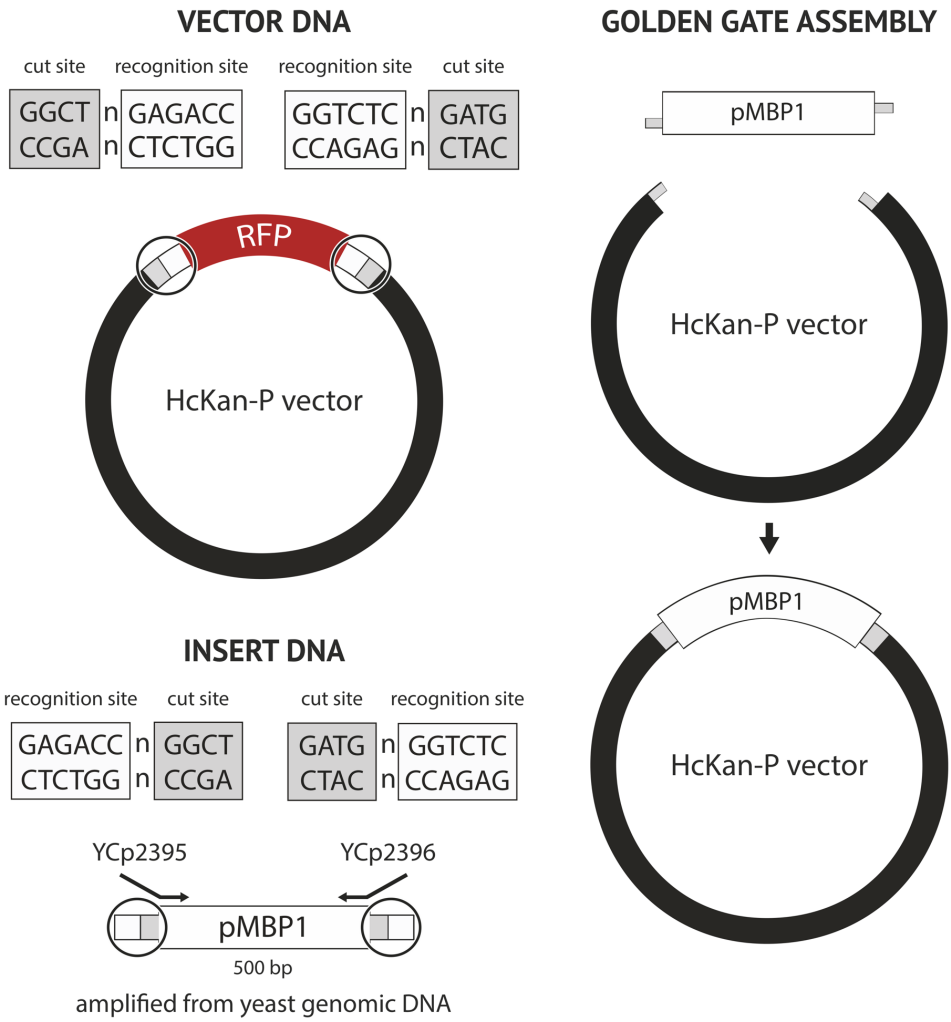


Figure 3.2 Golden Gate Assembly Methodology

The Golden Gate experiments aimed at cloning a budding yeast promoter into a plasmid DNA vector. The vector DNA harboured a DNA assembly slot, which encoded an RFP reading frame flanked by outward facing BsaI recognition sites and standardised DNA sticky ends. Upon successful DNA assembly, the RFP marker is replaced by the promoter DNA, which, likewise, was flanked by the restriction digestion sequences. These sequences were added through PCR and were compatible with the vector DNA flanking regions.

3.3.2 Results of the Nanolitre DNA Assembly

Using the proposed miniaturisation approach, correct DNA assemblies were observed at volumes as low as 50 and 250 nL, which led to proportional 150- and 80-fold reagent cost reductions (**Supp. Tables 2.1.1-3 in Appendix 2.1**) for the Golden Gate and Gibson methods, respectively (**Fig. 3.2 and 3.4**). These results were confirmed by sequencing of the assembly junctions (**Supp. Fig. 2.2.1 in Appendix 2.2**). The number of correct CFUs and the percentage efficiencies (for the Golden Gate miniaturisation experiments) however both decreased as the reactions were becoming smaller (2-fold dilutions of transformation mixes were plated). Moreover, the reliability of DNA assembly has also decreased and was particularly low for the smallest DNA assembly reactions, as indicated by considerably low mean CFU counts and large standard errors (exceeding the means), demonstrating that 50 nL Golden Gate and 250 nL Gibson DNA assemblies are not readily feasible (**Fig. 3.2 and 3.4**). In order to better estimate the success rate of the smallest DNA assembly reactions, and thus better dissect their feasibility, more biological replicate experiments should be set up in the future. Results presented in this manuscript represent data derived from three biological replicates.

The Gibson method uses a DNA polymerase which fills DNA gaps at the junction regions. Therefore, this DNA fabrication method can result in mutations. The proposed Golden Gate assembly methodology, on the other hand, posed a risk of obtaining false positive (white) colonies due to an illegitimate intracellular ligation of linear DNA under strong selective pressure (Shimizu *et al*., 1997). However, nucleic acid mutations and spurious DNA constructs were not observed in the plasmid DNA samples tested (1-2 colonies tested per successful DNA assembly experiment). More colonies however should have been tested to be able to conclude with confidence that these are not an issue at sub-microlitre reaction volumes.

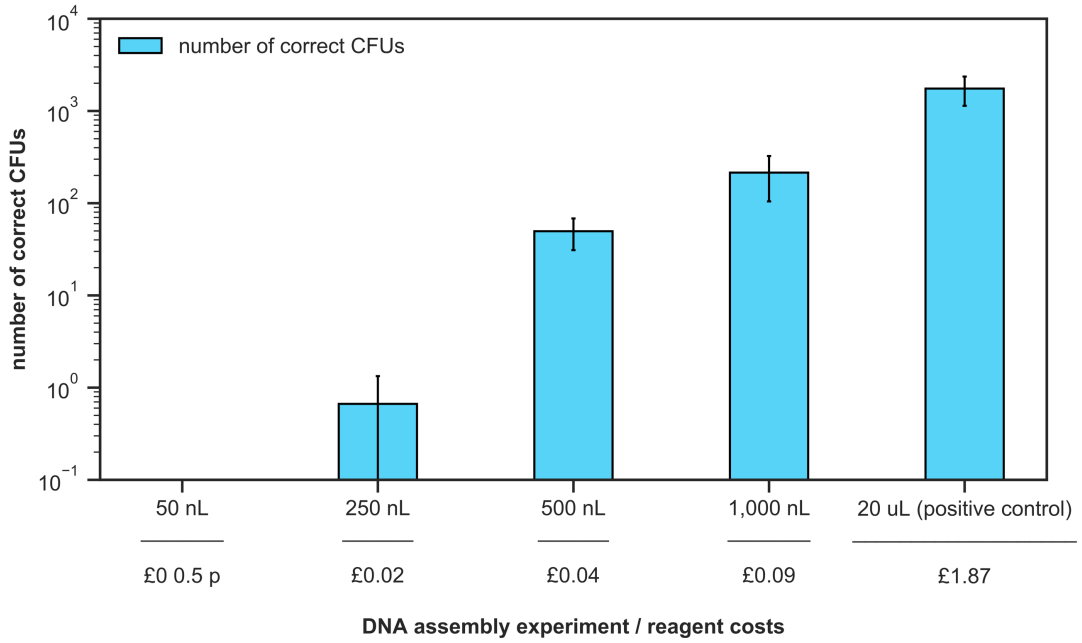


Figure 3.3 Results of Nanolitre Gibson Assemblies

(**A**) DNA assembly is feasible in volumes as low as 250 nL. The number of colony forming units propagating correct plasmid DNA decreases in proportion to the decreasing reaction volumes. Three biological replicates were set up in each nanolitre reaction experiment and the presented error bars represent standard error of the mean. The corresponding reagent costs per single reaction are provided. (**B**) Correct DNA assemblies led to reconstitution of the RFP reading frame and thus red colonies. No colonies were observed in the negative control experiments, missing either of the DNA assembly fragments.

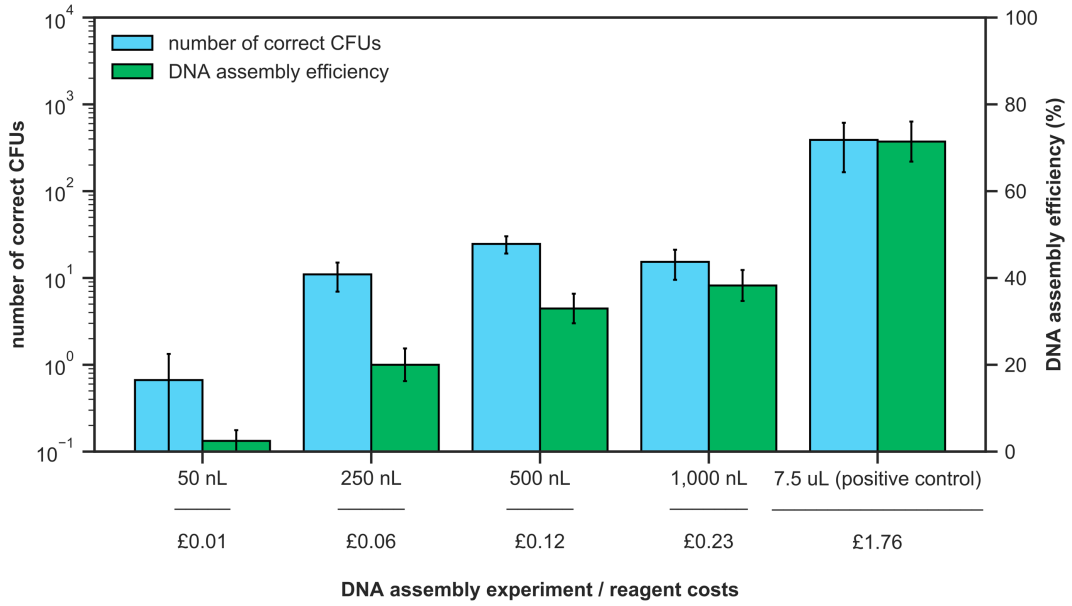


Figure 3.4 Results of Nanolitre Golden Gate Assemblies

(**A**) DNA assembly is feasible in volumes as low as 50 nL. Two measures of DNA assembly success are provided, i.e., number of colony forming units (CFUs) bearing correct DNA constructs and percentage of white (correct) CFUs in the colony population. Both are generally decreasing in proportion to the decreasing reaction volumes. Three biological replicates were set up in each nanolitre DNA assembly experiment and the presented error bars represent standard error of the mean. The corresponding reagent cost prices per single reaction are provided. (**B**) Correct DNA assemblies led to replacement of the RFP reading frame and thus gave rise to white colonies. Incorrect DNA assemblies, on the other hand, retained the RFP marker and thus led to red colonies.

3.3.3 Nanolitre PCR Reactions

In addition to miniaturising DNA construction, acoustic dispensing was also applied to downscale endpoint polymerase chain reactions (PCRs). Endpoint PCRs, along with other methods, such as restriction digestion, are typical assays used to test correctness of the DNA fabrication products, and are based on amplification of the DNA fragment junction regions. Acoustic liquid handling was therefore used to miniaturise these reactions down to 50, 250, 500, 750 and 1,000 nL. This work led to a successful and reproducible amplification of a ~1.4 kb DNA fragment at volumes as low as 250 nL (**Fig. 3.5**). All four biological replicates exhibited amplicons of the correct size. For all PCR experiments, including the positive control, the amount of amplicon DNA also decreased in proportion to the dropping reaction sizes (total nanolitre reaction volumes were loaded into the agarose gel).

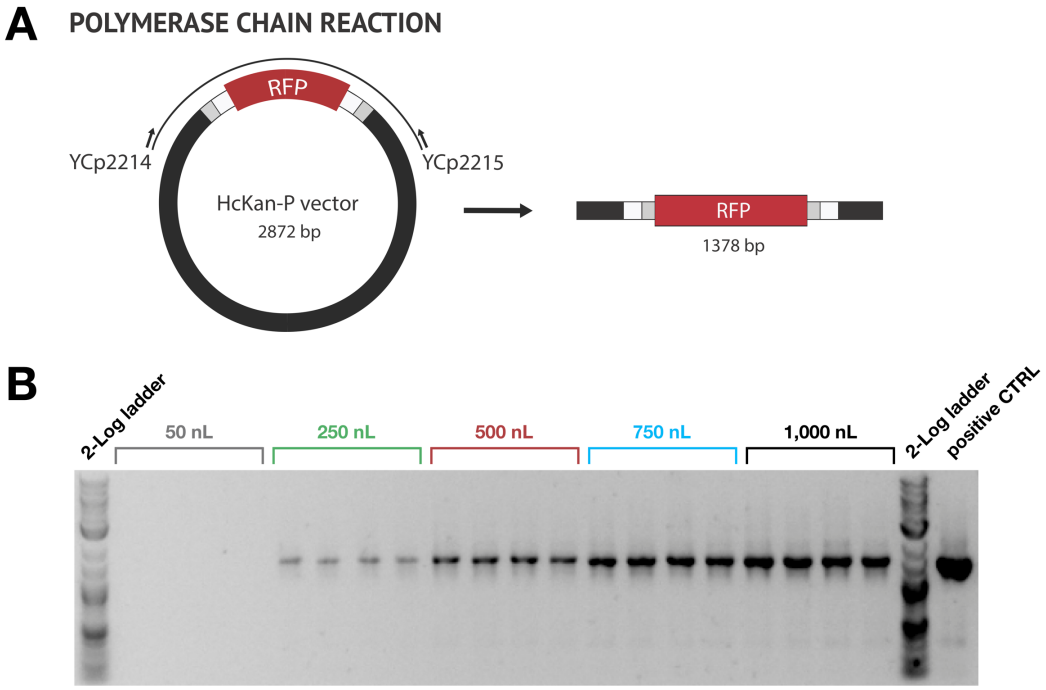


Figure 3.5 Endpoint PCR Miniaturisation Experiments

(**A**) Plasmid DNA PCR template and the amplicon DNA; (**B**) Results of the PCR miniaturisation experiments: DNA amplification was successful and reproducible at volumes as low as 250 nL. The amount of amplicon DNA decreased in proportion to the decreasing reaction volumes.

3.4 Discussion

In conclusion, the first successful application of the acoustic dispensing technology to downscale DNA fabrication beyond the microlitre threshold was demonstrated. The presented data shows that DNA assemblies can be successful in reaction volumes as low as 50 nL and can be as cheap as 1 British pence.

Despite the low reagent costs per single DNA assembly the overall reagent expenses can however turn higher as the DNA construction failure rate at the lowest reaction scales is significant (i.e., one successful DNA construction experiment out of three biological replicates). As the number of experiment repetitions upon a DNA fabrication failure is a geometric random variable, the expected overall DNA fabrication costs can be computed by multiplying the expected number of repetitions (i.e., 1 ÷ likelihood of success) and the reagent costs per single DNA assembly. Therefore, for a 50 nL Golden Gate DNA assembly reaction the expected reagent costs are 3 × 1 p, while for a 250 nL Gibson assembly they equal to 3 × 2 p (**Supp. Calculation in Appendix 2.3**). It is however important to emphasise that the expected total costs of DNA fabrication depend not only on the costs of reagents, but also on the costs of labware used as well as on the costs of personnel involved in setting up the corresponding automation procedures.

Volumes of 50 and 250 nL were the smallest functional Golden Gate and Gibson DNA construction scales. These reaction size boundaries could reflect a decreasing resolution of transferable liquid volumes. As the reaction components were dispensed in volumes which constituted multiples of 2.5 nL droplets, setting up equimolar mixtures of the DNA assembly fragments for the lowest reaction sizes tested was problematic. For instance, a 50 nL Golden Gate reaction was set up with 20 2.5 nL droplets. The volumes required to achieve equimolar amounts of insert and plasmid DNA were 30 and 3 nL, respectively. However, since 3 nL is not a multiple of 2.5 nL, the plasmid DNA volume had to be round down to 2.5 nL (1 droplet), which was a 20% smaller volume than the optimal amount.

Golden Gate reaction miniaturisation experiments relied on screening bacterial colonies using a colorimetric marker (RFP), which allowed to estimate DNA assembly efficiency. However, future applications of the presented nanolitre DNA fabrication methodology could use a *ccdB* gene instead. Failure to replace this gene by the insert DNA, through successful DNA restriction digestion and ligation, would lead to expression of a toxic CcdB protein and death of cells propagating an undigested vector (Wang *et al*., 2014).

As pointed out earlier, *E. coli* transfection with linear plasmid DNA can lead to growth of false positive bacterial colonies (Golden Gate experiments) (Shimizu *et al*., 1997). In the future, this issue could be mitigated by an additional treatment of the finished DNA assembly reactions with an exonuclease, which would digest the unligated plasmid (e.g., Plasmid-Safe ATP-dependent DNase from Epicentre). Such an enzymatic treatment would yet require an additional incubation period of the nanolitre reactions, which could lead to increased evaporation. Therefore, optimising duration of the Golden Gate thermal cycling protocol might be needed to accommodate this extra DNA digestion step.

PCR reaction miniaturisation work has also led to downscaling endpoint PCR reactions, down to 250 nL. It is however uncertain whether DNA amplification has failed below 250 nL (at a 50 nL reaction volume) or the detection method used failed to detect it. According to the manufacturer, the fluorescent DNA-intercalating dye used (SYBR Safe DNA Gel Stain from Thermo Fisher Scientific) has a detection limit of 500 pg per band for DNA fragments larger than 200 bp (when viewed on a 300 nm transilluminator), which corresponds to 336,161,661 molecules of 1,378 bp amplicon DNA ($MW_{amplicon\ DNA} = 1{,}378\ bp\ \times 650\frac{g}{mol} = 895{,}700\frac{g}{mol}$, where $650\frac{g}{mol}$ is an average MW of 1 base pair; 500 pg thus corresponds to 0.558 fmol of amplicon DNA or $0.558\ \times\ 10^{-15}\ mol\ \times\ 6.022 \times 10^{23}\ \frac{molecules}{mol} = 336{,}161{,}661$ molecules of amplicon DNA). The mass of template DNA used in the 50 nL reaction was 0.6 ng, which corresponds to 193,550,460 molecules of 2,872 bp template DNA (the number of template DNA molecules calculated as above). The ratio of the

detection threshold DNA amplicon molecule count to the starting amount of the template molecules used is thus 1.74. Such amplicon molecule count can be obtained in a single 86.8% efficient PCR cycle (the amount of DNA doubles at a 100% PCR efficiency, and therefore $Single\ PCR\ Cycle\ Efficiency = \frac{100\% \times 1.74}{2} = 86.8\%$). It is therefore unlikely that the lack of the amplification signal (visible DNA gel band) was caused by an insensitive detection method, given 32 PCR cycles performed in this study. In the future, a 500 pg amplicon DNA control could be run on the agarose gel, serving as a detection limit control further facilitating the following PCR efficiency analysis.

Chapter 4
CRISPR Deletion Library for Mapping Genotype-Phenotype Relationships

4.1 Work Contributions

Work presented in this chapter was a result of collaboration between the Cai Lab (The University of Manchester, UK), Tyers Lab (The University of Montreal, Canada) and the Tollervey Lab (The University of Edinburgh, UK).

Under supervision of Prof. Patrick Yizhi Cai and Prof. Michael Tyers, I outlined design of the deletion library, performed proof-of-concept experiments, cloned the library, conducted genome-scale screens in budding yeast, prepared NGS sequencing libraries, ran NGS experiments and analysed their results computationally. Dr. Jasmin Coulombe-Huntington (Tyers Lab) designed nucleic acid sequences of the genome editing constructs and performed additional computational analyses (not presented in this thesis). Dr. Ghada Ghazal (Tyers Lab) helped in establishing yeast screen protocols and performed additional genome-scale yeast experiments (not presented in this thesis). Dr. Tomasz Turowski (Tollervey Lab) filtered long non-coding RNA data, proposed the alternative NGS sequencing library preparation protocol and supervised setup of the NGS experiments. Dr. Thierry Bertomeu (Tyers Lab) designed genome-barcoding sequences.

4.2 Introduction

Saccharomyces cerevisiae is one of the main eukaryotic organisms used by synthetic biologists, with applications ranging from industrial biotechnology to synthetic genomics (Runguphan and Keasling, 2014; L. Wang *et al*., 2018). To date, yeast metabolism has been harnessed to cost-effectively produce an antimalarial drug precursor, artemisinic acid, and opioid compounds, while its model eukaryote status as well as its ability to efficiently recombine homologous DNA strands have encouraged its application as a platform to build the first "designer" set of eukaryotic chromosomes (i.e., the Sc2.0 Project) (Ro *et al*., 2006; Richardson *et al*., 2017). Therefore, efficient genome engineering tools continue to be required to facilitate understanding, debugging and evolution of custom phenotypes (Stovicek, Holkenbrink and Borodina, 2017).

Eukaryotic organisms, including *S. cerevisiae*, feature intricate genetic interaction networks which govern their phenotypic plasticity (Kuzmin *et al*., 2016). For years, genetic underpinnings of various phenotypes have been investigated through gene deletion experiments. Synthetic Genetic Array (SGA) methodology is one of the most established methodologies for conducting such studies genome-wide. SGA experiments however suffer from an extensive usage of auxotrophic and drug-resistance markers, which are known to lead to unintentional phenotypic perturbations and are limited to particular *S. cerevisiae* strain genetic backgrounds (S288C) (Kuzmin *et al*., 2016; Leng and Song, 2016).

This chapter describes development of a gene deletion library addressing these limitations. The library is suitable for genome editing of a broad range of wild-type *S. cerevisiae* strains as well as the Sc2.0 project strains and harnesses a novel genome editing approach – CRISPR. CRISPR leverages activity of an RNA-guided endonuclease, which introduces programmable double-stranded DNA breaks (DSBs). Unrepaired DSBs are lethal and necessitate repair by generally either of the two major DSB repair pathways – NHEJ (non-homologous end joining) or HDR (homology-directed repair). As opposed to NHEJ, which re-ligates broken DNA ends

without a reference DNA template (often leading to small indel mutations at the DSB sites), HDR uses a homologous DNA template ("donor DNA") to repair DSBs and thus allows introduction of custom changes at the DSB loci (e.g., premature termination codons) without the need for genetic marker selection (i.e., unrepaired DSBs lead to cell death). Moreover, HDR is a preferred *S. cerevisiae* DSB repair mechanism (Jasin and Rothstein, 2013). The described gene deletion library approach therefore leveraged mutagenic HDR donors to knockout *S. cerevisiae* genes genome-wide. The approach was first tested via deletion of example essential and non-essential genes, then assessed for correct identification of characterised *S. cerevisiae* essential genes genome-wide.

The library comprises 27,000 HDR donor – crRNA gene deletion DNA cassettes, constructed by a commercial array oligonucleotide synthesis. The HDR donor – crRNA pairs target every gene in the ~6,000 genes 12 Mb S. *cerevisiae* genome approximately 4-times (to maximise chances of successful gene knockout), including a set of appropriate positive and negative controls, as well as various characterised and uncharacterised RNA transcription units (TUs), e.g., the small nucleolar RNA (snoRNA), transfer RNA (tRNA) and long non-coding RNA (lncRNA) TUs, which are known to have significant impact on cellular phenotypes via guiding chemical modification of other RNAs, actively participate in the process of translation and regulate expression of other genes, respectively (Watkins and Bohnsack, 2012; Raina and Ibba, 2014; Wu, Yang and Chen, 2017). The target Cas protein cleavage sites were biased towards close proximity to transcription start sites (TSSs) by design to ensure efficient gene knockouts and prevent production of truncated proteins, which premature termination codon-bearing mRNAs escaped nonsense-mediated mRNA decay pathway surveillance, and designed to preserve the Sc2.0 genome synthetic DNA "watermark" sequences (PCRTags) as well as take into account its reduced stop codon set (Dymond *et al*., 2011).

Construction of a genome-wide gene deletion library necessitated a physical link between the HDR donor and the crRNA sequences, so that corresponding HDR donors and crRNAs could co-localise in a single cell. This was done by encoding

both on a single plasmid (within a "gene deletion cassette") and therefore using plasmid DNA as the HDR donor. While donor DNA does not require genetic expression to serve its function, it had to be included inside cRNA's RNA polymerase III transcription unit (in contrast with the analogous prokaryotic CREATE system). Eukaryotic promoters are longer than their prokaryotic counterparts and therefore their inclusion within the synthetic oligonucleotide gene deletion cassette constructs was not feasible (it would lead to exceeding the commercial synthesis length limit) (Kristiansson *et al*., 2009; Garst *et al*., 2016; Clore, 2018). Feasibility of this approach was supported by previous publications, which demonstrated that various nucleic acid sequences, including HDR donors and ribozymes, can be placed upstream of crRNAs, within their TUs, without interfering with their post-transcriptional function - and with some likely serving protective purposes, due to formation of secondary structures protecting crRNAs from nuclear 5' RNA exonucleases (e.g., Rat1) (Houseley and Tollervey, 2009; Ryan *et al*., 2014). Furthermore, previous work showed that plasmid-encoded HDR donors can be used to efficiently repair double-stranded DNA breaks (DSBs), provided that there are sufficient intracellular plasmid copies (Bao *et al*., 2015). However, prior to the library construction effort described by this manuscript, none of the previous studies have tested these findings on a genome-wide scale.

To efficiently delete genes and monitor behavior of the resulting knockout strains, HDR donors harboured internal "barcode" cassettes, integration of which at the Cas9 cleavage sites was designed to elicit formation of premature stop codons, as well as partial protospacer and complete protospacer-adjacent motif deletions, to prohibit the Cas9 enzyme from further endonucleolytic cleavages which tend to compromise strain fitness and are particularly problematic during non-mutagenic HDR and NHEJ (Cagney *et al*., 2006). Genomic barcodes allowed tracking, using NGS, abundance of individual single-gene mutants in several heterogenous yeast mutant pools in parallel as well as ensuring a one-to-one correspondence of the genomic barcode and the underlying mutant copy number (which cannot be guaranteed with, e.g., plasmid-encoded barcode sequences). Moreover, a second alternative method of sequencing barcode DNA was proposed and tested, relying on single-stranded DNA

circularisation, and aimed at facilitating unambiguous confirmation of successful CRISPR editing events. It has been shown that CRISPR can result in off-target genome editing (Tsai *et al.*, 2014). Therefore, an ideal deletion library sequencing methodology should be able to capture DNA sequences neighbouring barcode DNA, in order to confirm its correct insertion. Inability to differentiate between off-target and correct insertions can lead to experiment noise. The alternative barcode DNA sequencing method tried to address this issue.

4.3 Results

4.3.1 Deletion Library Design

27,000 deletion cassettes were encoded on separate plasmids, which harboured all components of the CRISPR machinery, i.e., the homology donor – crRNA transcription unit and the Cas9 endonuclease expression module. The plasmids also encoded a two micron origin of replication, to ensure their high-copy propagation and thus sufficient expression of the CRISPR system and abundance of the donor DNA (**Fig. 4.1**). Human codon-optimised *Streptococcus pyogenes* Cas9 endonuclease was used under control of a strong constitutive *TEF1* promoter. The homology donor - crRNA module was put under control of the RNA polymerase III transcriptional machinery, employing the often used *SNR52* promoter and *SUP4* terminator (Stovicek, Holkenbrink and Borodina, 2017). Polymerase III expression, as opposed to polymerase II, does not incorporate any undesirable RNA modifications, e.g., 3' polyadenylation, which can interfere with the guide RNA - Cas9 ribonucleoprotein complex formation (Bentley, 2014; Turowski and Tollervey, 2016). Polymerase III transcription can however be terminated by the presence of poly(T) stretches (Nielsen, Yuzenkova and Zenkin, 2013). The design parameters thus avoided more than 4 consecutive thymine residues in the genome targeting cassettes.

The HDR donor portions of the gene deletion cassettes comprised two flanking 45 bp homologous recombination arms, corresponding to the downstream crRNA sequence. The arms harboured an 8 bp deletion, deleting the last 3 bp of the

protospacer (tampering with the last 12 bp of the protospacer sequence negatively impacts the RNA-guided recognition (DiCarlo *et al*., 2013; Jiang *et al*., 2013)) and the protospacer adjacent motif (PAM) to, as mentioned earlier, lead to efficient knockout of a given gene and prevent further Cas9 recognition and the burden of resulting DNA cleavage.

The barcoding module was placed in between the 5' and 3' HDR donor sequences. It included a 16-21 base-pair barcode, identifying the target genomic locus, and two flanking Illumina NGS sequencing-compatible primer binding sites. Insertion of this module at a given cut site should mark the editing event and allow accurate tracking of the abundance of individual deletion mutants in complex mutant pools. As pointed out previously, the barcode sequences were also designed to trigger premature RNA translation termination by introduction of premature stop codons. Moreover, their varying 16-21 bp size was intentional and was designed to facilitate optical readout of the Illumina sequencing machine used by avoiding regions in which all sequencing clusters have an identical sequence at the same position (Mitra *et al*., 2015).

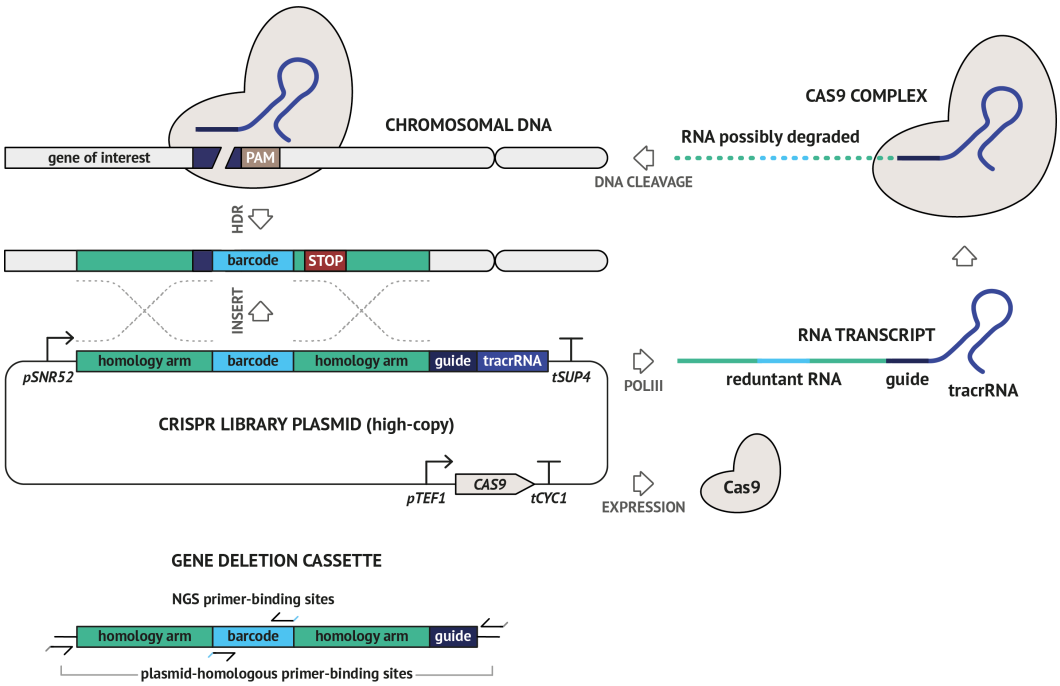


Figure 4.1 Design of the Plasmid Library and its Mechanism of Function

CRISPR library plasmids harboured all components of the CRISPR machinery as well as an homology-directed repair (HDR) donor sequence under control of the sgRNA (guide – tracrRNA) expression unit. The homology arm, barcode and guide DNA sequences constitute a gene deletion cassette targeting a given gene of interest. Every gene deletion cassette harbours two types of primer-binding sites, (1) plasmid-homologous primer-binding sites, allowing cassette amplification and subsequent cloning into the CRISPR library plasmid, and (2) NGS primer-binding sites, allowing amplification and sequencing of the plasmid-encoded and genome-integrated barcode sequences. Upon transformation of a given *S. cerevisiae* strain of interest, a single plasmid DNA molecule is introduced inside a cell and multiplies inside of it to a high-copy number. The Cas9 protein is expressed and the HDR donor - sgRNA sequence is transcribed by RNA polymerase III (Pol III). A Cas9::sgRNA ribonucleoprotein complex is formed and the redundant part of the RNA transcript (encoding the HDR donor sequence) is possibly degraded by host nucleases. The complex localises to the genomic site of interest via complementarity between the guide RNA sequence and the corresponding genomic DNA protospacer sequence, located downstream the PAM (protospacer-adjacent motif) sequence required for the complex recognition. Cas9 endonuclease introduces a DSB 3 bp downstream the PAM sequence. Plasmid DNA is used to repair the break. HDR repair leads to a partial deletion of the protospacer sequence and a complete deletion of the PAM sequence (both preventing further complex recognition and cleavage), as well as insertion of the barcode sequence at the site of the DSB, which leads to a nonsense mutation (premature stop codon generation) in the gene of interest. Location of the nonsense mutation is biased towards the TSS of the gene of interest to avoid production of aberrant truncated protein species due to the escape of the corresponding truncated mRNA species from the nonsense mediated mRNA decay pathway surveillance.

4.3.2 Deletion Library Composition

The plasmid pool included 24,854 DNA cassettes targeting 6,499 protein-coding genes. Each of the protein-coding genes was therefore targeted with ~4 different homology HDR donor - crRNA cassettes, maximizing the chance of obtaining at least one successful deletion mutant. The choice of cleavage loci was based on high on-target crRNA scores as well as their proximity to the start codon (median distance of 100 bp) (Bertomeu *et al*., 2018). The latter prerequisite was imposed to ensure fast knockouts and avoid possible aggregation of truncated and misfolded peptides (negatively impacting cell viability), in case premature stop codons were missed by the nonsense-mediated decay mRNA surveillance machinery (Decourty *et al*., 2014). Location of the cleavage loci was not biased towards sequences encoding characterised protein functional domains, which posed a danger of introducing measurement noise in the NGS assessment of yeast viability, caused by the NHEJ DSB repair pathway. NHEJ is a competing and less efficient DSB repair pathway in the budding *S. cerevisiae* yeast, which tends to introduce small DNA mutations at the DSB sites (Bétermier, Bertrand and Lopez, 2014). Thus, NHEJ-mediated mutations distal from DNA sequence regions encoding functional protein domains and, additionally, not disrupting the corresponding open reading frames are unlikely to cause gene knockouts. While this scenario does not interfere with NGS experiments based on amplification and sequencing of genomic DNA only, i.e., the barcoding cassette cannot be genome-integrated once the unwanted mutagenic NHEJ-mediated edit is made, it could prove problematic if a mixture of plasmid and genomic DNA, both bearing barcoding cassettes, was to be used as an NGS template (or if unwanted traces of plasmid DNA were present during the barcode amplification reaction), leading to possible false conclusions regarding yeast viability post-gene knockout. Bias of the cleavage sites towards functional *S. cerevisiae* protein domains can however be a complex task and does not guarantee efficient knockouts either, i.e., not all of the domains are fully characterised and some of their DNA sequence regions are additionally located far away from the TSSs, posing the aforementioned truncated protein production issue (Decourty *et al*., 2014). Therefore, it was decided that targeting of the functional domain-encoding DNA sequences will not be pursued

and that the NGS data analysis should be cautious of the possible measurement noise instead.

As pointed out earlier, the library targeted non-protein-coding genes as well. Long non-coding RNA (lncRNA), small nucleolar RNA (snoRNA) and transfer RNA (tRNA) genes were targeted. It was hypothesised that a larger (as compared with the protein-coding gene deletion cassettes) deletion of 50 bp anywhere in these genes would be sufficient to introduce a null mutation, in particular in the case of snoRNAs and tRNAs in which secondary structure is highly conserved (Watkins and Bohnsack, 2012; Raina and Ibba, 2014). To validate this strategy, a set of 100 essential gene 50 bp deletion control cassettes was included. However, it was anticipated that successful knockouts of lncRNAs might still be challenging. LncRNAs are ubiquitous cellular moieties, yet their function and modes of action are still largely unknown (Wu, Yang and Chen, 2017). Moreover, they exhibit lower sequence conservation, which makes it difficult to map nucleic acid regions susceptible to mutagenesis (Rivas, Clements and Eddy, 2017). To maximise chances of obtaining lncRNA null mutants, a list of putative lncRNAs was filtered in search for the most highly expressed RNAs which do not coincide with the neighbouring protein-coding reading frames (to circumvent any knockout effect misinterpretations) and resulted in a set of 200 gene targets. As with the protein-coding deletion cassettes, these, together with the snoRNA and tRNA sets, were each targeted by ~4 crRNAs.

The remaining portion of the pool was populated with different types of control constructs. Cassettes were designed with non-targeting random homology arms and/or crRNAs, which served as negative CRISPR and homologous recombination controls to investigate Cas9 cleavage in the absence of a repair template and homologous recombination in the absence of a dsDNA cut, respectively. Aside from these, 684 DNA cassettes targeted non-conserved non-coding genomic loci (negative DNA function knockout controls) and 182 cassettes were targeted at uncharacterised and conserved non-coding loci, which were considered promising novel function discovery candidates (**Supp. Table 3.1.3 in Appendix 3.1**).

4.3.3 Proof-of-concept Single-gene Deletion Experiments

Prior to conducting the genome-wide deletion experiments, the design was validated using example single-gene knockout experiments. In these proof-of-concept studies, deletions of two non-essential budding yeast genes (*ADE2* and *CAN1*) and one essential gene (*SEN34*) were investigated. *ADE2* is a protein-coding gene encoding a phosphoribosylaminoimidazole carboxylase, an enzyme catalysing a step in the *de novo* purine nucleotide biosynthesis pathway. Knockout of this gene leads to intracellular accumulation of a red pigment when insufficient amounts of adenine are present in the growth medium (**Fig. 4.2 B**). Therefore, successful deletion of the *ADE2* gene can be easily identified by looking at the yeast colony colour (Gedvilaite and Sasnauskas, 1994). Successful *CAN1* gene knockouts are also straightforward to identify. The *CAN1* gene encodes an arginine permease, involved in transmembrane transport of basic amino acids, and its knockout confers resistance to canavanine, a non-proteinogenic amino acid, which is a toxic analogue of arginine. *CAN1* mutants can thus be easily detected as they are able to grow in the presence of canavanine (Ahmad and Bussey, 1986). To test the capability of the CRISPR system to knock out essential genes, an HDR donor – crRNA cassette was constructed, which targeted the *SEN34* gene. The *SEN34* gene is an essential gene, which encodes a subunit of a tRNA splicing endonuclease (i.e., the SEN complex). However, its function has also been associated with, e.g., ribosomal RNA processing (Dhungel and Hopper, 2012).

At first, single-gene deletion cassettes did not harbour the barcoding sequences. These were incorporated in subsequent experiments, once efficient knockouts using the initial design were achieved (**Fig. 4.4 A**). 81 and 82% knockout efficiencies were achieved in the *ADE2* and *CAN1*, respectively, and were comparable to the positive control results (with a double-stranded oligo HDR donor), i.e., 82 and 91%, respectively (**Fig. 4.4 B and C**). As anticipated, control experiments, which did not harbour an HDR donor but crRNA and tracrRNA only (i.e., the single guide RNA), resulted in either no yeast colonies (the *CAN1* knockout experiments) or scarce colonies (2 and 3) and much lower gene deletion efficiencies - 3% (the *ADE2*

knockout experiments). In the absence of a mutagenic HDR donor, haploid yeast cells can take advantage of their duplicated genetic material (S and G2 cell cycle phases) and use it as an alternative HDR donor (Jasin and Rothstein, 2013). However, usage of endogenous chromosomal DNA leads to error-free DNA repair and thus recurring Cas9 cleavage events, which reduce cell viability (Cagney *et al.*, 2006). Double-stranded DNA breaks can also be repaired in G0 and G1 phases of the cell cycle, without the need for a homologous DNA donor. In these phases, the non-homologous end joining (NHEJ) pathway constitutes the predominant mode of the DNA break repair (Jasin and Rothstein, 2013). The NHEJ pathway has long been considered as error-prone. However, more recent data indicates that it is also capable of accurate DNA lesion repair, which leads to the aforementioned burdening cleavage cycles (Bétermier, Bertrand and Lopez, 2014). Therefore, the error-free repair mechanisms explain the observed low recovery of yeast cells and inefficient gene knockouts following transfection of control gene deletion constructs (**Fig. 4.3A and 4.4 C**).

Having demonstrated efficient non-essential gene knockouts using non-barcoded gene deletion constructs, the effect of the barcode sequence incorporation was investigated. An *ADE2* gene deletion cassette was constructed, which harboured a 56 bp sequence, including the unique barcode identifier and two flanking next-generation sequencing primer binding sites (**Fig. 4.4 A**). Transformation of the construct, cloned into the CRISPR plasmid vector, did not lead to a significant drop in the knockout efficiency (**Fig. 4.4 C**), as confirmed by an unpaired two-sample *t*-test (*t*-statistic = -0.054, p-value = 0.962). To further prove correct and efficient function of non-barcoded and barcoded CRISPR vectors, Sanger sequencing was used to sequence the corresponding genomic DNA loci. More than 15 samples were sequenced per plasmid vector design. The results validated the observed gene deletion phenotypes and confirmed insertion of the barcode cassettes (**Supp. Fig. 3.2.1 in Appendix 3.2**).

The gene deletion design was tested using an essential gene example. A ~3-4 orders of magnitude drop in the recovered yeast colonies, following transfection of the

CRISPR system, was observed, as compared to one of the non-essential gene examples (**Fig. 4.4 D**). This decrease implied successful knockout of the essential *SEN34* gene. Both the non-essential and essential gene examples exhibited lower CFU counts as compared to the negative control (missing the gene deletion cassette), although, as opposed to the essential gene example, the number of recovered CFUs for the non-essential gene knockout experiments, was much less different from the result obtained in the negative control experiment (**Fig. 4.4 D**). A small decrease in the number of recovered yeast colonies was expected in the non-essential gene knockout experiments. The recovered CFU count in a non-essential gene experiment depends on the speed of mutagenesis (also deleting the Cas9 recognition sequences), as Cas9 cleavages have a negative impact on the cell viability (Cagney *et al.*, 2006). Three yeast colonies were recovered, following transfection of the essential gene deletion construct. To investigate the reason behind their survival, their genomic DNA was extracted and the Cas9 cleavage locus sequenced. The results revealed that two yeast colonies did not have their genomic DNA mutagenised (due to the error-free DNA repair), while the third colony harboured a 3 bp insertion which did not disrupt the *SEN34* reading frame and either of the underlying protein active sites, leading to the observed cell survival (Trotta *et al.*, 1997). This DNA insertion could have been a product of the NHEJ repair pathway, exemplifying the aforementioned danger of measurement noise due to the competing NHEJ DSB repair pathway. Such measurement noise could be minimised by biasing the CRISPR targeting sites towards protein functional domain sites. Some functional domain sites are however distal from transcription start sites and their targeting was therefore compromising the other aforementioned design objective, i.e., minimizing the probability of long pre-maturely terminated transcripts escaping surveillance of the nonsense mediated decay machinery. It was therefore decided that this issue will not be addressed through biased CRISPR target site design. Instead, sequencing exclusively genome-inserted DNA barcodes aimed at avoiding the following experimental noise, which absence was vital for the success of the genome-wide proof of concept experiments (assessing depletion of the essential gene knockout strains).

Sequencing exclusively genomic DNA material can be achieved through counter-selection of the plasmid DNA (encoding the barcode sequences as well). The library plasmids were shuttle vectors, which encoded an auxotrophic *URA3* gene marker (**Fig. 4.1**). It was thus possible to counter-select plasmid DNA by adding 5-fluoroorotic acid (5-FOA) to the yeast growth medium. The *URA3* gene encodes orotidylate decarboxylase, which catalyses 5-FOA decarboxylation to 5-fluorouracil, a toxic metabolite which causes death of *URA3* gene-expressing cells (Boeke, La Croute and Fink, 1984) (**Fig. 4.2**).

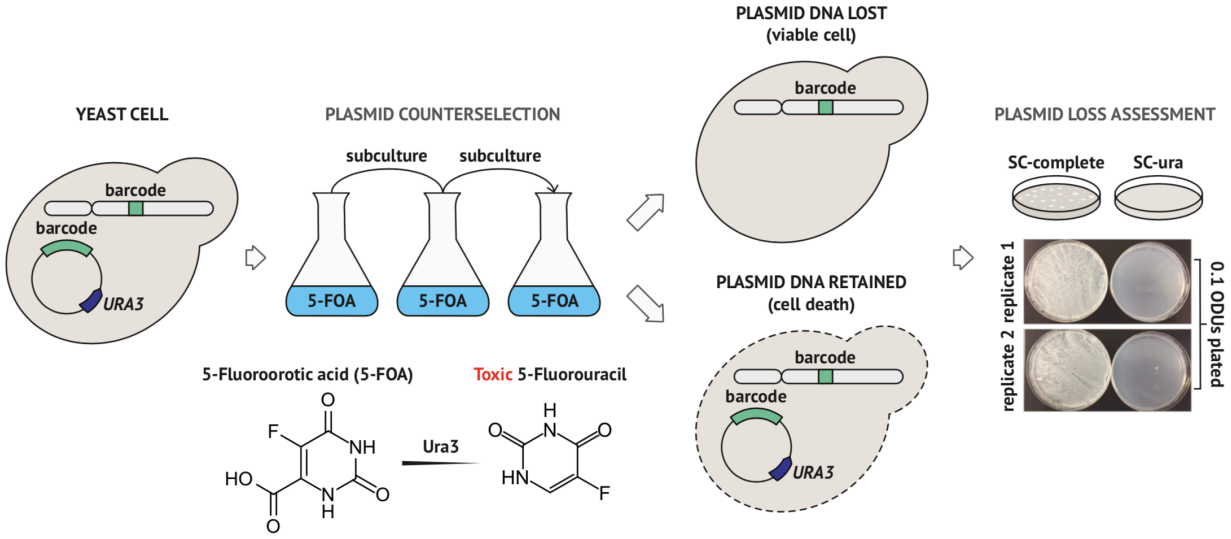


Figure 4.2 Plasmid DNA Counterselection Strategy

CRISPR library plasmids encoded a *URA3* auxotrophic marker, allowing their selection in *ura3* mutant cells and counterselection of plasmid DNA in the presence of 5-Fluoroorotic acid (5-FOA). Yeast cells harbouring a *URA3* gene express an orotidine 5'-phosphate decarboxylase, which decarboxylates 5-FOA (supplemented in the growth medium) giving rise to a toxic product – 5-Fluorouracil. Yeast cells which continue to propagate *URA3*-encoding plasmid DNA should therefore die and dropout from the cellular pool. Successful plasmid counterselection can be confirmed through lack of cellular growth in a uracil-deficient growth medium.

