Design, synthesis and characterization of the synthetic yeast genome

Yue Shen

Thesis submitted for the Degree of Doctor of Philosophy

School of Biological Sciences, The University of Edinburgh, UK
March 2018

Declaration

I declare that this thesis was composed by myself and all the works presented in the thesis is my own except where explicitly stated otherwise. This thesis has not been submitted for any other degree or professional qualification.

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly-authored publications has been included. The contributions to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

This thesis contains work of the following publications:

- Kanigowska P, **Shen Y**, Zheng Y, Rosser S, Cai Y. Smart DNA Fabrication Using Sound Waves: Applying Acoustic Dispensing Technologies to Synthetic Biology. J Lab Autom. July 2015:2211068215593754. doi:10.1177/2211068215593754. (Co-first author)
- **Shen Y**, Stracquadanio G, Wang Y, et al. SCRaMbLE generates designed combinatorial stochastic diversity in synthetic chromosomes. Genome Res. 2016;26(1):36-49. doi:10.1101/gr.193433.115.
- **Shen Y**, Wang Y, Chen T, et al. Deep functional analysis of synII, a 770-kilobase synthetic yeast chromosome. Science. 2017;355(6329). doi:10.1126/science.aaf4791.

The major part of research and the major part of writing of all three publications has been conducted by the candidate. The four co-authors (Paulina Kanigowska, Giovanni Stracquadanio, Yun Wang and Tai Chen) contributed by guiding the research and the publication writing process, as well as giving feedback and corrections for the publications.

Chapter Two Part 2.2 contains content from co-first author paper: Kanigowska P, **Shen Y**, Zheng Y, Rosser S, Cai Y. Smart DNA Fabrication Using Sound Waves: Applying Acoustic Dispensing Technologies to Synthetic Biology. J Lab Autom. July 2015:2211068215593754.
Chapter Two Part 2.3 contains content from first author paper: **Shen Y**, Wang Y, Chen T, et al. Deep functional analysis of synII, a 770-kilobase synthetic yeast chromosome. Science. 2017;355(6329).
Chapter Three Part 3.2 contains content from first author paper: **Shen Y**, Wang Y, Chen T, et al. Deep functional analysis of synII, a 770-kilobase synthetic yeast chromosome. Science. 2017;355(6329).

Chapter Four Part 4.2 contains content from first author paper: **Shen Y**, Stracquadanio G, Wang Y, et al. SCRaMbLE generates designed combinatorial stochastic diversity in synthetic chromosomes. Genome Res. 2016;26(1):36-49.

Chapter Five contains content from co-author paper: Wang Y, **Shen Y**, et al. Genome Writing: Current Progress and Related Applications. Genomics Proteomics Bioinformatics. 2018;16(1).

.

yue shen

Yue Shen

2018-08-17

Acknowledgements

This work would not have been possible without the constant help and support I received over the past few years. This journey has blessed me with the most wonderful experiences in my life, and I have had the pleasure of knowing many wonderful mentors, colleagues and friends who helped me and challenged me to grow and mature professionally and personally.

I must first thank my supervisors Prof. Patrick Cai and Prof. Christopher French for their outstanding supervision throughout my PhD training. Patrick offered me the opportunity to start my PhD study in Edinburgh. He has helped me in almost all ways that I can think about. In Cai Lab, I'm really lucky to be involved in so many different projects, which really broaden my knowledge and experiences. Patrick also encouraged me to attend seminars or conferences and helped me to improve my presentation skills. Even in the end of my fourth year, he has never stopped giving me suggestions on my future career. I would like to thank Chris for his insightful advice on my projects, as well as his patient instructions and comments throughout my study. Without them, the papers would not be published, nor would the thesis be completed in time.

I would also like to thank my PhD committee member, Prof. Adele Marston and Prof. Andrew Miller for their support on my PhD projects and professional suggestions over my PhD study. Without the help from Adele and her lab members (Bonnie Alver and Adrien Paul-Dubois-Taine), I wouldn't be able to finish the segregation experiment within a very short time-frame. I want to thank Andrew for the rich support that he has provided through the many discussions and encouragements that have enabled me to pursue this body of work and for teaching me to think more critically.

I owe many thanks to my mentor Prof. Huanming Yang, my leaders Dr. Xun Xu, Dr. Ye Yin and colleagues from BGI. The past four years have been very challenging for me, and I would not be able tackle all the challenges without their constant help and support.

I have to give many thanks as well to Prof. Jef Boeke and collaborators from the Sc2.0 consortium. Jef led me into the world of yeast and it is him who offered me the opportunity to have such great experiences in science. I thank Prof. Junbiao Dai, Prof. Yingjin Yuan, Prof. Joel Bader, Dr. Leslie Mitchell and others for their tremendous help and sharing over the years.

During my time in the Cai lab, I have been fortunate to work with many smart, supportive and entertaining friends and colleagues. I enjoyed working with Sarah Zheng and Paulina Kanigowska on automation related projects, which might greatly contribute to the obsessive-compulsive disorder symptoms to all three of us. I want to give the special thanks to Dr. Dariusz Abramczyk for the countless hours of help both on my work and being a close friend in my life. I still remember him helping me counting the cells for a whole day till the midnight during the dark depressing winter in Edinburgh. Sincere thanks also go to Yisha Luo, Wei Liu, Roy Walker and other Cai Lab members for making my time in Edinburgh an exciting experience. I would also like to thank the "Manchester gang" Jamie, Eva, Daniel and Sally for their great help to me in settling down in Manchester for the final half year. I enjoyed every second with them!

Finally, I am dedicating this work to my family. I feel blessed to have the constant love and supports from my Dad (Yonglin Shen) and Mum (Zhongli Ma), who have always put their child before themselves. It is them who taught me to never give up and persist till succeed. I'm particularly thankful to my husband Shida for being the most caring and inspiring person to me. I would never have made it this far without his encouragement, companionship and love. Flying all the way here from China to Manchester for keeping me company during the writing of my thesis is extraordinarily warm. For these things and so much more, I am eternally grateful.

Contents

Abstract

With the rapid development of DNA synthesis technologies, synthetic biology has made tremendous progress in the past 15 years, in particular for synthetic genomics. Synthetic genomics is a nascent field of synthetic biology, which aims to design new biological systems/organisms to satisfy human needs. Conventional synthetic biology focuses on the redesign, construction and modeling of biological parts, pathways or genomes that do not exist in nature, while synthetic genomics encompasses technologies that allow the generation of chemically synthesized larger parts of genomes or whole genomes, with simultaneous redesign of an organism's genetic material. Synthetic genomics is painting a blueprint for a new era of biology and holds great potential for a multitude of applications, such as pharmaceuticals, biofuels and rapid generation of vaccines against emerging diseases.

Chapter One gives an introduction of the current state of the art and challenges of synthetic genomics and the objectives of this study. Chapter Two demonstrates the design and construction strategy of two megabase-long synthetic yeast chromosomes, *SynII* and *SynVII*. Chapter Three describes the full characterization of *SynII* and *SynVII*. Chapter Four introduces the SCRaMbLE (Synthetic Chromosome Rearrangement and Modification by LoxPsym-mediated Evolution) system and its application in *SynII* and *SynVII*. Taken together, this work demonstrates the utility of synthetic yeast for understanding biological systems and its potential for industrial applications.

Lay Summary

"What I cannot create I do not understand" – Richard Feynman

Over the past centuries a great deal of effort has been dedicated to understand the innerworkings of life, leading to our current understanding on biological processes underpinning many life forms on earth. Despite these advances, it remains unresolved whether we can rebuild or create biological systems in a laboratory setting, the only way to demonstrate we fully grasp the gist of how life works. The question remains largely unanswered until two decades ago when synthetic biology was developed: by combining principles from biology, computation, engineering and physics, researchers were able to create biological systems from scratch. And the complexity of built systems went from prokaryotes to eukaryotes, from a few genes to a whole genome. As a nascent field of synthetic biology, synthetic genomics uses aspects of artificial gene synthesis to create entire lifeforms, with simultaneous redesign of an organism's genetic material.

In the past 15 years, thanks to the rapid development of DNA synthesis technologies, synthetic genomics has made tremendous progress, showing great potentials in various applications, such as pharmaceuticals, biofuels and rapid generation of vaccines against emerging diseases.

In this study, 1) new methods were developed to facilitate high efficient construction and characterization of synthetic chromosomes in yeast. 2) Two megabase-long yeast chromosomes, *SynII* and *SynVII*, were designed, constructed and fully characterized. And 3) an inducible genome rearrangement system was implemented in the synthetic yeast, which facilitates our understanding of fundamental biological systems and holds potential for industrial applications.

List of Abbreviations

AHLs	*N*-acylhomoserine lactones
AQs	2-alkyl-4-quinolones
ADE	Acoustic droplet ejection
BWA	Burrows-Wheeler Aligner
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
Cas	CRISPR-associated
CDK	cyclin-dependent kinase
CDS	Coding sequence
CREST	Clipping Reveals Structure
DSB	Double-strand break
E. coli	*Escherichia coli*
EDTA	Ethylenediaminetetraacetic acid
EBD	Estrogen binding domain
FDR	False discovery rate
FWER	Family Wise Error Rate
GO	Gene Ontology
GRAS	Generally Regarded As Safe
GFP	Green fluorescent protein
Gb	Gigabase
GWAS	Genome-side association studies
HSR	Heat shock response
HSF	Heat shock transcription factor
HHQ	2-heptyl-4(*1H*)-quinolones
H. polymorpha	*Hansenula polymorpha*
HPV	Human papillomavirus
HCl	Hydrochloric acid
HOG	High-osmolarity glycerol
HDR	Homology directed repair

IP	Isopentenyladenine
IAM	Indole-3-acetamide
K. lactis	*Kluyveromyces lactis*
KIV	2-ketoisovalerate
Kbp	Kilobase pair
LTR	Long terminal repeats
mRNA	massager RNA
M. mycoides	*Mycoplasma mycoides*
M. genitalium	*Mycoplasma genitalium*
MISO	Multichange ISOthermal
Mbp	Mega base pair
MMS	Methyl methanesulfonate
NaOH	Sodium hydroxide
ORF	Open reading frame
P. pastoris	*Pichia pastoris*
ROS	Reactive oxygen species
rtTA	responsive trans-activator
RQ	Respiratory quotient
RFP	Red fluorescent protein
S.cerevisiae	*Saccharomyces cerevisiae*
S.pombe	*Schizosaccharomyces pombe*
sgRNA	single-guide RNA
scRNA	scaffold RNA
SCRaMbLE	Synthetic Chromosome Rearrangement and Modification by LoxP-mediated Evolution
SwAP-In	Switching Auxotrophies Progressively for Integration
SynII	Synthetic chromosome II
SynVII	Synthetic chromosome VII

SNV	Single nucleotide variation
SV	Structural variation
SGD	Saccharomyces Genome Database
SC	Synthetic complete
SGA	Synthetic genetic array
SynIXR	Synthetic right arm of chromosome IX
tRNA	transfer RNA
TAE	Tris-acetate-EDTA
UTR	Untranslated region
uHTS	ultra-high-throughput screening
YEP	Yellow fluorescent protein
YPD	Yeast extract peptone dextrose
YPEG	Yeast extract peptone glycerol ethanol
5’-FOA	5’-Fluoroorotic Acid

List of Figures

List of Tables

Chapter One

Introduction

1.1 Yeast synthetic biology

1.1.1 The state-of-art in synthetic biology and synthetic genomics

One widely accepted definition refers to synthetic biology as the redesign of naturally existing biological systems, and *de novo* design and construction of new biological systems. The history of synthetic biology dates back to 1928 when Wöhler completed the synthesis of urea[1]. Since then and with the successful unraveling of the genetic code, synthetic biology was mainly focused on oligonucleotides or gene synthesis. One of the monumental accomplishments is the in vitro synthesis of a 207bp tyrosine suppressor tRNA gene by Khorana and coworkers in 1979[2].

Development of tools for efficient genetic refactoring of biological parts, pathways and host chassis is imperative for successful re-engineering of existing biological systems. For instance, libraries of well-characterized promoter and terminator elements are helping synthetic biologists to optimize fluxes of the heterologously expressed pathways[3], while genome-scale models help to properly incorporate synthetic gene networks within the intrinsic cellular metabolic pathways[4]. The development of DNA assembly technologies facilitates synthetic DNA construction. Methods such as enzymatic Gibson assembly[5], Golden Gate assembly[6] and MoClo assembly[7] and the homologous recombination-based YOGE[8] and VEGAS[9] protocols allow for *in vivo* construction of genetic circuits and pathways. TAR cloning enables rapid capture of genetic circuits and pathways directly from more complex genomes through recombination machinery[10], thus circumventing the need to de novo synthesize and amplify large DNA pieces. Finally, targeted DNA editing tools, such as the CRISPR-Cas system, make it possible to introduce multiplex modifications to the host genome[11].

The remarkably fast-developing DNA synthesis, assembly and manipulation techniques have expanded synthetic biology studies from the simplest prokaryotic model systems to more complex eukaryotic model systems and helped bring forth synthetic genomics, a nascent field of synthetic biology. Synthetic genomics aims to design new biological systems/organisms to satisfy human needs. Conventional synthetic biology focuses on the redesign, construction and modeling of biological parts, pathways or genomes typically at a smaller scale, while synthetic genomics encompasses technologies that allow the generation of chemically synthesized larger parts of genomes or whole genomes, with simultaneous redesign of an organism's genetic materials. Over the past decades progressive advancements have been made in synthetic genomics, which include the 7.5kb synthetic poliovirus in 2002[12], the 5.4kb synthetic φX174 phage in 2003[13], the 40kb refactored T7 phage in 2005[14] and the 1.1M "synthia" *Mycoplasma genitalium* genome in 2010[15]. More recently synthetic genomics has taken another giant step forward and celebrated the success of de novo design and assembly of first eukaryote chromosomes from yeast[16-22]. Building upon these success, synthetic biologists even start to conceptualize writing more complex multicellular chromosomes and even the human genome[23].

The combination of computational design and de novo DNA synthesis enables the construction of re-designed genetic material, which is inaccessible to conventional biotechnological approaches. Synthetic genomics is painting a blueprint for a new era of biology and holds great potential for a multitude of applications, such as pharmaceuticals, biofuels and rapid generation of vaccines against emerging diseases.