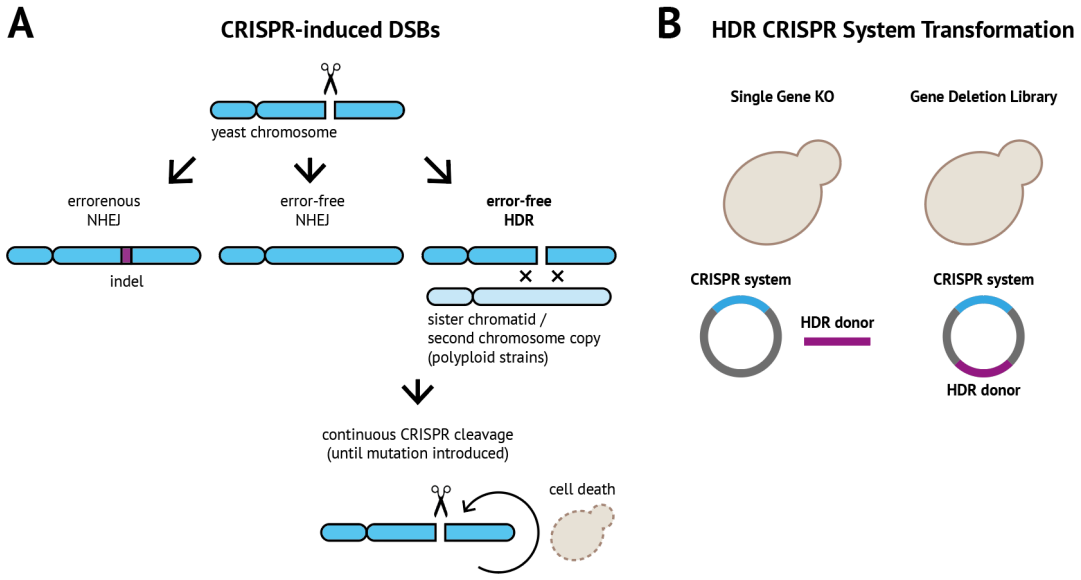


Figure 4.3 Library Design and Proof-of-concept Experiments Rationale

(**A**) Rationale for an HDR-based CRISPR deletion library; Upon a CRISPR-induced double-stranded DNA break (DSB), the cell has to repair the broken chromosomal DNA, otherwise it will die. There are two main DSB repair pathways in *S. cerevisiae*, the non-homologous end joining pathway (NHEJ) and the homology-director repair pathway (HDR). NHEJ can lead to erroneous repair (introduce indel mutations at the site of the break), but can also repair DNA without errors. HDR does not introduce errors, as it is guided by a homologous template (coming from newly replicated DNA or a second chromosome set). HDR (in bold) is a preferred DSB repair pathway in yeast (Jasin and Rothstein, 2013). Therefore, in the absence of a mutagenic homologous template CRISPR systems typically continue introducing DSBs for an extended amount of time, severely impacting cell viability and even leading to cell death (if NHEJ-induced indels are not introduced fast enough - mutating the Cas9 recognition sites). (**B**) Rationale for plasmid DNA-encoded HDR donors; Mutagenic HDR donors are typically introduced inside the cell through co-transformation with the CRISPR system plasmid DNA. This strategy however does not work on a library scale, as HDR cassettes have to be physically connected with the rest of the CRISPR system to appropriately co-localise with the corresponding sgRNAs inside single yeast cells. Proof-of-concept single-gene deletion experiments were performed (**Fig. 4.4**) in order to test whether such physical connection negatively impacts gene deletion efficiency.

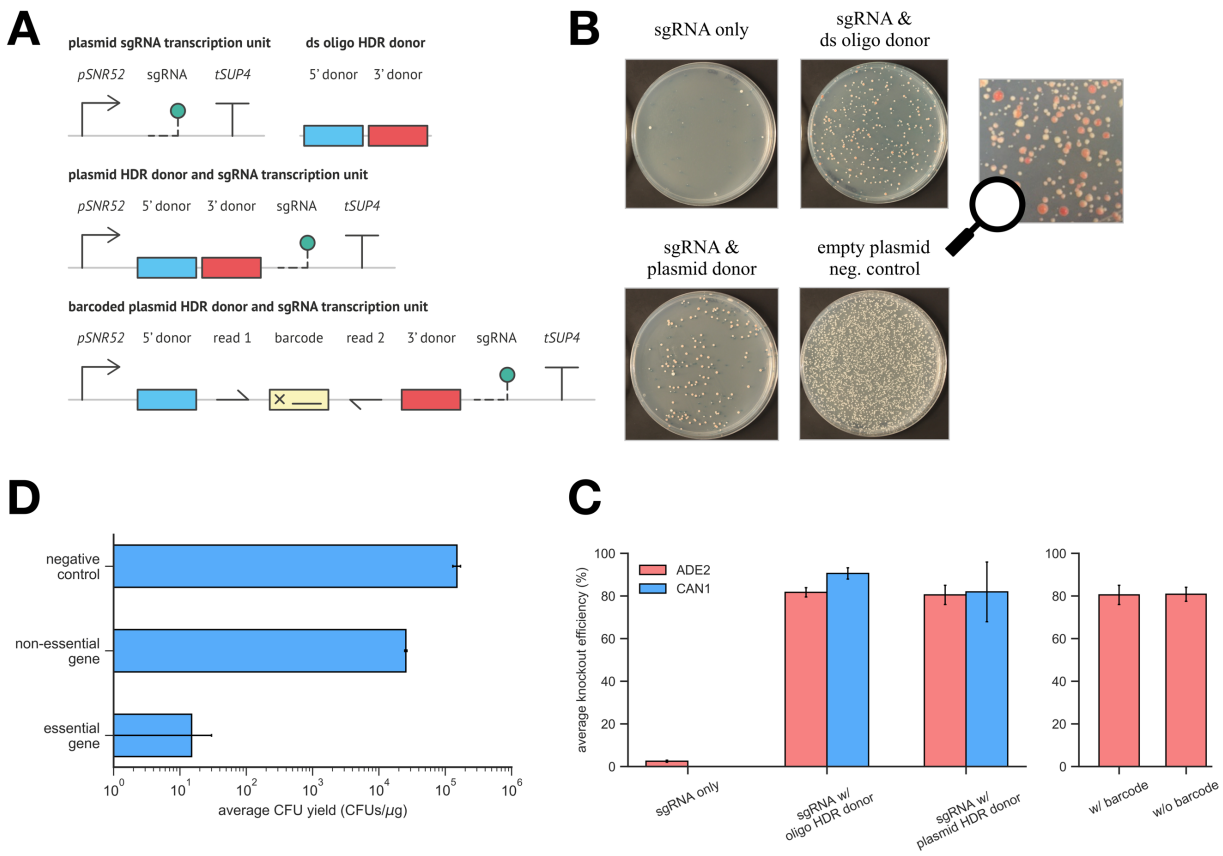


Figure 4.4 Proof-of-concept Single-gene Deletion Experiments

(**A**) CRISPR plasmid inserts tested: Function of several gene deletion constructs was assessed in proof-of-concept experiments to test whether plasmid-encoded HDR donors can repair DSBs as efficiently as co-transformed HDR donor DNA (**Fig 4.3B**). NHEJ repair (plasmid sgRNA transcription unit) and HDR positive controls (plasmid sgRNA transcription unit with the ds oligo HDR donor co-transformed) were included. Two versions of test constructs were tested, i.e., with and without the barcoding cassette (**B**) Example plate results of the *ADE2* gene knockout experiments: *ADE2* gene knockouts result in yeast cells exhibiting red pigmentation (**C**) Results of the single-gene knockout experiments: ~80% knockout efficiencies were observed in the *ADE2* and *CAN1* non-essential gene knockout experiments. Three biological replicates were tested per each experiment and error bars indicate standard error of the mean. (**D**) Results of the single-gene knockout experiments 2: Low recovery of cells was observed in the essential gene deletion experiments, implying successful gene knockout. Three biological replicates were tested per each experiment and error bars indicate standard error of the mean.

4.3.4 Cloning of the Deletion Library

Having validated the design, the next step was cloning and testing the entire genome-wide gene deletion library *en masse*. The deletion cassettes were cloned into the plasmid backbone using Gibson assembly (**Fig. 4.5**). To ensure presence of each of the deletion cassettes in the resulting plasmid DNA pool 4 DNA assemblies and 34 *E. coli* plasmid DNA electroporations were performed, which led to 20X CFU coverage of the library.

To investigate correctness of the obtained bacterial clones, 20 *E. coli* colonies were picked and their plasmid insert regions sequenced. 15 out of 20 colonies (75%) harboured correct full-length HDR donor – crRNA cassettes. 3 colonies propagated truncated DNA cassettes, while the remaining 2 sequences did not align against the reference deletion library (**Supp. Table 3.1.2 in Appendix 3.1**). Agilent quotes that 1 sequence error is to be expected per 300 nt of synthetic oligonucleotide DNA. This means that the likelihood of synthesis success at each nucleotide sequence position is 0.997 $((300-1)/300)$ and therefore at least one sequence error was expected in 50.2% of the ~209 nt oligonucleotide constructs $(1-0.997^{209})$. Thus, results of the initial Sanger sequencing assessment indicated a more moderate error rate.

Following the initial Sanger sequencing assessment, the entire plasmid pool was sequenced to investigate coverage of the genome-wide deletion cassettes. Next-generation Illumina sequencing was used to sequence the DNA barcode regions, so that the sequencing reads allowed to assess the abundance of individual deletion cassettes. Results indicated that a vast majority (~98%) of deletion cassettes was present in the plasmid pool (**Fig. 4.6**).

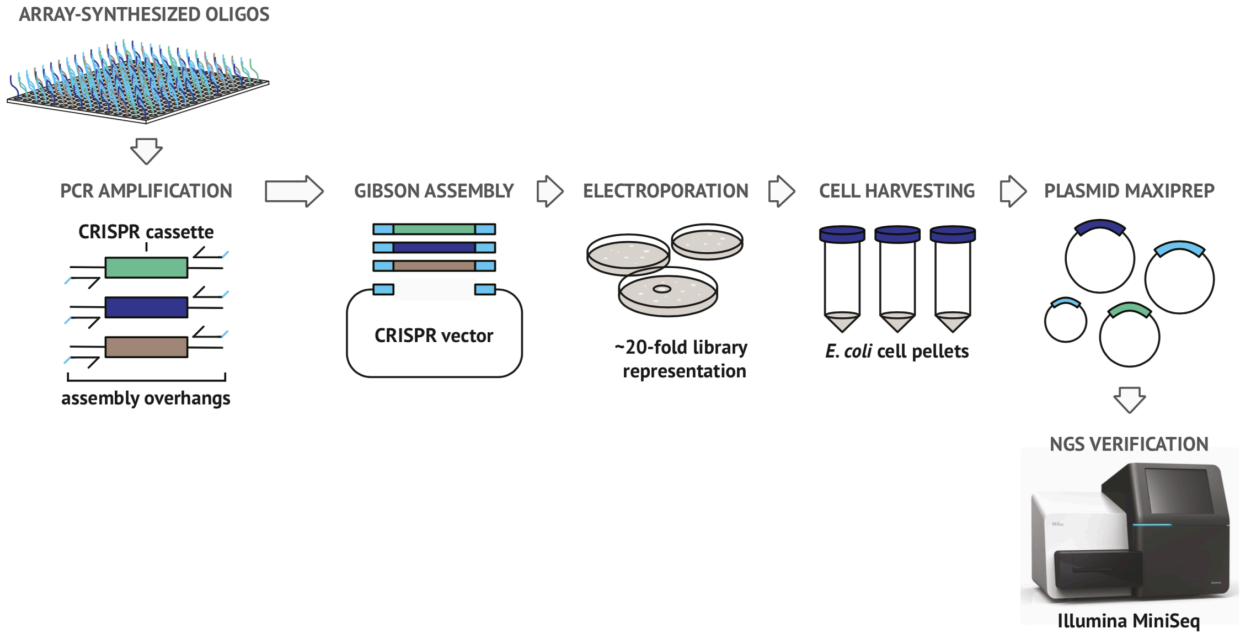


Figure 4.5 Cloning of the Genome-wide Plasmid Pool – Experimental Procedure

A commercial ssDNA oligonucleotide pool, harbouring 27,000 gene deletion cassettes, was amplified to yield double-stranded oligonucleotides with terminal plasmid homologies. These oligonucleotide fragments were later assembled into the library vector using Gibson Assembly. DNA assembly products were electroporated into *E. coli*. Multiple DNA assemblies and electroporations were performed to achieve a multiple library coverage. Amplified plasmid DNA material was purified and subjected to NGS verification.

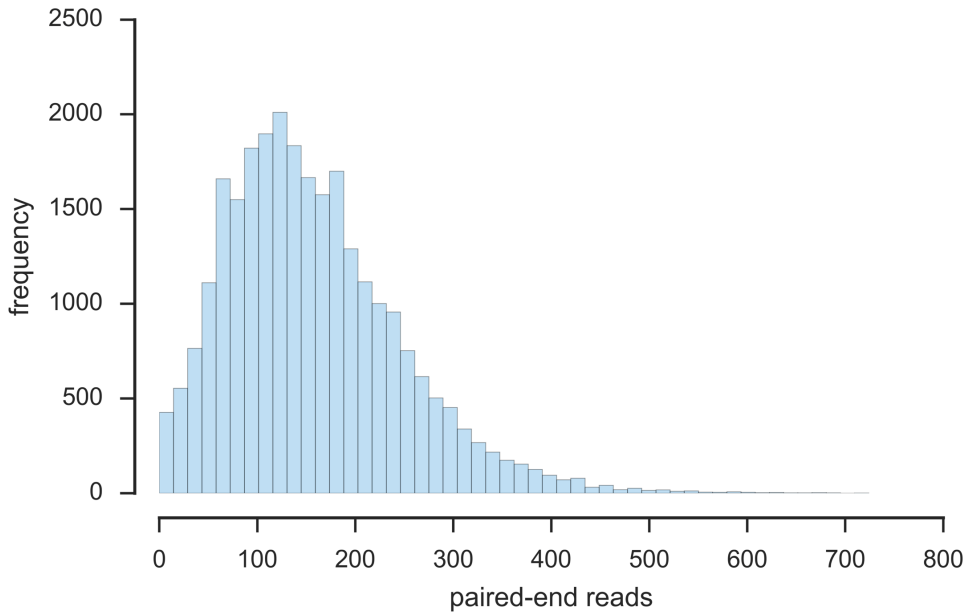


Figure 4.6 Cloning of the Genome-wide Plasmid Pool - Results

Distribution of the plasmid barcode cassettes' NGS paired-end read coverage. NGS verification of the cloned plasmid pool confirmed presence of nearly all library cassettes.

4.3.5 Genome-wide Gene Deletion Experiments

Having confirmed completeness of the plasmid library, it was decided to first test its function in a standard S288C-derivative laboratory yeast strain. This should result in strong depletion of the essential gene barcodes as compared to the non-essential ones, which would indicate correct genome-wide function of the CRISPR system. However, this was not seen in the first genome-wide experiments (**Fig. 4.7, 4.8, 4.9 and 4.10**). In these experiments, the plasmid deletion pool was first transformed into the wild-type strain, obtaining a ~50X CFU coverage of the library (each transformation procedure contained a set of appropriate controls, (1) a positive control: transformation of plasmid DNA without the CRISPR system, to monitor plasmid DNA transformation efficiency; and (2) a negative control: transformation of a blank water sample, to confirm no yeast growth on auxotrophic media in the absence of plasmid DNA). Following a plate outgrowth period, the yeast clones were harvested, pooled and stored in glycerol stocks, which were used in the subsequent experiments. One of the glycerol stocks was used to inoculated a rich liquid medium and sub-cultured into a plasmid counter-selection minimal medium (with 5-FOA) to compare barcode read depths for the plasmid-containing and plasmid-deficient yeast populations (to confirm successful genomic editing). Following the counter-selection, the mutant pool was further transferred into the rich non-selective medium and grown until mid-exponential phase (**Fig. 4.7**). While there was a significant difference between the essential and non-essential gene barcode read count distributions (as judged by a two-sample Kolmogorov-Smirnov test), no strong depletion of the essential gene barcodes was observed in either of the harvested solid and liquid media populations (**Fig. 4.9 and 4.10**). Both essential and non-essential gene read count distributions exhibited large standard deviations (sample 1 – 236/432, sample 2 – 690/1,612, sample 3 – 445/945 and sample 4 – 93/219 essential/non-essential gene reads) and comparable means (sample 1 – 173/188, sample 2 – 135/184, sample 3 – 118/156 and sample 4 – 40/61 essential/non-essential gene reads). Moreover, presence of the essential gene read signal in these populations indicated failure to knock out essential genes upon the barcoding sequence insertion. Possible reasons for this include, e.g., barcode mis-insertion.

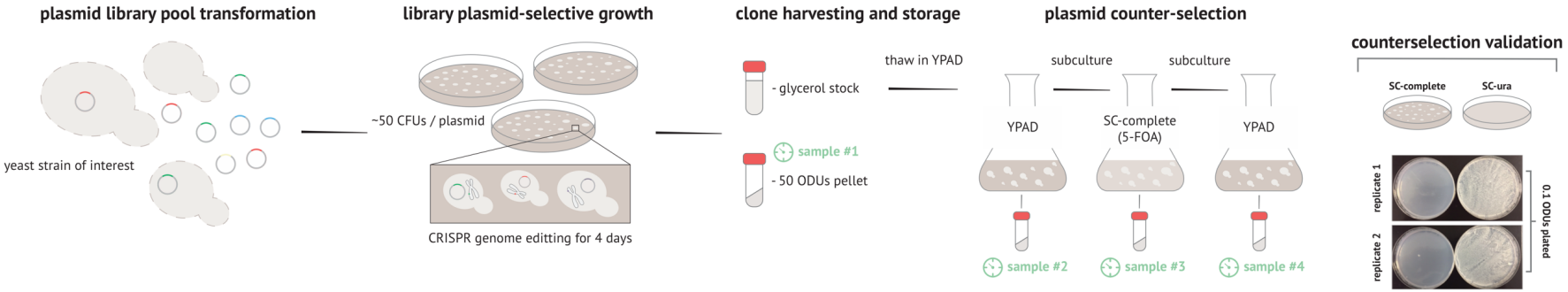


Figure 4.7 Setup of Genome-scale Knockout Experiments

The CRISPR plasmid pool was transformed into a standard laboratory yeast strain (BY4741). The transformation procedure was scaled up to obtain multiple coverage of the library. Yeast clones (grown on CRISPR system-selective media plates for several days) were harvested and stored in glycerol stocks. A single glycerol stock was later recovered in a rich non-selective YPDA liquid medium and the recovered cells were subjected to a chemical plasmid counter-selection treatment. Following plasmid counter-selection, plasmid loss was confirmed and the mutant population cultured in YPDA medium until exponential phase was reached. 50 ODUs samples were collected, following growth in every condition (samples 1-4). Total DNA material was extracted and subjected to further barcode copy number NGS analysis.

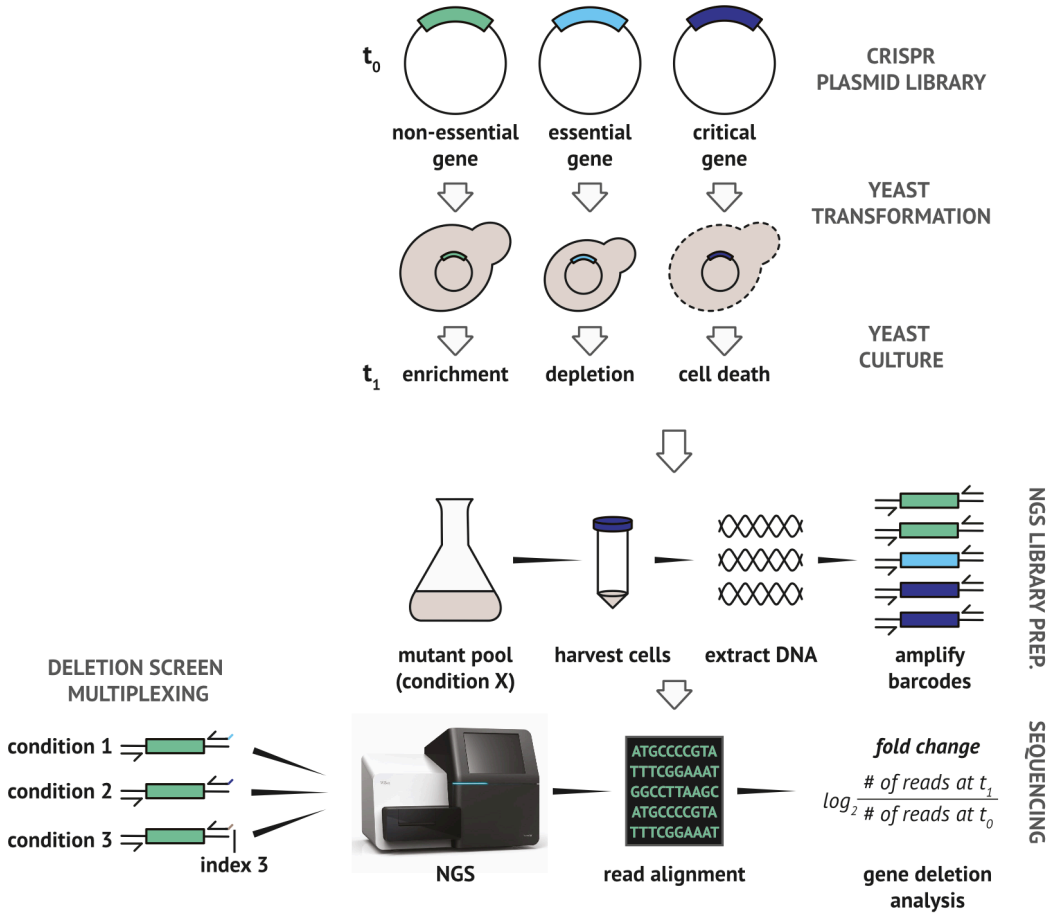


Figure 4.8 First Yeast Genome-wide Screens Workflow

The CRISPR plasmid library is cloned and library representation of individual barcoded gene deletion cassettes is determined by NGS (serving as a baseline time zero reference later on). The library harbours deletion cassettes targeting non-essential and essential genes. Upon transformation of an *Saccharomyces cerevisiae* strain of interest with the plasmid pool, significant depletion of strains harbouring the essential gene deletion cassettes is expected. In order to test this hypothesis, after a period of post-transformation outgrowth a yeast cell pool is harvested by centrifugation and total DNA is extracted from the pool. Barcode DNA is later amplified through PCR, quantified and subjected to NGS. Read count per individual non-essential and essential gene barcodes is obtained and its fold-change between the post-transformation (*t1*) and pre-transformation steps (*t0*) is computed. Because total read counts vary between sequencing runs, barcode read counts are normalised by the total sequencing run read counts before the fold-change computation. Different yeast deletion library outgrowth conditions can be tested in one sequencing run, as amplified barcode libraries can be assigned different NGS identifiers and mixed.

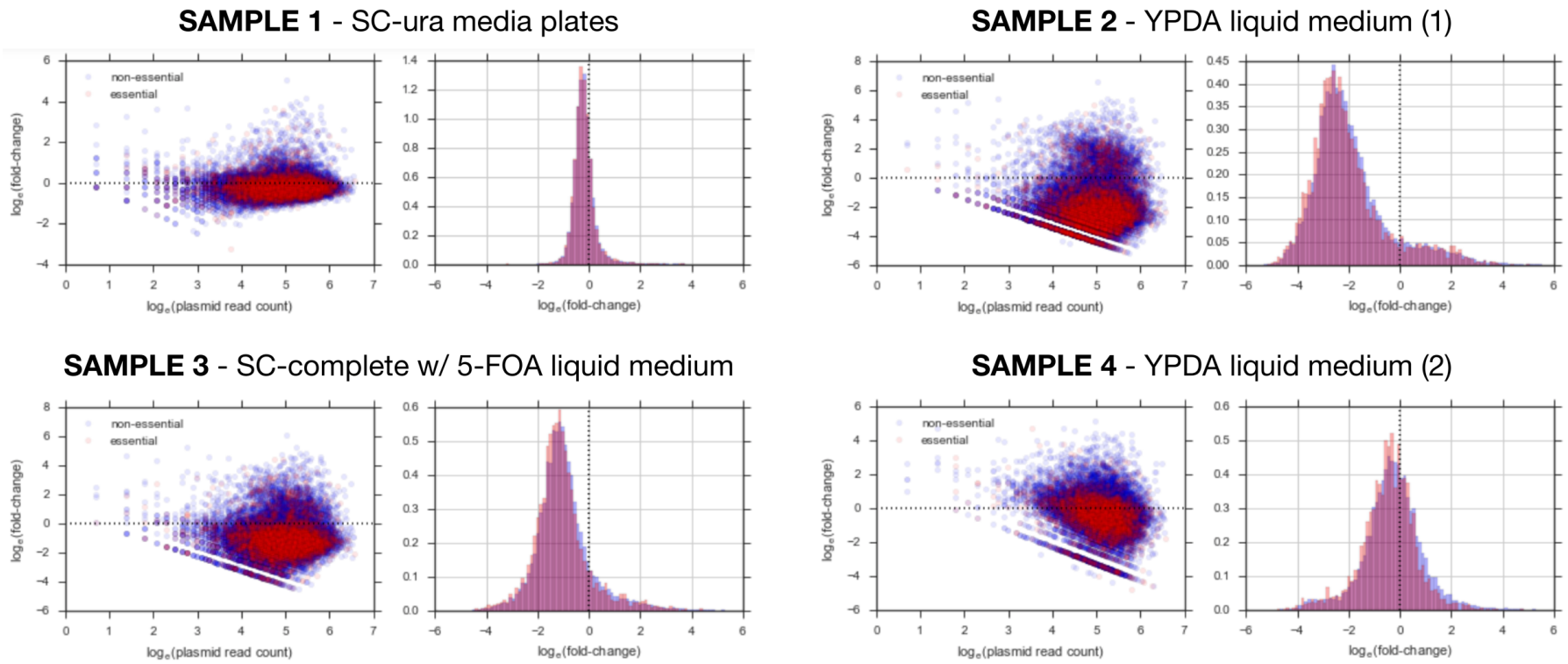


Figure 4.9 First Yeast Genome-wide Screens – Results

No strong essential gene mutant depletion was observed, as compared to non-essential gene mutants. No depletion was observed for both plasmid-selective (SC-ura media) and plasmid non-selective / counterselective growth conditions (YPDA and SC-complete w/ 5-FOA media, respectively). Fold-change data is presented, which describes change in barcode read counts relative to the starting plasmid read counts (CRISPR plasmids cloned in *E. coli*). Plasmid read counts are projected onto a logarithmic scale for better data spread. Histograms with normalised frequency are presented.

To investigate whether more time is needed to knock out yeast genes, a further experiment tested whether an additional plasmid-selective liquid outgrowth (48 and 96 hours), following growth on the plasmid-selective media plates, would lead to the anticipated strong essential gene mutant depletion. However, again, a strong reduction in the essential gene read signal was not observed. The read count standard deviations were considerable (48 hours: 1,164 and 1,970; 96 hours: 2,971 and 3,195 essential and non-essential gene reads) and the read count means counts similar (48 hours: 207 and 253; 96 hours: 271 and 314 essential and non-essential gene reads).

Aside from comparing the essential and non-essential gene read count distributions, the same analysis was conducted for the negative control non-targeting experiments, which used plasmids with random HDR donor and crRNA sequences. Results revealed a significant enrichment of the non-targeting construct barcodes, as compared with read counts of the genome-targeting constructs, in all growth condition experiments. While such enrichment was expected in the plasmid-selective screens, it was not anticipated in growth conditions lacking plasmid selection. In the plasmid selective screens, targeting constructs should lead to DSBs and barcode sequence insertions, while non-targeting constructs should lead to neither, therefore conferring growth advantage. In the non-selective conditions, non-targeting constructs should however not be enriched, as they are not designed to be integrated into the genome and thus they are only encoded by plasmid DNA, which is not selected for. Possible reasons for these results could be an inefficient counter-selection protocol (although, its correct function was assessed; **Fig 4.2 and 4.7**) or genomic insertions driven by events unrelated to CRISPR editing (Evert *et al.*, 2004).

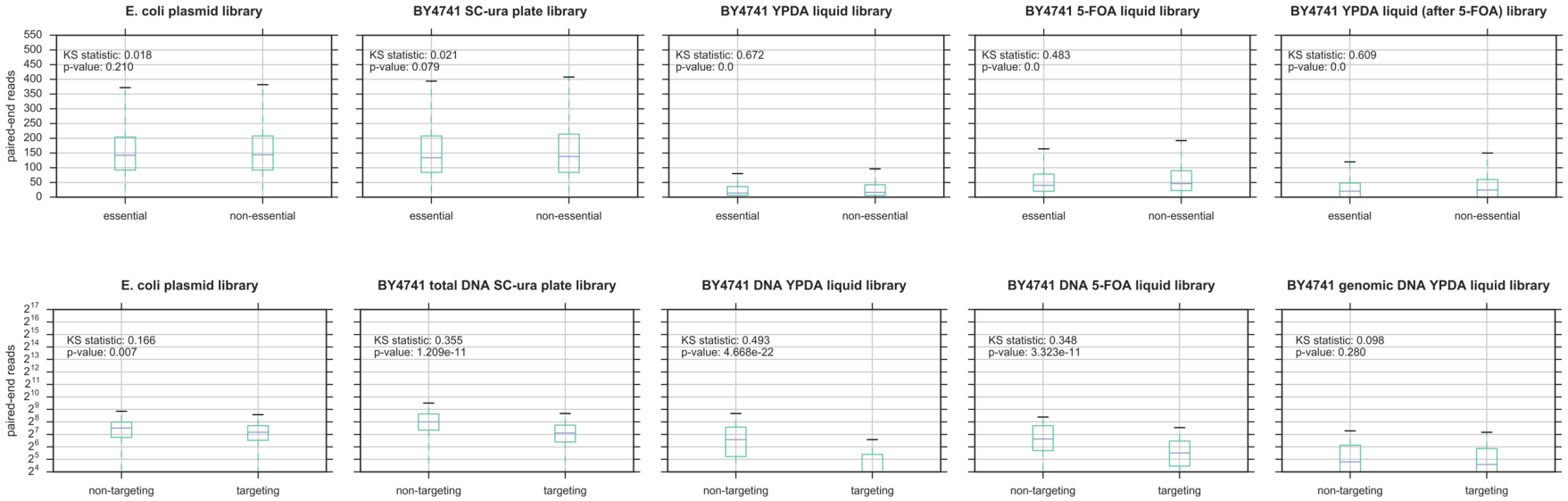


Figure 4.10 Genome-scale Knockout Experiments (including negative control data)

Significant difference between essential and non-essential gene read count distributions was observed, as judged by the two-sample Kolmogorov-Smirnov test. However, no strong depletion of the essential gene barcodes was detected. Significant enrichment of non-targeting construct barcode reads was observed as well in all growth condition experiments, and was not expected in screens lacking plasmid selection.

To further investigate the unexpected lack of a strong essential gene read depletion, individual barcode reads of three example genes were traced, including two non-essential genes (*YHR101C* and *YNL036W*) and one essential gene (*YNL151C*). These were followed at different experimental stages to look into the consistency of their NGS signal. For *YHR101C*, all 4 constructs exhibited a similar read count trend (with a noticeable population bottleneck at the cryopreservation recovery). However, for the remaining examples inconsistencies were observed (**Fig. 4.11**). For instance, reads of the fourth essential gene barcode exhibited a general decreasing tendency, to then increase ~2-fold at the last experimental stage. Read number of the third *YNL036W* gene barcode, on the other hand, reached 824 reads at the glycerol stock recovery, to then drop to 0 reads at the final experimental stage. It was therefore presumed that measurement noise was present in the gene deletion screen results, which could have obscured the NGS read analyses. Genetic suppression could have contributed to the following measurement noise. Namely, some cells could have harboured a second mutation, introduced by chance, which could have relieved the essential gene knockout phenotype, giving them a growth advantage and leading to their enrichment in the gene knockout strain pool. For instance, deletion phenotype of the *YNL151C* (*RPC31*) gene, encoding an RNase III C13 subunit, can be supressed by the *SSD1* gene deletion (Uesono, Toh-e and Kikuchi, 1997). Eliminating such experimental noise on a library scale is difficult and could have only been possible through additional whole genome sequencing of the gene deletion mutant pool, which has not been performed.

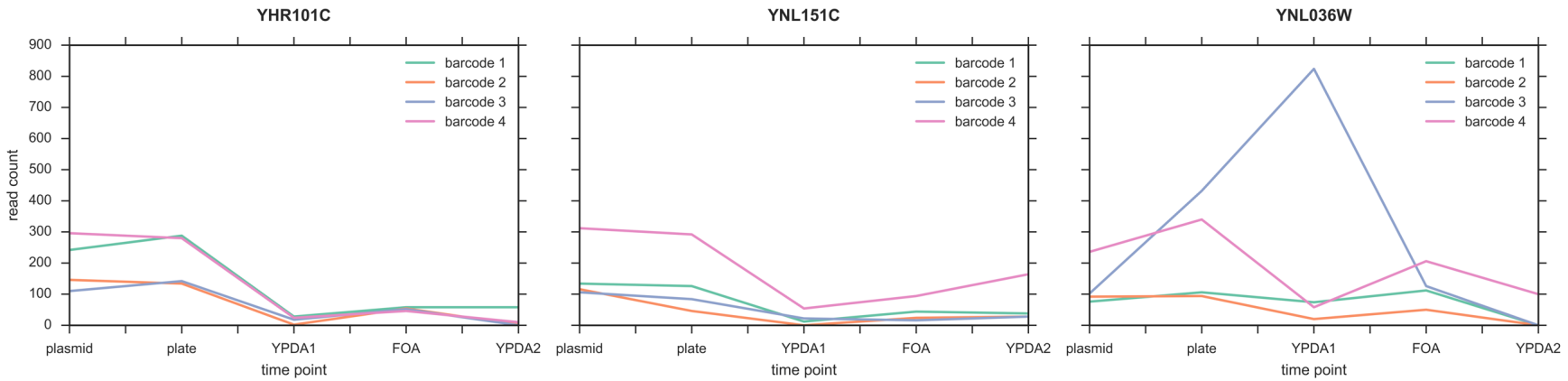


Figure 4.11 Barcode Read Count Anomalies

Some barcodes exhibited abnormal read counts, e.g., essential gene barcode read count increased following a decrease. Three example genes were investigated, including two non-essential genes (*YHR101C* and *YNL036W*) and one essential gene (*YNL151C*).

4.3.6 Establishing a New Method for Evaluating Correct Mutagenesis

The initial approach towards confirming successful genome editing relied on quantifying genomic barcode reads in the absence of the barcode-encoding plasmid DNA. However, this strategy relied on an indirect method for confirming plasmid loss (i.e., testing cell growth on a uracil-deficient minimal medium) and was unable to detect barcode mis-integrations. Therefore, given the unexpected results of the first genome-wide experiments, a new sequencing method was proposed, which would better distinguish true genome insertion events as well as spot mis-integration events.

To do that, a different sequencing sample preparation protocol was devised and was designed to capture DNA sequences proximal to the barcoding cassettes (**Fig. 4.12 B**). The new method relied on a sequence of enzymatic reactions processing biotinylated ssDNA bound to streptavidin-coupled magnetic beads. Immobilisation of the single-stranded DNA aimed at facilitating progression through consecutive enzymatic steps, e.g., by circumventing the need to precipitate and wash low DNA amounts to set up subsequent reaction steps.

To test the new strategy, a yeast strain was constructed, harbouring an *ADE2* gene deletion triggered by a barcoding sequence insertion. In proof-of-concept experiments, a different barcoding cassette design was used to make the first future NGS tests easier, i.e., without the need to design and use custom sequencing primers (**Fig. 4.12 A**). Two types of starting nucleic acid materials harbouring the barcoding cassette were tested - a pre-amplified ~600 bp genomic region and full-length genomic DNA. It was hypothesised that using a less complex DNA template (i.e., the amplicon DNA) would help prove that this approach is feasible and later optimise the new experimental protocol (i.e., the amplicon sequence being used to set up positive control experiments).

The barcode-neighbouring DNA sequences were captured through a single round of DNA polymerase extension (using a biotinylated primer), past the barcoding cassette

sequence and one of the homology arms. The resulting single-stranded DNA was next separated from the template through thermal denaturation to be then circularised, using a ssDNA ligase catalysing intramolecular ligation. Annealing of a single-stranded oligonucleotide complementary to a pre-designed BamHI restriction site allowed circular DNA to be linearised between outward facing NGS primer-binding sites. The resulting varied size population of linearised fragments was eluted and used as a heterogenous template in a PCR reaction harbouring NGS primers, thus generating an Illumina sequencing library (**Fig. 4.12 B**).

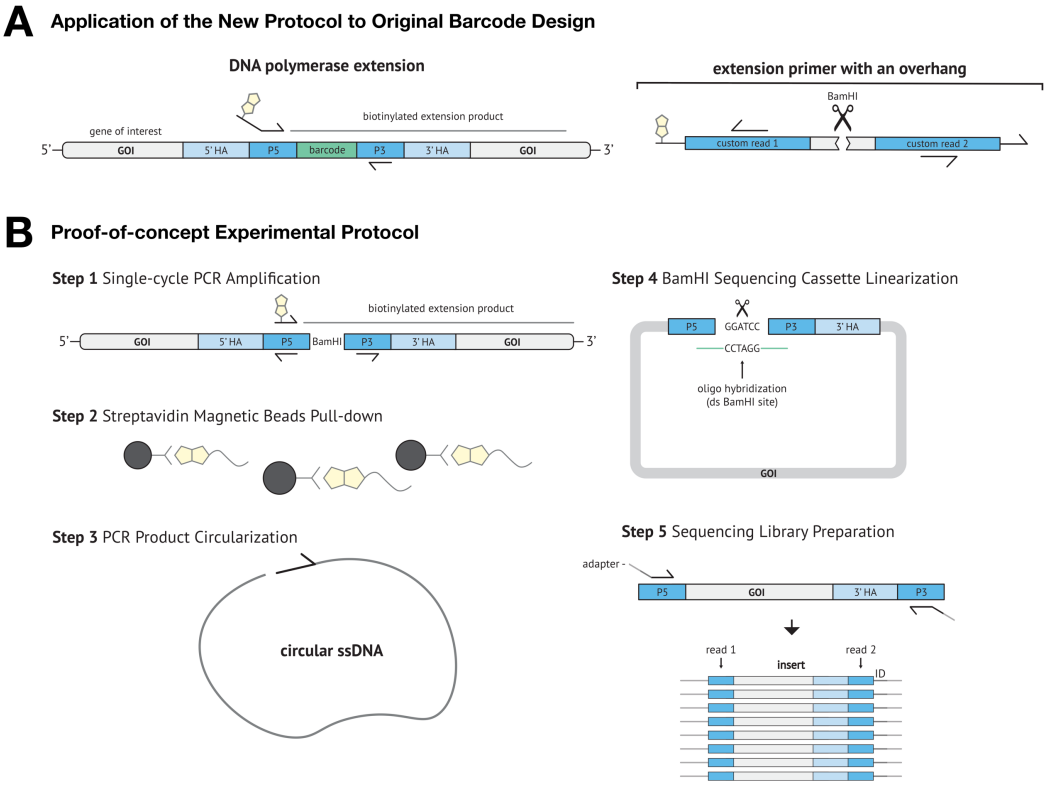
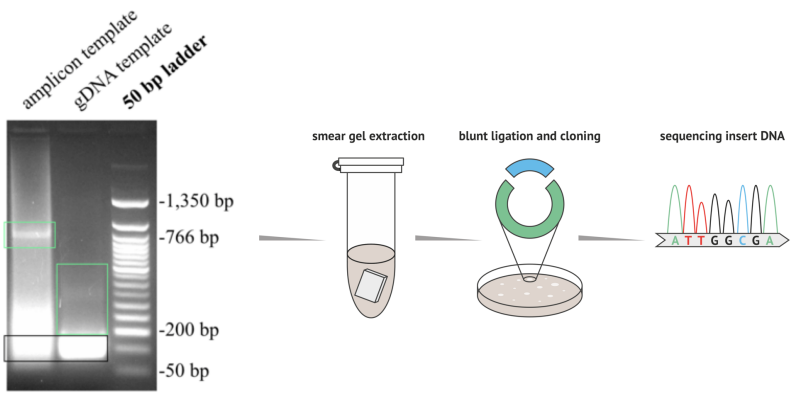


Figure 4.12 New Next-generation Sequencing Library Preparation Protocol

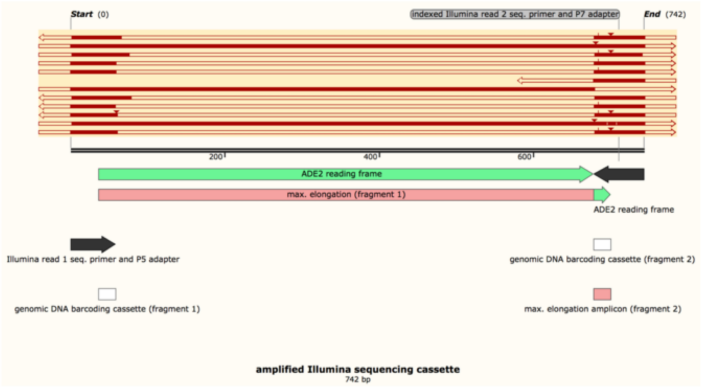
(**A**) Application of the new method to original barcoding cassette design: Application of the new method requires using an extension primer with an overhang region encoding outward facing custom read primer binding-sites and an internal restriction site (**B**) New method for generating NGS sequencing libraries: The second alternative NGS sequencing method includes a series of enzymatic reactions, processing a single-stranded DNA extension product bound to streptavidin magnetic beads, which captures genomic DNA regions neighbouring genomic barcode cassettes.

Successful generation of NGS sequencing cassettes containing portions of the starting nucleic acid templates was tested by cloning them into a commercial plasmid vector and sequencing the insert region of the plasmid DNA extracted from 12 *E. coli* clones (**Fig. 4.13 A and B**). For the amplicon template, 10 correct sequencing cassettes were identified (although, successful sequencing cassette generation is questionable in case of the 3 1 bp elongation products), with 3 of them capturing the complete amplicon nucleic acid sequence (**Fig. 4.13 B**). However, for the full-length genomic DNA template no correct sequencing cassette inserts were observed, with some inserts harbouring short sequences of random DNA. Therefore, further efforts are still required to optimise the new NGS protocol.

A



B



clone #	elongation (bp)
1	8
2	620 (full)
3	18
4	1
5	1
6	0
7	620 (full)
8	21
9	0
10	1
11	620 (full)
12	3

Figure 4.13 Testing the New NGS Protocol

(**A**) Sanger sequencing analysis of putative NGS inserts: Amplified DNA products were analysed by agarose gel electrophoresis. DNA fragment smear signal (indicated in green) above a putative NGS primer dimer region (indicated in black) indicates successful generation of NGS cassettes. In the amplicon template experiment, a distinct ~600 bp DNA band indicates NGS cassettes harbouring complete amplicon sequences (indicated in green). DNA was extracted from the smear regions and cloned, using a commercial vector system. The resulting plasmid insert regions were later verified by Sanger sequencing; (**B**) Sanger sequencing results: Generation of successful NGS cassettes was observed for the amplicon template.

4.4 Discussion

In conclusion, following successful single-gene deletion experiments using the proposed CRISPR system design, correct genome-wide function of the system was not demonstrated. There might be several causes of these unexpected results, the combined effects of which could have led to the lack of a strong essential gene read depletion.

As mentioned before, the oligonucleotide error rate stated by the manufacturer indicated that 50.2% of the purchased oligonucleotide constructs were expected to contain at least one sequence error. Errors in the single guide RNA could have reduced genome editing efficiency, by impeding DNA recognition and cleavage, therefore leading to a lack of selective pressure to incorporate the mutagenic barcode sequences through the homology-directed DNA repair mechanism. This issue should however be identifiable by efficient plasmid counter-selection, resulting in a weak non-essential gene read signal. Errors in the donor DNA could have also impeded barcode insertion, but also could have resulted in barcode mis-integrations, which could have confounded analyses of the essential gene barcodes. Barcode mis-incorporation at non-essential genomic loci should give rise to viable cells bearing essential gene identifiers, therefore leading to spurious results. Such mis-incorporation events are however unlikely in the absence of double-stranded DNA breaks at the corresponding genomic DNA loci (i.e., no selective pressure for DNA integration) and given the length of homologous arms used (Hua *et al*., 1997). Synthesis errors could have also been present in the barcode sequences and could have had two negative implications. First, they could have led to false clone identification. Second, they could have prevented introduction of premature stop codons. Here, even a small indel mutation could have led to undesired frameshifts or conversion of a programmed termination triplet into an amino acid-coding codon.

A recent publication by Roy *et al*., describing the MAGESTIC system, presented solutions to the synthesis error-related issues (Roy *et al*., 2018). The published genome-wide editing system tagged the already synthesised oligonucleotides with

unique 31 bp barcode sequences, which reduced the danger of using erroneous barcodes (through a more error-prone synthesis of longer oligonucleotides). Moreover, post-synthesis assignment of these resulted in identical editing cassettes bearing different barcode identifiers, which served as internal replicates. As opposed to the CRISPR system developed in this project, the MAGESTIC system barcoded the underlying gene mutants by integrating the entire barcoded HDR donor – crRNA construct, flanked by NGS-compatible primer-binding sites, at a separate (from the corresponding gene locus) engineered landing pad bearing a counter-selectable gene marker. Longer read next-generation sequencing of this insert therefore allowed identification of mutants propagating erroneous editing constructs. Separate genomic location of the barcoded sequence, unlinked to the target gene locus, is however not ideal, since the landing pad insertion does not provide direct proof of a successful gene editing event. Moreover, this methodology requires prior engineering of a landing pad for the barcoded editing cassette in every yeast strain of interest, leading to a perturbation of the original genetic background, which might be undesirable (see **Table 4.1** for detailed methodology comparison).

Genomic insertion of the barcoded gene editing cassette in the MAGESTIC editing pipeline was accomplished through a homology-directed CRISPR-Cas9 mechanism, yet using a different single-guide RNA sequence from the gene targeting sgRNA. This RNA sequence was encoded by the plasmid vector used to deliver the gene editing cassettes and targeted two genomic sites flanking the landing pad construct, as well as one plasmid DNA site at the downstream end of the gene editing cassette. Therefore, expression of this additional single-guide RNA led not only to the genomic barcoding event, but also to self-destruction of the plasmid DNA. Plasmid self-destruction could have been used as an alternative way to counter-select plasmid DNA in the presented CRISPR knockout experiments. Such an alternative plasmid elimination strategy would circumvent selective stress associated with 5-fluorouracil toxicity, which can alter the composition of a heterogenous mutant pool (i.e., cells which lose plasmid material faster have a competitive advantage over cells which are slower at losing their plasmid DNA). Moreover, the self-destruction mechanism can serve as an internal control for correct Cas9 function.

Insufficient plasmid copy numbers could have also been an issue in the gene deletion screens. Since efficient single-gene knockouts were observed in the proof-of-concept experiments, plasmids levels were not expected to be a limiting factor in the genome editing system, which harboured a two micron plasmid origin of replication leading to 10-40 plasmid copies per cell (Futcher and Cox, 1984). However, it might be possible that for some genomic loci it was in fact limiting. Repair of double-stranded breaks using an exogenous repair template (as opposed to, e.g., a sister chromatid template) requires a physical search for the donor DNA. For example, in budding yeast and other eukaryotes various proteins help in relocating damaged genomic loci to nuclear compartments which better assist in DNA repair (Seeber and Gasser, 2017). Therefore, increasing availability of the donor template may lead to increased genome editing efficiencies. Recently, Bao *et al*. boosted intracellular levels of their two micron CRISPR-Cas9 plasmid (to ~200 copies) by using a truncated auxotrophic marker promoter and later demonstrated successful genome-wide *S. cerevisiae* editing using this enhanced plasmid vector (Bao *et al*., 2018). The aforementioned MAGESTIC system, on the other hand, used an engineered LexA-Fkh1p fusion protein to recruit the plasmid-encoded repair templates to the double-stranded breaks. MAGESTIC plasmid vectors harboured several LexA DNA-binding sites, proximal to the donor DNA, to which the LexA portion of the fusion protein was able to bind. The Fkh1p side of the fusion construct interacted with proteins which accumulate at double-stranded breaks. These proteins are phosphorylated on particular threonine residues, to which the Fkh1p part binds, thus recruiting MAGESTIC donors to the cleavage sites, which led to a more than 5-fold increase in the editing efficiency.

In general, in genome editing experiments the non-homologous end joining (NHEJ) and the homology-directed repair (HDR) pathways are two major competing DNA repair routes. In this project, introduction of custom genomic alterations was achieved through the preferred HDR mechanism. NHEJ repair of DNA breaks could have however impeded introduction of these to some extent. In the proof-of-concept studies led to ~80% editing efficiencies. However, these could have been further

enhanced by inactivating NHEJ. Roy *et al*. achieved enhanced editing efficiencies using this strategy. The *NEJ1* gene encoding a protein involved in regulating non-homologous end joining was deleted, which led to nearly 100% editing efficiency (Roy *et al*., 2018). Nonetheless, like the aforementioned genomic landing pad engineering, the *NEJ1* knockout might be an undesirable modification of the original genotype in certain applications.

Table 4.1 Comparison of the constructed deletion library and other published *S. cerevisiae* CRISPR deletion libraries

Method Name	General Mutagenesis Mechanism	Strategies for Mutagenesis Confirmation	Strategies for Increasing Mutagenesis Efficiency	Strategies for Addressing Oligonucleotide Error Rates	General Advantages	General Disadvantages	Reference
Method described in this manuscript	transformation of a yeast strain of interest with a CRISPR plasmid library pool, so that one barcoded guide-donor DNA library plasmid is introduced into each cell; CRISPR system expression leads to target site mutagenesis, through mutagenic HDR ((1) PAM site and partial protospacer deletion, (2) insertion of a barcode cassette at the DSB site, and (3) pre-mature stop codon generation)	NGS sequencing of barcode cassettes integrated at the mutagenesis sites; prior plasmid DNA counterselection required (using 5-FOA) to consider genome-integrated barcodes only; alternative NGS sequencing method under development, potentially allowing unambiguous confirmation of correct mutagenesis through capture of DNA sequences downstream the barcoding cassettes (pre-eliminary tests demonstrate its potential viability on a genome-scale)	a high-copy library plasmid, increasing intracellular donor DNA copies (approx. 10-40)	none	one-to-one correspondence between genomic barcode counts and strain abundance; alternative NGS sequencing method under development, potentially able to unambiguously confirm correct mutagenesis; not many genotype requirements allow flexible choice of the strain background; library design compatible with genetic background of the Sc2.0 consortium synthetic yeast chromosome strains (i.e., preservation of the synthetic DNA watermark sequences and the reduced stop codon set)	cannot mutate DNA at a single-nucleotide resolution (the method aims at efficient gene knockouts); first attempts at genome-wide validation of the method (comparison of essential and non-essential gene knockout strain abundance) failed	none
MAGESTIC	transformation of a yeast strain of interest with a Cas9 expression plasmid and a plasmid library pool, so that one barcoded guide-donor DNA library plasmid is introduced into each cell; galactose-induced expression of Cas9 leads to simultaneous	NGS sequencing of the barcoded guide-donor plasmid inserts, integrated at pre-engineered genomic landing pads (distal to the corresponding target mutagenesis sites) via CRISPR with a separate guide RNA (leading to additional linearisation and	LexA-Fkh1p fusion protein-mediated recruitment of donor DNA to DSBs; a high-copy library plasmid, increasing intracellular donor DNA copies	post-library synthesis semi-random DNA barcode assignment to each guide-donor oligonucleotide and NGS of the resulting barcoded	single-nucleotide resolution mutagenesis; one-to-one correspondence between genomic barcode counts and strain abundance; multiple barcodes mapping to the same guide-donor combination provide internal replicates for a given mutagenesis experiment; temporal control of Cas9	unambiguous mutagenesis confirmation possible only through whole-genome sequencing; successful barcoded cassette integration does not prove successful mutagenesis at a different genomic locus; no genome-wide	Roy *et al.*, 2018

Method Name	General Mutagenesis Mechanism	Strategies for Mutagenesis Confirmation	Strategies for Increasing Mutagenesis Efficiency	Strategies for Addressing Oligonucleotide Error Rates	General Advantages	General Disadvantages	Reference
	(1) target site mutagenesis, through mutagenic HDR (using plasmid-encoded donor DNA), (2) integration of the barcoded guide-donor cassette at a different genomic locus (3) library plasmid destruction	degradation of plasmid DNA)	(approx. 10-40); knockout of the competing DSB repair pathway (NHEJ)	plasmid library allow identification of faulty constructs prior to mutagenesis experiments	expression (through galactose induction)	validation of the method (comparison of essential and non-essential gene knockout strain abundance); multiple strain genotype requirements limit strain choice flexibility	
CHAnGE	transformation of a yeast strain of interest with a CRISPR system plasmid library pool, so that one barcoded guide-donor DNA library plasmid is introduced into each cell and multiples inside it to an ultra high-copy level; CRISPR system expression leads to target site mutagenesis, through mutagenic HDR (using plasmid-encoded donor DNA)	ability to successfully mutagenise genomic DNA genome-wide inferred from pilot single- and double-mutant experiments (Sanger sequencing of the corresponding genomic DNA loci)	an ultra high-copy library plasmid with a 2 micron origin of replication and a truncated *URA3* promoter, increasing intracellular donor DNA copies (approx. 200); expression of a previously discovered higher efficiency Cas9 variant, i.e., iCas9	high colony coverage at the plasmid library cloning step (480- to 1,600-fold) ensures sufficient representation of DNA sequence-perfect guide-donor cassettes	single-nucleotide resolution mutagenesis; not many genotype requirements allow flexible choice of the strain background; supports iterative rounds of (1) plasmid library transformation, (2) mutagensis, (3) growth selection, (4) NGS analysis, and (5) plasmid counterselection (using 5-FOA), under increasing selective pressure	laborious plasmid library cloning procedure; unambiguous mutagenesis confirmation possible only through whole-genome sequencing; NGS sequencing of predominantly plasmid DNA barcodes does lead to a one-to-one correspondence between the barcode counts and strain abundance; no genome-wide validation of the method (comparison of essential and non-essential gene knockout strain abundance)	Bao *et al*., 2018
High-throughput	transformation of a yeast strain of interest	ability to successfully mutagenise genomic DNA	a high-copy library plasmid,	none	single-nucleotide resolution mutagenesis; approach	unambiguous mutagenesis	Guo *et al*., 2018

Method Name	General Mutagenesis Mechanism	Strategies for Mutagenesis Confirmation	Strategies for Increasing Mutagenesis Efficiency	Strategies for Addressing Oligonucleotide Error Rates	General Advantages	General Disadvantages	Reference
creation and functional profiling of DNA sequence variant libraries using CRISPR–Cas9 in yeast	with a CRISPR library pool, consisting of two linear DNA fragments sharing terminal homology (allowing plasmid circularisation via HDR); CRISPR system expression leads to target site mutagenesis, through mutagenic HDR (using plasmid-encoded donor DNA)	genome-wide inferred from Sanger sequencing experiments on a subset of yeast clones subjected to CRISPR, identifying their target mutagenesis loci via CRISPR plasmid insert sequencing and subsequently validating mutagenesis via sequencing of the corresponding genomic DNA regions	increasing intracellular donor DNA copies (approx. 10-40); intracellular plasmid DNA assembly increases efficiency of mutagenic HDR (from 0-30% to 80-100%)		validated through a partial (not all gene knockouts tested) genome-wide comparison of essential and non-essential gene knockout strain abundance	confirmation possible only through whole-genome sequencing; NGS sequencing of predominantly plasmid DNA barcodes does lead to a one-to-one correspondence between the barcode counts and strain abundance	

The presented CRISPR-Cas9 experiments concerned themselves with demonstrating successful essential gene knockouts using a reference list of 1,156 characterised essential genes (Giaever *et al*., 2002). However, accurate assessment of gene essentiality has to consider the cellular environment context (Zhang and Ren, 2015). Therefore, this reference list could have been further filtered to ensure that it corresponds to the investigated experimental conditions. Nevertheless, proper function of some essential genes is critical, largely irrespective of the environmental context. For instance, *ACT1* is an essential gene encoding actin, a conserved structural protein which controls proper cell polarisation, endocytosis and is involved in a number of other processes fundamental for the cell (Wertman, Drubin and Botstein, 1992). *ACT1* sequencing reads were therefore further scrutinised. Mean read count, corresponding to all 4 *ACT1* barcodes, however did not reveal a strong depletion (an average of 159 reads, i.e., 15% less than the non-essential mutant population sample mean).

Aside from the genome editing issues, confirmation of essential gene knockouts is challenging too. In the proof-of-concept experiments, it was concluded that a poor recovery of cells following transfection of the CRISPR system implied successful essential gene knockouts. However, it is also plausible that it was caused by inefficient mutagenesis, which is known to lead to burdening Cas protein cleavage cycles (Cagney *et al*., 2006). Therefore, in these initial studies, an example essential gene knockout, the lethal effect of which can be suppressed by a second mutation, could have been investigated instead. The essential gene deletion could have thus taken place in the presence and in the absence of genetic suppression, leading to more conclusive results.

To further investigate correctness of our genome editing, a novel NGS sequencing library preparation protocol was proposed, capable of confirming correct genomic barcode integrations in an unbiased manner. This alternative method relied on capturing and sequencing nucleic acid sequences neighbouring the barcoding cassettes and its application was expected to help detecting any incorrect barcode

insertions as well as any remaining plasmid vector background following the chemical counter-selection experiments.

This approach resembles some of the existing methodologies, e.g., inverse PCR or GUIDE-Seq (Tsai *et al*., 2014). However, these methodologies were incompatible with the barcoding cassettes used due to wrong primer-binding site orientation and lack of an internal restriction site, and were more prone to sequence amplification errors, respectively. In the initial experiments, feasibility of the proposed approach was demonstrated using a pre-amplified genomic DNA region as the extension template, but amplification signal was absent when unfragmented genomic DNA material was used as the extension template. Further optimisation of the first protocol step (i.e., the single-stranded DNA extension), in particular of the starting DNA template amount, could lead to improvement of these first results, since much greater molar amounts of the amplicon DNA were used owing to its much shorter length as compared to the genomic nucleic acid material. Moreover, genomic DNA sonication or enzymatic fragmentation could further help in obtaining NGS-compatible DNA amplicons with a more uniform size distribution.

Chapter 5
DNA Manufacturing Throughput Analysis via Probabilistic Simulations

5.1 Work Contributions

Most of the work presented in this chapter was conducted during an industrial placement at a Thermo Fisher Scientific DNA manufacturing facility (GeneArt; Regensburg, Germany).

Under supervision of Dr. Axel Trefzer and Dr. Michael Liss, I developed the Monte Carlo simulation software, conducted the proof-of-concept simulation studies and built a machine learning model to predict DNA sequence-specific manufacturing turnaround times. Dr. Phillip Kuhn provided additional guidance regarding process duration modelling as well as specifications for the software usage. The example manufacturing process information was obtained from Peter Poltnigg and Michaela Deinert, who also provided additional support in estimating parameters of the constructed example manufacturing process model. Tatiana Konovalova prepared the GeneArt Strings data mining dataset, which was used to train the machine learning classifier, and provided the raw manufacturing timestamp data.

5.2 Introduction

As DNA fabrication is becoming to an increasing extent high-throughput, development of tools for rational analysis of the expected DNA manufacturing costs and turnaround times is of growing importance. The costs of setting up and running high-throughput DNA production pipelines are substantial and therefore call for predictive methodologies capable of estimating DNA fabrication expenses and durations, as well as guiding researchers towards the most promising process optimisation targets (Chambers, Kitney and Freemont, 2016). However, decisions concerning setup, automation and planning of experimental procedures are often based on personal experience rather than data. Therefore, DNA manufacturing facilities could benefit from analytical methodologies used in other high-throughput industries (Jahangirian *et al.*, 2010; Cameron, Bashor and Collins, 2014).

However, to make these methodologies relevant to DNA fabrication the underlying DNA sequences should be considered. Different nucleic acid sequences are associated with unique levels of manufacturing complexity. Extreme GC content, stable secondary structure, DNA repeats and homopolymers are among the most problematic nucleic acid sequence features and often lead to manufacturing failures, alongside human error and equipment-related issues (Oberortner *et al.*, 2016). If possible, DNA characteristics should therefore be taken into account to ensure accurate prediction of the expected manufacturing costs and turnaround times as well as to assist in formulating DNA sequence constraint policies to recognise and reject DNA orders that are too challenging to complete within a given time frame.

This chapter therefore presents development and application of a probabilistic simulation framework to predict DNA manufacturing turnaround times and estimate the most significant manufacturing process parameters at one of the main synthetic DNA production facilities. To further improve predictions of the DNA production times, a classification model was built and used to estimate success probabilities of one of the investigated manufacturing steps, based on the underlying DNA sequence GC content and length. The proposed probabilistic simulation approach is evaluated

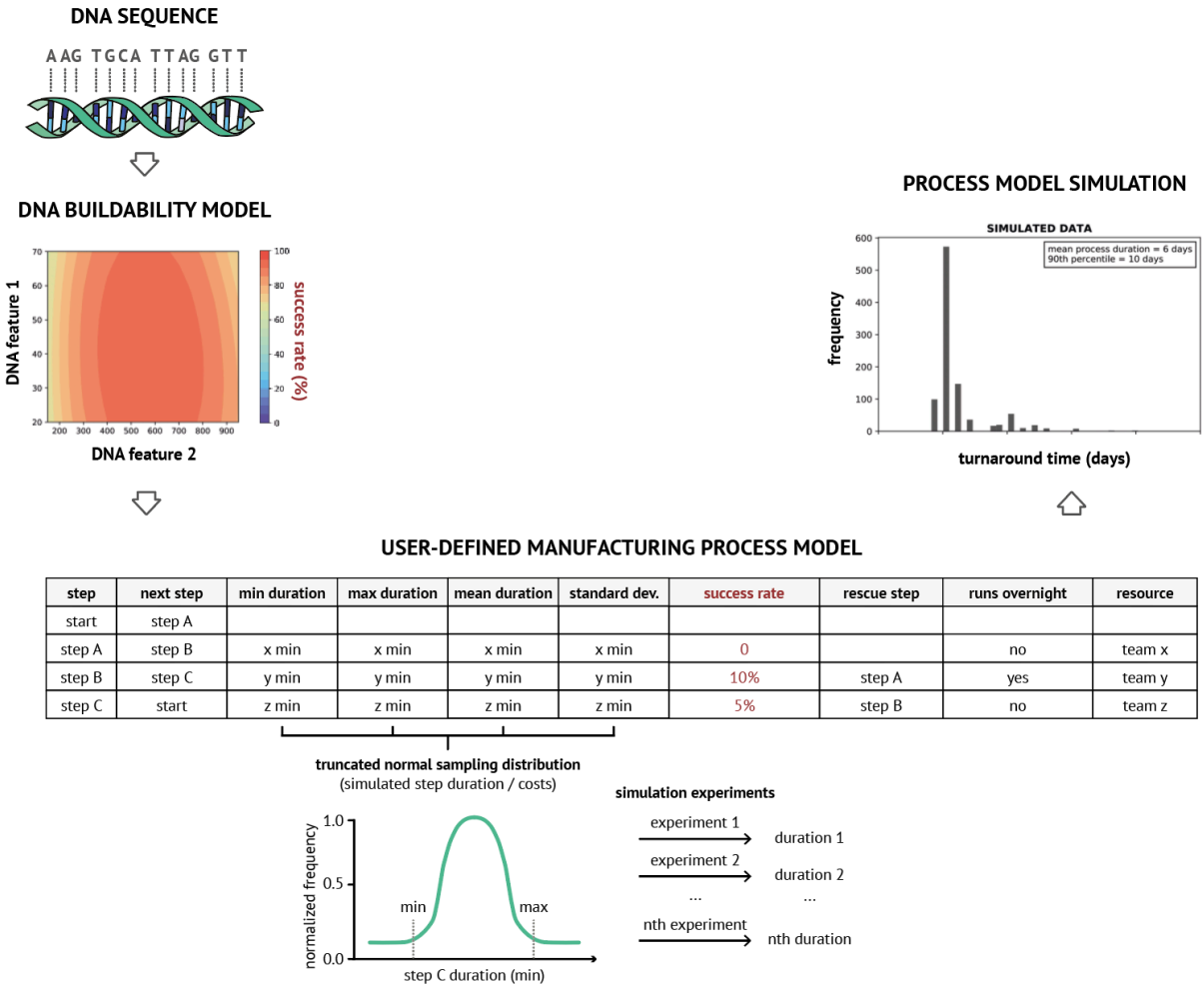
against real data from an example DNA manufacturing process. A statistical model of the example DNA fabrication process was built, using information obtained from the facility's personnel and manufacturing execution system, and simulation experiments were conducted using Python scripts to predict turnaround times in different manufacturing scenarios.

5.3 Results

Mathematical models of complex manufacturing processes can be too difficult to analyze using analytical approaches. Numerical methods, such as probabilistic Monte Carlo simulations, provide an alternative to computing analytical solutions of statistical models. Monte Carlo simulation studies use random number sampling from pre-defined distribution functions and statistical modelling to instead mimic behavior of such complex systems and estimate analytical solutions of their models (**Fig 5.1**).

5.3.1 The Probabilistic Simulation Methodology

Easy-to-use Monte Carlo simulation tools were therefore built to make computational analyses of the manufacturing turnaround times accessible to non-experts (Python code in **Appendix 4.3**). The tools constructed rely on a mathematical description of a given manufacturing process using a standard Microsoft Excel spreadsheet (**Supp. Table 4.2.1 in Appendix 4.2**), which is later analysed by Python simulation algorithms. The standardised spreadsheet requires users to define manufacturing steps and their order, as well as their failure probabilities and the corresponding manufacturing rescue routes.



step	next step	min duration	max duration	mean duration	standard dev.	success rate	rescue step	runs overnight	resource
start	step A								
step A	step B	x min	x min	x min	x min	0		no	team x
step B	step C	y min	y min	y min	y min	10%	step A	yes	team y
step C	start	z min	z min	z min	z min	5%	step B	no	team z

Figure 5.1 DNA Manufacturing Process Modelling Approach

The modelling work performed throughout this doctoral thesis project sought to develop tools for DNA-specific DNA manufacturing duration and costs simulation studies (process turnaround time simulation example shown above). In order to infer DNA-specific time and cost predictions, (1) statistical models have to be built and predict success of individual manufacturing steps, based on different DNA features (found to have significant impact on the success rates), and (2) a manufacturing step duration / cost simulation framework has to be developed. Success rates predicted in the first stage can be thus fed into the simulation analysis. The simulation framework developed throughout this thesis project allows the user to define the manufacturing process model, which will be later subjected to the simulation analysis. Each manufacturing step is thus assigned a statistical distribution which best reflects real life step durations / costs, and which will be used as a sampling distribution during the simulation studies (each simulation experiment samples one number from the sampling distribution; many simulation experiments are typically performed). Truncated normal distributions were chosen as the default sampling distributions (to avoid sampling extreme and negative time / cost values) and the user gets to define their parameters – minimum and maximum values, mean and standard deviation.

The tools developed use truncated normal distributions to model durations of consecutive DNA manufacturing steps. Truncated normal distributions are derived from the normal distribution and harbour two additional parameters defining strict minimal and maximal values (**Supp. Fig. 4.1.1 in Appendix 4.1**). These limit parameters can be used to prevent sampling of negative numbers, which is particularly probable for normal distributions with means close to zero and high standard deviations. Normal distribution is a common modelling choice for many phenomena. Various phenomena can be viewed as a composite of various independent probabilistic events, outcomes of which are described by independent random variables. According to the Central Limit Theorem and its variants, summation of these variables commonly tends towards a normal distribution, leading to the aforementioned observation (Brookes, 1955). In this work, the choice of truncated normal distributions in particular stemmed from an additional observation - that biological manufacturing step durations rarely exceed certain maximum values and are never negative values. For instance, duration of a polymerase chain reaction depends on a variable target amplicon length, impacting duration of the DNA extension phase. While PCR amplicon lengths are normally distributed, during gene synthesis they are frequently also bound by some minimum and maximum lengths, reflecting precise manufacturing needs. Additionally, PCR duration cannot fall below the minimum viable PCR protocol duration. Another example are bio-manufacturing steps involving microbial growth, e.g., if a certain bacterial incubation period is exceeded, cells will enter death phase and degradation of antibiotics will occur for experiments using antibiotic selection, leading to undesired bacterial growth (e.g., satellite colonies) (Hou and Poole, 1969). Too short incubation times, on the other hand, can lead to failure in obtaining microbial colonies or enough cells to extract sufficient quantities of plasmid DNA (Sezonov, Joseleau-Petit and D'Ari, 2007). Companies thus typically define minimum and maximum microbial outgrowth durations, in order to minimise manufacturing process failures.

Uniform, gamma distributions are two alternative statistical distributions allowing modelling non-negative values, as well as values bound by certain minimum to maximum ranges (uniform distribution), however none of them combines the ability

to model normally distributed event durations as well as minimum and maximum duration-bound events, thus leading to the final choice of truncated normal distributions for sampling random event durations. Simulation studies using truncated normal probability density functions should however carefully define the desired maximum and minimum value parameters. Namely, pre-setting the maximum and minimum value range in the tail regions of the Gaussian curve leads to non-normally distributed random variables and can additionally lead to slow random variates sampling (Botev and L'Ecuyer, 2017).

To account for staff availabilities, simulation software considers working schedules of people involved in running the DNA manufacturing operations. Specification of their working hours and weekdays as well as the simulated manufacturing start date and time is therefore required. The decision on starting a particular manufacturing step is guided by several factors. First, a computational algorithm checks whether a simulated process time and date fall within the working hours. Second, it evaluates whether a given step is likely to finish prior to the end of the working rota by calculating the 95th percentile of its duration distribution. If at least one of these conditions is not satisfied, the algorithm advances in time until both conditions are met. Some biological experiments as well as process steps harnessing automated procedures often do not require human supervision once set up. For instance, incubation of bacterial liquid cultures and media plates (following DNA transfection) can be performed overnight. Therefore, the modelling framework sought to distinguish such process steps by enforcing specification of their supervision requirements (i.e., can a process step run overnight). Consequently, special rules are applied to manufacturing steps, in that the simulation algorithm does not require their termination prior to the end the working schedule. Manufacturing steps to be started on the last working day are a special case for which the algorithm simulates a wait time equal to the sum of the remaining time of the last working day and duration of the non-working week period (e.g., a weekend). Such a modelling rule was imposed to prevent an excessive wait time, which in a real manufacturing setting could, e.g., lead to overgrown microbial cultures.

Following specification of a given process model, simulation algorithms parse the spreadsheet document, generate random durations of manufacturing steps and control their progression in the simulated process run time, so that it conforms with the pre-defined time constraints. A large number of computer simulations must be performed until the distribution of the simulated turnaround times converges. According to the law of large numbers (Sedor, 2015), the distribution profile thus obtained reflects the true probability density function of the turnaround time, under the assumptions of the probabilistic DNA manufacturing model. The number of simulation loops can therefore be specified by the user, based on the output simulation data, to ensure convergence.

The main process simulation tool returns three types of data automatically (**Fig. 5.3, 5.4 and 5.5 and Supp. Fig. 4.1.1 and 4.1.2 in Appendix 4.1**). First, a histogram of the simulated process turnaround times is returned, along with a cumulative distribution plot. A default 90th percentile cutoff is drawn on both plots to indicate process duration with a 90% likelihood of process completion. In addition to these plots, a text file is returned, which stores simulated process durations of every simulation experiment performed to allow for any other custom data processing. The second output includes data describing simulated manufacturing step start and end times. First, a timeline plot is returned, which represents overlaid simulated step durations in process run time. An alternative version of the simulation tool returns a stack plot, which instead represents percentage of simulated process step instances being completed in a particular time interval. A raw data file, describing simulated manufacturing step start and end timestamps, is also returned. The last output comprises plots illustrating all input truncated normal distributions used to model a given process, to provide researchers with a graphical representation of their chosen sampling probability density functions.

Aside from estimating distribution functions of process turnaround times, indicating the most significant process optimisation targets is yet another piece of information which is valuable throughout any process improvement efforts. Therefore, the simulation toolkit includes an additional computational tool which uses output of the

main simulation tool (i.e., the simulated process durations text file) and conducts a one-factor-at-a-time (OFAT) sensitivity analysis to indicate process model parameters (e.g., mean duration of gel electrophoresis) which have the most significant impact on the DNA manufacturing turnaround time. A one-factor-at-a-time method was used to compute sensitivities of a given process model to all of its parameters. One by one, each parameter of the model is thus increased by 1%, other parameters remaining unchanged. The resulting modified model is simulated *n* times, with *n* being the number of simulation loops computed in the initial simulation study. Expected process duration of the original model is later subtracted from the modified model's expected turnaround time, with the mathematical difference indicating how sensitive a given manufacturing process model is to a certain parameter. Results of this analysis are returned as a bar chart depicting sensitivities to the 20 most significant parameters (**Fig. 5.6**).

5.3.2 Modelling of an Example Industrial DNA Fabrication Process

An example Thermo Fisher Scientific DNA manufacturing process was studied to test the function of the Monte Carlo simulation tools (**Figure 5.2** illustrates the process model and the original DNA manufacturing process; also shown in **Supp. Fig. 4.1.6 in Appendix 4.1**). This process builds DNA fragments up to ~3 kb-long, which are later cloned in a plasmid vector. The target constructs are assembled from 2-3 sub-fragments, the size of which ranges from 240 bp to 1.1 kb. The sub-fragment pieces are in turn made from oligonucleotides, which are assembled using a polymerase chain assembly method (**Fig. 5.2 A**). To alleviate DNA synthesis errors, two rounds of sub-fragment and full-length fragment error correction are performed, after which the resulting populations of digested nucleic acid strands are stitched together enzymatically (Sequeira *et al.*, 2016). Following the double DNA correction, full-length DNA fragments are cloned into a plasmid DNA vector using a PCR-based protocol. Once transfected into *E. coli*, 8 bacterial colonies are picked and a colony PCR verification of the plasmid insert region is performed. Two clones exhibiting the correct amplification signal are later further assessed by DNA sequencing. Next-generation sequencing is the predominant DNA validation method. The remaining DNA orders are assessed using Sanger sequencing, due to, e.g.,

customer requirements for Sanger sequencing chromatogram files. Together with the last two DNA verification steps, the construction process harbours several additional control procedures following the gene synthesis, error correction and bacterial transformation manufacturing steps (**Fig. 5.2 and Supp. Fig 4.1.6 in Appendix 4.1**). For each of these control points, a specific rescue path is defined, which indicates alternative manufacturing protocols to be tested and experiments to be repeated in order to salvage failed DNA construction steps.

To build a mathematical model of the example process, probability distributions of its underlying manufacturing steps as well as failure rates at the quality control steps had to be estimated. Mean durations of most of the construction steps, except for the DNA design, oligonucleotide synthesis and DNA sequencing steps, were estimated by the personnel in charge of different experimental stages. Step durations were assumed to follow a normal distribution with maximal and minimal bounds. The remaining parameters of the step duration distributions, i.e., their variances and value limits, were assumed as well. A manufacturing data mining approach was taken to estimate turnaround time distributions and their parameters of the DNA design, oligonucleotide synthesis and DNA sequencing manufacturing stages. Timestamp data from the company's Manufacturing Execution System (MES), which logs start and termination of different manufacturing tasks, was analysed. Manufacturing stage duration distributions were inferred and Monte Carlo simulation experiments were used to recapitulate the observed distributions as well as deduce the underlying manufacturing sub-stages and their failure probabilities. This approach however led to predicted DNA manufacturing times which significantly differed from company's own observation. It was thus concluded that the MES data, which relies on curation by human operators, was not an ideal source of information. Details of this research are described in **Supp. Fig. 4.1.3** and **4.1.4** in **Appendix 4.1**. Thus, in the rest of this chapter durations of the DNA design, oligonucleotide synthesis and DNA sequencing manufacturing stages were assigned arbitrary durations, based on personal experience, although these could easily be replaced by data-based estimations in the future.

Information obtained from these studies was then used to simulate the constructed process model and compare its output with real manufacturing data. Due to significant discrepancies observed between the two manufacturing duration distributions, turnaround time distributions of the DNA design, oligonucleotide synthesis and DNA sequencing manufacturing steps, and their parameters were assumed instead (**Supp. Fig. 4.1.3 and 4.1.4 in Appendix 4.1 and Supp. Table 4.2.2 in Appendix 4.2**).

Failure rates at the individual biological sample control steps were likewise retrieved from the MES system. For the DNA sub-fragment gene synthesis, the obtained failure rate originated from the product of the underlying single sub-fragment construction failure probabilities. For the multiple bacterial clone assessment steps, on the other hand, i.e., the colony PCR and DNA sequencing steps, the acquired success rates referred to the likelihood of obtaining a sufficient number of correct bacterial clones (colony PCR – at least two, DNA sequencing – at least one). Upon manufacturing failure, rescue routes have to be initialised. The process model simplified some of these DNA sample salvage procedures. For instance, duration distributions of analogous sample rescue protocols and their failure probabilities were assumed to be the same as those of the original experiments (**Figure 5.2 and Supp. Fig. 4.1.6 in Appendix 4.1**).

Information on non-working weekdays was obtained from the company's personnel. Staff working schedules, on the other hand, were based on both true (the technicians' schedule) and assumed operational schedules (the IT team schedule). Information on manufacturing steps running overnight was based on communications with company staff, except for the insert sequencing step, which was assumed to run past working hours. As most of the example process DNA sequencing validation is done with NGS, it was also assumed that NGS was the only DNA sequencing procedure used to validate the finished DNA constructs. To model process turnaround times, given the time constraints stated beforehand, the simulation manufacturing start date and time had to be decided as well. The chosen simulation start date and time matched

the start of the first staff schedule and corresponded to the first working day of the week.

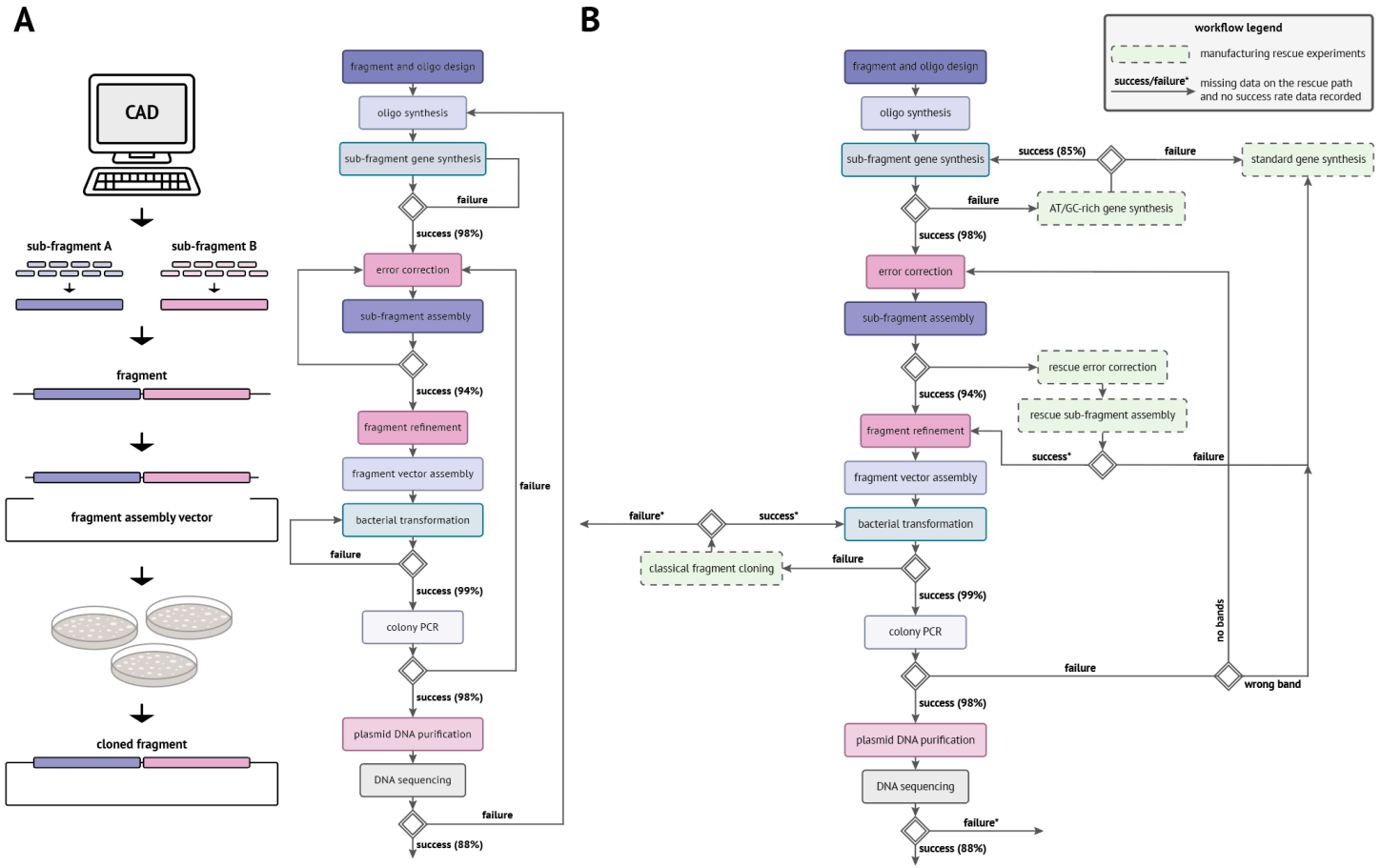


Figure 5.2 Case Study DNA Manufacturing Workflow

(**A**) The example DNA production pipeline manufactures ~3 kb DNA fragments from 2-3 sub-fragment DNA blocks, which are assembled from single-stranded oligonucleotides. Full-length fragment DNA is cloned into a plasmid vector (left panel). Graphical representation of the process model constructed (right panel), depicting all manufacturing rescue paths. (**B**) The original manufacturing process (reference for modelling simplifications / assumptions in **A**).

5.3.3 Proof-of-concept Monte Carlo Simulation Studies

The mathematical model of the example process (**Supp. Table 4.2.1 in Appendix 4.2**) was next subjected to the probabilistic simulation analysis to estimate the expected turnaround time of the DNA construction process as well as to infer the most significant process parameters. Results of the numerical analysis were compared with the observed manufacturing turnaround times (data collected throughout the first 5 months of 2018). However, to do that weekend data had to be first removed from the simulation results, as the real manufacturing data dataset described the manufacturing turnaround times in business days.

The Monte Carlo simulation experiments (1,000 simulation experiments performed) estimated that the expected manufacturing process duration totals 6 days, which equaled the observed mean duration. Both simulated and observed turnaround time distributions were non-Gaussian multimodal distributions with asymmetrical modes, which can be attributed to discrete manufacturing routes of different DNA orders, i.e., including none or different combinations of rescue procedures (**Fig 5.3**). The observed production time distribution however harboured a longer tail, describing probabilities of manufacturing procedures taking the most time to complete. This difference was reflected by 90% quantile values (simulated data – 10 days, observed data – 11 days), which differed by 1 day. Presence of distribution tails in both of the investigated datasets can be attributed to the elevated failure likelihood of the final DNA sequencing step (i.e., 12%), which corresponds to the success rate of at least 1 correct construct out of 2 and possibly stems from the fact that none of the preceding verification procedures assess correctness of the DNA sequences at a single nucleotide resolution (i.e., mainly gel electrophoresis assessments). In the future, failure rate of the last validation step could be alleviated by introduction of intermediate DNA sequencing steps or by sequencing more DNA constructs.

Two types of timeline plots were generated to illustrate completion of the simulated manufacturing steps in time (**Fig 5.4 and 5.5**). Both graphical representations indicated simulation instances which included execution of the manufacturing rescue

paths. Both graphs also reflected the imposed time constraints, i.e., the simulated DNA fabrication did not take place on weekends and run beyond the working hours, except for the overnight manufacturing steps, which ran past the staff working schedule.

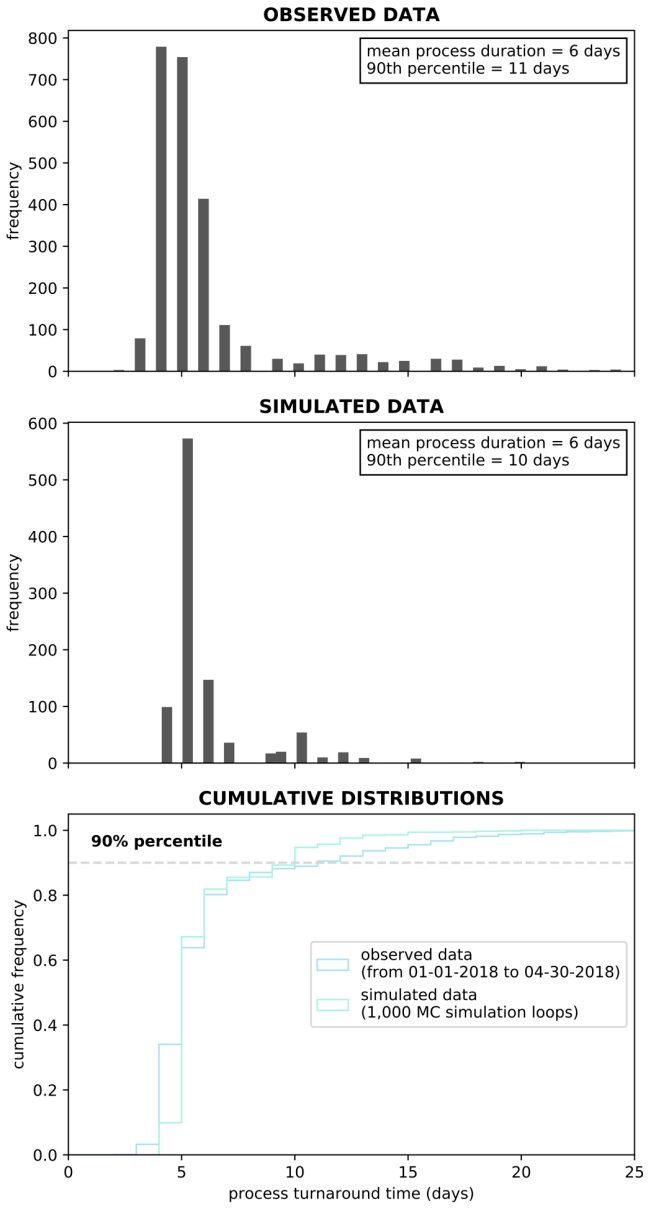


Figure 5.3 Comparison of the Simulated Process Durations and the Observed Duration Data

Observed and simulated process durations datasets are illustrated with histogram and cumulative distribution plots. 1,000 Monte Carlo simulation experiments were carried out to obtain the simulated process durations dataset. The simulated and observed mean process turnaround times (in days) are equal. Minor discrepancies were observed between the 90th percentiles, i.e., 10 and 11 days, respectively, and reflected a lighter distribution tail in the simulated dataset. Both of the investigated distributions harboured several modes, or maxima, which could have indicated different DNA manufacturing rescue profiles.

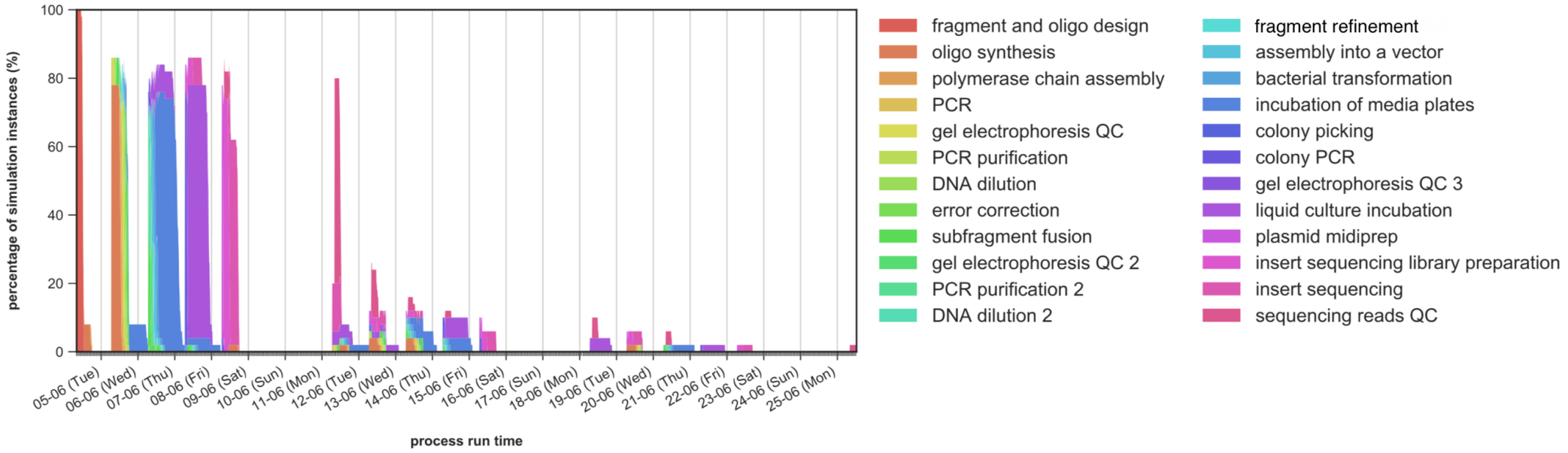


Figure 5.4 Simulation Instances in the Simulated Process Run Time - Stack Plot

For every simulation experiment, start and end times of the consecutive manufacturing steps were recorded. The stack plot illustrates percentage of simulated manufacturing steps taking place at a certain time. No manufacturing steps took place during the pre-defined free weekdays (here, the weekend) and outside of the pre-defined working hours, except for those manufacturing steps which were allowed to run overnight. Hence, the simulated manufacturing processes are compliant with the time constraints set.