1.1.2 The state-of-art in yeast synthetic biology

Among all the model organisms that are amenable to de novo synthesis from scratch, yeast stands out as one of the most powerful eukaryotic model organisms for several reasons. First, yeast is a simple unicellular organism whose molecular biology and genetics can often be readily related to higher eukaryotes, including humans. Second, most of the model yeast species possess the "Generally Regarded As Safe" (GRAS) status, and exhibit robust growth on inexpensive carbon sources. In addition, an increasing number of genetic engineering tools are readily available for these single-celled fungi. Finally, yeast genomes are much more compact than their mammalian counterparts and thus yeast cells are more convenient for studying complex cellular processes, excepting developmental studies.

Among different species of yeasts, baker's yeast *S. cerevisiae* is undoubtedly the most well-established eukaryotic model organism, whose remarkable ability for efficient homologous recombination has made it one of the main synthetic biology chassis. But *S. cerevisiae* is not the only model yeast to answer the fundamental cell biology questions. Fission yeast *Schizosaccharomyces pombe* (*S. pombe*), for example, has been extensively studied to investigate cell cycle control mechanisms[24,25]. In the following parts, a few examples from different angles are shown to elaborate on the current status of applying synthetic biology approach to yeast models.

Synthetic yeasts reveal natural cellular functions

In the past few years synthetic biology has greatly enhanced our understanding of natural biological processes mainly by constructing in-vitro/in-vivo pathways and networks to interrogate specific biological functions[26,27]. Direct testing custom-built networks or systems in living cells enables the researches to obtain readouts of synthetic biological elements in the cellular contexts, by which also enriches our understanding of natural systems. To date, synthetic biology studies in bacterial system have made tremendous advances largely due to its well-defined cellular components and simplicity[28]. More recently synthetic biologists have also successfully demonstrated the feasibility and utility of constructing synthetic regulatory networks to recapitulate various cellular functions in the more complex eukaryotic system, the yeast.

Synthetic gene expression programs

Modulation of gene expression levels is essential in metabolic engineering for optimized production[29]. Eukaryotic cells possess a sophisticated program evolved naturally for precise control over gene activation/suppression in various cellular processes and/or in response to external stimuli. Design and implementation of novel gene regulatory programs will not only provide a new level of understanding for many natural biological processes but also facilitate engineered designer cells for biotechnological applications. In general, gene expression can be regulated at the levels of transcription, post-transcriptional modification (such as RNA splicing and histone mRNA processing) and translation. The following brief survey on gene expression regulation is accompanied by a discussion of synthetic approaches via transcriptional and translational engineering (Figure 1.1). The *cis* regulatory elements including promoters, 5’ or 3’ UTRs and introns are involved in the regulation of gene expression. Many studies have generated synthetic promoter libraries to fine-tune gene expression. In eukaryotic cells, RNA splicing also plays a key role in controlling the levels of gene expression. Yofe et al. constructed a *S. cerevisiae* intron reporter library and developed a model based on this library to study the rules of intron-mediated regulation of gene expression[30]. In their study, hundreds of native yeast introns were individually cloned in front of a yellow fluorescent protein (YFP) cassette to access their effects on the regulation of YFP expression. In agreement with previous studies, the authors found that intron-exon junction structure (RNA secondary structure), transcript GC content and splicing regulatory elements have impacts on gene expression levels[31,32]. Intriguingly, they observed no correlation between intron length and gene expression levels.

Apart from engineering transcription regulatory elements, locus-specific regulation can also be achieved by repurposing CRISPR single-guide RNA (sgRNA) as scaffolding molecules with protein recruitment capabilities[33]. Zalatan et al. designed a scaffold RNA (scRNA) by individually connecting RNA hairpin domain to the CRISPR sgRNA at the 3’ end with a linker sequence. Three well-characterized viral RNA sequences MS2, PP7 and com were selected as recruitment RNA modules, while their corresponding RNA binding proteins MCP, PCP and Com were also individually engineered by fusing with transcriptional activation domain VP64. The transcriptional activation capability of this system was tested in a dCas9 expressing yeast strain with a fluorescent protein reporter driven by a tet-promoter. The results showed the scRNA-protein recruitment modules could strongly activate reporter gene expression in an orthogonal manner. In addition, they also found that by manipulating the numbers of recruitment motif and combining a mixture of recruitment RNA modules, their CRISPR scRNA platform can achieve site-specific tunable multiplexing gene activation.

Riboswitches are RNA-only molecules, which constitute an attractive tool for translational regulation[34]. There are two functional domains in a riboswitch - an ligand-binding aptamer domain and an expression platform that determines system output[35]. By engineering aptamer ligand complex, programmable transcriptional and translational regulation can be achieved[36,37]. Grate et al. engineered a synthetic cyclin transcript *CLB2* in budding yeast by inserting a malachite green binding aptamer into *CLB2* 5'-UTR to regulate expression of CLB2 protein, and thereby controlling the yeast cell cycle[38]. RT-PCR and Western blot analysis revealed that aptamer insertion limited the binding of 40S ribosomal submit to the *CLB2* 5'-UTR, which led to a reduced rate of translational initiation. These results indicate that aptamer-mediated mRNA translation efficiency was affected rather than mRNA stability[36]. In another study, Hanson et al. further investigated the correlation between translation efficiency and insertion positions within the 5'-UTR using a tetracycline-aptamer-mediated translational regulation system[37]. In their study, the tetracycline-binding aptamer was inserted into the 5'-UTR of a luciferase gene *LUC* either at a cap-proximal position (9bp from the cap site) or a cap-distal position (5bp preceding the start codon). The results showed that tetracycline binding to its cognate aptamer located in *LUC* mRNA could trigger strong inhibition of translation, while aptamer insertion in cap-distal position was more effective than the cap-proximal position. In sum, these studies all successfully demonstrate the robustness and versatility of synthetic riboswitches in regulating eukaryotic gene expression.

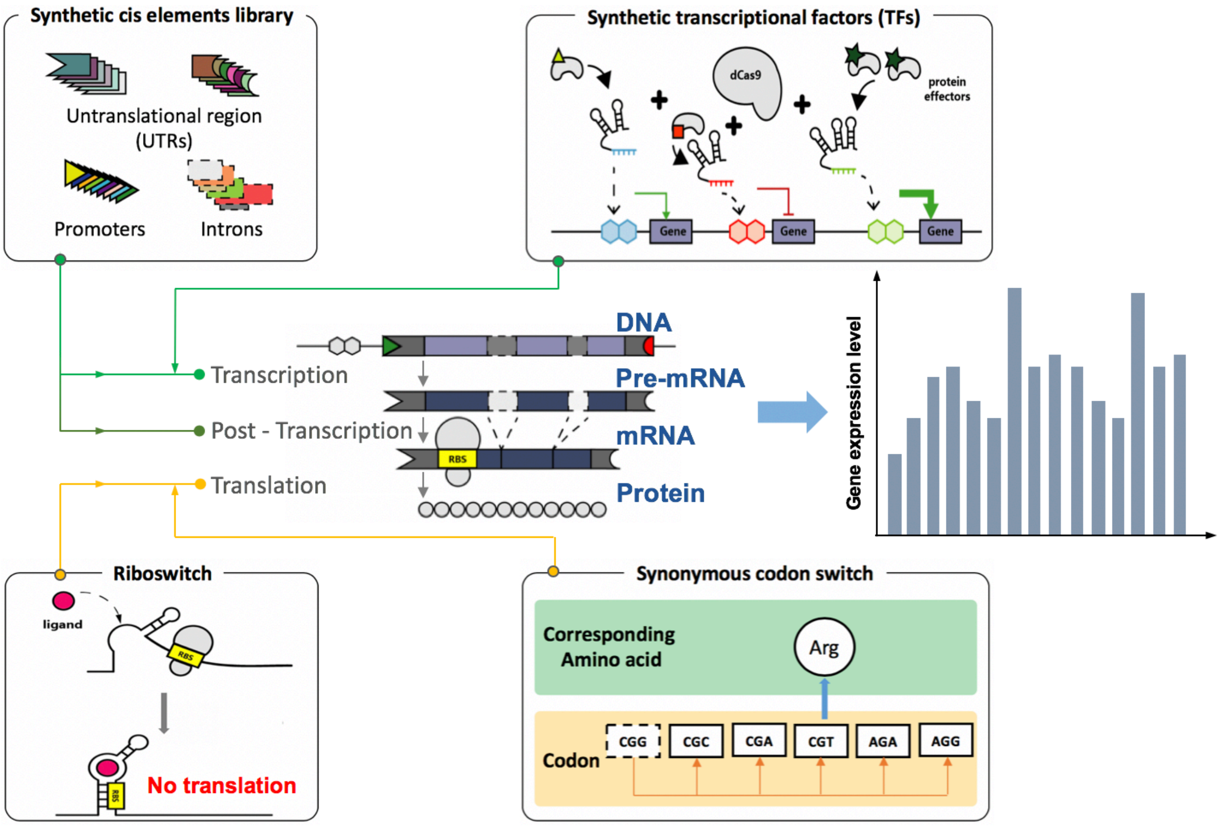


Figure 1.1 Gene expression regulation through synthetic functional parts

Gene expression can be regulated through following approaches: Synthetic promoters, UTRs have been engineered to regulate gene expression level during transcription process, while synthetic introns have been used for the same purpose during post-transcription process. CRISPR/dCas9 system has also been engineered and used as synthetic transcriptional factors for gene expression regulation both positively and negatively. Riboswitch has been developed to stop translation by folding and closing ribosome binding site (RBS) after its aptamer binding different ligands. Codon optimization has also been wildly used for gene expression regulation. Green arrow indicates promotion, while red blocked line indicates repression.

Synthetic systems for emulating fundamental biological processes

Design and testing of synthetic systems in living cells represents a unique approach to reveal key biological principles to fill the gaps of our understanding of natural systems. Prokaryotic synthetic circuits have been well established for studying biological processes with rationally predefined parts[26]. More recently, synthetic systems that function properly in eukaryotic cells have also been developed for emulating some of the fundamental biological processes, such as cell differentiation.

Cell differentiation involves the generation of binary responses from positive feedbacks of transcriptional regulators. Two factors pose the challenge of understanding the regulatory mechanisms underlying positive feedbacks of eukaryotic transcription: unidentified transcription factors and poorly understood the model of "enhancer" action. To address this issue, Becskei et al. deployed a synthetic eukaryotic gene switch in *S. cerevisiae* using well-defined constitutive promoters and tetracycline-responsive trans-activator (rtTA) to activate the expression of a chromosomal integrated reporter construct[39] (Figure 1.2). They took advantage of rtTA, which is capable of generating a graded response in constitutive system by either changing the gene copy number or adjusting the concentration of inducer doxycycline, to investigate positive feedback with a mathematical model they built. Two distinct subpopulations of cells, "on" and "off" cells, could be identified in this simplified gene activation system. With increased inducer concentrations the number of "on" cells marked by fluorescence goes up, while the number of "off" cells that don't have fluorescent signal goes down. As another example, Ajo-Franklin et al. successfully applied a rationally designed synthetic positive feedback network to achieve sustained transcriptional activation upon a transient stimulus[40]. The synthetic device consists two modules that are both under the control of an inducible promoter: synthetic factors labeled by fluorescent protein (activator) and their cognate transcriptional targets (reporter) (Figure 1.2). They evaluated the functionality of the introduced activator and reporter upon transcriptional induction and over several cell generations following sustained gene activation, and also quantified their effects on gene expression. The results showed that a simple well-defined synthetic circuit with positive feedback loop could generate long-term cellular memory, which holds potential for mechanistic understanding of positive feedback regulation during cellular differentiation.

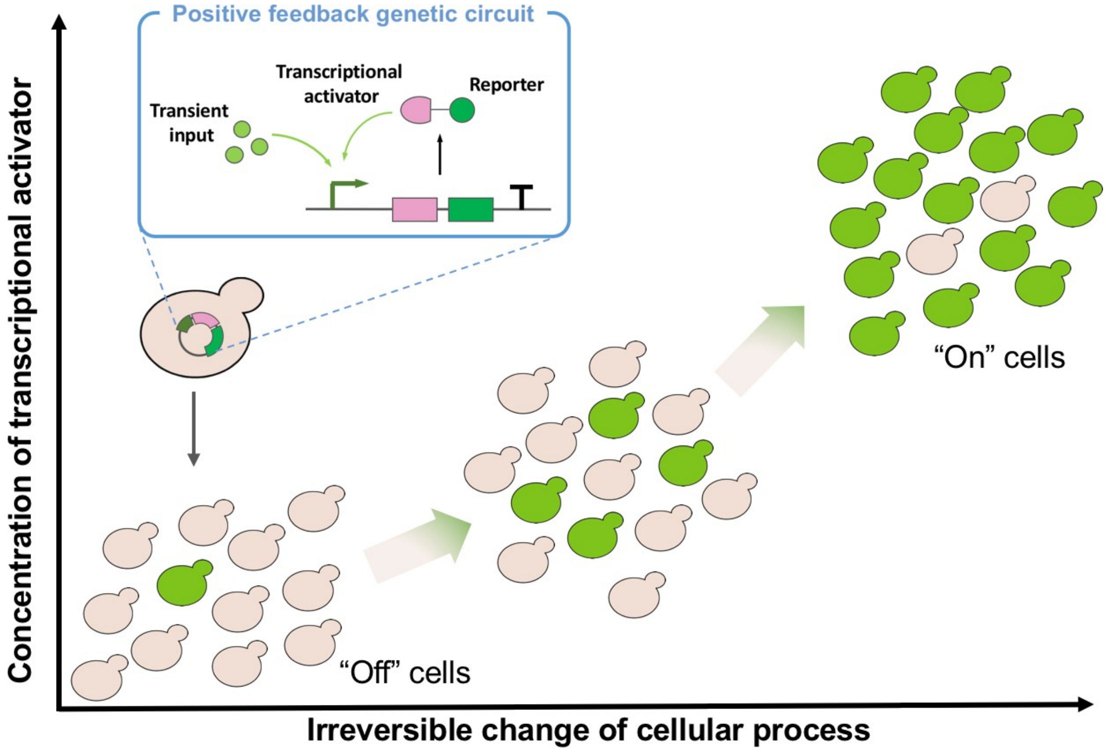

Positive feedback genetic circuit
Transcriptional activator
Reporter
Transient input
"Off" cells
"On" cells
Concentration of transcriptional activator
Irreversible change of cellular process

Figure 1.2 Synthetic network for emulating cell differentiation process

A positive feedback genetic circuit, which could generate binary responses from positive feedbacks of transcriptional activator, is designed and incorporated into engineered yeast. This system is capable of generating a graded response: with increased inducer concentrations the number of "on" cells marked by GFP fluorescence goes up, while the number of "off" cells that don't have fluorescent signal goes down.