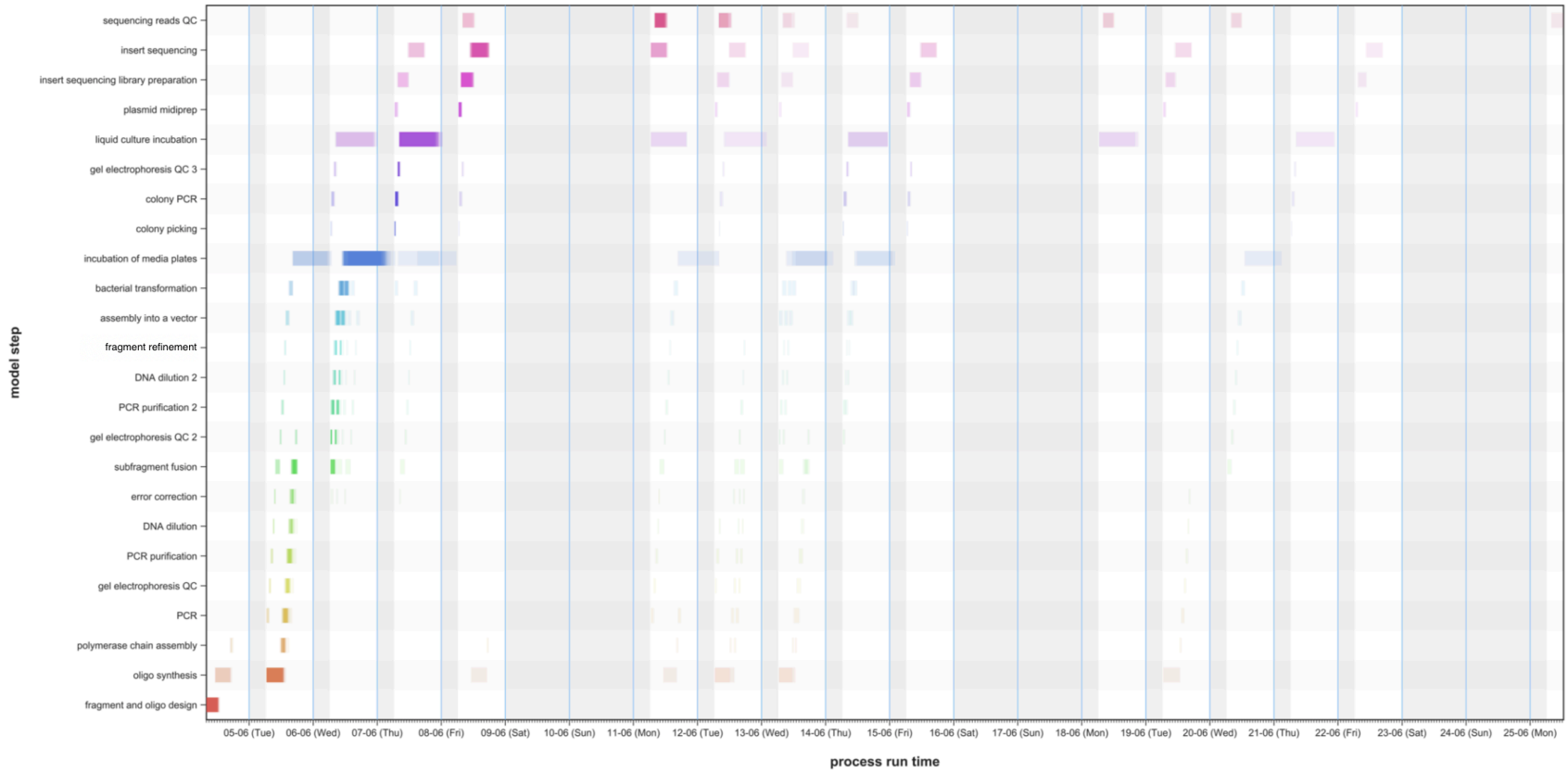


Figure 5.5 Simulation Instances in the Simulated Process Run Time - Timeline Plot

Individual simulated manufacturing step durations were overlaid in the timeline plot. Colour strength of the horizontal step duration bars correlates with the number of simulated steps taking place at a certain simulated process time. Non-working hours and days were shaded light grey. No manufacturing steps took place during the pre-defined free weekdays and outside of the staff working hours, except for the overnight steps, e.g., the incubation of media plates. Therefore, the simulated manufacturing processes conform with the pre-defined time constraints.

To investigate the most significant process model parameters, sensitivities of the process model to all of its parameters were calculated. The sensitivity analysis tool estimated that the most significant model parameter is the minimal duration of plasmid insert sequencing (**Fig 5.6**). Insert sequencing is one of the longest steps (~5 hours duration) in the DNA manufacturing pipeline, therefore it was expected that a slight increase in its duration will have a significant impact on the total manufacturing process duration. However, some of the results were surprising. For instance, considering the oligo synthesis step's duration distribution with small variance, it is surprising that its minimum duration was highlighted as an important parameter instead of its mean, which increase would have moved more probability density towards the longer step durations. More simulation experiments (> 100) should therefore be performed in order to minimise analysis' stochasticity and improve its accuracy. Performing such simulation analysis is expected to take more than 2 hours, using a 2.8 GHz Intel Core i7 processor.

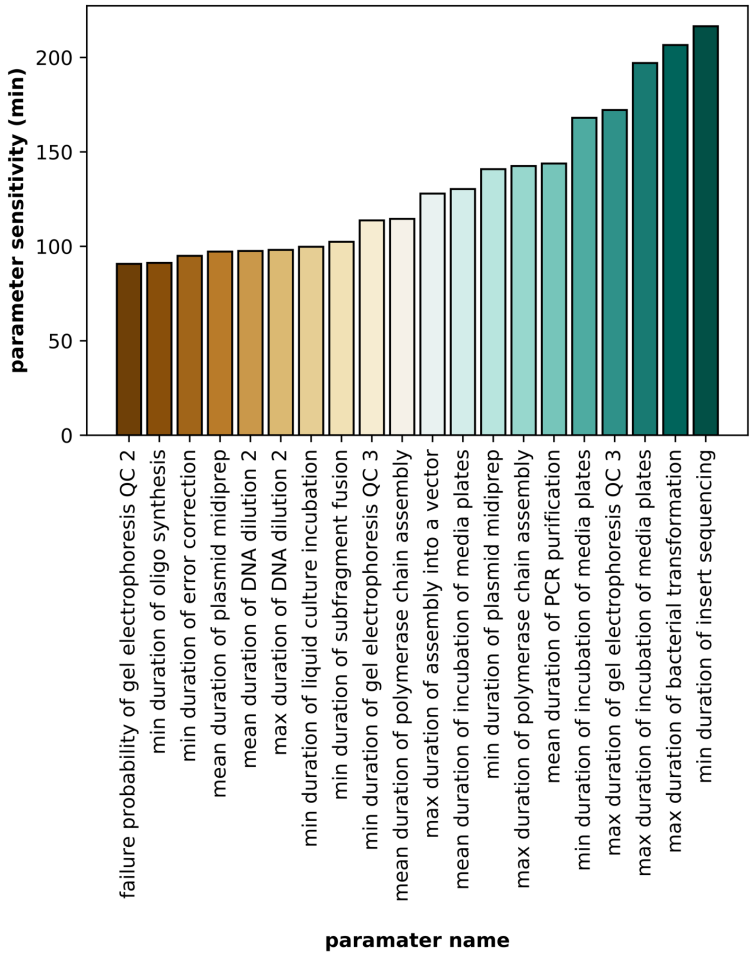


Figure 5.6 Sensitivity Analysis of the Process Model

A one-factor-at-a-time sensitivity analysis was conducted to identify the most significant parameters of the example process model. In every model sensitivity experiment, one of the model parameters (e.g., mean duration of oligo synthesis) was increased by 1% and the resulting new process model was simulated the same number of times as the original model. Expected process turnaround time of the initial model was next subtracted from the expected process duration of the new model, yielding a measure of model's sensitivity to a given parameter (i.e., the parameter sensitivity). Model sensitivities to 20 most significant parameters are illustrated by the bar chart and their calculations were based on a hundred simulation experiments. According to the analysis performed, minimal duration of insert sequencing is the most significant model parameter.

5.3.4 Classification Model for Predicting Sequence-specific Turnaround Times

As mentioned earlier, several nucleic acid sequence characteristics are known to cause problems throughout the DNA manufacturing process (Oberortner *et al*., 2016). Thus, the second goal of the project was to build a predictive tool capable of estimating manufacturing failure probabilities of different DNA sequences based on the nucleic acid sequence features which are known to have a negative impact on the DNA fabrication failure rates. To accomplish that, data mining studies were conducted to find such problematic DNA sequence characteristics. Information obtained from these studies was later used to build a supervised machine learning model able to predict DNA sequence-specific manufacturing failure probabilities. Output of this model was also harnessed to estimate DNA sequence-specific turnaround times of the example manufacturing process (**Table 5.1**)

The first data mining attempts used a DNA sequence dataset obtained from an academic project aiming at PCR amplification and cloning of standardised yeast transcription unit DNA parts, using Golden Gate assembly (Guo *et al*., 2015). However, no significant difference was observed between the successful and failed DNA sequence datasets when impact of several nucleic acid sequence features on DNA amplification was studied (e.g., impact of the amplicon minimal free energies) (**Supp. Fig. 4.1.5 in Appendix 4.1**). This data was also not directly relatable to the example manufacturing process.

Therefore, a more relevant industrial dataset was acquired, which comprised manufacturing data of GeneArt Strings (from 2017-2018), i.e., one of the Thermo Fisher Scientific synthetic DNA products. As in the example process sub-fragment DNA constructs, GeneArt Strings are built from single-stranded oligonucleotides. Thus, data comprising their gene synthesis failures can be better related to the example manufacturing process and harnessed to predict its DNA sequence-specific sub-fragment construction failure probabilities. The acquired dataset harboured nucleic acid sequence information on ~17,000 DNA orders and their corresponding gene synthesis failure rates. Size of the DNA sequences obtained did not exceed

maximal length of the example process sub-fragments and therefore corresponded to a single round of oligonucleotide assembly. Customer orders were classified either as ‘easy’, ‘difficult’ or ‘very difficult’. The first class referred to orders which never failed the gene synthesis step, while the latter two classes concerned orders which involved at least one gene synthesis failure. Two problematic DNA sequence features were investigated, i.e., the DNA sequence length and its global GC content (Oberortner *et al*., 2016). For both of them, significant differences were observed between the successful and failed DNA construct distributions. Moreover, a ~50% decrease in the gene synthesis success rate was observed for DNA sequences with lengths < 200 bp and DNA fragments with GC contents > 60% (**Fig. 5.7**).

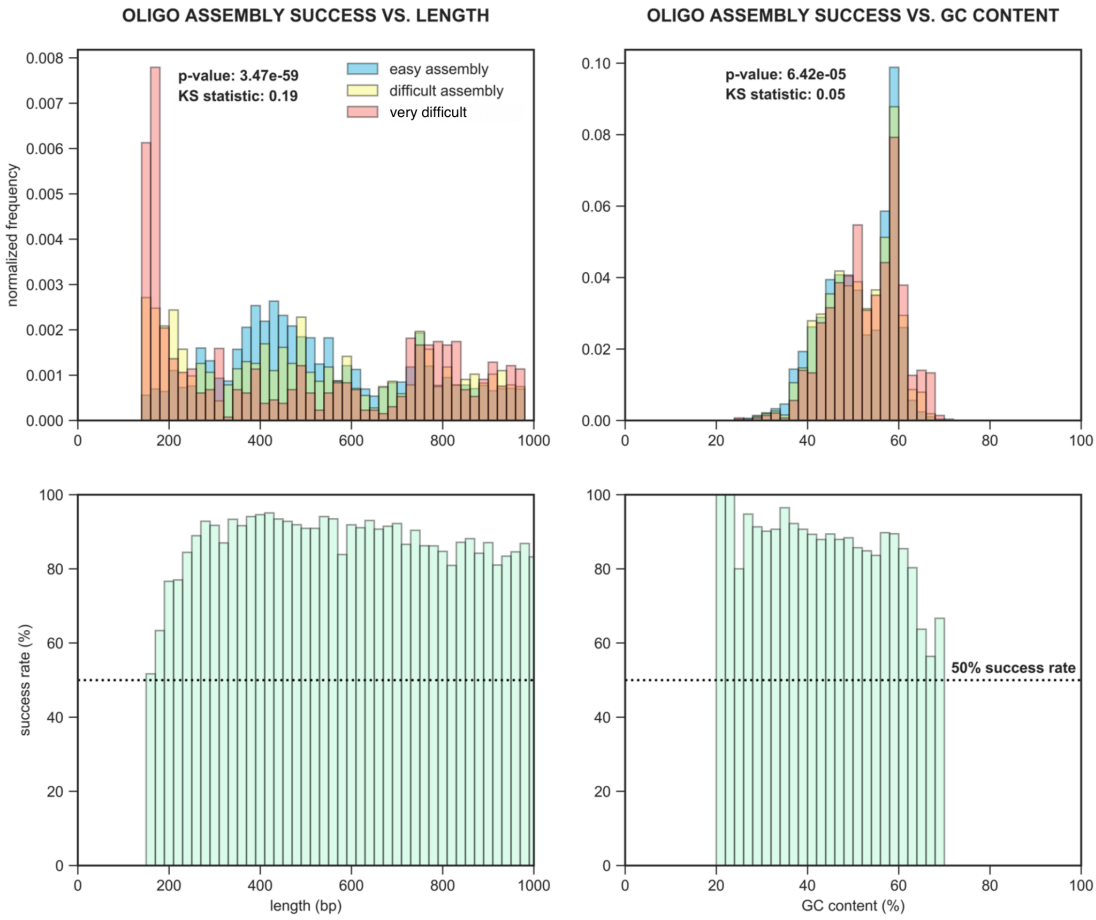


Figure 5.7 Impact of DNA Sequence Features on Failure of Gene Synthesis

Data mining studies were conducted to identify how global GC content and DNA sequence length impact failure of oligonucleotide assemblies. GeneArt Strings manufacturing data was analysed. "Easy" assemblies refer to oligonucleotide assemblies which did not have to be repeated. "Difficult" and "very difficult" assemblies had to be repeated (at least once). Histograms of the "easy", "difficult" and "very difficult" are overlaid. A statistically significant difference (as judged by a two-sample Kolmogorov-Smirnov test) was observed between the successful and failed sample (the "difficult" and "very difficult" datasets) GC content and DNA sequence length data. A decrease in the manufacturing success rate was observed for DNA fragments with short lengths and high GC contents.

The GC content and DNA sequence length data was next used to build a predictive classification machine learning model, predicting gene synthesis success rates (based on these two DNA sequence features). Various machine learning algorithms were considered, e.g., logistic regression, k-nearest neighbours (k-NN) and support vector machine (SVM) algorithms. Despite its ability to generate predictions readily interpretable as class probabilities, the logistic regression algorithm was ruled out early on due to its inability to work well with non-linearly distributed data, necessitating modelling of non-linear decision boundaries. The k-NN algorithm was another alternative, able to model non-linearly distributed data. However, the k-NN algorithm is known to be sensitive to outliers and therefore was not chosen. Kernel SVM algorithms (e.g., the RBF kernel SVM) work well with outliers, and thus do not require any data pre-processing, are able to model non-linear data structures and, additionally, work well with small high-dimensionality datasets, which makes the resulting SVM models scalable in terms of adding additional features to the machine learning analysis (King, Feng and Sutherland, 1995).

A support vector classifier (SVC) was thus trained and used to predict gene synthesis success rates (**Fig. 5.8**). Its optimal parameters (C and gamma) were found using a hyperparameter tuning tool from the scikit-learn Python machine learning package, evaluating all possible combinations of pre-defined model parameters by cross validation. Cross-validation is a standard statistical technique, which evaluates the ability of a given model to infer predictions on one part of the dataset after being trained using its other part. A classification method with a higher cross-validation score is therefore better adapted to the problem at hand and does not suffer from problems such as over- or -under-fitting the data (Kohavi, 1995). The C parameter is a regularisation parameter, which describes the impact of misclassification on the algorithm objective function. For instance, an SVM model with a large C value will try to avoid misclassification at a sacrifice of a smaller data points' distance from the classification hyperplane. A small C value SVM model on the other hand will allow more misclassification to find a hyperplane with the largest minimum margin. The gamma parameter describes the standard deviation of the Gaussian radial basis function (RBF), namely equals to the inverse of its standard deviation. Thus, small

gamma parameter values lead to a large RBF standard deviation, which points the machine learning algorithm towards considering a large number of local data points with similar importance (even if they are far away from each other) when making the local data classification decision. A small RBF standard deviation (large values of the gamma parameter) on the other hand points the RBF kernel SVM algorithm towards considering a small number of local data points, when taking the local classification decision, where two data points are only considered similarly important when close to each other (Syarif, Prugel-Bennett and Wills, 2016).

To evaluate impact of the predicted success probabilities on the expected DNA manufacturing turnaround time estimations, a set of example GC content and length DNA sequence features was used to predict success rates of the example process sub-fragment gene syntheses using the SVC model. In order to obtain success probabilities of all oligonucleotide assemblies combined, involved in a given DNA fragment construction, the underlying sub-fragment success rates were next multiplied and the corresponding failure rates were used in simulating the process model. As anticipated, optimal sub-fragment GC contents and lengths led to the shortest expected turnaround times (accounting for both business and non-business days), while extreme DNA sequence features led to longer expected DNA construction durations (**Table 5.1**).

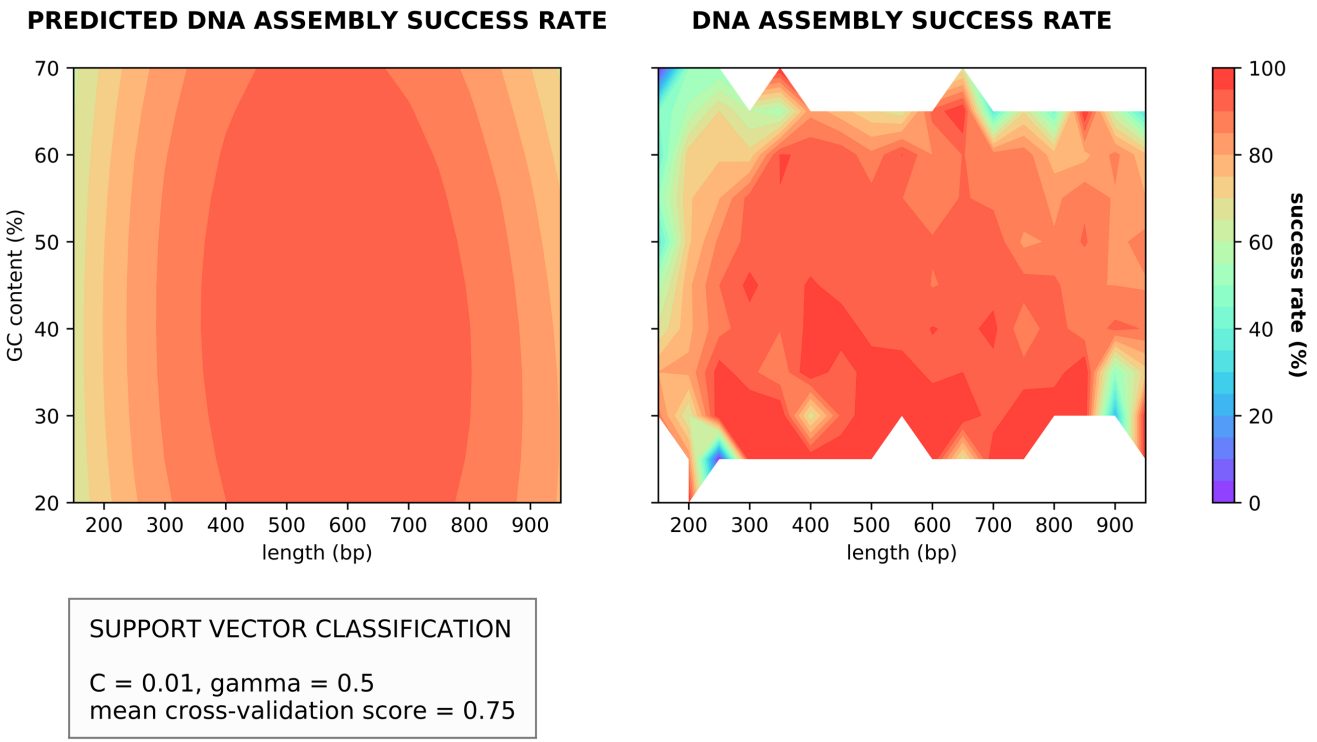


Figure 5.8 Classification Model to Predict Success Probability of Gene Synthesis

The GeneArt Strings GC content and DNA sequence length data was used to train a predictive model, using a support vector machine (SVM) algorithm. The constructed model was able to predict success probabilities of oligonucleotide assemblies, based on the underlying DNA GC content and length. GeneArt Strings oligonucleotide assemblies follow an analogous experimental procedure to the example process sub-fragment assemblies and therefore can be related to the process case study. A hyperparameter tuning computational tool was used to choose the best classification model (i.e., with the highest mean cross-validation score). Hyperparameters (C and gamma) and the cross-validation score of the best machine learning model are indicated. Predicted (left plot) and observed (right plot) DNA assembly success rates are illustrated with heat maps.

Table 5.1 DNA Sequence-specific Manufacturing Turnaround Time Estimations

fragment ID	sub-fragment A GC content and length	sub-fragment B GC content and length	sub-fragment A probability of success	sub-fragment B probability of success	fragment probability of success	expected manufacturing turnaround time
fragment 1	40%, 500 bp	40%, 500 bp	0.940	0.940	0.884	7 days 23 hours
fragment 2	70%, 500 bp	70%, 500 bp	0.909	0.909	0.826	8 days 9 hours
fragment 3	70%, 500 bp	40%, 150 bp	0.909	0.673	0.612	8 days 16 hours
fragment 4	70%, 150 bp	40%, 150 bp	0.638	0.673	0.429	8 days 20 hours
fragment 5	70%, 150 bp	70%, 150 bp	0.638	0.638	0.407	9 days 3 hours
simulations w/o DNA information	N/A	N/A	N/A	N/A	0.980	7 days 9 hours

5.4 Discussion

The developed Monte Carlo simulation tools are capable of accurately estimating DNA manufacturing process turnaround times, with results of the example simulation studies being comparable to the observed production durations (although minor discrepancies between the simulated and real data were also present). Furthermore, the software developed is capable of estimating the most significant manufacturing parameters, which can be expected to be the most promising process optimisation targets. Methods to conduct DNA sequence-specific simulation studies were also developed and allowed computation of customised gene synthesis success rate predictions, which were later used to estimate DNA sequence-specific manufacturing turnaround time distributions. Several modelling methods could however be improved and some limitations of the analyses performed have to be considered when interpreting results of the simulation studies.

First, construction of the example manufacturing process model was based on estimating and assuming the underlying manufacturing step distribution functions and their parameters based on the information obtained from the company's personnel. However, data describing real manufacturing step durations would have been more supportive in defining the input distribution functions. I tried to obtain such data by mining the manufacturing step start and end timestamp records (stored in the company's MES system). However, these attempts did not lead to improved manufacturing turnaround time estimations, possibly due to data artefacts present in the timestamp datasets (**Supp. Fig. 4.1.3 in Appendix 4.1**). More accurate variance estimations could have also been obtained, e.g., by collecting additional 10% and 90% quantile estimations from the staff, which can be used to calculate standard deviations, given known distribution means (Cook, 2010).

Second, simplification of the biological sample rescue manufacturing routes in the constructed statistical model could have contributed to the observed minor data discrepancies. For example, some manufacturing rescue paths of the example DNA fabrication process involved different experimental protocols (**Supp. Fig. 4.1.6 in**

Appendix 4.1), durations of which could have differed from those of the original procedures. However, such rescue scenarios were not accounted for in the example process model. The simulation tools also do not support modelling of multiple rescue options, which are possible in the example manufacturing process (i.e., the colony PCR step), and therefore sample control steps involving these had to have their rescue paths simplified.

Level of modelling detail could have also had an impact on the simulation results. In principle, manufacturing steps can be decomposed into a large number of sub-steps. Therefore, decisions regarding how granular a particular process model is are subjective and often guided by a certain level of detail required to assess some possible targets for process optimisation. However, despite the freedom to choose the required level of detail, some manufacturing stages can be difficult to decompose into more fine-grained steps. In this project, the oligonucleotide and DNA fragment design step posed such difficulties. Modelling duration of the design procedure was needed to benchmark simulation results against the observed data, referring to complete customer DNA order lifecycles, i.e., from the time of the order placement until the time of its shipment. However, deconvolution of the DNA design steps exclusive to the example DNA production process was challenging, as the design procedure involves simultaneous evaluation of multiple possible manufacturing routes to meet various customer, manufacturing and DNA sequence constraints. Therefore, modelling assumptions were made regarding the DNA sequence design step.

Modelling tools developed by this project allowed certain properties to be assigned to the manufacturing steps (e.g., the “overnight” attribute). However, an additional attribute could have been considered to make it possible to “chain” some of the consecutive manufacturing steps. Such a step characteristic could, e.g., prevent setting up bacterial transformations, if there is not enough time to plate the transfected cells, while maintaining the desired level of modelling detail. Furthermore, it could be used to investigate the impact of automation and integration of manufacturing steps.

Biological sample batching is another important aspect of DNA manufacturing. In high-throughput DNA manufacturing facilities DNA fabrication often takes place in large sample batches, comprising 96 samples or more, which makes it difficult to model such processes with single sample resolution. Such resolution could be useful, e.g., in conducting accurate manufacturing cost simulation studies, which account for repetitions of fractions of sample batches only. However, the modelling approach presented in this manuscript does not support investigation of such scenarios.

The simulation software developed in this project also does not account for manufacturing steps which happen in parallel. While this limitation does not pose major issues for simulating manufacturing turnaround times of the example manufacturing process, it makes it difficult to accurately model costs of the sub-fragment gene synthesis manufacturing stage. For instance, for three example parallel sub-fragment gene synthesis manufacturing steps, A, B and C, for which manufacturing costs are independent and distributed identically, however, which gene synthesis success rates differ, example repetition of step A only is associated with approximately one third of the costs of running the parallel A, B, C sub-process. In the case of modelling manufacturing step durations, duration of step A repetition could be similar to the parallel A, B, C sub-process duration, which makes modelling of production turnaround times easier as compared to the simulation analysis of manufacturing costs.

Last but not least, incorporation of the DNA sequence information into the process simulation studies sought to make the manufacturing turnaround time estimations more relevant to the production of complex nucleic acid sequences. A classification model was therefore built to tailor gene synthesis failure probabilities to specific sets of sub-fragment DNA sequences and accounted for two problematic DNA sequence features, i.e., extreme GC content and DNA sequence length. Nevertheless, information about several additional DNA sequence features could have been used to train the model in an effort to improve its predictive power. For instance, DNA repeats are known to pose problems throughout the gene synthesis process and

therefore information about their abundance could have improved accuracy of the predictions made (Oberortner *et al*., 2016). Moreover, information about the DNA sequence length could have been used to improve modelling of the DNA amplification steps, duration of which is DNA sequence length-dependent and would therefore become deterministic. This would lead to less randomness in the simulation data and therefore less variance in the predicted turnaround times, yielding more definitive predictions after shorter simulation times.

Chapter 6
Conclusion

Scientific discovery, technological innovation and development of rapid prototyping methodologies are three factors which drive the progress of engineering disciplines. For instance, invention of the steam engine or the power loom are examples of technological innovations which sparked the first Industrial Revolution, leading to mechanisation of various manufacturing processes (Ó Gráda, 2016). More recently, further standardisation and automation of different operations have been making it possible to rapidly *Design*, *Build* and *Test* mechanical and electrical systems using standardised engineering components, computer-aided design (CAD) and rapid prototyping methods (e.g., 3D printing) (Deming, 2000; Gross *et al*., 2014). The latest advances in machine learning and data science on the other hand have been providing means to additionally *Learn* from the consecutive *Design*, *Build* and *Test* cycle iterations, thus allowing researchers to capitalise on the knowledge gained throughout the engineering process (LeCun, Bengio and Hinton, 2015).

Today, biology is maturing into an engineering discipline thanks to the same enabling factors. Namely, the improving DNA *reading*, *writing* and *editing* technologies and laboratory automation solutions are the main driving forces of this transition, facilitating rapid construction and prototyping of novel biological systems (Check Hayden, 2014; Kosuri and Church, 2014; Libbrecht and Noble, 2015; Chao *et al*., 2017; Shendure *et al*., 2017; Clore, 2018). This doctoral thesis project therefore aimed at harnessing this technological progress to further assist in accelerating biological research and rationalizing function of living organisms.

In order to do that, work presented in this thesis sought to address several remaining challenges which are limiting aspects of the genetic engineering *Design*, *Build*, *Test* and *Learn* cycle: first, the demand for cheaper and faster DNA fabrication; second, the need for efficient genome-scale DNA editing tools for rapid screening of engineered microbes; third, the need for predictive methodologies able to improve planning of high-throughput DNA manufacturing.

In the first results chapter of this thesis, application of an automated acoustic dispensing technology was proposed. The experimental work presented showed that

it is possible to assemble DNA in volumes orders of magnitude lower than previously, leading to substantial reductions in reagent costs. Currently, a growing number of high-throughput industrial and academic facilities harnesses acoustic dispensing robots to miniaturise their DNA manufacturing assays (Shapland *et al*., 2015; Chao *et al*., 2017). Moreover, since its publication, the miniaturised DNA fabrication approach developed has been already adopted by automated biofoundries, seeking to increase their DNA manufacturing capacities in a cost-efficient manner (Johnson *et al*., 2016; Chao *et al*., 2017). One limitation of applying acoustic dispensing in DNA fabrication is however the high initial cost of purchasing an acoustic liquid handler, which makes this type of technology accessible mostly to larger synthetic DNA manufacturing facilities rather than individual research laboratories. Droplet microfluidics provides an alternative solution to miniaturisation of biological assays and is becoming more accessible to researchers as 3D printing services and devices are becoming commodities (Gach *et al*., 2017). To date, Gibson and Golden Gate assembly protocols have been downscaled using this technology. For instance, Patrick *et al*. miniaturised a two-fragment Golden Gate reaction down to 490 nL, while Khilko *et al*. assembled 12 oligonucleotides using 600 nL Gibson assembly and performed an additional nanolitre DNA error correction reaction (Patrick *et al*., 2015; Khilko *et al*., 2018). Design and operation of microfluidic instruments however still requires relevant expertise and therefore this technology has not been yet widely applied by individual research laboratories (Gach *et al*., 2017).

In the second results chapter, development of a genetic tool for rapid prototyping of wild-type and synthetic *S. cerevisiae* strains was described. This section detailed efforts to build a genome-scale gene deletion library based on the state-of-the-art CRISPR DNA editing technology. Completion of this work holds promise for functional genetic screens, in search for essential gene sets ensuring yeast survival under different environmental conditions. Moreover, application of the proposed gene deletion approach to synthetic budding yeast strains might facilitate identification of genetic defects as well as novel genetic interactions caused by refactoring of their genetic code (Shen *et al*., 2017).

Last but not least, in the last results chapter, probabilistic Monte Carlo simulation software was developed to estimate speed of DNA manufacturing, and thus help plan DNA fabrication in high-throughput DNA synthesis facilities. Benchmarking the DNA order turnaround time predictions against real manufacturing data from a commercial DNA provider revealed accurate estimates of DNA production times. Moreover, a machine learning classification method was established to estimate nucleic acid sequence-specific customer order turnaround times which accounted for difficulties in synthesizing DNA fragments with extreme GC contents and lengths (Oberortner *et al.*, 2016). The computational studies presented therefore demonstrate the potential of business intelligence strategies to be used for analysis of high-throughput DNA production pipelines, e.g., those of automated biofoundries.

In conclusion, results of this project show that current limitations of the synthetic biology *Design*, *Build*, *Test* and *Learn* cycle can be tackled with a range of interdisciplinary approaches, including methodologies derived from industrial process engineering (Mourtzis, Doukas and Bernidaki, 2014). Application of automation technologies constitutes the common denominator of the work presented in this thesis manuscript, as these technologies have proven their potential in establishing robust and reliable high-throughput DNA synthesis procedures. However, this thesis also stresses the importance of harnessing high-throughput data for *learning* purposes, using computer-aided manufacturing software to collect manufacturing data and predictive analytics to predict unknown future events.

Appendix 1

Appendix 1.1 Supplementary Tables

Supp. Table 1.1.1 PCR Primers

ID	direction	DNA sequence (5' – 3')	application	reference DNA template
YCp2391*	forward	GAGATCCAGTTCGATGTAACC	amplification of the nanolitre Gibson Assembly fragment (1)	pPC025*
YCp2392*	reverse	AGGATGTCCCAAGCGAAC	amplification of the nanolitre Gibson Assembly fragment (1)	pPC025*
YCp2393*	forward	TTACCAAAGGTGGTCCGCTG	amplification of the nanolitre Gibson Assembly fragment (2)	pPC025*
YCp2394*	reverse	TCAGTTGGGTGCACGAGTG	amplification of the nanolitre Gibson Assembly fragment (2)	pPC025*
YCp2395*	forward	AGCGTGGGTCTCGGGCTACTAGTAGTTGATCTAATTATGGAATACC	*pMBP1* promoter part amplification	BY4741 genomic DNA
YCp2396*	reverse	GTGCTGGGTCTCACATCGCTTGTGTTTCTGGGATTTACGTTGTGTC	*pMBP1* promoter part amplification	BY4741 genomic DNA
YCp2214*	forward	GATCCTTTGATTTTCTACCG	nanolitre PCR experiments	HcKan_P*
YCp2215*	reverse	CTCGATAACTCAAAAAATACG	nanolitre PCR experiments	HcKan_P
PKp001	forward	TGCGCATGTTTCGGCGTTCGAAACTTCTCCGCAGTGAAAGATAAATGATC	DNA insert cloning into the CRISPR system vector	pGZ110 (cloned in PKe063*)
PKp002	reverse	GTTGATAACGGACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC	DNA insert cloning into the CRISPR system vector	pGZ110 (cloned in PKe063*)
PKp003	forward	AGCAACTCCAATGACCACG	Sanger sequencing verification of *ADE2* gene mutations	BY4741 genomic material

ID	direction	DNA sequence (5’ – 3’)	application	reference DNA template
PKp004	reverse	GTGACGCAAGCATCAATGG	Sanger sequencing verification of *ADE2* gene mutations	BY4741 genomic material
P5	forward	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT	Illumina NGS sequencing	DNA w/ Illumina primer-binding sites
P7-B1	reverse	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCT	Illumina NGS sequencing (primer w/ a sequencing index)	DNA w/ Illumina primer-binding sites
P7-B2	reverse	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCT	Illumina NGS sequencing (primer w/ a sequencing index)	DNA w/ Illumina primer-binding sites
P7-B3	reverse	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCT	Illumina NGS sequencing (primer w/ a sequencing index)	DNA w/ Illumina primer-binding sites
P7-B4	reverse	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCT	Illumina NGS sequencing (primer w/ a sequencing index)	DNA w/ Illumina primer-binding sites
P7-B5	reverse	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCT	Illumina NGS sequencing (primer w/ a sequencing index)	DNA w/ Illumina primer-binding sites
P7-B6	reverse	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCT	Illumina NGS sequencing (primer w/ a sequencing index)	DNA w/ Illumina primer-binding sites
P7-B7	reverse	CAAGCAGAAGACGGCATACGAGATCAGATCGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCT	Illumina NGS sequencing (primer w/ a sequencing index)	DNA w/ Illumina primer-binding sites
P7-B8	reverse	CAAGCAGAAGACGGCATACGAGATTAGCTTGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCT	Illumina NGS sequencing (primer w/ a sequencing index)	DNA w/ Illumina primer-binding sites
P7-B9	reverse	CAAGCAGAAGACGGCATACGAGATGATCAGGTGACTGGAGTTCAGACGTGTGCTCTT CCGATCT	Illumina NGS sequencing (primer w/ a sequencing index)	DNA w/ Illumina primer-binding sites

ID	direction	DNA sequence (5’ – 3’)	application	reference DNA template
P7-B10	reverse	CAAGCAGAAGACGGCATACGAGATATCACGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT	Illumina NGS sequencing (primer w/ a sequencing index)	DNA w/ Illumina primer-binding sites
PKp005	forward	AGCAACTCCAATGACCACG	amplification of the barcoded genomic frag.	BY4741 genomic material
PKp006	reverse	GTGACGCAAGCATCAATGG	amplification of the barcoded genomic frag.	BY4741 genomic material
PKp007	forward	AGATCGGAAGAGCACACGTCTG	biotinylated ssDNA elongation	barcoded BY4741 genomic material
PKp008	reverse	GCGTCGTGTAGGATCCCAGACGTGTG	generation of a dsDNA BamHI restriction site	barcoded BY4741 genomic material

*more information in the Cai Lab database

Supp. Table 1.1.2 Nanolitre Gibson DNA Assemblies

reagent	volume (Echo 550)	volume (Echo 550)	volume (Echo 550)	volume (Echo 550)	volume (manual addition)
master mix	37.5 nL	187.5 nL	375.0 nL	750 nL	15,000 nL
fragment 1 (113.8 ng/μL)	5 nL	20 nL	40 nL	80 nL	2,500 nL
fragment 2 (86 ng/μL)	7.5 nL	42.5 nL	85 nL	170 nL	2,500 nL
total volume	**50 nL**	**250 nL**	**500 nL**	**1,000 nL**	**20,000 nL**

Supp. Table 1.1.3 Nanolitre Golden Gate DNA Assemblies

reagent	volume (Echo 550)	volume (Echo 550)	volume (Echo 550)	volume (Echo 550)	volume (manual addition)
master mix	17.5 nL	82.5 nL	167.5 nL	332.5 nL	2,500 nL
pMBP1 (20 ng/μL)	30 nL	150 nL	300 nL	600 nL	4,500 nL
HcKan_P (10 ng/μL)	2.5 nL	17.5 nL	32.5 nL	67.5 nL	500 nL
total volume	**50 nL**	**250 nL**	**500 nL**	**1,000 nL**	**7,500 nL**

Supp. Table 1.1.4 Nanolitre PCR Reactions

reagent	volume (Echo)	volume (Echo)	volume (Echo)	volume (Echo)	volume (Echo)	volume (manual addition)
10 μM primer (YCp2214)	2.5 nL	12.5 nL	25 nL	37.5 nL	50 nL	500 nL
10 μM primer (YCp2215)	2.5 nL	12.5 nL	25 nL	37.5 nL	50 nL	500 nL
DNA template	5 nL	25 nL	50 nL	75 nL	100 nL	1,000 nL
GoTaq Green Master Mix	25 nL	125 nL	250 nL	375 nL	500 nL	5,000 nL
nuclease-free water	15 nL	75 nL	150 nL	225 nL	300 nL	3,000 nL
total volume	**50 nL**	**250 nL**	**500 nL**	**750 nL**	**1,000 nL**	**10,000 nL**

Appendix 2

Appendix 2.1 Supplementary Tables

Supp. Table 2.1.1 DNA Assembly Reagent Costs

reagent name	supplier	catalogue number	units	pack size	pack price (£)	price/unit (£)
UltraPure 1 M Tris-HCI Buffer, pH 7.5	Life Technologies	15567027	µL	1,000,000	47.38	0.000047
Magnesium chloride solution, for molecular biology, 1.00 M±0.01 M	Sigma-Aldrich	M1028-100ML	µL	100,000	50.1	0.000501
Deoxynucleotide Mix, PCR-Grade, 400 µL	Agilent Technologies	200415	µL	400	125.1	0.312750
DL-Dithiothreitol, for molecular biology, =98% (TLC), =99% (titration)	Sigma-Aldrich	D9779-5G	grams	5	86.77	17.3540
Poly(ethylene glycol), BioUltra, 8,000	Sigma-Aldrich	89510-250G-F	grams	250	17.9	0.071600
ß-Nicotinamide adenine dinucleotide hydrate, =99%	Sigma-Aldrich	N1511-1G	grams	1	85.1	85.1000
T5 Exonuclease	NEB	M0363L	units	5,000	163.2	0.03264
Phusion High-Fidelity DNA Polymerase	NEB	M0530L	units	500	250.4	0.50080
Thermus Aquaticus (Taq) DNA Ligase	NEB	M0208L	units	10,000	195.2	0.01952
T4 DNA Ligase	NEB	M0202M	units	100,000	166.4	0.00166
BsaI-HF	NEB	R3535L	units	5,000	169.6	0.03392
BSA, Molecular Biology Grade	NEB	B9000S	mg	12	15.2	1.26667

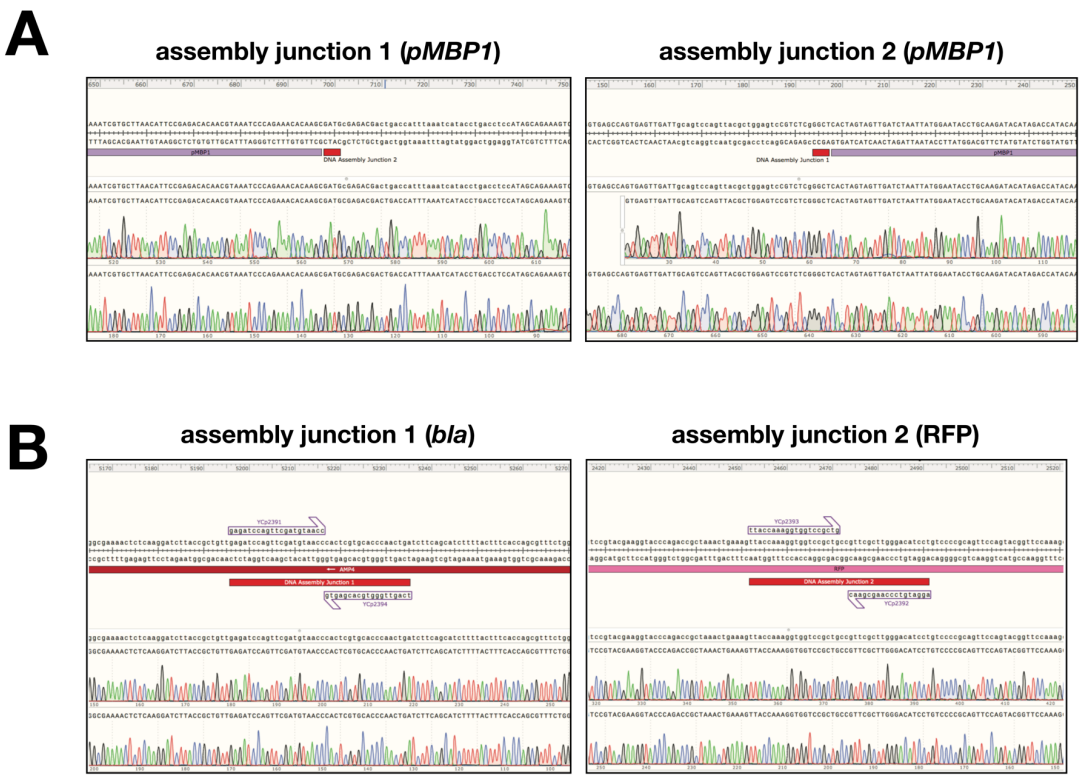
Supp. Table 2.1.2 DNA Assembly Reagent Costs - Gibson Method

reagent name	reaction volume (nL)	reagent cost (£)
UltraPure 1 M Tris-HCl Buffer, pH 7.5	50	2.369×10^{-7}
	250	1.1845×10^{-6}
	500	0.000002369
	1,000	0.000004738
	20,000 (positive control)	0.00009476
Magnesium chloride solution, for molecular biology, 1.00 M±0.01 M	50	2.505×10^{-7}
	250	1.2525×10^{-6}
	500	0.000002505
	1,000	0.00000501
	20,000 (positive control)	0.0001002
Deoxynucleotide Mix, PCR-Grade, 400 µL	50	0.0001251
	250	0.0006255
	500	0.001251
	1,000	0.002502
	20,000 (positive control)	0.05004
DL-Dithiothreitol, for molecular biology, =98% (TLC), =99% (titration)	50	6.68997×10^{6}
	250	3.34498×10^{-5}
	500	6.68997×10^{-5}
	1,000	0.000133799
	20,000 (positive control)	0.002675987
Poly(ethylene glycol), BioUltra, 8,000	50	0.000000179
	250	0.000000895
	500	0.00000179
	1,000	0.00000358
	20,000 (positive control)	0.0000716
ß-Nicotinamide adenine dinucleotide hydrate, =99%	50	2.82715×10^{-6}
	250	1.41357×10^{-5}
	500	2.82715×10^{-5}
	1,000	5.6543×10^{-5}
	20,000 (positive control)	0.00113086
T5 Exonuclease	50	0.000006528
	250	0.00003264
	500	0.00006528
	1,000	0.00013056
	20,000 (positive control)	0.0026112
Phusion High-Fidelity DNA Polymerase	50	0.000626
	250	0.00313
	500	0.00626
	1,000	0.01252
	20,000 (positive control)	0.2504
Thermus Aquaticus (Taq) DNA Ligase	50	0.003904
	250	0.01952
	500	0.03904
	1,000	0.07808
	20,000 (positive control)	1.5616
total reagent costs (£)	50	0.004671812
	250	0.023359058
	500	0.046718115
	1,000	0.09343623
	20,000 (positive control)	1.868724607

Supp. Table 2.1.3 DNA Assembly Reagent Costs – Golden Gate Method

reagent name	reaction volume (nL)	reagent cost (£)
T4 DNA Ligase	50	0.011093333
	250	0.055466667
	500	0.110933333
	1,000	0.221866667
	7,500 (positive control)	1.664
BsaI-HF	50	0.000565333
	250	0.002826667
	500	0.005653333
	1,000	0.011306667
	7,500 (positive control)	0.0848
BSA, Molecular Biology Grade	50	6.02933×10^{-5}
	250	0.000301467
	500	0.000602933
	1,000	0.001205867
	7,500 (positive control)	0.009044
total reagent costs (£)	50	0.01171896
	250	0.0585948
	500	0.1171896
	1,000	0.2343792
	7,500 (positive control)	1.757844

Appendix 2.2 Supplementary Figures



Supp. Figure 2.2.1 Sanger Sequencing Verification of DNA Assembly Junctions

(**A**) DNA junctions of Golden Gate assemblies; There are two junctions, upstream and downstream of the cloned *pMBP1* yeast promoter. (**B**) DNA junctions of Gibson assemblies; There are two junctions, one inside the *bla* gene open reading frame (conferring resistance to ampicillin and carbenicillin antibiotics) and the second one inside the red fluorescent protein (RFP) reading frame, allowing bacterial clone antibiotic selection and colorimetric screening, respectively.

Appendix 2.3 Supplementary Calculations

This section details computations of the number of expected experimental trials required until a successful DNA assembly. Calculations of the expected DNA fabrication reagent costs are subsequently shown.

The number of DNA fabrication trials until a DNA construct is made is a geometrically distributed random variable K, which expected value equals

$$E(K) = \frac{1}{p}$$

For both Gibson and Golden Gate assemblies, the observed success rate of DNA assembly was 1 successful DNA fabrication experiment (at least 1 correct clone obtained) out of 3. Hence,

$$p = \frac{1}{3}$$

The expected reagent costs of Golden Gate and Gibson assemblies are a multiple of the number of expected experimental trials and reagents costs per one reaction.

Therefore, the expected reagent costs of a 50 nL Golden Gate DNA fabrication equal

$$E(K_{Golden\ Gate\ method}) = \frac{1}{1/3} \times 1 \text{ British pence} = 3 \text{ British pence}$$

while the expected reagent costs of a 250 nL Gibson DNA fabrication equal

$$E(K_{Gibson\ method}) = \frac{1}{1/3} \times 2 \text{ British pence} = 6 \text{ British pence}$$

Appendix 2.4 Published Work

Journal of Laboratory Automation
2016, Vol. 21(1) 49–56

DOI: 10.1177/2211068215593754
jala.sagepub.com
SAGE

Smart DNA Fabrication Using Sound Waves: Applying Acoustic Dispensing Technologies to Synthetic Biology

Paulina Kanigowska[1†], Yue Shen[1,2†], Yijing Zheng[1], Susan Rosser[1], and Yizhi Cai[1]

Abstract

Acoustic droplet ejection (ADE) technology uses focused acoustic energy to transfer nanoliter-scale liquid droplets with high precision and accuracy. This noncontact, tipless, low-volume dispensing technology minimizes the possibility of cross-contamination and potentially reduces the costs of reagents and consumables. To date, acoustic dispensers have mainly been used in screening libraries of compounds. In this paper, we describe the first application of this powerful technology to the rapidly developing field of synthetic biology, for DNA synthesis and assembly at the nanoliter scale using a Labcyte Echo 550 acoustic dispenser. We were able to successfully downscale PCRs and the popular one-pot DNA assembly methods, Golden Gate and Gibson assemblies, from the microliter to the nanoliter scale with high assembly efficiency, which effectively cut the reagent cost by 20- to 100-fold. We envision that acoustic dispensing will become an instrumental technology in synthetic biology, in particular in the era of DNA foundries.

Keywords

synthetic biology, DNA assembly, acoustic dispensing

Introduction

Synthetic biology is a nascent interdisciplinary research field that leverages rational design approaches based on engineering principles.[1,2] Synthetic biology distinguishes itself from traditional genetic engineering in several ways: (1) synthetic biology takes advantage of de novo DNA synthesis technologies, rather than relying on the existing natural templates; (2) synthetic biologists use standardized genetic parts not only to facilitate the assembly of novel sequences, but also to more predictably construct the biological system based on the characterization of individual parts[3]; and (3) similar to other engineering disciplines, computer-assisted designers (CADs) and mathematical modeling are instrumental in synthetic biology to effectively help synthetic biologists navigate the design space.[4]

Although synthetic biology is still in an early stage, several breakthroughs in the past decade have already demonstrated its great potential for society; for instance, Keasling's group used a synthetic biology approach to engineer the baker's yeast *Saccharomyces cerevisiae* to produce artemisinin, an important antimalarial drug.[5] Lu and Collins engineered bacteriophage for an antibiotic therapy[6] and, more recently, also developed a paper-based cell-free methodology to rapidly detect Ebola viruses.[7]

The enabling technology for synthetic biology is the development of a suite of advanced DNA synthesis and assembly methods, such as Golden Gate assembly,[8] Gibson assembly,[9] circular polymerase extension cloning (CPEC),[10] transformation-assisted recombination (TAR) cloning,[11] and PaperClip assembly[12] (for a comprehensive review on DNA assembly methods, refer to Ellis et al.[13]). Collectively, these technologies open up the possibility to redesign and resynthesize DNA at the genome scale. Poliovirus cDNA was synthesized without a natural template in 2002 by Cello et al.[14] Itaya's group pioneered the combination of two genomes in one cell in vivo.[15] The J. Craig Venter Institute chemically resynthesized a bacterial genome[16] and developed the genome transplantation technology to reboot the cell with the synthetic bacterial genome.[17] Together with several other groups across the world, our group is part of the international synthetic yeast consortium (www.syntheticyeast.org), which aims to redesign and resynthesize the

[1]School of Biological Sciences, University of Edinburgh, The King's Buildings, Edinburgh, UK
[2]BGI-Shenzhen, Shenzhen, China
[†]These authors contributed equally and are listed alphabetically.

Received March 3, 2015.

Supplementary material for this article is available on the *Journal of Laboratory Automation* Web site at http://jla.sagepub.com/supplemental.

Corresponding Author:
Yizhi Cai, PhD, School of Biological Sciences, University of Edinburgh, The King's Buildings, Edinburgh EH9 3BF, UK.
Email: yizhi.cai@ed.ac.uk

world's first eukaryotic genome. We recently reported the completion of the first synthetic yeast chromosome arm[18] and the first fully synthetic eukaryotic chromosome.[19] As a safety measurement and responsible innovation in synthetic biology, efficient biocontainment technologies have been developed to restrict the viability of engineered microbes to prevent the dual use of synthetic biology technologies.[20]

Traditional liquid handling technology has enabled increased throughput of many Life Sciences (Lowell, MA) protocols and assays by (1) increasing operational speeds, (2) reducing working volumes (down to a microliter range), and (3) reducing the need for a generally error-prone human handling, and ultimately contributed to substantial workflow cost savings. Despite an already big "leap forward," the demand for further protocol miniaturization continues to increase, in particular in ultra-high-throughput screening (uHTS).[21] Traditional tips/nozzles-based robotic platforms struggle to precisely dispense liquid droplets below the microliter threshold. Pin tools can be used to transfer nanoliter to microliter liquid from source plates to destination plates; however, because they are contact based, the pin tools usually require washing and drying between transfers to avoid cross-contamination. Also, the delivery volume of pin tools is difficult to control, as it is due to a combination of many factors, such as the shape of the pin, the diameter of the pin, the coating of the pin, and the speed of dipping and removing of the pin. Finally, pin tools are usually made in 96, 384, and 1536 formats, which limits their flexibility of usage, e.g., in setting up different reaction volumes in the same plate. Another technology allowing reaction miniaturization is the microfluidic chip technology.[22] Kong and others have successfully used microfluidic chips to synthesize DNA sequences up to 1 kb,[23] and Tewhey et al. used microfluidic chips to run 1.5 million PCRs in parallel.[24] The main disadvantage of the microfluidic chip approach is that the master molds and the control layer need to be custom designed and fabricated for different reactions; however, the de novo DNA synthesis using microfluidic chips is very complementary with the miniaturized assembly methods described in this paper.

First described in 1927, the acoustic droplet ejection (ADE) phenomenon utilizes acoustic energy to rapidly move low-volume nanoliter to picoliter droplets without any physical contact.[25] Before it reached the laboratory setting in the 2000s, the drop-on-demand technology was first exploited in a number of other fields, including the ink-jet printing industry. Today, Labcyte, Inc. (Sunnyvale, CA) is pioneering the acoustic dispensing technology for Life Sciences, with its Echo series robotic platforms being able to transfer multiple 2.5 or 25 nL droplets from the 384- and 1536-well sources to the various (inverted) destination plates. Unlike traditional robotic liquid transfer methods, laboratory acoustic dispensing has been shown to be highly precise at the nanoliter volume range (as demonstrated by its low coefficients of variation), therefore enabling the desired further miniaturization of current protocols and assays. The acoustic dispenser is flexible enough to set up any-to-any configurations between the source plate and the destination plate, and the reaction volumes can vary from well to well in the same reaction plate.

Here, for the first time, we report yet another exciting acoustic dispensing application: nanoliter-scale DNA assembly. The majority of assembly expenses are enzymes, including DNA polymerases, Therefore, downscaling the reaction volume from the microliter to the nanoliter scale while maintaining high assembly efficiency, will make DNA synthesis and assembly more accessible to synthetic biologists.

Materials and Methods

Echo PCR

Conventional endpoint PCR is instrumental in making synthetic DNA. To test the minimal volume of regular PCR using Echo, we set up PCRs of various volumes. The plasmid HcKan_P vector (120 ng/μl) was used as the DNA template, and a pair of primers YCp2214 and YCp2215 were designed to amplify a targeted DNA fragment of 1378 bp (**Fig. 1A**; all primers used in this paper are listed in **Suppl. Table S1**). The GoTaq Green Master Mix (Promega, Madison, WI) was used in the PCR. Five reaction volumes ranging from 50 to 1000 nL were set up (**Table 1**), and each reaction was performed in four replicates. All PCRs were set up using the following cycling conditions: preheat the PCR machine and then put in the PCR plate, 2 min at 95 °C, 32 cycles of 10 s at 95 °C, 30 s at 50 °C and 2 min at 72 °C, 7 min at 72 °C, and hold at 4° C. GoTaq Green Master Mix (35 μL) and double-distilled water (ddH_2O; 30 μL) were added separately to source plate 1, which was an Echo 384-well polypropylene plate (Labcyte). YCp2214 and YCp2215 (10 μL each) and template DNA (10 μL) were added separately to source plate 2, which was an Echo 384-well low-dead-volume plate (Labcyte). The destination plate used in this paper was MicroAmp EnduraPlate (Life Technologies, Carlsbad, CA).

Gibson DNA Assembly

First described in 2009, the Gibson DNA assembly method[9] belongs to a group of overlap-directed DNA assembly techniques such as CPEC,[10] SLiCE,[26] and SLIC[27] assemblies. The Gibson assembly method is one of the most used in synthetic biology, and it can assemble DNA sequences up to small genome sizes from overlapping DNA fragments in an isothermal one-pot reaction. The advantage of Gibson assembly is that it is sequence independent and generates scarless final assembled DNA products. Typically, the Gibson assembly requires about a 40 bp homologous region between two adjacent DNA fragments, and these homologous regions are usually added to the fragments by a high-fidelity PCR. Briefly, the assembly reaction takes place in a cocktail of enzymes (termed

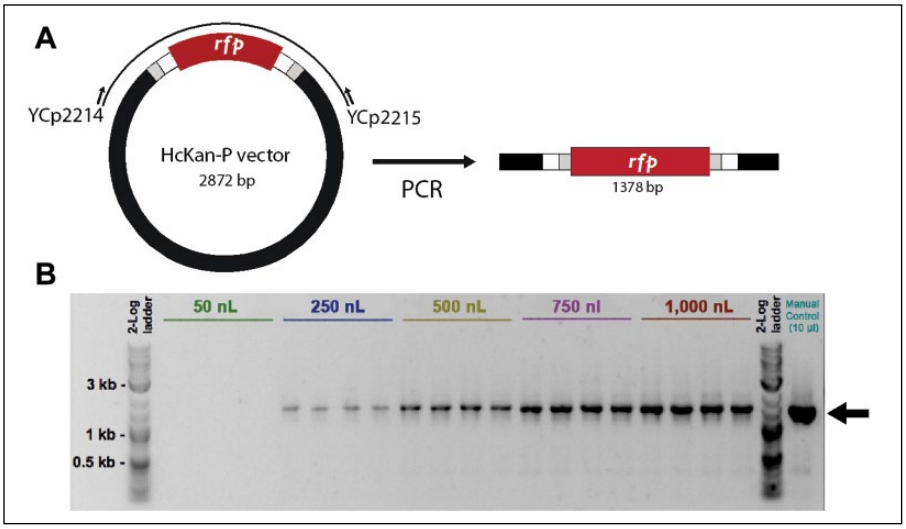


Figure 1. PCR setup by Echo. (**A**) A pair of primers was designed to amplify a fragment of 1.3 kb, and PCRs of various volumes were set up by the Echo machine. (**B**) Gel electrophoresis confirms that PCR can work at the 250 nL scale.

Table 1. PCR Setup.

Reagent/nl	Echo	Echo	Echo	Echo	Echo	Manual
Primer YCp2214	2.5	12.5	25.0	37.5	50.0	500.0
Primer YCp2215	2.5	12.5	25.0	37.5	50.0	500.0
Template DNA	5.0	25.0	50.0	75.0	100.0	1000.0
ddH_2O	15.0	75.0	150.0	225.0	300.0	3000.0
GoTaq Green Master Mix	25.0	125.0	250.0	375.0	500.0	5000.0
Total	50 nL	250 nL	500 nL	750 nL	1000 nL	10,000 nL

Gibson master mix) at 50 °C for 60 min: (1) First, T5 exonulease chews back the DNA in a 5′ to 3′ direction from the homologous terminal ends to reveal reverse complementary single-stranded sequences between two adjacent fragments. (2) While the 5′ to 3′ DNA digestion proceeds, a high-fidelity DNA polymerase fills in the single-stranded DNA region. (3) Finally, Taq DNA ligase seals the nicked DNA strands, which yields the final assembled product.

Gibson Reaction Setup by Echo. Two pairs of primers (YCp2391 and YCp2392 for fragment 1, YCp2393 and YCp2394 for fragment 2) were designed to amplify two fragments with 40 bp end homology from a red fluorescent protein (RFP)–containing plasmid pPC025, thus allowing subsequent Gibson reassembly of the plasmid. The two homologous junctions were placed within the ampicillin resistance gene and the RFP open reading frame (ORF) to reduce the overall false positive rate and to allow phenotypic screening for successful assembly isolates, respectively. In contrast to Golden Gate assembly (see below), here RFP serves as a positive screen for correct assemblies. PCR products were gel purified using the QIAquick gel extraction kit (Qiagen, Valencia, CA). The standard 15 µL Gibson assembly master mix was prepared as described in the original Gibson assembly paper.[9] Gibson master mix (40 µL) was added to source plate 1, which is an Echo 384 polypropylene plate (Labcyte). Each DNA fragment (10 µL) was added to source plate 2, which is an Echo 384 low-dead-volume plate (Labcyte). One-pot Gibson assembly was incubated at 50 °C for 60 min in a preheated PCR thermal cycler (**Table 2**).

Golden Gate Assembly

The Golden Gate DNA assembly method utilizes a combination of a TypeIIS restriction enzyme and a ligase to assemble the DNA fragments.[8] TypeIIS enzymes (e.g., BsaI and BsmBI enzymes) are endonucleases that cut outside their recognition sites, creating 4 bp DNA overhangs. By carefully designing the 4 bp overhangs, one can use the Golden Gate reaction to directionally assemble DNA fragments. The Golden Gate DNA assembly reaction starts with a given TypeIIS endonuclease DNA digestion, leaving behind staggered cuts in the backbone and the fragment DNA. The design-imposed DNA complementarity allows annealing of the resulting "sticky ends," creating the desired plasmid construct. In the final reaction step, the T4 DNA ligase repairs the nicks to complete the DNA construction phase.

Golden Gate Reaction Setup by Echo. The HcKan_P plasmid (2.8 kb, diluted to 10 ng/µl) was used as the acceptor vector. This plasmid carries a KanR selectable marker, along with a RFP cassette flanked by a pair of outward-facing BsaI

Table 2. Gibson Assembly Reactions.

Reagent/nl	Echo	Echo	Echo	Echo	Manual
Gibson master mix	37.5	187.5	375.0	750.0	15,000.0
Fragment 1 (113.8 ng/µl)	5.0	20.0	40.0	80.0	2,500.0
Fragment 2 (86.8 ng/µl)	7.5	42.5	85.0	170.0	2,500.0
Total	50 nL	250 nL	500 nL	1000 nL	20,000 nL

sites. We amplified the promoter *pMBP1* (500 bp) directly from yeast BY4741 (MATa, *leu2Δ0 met15Δ0 ura3Δ0 his3Δ1*) genomic DNA with primers YCp2395 and YCp2396 and added a pair of inward-facing BsaI sites to flank the promoter part (**Fig. 2A**). The PCR product was purified using a PureLink PCR purification kit (Life Technologies) and diluted to 20 ng/µl. The 4 bp overhangs were designed in such a way that the promoter can be efficiently assembled into the acceptor vector. Bacteria carrying the residual RFP plasmid will give a bright red pigment, which would facilitate the visual identification of correct assembled clones (white colonies; see **Fig. 2C**).

The Golden Gate master mix was made of 35 µL T4 ligase (2000 U/µl, New England Biolabs, NEB), 35 µL BsaI-HF (NEB), 52.5 µL 10× T4 buffer (NEB), and 25 µL 200× BSA (NEB). Golden Gate assembly reactions were set up using the following cycling conditions: 15 cycles of 5 min at 37 °C and 10 min at 16 °C, 5 min at 50 °C, 10 min at 80 °C, and hold at 4 °C. Five reaction volumes arranging from 50 to 1000 nL were set up (**Table 3**), and each reaction was performed in triplicate. A manual positive control reaction of 7.5 µL was also set up to confirm the fidelity of the reagents. Golden Gate master mix (30 µL) was added to source plate 1, which is an Echo 384 polypropylene plate (Labcyte). *pMBP1* PCR product (10 µL) and HcKan_P vector (10 µL) were added to source plate 2, which is an Echo 384 low-dead-volume plate (Labcyte).

Bacterial Transformation

As the assembly reactions set up by Echo were at the nanoliter scale, it is difficult to take out the assembled DNA using pipets and transform them into bacterial competent cells. Instead, bacterial competent cells were added to each well containing an assembled product. Competent *Escherichia coli* (20 µL; MAX Efficiency DH5α, Life Technologies) was added to each well of the reaction plate. The PCR plate was incubated on ice for 20 min and then placed in a heat block at 42 °C for 45 s. The plate was placed back on ice to incubate for 5 min, before adding 200 µL of room temperate super Optimal Catabolite repression (SOC) medium to each well. The plate was incubated at 37 °C with shaking at 200 rpm for 1 h. A multichannel pipet was used to slowly drip 40 µL of each transformation mixture onto an omnitray containing selective solid agar medium (LB—Kan). Alternatively, 100 µL of transformation mixture was plated on individual petri dishes with selective solid agar medium (Golden Gate assembly, LB—Kan; Gibson assembly, LB—Amp). Plates were incubated overnight at 37 °C until single colonies appeared.

Gel Electrophoresis

Gel electrophoresis was performed to analyze the PCR products (120 V, 30 min; 1% w/v agarose in Tris-acetate-EDTA (TAE) buffer with 1× SYBR Safe DNA stain). Each PCR product was first diluted with ddH_2O to a final volume of 5 µL when the PCR volume was smaller than 5 µL.

Sanger Sequencing

A BigDye Terminator v3.1 Cycle Sequencing Kit (Life Technologies) was used to verify the DNA assembly clones according to the manufacturer's instructions, and the Sanger sequencing reactions were carried out by Edinburgh Genomics.

Results and Discussion

We used the Echo machine to set up PCRs in total volumes ranging from 50 nL to 1 µL (**Fig. 1** and **Table 1**). Starting from 250 nL, a band of the correct size could be detected in the gel electrophoresis. Because we diluted the PCR product to 5 µL in order to run the gel electrophoresis, it is possible that PCRs at 50 nL scale were successful, but the gel electrophoresis was not sensitive enough to detect the signal. Alternatively, it would be possible to use the Caliper Labchip GX instrument that can detect DNA concentrations as low as 5 ng/µL. Downsizing the PCR from 50 µL or higher to 250 nL already effectively cuts the reagent cost by 200-fold. Miniaturized PCR is ideal for diagnostic purposes such as fast genotyping and colony-screening PCR, but it is less suitable for applications requiring use of the PCR product for downstream procedures, such as cloning, because the yield of double-stranded DNA may not be sufficient.

Gibson assembly worked extremely well in this experiment. Correct assembly was observed from as low as the 250 nL reaction volumes, and at 500 and 1000 nL the assembly efficiencies are comparable with or better than the manual control of the 20 µL reaction, but with a significant standard deviation. This allows us to cut the reagent cost by 20-fold or more. Even more encouraging, we observed no background (**Fig. 2C**) and 100% correct assembly through

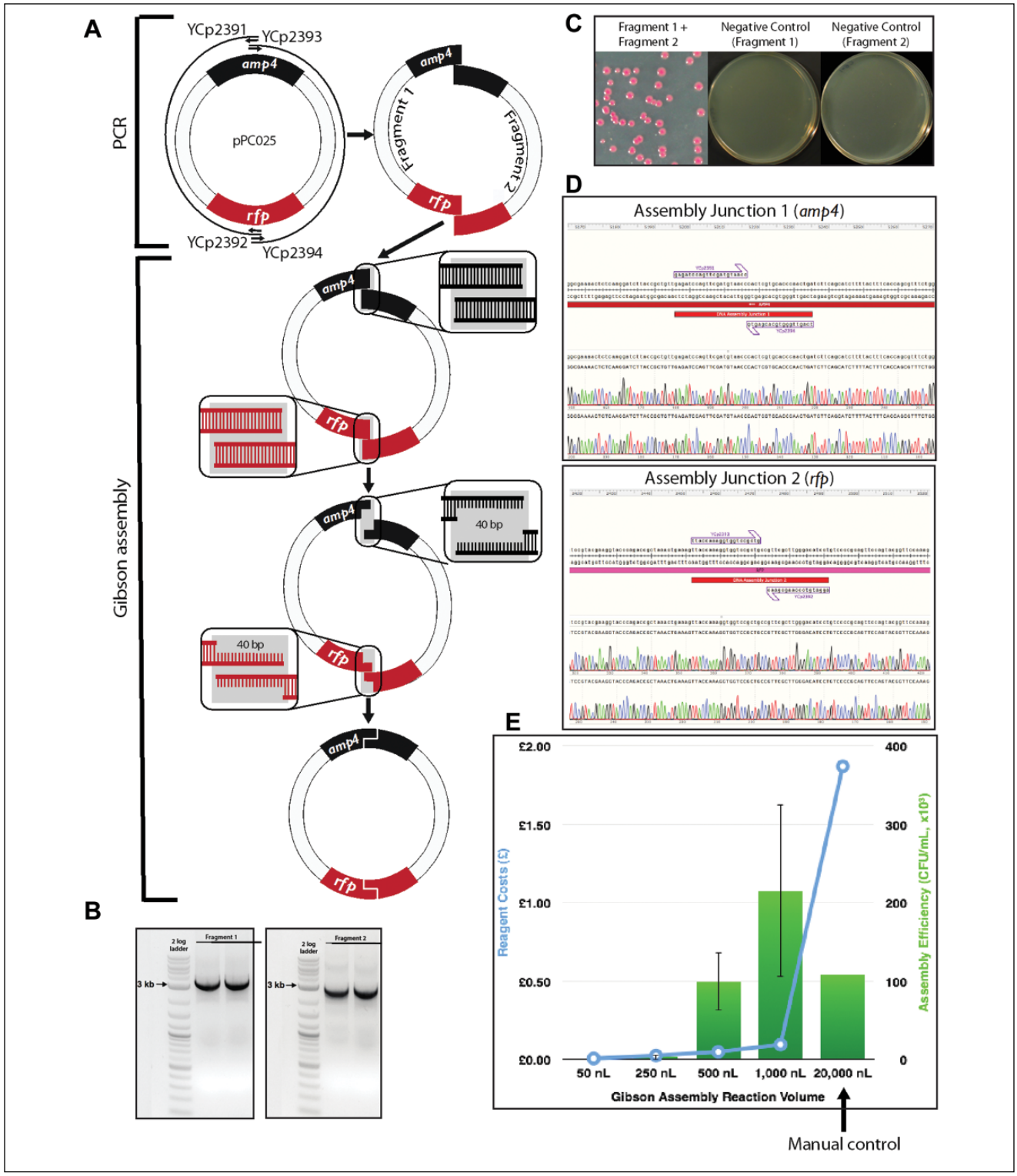


Figure 2. Gibson assembly reaction setup by Echo. (**A**) The pPC025 plasmid was split into two overlapping fragments at the middle of the ampicillin resistance gene and the RFP ORF. Two fragments were generated with 40 bp overlap at both ends and then assembled by the Gibson assembly reaction. (**B**) Gel electrophoresis confirming the successful PCR amplification of both fragments. (**C**) Successful Gibson assembly product gives rise to red bacterial colonies. The assembly efficiency was so high that no background colonies (white) were observed. Negative control reactions, which had only one fragment in the reactions, yielded no colonies. (**D**) Sequencing verification of both assembly junctions shows 100% assembly accuracy. (**E**) Cost-effectiveness and assembly efficiency comparison of different reaction volumes for Gibson assembly.

Table 3. Golden Gate Assembly Reactions.

Reagent/nl	Echo	Echo	Echo	Echo	Manual
Golden Gate master mix	17.5	82.5	167.5	332.5	2500
pMBP1 (20 ng/µl)	30.0	150.0	300.0	600.0	4500
HcKan_P (10 ng/µl)	2.5	17.5	32.5	67.5	500
Total	50 nL	250 nL	500 nL	1000 nL	7500 nL

Sanger sequencing across the assembly junctions (six red colonies were sequenced; **Fig. 2D**), and this will be highly beneficial for future automation plans, as it will greatly reduce the colony screening effort.

With Golden Gate assembly, we successfully assembled DNA at a 50 nL reaction volume (typically 15 µL reactions when performed manually), and at the 250 and 500 nL scales the assembly efficiencies are higher than those of the manual control. This leads to at least a 30-fold reduction in reagent use when performing Golden Gate reactions using Echo. We did observe vector background in the assembly (red colonies, as shown in **Fig. 3C**). There are several ways we can overcome this problem. First, instead of using RFP for screening, we can use the toxic ccdB gene, which cannot give rise to background colonies in a nonpermissive transformation host. Second, we can add a higher concentration of the BsaI enzyme in the Golden Gate master mix to further digest the residual acceptor vector. Finally, we may be able to reduce the background by extending the BsaI digestion step in the incubation.

With further optimization, it should be possible to downsize the reaction volume even further. For instance, in this paper, individual components were shot off to the destination well one by one (in the case of 50 nL PCRs, only 1 droplet of primer was shot), and it is possible that some components were not sent into the reaction pool due to slight misalignment of the acoustic dispenser, meaning the reactants simply didn't mix. In this case, it would be of advantage to premix as many components as possible and then shoot more droplets altogether. We also suggest dispensing the master mix using a bulk dispenser or liquid handler, so that the destination well has a larger liquid surface to uptake the incoming droplet and minimize chances for the droplets hitting the well wall. It is always a good practice to centrifuge the PCR plate when appropriate before putting it into the PCR thermal cycler to start the reaction. To prevent the nanoliter droplet from evaporating before the chemical reaction starts, we always preheat the PCR machine before putting in the reaction plate. Finally, it is more economical to use low-dead-volume plates as the source plate for expensive reagents such as enzymes and polymerases.

In the world of laboratory automation, efficiency and robustness are as important as cost saving. With this in mind, we overlaid the number of correct assemblies (efficiency) with standard deviation (robustness) in the same plot with the cost of reactions for the Gibson assembly (**Fig. 2E**) and the Golden Gate assembly (**Fig. 3E**). The intersections of the two curves indicate the "sweet spots" for choosing desired reaction volumes, which are of high efficiency, low standard deviation, and relatively low cost. It should be noted that our cost calculation did not take into account the dead volume of reagents, and logically it can be assumed that the dead-volume cost per reaction would decrease as more reactions are set up by Echo in one experiment. Whenever possible, low-dead-volume plates should be used for expensive reagents to save cost. Conversely, we did not include the tip cost in the manual control experiments, which increase substantially when the number of reactions is scaled up. Continuously monitoring DNA assembly efficiency along with the assembly cost is critical to successful operation of a large DNA synthesis and assembly automation facility, such as the UK DNA foundries.

The acoustic dispensing has great potential in automating other molecular biology operations. We also used the Echo to purify single colonies from bacterial and yeast cultures, which is traditionally challenging to automate. As Echo is capable of dispensing nanoliter droplets with high precision, it is also ideal for generating high-density assembly libraries through combinatorial assembly methods. In conclusion, the work described here is the first report on use of the acoustic dispenser in the area of synthetic biology, and we envision that this technology will be instrumental in lab automation, in particular in the era of DNA foundries.

Acknowledgments

We acknowledge Edinburgh Genome Foundry for access to a Labcyte Echo 550 instrument. We thank Jamie Auxillos for assistance with the figures and Roy Walker and Dr. Chris French for proofreading the manuscript.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: P.K., Y.S, Y.Z., and Y.C. were supported by a Chancellor's fellowship from the University of Edinburgh, a start-up fund from SULSA (to Y.C.), and BBSRC grants BB/M005690/1 and BB/

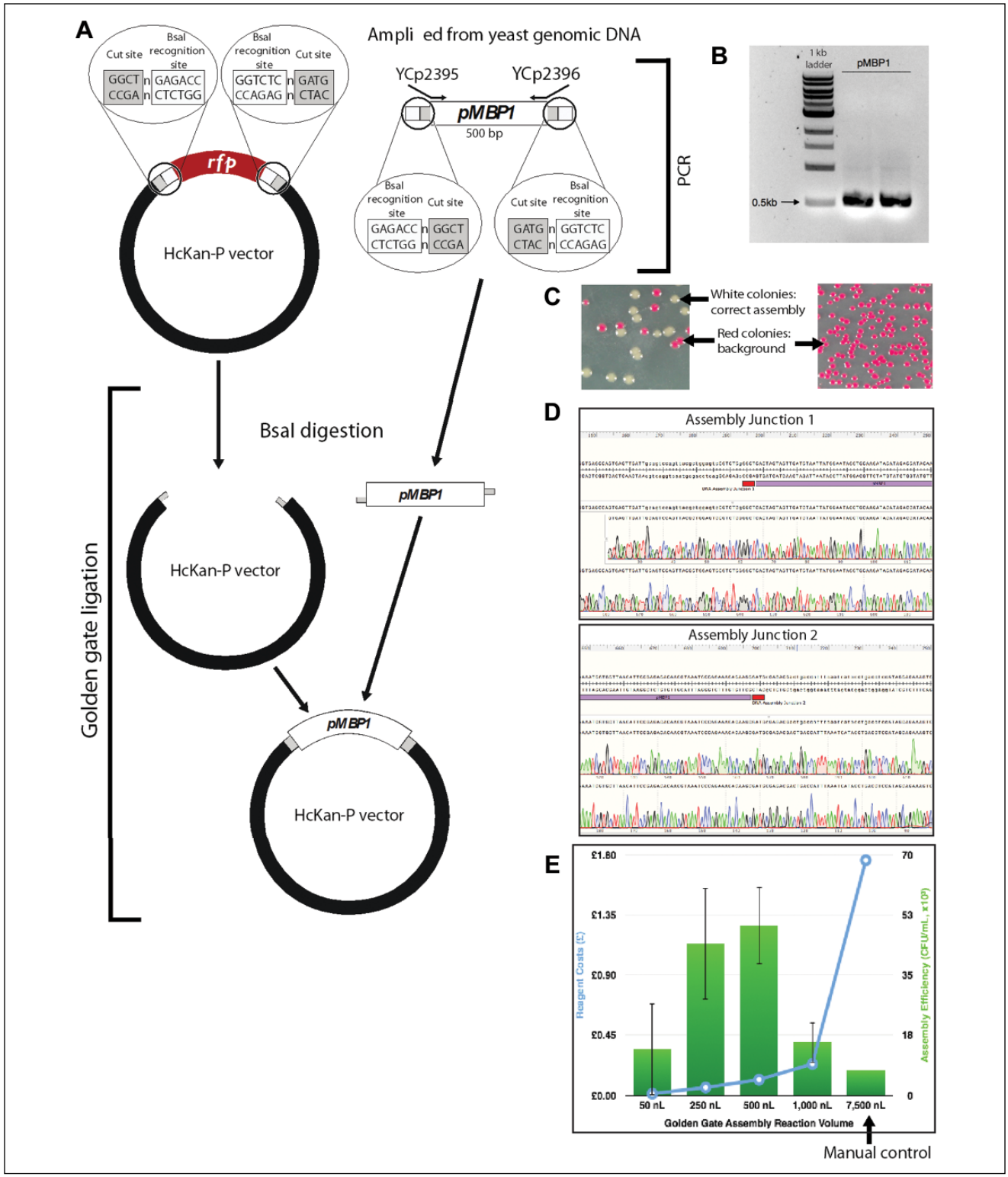


Figure 3. Golden Gate assembly setup by Echo. (**A**) A promoter pMBP1 was amplified from the yeast genome to add appropriate Golden Gate sequences (BsaI recognition sites + 4 bp overhangs). The acceptor vector HcKan_P plasmid carries a RFP cassette, which is flanked by corresponding Golden Gate sequences to uptake the pMBP1 part in the Golden Gate reaction. (**B**) Gel electrophoresis indicates successful amplification of pMBP1. (**C**) Left: Successful assembled DNA gives rise to white colonies, while the residual acceptor vector yields red colonies. Right: Negative control, which contained only the acceptor vector in the Golden Gate reaction, yielded only red colonies. (**D**) Sequencing verification of both assembly junctions shows 100% assembly accuracy. (**E**) Cost-effectiveness and assembly efficiency comparison of different reaction volumes for Golden Gate assembly.

M025640/1 (to Y.C. and S.R.). P.K. is also supported by an EPSRC CASE studentship EP/M506515/1 with Thermo Fisher (to Y.C.). The open-access charge is supported by the Research Councils UK Open Access Fund.

References

1. (a) Andrianantoandro, E.; Basu, S.; Karig, D. K.; et al. Synthetic Biology: New Engineering Rules for An Emerging Discipline. *Mol. Syst. Biol*. **2006**, *2*, 2006.0028. (b) Ball, P. Synthetic Biology: Starting from Scratch. *Nature* **2004**, *431*, 624–626. (c) Endy, D. Foundations for Engineering Biology. *Nature* **2005**, *438*, 449–453.
2. (a) Bio, F. A. B. G.; Baker, D.; Church, G.; et al. Engineering Life: Building a Fab for Biology. *Sci. Am*. **2006**, *294*, 44–51. (b) Cameron, D. E.; Bashor, C. J.; Collins, J. J. A Brief History of Synthetic Biology. *Nat. Rev. Microbiol.* **2014**, *12*, 381–390.
3. Mutalik, V. K.; Guimaraes, J. C.; Cambray, G.; et al. Precise and Reliable Gene Expression via Standard Transcription and Translation Initiation Elements. *Nat. Methods* **2013**, *10*, 354–360.
4. (a) Cai, Y.; Hartnett, B.; Gustafsson, C.; et al. A Syntactic Model to Design and Verify Synthetic Genetic Constructs Derived from Standard Biological Parts. *Bioinformatics* **2007**, *23*, 2760–2767. (b) Czar, M. J.; Cai, Y.; Peccoud, J. Writing DNA with GenoCAD. *Nucl. Acids Res*. **2009**, *37*, W40–W47. (c) Hillson, N. J. j5 DNA Assembly Design Automation. *Methods Mol. Biol*. **2014**, *1116*, 245–269. (d) Hillson, N. J.; Rosengarten, R. D.; Keasling, J. D. j5 DNA Assembly Design Automation Software. *ACS Synth. Biol.* **2012**, *1*, 14–21. (e) Xia, B.; Bhatia, S.; Bubenheim, B.; et al. Developer's and User's Guide to Clotho v2.0: A Software Platform for the Creation of Synthetic Biological Systems. *Methods Enzymol.* **2011**, *498*, 97–135.
5. Ro, D. K.; Paradise, E. M.; Ouellet, M.; et al. Production of the Antimalarial Drug Precursor Artemisinic Acid in Engineered Yeast. *Nature* **2006**, *440*, 940–943.
6. Lu, T. K.; Collins, J. J. Engineered Bacteriophage Targeting Gene Networks as Adjuvants for Antibiotic Therapy. *Proc. Natl. Acad. Sci. U.S.A*. **2009**, *106*, 4629–4634.
7. Pardee, K.; Green, A. A.; Ferrante, T.; et al. Paper-Based Synthetic Gene Networks. *Cell* **2014**, *159*, 940–954.
8. Engler, C.; Gruetzner, R.; Kandzia, R.; et al. Golden Gate Shuffling: A One-Pot DNA Shuffling Method Based on Type IIs Restriction Enzymes. *PloS One* **2009**, *4*, e5553.
9. Gibson, D. G.; Young, L.; Chuang, R. Y.; et al. Enzymatic Assembly of DNA Molecules up to Several Hundred Kilobases. *Nat. Methods* **2009**, *6*, 343–345.
10. Quan, J.; Tian, J. Circular Polymerase Extension Cloning of Complex Gene Libraries and Pathways. *PloS One* **2009**, *4*, e6441.
11. Larionov, V.; Kouprina, N.; Graves, J.; et al. Specific Cloning of Human DNA as Yeast Artificial Chromosomes by Transformation-Associated Recombination. *Proc. Natl. Acad. Sci. U.S.A*. **1996**, *93*, 491–496.
12. Trubitsyna, M.; Michlewski, G.; Cai, Y.; et al. PaperClip: Rapid Multi-Part DNA Assembly from Existing Libraries. *Nucl. Acids Res*. **2014**, *42*, e154.
13. Ellis, T.; Adie, T.; Baldwin, G. S. DNA Assembly for Synthetic Biology: From Parts to Pathways and Beyond. *Integr. Biol*. **2011**, *3*, 109–118.
14. Cello, J.; Paul, A. V.; Wimmer, E. Chemical Synthesis of Poliovirus cDNA: Generation of Infectious Virus in the Absence of Natural Template. *Science* **2002**, *297*, 1016–1018.
15. Itaya, M.; Tsuge, K.; Koizumi, M.; et al. Combining Two Genomes in One Cell: Stable Cloning of the Synechocystis PCC6803 Genome in the *Bacillus subtilis* 168 Genome. *Proc. Natl. Acad. Sci. U.S.A*. **2005**, *102*, 15971–15976.
16. Gibson, D. G.; Benders, G. A.; Andrews-Pfannkoch, C.; et al. Complete Chemical Synthesis, Assembly, and Cloning of a *Mycoplasma genitalium* Genome. *Science* **2008**, *319*, 1215–1220.
17. Gibson, D. G.; Glass, J. I.; Lartigue, C.; et al. Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. *Science* **2010**, *329*, 52–56.
18. Dymond, J. S.; Richardson, S. M.; Coombes, C. E.; et al. Synthetic Chromosome Arms Function in Yeast and Generate Phenotypic Diversity by Design. *Nature* **2011**, *477*, 471–476.
19. Annaluru, N.; Muller, H.; Mitchell, L. A.; et al. Total Synthesis of a Functional Designer Eukaryotic Chromosome. *Science* **2014**, *344*, 55–58.
20. (a) Cai, Y.; Agmon, N.; Choi, W. J.; et al. Intrinsic Biocontainment: Multiplex Genome Safeguards Combine Transcriptional and Recombinational Control of Essential Yeast Genes. *Proc. Natl. Acad. Sci. U.S.A*. **2015**, *112*, 1803–1808. (b) Mandell, D. J.; Lajoie, M. J.; Mee, M. T.; et al. Biocontainment of Genetically Modified Organisms by Synthetic Protein Design. *Nature* **2015**, *518*, 55–60. (c) Gallagher, R. R.; Patel, J. R.; Interiano, A. L.; et al. Multilayered Genetic Safeguards Limit Growth of Microorganisms to Defined Environments. *Nucl. Acids Res*. **2015**, *43*, 1945–1954. (d) Rovner, A. J.; Haimovich, A. D.; Katz, S. R.; et al. Recoded Organisms Engineered to Depend on Synthetic Amino Acids. *Nature* **2015**, *518*, 89–93.
21. Dunn, D. A.; Feygin, I. Challenges and Solutions to Ultra-High-Throughput Screening Assay Miniaturization: Submicroliter Fluid Handling. *Drug Discov. Today* **2000**, *5*, 84–91.
22. Szita, N.; Polizzi, K.; Jaccard, N.; et al. Microfluidic Approaches for Systems and Synthetic Biology. *Curr. Opin. Biotechnol*. **2010**, *21*, 517–523.
23. (a) Kong, D. S.; Carr, P. A.; Chen, L.; et al. Parallel Gene Synthesis in a Microfluidic Device. *Nucl. Acids Res*. **2007**, *35*, e61. (b) Huang, M. C.; Ye, H.; Kuan, Y. K.; et al. Integrated Two-Step Gene Synthesis in a Microfluidic Device. *Lab Chip* **2009**, *9*, 276–285. (c) Lee, C. C.; Snyder, T. M.; Quake, S. R. A Microfluidic Oligonucleotide Synthesizer. *Nucl. Acids Res.* **2010**, *38*, 2514–2521.
24. Tewhey, R.; Warner, J. B.; Nakano, M.; et al. Microdroplet-Based PCR Enrichment for Large-Scale Targeted Sequencing. *Nat. Biotechnol*. **2009**, *27*, 1025–1031.
25. Ellson, R.; Mutz, M.; Browning, B.; et al. Transfer of Low Nanoliter Volumes between Microplates Using Focused Acoustics—Automation Considerations. *JALA* **2003**, *8*, 29–34.
26. Zhang, Y.; Werling, U.; Edelmann, W. SLiCE: A Novel Bacterial Cell Extract-Based DNA Cloning Method. *Nucl. Acids Res*. **2012**, *40*, e55.
27. Li, M. Z.; Elledge, S. J. Harnessing Homologous Recombination In Vitro to Generate Recombinant DNA via SLIC. *Nat. Methods* **2007**, *4*, 251–256.