Engineered yeasts with novel functions

The above-mentioned pioneering studies applying simple synthetic systems to yeast aim to describe and understand intrinsic cellular properties and functions. In addition, there are also efforts attempting to rewire cellular physiology by implementing completely novel functions, such as artificial cell-to-cell communication, which offer the possibility to control cellular behavior at the population level. Cell-cell communication is a widely adopted natural process in both prokaryotes and eukaryotes: in prokaryotes cell-cell communication is mostly based on quorum sensing of small diffusible molecules[41], while in eukaryotes like yeast is mostly mediated by pheromones[42]. In recent studies, both approaches have been successfully re-engineered and implemented within host yeast cells to achieve artificial cellular communication for various purposes[43]:

Engineered yeasts with quorum sensing function

Bacterial cell-cell communication driven by quorum sensing depends on the utilization of small diffusible molecules (also referred as autoinducers), such as *N*-acylhomoserine lactones (AHLs), 2-alkyl-4-quinolones (AQs) and 2-heptyl-4(*1H*)-quinolones (HHQ)[41]. Sender cells signed receiver cells generating a density-dependent response within the entire population. A similar approach has also been applied to yeast cells by incorporating signaling pathway components from other species for quorum sensing behaviors in yeast. Chen and Weiss used the plant hormone isopentenyladenine (IP) originated from *Arabidopsis thaliana* as an intercellular signaling molecule and its corresponding receptors to achieve positive feedback quorum sensing function in *S. cerevisiae* yeast[44] (Figure 1.2a). In this system, “sender” cells are engineered to synthesize diffusible cytokinin IP, which can activate the expression of IP receptor AcCRE1 and subsequently activate its downstream nuclear aspartate response regulator SKN7 in neighboring engineered “receiver” cells, which can further activate the synthetic SKN7 response promoter driven GFP reporter in a density-dependent manner (that is, quorum sensing). As another example, Khakhar et al. utilized plant hormone auxin and CRISPR transcription factors to engineer “sender” cells and “receiver” cells for orthogonal and tunable cell-cell communication[45]. A large part of the indole-3-acetamide (IAM) pathway from *Agrobacterium tumefaciens* was integrated into “sender” yeast cells so that the cells can produce auxin by converting IAM provided in the culture medium, while “receiver” cells were engineered to co-express a fusion protein, which contains dCas9, an auxin-sensitive degron from *Arabidopsis*[46], a transcriptional activator, and a CRISPR guide RNA specific to a GFP reporter. When exposed to IAM, “receiver” cells express GFP in the absence of “sender” cells, and during co-culturing of both cells, auxin produced by “sender” cells can trigger the degradation of the fusion protein and subsequently lead to a reduction of GFP expression in “receiver” cells. Both examples mentioned above share several important features such as orthogonality to host cellular machinery, tunability and easy-to-reconfigure. These features make synthetic quorum sensing system a unique approach to investigate highly complex population behaviors.

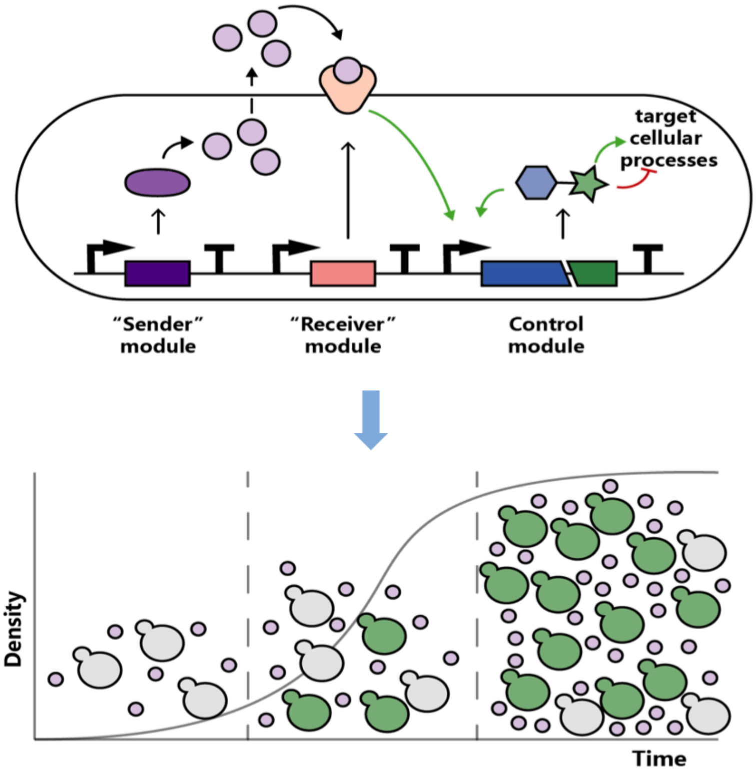


Figure 1.3 Engineered yeasts with quorum sensing function

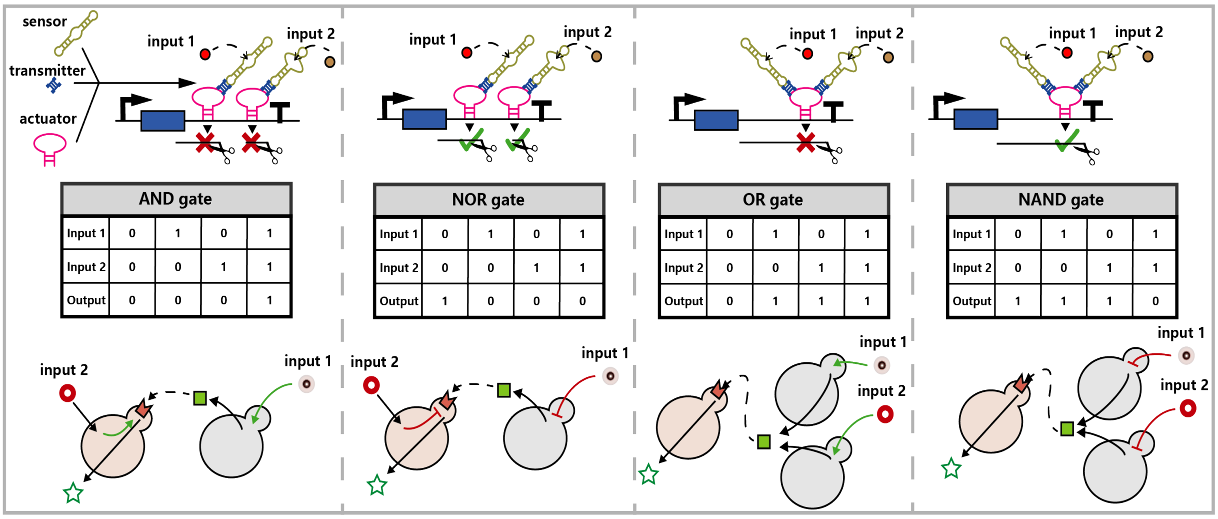
Engineered "Sender" module enables the yeast to produce autoinducers, while engineered "Receiver" module enables the cells to sense the autoinducers and activate the downstream pathways encoded by the control module (positive feedback loop and GFP reporter as example here). These three modules incorporated in yeast cells can lead to density-dependent population behaviors. Each module was simplified and represented by one transcriptional unit.

<u>Engineered yeasts with biological computation function</u>

Computation by engineered biological system has also become an emerging research subject in synthetic biology[47]. AND gates[48], multi-input logic gates[49], counting[50] and oscillators[51] have been achieved in re-engineered prokaryotic cells. Similarly, synthetic biologists also successfully re-engineered eukaryotic cells in single population, as well as multiple subpopulations, to perform logic operations.

Win and Smolke engineered RNA-based devices and developed a framework for using these RNA devices to perform cellular information processing, transduction and control operations in a single yeast population[52] (Figure 1.4). The single-input/output RNA devices are assembled with three modules: a sensing module (an RNA aptamer), an actuation module (a hammerhead ribozyme) and a transmitter module designed to couple both sensing and actuation modules. The RNA device is imbedded in the 3'UTR region of the target gene. By adding input molecules, the input-sensor binding will lead to the activation of actuation module, i.e. self-cleavage of ribozyme and subsequently inactivate the target gene expression. The authors also succeeded in extending the utility of the engineered RNA device by using different module assembly strategies and different integration sites so that the assembled devices can perform logic operations (AND, NOR, OR and NAND gates) and signal filtering in a cooperative manner. When it comes to highly complex computational tasks, engineering in single cell population is challenging as it would require functional integration of many more complex regulatory components, which often introduces metabolic burdens to the cells. To overcome this one possible solution is to implement different computation modules separately into each subpopulation of the microbial consortium (also called as distributed computation) to simplify the design and optimization of logic functions and reduce potential metabolic burdens. In general, distributed computation can be achieved by combining logic function devices and efficient cell-cell communication system. As a proof-of-principle, Regot et al. managed to perform a broad variety of computational functions with only a small library of engineered yeast cells with each subpopulation only responds to one external input and/or one single diffusible molecule for wiring purpose[53]. Another advantage of this strategy is that instead of engineering a highly complicated single-cell type derived system, each engineered cell type is reusable and can act as functional modules to complete more complicated computing task.

To summarize, the design and implementation of biological computing systems in either single or multiple cell populations can greatly enhance our ability to learn and further control existing biological systems, and hold great potential for biotechnological and medical applications.



AND gate				
Input 1	0	1	0	1
Input 2	0	0	1	1
Output	0	0	0	1

NOR gate				
Input 1	0	1	0	1
Input 2	0	0	1	1
Output	1	0	0	0

OR gate				
Input 1	0	1	0	1
Input 2	0	0	1	1
Output	0	1	1	1

NAND gate				
Input 1	0	1	0	1
Input 2	0	0	1	1
Output	1	1	1	0

Figure 1.4 Engineered yeasts with biological computation function

Novel biological computation function created by RNA devices (upper panel) and distributed computation (middle panel) were illustrated. The RNA devices used to create biological computation function is composed with different sensor molecules, transmitter molecules and actuator molecules. When input signal molecules are sensed by the sensors, the signal will transmit to actuator by the transmitter, and the devices located at the 3'UTR will transform into either "ribozyme-active" state that results in self-cleavage (indicate by scissor icon) of transcript or "ribosome-inactive" state that stop self-cleavage. The distributed computation consists of different subpopulation cell (lower panel). Different subpopulations were engineered differently so that they played different roles in the computation.

Engineered yeasts with versatile social behaviors

Natural ecosystems are highly complex and not well understood due to challenges in multi-parameters analysis and lack of manipulation tools. Synthetic ecosystems designed to recapitulate key features of natural ecological habitats, on the other hand, hold the promise to dissect major factors underlying specific ecosystem behaviors and therefore has led to a growing interest in studying prokaryotic and eukaryotic consortium[54,55].

One important feature often observed in natural ecosystems is symbiosis, the co-existence of organisms from the same or different species in a cooperative manner such that each species benefits from the other's activity. Symbiosis has also been studied in a synthetic biology context. Shou et al. studied cooperative interactions between two engineered yeast populations that have no prior known interaction by constructing a synthetic obligatory cooperative system[56] (Figure 1.5). Two auxotrophic yeast strains with same mating type were selected: one requires adenine and the other requires lysine. Co-culturing of these two strains in the absence of adenine and lysine will lead to eventually death of both strains. To create a cooperative behavior, each strain was genetically modified with one requiring adenine but overproducing lysine, while the other required lysine but overproduced adenine. By co-culturing the two mutant yeast strains together, nutritional cross feeding is established. This is reflective of the obligate mutualistic system that both populations need to constantly cooperate to survive, which provide insights of population dynamics of co-habituating communities.

The above example focuses on how cooperative entities sacrifice to survive in an ecosystem at a cost to themselves. In nature, there is another cooperative interaction with nonlinear benefits been incorporated when exploiters or cheaters reap the benefits from the cooperators and forego costs. Investigating this cooperative behavior in a multi-agent competing system, also known as snowdrift game, will provide insights into the origin of coexistence of cooperators and cheaters during evolution. Gore et al. investigated the snowdrift game dynamics with a synthetic ecosystem in which histidine auxotrophic strains serve as the cooperator, which secretes invertase to hydrolyse sucrose outside the cytoplasm, and a mutant strain lacking an invertase gene as the cheater[57] (Figure 1.5). The cooperator cells can utilize sucrose, the only carbon source provided in the culture, to grow, while the cheater cells cannot grow on their own but instead benefit from the sugars produced by cooperator cells. By adjusting the amount of histidine supplement as well as the glucose concentration, the authors were able to alter the competition outcomes, which shed light on the mechanisms underpinning cooperation in nature.

In a recent study, the sociability of yeast cells, i.e. self-communication (asocial behavior) and neighbor-communication (social behavior), was also successfully manipulated by engineering yeast cells to secrete and sense the mating pheromone[58] (Figure 1.5). Youk and Lim imbedded a synthetic secrete-and-sense circuit in the budding yeast cells, which enabled the cells to secrete and sense α-factor. As a response after sensing the mating pheromone, the engineered cell would express GFP under the control of the α-factor-responsive FUS1 promoter. By co-culturing with “sense-only” cells that can only sense α-factor but cannot produce α-factor, self-communication and neighbor-communication could be distinguished. The authors also studied the key parameters affecting the social mode such as cell density and α-factor secretion rate. Finally by engineering a positive feedback strain expressing Bar1 protease, that degrades α-factor, they achieved tunable ratios of biomodal population. This impressive approach successfully measured the effects of synthetic circuit on intra/inter-cellular interactions and population-level behaviors, which could further help understand higher-order multicellular behaviors.

It is exciting to see that synthetic ecosystems have been widely adopted to help us understand the key factors affecting the microbial communities and study ecological theories. Benefiting from their reduced complexity and improved manageability when compared to natural ecosystems, synthetic ecosystems found their way into not only fundamental research, but also practical applications such as industrial production of chemicals and bioremediation of contaminated areas[59].