Appendix 3

Appendix 3.1 Supplementary Tables

Supp. Table 3.1.1 crRNAs Used in Single-gene Deletion Experiments

gene name	crRNA sequence (5' – 3')	PAM sequence	specificity score (Hsu *et al*., 2013)	efficiency score (Doench *et al*., 2014)
ADE2 (*YOR128C*)	TAACTTCGTTGTAAAGAATA	AGG	50.0	47.5
CAN1 (*YEL063C*)	CACAAACACACCACAGACGT	GGG	50.0	71.8
SEN34 (*YAR008W*)	ATCAAACTTCTAAGGAAATG	GGG	99.9	55.1

Supp. Table 3.1.2 Sanger Sequencing of Sample CRISPR Pool Bacterial Clones' Plasmid DNA

clone number	plasmid insertion length (bp)	comments
1	167	correct full oligonucleotide length insertion
2	167	correct full oligonucleotide length insertion
3	171	correct full oligonucleotide length insertion
4	23	truncated oligonucleotide insertion
5	170	correct full oligonucleotide length insertion
6	168	correct full oligonucleotide length insertion
7	167	correct full oligonucleotide length insertion
8	169	correct full oligonucleotide length insertion
9	166	correct full oligonucleotide length insertion
10	106	truncated oligonucleotide insertion
11	0	no alignment
12	169	correct full oligonucleotide length insertion
13	170	correct full oligonucleotide length insertion
14	169	correct full oligonucleotide length insertion
15	53	truncated oligonucleotide insertion
16	166	correct full oligonucleotide length insertion
17	0	no alignment
18	166	correct full oligonucleotide length insertion
19	171	correct full oligonucleotide length insertion
20	170	correct full oligonucleotide length insertion

Supp. Table 3.1.3 Conserved and Non-conserved Regions Targeted

chromosome number	region coordinates (bp)	control type
IV	620666-621411	non-conserved
XV	481945-482019	non-conserved
II	28859-28918	non-conserved
V	281218-281493	non-conserved
XIV	495207-495385	non-conserved
II	95942-95973	non-conserved
XI	224685-224725	non-conserved
II	60247-60653	non-conserved
IV	582499-582789	non-conserved
VII	478967-479395	non-conserved
II	431106-431124	non-conserved
XV	851843-851940	non-conserved
XVI	14840-15051	non-conserved
XV	329302-329403	non-conserved
VIII	266917-266951	non-conserved
XII	1064955-1065023	non-conserved
XVI	920702-921212	non-conserved
XII	84174-84264	non-conserved
XII	981702-982094	non-conserved
XIII	317097-317163	non-conserved
XV	592809-592873	non-conserved
IX	226079-226430	non-conserved
XVI	620053-620317	non-conserved
IV	722103-722354	non-conserved
II	45731-45922	non-conserved
XIV	495207-495385	non-conserved
III	308313-308558	non-conserved
III	289404-289616	non-conserved
IV	1477200-1477237	non-conserved
I	203149-203310	non-conserved
III	123232-123575	non-conserved
XIV	569941-570023	non-conserved
XIII	759228-759271	non-conserved
XI	527363-527670	non-conserved
III	116734-117312	non-conserved
XII	137875-137938	non-conserved
VIII	423698-423722	non-conserved

chromosome number	region coordinates (bp)	control type
II	143042-143160	non-conserved
IV	217209-217306	non-conserved
II	589120-589237	non-conserved
XV	80015-80191	non-conserved
V	61058-61141	non-conserved
XIV	210183-210231	non-conserved
II	658393-658598	non-conserved
XV	908667-908898	non-conserved
XII	337415-337459	non-conserved
VIII	423673-423697	non-conserved
XII	283734-283870	non-conserved
X	462555-462719	non-conserved
XII	449807-449876	non-conserved
XVI	829071-829168	non-conserved
II	334024-334149	non-conserved
XI	666437-666548	non-conserved
XV	31818-31847	non-conserved
XI	418186-418305	non-conserved
XVI	921683-921858	non-conserved
VII	751103-751225	non-conserved
VI	226771-226851	non-conserved
XIII	77771-78272	non-conserved
X	703207-703395	non-conserved
XII	130725-130760	non-conserved
XII	554147-554307	non-conserved
XVI	85543-85580	non-conserved
II	536452-536555	non-conserved
XVI	572342-572407	non-conserved
V	260857-260931	non-conserved
XIV	525385-525535	non-conserved
IV	1206555-1206674	non-conserved
X	726899-726934	non-conserved
VI	256998-257462	non-conserved
XV	903780-903837	non-conserved
XI	527363-527670	non-conserved
VII	478967-479395	non-conserved
VIII	184836-184873	non-conserved
IV	741842-741854	non-conserved
IX	142032-142548	non-conserved

chromosome number	region coordinates (bp)	control type
III	127011-127143	non-conserved
M	494-550	non-conserved
XIII	124357-124496	non-conserved
XVI	47105-47171	non-conserved
V	498852-499077	non-conserved
XV	167439-168003	non-conserved
II	565531-565546	non-conserved
XI	462997-463199	non-conserved
XIV	356727-356792	non-conserved
XII	1001725-1001926	non-conserved
V	422867-422956	non-conserved
VII	599208-599242	non-conserved
XIV	632377-632565	non-conserved
III	78290-78345	non-conserved
V	570137-570322	non-conserved
IV	158870-159350	non-conserved
XVI	880725-881010	non-conserved
VII	433580-433721	non-conserved
XI	21080-21165	non-conserved
XVI	627769-627858	non-conserved
XVI	261318-261335	non-conserved
II	668553-668661	non-conserved
X	543766-544060	non-conserved
IX	36558-36664	non-conserved
XI	151239-151621	non-conserved
XV	384977-385116	non-conserved
XIII	512201-512744	non-conserved
VIII	297098-297150	non-conserved
IV	507787-507836	non-conserved
IX	432714-432904	non-conserved
VI	104471-104503	non-conserved
IV	1447327-1447705	non-conserved
XV	968475-968992	non-conserved
IV	147931-147952	non-conserved
VII	788319-788529	non-conserved
XII	1041367-1041515	non-conserved
XI	392404-392433	non-conserved
X	623279-623349	non-conserved
XII	391945-392710	non-conserved

chromosome number	region coordinates (bp)	control type
VII	1065663-1067146	non-conserved
XV	28048-28162	non-conserved
XI	134217-134240	non-conserved
XIII	318483-318496	non-conserved
VII	547996-548034	non-conserved
XIII	809026-809051	non-conserved
IV	829059-829139	non-conserved
XIII	135348-135374	non-conserved
XV	21312-21426	non-conserved
XIV	63333-63411	non-conserved
XI	200576-200833	non-conserved
XV	458793-458890	non-conserved
XIII	503324-503739	non-conserved
XII	898895-899023	non-conserved
XIV	495123-495205	non-conserved
IX	35651-35900	non-conserved
IX	428006-428307	non-conserved
XIV	439284-439353	non-conserved
XI	221281-221342	non-conserved
VII	23365-23449	non-conserved
XIV	392497-392514	non-conserved
II	720723-721240	non-conserved
XIV	499057-499415	non-conserved
XVI	339440-339552	non-conserved
XIV	185556-185585	non-conserved
XV	989298-989352	non-conserved
VII	270776-270802	non-conserved
XII	949196-949516	non-conserved
XII	377465-377480	non-conserved
X	445062-445193	non-conserved
VIII	255890-256214	non-conserved
XIII	922445-922541	non-conserved
XVI	22017-22580	non-conserved
XII	201622-201735	non-conserved
IX	142032-142548	non-conserved
XIII	917052-917577	non-conserved
XVI	128284-128314	non-conserved
XIII	793657-793720	non-conserved
IV	1334154-1334344	non-conserved

chromosome number	region coordinates (bp)	control type
XII	783956-784166	non-conserved
II	498801-498837	non-conserved
IV	509560-509669	non-conserved
XII	289479-289532	non-conserved
VII	726593-726821	non-conserved
XVI	860460-860621	non-conserved
XI	137104-137850	non-conserved
V	257963-258034	non-conserved
XI	229881-229987	non-conserved
VIII	14579-14899	non-conserved
XIII	463216-463291	non-conserved
XIV	746565-746791	non-conserved
III	126731-126909	non-conserved
VII	414029-414100	non-conserved
XIII	789103-789220	non-conserved
IV	765552-765704	non-conserved
X	223575-223782	non-conserved
X	374589-374846	non-conserved
II	40528-40579	non-conserved
XV	468094-468210	non-conserved
VIII	483969-484026	non-conserved
XV	343082-343431	non-conserved
II	808361-808598	non-conserved
IV	1493991-1494009	non-conserved
IV	454905-455028	non-conserved
XV	312734-312755	non-conserved
VII	508702-508891	non-conserved
XVI	183057-183198	non-conserved
X	745262-745340	non-conserved
VIII	242024-242347	non-conserved
X	719065-719183	non-conserved
XIV	330980-331210	non-conserved
VII	478967-479395	non-conserved
I	13226-13361	non-conserved
IV	1189572-1189838	non-conserved
XV	960397-960730	non-conserved
XI	108911-109116	non-conserved
XIII	10029-10196	non-conserved
XVI	272816-273002	non-conserved

chromosome number	region coordinates (bp)	control type
IV	19689-19981	non-conserved
II	555447-555664	non-conserved
II	196222-196559	non-conserved
XII	1041676-1041797	non-conserved
XII	368125-368242	non-conserved
VII	35235-35405	non-conserved
XIII	91727-91980	non-conserved
XV	818559-818570	non-conserved
XVI	129738-129795	non-conserved
XVI	246374-246579	non-conserved
V	254429-254544	non-conserved
IV	1182755-1182823	non-conserved
VII	16395-17053	non-conserved
III	122464-122483	non-conserved
VII	810962-811065	non-conserved
XV	216345-216430	non-conserved
XV	40168-40299	non-conserved
I	229752-229892	non-conserved
VII	962062-962159	non-conserved
XIV	546430-546468	non-conserved
IV	1005338-1005367	non-conserved
VII	1069616-1069864	non-conserved
IX	316469-316603	non-conserved
XVI	775846-776015	non-conserved
V	53497-53589	non-conserved
XIV	355507-355949	non-conserved
XVI	929803-930343	non-conserved
XIII	634562-634687	non-conserved
IV	25877-25937	non-conserved
II	714637-714744	non-conserved
II	415424-415644	non-conserved
VI	113986-114652	non-conserved
XVI	568997-569134	non-conserved
XV	170507-170558	non-conserved
XIV	405814-405897	non-conserved
IV	477362-477724	non-conserved
XII	898363-898406	non-conserved
XI	304399-304484	non-conserved
XV	1078213-1078478	non-conserved

chromosome number	region coordinates (bp)	control type
IX	93119-93297	non-conserved
IV	443725-443855	non-conserved
VII	1083912-1083968	non-conserved
IV	1013965-1014151	non-conserved
IV	158870-159350	non-conserved
X	724098-724273	non-conserved
XVI	488294-488542	non-conserved
IX	99158-99344	non-conserved
VI	181208-181402	non-conserved
XVI	645550-645589	non-conserved
VII	508642-508703	non-conserved
XV	487521-487571	non-conserved
XIII	923147-923388	non-conserved
XII	292176-292342	non-conserved
VI	258038-258299	non-conserved
XII	949196-949516	non-conserved
X	615332-615574	non-conserved
II	581389-581417	non-conserved
VII	377325-377342	non-conserved
IV	569955-570410	non-conserved
VII	413350-413454	non-conserved
XI	207921-207951	non-conserved
VII	257007-257230	non-conserved
I	87147-87179	non-conserved
XIV	250316-250816	non-conserved
IX	7372-7509	non-conserved
II	527860-527879	non-conserved
IV	1362189-1362307	non-conserved
VIII	412521-412795	non-conserved
X	339488-339519	non-conserved
XIII	512201-512744	non-conserved
VII	202386-202602	non-conserved
IX	289137-289224	non-conserved
XII	980987-981056	non-conserved
VI	245339-245990	non-conserved
II	558199-558265	non-conserved
V	468261-468361	non-conserved
XIV	479569-479587	non-conserved
IX	439609-439667	non-conserved

chromosome number	region coordinates (bp)	control type
XI	325521-325650	non-conserved
III	273666-274101	non-conserved
VII	56058-56233	non-conserved
II	45338-45459	non-conserved
I	158806-158964	non-conserved
XIII	31733-32134	non-conserved
XV	80015-80191	non-conserved
XVI	209971-209983	non-conserved
IV	475990-476554	non-conserved
X	521740-521781	non-conserved
IV	926935-926984	non-conserved
IV	1169013-1169168	non-conserved
XI	457055-457182	non-conserved
XII	823480-823511	non-conserved
XIII	296745-296874	non-conserved
XIV	441113-441274	non-conserved
XIII	795565-795645	non-conserved
XVI	857320-857396	non-conserved
VII	589747-589805	non-conserved
XVI	834505-834559	non-conserved
XII	581939-582004	non-conserved
X	617540-617907	non-conserved
XIV	374961-375115	non-conserved
VII	74965-75137	non-conserved
II	785701-786131	non-conserved
I	181-261	non-conserved
VII	787932-788171	non-conserved
XIV	274080-274209	non-conserved
X	227933-228081	non-conserved
XIV	135812-135938	non-conserved
XV	701313-701392	non-conserved
II	575999-576132	non-conserved
IV	715380-715452	non-conserved
VI	207087-207318	non-conserved
IV	1341992-1342065	non-conserved
VII	857838-858144	non-conserved
VIII	498860-498889	non-conserved
IV	591950-592218	non-conserved
XII	407195-407257	non-conserved

chromosome number	region coordinates (bp)	control type
IV	1050383-1050420	non-conserved
V	53604-53736	non-conserved
XVI	544159-544465	non-conserved
VII	767000-767342	non-conserved
VI	260381-260480	non-conserved
XIII	411288-411349	non-conserved
III	8669-8761	non-conserved
VIII	455706-455746	non-conserved
XV	702585-702680	non-conserved
XV	665146-665335	non-conserved
XIV	87457-87499	non-conserved
XI	615066-615127	non-conserved
VI	174001-174015	non-conserved
XIV	663914-663947	non-conserved
VI	169470-169732	non-conserved
IX	249136-249162	non-conserved
XIV	63104-63257	non-conserved
IV	1494446-1494510	non-conserved
VII	10274-10637	non-conserved
IV	50680-50982	non-conserved
XII	20577-21002	non-conserved
XV	837228-837468	non-conserved
X	154937-154983	non-conserved
V	570137-570322	non-conserved
XII	84089-84139	non-conserved
XIV	59937-60021	non-conserved
III	78618-78945	non-conserved
IV	1162461-1162733	non-conserved
XI	270825-271775	non-conserved
XVI	14840-15051	non-conserved
X	639431-639544	non-conserved
II	604781-604802	non-conserved
XI	21080-21165	non-conserved
XI	374357-374503	non-conserved
VII	857838-858144	non-conserved
XIII	547633-547712	non-conserved
I	212162-212354	non-conserved
IV	1116833-1116961	non-conserved
II	496893-496981	non-conserved

chromosome number	region coordinates (bp)	control type
XVI	287254-287512	non-conserved
IV	103503-103547	non-conserved
VI	11125-11255	non-conserved
IX	27820-28625	non-conserved
IV	2954-3017	non-conserved
VIII	501144-501930	non-conserved
IV	48633-48679	non-conserved
XI	567367-567702	non-conserved
VIII	282383-282477	non-conserved
IV	117484-117531	non-conserved
V	106786-107006	non-conserved
III	178321-178451	non-conserved
XIII	601254-601560	non-conserved
VII	421267-421299	non-conserved
XIII	651227-651588	non-conserved
V	362374-362580	non-conserved
II	592921-593068	non-conserved
V	258316-258735	non-conserved
IV	964893-964929	non-conserved
VIII	258173-258237	non-conserved
XII	510592-510715	non-conserved
XIV	723114-723205	non-conserved
XII	928066-928291	non-conserved
V	504151-504825	non-conserved
XIII	133054-133119	non-conserved
V	39746-39758	non-conserved
VIII	522032-524787	non-conserved
XV	619357-619398	non-conserved
V	263884-264012	non-conserved
VII	959060-959232	non-conserved
VII	107915-107974	non-conserved
VI	255447-255495	non-conserved
XII	84174-84264	non-conserved
XVI	129808-129880	non-conserved
XIV	553119-553334	non-conserved
XV	304256-304413	non-conserved
IV	407760-408138	non-conserved
XI	361980-362326	non-conserved
IV	524918-525039	non-conserved

chromosome number	region coordinates (bp)	control type
IV	141724-141836	non-conserved
V	541174-541552	non-conserved
XVI	264479-264600	non-conserved
V	552228-552407	non-conserved
X	177792-177916	non-conserved
XII	390701-390752	non-conserved
VII	17314-17659	non-conserved
XV	766795-766867	non-conserved
XIII	574016-574469	non-conserved
V	132720-132827	non-conserved
IV	620666-621411	non-conserved
XII	48914-49245	non-conserved
XIII	221335-221394	non-conserved
IV	80606-82437	non-conserved
X	417045-417121	non-conserved
II	410669-410822	non-conserved
VIII	516628-517275	non-conserved
I	220498-220814	non-conserved
VII	1011458-1011604	non-conserved
IV	544754-544875	non-conserved
IV	1005217-1005332	non-conserved
I	119970-120091	non-conserved
V	56296-56421	non-conserved
VI	26101-26811	non-conserved
XVI	582993-583060	non-conserved
VIII	1873-1975	non-conserved
XV	40168-40299	non-conserved
VII	534916-534991	non-conserved
XVI	822303-822371	non-conserved
XIV	65142-65273	non-conserved
X	538424-538542	non-conserved
I	199087-199140	non-conserved
II	186637-186842	non-conserved
M	43795-44046	non-conserved
III	133234-133273	non-conserved
VI	222962-223040	non-conserved
X	727018-727118	non-conserved
I	222103-222378	non-conserved
XIII	512201-512744	non-conserved

chromosome number	region coordinates (bp)	control type
XVI	281771-281926	non-conserved
V	221690-221713	non-conserved
IV	292141-292358	non-conserved
XVI	454991-455040	non-conserved
II	288720-289139	non-conserved
IV	1166558-1166632	non-conserved
I	215708-215877	non-conserved
XIV	335181-335202	non-conserved
X	397116-397353	non-conserved
XII	1064955-1065023	non-conserved
II	476805-477235	non-conserved
XIV	777046-777318	non-conserved
XIV	745789-745854	non-conserved
V	423598-423914	non-conserved
IV	1140956-1141111	non-conserved
XII	20397-20578	non-conserved
XII	539634-539709	non-conserved
XIV	349766-349813	non-conserved
XV	891606-891690	non-conserved
XVI	627503-627525	non-conserved
VII	1069616-1069864	non-conserved
VII	634252-634302	non-conserved
XI	302062-302332	non-conserved
VII	318080-318707	non-conserved
III	316255-316330	non-conserved
IV	1506101-1506587	non-conserved
IV	68930-68995	non-conserved
VI	221643-221712	non-conserved
IX	130084-130099	non-conserved
VI	82059-82344	non-conserved
IV	1090264-1090414	non-conserved
XIII	396591-396906	non-conserved
V	242655-242996	non-conserved
VII	990636-991077	non-conserved
I	82134-82600	non-conserved
IV	132784-132817	non-conserved
II	36728-36753	non-conserved
XI	645918-645934	non-conserved
XV	581492-581736	non-conserved

chromosome number	region coordinates (bp)	control type
IV	1362189-1362307	non-conserved
IV	1270898-1270922	non-conserved
XIV	784024-784073	non-conserved
VII	581738-581770	non-conserved
VIII	48356-48374	non-conserved
II	720605-720653	non-conserved
XVI	98892-99215	non-conserved
VIII	335840-336192	non-conserved
XII	872729-872762	non-conserved
IV	1447179-1447326	non-conserved
XV	463904-464438	non-conserved
VII	323584-323696	non-conserved
XII	786160-786210	non-conserved
VIII	370603-370720	non-conserved
XIII	325480-325627	non-conserved
XII	822870-822911	non-conserved
IV	1365912-1366222	non-conserved
XV	681159-681350	non-conserved
IX	27820-28625	non-conserved
XI	666547-666653	non-conserved
IV	1031583-1031750	non-conserved
III	273666-274101	non-conserved
V	362374-362580	non-conserved
VII	122133-122255	non-conserved
XIII	619748-619841	non-conserved
VIII	123858-123984	non-conserved
VII	905860-905923	non-conserved
XIII	225366-225536	non-conserved
XII	874793-874944	non-conserved
XI	506453-506515	non-conserved
XIII	731636-732154	non-conserved
IX	426715-426811	non-conserved
IV	320528-320666	non-conserved
V	554027-554629	non-conserved
XV	303088-303230	non-conserved
XII	792684-792889	non-conserved
VII	1007362-1007406	non-conserved
IV	477362-477724	non-conserved
XVI	825129-825148	non-conserved

chromosome number	region coordinates (bp)	control type
IX	278140-278183	non-conserved
V	363849-364064	non-conserved
XII	877549-877740	non-conserved
IV	1383640-1383798	non-conserved
II	269759-269892	non-conserved
II	288225-288532	non-conserved
VII	788540-788798	non-conserved
IV	494486-494572	non-conserved
VII	1078916-1079517	non-conserved
XIII	551323-551429	non-conserved
VII	330857-330965	non-conserved
IV	1005145-1005165	non-conserved
XI	638393-638753	non-conserved
II	801930-802564	non-conserved
XIV	635543-635817	non-conserved
X	94711-94850	non-conserved
XV	238515-238617	non-conserved
XV	170507-170558	non-conserved
XVI	205091-205136	non-conserved
VIII	123591-123702	non-conserved
XI	352625-352888	non-conserved
V	264649-264799	non-conserved
XIII	432435-432573	non-conserved
II	349931-350096	non-conserved
XIII	119521-119546	non-conserved
V	504151-504825	non-conserved
XVI	739729-739872	non-conserved
XIII	689011-689081	non-conserved
IV	859303-859336	non-conserved
II	151359-151491	non-conserved
XVI	793805-793909	non-conserved
XVI	843200-843260	non-conserved
XVI	173052-173149	non-conserved
X	639431-639544	non-conserved
VIII	131595-131900	non-conserved
XI	260201-260507	non-conserved
VII	522840-523435	non-conserved
VII	951626-951888	non-conserved
IV	80606-82437	non-conserved

chromosome number	region coordinates (bp)	control type
XV	123762-123860	non-conserved
XVI	321954-322069	non-conserved
XVI	227578-227624	non-conserved
XV	310150-310172	non-conserved
VII	22201-22302	non-conserved
I	210984-211021	non-conserved
XIII	91727-91980	non-conserved
IV	321835-322009	non-conserved
VIII	138282-138402	non-conserved
VII	1083912-1083968	non-conserved
XI	67470-67553	non-conserved
X	629169-629237	non-conserved
IV	436924-437011	non-conserved
II	414818-415197	non-conserved
VII	180611-180644	non-conserved
VII	842639-842744	non-conserved
VI	90790-90860	non-conserved
IV	130610-130646	non-conserved
IV	699828-700017	non-conserved
IV	321835-322009	non-conserved
XVI	525066-525597	non-conserved
IV	1466157-1466229	non-conserved
VII	700303-700604	non-conserved
XII	708767-708794	non-conserved
XIII	271692-271816	non-conserved
IV	410811-410823	non-conserved
VI	153125-153211	non-conserved
XVI	418028-418123	non-conserved
V	271311-271433	non-conserved
VI	255646-255661	non-conserved
X	30365-30611	non-conserved
VI	207087-207318	non-conserved
V	512362-512523	non-conserved
XII	931979-932204	non-conserved
XI	389316-389378	non-conserved
XV	892171-892314	non-conserved
VII	770479-770567	non-conserved
IX	21322-22053	non-conserved
M	56095-56170	non-conserved

chromosome number	region coordinates (bp)	control type
IV	83144-83546	non-conserved
XIV	186886-187019	non-conserved
XI	262030-262273	non-conserved
V	396389-396608	non-conserved
XV	161372-161519	non-conserved
IX	10850-11254	non-conserved
V	362374-362580	non-conserved
II	555447-555664	non-conserved
II	203267-203323	non-conserved
IV	829059-829139	non-conserved
XII	745103-745430	non-conserved
VII	584336-584645	non-conserved
XVI	163803-163920	non-conserved
IV	387176-387324	non-conserved
IV	1523037-1523147	non-conserved
V	53604-53736	non-conserved
IV	506976-506995	non-conserved
XIV	778952-779269	non-conserved
XV	773628-773756	conserved
XI	255096-255189	conserved
XIV	559247-559360	conserved
X	337773-337879	conserved
XV	773755-773854	conserved
VII	1028796-1028905	conserved
XV	217033-217124	conserved
V	483171-483269	conserved
XV	16899-17006	conserved
XVI	572490-572644	conserved
XIII	662107-662289	conserved
XV	947657-947887	conserved
XIII	434492-434588	conserved
XV	725193-725327	conserved
XIII	20562-20687	conserved
III	2571-2700	conserved
XI	100380-100669	conserved
IV	1517850-1518260	conserved
IV	1165026-1165164	conserved
XIII	408822-408915	conserved
VII	856632-856750	conserved

chromosome number	region coordinates (bp)	control type
XVI	592144-592313	conserved
II	336896-337058	conserved
IV	1516676-1517060	conserved
V	140077-140230	conserved
II	363626-363743	conserved
VIII	484655-484767	conserved
XIV	447517-447609	conserved
V	88949-89093	conserved
II	364403-364520	conserved
VII	371385-371493	conserved
IV	867926-868083	conserved
VII	993761-993912	conserved
IV	1005424-1005523	conserved
VII	1000608-1000717	conserved
IV	1514076-1514221	conserved
IV	811949-812044	conserved
V	89657-89749	conserved
VIII	98120-98257	conserved
XI	164571-164684	conserved
II	145318-145493	conserved
IV	1514799-1515064	conserved
II	145077-145319	conserved
VIII	244492-244620	conserved
VII	1029004-1029123	conserved
XI	171364-171474	conserved
V	88278-88554	conserved
XVI	520613-520745	conserved
XIV	335953-336063	conserved
IV	848354-848509	conserved
V	9202-9296	conserved
III	3820-3999	conserved
IV	927058-927166	conserved
XII	645994-646140	conserved
XIV	634737-634995	conserved
XI	338928-339027	conserved
XV	947986-948171	conserved
IV	1515154-1515293	conserved
XVI	585905-586003	conserved
II	164167-164306	conserved

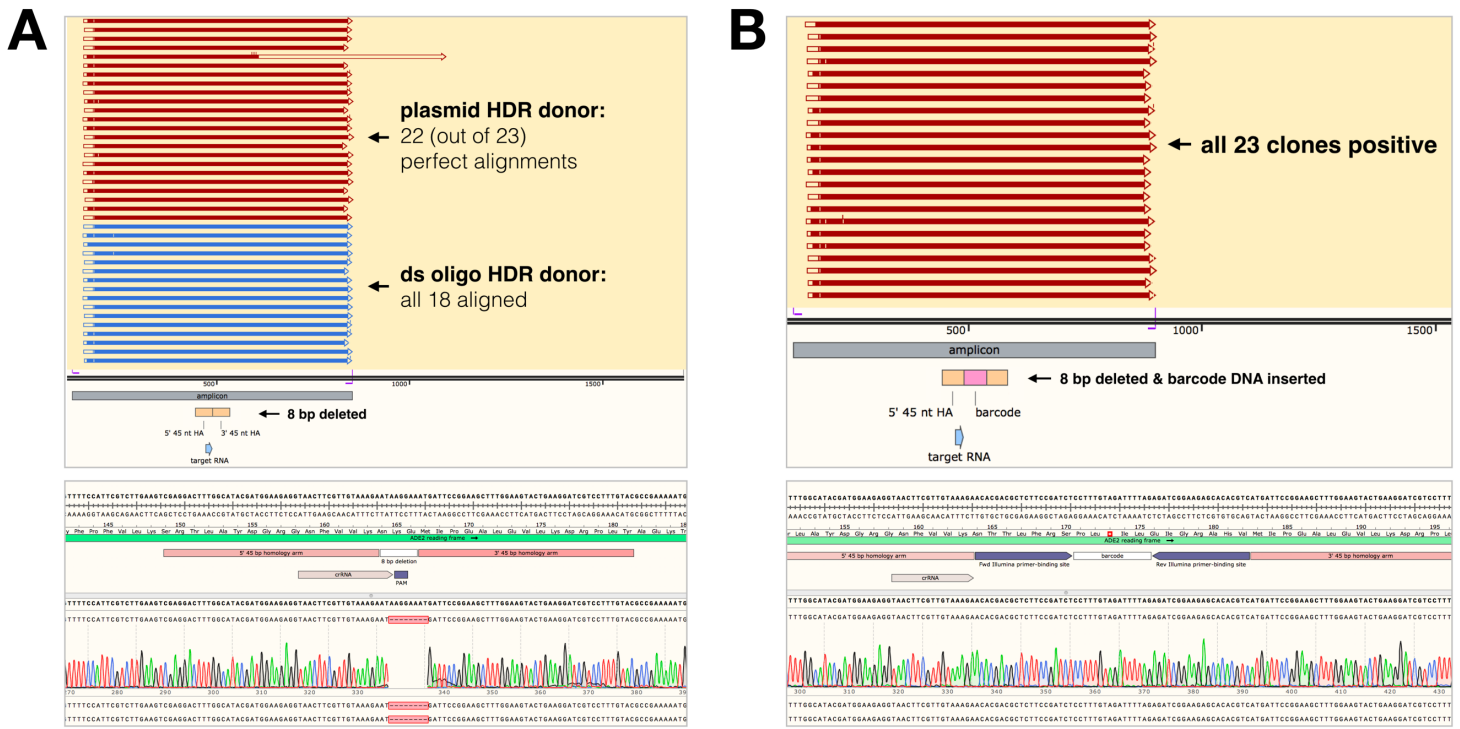
chromosome number	region coordinates (bp)	control type
XVI	497847-497972	conserved
II	165829-165921	conserved
XII	510876-510974	conserved
V	87683-87861	conserved
VIII	67564-67669	conserved
XIV	642506-642605	conserved
XVI	412821-412967	conserved
XVI	863278-863369	conserved
VII	483108-483235	conserved
XV	104019-104113	conserved
X	180384-180502	conserved
XVI	863435-863677	conserved
VII	1029793-1029891	conserved
XII	447675-447920	conserved
XV	17086-17188	conserved
V	338652-338844	conserved
V	139919-140033	conserved
I	227106-227214	conserved
XV	948274-948620	conserved
IV	1515400-1515658	conserved
XIII	624310-624414	conserved
XVI	551786-551896	conserved
XII	826282-826401	conserved
XIV	263504-263609	conserved
VIII	190297-190441	conserved
XVI	862340-862560	conserved
XVI	829273-829434	conserved
V	339119-339385	conserved
XIV	107190-107299	conserved
XIV	662052-662152	conserved
XIII	445289-445393	conserved
XVI	874978-875103	conserved
VIII	509163-509256	conserved
XV	947285-947455	conserved
XV	946688-946810	conserved
IV	416953-417053	conserved
XV	978345-978506	conserved
II	364171-364362	conserved
IV	1515670-1515811	conserved

chromosome number	region coordinates (bp)	control type
XIV	87553-87653	conserved
VIII	56506-56597	conserved
XII	212230-212323	conserved
II	482485-482817	conserved
VII	857640-857732	conserved
IV	49150-49258	conserved
XIII	661992-662095	conserved
IV	411442-411639	conserved
II	342578-342676	conserved
XII	825281-825426	conserved
VII	993904-993996	conserved
X	393288-393557	conserved
VI	164211-164351	conserved
VIII	4803-4928	conserved
X	526030-526153	conserved
IV	867571-867678	conserved
VII	398396-398498	conserved
IV	915929-916023	conserved
XVI	412073-412252	conserved
XII	492926-493115	conserved
VII	277865-278388	conserved
XVI	67124-67223	conserved
III	2700-2822	conserved
II	408772-408873	conserved
IV	1519098-1519267	conserved
IV	1515832-1515949	conserved
II	740720-740825	conserved
XV	768174-768265	conserved
VIII	119813-119926	conserved
VII	150698-150795	conserved
II	482950-483424	conserved
VIII	470673-470780	conserved
XIII	445103-445205	conserved
XI	336322-336427	conserved
III	293033-293177	conserved
VI	4686-4813	conserved
IV	173512-173769	conserved
II	36811-36914	conserved
IV	889097-889216	conserved

chromosome number	region coordinates (bp)	control type
XVI	464052-464168	conserved
VII	610118-610237	conserved
VII	1011826-1011933	conserved
XIII	306111-306222	conserved
VII	610251-610342	conserved
XVI	412282-412394	conserved
VIII	2238-2595	conserved
XVI	280027-280181	conserved
V	107010-107101	conserved
VI	164564-164742	conserved
VI	261312-261463	conserved
V	201772-201911	conserved
XIII	810589-810721	conserved
XII	921681-921778	conserved
XVI	863171-863264	conserved
IV	1071542-1071636	conserved
XII	825500-825606	conserved
IV	461402-461495	conserved
XI	381974-382070	conserved
XIII	439011-439108	conserved
II	680107-680241	conserved
XII	416819-416936	conserved
XIV	447421-447513	conserved
XIV	692062-692159	conserved
XII	286619-286729	conserved
VII	884069-884164	conserved
IV	1459419-1459536	conserved
VI	54947-55043	conserved
XIV	174438-174591	conserved
XV	947455-947546	conserved
XV	128034-128174	conserved
XII	341163-341263	conserved
IV	1514330-1514521	conserved
XIV	452140-452246	conserved
X	521786-521878	conserved
IX	93422-93514	conserved
XI	175051-175301	conserved
X	623143-623234	conserved
XI	159092-159212	conserved

chromosome number	region coordinates (bp)	control type
XII	441289-441466	conserved
XVI	297911-298031	conserved
XIV	547726-547835	conserved

Appendix 3.2 Supplementary Figures

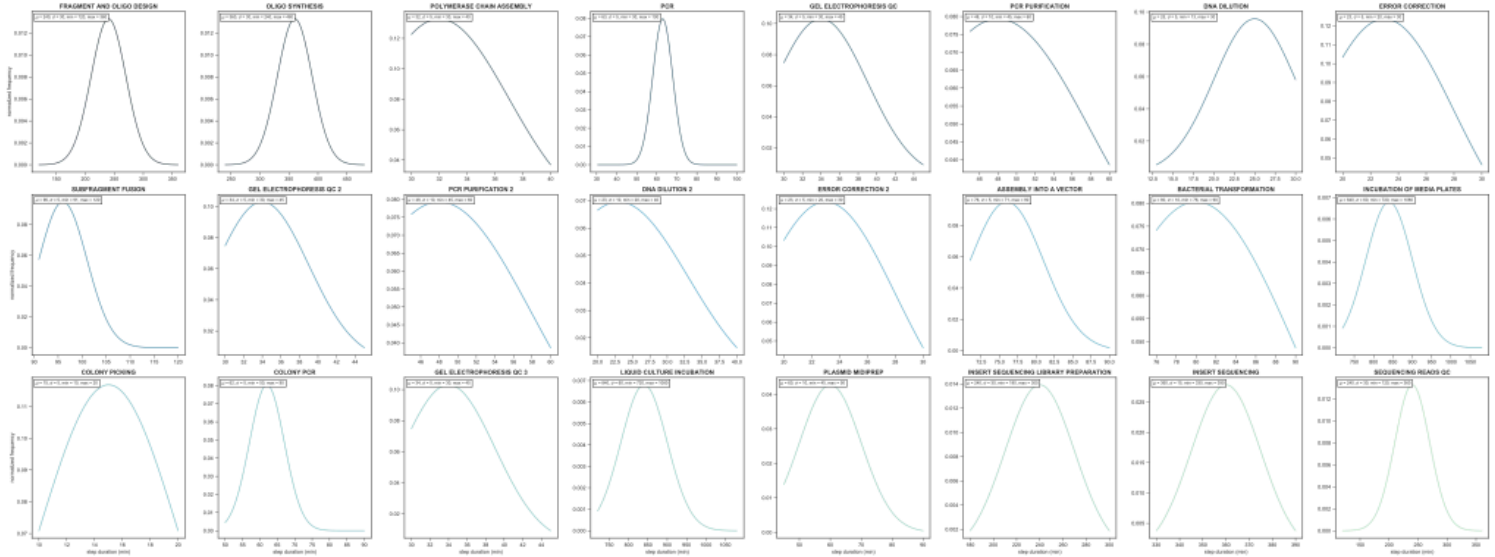


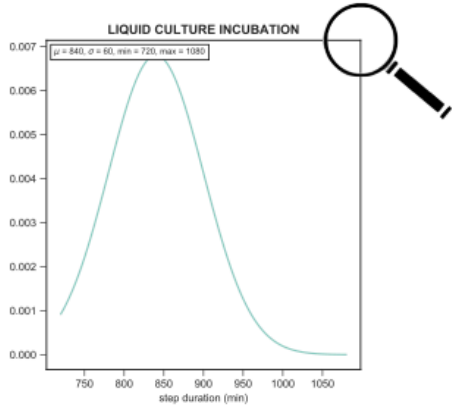
Supp. Figure 2.2.1 Sanger Sequencing Verification of Single-gene CRISPR Deletion Experiments

(**A**) *ADE2* deletion experiments introducing an 8 bp frameshift-triggering mutation, deleting the Cas9::sgRNA recognition sequences (i.e., the protospacer and protospacer-adjacent motif sequences); Plasmid DNA from > 15 red yeast colonies (putative *ADE2* mutants) was sequenced and the sequencing results confirmed successful non-essential gene deletions. Test (plasmid HDR donor) and positive control (dsDNA HDR donor) experiments were DNA sequencing-verified. (**B**) *ADE2* deletion experiments introducing the 8 bp deletions and additional barcode DNA insertions, causing pre-mature stop codon generation; Plasmid DNA from > 15 red yeast colonies was sequenced and the sequencing results confirmed *ADE2* gene knockouts.

Appendix 4

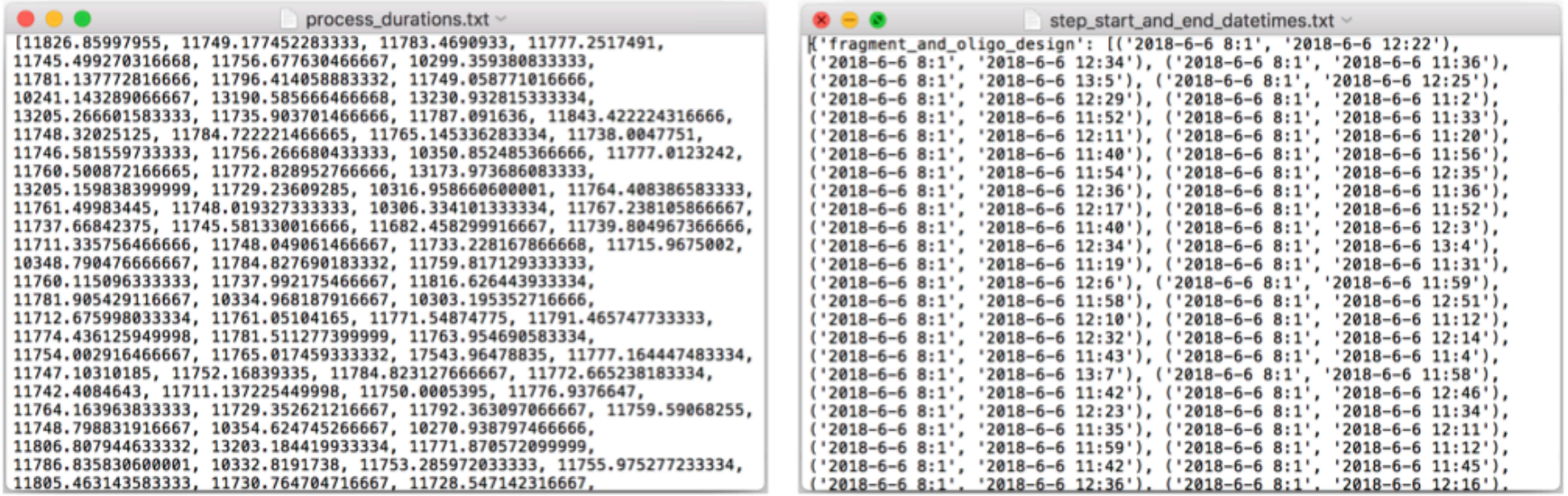
Appendix 4.1 Supplementary Figures





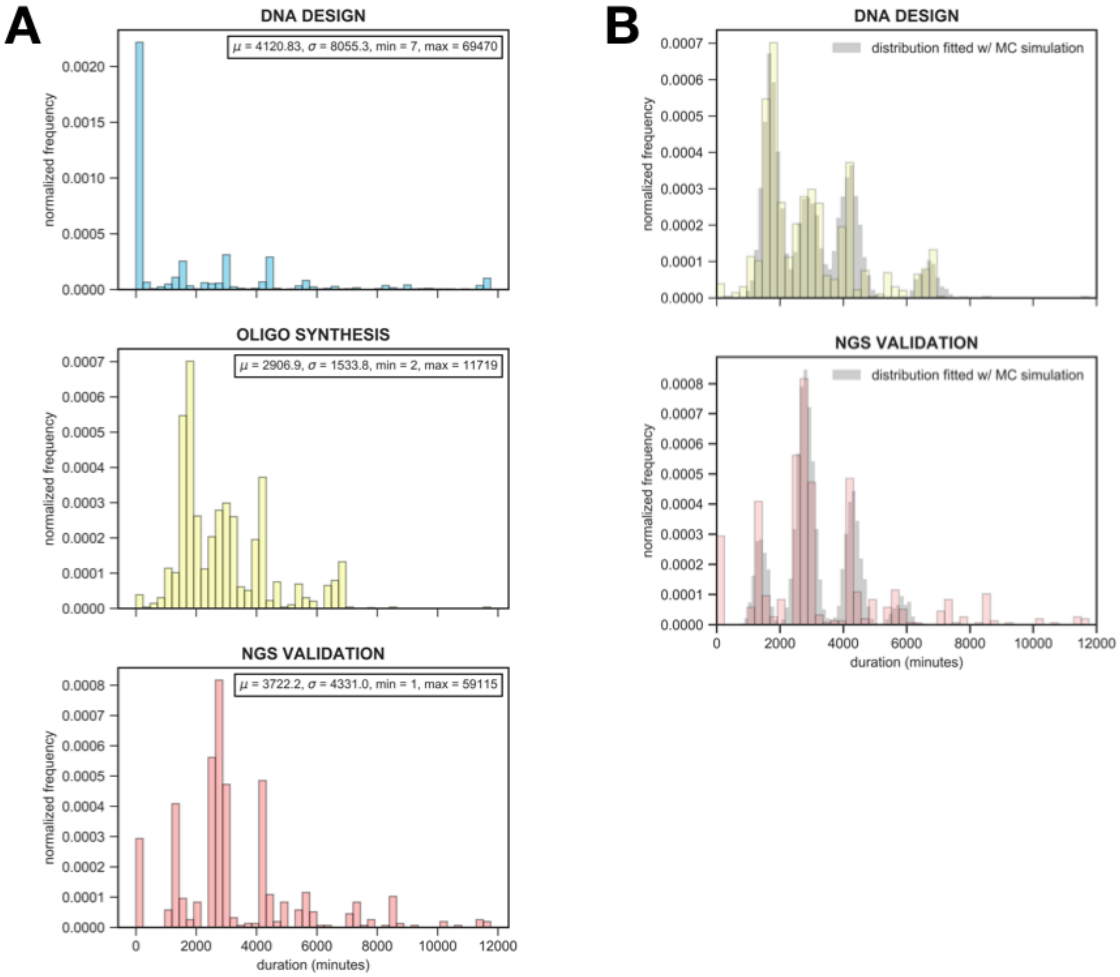
Supp. Figure 4.1.1 Automatically-generated Plot of Input Manufacturing Step Duration Distributions

Probability density functions of the input truncated normal distributions, with their mean, standard deviation, minimum and maximum bound parameters indicated, are one of the automatically-generated plots returned by the Monte Carlo simulation tool.



Supp. Figure 4.1.2 Automatically-generated Simulation Results Files

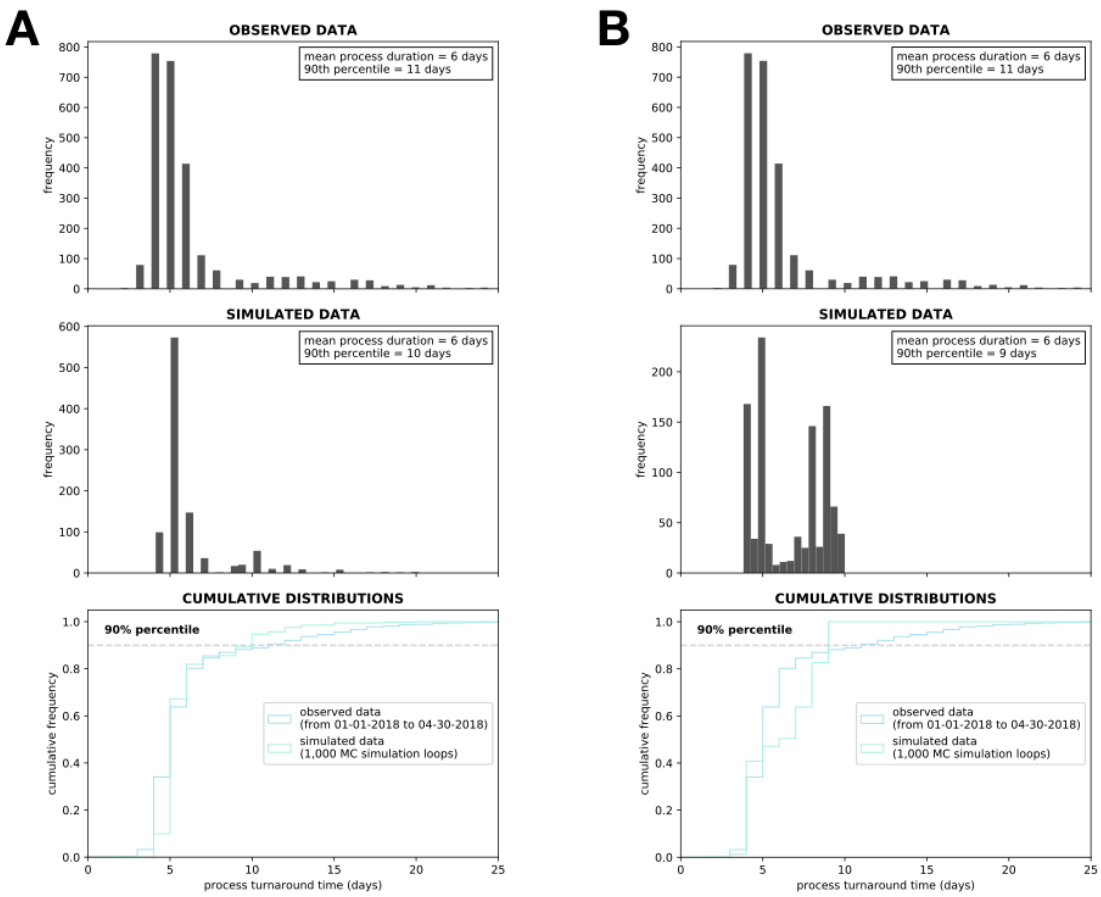
Two simulation results files are returned by the Monte Carlo simulation tool. The “process_durations.txt” file stores all simulated process durations (in minutes), while the “step_start_and_end_datetimes.txt” file stores information on all simulated start and end times of the individual model manufacturing steps. Data from the “process_durations.txt” file can be used to perform model sensitivity analysis using the second simulation tool developed. The raw data output can be used to perform other custom analyses as well.



Supp. Figure 4.1.3 Manufacturing Timestamp Data Mining and Simulation-aided Fitting of Turnaround Time Distributions

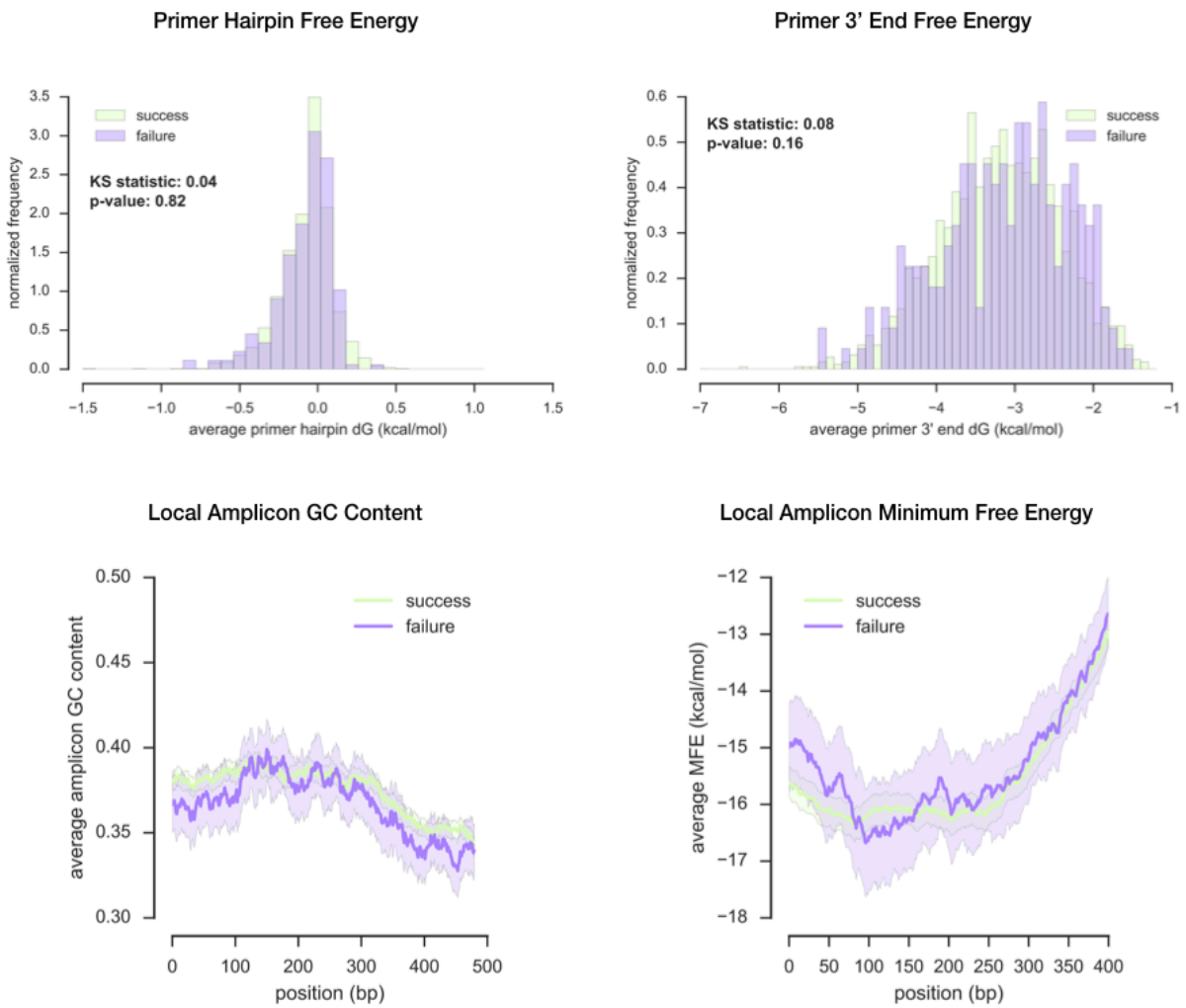
(**A**) To estimate turnaround time distributions of DNA design and the oligo synthesis and NGS manufacturing stages, start and end timestamps of these manufacturing phases were retrieved from the company's manufacturing execution system, which stores information on individual DNA orders. The time difference between the end and start times was computed to yield durations of the investigated production phases. Durations corresponding to non-working days (i.e., weekends) were subtracted from this data. Distributions of the turnaround times obtained are illustrated. All three distributions are multimodal and thus not Gaussian. For the oligo synthesis and NGS validation manufacturing stages, this data suggests different manufacturing rescue routes involved in the underlying DNA orders. Possible data collection artefacts were observed in the last two turnaround time distributions, e.g., DNA orders with oligo syntheses and NGS validations shorter than 5 min. (**B**) The oligo synthesis and NGS turnaround time distributions were recapitulated with Monte Carlo simulation experiments. The two manufacturing stages were split into a number of manufacturing steps. Their number was based on the number of distribution modes. Durations of these steps were next modelled with normal distributions (which complied with the modelling framework developed). Parameters of the step

duration distributions, as well as their failure probabilities, were adjusted through simulation experiments so that the output manufacturing stage distributions resembled the observed duration data. Fitting of a mathematical distribution to the DNA design duration data was not successful and therefore is not presented.