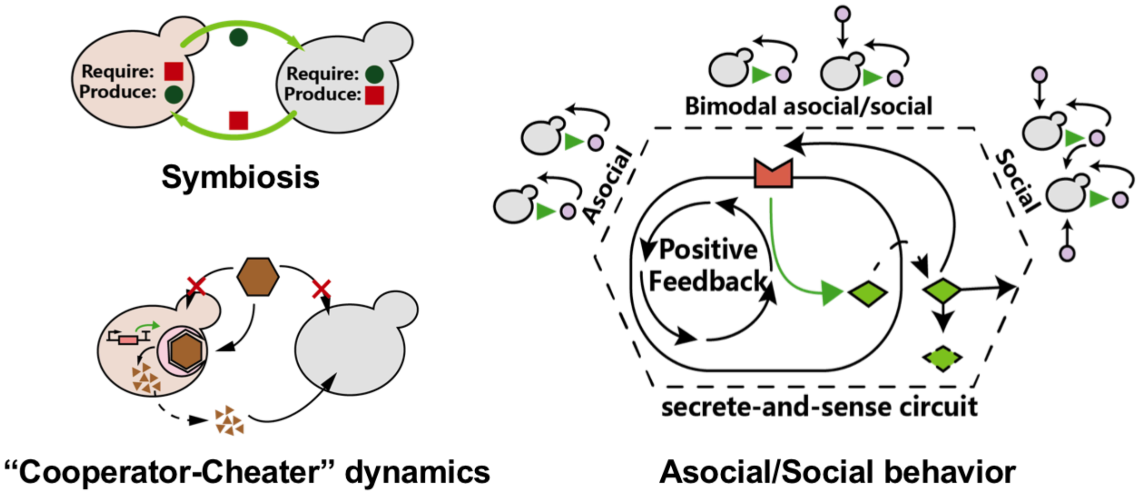


Figure 1.5 Engineered yeasts with versatile social behaviors

“Symbiosis” behavior has been achieved by engineering two yeast populations carrying two different auxotrophic marker to establish nutritional cross feeding. “cooperator-cheater” dynamics has been accomplished by modulating the amount of nutrition supplement and carbon source of culture, in which two different engineered yeast populations were cultured. One population can utilize the carbon source directly to produce sugar, while the other can only survive with the produced sugar. Asocial/Social behavior has been manipulated by engineering yeast cells to secrete and sense the mating pheromone.

Apart from being used as tools for answering fundamental biological questions, yeast has also served as the industrial production workhorse for centuries. In addition to its traditional use in alcoholic beverages manufacturing, yeast has now become specialized cellular "factories" whose genomes and native metabolism pathways are amenable to extensive re-engineering. One prominent example is the production of artemisinic acid in baker's yeast[60]. By modifying the native yeast mevalonate pathway and expression of the *A. annua* amorphadiene synthase and cytochrome P450 monooxygenase genes, researchers successfully re-directed the yeast metabolic flux towards a high-yield production of the drug precursor, achieving titers of up to 100 mg per liter. The artemisinic acid could be secreted from the engineered yeast cells, which allowed for an easy and low-cost purification of the product, opening a possibility for scale-up industrial productions.

To summarize, in this section, several successful examples of synthetic biology-driven approaches using yeast have been described. These covered a wide range of diverse scientific initiatives, which include: building custom genetic circuits to unravel native cellular mechanisms, introducing novel functions to yeast chassis as well as re-wiring the host metabolism to produce value-added chemicals and proteins. The capability of using synthetic biology tools in yeast is based on quantitatively predicting the synthetic system's behavior and implementing cellular modifications. By applying tools including DNA assembly techniques, recombinant DNA techniques, pathway modification and analysis, genome-scale engineering and modeling, synthetic yeast has been widely used for improving our understanding of fundamental biology as well as developing a multitude of practical applications.

1.2 Current challenges of synthetic genomics

As reviewed in the previous section, synthetic biology has shown great potential in integrating engineered parts or modules into existing biological systems for novel functions. As for the ultimate goal of synthetic genomics, however, currently there are, among others, four grand challenges need to be addressed:

First, following the idea of "build to understand", we have seen how synthetic biology can transform our basic understanding of life forms. However, we also need to admit that it is still far fetched de novo design and create artificial life forms. And therefore most, if not all, synthetic biology studies are largely designed to facilitate our understanding of existing life form. Taking the first synthetic life form Synthia (Syn1.0) as an example, the 1079–kilobase pair (kbp) synthetic genome of *M. mycoides* contains 901 genes[61]. Recently Synthia genome was further reduced to a 531kbp containing 473 genes through several cycles of design, synthesis, and testing [62]. And surprisingly, there are still 149 genes, 30% of the whole genome, have unknown biological functions that are essential for life. Creating a truly living and dividing cell from scratch by de novo synthesis requires the comprehensive knowledge of genes essential for life and how each gene functions and interacts. Despite the fact that model prokaryotic and eukaryotic organisms such as *E. coli* and budding yeast *S.cerevisiae* have been extensively studied over decades, our understanding is still limited. In addition, many widely used industrial strains were chosen based on historical rather than scientific grounds. Therefore, extensive genetic and phenotypic studies of nonconventional organisms are warranted to improve our knowledge and provide more guidance during synthetic genomic design.

Second, the entire *Mycoplasma genitalium* (*M. genitalium*) genome was synthesized from oligonucleotides, assembled in budding yeast into a complete genome. This genome transplantation method heavily depends on accurate design of a viable genome, and the low success rate of the transplantation step makes it challenging to restore fitness when defects arise. The issue here can be even more problematic if the complexity and size of the target genome increases significantly. Therefore, highly-efficient construction strategies, that can be applied to various systems, are crucial. In addition, establishing high-efficient methods for detecting bugs and debugging is also crucial since unforeseen design flaws or unpredictable events many occur during the construction of designer genome.

Third, for a designer genome, maintaining a near wild type viability is one of the most essential requirements, which is important for further engineering, evolution and applications in subsequent studies. Apart from genotype confirmation by sequencing, phenotypic assays, such as monitoring growth and cell morphology, are normally applied to test the viability of the synthetic cells. However, this strategy is only applied when a fitness defect has been detected. Again, considering the increase in complexity of designer genome in the future, the lack of in depth characterization strategies could potentially lead to great amount of unnecessary troubleshooting efforts. Therefore, comprehensive and systematic characterization strategies are required which thoroughly interrogate of the function of designer cells. Such strategies will identify subtle changes not detected by phenotypic analysis and also reveal functions of certain synthetic elements.

Last but not least, from Synthia to the world first synthetic eukaryotic chromosome then to the announcement of The Genome Project-Write[23], we foresee an ever-increasing demand of DNA synthesis for synthetic genomics. Even though the cost of DNA synthesis has been steadily decreasing over the last few years, it is still exceedingly expensive and time consuming to construct an entire eukaryotic genome. Therefore, future efforts to overcome this limitation are necessary. In this regard, automated DNA synthesis and assembly technologies and platforms constitute a very promising direction to further reduce the cost as well as increase the throughput.

Currently, as mentioned above, synthetic genomics is faced with various mounting challenges both technically and scientifically. With dedicated efforts, however, these difficulties will eventually be overcome and the full potential of synthetic genomics will be unlocked in the future, which consequently will bring about a paradigm shift in biotechnological research and applications.

1.3 Introduction of *Sc2.0*

Apart from application-driven genome scale manipulations with genome engineering tools such as CRISPR-Cas systems[63-65], synthetic genomics can also be used to answer a wide variety of questions about fundamental properties of genome structure and organization, gene content, and genome evolution. Re-engineering the whole genome has been achieved in virus and bacteria[12,15]. Re-engineering eukaryotic genome poses more challenges due to the increased complexity. Regardless however, significant progresses have already been made recently with the eukaryotic organism yeast.

Background of *Sc2.0* project

The *Sc2.0* project, which aims to re-design and synthesize all 16 chromosomes to build an entire synthetic *S. cerevisiae* genome, is considered to be one of the most ambitious current synthetic genomics projects. Three principles are developed to guide the genome design, namely genome viability, stability and flexibility. Based on these principles, several design elements are incorporated to generate the synthetic genomes (Figure 1.6). First of all, a number of non-essential features, such as transposons, introns with annotation of non-essential functions, sub-telomeric repeats, are deleted from the blueprint genome. Furthermore, one of the three stop codons is eliminated from the genome to allow the incorporation of an alternative artificial amino acid, by which protein properties can be altered. tRNAs are known regions of genome instability[66,67], therefore, in order to isolate the hotspots for transposition and genome rearrangements from synthetic chromosomes, all tRNAs are relocated to a separate "neochromosome" for overall increased genome stability. Together, these modifications lead to a roughly 4-21% reduction of chromosome size in Sc2.0 yeast (Table 1.1). The most novel element in *Sc2.0* is the introduction of LoxPsym site that flanks each non-essential gene in all synthetic chromosomes. By doing so, it could, in theory, generate combinatorial genomic diversity through rearrangements at designed sites upon expression of Cre recombinase, which is designated as "SCRaMbLE" (Synthetic Chromosome Rearrangement and Modification by LoxP-mediated Evolution). The power of this build-in tool has been demonstrated by Shen et al. on Synthetic chromosome arm SynIXR, which enables dramatic structural diversity after SCRaMbLE[68].

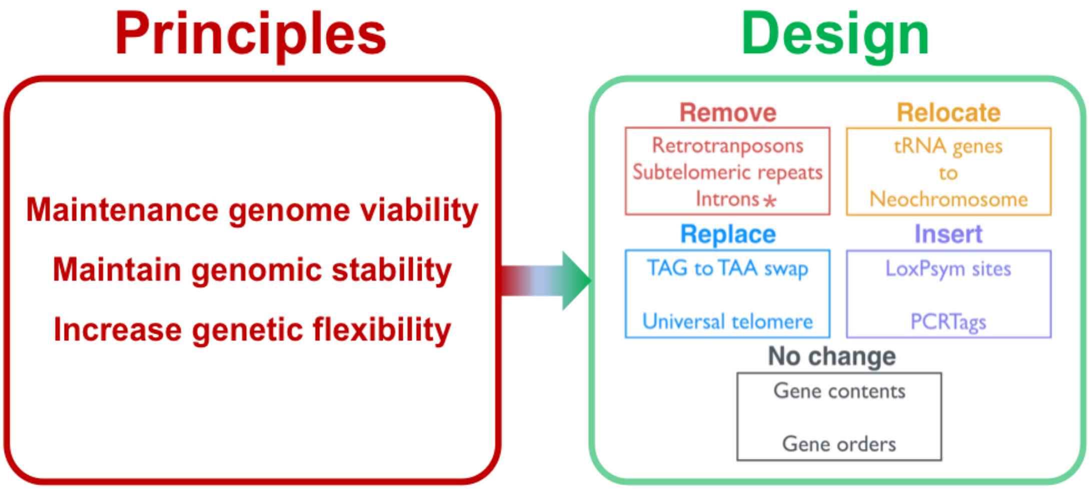


Figure 1.6 Sc2.0 Design principles and design summary

Four types of designs are introduced: retrotransposons, subtelomeric repeats and introns with annotation of non-essential function (* ribosomal related introns are not deleted in designer sequence due to their role in maintaining cell viability[69]) are removed. All tRNA genes are relocated to a separate "neochromosome" for overall increased genome stability. TAG stop codons are replaced by TAA stop codon to allow the incorporation of an alternative artificial amino acid and universal telomere of the simple ~350bp sequence repeat $(TG_{1-3})_n$ is incorporated to replace original telomeres. PCRTags are alterations incorporated into most open reading frames (ORFs) and served as water marker. LoxPsym site that flanks each non-essential gene in all synthetic chromosomes is used as a tool to induce designated genome rearrangement.

All above mentioned design principles are incorporated into the design sequence *in silico* using the in-house developed software "Biostudio"[70]. Each designer chromosome sequence is then divided in a hierarchical order into megachunks (~30-50kb), chunks (~10-15kb), minichunks (~2-3kb), building blocks (~750bp) and oligos (~60-80bp) to facilitate a step-wise construction. To accelerate the completion of *Sc2.0* project, 9 institutes from US, UK, China, Australia and Singapore were assigned the task of different chromosomes and proceed independently. The general construction strategy is called "SwAP-In" (Switching Auxotrophies Progressively for Integration), which takes the advantage of the high-efficient homologous recombination in yeast to facilitate the replacement of native sequence with corresponding synthetic sequence. In order to further accelerate the progress, different construction strategies are also developed in the *Sc2.0* consortium. For example, *SynII* and

SynVII are constructed from two initial strains in parallel and combined by I-*Sce*I mediated method[18] (detailed method can be found in Chapter Two), while *SynXII* was built from six initial strains and combined by meiotic recombination-mediated assembly strategy[22]. To generate the final synthetic yeast, carrying all synthetic designer chromosomes, a meiotic-mediated method called "endoreduplication backcross" has also been developed[20].

	WT size	SYN Size	No. of stop codon swaps	No. of loxP sites added	bp of PCRTag recoded	No. of tRNA deleted	bp of repeats deleted
chr01	230208	181030	19	62	3535	4	3987
chr02	813184	770035	93	271	13651	13	7030
chr03	316617	272195	44	100	5272	10	7358
chr04	1531933	1454671	183	479	25398	28	11674
chr05	576874	536024	61	174	8760	20	11181
chr06	270148	242745	30	69	4553	10	9297
chr07	1090940	1028952	126	380	17910	36	13284
chr08	562643	506705	61	186	9980	11	19019
chr09	439885	405513	54	142	7943	10	11632
chr10	745751	707459	85	249	12582	24	7523
chr11	666816	659617	68	199	11769	15	4214
chr12	1078177	999406	122	291	15129	19	10843
chr13	924431	883749	100	337	15911	21	7673
chr14	784333	753096	96	260	13329	14	5115
chr15	1091291	1048343	147	399	18015	20	9542
chr16	948066	902994	127	334	15493	17	10048
Total	**12071297**	**11352534**	**1416**	**3932**	**199230**	**272**	**149420**

Table 1.1. Summary statistics for design of Sc2.0

(WT, wild type; SYN, synthetic.)

To date, 6.5 chromosomes have been constructed, with 2.5 chromosomes consolidated into one strain. Once finished the hope of *Sc2.0* synthetic genome will serve as a powerful analytical tool to enhance our knowledge of yeast biology.

1.4 Objectives of this study – *SynII* & *SynVII*

Among the 16 chromosomes, *SynII* and *SynVII* are considered relatively long chromosomes. The native chromosome *II* is 807 kb in length and includes 410 open reading frames, 13 tRNAs, and 30 introns(*13*), while the native chromosome VII is 1090 kb in length, carrying 583 coding genes, 36 tRNAs and 26 introns[71]. *SynII* and *SynVII* were designed based on the native sequence following rules mentioned above, resulting in two “designer” chromosomes 770 kb and 1029 kb in length. Both are 5% shorter than their native counterparts. The current study focuses on using *SynII* and *SynVII* to explore the following topics:

Construction strategy

Current construction strategy “SwAP-In” works very well for small chromosomes, such as chromosome I or III, as it only requires a few runs to complete the construction. However, it does not work efficiently for megabase long chromosomes like *II* or *VII*. Any unexpected problem during the construction will stall the entire process. Therefore, a high-efficient construction strategy, especially for large chromosomes, is necessary and could be potentially applied to other large chromosome constructions in the future.

Comprehensive phenotypic assays

For each designer chromosome in Sc2.0 projects, thousands of changes have been introduced, which naturally raises the questions whether there are any phenotypic impacts and how to quickly identify the cause. Classical phenotypic assays would not be enough to reveal subtle changes. Therefore, more in-depth analysis should be established, e.g. examining DNA, RNA, protein and metabolic profiles, checking 3D genome structure and monitoring cell replication and segregation. For defect mapping, the modular construction approach applied in Sc2.0 project provides an important mechanism to systematically discover and repair phenotypic defects during construction. Therefore, it will greatly reduce the workload of debugging by narrowing down the target into one 30-50 kb megachunk region. However, additional efforts are still needed for efficient detection and targeting defect regions.

Use of SCRaMbLE to select novel properties

One of the most novel elements in *Sc2.0* is the SCRaMbLE system. By incorporating SCRaMbLE, the designer Sc2.0 yeast genome holds the potential to be further evolved, generating strains that encode various phenotypes. It will be interesting to see if SCRaMbLE allows designer yeast cells to quickly adapt to specific stress conditions and how the genome diversity introduced by SCRaMbLE can be used as a resource to study scientific questions.

1.5 Thesis outline

This thesis is structured as follows. Chapter Two presents the construction of *SynII* and *SynVII*. Section 2.2 introduces the use of acoustic dispensing technology to facilitate DNA assembly at very low volumes. The construction and variation repair of both synthetic chromosomes are described in Section 2.3 and 2.4. Chapter Three focuses on the full fitness characterization and fitness defect rescue of *SynII* and *SynVII.* Chapter Four is devoted to the introduction of SCRaMbLE system. The sequence reconstruction workflow using deep sequencing data is described in Section 4.1. Next, using SCRaMbLE to rescue the fitness defect in *SynII* is demonstrated in Section 4.2. Finally, Chapter Five points the conclusions and possible future research directions.

1.6 Summary of contributions

The design of *SynII* and *SynVII* was accomplished by team led by Dr. Jef Boeke and Dr. Joel Bader. The majority of work for *SynII* and *SynVIIR* construction was done by Tai Chen and Yang Deng from BGI. Construction of *SynVIIL*, structure variation repairing of *SynII* and *SynVII*, and biological characterization were accomplished by myself. Work of DNA assembly methodology with acoustic dispensing technology was led by Paulina Kanigowska, myself and Sarah Zheng. Bioinformatics analysis workflow was co-developed by me, Yun Wang and Giovanni Stracquadanio. SCRaMbLE studies of *SynII* and *SynVII* were accomplished by myself.

<u>**Chapter Two**</u>

Part 1: Build - the design and construction of two megabase long yeast chromosomes

2.1 Overview of Sc2.0 chromosome design and construction strategy

The Sc2.0 chromosome design was specific to *S. cerevisiae* reference sequence based on the S288c wild type strain derivatives [reference version updated by the *Saccharomyces* Genome Database (SGD) on February 3rd 2011]. A series of edits including deletions, insertions and base substitutions were introduced to the reference sequence of each chromosome by BioStudio design suite[70] followed by manual check. For each step, a version number was created. Take *SynII* as example, version 2_0_00 refers to the wild type sequence, while 2_1_00 indicates an intermediate *SynII* sequence with PCRTags added. Then stop codon swap, loxpsym site insertion and intermediate editing steps were introduced, generating the final version 2_3_22.

Complete chromosome re-design was performed using the BioStudio design suite[70] conforming to the Sc2.0 project standards to yield the synthetic yeast chromosome *II* and *VII* (termed *SynII* and *SynVII*). Relative to the wild-type reference sequences, *SynII* has 33 deletions (53,605bp in total), 269 insertions (10,456 bp in total) and 14,949 single nucleotide substitutions, while *SynVII* has 93 deletions (69,292 bp in total), 380 insertions (12,920 bp in total) and 20,745 single nucleotide substitutions (Table 2.1, Table 2.2). Although *SynII* and *SynVII* reflect 5.3% and 5.7% reductions in size compared to the native chromosomes, hundreds of designer features were incorporated throughout (Table 2.3).

		Number	Base alteration
Replacement	wild type telomere > Universal Telomere Cap[a]	2	24714 > 1378
	532 pairs of WT PCR tags and 532 pairs of SYN PCRTags	1064	13623
	stop codon TAG > TAA	93	93
	“landmark” restriction sites (removal or introduction)[b]	190	477
	repeatsmash of gene	1	862
Deletion	transposable element region[c]	318	21887
	gene	19	14910
	tRNA	13	993
	intron[d]	22	3450
Insertion	LoxPsym site[e]	267	9078

Table 2.1 Summary of *SynII* design

a. Wild-type telomeres together with its adjacent subtelomeric regions are replaced with universal telomere caps. b. Unique restriction sites were generated in *SynII* either by introduction of synonymous mutations in ORFs, or by removal of redundant sites to leave only one pre-existing restriction site in the *SynII* sequence. c. Transposable element region include the functional features of LTR retrotransposons, transposable element genes and long terminal repeats. d. Eight introns were retained as they reside in ribosomal subunit coding genes.

		Number	Base alteration
Replacement	wild type telomere > Universal Telomere Cap[a]	2	21094 > 1378
	705 pairs of WT PCR tags and 705 pairs of SYN PCRTags	1410	17876
	stop codon TAG > TAA	126	126
	“landmark” restriction sites (removal or introduction)[b]	261	596
	repeatsmash of gene	3	2147
Deletion	transposable element region[c]	20	51111
	gene	14	12232
	tRNA	36	2887
	intron[d]	23	3062
Insertion	LoxPsym site[e]	380	12920

Table 2.2 Summary of *SynVII* design

Description of a. to c. is the same with Table 2.1. d. Eleven introns were retained as they reside in ribosomal subunit coding genes.

Chromosome version		chr02_0_0	chr02_3_22	chr07_0_0	chr07_3_57
protein coding genes	protein coding genes	456	437	583	569
	protein coding essential genes	70	70	111	111
	protein coding genes required for fast growth	33	33	55	55
	protein coding genes with introns	29	8	24	10
	introns within protein coding genes	30	8	26	11
	PCRTags	0	1064	0	1410
	total number of TAG stop codons	98	8	131	1
	stop retain variants	0	90	0	126
	synonymous codons (modified to accommodate TAG to TAA)	0	5	0	8
	non synonymous codons (modified to accommodate TAG to TAA)	0	0	0	3
non-protein coding genes	tRNAs	13	0	36	0
	snoRNAs	2	2	6	6
	snRNAs	1	1	2	2
	ncRNA	1	1	0	0
transposons & repeats	LTR retrotransposons	3	0	6	0
	transposable element genes	6	0	10	0
	long terminal repeats	22	0	44	0
	repeat regions	1	0	1	0
chromosome features	length	813184	770035	1090940	1028952
	site specific recombination target regions	0	267	0	380
	ARSs	18	18	28	26
	gene repeat-smashed	0	1	0	3
	telomeres	2	2	2	2
	X element combinatorial repeats	2	0	2	0
	X elements	2	0	2	0
	Y prime element	1	0	1	0
	telomeric repeat	1	0	1	0
	centromere	1	1	1	1
	centromere DNA Element I	1	1	1	1
	centromere DNA Element II	1	1	1	1
	centromere DNA Element III	1	1	1	1
	megachunks	0	25	0	25
	chunks	0	103	0	129

Table 2.3 Comparison between wild-type and synthetic chromosomes II and VII

"chr02_0_0" and "chr07_0_0" refer to wild-type chromosome *II* and *VII* respectively, and "chr02_3_22" and "chr07_3_57" refer to *SynII* and *SynVII* respectively.

Compared with previous strategies[16,17], an alternative modular assembly method was devised for the construction of *SynII* and *SynVII*. This new approach enables parallel integration to speed assembly, while at the same time also provides an effective way to debug potential phenotypic defects (see below in section 2.3.2). *SynII* and *SynVII* were divided into 25 megachunks (~30 kb each) using BioStudio, and each megachunk was then segmented using in-house software "Segman" into minichunks (~3 kb, Figure 2.1), which are compatible with the Gibson assembly method[5] (Figure 2.2A) (Detailed information on megachunks, chunks and minichunk can be found at www.syntheticyeast.org).

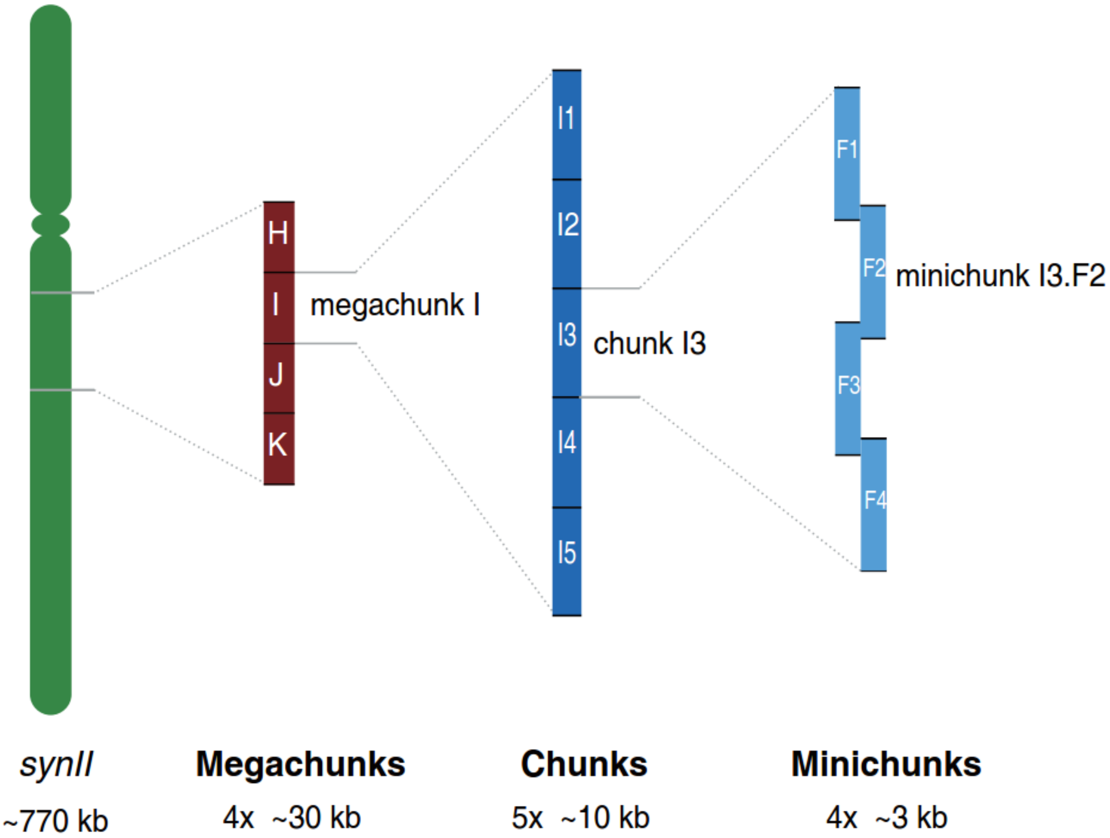


Figure 2.1 Synthetic chromosome hierarchical nomenclature

An entire synthetic chromosome (green) is segmented stepwise into ~30 kb megachunks (dark red), ~10 kb chunks (dark blue) and ~3 kb minichunks (light blue) according to the experimental design.

Instead of performing step-by-step megachunk integration (SwAP-In)[70] from one end (left or right arm), I performed integration in parallel from both ends with two separate parental strains *YS000-L* and *YS000-R* (Figure 2.2B) bearing *SynII-L/SynVII-L* and *SynII-R/SynVII-R* respectively, which overlap by 30 kb, effectively reducing the overall integration time by almost 50%. These two strains were crossed to produce a heterozygous diploid. Regions bearing substantial sequence similarity to the native chromosome remain and might trigger various unwanted meiotic recombination events, which would result in tremendous screening efforts when combining semi-synthetic chromosomes via homology-directly recombination. Therefore, to increase the frequency of mitotic recombination between two semi-synthetic chromosomes (*SynII-L/SynVII-L* and *SynII-R/SynVII-R*), an I-*Sce*I mediated strategy[72] was adopted to break both semi-synthetic chromosomes at a designed site, promoting mitotic recombination between them to generate the fully synthetic chromosome (Figure 2.2C). Induction of the lethal I-*Sce*I-mediated DNA double-strand breaks at the junctions of synthetic and wild-type sequence introduced a strong selective pressure for the two semi-synthetic segments to recombine, via a HDR mechanism. Since our recombination strategy is designed to generate a marker-less synthetic chromosome and leave the *URA3* marker on wild-type chromosome, a 10-fold increase was observed in the *URA3*-deficient colony numbers upon the I-*Sce*I induction, suggesting our strategy greatly bias the crossover events at the designated regions.