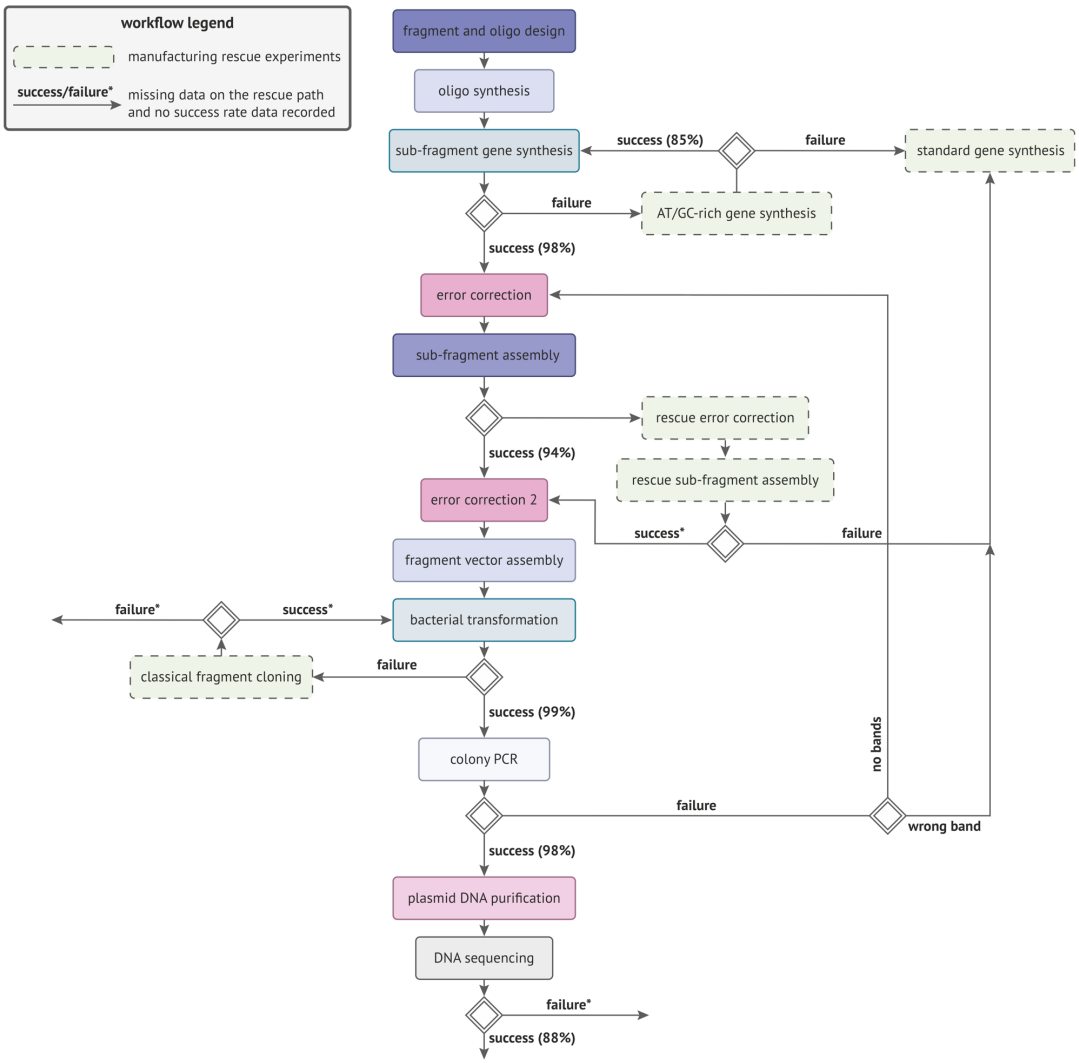
Supp. Figure 4.1.4 Using the Simulation-estimated Turnaround Time Distributions to Simulate the Example Process Model

(**A**) Results of simulation experiments based on data including assumptions regarding durations of DNA design, oligonucleotide synthesis and NGS (**B**) Results of simulation experiments based on the simulation-estimated turnaround time distributions of the DNA oligonucleotide synthesis and NGS manufacturing stages; The DNA design stage was not considered by the alternative process model as it appeared to take a negligible amount of time (according to the timestamp-derived turnaround time distribution). In order to use the simulation-derived distribution data, the corresponding manufacturing steps were assigned an additional attribute called “run whenever”, which allowed them to be unconstrained by the pre-defined working schedules as information on staff availabilities was not applicable to modelling of these manufacturing steps. A larger difference between the simulated and observed mean production times (2 days) was observed. Similarly to the observed data distribution, distribution of the simulated DNA manufacturing times was multimodal, however with a much heavier tail. Due to these discrepancies, timestamp data-derived distributions were thus not used to model the example manufacturing process.



Supp. Figure 4.1.5 Impact of DNA Sequence Features on Failure of DNA Amplification

Data regarding DNA amplification failures and the underlying DNA sequences (500 bp-long) was obtained from the YeastFab project database (Guo *et al*., 2015). Impact of different global and local nucleic acid sequence features on the amplification failure was investigated. No significant difference (as judged by a two-sample Kolmogorov-Smirnov test) was observed between distributions of the successful and failed DNA amplifications for the global features. Similarly, no significant difference was observed when the local features were investigated in 100 bp DNA windows. Mean GC content and minimum free energy are indicated by the plotted lines. Shaded areas represent the corresponding 95% confidence intervals.



Supp. Figure 4.1.6 Diagram of the Example DNA Manufacturing Workflow

Diagram of the case study DNA manufacturing workflow; Manufacturing rescue paths are indicated and include manufacturing steps with multiple rescue options (i.e., the colony PCR step); Several manufacturing success rates were not recorded by the MES system.

Appendix 4.2 Supplementary Tables

Supp. Table 4.2.1 Model of the Example DNA Manufacturing Process

step no.	step name	minimum duration (min)	maximum duration (min)	mean duration (min)	standard deviation (min)	duration distribution	team resource	next step	step failure probability	rescue step	runs overnight
0	completed	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
1	fragment and oligo design	120	360	240	30	truncated normal	IT	2	0	N/A	no
2	oligo synthesis	240	480	360	30	truncated normal	technicians	3	0	N/A	no
3	polymerase chain assembly	30	40	32	5	truncated normal	technicians	4	0	N/A	no
4	PCR	30.5	100	63	5	truncated normal	technicians	5	0	N/A	no
5	gel electrophoresis QC	30	45	34	5	truncated normal	technicians	6	2	3	no
6	PCR purification	45	60	48	10	truncated normal	technicians	7	0	N/A	no
7	DNA dilution	13	30	25	5	truncated normal	technicians	8	0	N/A	no
8	error correction	20	30	23	5	truncated normal	technicians	9	0	N/A	no
9	sub-fragment fusion	91	120	96	5	truncated normal	technicians	10	0	N/A	no
10	gel electrophoresis QC 2	30	45	34	5	truncated normal	technicians	11	6	8	no
11	PCR purification 2	45	60	48	10	truncated normal	technicians	12	0	N/A	no
12	DNA dilution 2	20	40	23	10	truncated normal	technicians	13	0	N/A	no
13	fragment refinement	20	30	23	5	truncated normal	technicians	14	0	N/A	no
14	assembly into a vector	71	90	76	5	truncated normal	technicians	15	0	N/A	no
15	bacterial transformation	76	90	80	10	truncated normal	technicians	16	0	N/A	no
16	incubation of media plates	720	1080	840	60	truncated normal	technicians	17	1	15	yes
17	colony picking	10	20	15	5	truncated normal	technicians	18	0	N/A	no
18	colony PCR	50	90	62	5	truncated normal	technicians	19	0	N/A	no
19	gel electrophoresis QC 3	30	45	34	5	truncated normal	technicians	20	2	8	no
20	liquid culture incubation	720	1080	840	60	truncated normal	technicians	21	0	N/A	yes
21	plasmid midiprep	45	90	60	10	truncated normal	technicians	22	0	N/A	no
22	insert sequencing library prep.	180	300	240	30	truncated normal	technicians	23	0	N/A	no
23	insert sequencing	330	390	360	15	truncated normal	technicians	24	0	N/A	yes
24	sequencing reads QC	120	360	240	30	truncated normal	IT	0	12	2	no

Supp. Table 4.2.2 Process Model – Incorporating Timestamp Data-derived Distribution Estimations

step no.	step name	minimum duration (min)	maximum duration (min)	mean duration (min)	standard deviation (min)	duration distribution	team resource	next step	step failure probability	rescue step	runs overnight	run whenever
0	completed	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
1	oligo synthesis	0	5000000	1700	250	truncated normal	N/A	5	60	2	N/A	yes
2	oligo synthesis rescue 1	0	5000000	1200	200	truncated normal	N/A	5	60	3	N/A	yes
3	oligo synthesis rescue 2	0	5000000	1300	1	truncated normal	N/A	5	20	4	N/A	yes
4	oligo synthesis rescue 3	0	5000000	2500	1	truncated normal	N/A	5	0	5	N/A	yes
5	PCA	30	40	32	5	truncated normal	technicians	6	0	N/A	no	no
6	PCR	30.5	100	63	5	truncated normal	technicians	7	0	N/A	no	no
7	gel electrophoresis QC	30	45	34	5	truncated normal	technicians	8	2	3	no	no
8	PCR purification	45	60	48	10	truncated normal	technicians	9	0	N/A	no	no
9	DNA dilution	13	30	25	5	truncated normal	technicians	10	0	N/A	no	no
10	error correction	20	30	23	5	truncated normal	technicians	11	0	N/A	no	no
11	sub-fragment fusion	91	120	96	5	truncated normal	technicians	12	0	N/A	no	no
12	gel electrophoresis QC 2	30	45	34	5	truncated normal	technicians	13	6	10	no	no
13	PCR purification 2	45	60	48	10	truncated normal	technicians	14	0	N/A	no	no
14	DNA dilution 2	20	40	23	10	truncated normal	technicians	15	0	N/A	no	no
15	fragment refinement	20	30	23	5	truncated normal	technicians	16	0	N/A	no	no
16	assembly into a vector	71	90	76	5	truncated normal	technicians	17	0	N/A	no	no
17	bacterial transformation	76	90	80	10	truncated normal	technicians	18	0	N/A	no	no
18	incubation of plates	720	1080	840	60	truncated normal	technicians	19	1	16	yes	no
19	colony picking	10	20	15	5	truncated normal	technicians	20	0	N/A	no	no
20	colony PCR	50	90	62	5	truncated normal	technicians	21	0	N/A	no	no
21	gel electrophoresis QC 3	30	45	34	5	truncated normal	technicians	22	2	10	no	no
22	liquid culture incubation	720	1080	840	60	truncated normal	technicians	23	0	N/A	yes	no
23	plasmid midiprep	45	90	60	10	truncated normal	technicians	24	0	N/A	no	no
24	NGS	0	5000000	1400	200	truncated normal	N/A	0	85	25	N/A	yes
25	NGS rescue 1	0	5000000	1400	150	truncated normal	N/A	0	35	26	N/A	yes
26	NGS rescue 2	0	5000000	1500	1	truncated normal	N/A	0	15	27	N/A	yes
27	NGS rescue 3	0	5000000	1500	1	truncated normal	N/A	0	0	0	N/A	yes

Appendix 4.3 Supplementary Code

4.3.1 Monte Carlo Simulation Tool

```
import pandas as pd
import numpy as np
import math
import operator
from collections import defaultdict
import datetime as dt
import matplotlib.pyplot as plt
import matplotlib.dates as mdates
from mpl_toolkits.axes_grid.anchored_artists import AnchoredText
import seaborn as sns
from scipy.stats import truncnorm

from subroutines_v2 import (simulate_model_n_times_w_time_constraints)

from plotting_v2 import (
plot_process_durations_pdf_cdf,
plot_simulated_model_steps_in_process_runtime,
plot_input_pdfs)

def mc_simulation(file_name, nloops):
    """Returns and saves 3 figures, a .txt file with n simulated process durations and a .txt file with n simulated step start and end datetimes"""

    model = pd.read_excel(file_name, sheet_name="process_model") #file name has to be in parentheses

    start_date_and_time = pd.read_excel(file_name, sheet_name="start_datetime")
    start_datetime = dt.datetime.strptime(
```

```
    start_date_and_time.iloc[0].start_datetime, "%Y.%m.%d.%H.%M"
)

working_hours = pd.read_excel(file_name, sheet_name="working_hours")
free_weekdays = pd.read_excel(file_name, sheet_name="free_weekdays")

model["step_number"] = model.step_number.astype(int)
model["next_step"] = model.next_step.astype(int)
model["next_rescue_step"] = model.next_rescue_step.astype(int)

step_datetimes, process_durations = simulate_model_n_times_w_time_constraints(model, start_datetime, working_hours, free_weekdays, nloops)

model_steps = []

for key, value in step_datetimes.items():
    for i in range(len(value)):
        model_steps.append(key)

start_datetimes = []

for key, value in step_datetimes.items():
    for item in value:
        start_datetimes.append(dt.datetime.strptime(item[0], "%Y-%m-%d %H:%M"))

end_datetimes = []

for key, value in step_datetimes.items():
    for item in value:
        end_datetimes.append(dt.datetime.strptime(item[1], "%Y-%m-%d %H:%M"))

max_end_datetime = max(end_datetimes)
```

```
fig1 = plot_process_durations_pdf_cdf(process_durations, nloops)
fig2 = plot_simulated_model_steps_in_process_runtime(model,
                                                     nloops,
                                                     start_datetime,
                                                     model_steps,
                                                     start_datetimes,
                                                     end_datetimes,
                                                     max_end_datetime)
fig3 = plot_input_pdfs(model)

with open("process_durations.txt", "w") as output:
    output.write(";".join([str(d) for d in process_durations]))

with open("step_start_and_end_datetimes.txt", "w") as output:
    output.write(str(step_datetimes))

fig1.savefig("fig1_process_durations_pdf_cdf.pdf", format="pdf", bbox_inches="tight")
fig2.savefig("fig2_simulated_model_steps_in_process_runtime.png", format="png", bbox_inches="tight", dpi=1000)
fig3.savefig("fig3_input_pdfs.pdf", format="pdf", bbox_inches="tight")

return fig1, fig2, fig3
```

4.3.2 Subroutines of the Simulation Tool

```
import pandas as pd
import numpy as np
import math
import operator
from collections import defaultdict
import datetime as dt
import matplotlib.pyplot as plt
import matplotlib.dates as mdates
from mpl_toolkits.axes_grid.anchored_artists import AnchoredText
import seaborn as sns
from scipy.stats import truncnorm

def random_step_duration(model_step):
    """Returns a random number sampled from a truncated normal distribution"""

    mean = model_step.mean_duration
    stdev = model_step.standard_deviation
    minimum = model_step.minimum_duration
    maximum = model_step.maximum_duration
    a, b = (minimum - mean)/stdev, (maximum - mean)/stdev
    return truncnorm.rvs(a, b, loc=mean, scale=stdev)

def step_is_failing(model_step):
    """Returns True if model_step failing"""

    failure_probability = model_step.step_failure_probability
    return (np.random.randint(0, 100) <= failure_probability)

def step_p95(model_step):
```

```
    """Returns model_step 95th percentile"""

    mean = model_step.mean_duration
    stdev = model_step.standard_deviation
    minimum = model_step.minimum_duration
    maximum = model_step.maximum_duration
    a, b = (minimum - mean)/stdev, (maximum - mean)/stdev
    return truncnorm(a, b, loc=mean, scale=stdev).ppf(0.95)

def compute_percentile(results_list, n):
    """Return simulation result_list nth percentile"""

    return np.percentile(
        np.asarray(results_list),
        n
    )

def can_start_at_t(model_step, t, human_start, human_end, free_days):
    """Returns True if model_step can be executed at time t"""

    weekday_dict = {
        "Mon": 0,
        "Tue": 1,
        "Wed": 2,
        "Thu": 3,
        "Fri": 4,
        "Sat": 5,
        "Sun": 6
    }

    if t.strftime('%a') in free_days: #is it weekend?
```

```
        return False

    if not (human_start < t.time() < human_end): #is process current time within working hours?
        return False

    step_end = t + dt.timedelta(minutes=step_p95(model_step)) #can we finish a step before end of working hours?

    if model_step.runs_overnight == "no":
        next_evening = t.replace(hour=human_end.hour,
                                 minute=human_end.minute)
        return step_end < next_evening

    else:
        next_last_day_evening = t + dt.timedelta(days=max(0, (weekday_dict[free_days[0]] - 1) - t.weekday()))
        next_last_day_evening = next_last_day_evening.replace(hour=human_end.hour,
                                                              minute=human_end.minute)
        return step_end < next_last_day_evening #don't do overnight steps on the last working day

def simulate_model_w_time_constraints(model, start_datetime, working_hours, free_weekdays):
    """Returns a dictionary with step_name keys and their values,
    as tuples of string start and end date and time results;
    and a total process turnaround time (in min)"""

    step_number = 1
    current_datetime = start_datetime
    simulation_result = {}

    while (step_number != 0):

        model_step = model.iloc[step_number]
```

```
#human working schedule for model_step
human_working_schedule = working_hours[working_hours.team_name == model_step.team_resource]
human_start = dt.time(hour=int(human_working_schedule.start_hour),
                      minute=int(human_working_schedule.start_minute))
human_end = dt.time(hour=int(human_working_schedule.end_hour),
                    minute=int(human_working_schedule.end_hour))

#free weekdays
free_days = list(free_weekdays.free_weekdays.values)

#can model_step start at current_datetime? if not, keep checking every 1 min and wait until it can start
while not can_start_at_t(model_step, current_datetime, human_start, human_end, free_days):
    current_datetime = current_datetime + dt.timedelta(minutes=1)

#now, model_step can be executed (save its start and end datetimes)
step_start = "%d-%d-%d %d:%d" % (
    current_datetime.year,
    current_datetime.month,
    current_datetime.day,
    current_datetime.hour,
    current_datetime.minute)

current_datetime += dt.timedelta(minutes=random_step_duration(model_step))

step_end = "%d-%d-%d %d:%d" % (
    current_datetime.year,
    current_datetime.month,
    current_datetime.day,
    current_datetime.hour,
    current_datetime.minute)
```

```
        simulation_result[model_step.step_name] = (step_start, step_end)

        #did it fail?
        if step_is_failing(model_step):
            step_number = model_step.next_rescue_step
        else:
            step_number = model_step.next_step

    return simulation_result, (current_datetime - start_datetime).total_seconds()/60 #total turnaround time in minutes

def simulate_model_n_times_w_time_constraints(model, start_datetime, working_hours, free_weekdays, ntimes):
    """Returns a dictionary of model_step keys and values - lists of n (start, end datetime) tuples;
    and a list of n total process durations (in min)"""

    simulation_results = [
        simulate_model_w_time_constraints(model, start_datetime, working_hours, free_weekdays) for i in range(ntimes)]

    step_datetimes = {
        key: [dictionary[key] for dictionary in [item[0] for item in simulation_results]]
        for key in [item[0] for item in simulation_results][0]
    }
    process_durations = [item[1] for item in simulation_results]
    return step_datetimes, process_durations

def compute_process_sensitivity(model, start_datetime, working_hours, free_weekdays, process_durations, nloops):
    """Returns an absolute value-sorted dictionary with sensitivity analysis results (values) for each model parameter (key)"""

    model_parameters_dictionary = {}

    for i, row in model.iterrows():
```

```
failure_probability_df = model.set_value(i, "step_failure_probability", 1.01*row.step_failure_probability)
mean_duration_df = model.set_value(i, "mean_duration", 1.01*row.mean_duration)
min_duration_df = model.set_value(i, "minimum_duration", 1.01*row.minimum_duration)
max_duration_df = model.set_value(i, "maximum_duration", 1.01*row.maximum_duration)
standard_deviation_df = model.set_value(i, "standard_deviation", 1.01*row.standard_deviation)

if i != 0:
    if row.step_failure_probability != 0:
        failure_probability_dict = {
            (row.step_name, "failure probability"):
            np.mean(simulate_model_n_times_w_time_constraints(failure_probability_df,
                                                              start_datetime,
                                                              working_hours,
                                                              free_weekdays, nloops)[1])-np.mean(process_durations)
        }
        mean_duration_dict = {
            (row.step_name, "mean duration"):
            np.mean(simulate_model_n_times_w_time_constraints(mean_duration_df,
                                                              start_datetime,
                                                              working_hours,
                                                              free_weekdays, nloops)[1])-np.mean(process_durations)
        }
        min_duration_dict = {
            (row.step_name, "min duration"):
            np.mean(simulate_model_n_times_w_time_constraints(min_duration_df,
                                                              start_datetime,
                                                              working_hours,
                                                              free_weekdays, nloops)[1])-np.mean(process_durations)
        }
        max_duration_dict = {
            (row.step_name, "max duration"):
```

```
        np.mean(simulate_model_n_times_w_time_constraints(max_duration_df,
                                                          start_datetime,
                                                          working_hours,
                                                          free_weekdays, nloops)[1])-np.mean(process_durations)
    }
    standard_deviation_dict = {
        (row.step_name, "max duration"):
        np.mean(simulate_model_n_times_w_time_constraints(standard_deviation_df,
                                                          start_datetime,
                                                          working_hours,
                                                          free_weekdays, nloops)[1])-np.mean(process_durations)
    }

    model_parameters_dictionary.update(failure_probability_dict)
    model_parameters_dictionary.update(mean_duration_dict)
    model_parameters_dictionary.update(min_duration_dict)
    model_parameters_dictionary.update(max_duration_dict)
    model_parameters_dictionary.update(standard_deviation_dict)

else:
    mean_duration_dict = {
        (row.step_name, "mean duration"):
        np.mean(simulate_model_n_times_w_time_constraints(mean_duration_df,
                                                          start_datetime,
                                                          working_hours,
                                                          free_weekdays, nloops)[1])-np.mean(process_durations)
    }
    min_duration_dict = {
        (row.step_name, "min duration"):
        np.mean(simulate_model_n_times_w_time_constraints(min_duration_df,
                                                          start_datetime,
```

```
                                                        working_hours,
                                                        free_weekdays, nloops)[1])-np.mean(process_durations)
            }
            max_duration_dict = {
                (row.step_name, "max duration"):
                np.mean(simulate_model_n_times_w_time_constraints(max_duration_df,
                                                        start_datetime,
                                                        working_hours,
                                                        free_weekdays, nloops)[1])-np.mean(process_durations)
            }
            standard_deviation_dict = {
                (row.step_name, "max duration"):
                np.mean(simulate_model_n_times_w_time_constraints(standard_deviation_df,
                                                        start_datetime,
                                                        working_hours,
                                                        free_weekdays, nloops)[1])-np.mean(process_durations)
            }

            model_parameters_dictionary.update(mean_duration_dict)
            model_parameters_dictionary.update(min_duration_dict)
            model_parameters_dictionary.update(max_duration_dict)
            model_parameters_dictionary.update(standard_deviation_dict)
    else:
        pass

sorted_dict = sorted(model_parameters_dictionary.items(), key=lambda it: abs(it[1])) #list output
sorted_dict_keys = [sorted_dict[entry][0] for entry in range(len(sorted_dict))]
sorted_dict_values = [sorted_dict[entry][1] for entry in range(len(sorted_dict))]

return sorted_dict, sorted_dict_keys, sorted_dict_values
```

4.3.3 Sensitivity Analysis Tool

```
import pandas as pd
import numpy as np
import math
import operator
from collections import defaultdict
import datetime as dt
import matplotlib.pyplot as plt
import matplotlib.dates as mdates
from mpl_toolkits.axes_grid.anchored_artists import AnchoredText
import seaborn as sns
from scipy.stats import truncnorm

from subroutines_v2 import (compute_process_sensitivity)
from plotting_v2 import (plot_top_parameters)

def sensitivity_analysis(model_file_name, simulation_results_file_name):
    """Returns a figure of process sensitivity to top 20 most significant model parameters"""

    model = pd.read_excel(model_file_name, sheet_name="process_model") #file name has to be in parentheses

    start_date_and_time = pd.read_excel(model_file_name, sheet_name="start_datetime")
    start_datetime = dt.datetime.strptime(
        start_date_and_time.iloc[0].start_datetime, "%Y.%m.%d.%H.%M"
    )

    working_hours = pd.read_excel(model_file_name, sheet_name="working_hours")
    free_weekdays = pd.read_excel(model_file_name, sheet_name="free_weekdays")

    model["step_number"] = model.step_number.astype(int)
```

```
model["next_step"] = model.next_step.astype(int)
model["next_rescue_step"] = model.next_rescue_step.astype(int)

with open(simulation_results_file_name, 'r') as f:
    content = f.read()

process_durations = [float(v) for v in content.split(';')]
nloops = len(process_durations)

sorted_dict, sorted_dict_keys, sorted_dict_values = compute_process_sensitivity(
    model,
    start_datetime,
    working_hours,
    free_weekdays,
    process_durations,
    nloops
)

fig = plot_top_parameters(sorted_dict, sorted_dict_keys, sorted_dict_values)
fig.savefig("fig1_sensitivity_analysis_results.pdf", format="pdf", bbox_inches="tight")

return fig
```

Bibliography

Abadi, S. *et al.* (2017) ‘A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action’, *PLOS Computational Biology*. Edited by H. Xu. Public Library of Science, 13(10), p. e1005807. doi: 10.1371/journal.pcbi.1005807.

Agarwal, K. L. *et al.* (1976) ‘Total synthesis of the structural gene for the precursor of a tyrosine suppressor transfer RNA from Escherichia coli. 6. Synthesis of the deoxyribopolynucleotide segments corresponding to the nucleotide sequence 100-126.’, *The Journal of biological chemistry*, 251(3), pp. 624–33.

Ahmad, M. and Bussey, H. (1986) ‘Yeast arginine permease: nucleotide sequence of the CAN1 gene.’, *Current genetics*, 10(8), pp. 587–92.

Aird, D. *et al.* (2011) ‘Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries.’, *Genome biology*. BioMed Central, 12(2), p. R18. doi: 10.1186/gb-2011-12-2-r18.

Allentoft, M. E. *et al.* (2012) ‘The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils.’, *Proceedings. Biological sciences*. The Royal Society, 279(1748), pp. 4724–33. doi: 10.1098/rspb.2012.1745.

Andreou, A. I. and Nakayama, N. (2018) ‘Mobius Assembly: A versatile Golden-Gate framework towards universal DNA assembly’. doi: 10.1371/journal.pone.0189892.

Annaluru, N. *et al.* (2014) ‘Total Synthesis of a Functional Designer Eukaryotic Chromosome’, *Science*, 344(6179), pp. 55–58. doi: 10.1126/science.1249252.

Appleton, E. *et al.* (2014) ‘Interactive assembly algorithms for molecular cloning’, *Nature Methods*, 11(6), pp. 657–662. doi: 10.1038/nmeth.2939.

Appleton, E. *et al.* (2017) ‘Design Automation in Synthetic Biology.’, *Cold Spring Harbor perspectives in biology*. Cold Spring Harbor Laboratory Press, 9(4), p. a023978. doi: 10.1101/cshperspect.a023978.

Astle, T. W. (2016a) ‘Microplate Standardization Report’:, *https://doi.org/10.1177/221106829800300112*. SAGE PublicationsSage CA: Los

Angeles, CA. doi: 10.1177/221106829800300112.
Astle, T. W. (2016b) 'Recollections of Early Microplate Automation':, *https://doi.org/10.1016/S1535-5535-04-00103-0*. SAGE PublicationsSage CA: Los Angeles, CA. doi: 10.1016/S1535-5535-04-00103-0.
Aubel, D. and Fussenegger, M. (2010) 'Mammalian synthetic biology - from tools to therapies', *BioEssays*. Wiley-Blackwell, 32(4), pp. 332–345. doi: 10.1002/bies.200900149.
Bao, Z. *et al.* (2015) 'Homology-Integrated CRISPR–Cas (HI-CRISPR) System for One-Step Multigene Disruption in *Saccharomyces cerevisiae*', *ACS Synthetic Biology*, 4(5), pp. 585–594. doi: 10.1021/sb500255k.
Bao, Z. *et al.* (2018) 'Genome-scale engineering of Saccharomyces cerevisiae with single-nucleotide precision', *Nature Biotechnology*. Nature Publishing Group, 36(6), pp. 505–508. doi: 10.1038/nbt.4132.
Beller, H. R., Lee, T. S. and Katz, L. (2015) 'Natural products as biofuels and bio-based chemicals: fatty acids and isoprenoids', *Natural Product Reports*. Royal Society of Chemistry, 32(10), pp. 1508–1526. doi: 10.1039/C5NP00068H.
Bentley, D. L. (2014) 'Coupling mRNA processing with transcription in time and space', *Nature Reviews Genetics*. Nature Publishing Group, 15(3), pp. 163–175. doi: 10.1038/nrg3662.
Bertomeu, T. *et al.* (2018) 'A High-Resolution Genome-Wide CRISPR/Cas9 Viability Screen Reveals Structural Features and Contextual Diversity of the Human Cell-Essential Proteome.', *Molecular and cellular biology*. American Society for Microbiology (ASM), 38(1). doi: 10.1128/MCB.00302-17.
Bétermier, M., Bertrand, P. and Lopez, B. S. (2014) 'Is Non-Homologous End-Joining Really an Inherently Error-Prone Process?', *PLOS Genetics*. Public Library of Science, 10(1), p. e1004086. doi: 10.1371/JOURNAL.PGEN.1004086.
Blakes, J. *et al.* (2014) 'Heuristic for Maximizing DNA Reuse in Synthetic DNA Library Assembly', *ACS Synthetic Biology*. American Chemical Society, 3(8), pp. 529–542. doi: 10.1021/sb400161v.
Boeke, J. D. *et al.* (2016) 'GENOME ENGINEERING. The Genome Project-Write.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 353(6295), pp. 126–7. doi: 10.1126/science.aaf6850.

Boeke, J. D., La Croute, F. and Fink, G. R. (1984) ‘A positive selection for mutants lacking orotidine-5′-phosphate decarboxylase activity in yeast: 5-fluoro-orotic acid resistance’, *Molecular and General Genetics MGG*. Springer-Verlag, 197(2), pp. 345–346. doi: 10.1007/BF00330984.

Botev, Z. and L’Ecuyer, P. (2017) ‘Simulation from the Normal Distribution Truncated to an Interval in the Tail’, in *10th EAI International Conference on Performance Evaluation Methodologies and Tools*.

Brookes, B. C. (1955) ‘Limit Distributions for Sums of Independent Random Variables. By B.V. Gnedenko and A.N. Kolmogorov. Translated by K.L. Chung. Pp. ix, 264. $7.50. 1954. (Addison-Wesley, Cambridge, Mass.)’, *The Mathematical Gazette*. Cambridge University Press, 39(330), pp. 342–343. doi: 10.2307/3608621.

Cagney, G. *et al.* (2006) ‘Functional genomics of the yeast DNA-damage response.’, *Genome Biology*, 7(9), p. 233. doi: 10.1186/gb-2006-7-9-233.

Cameron, D. E., Bashor, C. J. and Collins, J. J. (2014) ‘A brief history of synthetic biology’, *Nature Reviews Microbiology*. Nature Publishing Group, 12(5), pp. 381–390. doi: 10.1038/nrmicro3239.

Carbonell, P. *et al.* (2018) ‘An automated Design-Build-Test-Learn pipeline for enhanced microbial production of fine chemicals’, *Communications Biology*. Nature Publishing Group, 1(1), p. 66. doi: 10.1038/s42003-018-0076-9.

Carr, P. A. *et al.* (2004) ‘Protein-mediated error correction for de novo DNA synthesis.’, *Nucleic acids research*. Oxford University Press, 32(20), p. e162. doi: 10.1093/nar/gnh160.

Caruthers, M. H. *et al.* (1987) ‘[15] Chemical synthesis of deoxyoligonucleotides by the phosphoramidite method’, *Methods in Enzymology*. Academic Press, 154, pp. 287–313. doi: 10.1016/0076-6879(87)54081-2.

Ceccaldi, R., Rondinelli, B. and D’Andrea, A. D. (2016) ‘Repair Pathway Choices and Consequences at the Double-Strand Break’, *Trends in Cell Biology*. Elsevier, 26(1), pp. 52–64. doi: 10.1016/j.tcb.2015.07.009.

Cello, J., Paul, A. V and Wimmer, E. (2002) ‘Chemical Synthesis of Poliovirus cDNA: Generation of Infectious Virus in the Absence of Natural Template’, *Science*, 297(5583), pp. 1016–1018. doi: 10.1126/science.1072266.

Chambers, S., Kitney, R. and Freemont, P. (2016) ‘The Foundry: the DNA synthesis

and construction Foundry at Imperial College.', *Biochemical Society transactions*. Portland Press Ltd, 44(3), pp. 687–8. doi: 10.1042/BST20160007.

Chandran, S. (2017) 'Rapid Assembly of DNA via Ligase Cycling Reaction (LCR)', in *Methods in molecular biology (Clifton, N.J.)*, pp. 105–110. doi: 10.1007/978-1-4939-6343-0_8.

Chao, R. *et al.* (2017) 'Engineering biological systems using automated biofoundries', *Metabolic Engineering*. Academic Press, 42, pp. 98–108. doi: 10.1016/J.YMBEN.2017.06.003.

Check Hayden, E. (2014) 'The automated lab', *Nature*, 516(7529), pp. 131–132. doi: 10.1038/516131a.

Cheng, A. A. and Lu, T. K. (2012) 'Synthetic Biology: An Emerging Engineering Discipline', *Annual Review of Biomedical Engineering*. Annual Reviews , 14(1), pp. 155–178. doi: 10.1146/annurev-bioeng-071811-150118.

Chou, H. H. and Keasling, J. D. (2013) 'Programming adaptive control to evolve increased metabolite production', *Nature Communications 2013 4*. Nature Publishing Group, 4, p. 2595. doi: 10.1038/ncomms3595.

Church, G. M. *et al.* (2012) 'Next-generation digital information storage in DNA.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 337(6102), p. 1628. doi: 10.1126/science.1226355.

Church, G. M., Gao, Y. and Kosuri, S. (2012) 'Next-Generation Digital Information Storage in DNA', *Science*, 337(6102), pp. 1628–1628. doi: 10.1126/science.1226355.

Cleary, M. A. *et al.* (2004) 'Production of complex nucleic acid libraries using highly parallel in situ oligonucleotide synthesis', *Nature Methods*, 1(3), pp. 241–248. doi: 10.1038/nmeth724.

Cleveland, P. H. and Koutz, P. J. (2005) 'Nanoliter Dispensing for uHTS Using Pin Tools', *https://home.liebertpub.com/adt*. Mary Ann Liebert, Inc. 2 Madison Avenue Larchmont, NY 10538 USA . doi: 10.1089/ADT.2005.3.213.

Clore, A. (2018) 'A new route to synthetic DNA', *Nature Biotechnology*. Nature Publishing Group, 36(7), pp. 593–595. doi: 10.1038/nbt.4185.

Coleman, J. R. *et al.* (2008) 'Virus attenuation by genome-scale changes in codon pair bias.', *Science (New York, N.Y.)*. NIH Public Access, 320(5884), pp. 1784–7.

doi: 10.1126/science.1155761.
Cook, J. D. (2010) *Determining distribution parameters from quantiles*.
Costanzo, M. *et al.* (2016) 'A global genetic interaction network maps a wiring diagram of cellular function.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 353(6306), p. aaf1420. doi: 10.1126/science.aaf1420.
Craig, T. *et al.* (2017) 'Leaf LIMS: A Flexible Laboratory Information Management System with a Synthetic Biology Focus', *ACS Synthetic Biology*. American Chemical Society, 6(12), pp. 2273–2280. doi: 10.1021/acssynbio.7b00212.
Czar, M. J., Cai, Y. and Peccoud, J. (2009) 'Writing DNA with GenoCAD.', *Nucleic acids research*. Oxford University Press, 37(Web Server issue), pp. W40-7. doi: 10.1093/nar/gkp361.
Dawes, T. D. *et al.* (2016) 'Compound Transfer by Acoustic Droplet Ejection Promotes Quality and Efficiency in Ultra-High-Throughput Screening Campaigns', *Journal of Laboratory Automation*. SAGE PublicationsSage CA: Los Angeles, CA, 21(1), pp. 64–75. doi: 10.1177/2211068215590588.
Decourty, L. *et al.* (2014) 'Long Open Reading Frame Transcripts Escape Nonsense-Mediated mRNA Decay in Yeast', *Cell Reports*. Cell Press, 6(4), pp. 593–598. doi: 10.1016/J.CELREP.2014.01.025.
Deming, W. E. (William E. (2000) *Out of the crisis*. MIT Press.
Dhungel, N. and Hopper, A. K. (2012) 'Beyond tRNA cleavage: novel essential function for yeast tRNA splicing endonuclease unrelated to tRNA processing.', *Genes & development*. Cold Spring Harbor Laboratory Press, 26(5), pp. 503–14. doi: 10.1101/gad.183004.111.
DiCarlo, J. E. *et al.* (2013) 'Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems.', *Nucleic acids research*. Oxford University Press, 41(7), pp. 4336–43. doi: 10.1093/nar/gkt135.
Doench, J. G. *et al.* (2014) 'Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation', *Nature Biotechnology*. Nature Publishing Group, 32(12), pp. 1262–1267. doi: 10.1038/nbt.3026.
Douglas, S. M. *et al.* (2009) 'Self-assembly of DNA into nanoscale three-dimensional shapes', *Nature 2009 459:7245*. Nature Publishing Group, 459(7245), p. 414. doi: 10.1038/nature08016.

Duina, A. A., Miller, M. E. and Keeney, J. B. (2014) 'Budding yeast for budding geneticists: a primer on the Saccharomyces cerevisiae model system.', *Genetics*. Genetics Society of America, 197(1), pp. 33–48. doi: 10.1534/genetics.114.163188.
Duportet, X. *et al.* (2014) 'A platform for rapid prototyping of synthetic gene networks in mammalian cells', *Nucleic Acids Research*. Oxford University Press, 42(21), pp. 13440–13451. doi: 10.1093/nar/gku1082.
Dymond, J. S. *et al.* (2011) 'Synthetic chromosome arms function in yeast and generate phenotypic diversity by design', *Nature*, 477(7365), pp. 471–476. doi: 10.1038/nature10403.
Egeland, R. D. and Southern, E. M. (2005) 'Electrochemically directed synthesis of oligonucleotides for DNA microarray fabrication', *Nucleic Acids Research*. Oxford University Press, 33(14), pp. e125–e125. doi: 10.1093/nar/gni117.
Ellis, T., Adie, T. and Baldwin, G. S. (2011) 'DNA assembly for synthetic biology: from parts to pathways and beyond', *Integrative Biology*, 3(2), p. 109. doi: 10.1039/c0ib00070a.
Ellson, R. *et al.* (2016) 'Transfer of Low Nanoliter Volumes between Microplates Using Focused Acoustics—Automation Considerations':, *https://doi.org/10.1016/S1535-5535-03-00011-X*. SAGE PublicationsSage CA: Los Angeles, CA. doi: 10.1016/S1535-5535-03-00011-X.
Elowitz, M. B. and Leibler, S. (2000) 'A synthetic oscillatory network of transcriptional regulators', *Nature*. Nature Publishing Group, 403(6767), pp. 335–338. doi: 10.1038/35002125.
Endy, D. (2005) 'Foundations for engineering biology'. doi: 10.1038/nature04342.
Engler, C., Kandzia, R. and Marillonnet, S. (2008) 'A One Pot, One Step, Precision Cloning Method with High Throughput Capability', *PLoS ONE*. Edited by H. A. El-Shemy. Public Library of Science, 3(11), p. e3647. doi: 10.1371/journal.pone.0003647.
Eric LeProust *et al.* (2000) 'Digital Light-Directed Synthesis. A Microarray Platform That Permits Rapid Reaction Optimization on a Combinatorial Basis'. American Chemical Society . doi: 10.1021/CC000009X.
Eroshenko, N. *et al.* (2012) 'Gene Assembly from Chip-Synthesized Oligonucleotides.', *Current protocols in chemical biology*. NIH Public Access, 2012.

doi: 10.1002/9780470559277.ch110190.
Esvelt, K. M. and Wang, H. H. (2014) 'Genome-scale engineering for systems and synthetic biology', *Molecular Systems Biology*, 9(1), pp. 641–641. doi: 10.1038/msb.2012.66.
Evert, B. A. *et al.* (2004) 'Spontaneous DNA Damage in *Saccharomyces cerevisiae* Elicits Phenotypic Properties Similar to Cancer Cells', *Journal of Biological Chemistry*, 279(21), pp. 22585–22594. doi: 10.1074/jbc.M400468200.
Extance, A. (2016) 'How DNA could store all the world's data', *Nature*, 537(7618), pp. 22–24. doi: 10.1038/537022a.
Fang, J. and Dorrestein, P. C. (2014) 'Emerging mass spectrometry techniques for the direct analysis of microbial colonies.', *Current opinion in microbiology*. NIH Public Access, 19, pp. 120–129. doi: 10.1016/j.mib.2014.06.014.
Fiers, W. *et al.* (1976) 'Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene', *Nature*. Nature Publishing Group, 260(5551), pp. 500–507. doi: 10.1038/260500a0.
Fleischmann, R. D. *et al.* (1995) 'Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.', *Science (New York, N.Y.)*, 269(5223), pp. 496–512.
Florea, M. *et al.* (2016) 'Engineering control of bacterial cellulose production using a genetic toolkit and a new cellulose-producing strain', *Proceedings of the National Academy of Sciences*, 113(24), pp. E3431–E3440. doi: 10.1073/pnas.1522985113.
Fodor, S. P. *et al.* (1993) 'Multiplexed biochemical assays with biological chips.', *Nature*, 364(6437), pp. 555–6. doi: 10.1038/364555a0.
FREGEAU, C. *et al.* (2007) 'Optimized Configuration of Fixed-Tip Robotic Liquid-Handling Stations for the Elimination of Biological Sample Cross-Contamination', *Journal of the Association for Laboratory Automation*. SAGE PublicationsSage CA: Los Angeles, CA, 12(6), pp. 339–354. doi: 10.1016/j.jala.2007.08.001.
Friedland, A. E. *et al.* (2009) 'Synthetic gene networks that count.', *Science (New York, N.Y.)*. NIH Public Access, 324(5931), pp. 1199–202. doi: 10.1126/science.1172005.
Futcher, A. B. and Cox, B. S. (1984) 'Copy number and the stability of 2-micron circle-based artificial plasmids of Saccharomyces cerevisiae.', *Journal of bacteriology*. American Society for Microbiology (ASM), 157(1), pp. 283–90.