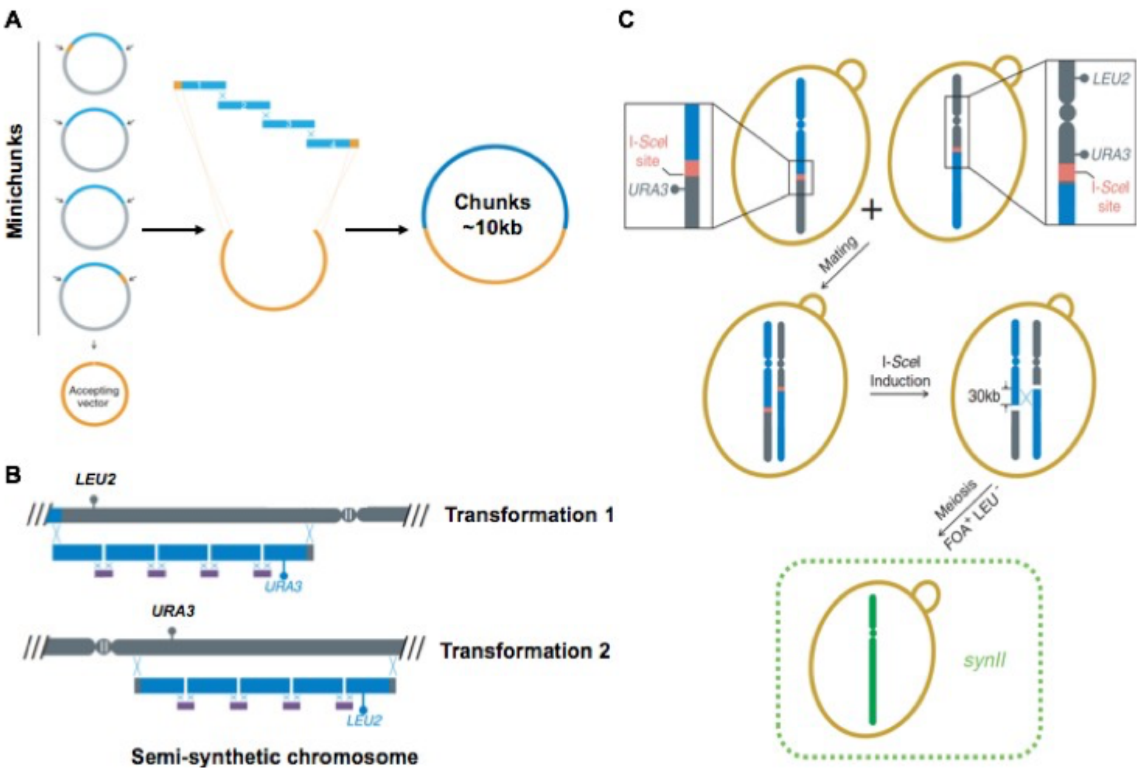


Figure 2.2 Synthetic chromosome construction strategy

A. Minichunk assembly. After restriction enzyme digestion, ~3 kb minichunks (light blue) were assembled into ~10 kb chunks (dark blue) using the Gibson assembly method. Each minichunk was designed to carry a 40 bp overlapping region with accepting vector (orange) or adjacent minichunks. B. Native chromosome replacement with synthetic chunks. On average 5 chunks were used to replace the native segments of chromosome *II* (dark gray). Step-by-step replacements were carried out in parallel from the end of the left arm and right arm respectively with iterative selectable markers (*LEU2* or *URA3*) through homologous recombination in yeast with homologous regions around 500 bp. ~1 kb junctions (purple) produced by PCR using ligated chunks as template were designed to have 500 bp regions overlapping with two adjacent chunks to improve the replacement efficiency. C. I-*Sce*I mediated *synIIA-R* and *synIIR-Y* integration. An I-*Sce*I site (red) was introduced on both semi synthetic chromosomes. After mating, I-*Sce*I induction was performed to generate double strand breaks on both semi synthetic chromosomes. 30 kb homologous region on both semi synthetic chromosomes enabled the integration of complete synthetic chromosome.

2.2 Acoustic dispensing technologies mediated nanoliter scale DNA assembly

To accelerate completion of Sc2.0, the project was decentralized by parceling out the construction of individual chromosomes to different teams. Each team is also trying various strategies to further facilitate the construction process. However, the construction effort of each chromosome, especially for megabase long chromosomes, are still labor-intensive. Take *SynVII* as example, 435 minichunks need to be assembled into 129 chunks, then followed by 25 runs of megachunk integration through homologous recombination in yeast. With current synthesis costs for Sc2.0 averaging approximately US$0.10 per base pair, the overall cost for just the Sc2.0 DNA is around US$1.25 million. The total costs of the project, including labor for assembly, genotyping, sequencing, fitness and phenotypes evaluation, debugging and bug corrections, developing and maintaining software and servers, and other activities and associated indirect costs will be considerably higher. Therefore, methods that can reduce the labor and significantly decrease the cost are much appreciated.

Traditional liquid handling technology has enabled increased throughput of many experimental protocols and assays by (1) increasing operational speeds, (2) reducing working volumes (down to a microliter range), and (3) reducing the need for error-prone manual handling, and ultimately contributed to substantial workflow cost savings. Although a big "leap forward," the demand for further protocol miniaturization continues to increase, particularly in ultra-high-throughput screening (uHTS)[73]. 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 many factors are involved, such as the shape, the diameter, and the coating of the pin, as well as 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[74]. Kong et al. have successfully used microfluidic chips to synthesize DNA sequences up to 1 kb, and Tewhey et al. used microfluidic chips to run 1.5 million PCRs in parallel[75,76]. 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; regardless, the *de novo* DNA synthesis using microfluidic chips is very complementary to the miniaturized assembly methods applied.

First described in 1927, the acoustic droplet ejection (ADE) method utilizes acoustic energy to rapidly move low-volume nanoliter to picoliter droplets without any physical contact[77]. Before it reached the laboratory setting in the 2000s, the drop-on-demand technology was first exploited in a number of other fields, such as the ink-jet printing industry and pharmaceutical companies. 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), thereby enabling 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.

Frequently used experiment protocols during the synthetic chromosome construction and verification, such as PCR, Gibson assembly, Goldengate assembly, bacterial transformation and PCRTag analysis were designed and adapted accordingly to the Echo550 robotic platform. In general, the preliminary data is very promising, showing high efficient PCR, assembly and verification can also be achieved on nanoliter scale, which would significantly reduce the cost and hold the potential to be further automated to reduce the labor dependence.

Echo PCR

Conventional endpoint PCR is instrumental in making synthetic DNA. To test the minimal volume of regular PCR using Echo, PCRs of various volumes was set up. The plasmid HcKan_P vector (concentration at 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 (Figure 2.3A). Five reaction volumes ranging from 50 to 1000 nL were set up, with a manual control at 10µL (Table 2.4). Each reaction was performed in four replicates.

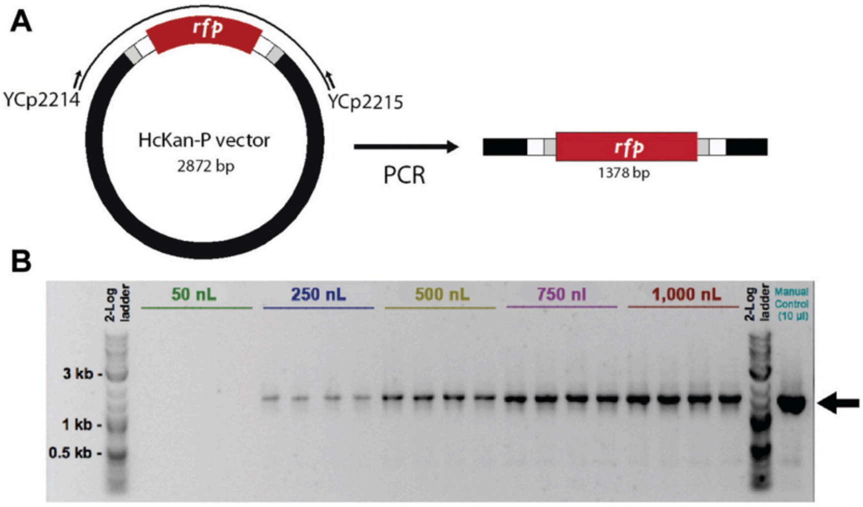


Figure 2.3 PCR setup by Echo

A. A pair of primers was designed to amplify a 1.3 kb fragment, 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. Gel stained with 1× SYBR Safe DNA stain.

We used the Echo machine to set up PCRs in total volumes ranging from 50 nL to 1 µL (Figure 2.3B and Table 2.4). Starting from 250 nL, a band of the correct size could be detected by gel electrophoresis. Because the PCR product was diluted to 5 µL for gel electrophoresis, it is possible that PCRs at 50 nL scale were successful, but the detect by gel electrophoresis was not sensitive enough to detect a 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.

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

Table 2.4 Echo PCR setup

Gibson DNA assembly

First described in 2009, the Gibson DNA assembly method[5] belongs to a group of overlap-directed DNA assembly techniques such as CPEC[78], SLiCE[79], and SLIC[80] assemblies. The Gibson assembly method is one of the most used assembly methods 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 advantages of Gibson assembly include sequence independency and the ability to generate scarless final assembled DNA products. Typically, the Gibson assembly requires 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 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.

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 (Figure 2.4A,B). 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 selection for successful assembly clones isolates, respectively. Here, RFP serves as a positive screen for correct assemblies. Four reaction volumes ranging from 50 to 1000 nL were set up, with a manual control at 20µL (Table 2.5), and each reaction was performed in triplicate.

Our result shows that Gibson assembly worked extremely well in this setting. Correct assembly was observed from as low as the 250 nL reaction volumes, and at 500 and 1000 nL the assembly efficiencies are comparable to or better than the manual control, but with notable standard deviations. This would result in cutting the reagent cost by 20-fold or more (Figure 2.4D). Even more encouraging, no background was observed (Figure 2.4C) and 100% correct assembly through Sanger sequencing across the assembly junctions, and this will be highly beneficial for future automation plans, as it will greatly reduce the colony screening effort.

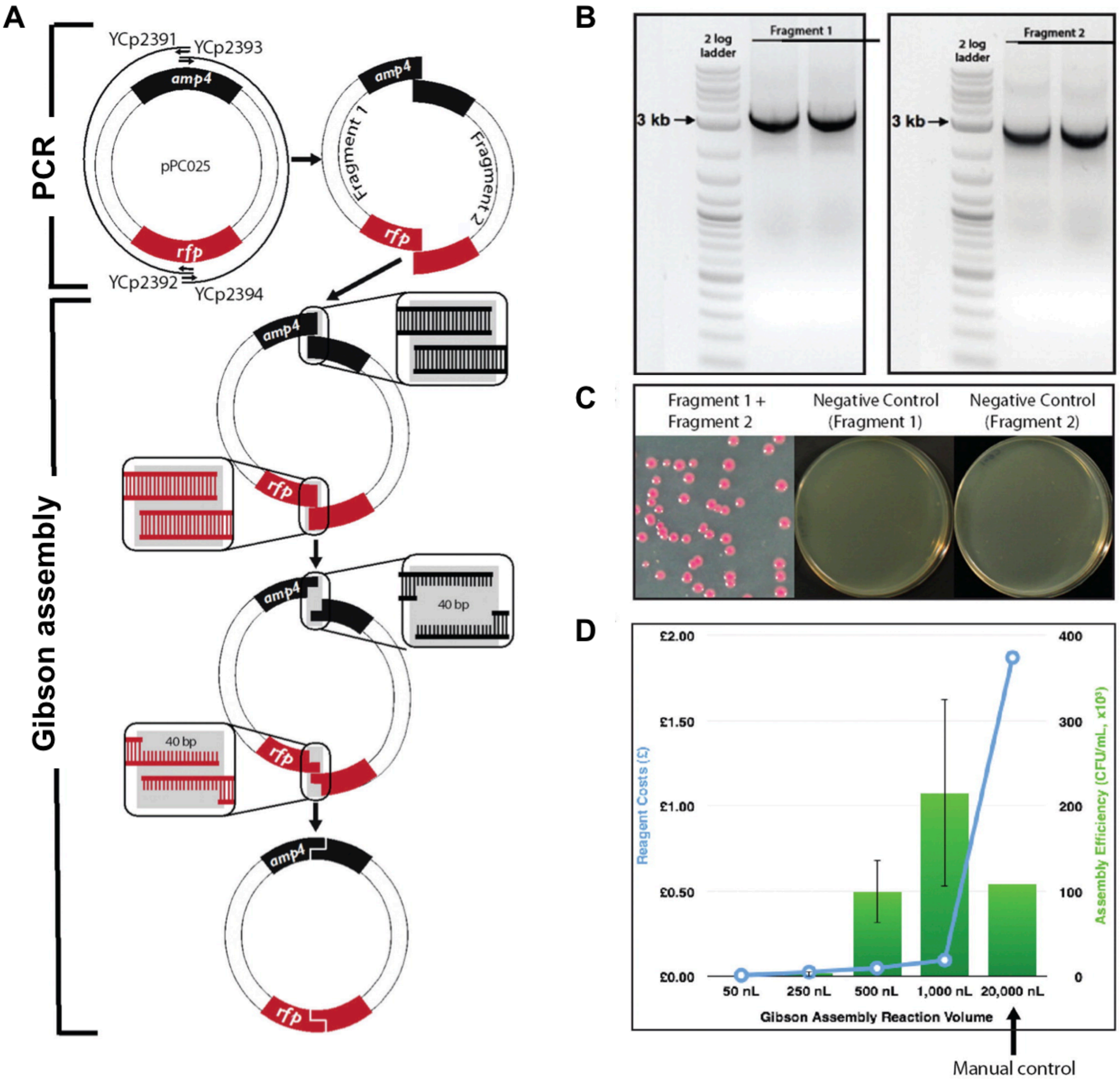


Figure 2.4 Gibson assembly reaction setup by Echo

A. The pPC025 plasmid was split into two overlapping fragments in 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 confirms the successful PCR amplification of both fragments. Gel stained with 1× SYBR Safe DNA stain. C. Successful Gibson assembly product gives rise to red bacterial colonies. The assembly efficiency was high and no background colonies (white) were observed. Negative control reactions, which had only one fragment in the reactions, yielded no colonies. D. Cost-effectiveness and assembly efficiency comparison of different reaction volumes for Gibson assembly.

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/uL)	5.0	20.0	40.0	80.0	2,500.0
Fragment 2 (86.8 ng/uL)	7.5	42.5	85.0	170	2,500.0
Total	50 nL	250 nL	500 nL	1000 nL	20,000 nL

Table 2.5 Gibson Assembly Reactions

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[81]. 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.

In this study, the HcKan_P plasmid (2.8 kb) 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 sites. The promoter pMBP1 (500 bp) was amplified directly from yeast BY4741 (*MAT***a**, 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 (Figure 2.5A,B). 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 fluorescence, which would facilitate the visual identification of correct assembled clones (white colonies; Figure 2.5C). Five reaction volumes arranging from 50 to 1000 nL were set up (Table 2.6), 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.

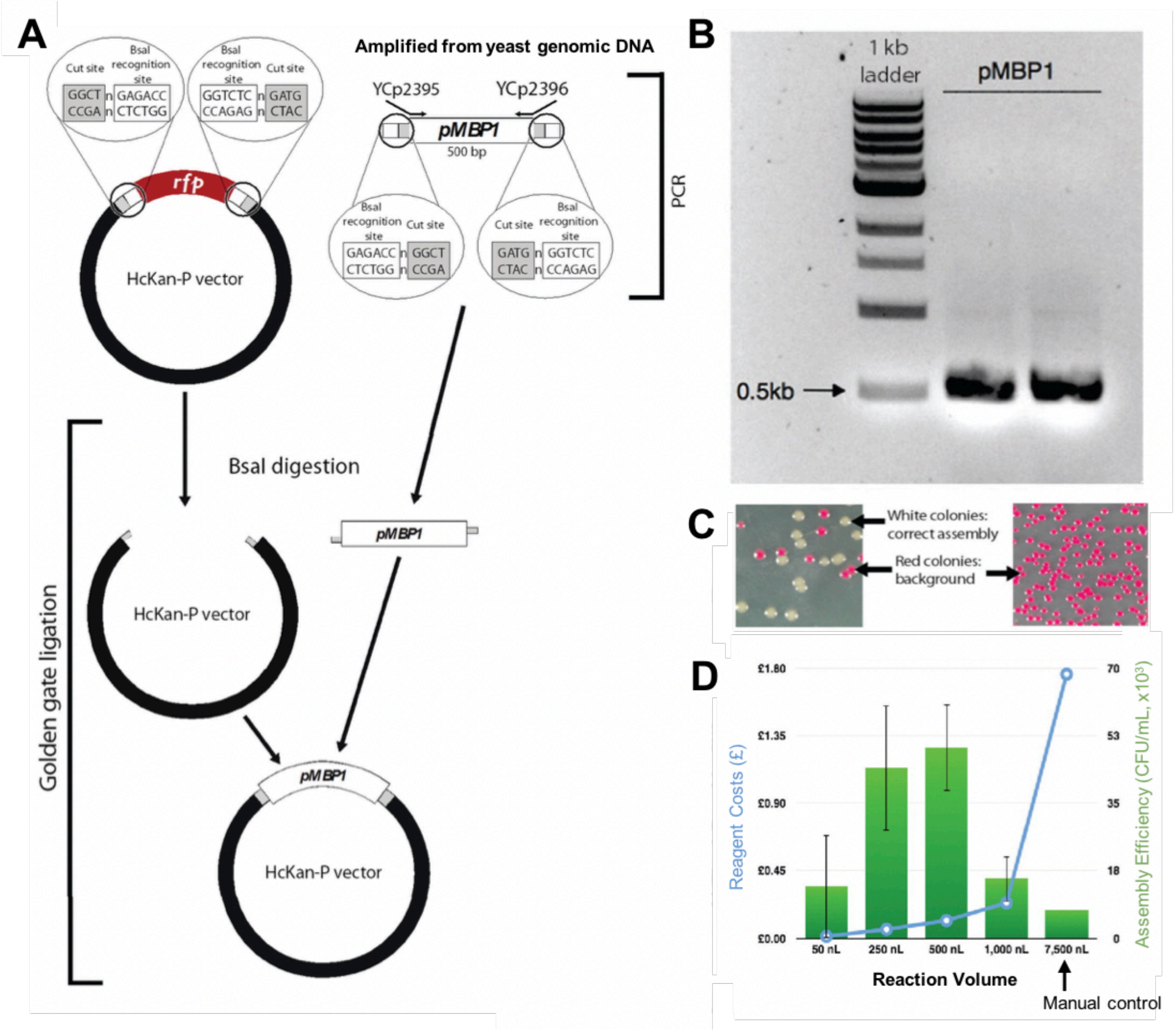


Figure 2.5 Golden Gate assembly setup by Echo

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. Gel stained with 1× SYBR Safe DNA stain. 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. Cost-effectiveness and assembly efficiency comparison of different reaction volumes for Golden Gate assembly.

With Golden Gate assembly, DNA was successfully assembled in a 50 nL reaction volume (typically 15 µL reactions when performed manually), and at the 250 and 500 nL scales the assembly efficiencies were found higher than those of the manual control. This would lead to at least a 30-fold reduction in reagent use when performing Golden Gate reactions using Echo (Figure 2.5D). Uncut vector background in the assembly was observed (red colonies, as shown in (Figure 2.5C). There are several ways to overcome this. 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.

Reagent/nL	**Echo**	**Echo**	**Echo**	**Echo**	**Manual**
Golden Gate master mix	17.5	82.5	167.5	332.5	2,500.0
pMBPI (20 ng/uL)	30.0	150.0	300.0	600.0	4,500.0
HcKan_P (10 ng/uL)	2.5	17.5	32.5	67.5	500
Total	50 nL	250 nL	500 nL	1000 nL	7,500 nL

Table 2.6 Golden Gate Assembly Reactions

In general, our preliminary data shows nanoliter PCR and DNA assemblies can be as equally efficient as manual reaction volumes. With further optimization, it should be possible to downsize the reaction volume even further. For instance, in this study, individual reaction components were transferred 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 reagents were not properly added to the reaction pool due to slight misalignment of the acoustic dispenser. In this case, it would be advantageous to premix as many reaction components as possible and then transfer 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 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 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.

qPCRTag assay of wild type (BY4741) and *SynII*

One of the important design feature of Sc2.0 is the introduction of PCRTags, which are short, synonymously recoded sequences within the ORFs that enable differentiation of synthetic chromosomes from their wild type counterparts. The standard readout of PCRTag assay is to identify the presence or absence of PCRTag amplicons by agarose gel electrophoresis. However, there will be over 8000 PCRTags in the final complete Sc2.0 genome, a high throughput and easy handling approach is necessary. The method of Mitchell, et al. has improved PCRTag genotyping by developing a real time PCR-based detection assay (qPCRTag assay) [82]. Echo was used to distribute qPCR mastermix, template DNA and PCRTag primers into each well of a 1536 multiwell plate. Then a qPCR thermal cycler was used to miniaturize reactions at 500nL scale and maximize throughput. And both steps can be further automated. Here in this study, this improved method was tested and applied using 768 pairs of wild type and synthetic chromosome II PCRTag primers. Primers were arrayed identically in each quadrant of the multiwall plate for easy visual comparison.

From our result, for the most part, amplification was as expected, whereby synthetic primers exclusively amplified synthetic DNA and *vice versa* (Figure 2.6). However, we also observed several deviations from the expected pattern, suggesting false negatives and false positives. In all false negative results observed, 10 of the 11 are known to fail due to unspecific amplification or primer dimers (Figure 2.7). The remaining one is showing undetermined feedback, which could arise from a lack of transfer of mastermix or gDNA template. Overall, our result shows a relative low false positive and false negative rate, and careful parameter optimization will help to further eliminate some of the noise.

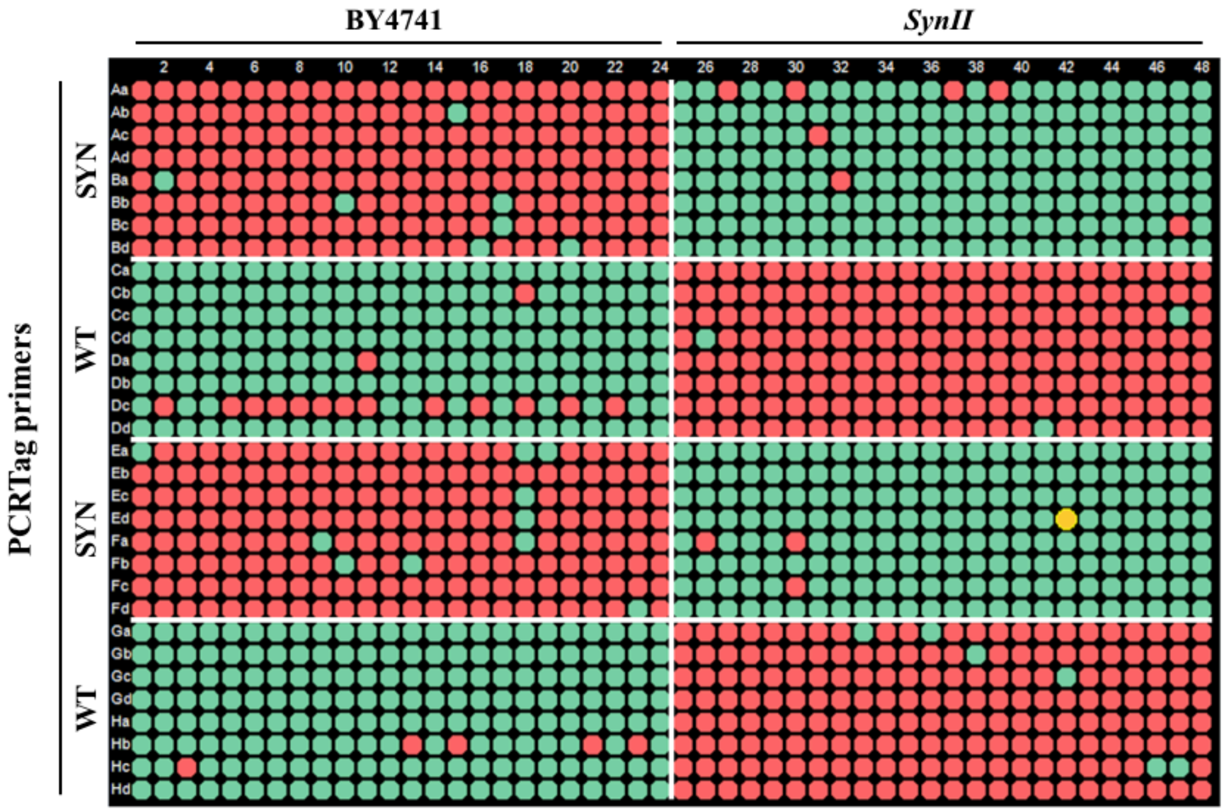


Figure 2.6 Plate heat map displaying presence/absence call for *SynII* qPCRTag analysis

Two different types of genomic DNA were subjected to PCRTag analysis using synthetic (SYN) and wild type (WT) chromosome II PCRTag primers. BY4741 and *SynII* encode wild type and synthetic chromosome II, yielding amplification with WT PCRTag primers and SYN PCRTag primers respectively (Green: positive; Red: negative; Yellow: undetermined, signal did not resemble a sigmoidal amplification curve).

2.3 Building 770-kilobase *SynII*

The sequence of native *Saccharomyces cerevisiae* chromosome *II* was determined over two decades ago. This native chromosome is 807,888 bp in length and includes 410 open reading frames, 13 tRNAs, and 30 introns[83]. *SynII* was designed based on the native chromosome *II* following previously reported rules introduced in Chapter One, resulting in a "designer" chromosome 770,035 bp in length, 43,149 bp shorter than the native sequence. *SynII* was initially synthesized as minichunks (~3 kb), assembled into chunks (~10 kb) and integrated into the yeast genome to replace the native chromosome. The construction was done previously by the team from BGI. The resulting strain was identified by the PCRTag method. PCRTags are pairs of ~20bp synonymously recoded segments of the coding region of ORFs. Most ORFs are incorporated with PCRTags, with one PCRTag per ORF on average. Therefore, after each run of synthetic DNA integration, pairs of wild type and synthetic PCRTags in that corresponding region can be used to quickly scan to ascertain whether a complete substitution has occurred. Here, PCRTag analysis confirmed the complete construction of *SynII* (Figure 2.7, Figure 2.8). Although PCRTags can help easily identify the presence of synthetic segment, they cannot distinguish any structure variations such as inversion or duplication. Therefore, whole genome sequencing using the Ion PGM™ sequencing platform (Life Tech) was performed to further verify *SynII* strain.

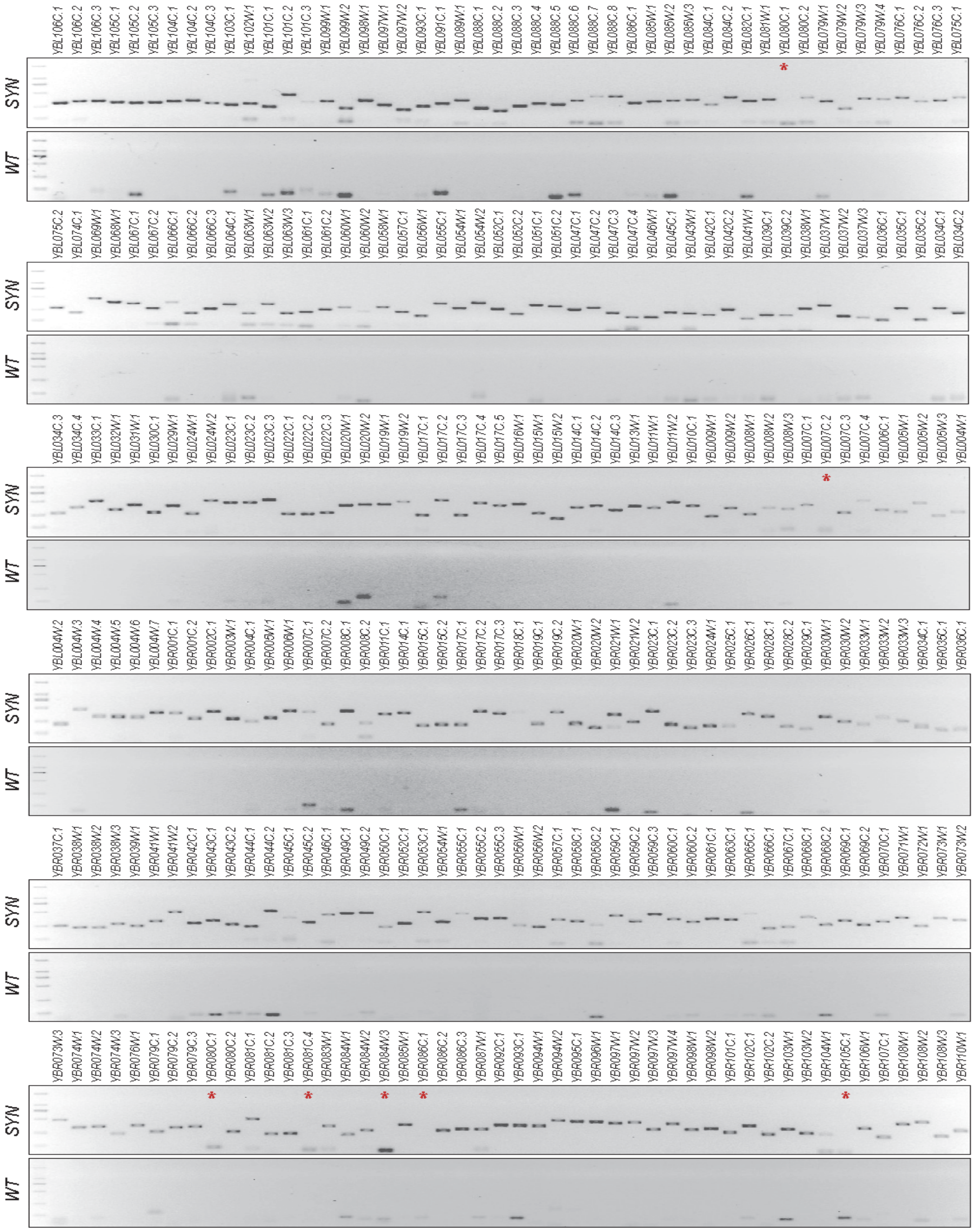


Figure 2.7 *SynII* full PCRTag analyses

The presence of synthetic PCRTags and absence of wild-type PCRTags indicate the successful replacement of native sequence by synthetic sequence. Red asterisk: amplification failure of synthetic PCRTags, which might be caused by PCRTag design failure or PCR failure. Blue asterisk: PCRTags that can amplify both synthetic and wild type genome, which were found to be homologous regions on other chromosomes.