Gach, P. C. *et al.* (2017) 'Droplet microfluidics for synthetic biology', *Lab on a Chip*. The Royal Society of Chemistry, 17(20), pp. 3388–3400. doi: 10.1039/C7LC00576H.
Gaj, T. *et al.* (2016) 'Genome-Editing Technologies: Principles and Applications.', *Cold Spring Harbor perspectives in biology*. Cold Spring Harbor Laboratory Press, 8(12), p. a023754. doi: 10.1101/cshperspect.a023754.
Galanie, S. *et al.* (2015) 'Complete biosynthesis of opioids in yeast.', *Science (New York, N.Y.)*. NIH Public Access, 349(6252), pp. 1095–100. doi: 10.1126/science.aac9373.
Galdzicki, M. *et al.* (2014) 'The Synthetic Biology Open Language (SBOL) provides a community standard for communicating designs in synthetic biology', *Nature Biotechnology*. Nature Publishing Group, 32(6), pp. 545–550. doi: 10.1038/nbt.2891.
Gao, X. *et al.* (2001) 'A flexible light-directed DNA chip synthesis gated by deprotection using solution photogenerated acids.', *Nucleic acids research*, 29(22), pp. 4744–50.
Gardner, T. S. (2013) 'Synthetic biology: from hype to impact', *Trends in Biotechnology*, 31, pp. 123–125. doi: 10.1016/j.tibtech.2012.12.001.
Gardner, T. S., Cantor, C. R. and Collins, J. J. (2000) 'Construction of a genetic toggle switch in Escherichia coli', *Nature*. Nature Publishing Group, 403(6767), pp. 339–342. doi: 10.1038/35002131.
Garst, A. D. *et al.* (2016) 'Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering', *Nature Biotechnology*. Nature Publishing Group, 35(1), pp. 48–55. doi: 10.1038/nbt.3718.
Gedvilaite, A. and Sasnauskas, K. (1994) 'Control of the expression of the ADE2 gene of the yeast Saccharomyces cerevisiae.', *Current genetics*, 25(6), pp. 475–9.
Giaever, G. *et al.* (2002) 'Functional profiling of the Saccharomyces cerevisiae genome', *Nature*, 418(6896), pp. 387–391. doi: 10.1038/nature00935.
Giaever, G. and Nislow, C. (2014) 'The yeast deletion collection: a decade of functional genomics.', *Genetics*. Genetics Society of America, 197(2), pp. 451–65. doi: 10.1534/genetics.114.161620.
Gibson, D. G. *et al.* (2008) 'Complete Chemical Synthesis, Assembly, and Cloning of a Mycoplasma genitalium Genome', *Science*, 319(5867), pp. 1215–1220. doi:

10.1126/science.1151721.
Gibson, D. G. *et al.* (2009) 'Enzymatic assembly of DNA molecules up to several hundred kilobases', *Nature Methods 2009 6:5*. Nature Publishing Group, 6(5), p. 343. doi: 10.1038/nmeth.1318.
Gibson, D. G. *et al.* (2010) 'Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome', *Science*, 329(5987), pp. 52–56. doi: 10.1126/science.1190719.
Goffeau, A. *et al.* (1996) 'Life with 6000 genes.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 274(5287), pp. 546, 563–7. doi: 10.1126/SCIENCE.274.5287.546.
Goldberg, M. (2013) 'BioFab: Applying Moore's Law to DNA Synthesis', *Industrial Biotechnology*. Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA , 9(1), pp. 10–12. doi: 10.1089/ind.2012.1552.
Goldman, N. *et al.* (2013) 'Towards practical, high-capacity, low-maintenance information storage in synthesized DNA', *Nature*. Nature Publishing Group, 494(7435), pp. 77–80. doi: 10.1038/nature11875.
Goler, J. A. 1980- (2004) 'BioJADE : a design and simulation tool for synthetic biological systems'. Massachusetts Institute of Technology.
Green, E. D., Rubin, E. M. and Olson, M. V. (2017) 'The future of DNA sequencing', *Nature*, 550(7675), pp. 179–181. doi: 10.1038/550179a.
Greiner, E. (1894) 'New Automatic Pipette', *J. Am. Chem. Soc.*, 16(9), pp. 643–644.
Gross, B. C. *et al.* (2014) 'Evaluation of 3D Printing and Its Potential Impact on Biotechnology and the Chemical Sciences'. American Chemical Society. doi: 10.1021/AC403397R.
Guo, X. *et al.* (2018) 'High-throughput creation and functional profiling of DNA sequence variant libraries using CRISPR-Cas9 in yeast.', *Nature biotechnology*, 36(6), pp. 540–546. doi: 10.1038/nbt.4147.
Guo, Y. *et al.* (2015) 'YeastFab: the design and construction of standard biological parts for metabolic engineering in *Saccharomyces cerevisiae*', *Nucleic Acids Research*, 43(13), pp. e88–e88. doi: 10.1093/nar/gkv464.
Haeussler, M. *et al.* (2016) 'Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR', *Genome Biology*.

BioMed Central, 17(1), p. 148. doi: 10.1186/s13059-016-1012-2.
Ham, T. S. *et al.* (2012) 'Design, implementation and practice of JBEI-ICE: an open source biological part registry platform and tools', *Nucleic Acids Research*, 40(18), pp. e141–e141. doi: 10.1093/nar/gks531.
Haurwitz, R. E. *et al.* (2010) 'Sequence- and Structure-Specific RNA Processing by a CRISPR Endonuclease', *Science*. American Association for the Advancement of Science, 329(5997), pp. 1355–1358. doi: 10.1126/science.1192272.
Hou, J. P. and Poole, J. W. (1969) 'Kinetics and Mechanism of Degradation of Ampicillin in Solution', *Journal of Pharmaceutical Sciences*. Wiley-Blackwell, 58(4), pp. 447–454. doi: 10.1002/jps.2600580412.
Houseley, J. and Tollervey, D. (2009) 'The Many Pathways of RNA Degradation', *Cell*, 136(4), pp. 763–776. doi: 10.1016/j.cell.2009.01.019.
Hsu, P. D. *et al.* (2013) 'DNA targeting specificity of RNA-guided Cas9 nucleases', *Nature Biotechnology*. Nature Publishing Group, 31(9), pp. 827–832. doi: 10.1038/nbt.2647.
Hua, S. *et al.* (1997) 'Minimum Length of Sequence Homology Required forin VivoCloning by Homologous Recombination in Yeast', *Plasmid*, 38(2), pp. 91–96. doi: 10.1006/plas.1997.1305.
Huang, T. *et al.* (2011) 'Analysis and prediction of translation rate based on sequence and functional features of the mRNA.', *PloS one*. Public Library of Science, 6(1), p. e16036. doi: 10.1371/journal.pone.0016036.
Hutchison, C. A. *et al.* (2016) 'Design and synthesis of a minimal bacterial genome.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 351(6280), p. aad6253. doi: 10.1126/science.aad6253.
Ibrahim, E. *et al.* (2018) 'Recombinant *E. coli* Cellulases, β-Glucosidase, and Polygalacturonase Convert a Citrus Processing Waste into Biofuel Precursors', *ACS Sustainable Chemistry & Engineering*. American Chemical Society, 6(6), pp. 7304–7312. doi: 10.1021/acssuschemeng.7b04518.
Itaya, M. *et al.* (2008) 'Bottom-up genome assembly using the Bacillus subtilis genome vector', *Nature Methods*, 5(1), pp. 41–43. doi: 10.1038/nmeth1143.
Iverson, S. V. *et al.* (2016) 'CIDAR MoClo: Improved MoClo Assembly Standard and New *E. coli* Part Library Enable Rapid Combinatorial Design for Synthetic and

Traditional Biology', *ACS Synthetic Biology*. American Chemical Society, 5(1), pp. 99–103. doi: 10.1021/acssynbio.5b00124.

Jahangirian, M. *et al.* (2010) 'Simulation in manufacturing and business: A review', *European Journal of Operational Research*. North-Holland, 203(1), pp. 1–13. doi: 10.1016/J.EJOR.2009.06.004.

Jasin, M. and Rothstein, R. (2013) 'Repair of strand breaks by homologous recombination.', *Cold Spring Harbor perspectives in biology*. Cold Spring Harbor Laboratory Press, 5(11), p. a012740. doi: 10.1101/cshperspect.a012740.

Jeschek, M., Gerngross, D. and Panke, S. (2016) 'Rationally reduced libraries for combinatorial pathway optimization minimizing experimental effort', *Nature Communications*. Nature Publishing Group, 7, p. 11163. doi: 10.1038/ncomms11163.

Jessop-Fabre, M. M. *et al.* (2016) 'EasyClone-MarkerFree: A vector toolkit for marker-less integration of genes into *Saccharomyces cerevisiae* via CRISPR-Cas9', *Biotechnology Journal*. Wiley-Blackwell, 11(8), pp. 1110–1117. doi: 10.1002/biot.201600147.

Jiang, Q. *et al.* (2012) 'DNA Origami as a Carrier for Circumvention of Drug Resistance', *Journal of the American Chemical Society*. American Chemical Society, 134(32), pp. 13396–13403. doi: 10.1021/ja304263n.

Jiang, W. *et al.* (2013) 'RNA-guided editing of bacterial genomes using CRISPR-Cas systems', *Nature Biotechnology*. Nature Publishing Group, 31(3), pp. 233–239. doi: 10.1038/nbt.2508.

Johnson, J. R. *et al.* (2016) 'GeneMill: A 21st century platform for innovation.', *Biochemical Society transactions*. Portland Press Limited, 44(3), pp. 681–3. doi: 10.1042/BST20160012.

Joung, J. K. and Sander, J. D. (2012) 'TALENs: a widely applicable technology for targeted genome editing', *Nature Reviews Molecular Cell Biology 2012 14:1*. Nature Publishing Group, 14(1), p. 49. doi: 10.1038/nrm3486.

Juhas, M. and Ajioka, J. W. (2016) 'High molecular weight DNA assembly in vivo for synthetic biology applications', *http://dx.doi.org/10.3109/07388551.2016.1141394*. Taylor & Francis. doi: 10.3109/07388551.2016.1141394.

Kanigowska, P. *et al.* (2016) 'Smart DNA Fabrication Using Sound Waves', *Journal of Laboratory Automation*, 21(1), pp. 49–56. doi: 10.1177/2211068215593754.
Katre, G. *et al.* (2018) 'Optimization of the in situ transesterification step for biodiesel production using biomass of Yarrowia lipolytica NCIM 3589 grown on waste cooking oil', *Energy*. Pergamon, 142, pp. 944–952. doi: 10.1016/J.ENERGY.2017.10.082.
Katz, L. *et al.* (2018) 'Synthetic biology advances and applications in the biotechnology industry: a perspective', *Journal of Industrial Microbiology & Biotechnology*. Springer International Publishing, 45(7), pp. 449–461. doi: 10.1007/s10295-018-2056-y.
Kelwick, R. *et al.* (2015) 'Promoting microbiology education through the iGEM synthetic biology competition', *FEMS Microbiology Letters*. Edited by B. Fahnert. Oxford University Press, 362(16), p. fnv129. doi: 10.1093/femsle/fnv129.
Kelwick, R. *et al.* (2017) 'Cell-free prototyping strategies for enhancing the sustainable production of polyhydroxyalkanoates bioplastics', *bioRxiv*. Cold Spring Harbor Laboratory, p. 225144. doi: 10.1101/225144.
Khilko, Y. *et al.* (2018) 'DNA assembly with error correction on a droplet digital microfluidics platform.', *BMC biotechnology*. BioMed Central, 18(1), p. 37. doi: 10.1186/s12896-018-0439-9.
Kim, E., Moore, B. S. and Yoon, Y. J. (2015) 'Reinvigorating natural product combinatorial biosynthesis with synthetic biology', *Nature Chemical Biology 2015 11:9*. Nature Publishing Group, 11(9), p. 649. doi: 10.1038/nchembio.1893.
King, R. D. *et al.* (2009) 'The automation of science.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 324(5923), pp. 85–9. doi: 10.1126/science.1165620.
King, R. D., Feng, C. and Sutherland, A. (1995) 'STATLOG: COMPARISON OF CLASSIFICATION ALGORITHMS ON LARGE REAL-WORLD PROBLEMS', *Applied Artificial Intelligence*. Taylor & Francis, 9(3), pp. 289–333. doi: 10.1080/08839519508945477.
Knight, T. (2003) 'Idempotent Vector Design for Standard Assembly of Biobricks'. MIT Artificial Intelligence Laboratory; MIT Synthetic Biology Working Group.
Kobayashi, H. *et al.* (2004) 'Programmable cells: interfacing natural and engineered

gene networks.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 101(22), pp. 8414–9. doi: 10.1073/pnas.0402940101.

Kohavi, R. (1995) *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*.

Kong, F. *et al.* (2012) 'Automatic Liquid Handling for Life Science', *Journal of Laboratory Automation*. SAGE PublicationsSage CA: Los Angeles, CA, 17(3), pp. 169–185. doi: 10.1177/2211068211435302.

Kosuri, S. *et al.* (2010) 'Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips', *Nature Biotechnology*, 28(12), pp. 1295–1299. doi: 10.1038/nbt.1716.

Kosuri, S. and Church, G. M. (2014) 'Large-scale de novo DNA synthesis: technologies and applications', *Nature Methods*. Nature Publishing Group, 11(5), pp. 499–507. doi: 10.1038/nmeth.2918.

Kouprina, N. and Larionov, V. (2016) 'Transformation-associated recombination (TAR) cloning for genomics studies and synthetic biology.', *Chromosoma*. NIH Public Access, 125(4), pp. 621–32. doi: 10.1007/s00412-016-0588-3.

Kristiansson, E. *et al.* (2009) 'Evolutionary Forces Act on Promoter Length: Identification of Enriched Cis-Regulatory Elements', *Molecular Biology and Evolution*, 26(6), pp. 1299–1307. doi: 10.1093/molbev/msp040.

Kroese, D. P. *et al.* (2014) 'Why the Monte Carlo method is so important today', *Wiley Interdisciplinary Reviews: Computational Statistics*. Wiley-Blackwell, 6(6), pp. 386–392. doi: 10.1002/wics.1314.

Kurata, M. *et al.* (2017) 'CRISPR/Cas9 library screening for drug target discovery', *Journal of Human Genetics 2017 63:2*. Nature Publishing Group, 63(2), p. 179. doi: 10.1038/s10038-017-0376-9.

Kuzmin, E. *et al.* (2016) 'Synthetic Genetic Array Analysis', *Cold Spring Harbor Protocols*, 2016(4), p. pdb.prot088807. doi: 10.1101/pdb.prot088807.

Kuzmin, E. *et al.* (2018) 'Systematic analysis of complex genetic interactions', *Science*. American Association for the Advancement of Science, 360(6386), p. eaao1729. doi: 10.1126/SCIENCE.AAO1729.

Kwok, R. (2010) 'Five hard truths for synthetic biology', *Nature*. Nature Publishing

Group, 463(7279), pp. 288–290. doi: 10.1038/463288a.
Lander, E. S. *et al.* (2001) 'Initial sequencing and analysis of the human genome', *Nature*. Nature Publishing Group, 409(6822), pp. 860–921. doi: 10.1038/35057062.
LeCun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', *Nature*. Nature Publishing Group, 521(7553), pp. 436–444. doi: 10.1038/nature14539.
Lee, M. E. *et al.* (2015) 'A Highly Characterized Yeast Toolkit for Modular, Multipart Assembly'. doi: 10.1021/sb500366v.
Leng, G. and Song, K. (2016) 'Watch out for your *TRP1* marker: the effect of *TRP1* gene on the growth at high and low temperatures in budding yeast', *FEMS Microbiology Letters*. Edited by D. Jamieson, 363(10), p. fnw093. doi: 10.1093/femsle/fnw093.
Li, M. Z. and Elledge, S. J. (2012) 'SLIC: A Method for Sequence- and Ligation-Independent Cloning', in *Methods in molecular biology (Clifton, N.J.)*, pp. 51–59. doi: 10.1007/978-1-61779-564-0_5.
Li, Xue *et al.* (2018) 'Rapid Detection of Respiratory Pathogens for Community-Acquired Pneumonia by Capillary Electrophoresis-Based Multiplex PCR':, *https://doi.org/10.1177/2472630318787452*. SAGE PublicationsSage CA: Los Angeles, CA. doi: 10.1177/2472630318787452.
Liang, J. *et al.* (2013) 'FairyTALE: A High-Throughput TAL Effector Synthesis Platform'. American Chemical Society. doi: 10.1021/SB400109P.
Libbrecht, M. W. and Noble, W. S. (2015) 'Machine learning applications in genetics and genomics', *Nature Reviews Genetics*. Nature Publishing Group, 16(6), pp. 321–332. doi: 10.1038/nrg3920.
Liebsch, A. (2014) 'Renewable hydrocarbons from sugarcane', *BMC Proceedings*. BioMed Central, 8(Suppl 4), p. O35. doi: 10.1186/1753-6561-8-S4-O35.
Lienert, F. *et al.* (2014) 'Synthetic biology in mammalian cells: next generation research tools and therapeutics', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 15(2), pp. 95–107. doi: 10.1038/nrm3738.
Linshiz, G. *et al.* (2013) 'PaR-PaR Laboratory Automation Platform', *ACS Synthetic Biology*, 2(5), pp. 216–222. doi: 10.1021/sb300075t.
Liu, W. and Stewart, C. N. (2015) 'Plant synthetic biology.', *Trends in plant science*. Elsevier, 20(5), pp. 309–317. doi: 10.1016/j.tplants.2015.02.004.

de los Santos, E. L. C. *et al.* (2016) 'Engineering Transcriptional Regulator Effector Specificity Using Computational Design and *In Vitro* Rapid Prototyping: Developing a Vanillin Sensor', *ACS Synthetic Biology*. American Chemical Society, 5(4), pp. 287–295. doi: 10.1021/acssynbio.5b00090.

Madsen, C. *et al.* (2016) 'The SBOL Stack: A Platform for Storing, Publishing, and Sharing Synthetic Biology Designs', *ACS Synthetic Biology*. American Chemical Society, 5(6), pp. 487–497. doi: 10.1021/acssynbio.5b00210.

Magi, A. *et al.* (2010) 'Bioinformatics for next generation sequencing data.', *Genes*. Multidisciplinary Digital Publishing Institute (MDPI), 1(2), pp. 294–307. doi: 10.3390/genes1020294.

Martella, A. *et al.* (2016) 'Mammalian Synthetic Biology: Time for Big MACs', *ACS Synthetic Biology*. American Chemical Society, 5(10), pp. 1040–1049. doi: 10.1021/acssynbio.6b00074.

Martella, A. *et al.* (2017) 'EMMA: An Extensible Mammalian Modular Assembly Toolkit for the Rapid Design and Production of Diverse Expression Vectors.', *ACS synthetic biology*, 6(7), pp. 1380–1392. doi: 10.1021/acssynbio.7b00016.

Matzas, M. *et al.* (2010) 'High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing.', *Nature biotechnology*, 28(12), pp. 1291–4. doi: 10.1038/nbt.1710.

McCulloch, E. S. (2013) 'Harnessing the Power of Big Data in Biological Research', *BioScience*. Oxford University Press, 63(9), pp. 715–716. doi: 10.1525/bio.2013.63.9.4.

Mendez-Perez, D. *et al.* (2017) 'Production of jet fuel precursor monoterpenoids from engineered *Escherichia coli*', *Biotechnology and Bioengineering*. Wiley-Blackwell, 114(8), pp. 1703–1712. doi: 10.1002/bit.26296.

Meng, H. *et al.* (2013) 'Quantitative Design of Regulatory Elements Based on High-Precision Strength Prediction Using Artificial Neural Network', *PLoS ONE*. Edited by S. Semsey, 8(4), p. e60288. doi: 10.1371/journal.pone.0060288.

Metzakopian, E. *et al.* (2017) 'Enhancing the genome editing toolbox: genome wide CRISPR arrayed libraries', *Scientific Reports*. Nature Publishing Group, 7(1), p. 2244. doi: 10.1038/s41598-017-01766-5.

Michelson, A. M. and Todd, A. R. (1955) 'Nucleotides part XXXII. Synthesis of a

dithymidine dinucleotide containing a 3′: 5′-internucleotidic linkage', *J. Chem. Soc.* The Royal Society of Chemistry, 0(0), pp. 2632–2638. doi: 10.1039/JR9550002632.
Mikkilä, J. *et al.* (2014) 'Virus-Encapsulated DNA Origami Nanostructures for Cellular Delivery'. American Chemical Society. doi: 10.1021/NL500677J.
Mitchell, L. A. *et al.* (2015) 'qPCRTag Analysis--A High Throughput, Real Time PCR Assay for Sc2.0 Genotyping.', *Journal of visualized experiments : JoVE*, (99), p. e52941. doi: 10.3791/52941.
Mitchell, L. A. *et al.* (2017) 'Synthesis, debugging, and effects of synthetic chromosome consolidation: synVI and beyond', *Science*, 355(6329), p. eaaf4831. doi: 10.1126/science.aaf4831.
Mitra, A. *et al.* (2015) 'Strategies for Achieving High Sequencing Accuracy for Low Diversity Samples and Avoiding Sample Bleeding Using Illumina Platform', *PLOS ONE*. Edited by C. Oudejans. Public Library of Science, 10(4), p. e0120520. doi: 10.1371/journal.pone.0120520.
Mojica, F. J., Juez, G. and Rodríguez-Valera, F. (1993) 'Transcription at different salinities of Haloferax mediterranei sequences adjacent to partially modified PstI sites.', *Molecular microbiology*, 9(3), pp. 613–21.
Mojica, F. J. M. *et al.* (2005) 'Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements', *Journal of Molecular Evolution*, 60(2), pp. 174–182. doi: 10.1007/s00239-004-0046-3.
Moore, G. E. (2006) 'Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114 ff.', *IEEE Solid-State Circuits Society Newsletter*, 11(3), pp. 33–35. doi: 10.1109/N-SSC.2006.4785860.
Moore, S. J. *et al.* (2016) 'EcoFlex: A Multifunctional MoClo Kit for *E. coli* Synthetic Biology', *ACS Synthetic Biology*. American Chemical Society, 5(10), pp. 1059–1069. doi: 10.1021/acssynbio.6b00031.
de Mora, K. *et al.* (2011) 'A pH-based biosensor for detection of arsenic in drinking water', *Analytical and Bioanalytical Chemistry*. Springer-Verlag, 400(4), pp. 1031–1039. doi: 10.1007/s00216-011-4815-8.
Mourtzis, D., Doukas, M. and Bernidaki, D. (2014) 'Simulation in Manufacturing: Review and Challenges', *Procedia CIRP*. Elsevier, 25, pp. 213–229. doi:

10.1016/J.PROCIR.2014.10.032.
Mühlmann, M. *et al.* (2017) 'Cellulolytic RoboLector – towards an automated high-throughput screening platform for recombinant cellulase expression', *Journal of Biological Engineering*. BioMed Central, 11(1), p. 1. doi: 10.1186/s13036-016-0043-2.
Nesbeth, D. N. *et al.* (2016) 'Synthetic biology routes to bio-artificial intelligence.', *Essays in biochemistry*. Portland Press Ltd, 60(4), pp. 381–391. doi: 10.1042/EBC20160014.
Nielsen, A. A. K. *et al.* (2016) 'Genetic circuit design automation.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 352(6281), p. aac7341. doi: 10.1126/science.aac7341.
Nielsen, S., Yuzenkova, Y. and Zenkin, N. (2013) 'Mechanism of Eukaryotic RNA Polymerase III Transcription Termination', *Science*, 340(6140), pp. 1577–1580. doi: 10.1126/science.1237934.
Nishikata, K. *et al.* (2014) 'Database Construction for PromoterCAD: Synthetic Promoter Design for Mammals and Plants', *ACS Synthetic Biology*, 3(3), pp. 192–196. doi: 10.1021/sb400178c.
Ó Gráda, C. (2016) 'Did Science Cause the Industrial Revolution?', *Journal of Economic Literature*, 54(1), pp. 224–239. doi: 10.1257/jel.54.1.224.
Oberortner, E. *et al.* (2016) 'Streamlining the Design-to-Build Transition with Build-Optimization Software Tools'. American Chemical Society. doi: 10.1021/ACSSYNBIO.6B00200.
Olsen, K. (2012) 'The First 110 Years of Laboratory Automation', *Journal of Laboratory Automation*, 17(6), pp. 469–480. doi: 10.1177/2211068212455631.
Palluk, S. *et al.* (2018) 'De novo DNA synthesis using polymerase-nucleotide conjugates', *Nature Biotechnology*. Nature Publishing Group, 36(7), pp. 645–650. doi: 10.1038/nbt.4173.
Pan, X. *et al.* (2007) 'dSLAM analysis of genome-wide genetic interactions in Saccharomyces cerevisiae.', *Methods (San Diego, Calif.)*. NIH Public Access, 41(2), pp. 206–21. doi: 10.1016/j.ymeth.2006.07.033.
Pardee, K. *et al.* (2014) 'Paper-Based Synthetic Gene Networks'. doi: 10.1016/j.cell.2014.10.004.

Patrick, W. G. *et al.* (2015) 'DNA Assembly in 3D Printed Fluidics', *PLOS ONE*. Edited by M. Wanunu. Public Library of Science, 10(12), p. e0143636. doi: 10.1371/journal.pone.0143636.
Patron, N. J. *et al.* (2015) 'Standards for plant synthetic biology: a common syntax for exchange of DNA parts', *New Phytologist*. Wiley/Blackwell (10.1111), 208(1), pp. 13–19. doi: 10.1111/nph.13532.
Pattanayak, V. *et al.* (2013) 'High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity', *Nature Biotechnology*. Nature Publishing Group, 31(9), pp. 839–843. doi: 10.1038/nbt.2673.
Paytubi, S. *et al.* (2017) 'A High-Throughput Screening Platform of Microbial Natural Products for the Discovery of Molecules with Antibiofilm Properties against Salmonella', *Frontiers in Microbiology*. Frontiers, 8, p. 326. doi: 10.3389/fmicb.2017.00326.
Pease, A. C. *et al.* (1994) 'Light-generated oligonucleotide arrays for rapid DNA sequence analysis.', *Proceedings of the National Academy of Sciences of the United States of America*, 91(11), pp. 5022–6.
Peris, D. *et al.* (2017) 'Hybridization and adaptive evolution of diverse Saccharomyces species for cellulosic biofuel production', *Biotechnology for Biofuels*. BioMed Central, 10(1), p. 78. doi: 10.1186/s13068-017-0763-7.
Petzold, C. J. *et al.* (2015) 'Analytics for Metabolic Engineering.', *Frontiers in bioengineering and biotechnology*. Frontiers Media SA, 3, p. 135. doi: 10.3389/fbioe.2015.00135.
Pfleger, B. F. *et al.* (2007) 'Microbial sensors for small molecules: Development of a mevalonate biosensor', *Metabolic Engineering*, 9(1), pp. 30–38. doi: 10.1016/j.ymben.2006.08.002.
Piccinini, F. *et al.* (2017) 'Advanced Cell Classifier: User-Friendly Machine-Learning-Based Software for Discovering Phenotypes in High-Content Imaging Data.', *Cell systems*. Elsevier, 4(6), pp. 651-655.e5. doi: 10.1016/j.cels.2017.05.012.
Plesa, C. *et al.* (2018) 'Multiplexed gene synthesis in emulsions for exploring protein functional landscapes.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 359(6373), pp. 343–347. doi: 10.1126/science.aao5167.
Purnick, P. E. M. and Weiss, R. (2009) 'The second wave of synthetic biology: from

modules to systems', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 10(6), pp. 410–422. doi: 10.1038/nrm2698.

Qi, L. S. *et al.* (2013) 'Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression.', *Cell*. NIH Public Access, 152(5), pp. 1173–83. doi: 10.1016/j.cell.2013.02.022.

Quan, J. *et al.* (2011) 'Parallel on-chip gene synthesis and application to optimization of protein expression', *Nature Biotechnology*, 29(5), pp. 449–452. doi: 10.1038/nbt.1847.

Quan, J. and Tian, J. (2009) 'Circular polymerase extension cloning of complex gene libraries and pathways.', *PloS one*. Public Library of Science, 4(7), p. e6441. doi: 10.1371/journal.pone.0006441.

Raina, M. and Ibba, M. (2014) 'tRNAs as regulators of biological processes.', *Frontiers in genetics*. Frontiers Media SA, 5, p. 171. doi: 10.3389/fgene.2014.00171.

Rajendran, A. *et al.* (2011) 'Programmed Two-Dimensional Self-Assembly of Multiple DNA Origami Jigsaw Pieces', *ACS Nano*, 5(1), pp. 665–671. doi: 10.1021/nn1031627.

Ran, F. A. *et al.* (2013) 'Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity', *Cell*, 154(6), pp. 1380–1389. doi: 10.1016/j.cell.2013.08.021.

Richardson, S. M. *et al.* (2017) 'Design of a synthetic yeast genome', *Science*, 355(6329), pp. 1040–1044. doi: 10.1126/science.aaf4557.

Rivas, E., Clements, J. and Eddy, S. R. (2017) 'A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs.', *Nature methods*. Howard Hughes Medical Institute, 14(1), pp. 45–48. doi: 10.1038/nmeth.4066.

Ro, D.-K. *et al.* (2006) 'Production of the antimalarial drug precursor artemisinic acid in engineered yeast', *Nature*. Nature Publishing Group, 440(7086), pp. 940–943. doi: 10.1038/nature04640.

Robinson, D. G. *et al.* (2014) 'Design and analysis of Bar-seq experiments.', *G3 (Bethesda, Md.)*. Genetics Society of America, 4(1), pp. 11–8. doi: 10.1534/g3.113.008565.

Roehner, N. *et al.* (2016) 'Double Dutch: A Tool for Designing Combinatorial Libraries of Biological Systems.', *ACS synthetic biology*, 5(6), pp. 507–17. doi:

10.1021/acssynbio.5b00232.
Rogers, J. K. and Church, G. M. (2016) ‘Genetically encoded sensors enable real-time observation of metabolite production’, *Proceedings of the National Academy of Sciences*, 113(9), pp. 2388–2393. doi: 10.1073/pnas.1600375113.
Rogers, J. K., Taylor, N. D. and Church, G. M. (2016) ‘Biosensor-based engineering of biosynthetic pathways’, *Current Opinion in Biotechnology*. Elsevier Current Trends, 42, pp. 84–91. doi: 10.1016/J.COPBIO.2016.03.005.
Roy, K. R. *et al.* (2018a) ‘Multiplexed precision genome editing with trackable genomic barcodes in yeast’, *Nature Biotechnology*. Nature Publishing Group. doi: 10.1038/nbt.4137.
Roy, K. R. *et al.* (2018b) ‘Multiplexed precision genome editing with trackable genomic barcodes in yeast’, *Nature Biotechnology*, 36(6), pp. 512–520. doi: 10.1038/nbt.4137.
Roy, S. and Caruthers, M. (2013) ‘Synthesis of DNA/RNA and Their Analogs via Phosphoramidite and H-Phosphonate Chemistries’, *Molecules*, 18(11), pp. 14268–14284. doi: 10.3390/molecules181114268.
Runguphan, W. and Keasling, J. D. (2014) ‘Metabolic engineering of Saccharomyces cerevisiae for production of fatty acid-derived biofuels and chemicals.’, *Metabolic engineering*, 21, pp. 103–13. doi: 10.1016/j.ymben.2013.07.003.
La Russa, M. F. and Qi, L. S. (2015) ‘The New State of the Art: Cas9 for Gene Activation and Repression.’, *Molecular and cellular biology*. American Society for Microbiology, 35(22), pp. 3800–9. doi: 10.1128/MCB.00512-15.
Rutkin, A. (2015) ‘Yeast’s heavenly potential’, *New Scientist*. Reed Business Information, 225(3011), pp. 8–9. doi: 10.1016/S0262-4079(15)60425-0.
Ryan, D. E. *et al.* (2018) ‘Improving CRISPR-Cas specificity with chemical modifications in single-guide RNAs.’, *Nucleic acids research*. Oxford University Press, 46(2), pp. 792–803. doi: 10.1093/nar/gkx1199.
Ryan, O. W. *et al.* (2014) ‘Selection of chromosomal DNA libraries using a multiplex CRISPR system’, *eLife*, 3. doi: 10.7554/eLife.03703.
Salis, H. M., Mirsky, E. A. and Voigt, C. A. (2009) ‘Automated design of synthetic ribosome binding sites to control protein expression.’, *Nature biotechnology*, 27(10),

pp. 946–50. doi: 10.1038/nbt.1568.
Sanchez, C. *et al.* (1999) *Grasping at molecular interactions and genetic networks in Drosophila melanogaster using FlyNets, an Internet database*, *Nucleic Acids Research*.
Sarrion-Perdigones, A. *et al.* (2013) 'GoldenBraid 2.0: A Comprehensive DNA Assembly Framework for Plant Synthetic Biology', *PLANT PHYSIOLOGY*, 162(3), pp. 1618–1631. doi: 10.1104/pp.113.217661.
Sauer, B. (1994) 'Recycling selectable markers in yeast.', *BioTechniques*, 16(6), pp. 1086–8.
Sedor, K. (no date) *The Law of Large Numbers and its Applications*.
Seeber, A. and Gasser, S. M. (2017) 'Chromatin organization and dynamics in double-strand break repair', *Current Opinion in Genetics & Development*. Elsevier Current Trends, 43, pp. 9–16. doi: 10.1016/J.GDE.2016.10.005.
Sequeira, A. F. *et al.* (2016) 'T7 Endonuclease I Mediates Error Correction in Artificial Gene Synthesis', *Molecular Biotechnology*, 58(8–9), pp. 573–584. doi: 10.1007/s12033-016-9957-7.
Sezonov, G., Joseleau-Petit, D. and D'Ari, R. (2007) 'Escherichia coli physiology in Luria-Bertani broth.', *Journal of bacteriology*. American Society for Microbiology, 189(23), pp. 8746–9. doi: 10.1128/JB.01368-07.
Shapland, E. B. *et al.* (2015) 'Low-Cost, High-Throughput Sequencing of DNA Assemblies Using a Highly Multiplexed Nextera Process', *ACS Synthetic Biology*, 4(7), pp. 860–866. doi: 10.1021/sb500362n.
Shen, Y. *et al.* (2017) 'Deep functional analysis of synII, a 770-kilobase synthetic yeast chromosome.', *Science (New York, N.Y.)*. Europe PMC Funders, 355(6329). doi: 10.1126/science.aaf4791.
Shendure, J. *et al.* (2017) 'DNA sequencing at 40: past, present and future', *Nature 2017 550:7676*. Nature Publishing Group, 550(7676), p. 345. doi: 10.1038/nature24286.
Shetty, R. *et al.* (2011) 'Assembly of BioBrick Standard Biological Parts Using Three Antibiotic Assembly', in *Methods in enzymology*, pp. 311–326. doi: 10.1016/B978-0-12-385120-8.00013-9.
Shetty, R. P., Endy, D. and Knight, T. F. (2008) 'Engineering BioBrick vectors from

BioBrick parts.', *Journal of biological engineering*, 2, p. 5. doi: 10.1186/1754-1611-2-5.
Shi, S. *et al.* (2017) 'Discovery and engineering of a 1-butanol biosensor in Saccharomyces cerevisiae h i g h l i g h t s'. doi: 10.1016/j.biortech.2017.06.114.
Shimizu, H. *et al.* (1997) 'Short-homology-independent illegitimate recombination in Escherichia coli: distinct mechanism from short-homology-dependent illegitimate recombination.', *Journal of molecular biology*, 266(2), pp. 297–305.
Shipman, S. L. *et al.* (2017) 'CRISPR–Cas encoding of a digital movie into the genomes of a population of living bacteria', *Nature*. Nature Publishing Group, 547(7663), pp. 345–349. doi: 10.1038/nature23017.
Si, T. *et al.* (2015) 'RNAi-Assisted Genome Evolution in *Saccharomyces cerevisiae* for Complex Phenotype Engineering', *ACS Synthetic Biology*. American Chemical Society, 4(3), pp. 283–291. doi: 10.1021/sb500074a.
Si, T. *et al.* (2017) 'Automated multiplex genome-scale engineering in yeast', *Nature Communications*. Nature Publishing Group, 8, p. 15187. doi: 10.1038/ncomms15187.
Sinclair, I. *et al.* (2016) 'Novel Acoustic Loading of a Mass Spectrometer', *Journal of Laboratory Automation*, 21(1), pp. 19–26. doi: 10.1177/2211068215619124.
Singh, N. *et al.* (2018) 'Bioethanol production by a xylan fermenting thermophilic isolate Clostridium strain DBT-IOC-DC21', *Anaerobe*. Academic Press, 51, pp. 89–98. doi: 10.1016/J.ANAEROBE.2018.04.014.
Smanski, M. J. *et al.* (2014) 'Functional optimization of gene clusters by combinatorial design and assembly', *Nature Biotechnology*, 32(12), pp. 1241–1249. doi: 10.1038/nbt.3063.
Smanski, M. J. *et al.* (2016) 'Synthetic biology to access and expand nature's chemical diversity', *Nature Reviews Microbiology*. Nature Publishing Group, 14(3), pp. 135–149. doi: 10.1038/nrmicro.2015.24.
Smith, J. D. *et al.* (2016) 'Quantitative CRISPR interference screens in yeast identify chemical-genetic interactions and new rules for guide RNA design', *Genome Biology*. BioMed Central, 17(1), p. 45. doi: 10.1186/s13059-016-0900-9.
Stein, L. D. (2010) 'The case for cloud computing in genome informatics', *Genome Biology*. BioMed Central, 11(5), p. 207. doi: 10.1186/gb-2010-11-5-207.

Stevens, T. M. (1875) ‘Rapid and Automatic Filtration’, *Am. Chemist*, 6(3), p. 102.
Stewart, D. and Wilson-Kanamori, J. R. (2011) ‘Modular Modelling in Synthetic Biology: Light-Based Communication in E. coli’, *Electronic Notes in Theoretical Computer Science*. Elsevier, 277, pp. 77–87. doi: 10.1016/J.ENTCS.2011.09.037.
Stovicek, V., Holkenbrink, C. and Borodina, I. (2017) ‘CRISPR/Cas system for yeast genome engineering: advances and applications.’, *FEMS yeast research*. Oxford University Press, 17(5). doi: 10.1093/femsyr/fox030.
Stricker, J. *et al.* (2008) ‘A fast, robust and tunable synthetic gene oscillator’, *Nature*. Nature Publishing Group, 456(7221), pp. 516–519. doi: 10.1038/nature07389.
Syarif, I., Prugel-Bennett, A. and Wills, G. (2016) ‘SVM Parameter Optimization Using Grid Search and Genetic Algorithm to Improve Classification Performance’, *TELKOMNIKA*, 14(4), pp. 1502–1509.
Tao, Q. *et al.* (2017) ‘Enhanced biomass/biofuel production and nutrient removal in an algal biofilm airlift photobioreactor’, *Algal Research*. Elsevier, 21, pp. 9–15. doi: 10.1016/J.ALGAL.2016.11.004.
Taylor, B. Hugh, S. (1922) ‘Automatic Volumetric Analysis Carbon Monoxide Recorder’, *J. Ind. Eng. Chem.*, 14(11), p. 1008.
TerMaat, J. R. *et al.* (2009) ‘Gene synthesis by integrated polymerase chain assembly and PCR amplification using a high-speed thermocycler.’, *Journal of microbiological methods*. NIH Public Access, 79(3), pp. 295–300. doi: 10.1016/j.mimet.2009.09.015.
Trotta, C. R. *et al.* (1997) ‘The yeast tRNA splicing endonuclease: a tetrameric enzyme with two active site subunits homologous to the archaeal tRNA endonucleases.’, *Cell*, 89(6), pp. 849–58.
Tsai, S. Q. *et al.* (2014) ‘GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases’, *Nature Biotechnology 2014 33:2*. Nature Publishing Group, 33(2), p. 187. doi: 10.1038/nbt.3117.
Turowski, T. W. and Tollervey, D. (2016) ‘Transcription by RNA polymerase III: insights into mechanism and regulation.’, *Biochemical Society transactions*. Portland Press Ltd, 44(5), pp. 1367–1375. doi: 10.1042/BST20160062.
Uesono, Y., Toh-e, A. and Kikuchi, Y. (1997) ‘Ssd1p of Saccharomyces cerevisiae Associates with RNA *’, 272(26), pp. 16103–16109.

Urnov, F. D. *et al.* (2010) 'Genome editing with engineered zinc finger nucleases', *Nature Reviews Genetics*. Nature Publishing Group, 11(9), pp. 636–646. doi: 10.1038/nrg2842.

Vasilev, V. *et al.* (2011) *A Software Stack for Specification and Robotic Execution of Protocols for Synthetic Biological Engineering*.

Vazquez-Vilar, M. *et al.* (2017) 'GB3.0: a platform for plant bio-design that connects functional DNA elements with associated biological data', *Nucleic Acids Research*, 45(4), p. gkw1326. doi: 10.1093/nar/gkw1326.

Veneziano, R. *et al.* (2018) 'In vitro synthesis of gene-length single-stranded DNA', *Scientific Reports*. Nature Publishing Group, 8(1), p. 6548. doi: 10.1038/s41598-018-24677-5.

Villalobos, A. *et al.* (2006) 'Gene Designer: a synthetic biology tool for constructing artificial DNA segments.', *BMC bioinformatics*. BioMed Central, 7, p. 285. doi: 10.1186/1471-2105-7-285.

Wang, H. *et al.* (2014) 'Improved seamless mutagenesis by recombineering using ccdB for counterselection.', *Nucleic acids research*. Oxford University Press, 42(5), p. e37. doi: 10.1093/nar/gkt1339.

Wang, L. *et al.* (2018) 'Synthetic Genomics: From DNA Synthesis to Genome Design', *Angewandte Chemie International Edition*, 57(7), pp. 1748–1756. doi: 10.1002/anie.201708741.

Wang, P. *et al.* (2017) 'The Beauty and Utility of DNA Origami', *Chem*. Cell Press, 2(3), pp. 359–382. doi: 10.1016/J.CHEMPR.2017.02.009.

Wang, T. *et al.* (2018) 'Pooled CRISPR interference screening enables genome-scale functional genomics study in bacteria with superior performance', *Nature Communications*. Nature Publishing Group, 9(1), p. 2475. doi: 10.1038/s41467-018-04899-x.

Wang, Z.-W. and Chen, S. (2009) 'Potential of biofilm-based biofuel production', *Applied Microbiology and Biotechnology*. Springer-Verlag, 83(1), pp. 1–18. doi: 10.1007/s00253-009-1940-9.

Watanabe, L. *et al.* (2018) 'i B io S im 3: A Tool for Model-Based Genetic Circuit Design', *ACS Synthetic Biology*. American Chemical Society, p. acssynbio.8b00078. doi: 10.1021/acssynbio.8b00078.

Watkins, N. E. *et al.* (2005) 'Nearest-neighbor thermodynamics of deoxyinosine pairs in DNA duplexes.', *Nucleic acids research*. Oxford University Press, 33(19), pp. 6258–67. doi: 10.1093/nar/gki918.

Watkins, N. J. and Bohnsack, M. T. (2012) 'The box C/D and H/ACA snoRNPs: key players in the modification, processing and the dynamic folding of ribosomal RNA', *Wiley Interdisciplinary Reviews: RNA*, 3(3), pp. 397–414. doi: 10.1002/wrna.117.

Weber, E. *et al.* (2011) 'A Modular Cloning System for Standardized Assembly of Multigene Constructs', *PLoS ONE*. Edited by J. Peccoud. Public Library of Science, 6(2), p. e16765. doi: 10.1371/journal.pone.0016765.

Werner, S. *et al.* (2012) 'Fast track assembly of multigene constructs using Golden Gate cloning and the MoClo system', *Bioengineered*, 3(1), pp. 38–43. doi: 10.4161/bbug.3.1.18223.

Wertman, K. F., Drubin, D. G. and Botstein, D. (1992) 'Systematic mutational analysis of the yeast ACT1 gene.', *Genetics*, 132(2), pp. 337–50.

Wickham, S. F. J. *et al.* (2012) 'A DNA-based molecular motor that can navigate a network of tracks', *Nature Nanotechnology*. Nature Publishing Group, 7(3), pp. 169–173. doi: 10.1038/nnano.2011.253.

Wilson-Kanamori, J. *et al.* (2015) 'Kappa rule-based modeling in synthetic biology.', *Methods in molecular biology (Clifton, N.J.)*, 1244, pp. 105–35. doi: 10.1007/978-1-4939-1878-2_6.

Wong, A. S. L. *et al.* (2015) 'Massively parallel high-order combinatorial genetics in human cells.', *Nature biotechnology*. NIH Public Access, 33(9), pp. 952–61. doi: 10.1038/nbt.3326.

Wong, A. S. L., Choi, G. C. G. and Lu, T. K. (2016) 'Deciphering Combinatorial Genetics', *Annual Review of Genetics*. Annual Reviews , 50(1), pp. 515–538. doi: 10.1146/annurev-genet-120215-034902.

Wu, H.-C. *et al.* (2016) 'A Demonstration of High-Throughput Electrophoresis':, *https://doi.org/10.1016/S1535-5535-04-00238-2*. SAGE PublicationsSage CA: Los Angeles, CA. doi: 10.1016/S1535-5535-04-00238-2.

Wu, H., Yang, L. and Chen, L.-L. (2017) 'The Diversity of Long Noncoding RNAs and Their Generation.', *Trends in genetics : TIG*. Elsevier, 33(8), pp. 540–552. doi: 10.1016/j.tig.2017.05.004.

Wu, Y. *et al.* (2017) ‘Bug mapping and fitness testing of chemically synthesized chromosome X’, *Science*, 355(6329), p. eaaf4706. doi: 10.1126/science.aaf4706.
Xia, B. *et al.* (2011) ‘Developer’s and User’s Guide to Clotho v2.0’, in *Methods in enzymology*, pp. 97–135. doi: 10.1016/B978-0-12-385120-8.00005-X.
Xie, Z.-X. *et al.* (2017) ‘“Perfect” designer chromosome V and behavior of a ring derivative’, *Science*, 355(6329), p. eaaf4704. doi: 10.1126/science.aaf4704.
Young, B. P. and Loewen, C. J. (2013) ‘Balony: a software package for analysis of data generated by synthetic genetic array experiments’, *BMC Bioinformatics*. BioMed Central, 14(1), p. 354. doi: 10.1186/1471-2105-14-354.
Zhang, W. *et al.* (2017) ‘Engineering the ribosomal DNA in a megabase synthetic chromosome’, *Science*, 355(6329), p. eaaf3981. doi: 10.1126/science.aaf3981.
Zhang, X.-H. *et al.* (2015) ‘Off-target Effects in CRISPR/Cas9-mediated Genome Engineering’, *Molecular Therapy - Nucleic Acids*. Cell Press, 4. doi: 10.1038/MTNA.2015.37.
Zhang, Y., Werling, U. and Edelmann, W. (2014) ‘Seamless Ligation Cloning Extract (SLiCE) cloning method.’, *Methods in molecular biology (Clifton, N.J.)*. NIH Public Access, 1116, pp. 235–44. doi: 10.1007/978-1-62703-764-8_16.
Zhang, Z. and Ren, Q. (2015) ‘Why are essential genes essential? - The essentiality of Saccharomyces genes.’, *Microbial cell (Graz, Austria)*. Shared Science Publishers, 2(8), pp. 280–287. doi: 10.15698/mic2015.08.218.
Zheng, Y. *et al.* (2018) ‘CRISPR interference-based specific and efficient gene inactivation in the brain’, *Nature Neuroscience*. Nature Publishing Group, 21(3), pp. 447–454. doi: 10.1038/s41593-018-0077-5.
Zimin, A. *et al.* (2014) ‘Sequencing and assembly of the 22-gb loblolly pine genome.’, *Genetics*. Genetics Society of America, 196(3), pp. 875–90. doi: 10.1534/genetics.113.159715.