[Figure 2.7 *SynII* full PCRTag analyses - Continued]

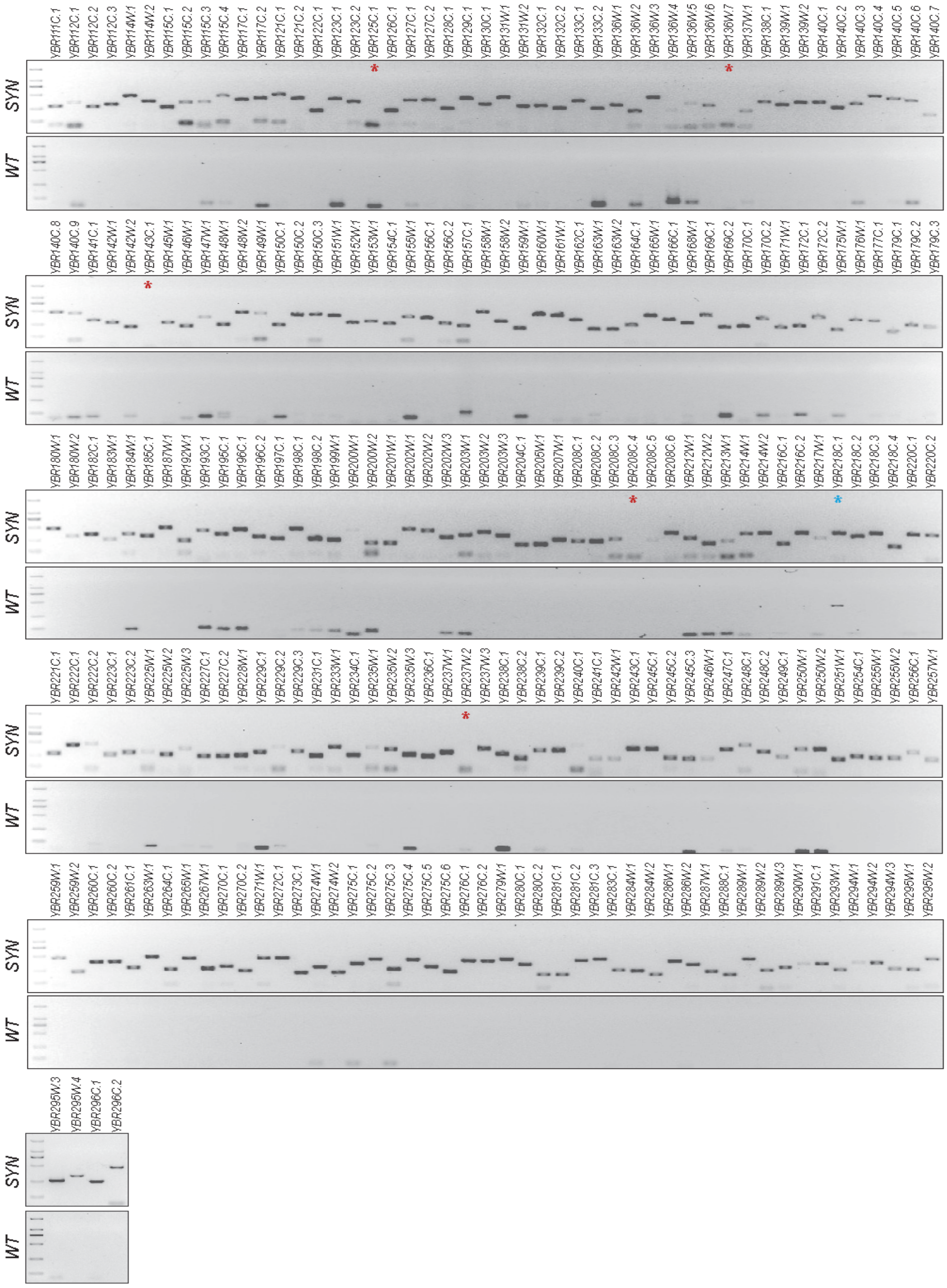


Figure 2.7 *SynII* full PCRTag analyses

The presence of synthetic PCRTags and absence of wild-type PCRTags indicate the successful replacement of native sequence by synthetic sequence. Red asterisk: amplification failure of synthetic PCRTags, which might be caused by PCRTag design failure or PCR failure. Blue asterisk: PCRTags that can amplify both synthetic and wild type genome, which were found to be homologous regions on other chromosomes.

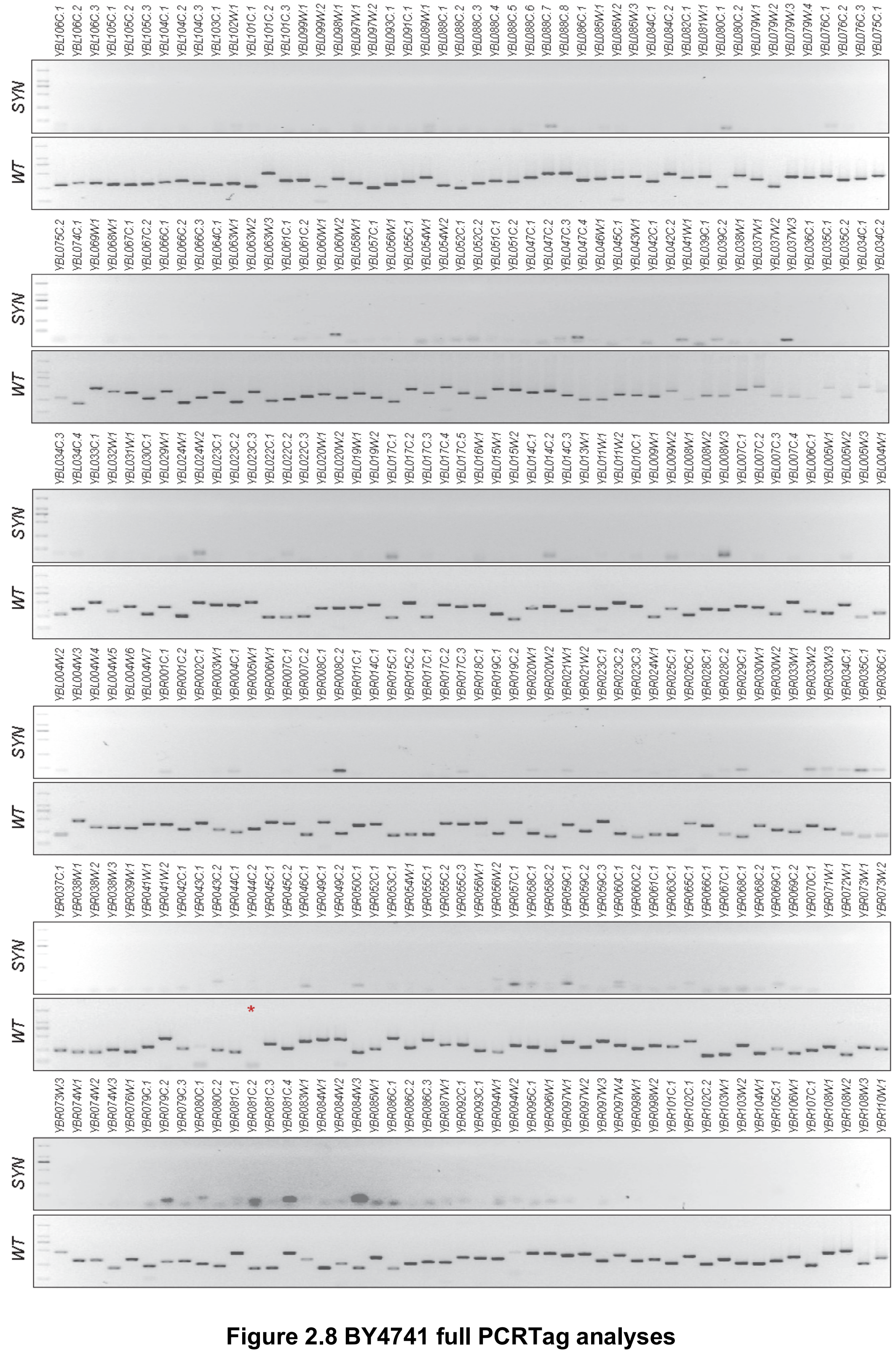


Figure 2.8 BY4741 full PCRTag analyses

The presence of only wild type PCRTags amplicons means high specificity of both synthetic and wild type PCRtag primers. Red asterisk: failure amplification of wild-type PCRTag may be caused by PCR failure.

[Figure 2.8 BY4741 full PCRTag analyses - Continued]

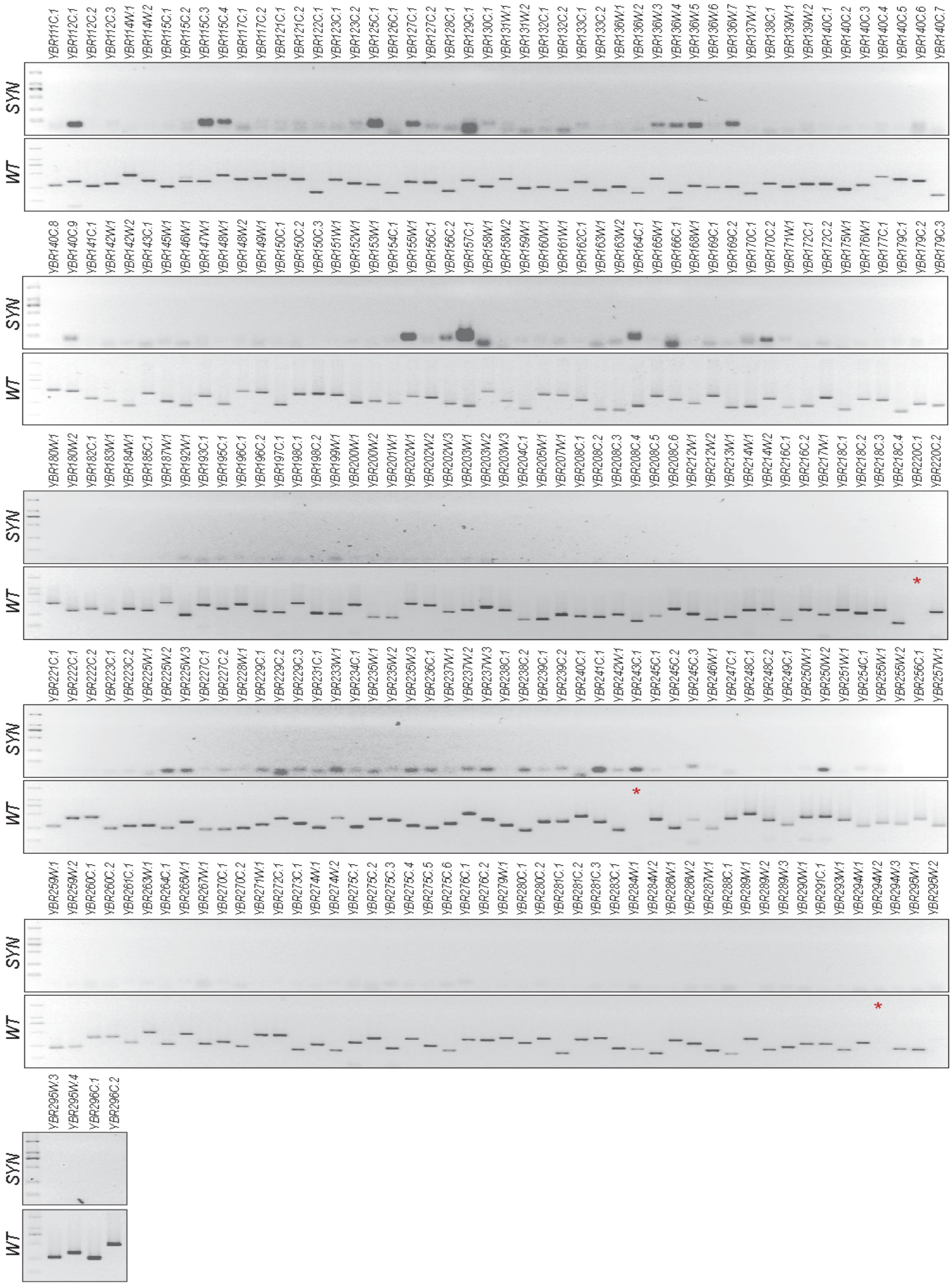


Figure 2.8 BY4741 full PCRTag analyses

The presence of only wild type PCRTags amplicons means high specificity of both synthetic and wild type PCRtag primers. Red asterisk: failure amplification of wild-type PCRTag may be caused by PCR failure.

2.3.1 Whole genome sequencing revealed structural variations in *SynII*

In total, 455M bp clean ~400bp reads, with 37.6-fold sequencing depth of the genome were generated (Figure 2.9). Clean reads were mapped to BY4741 and BY4742 yeast reference sequences (the wildtype sequence of chromosome II is replaced by synthetic chromosome II sequence) using bowtie2-2.0.0[84] with standard settings. For each alignment result, local realignment was performed with GATK 2.7-2[85] RealignerTargetCreator and IndelRealigner tools to clean up mapping artifacts generated during reads mapping on the edges of indels. The resulting files with BAM format were then prepared for variation calling.

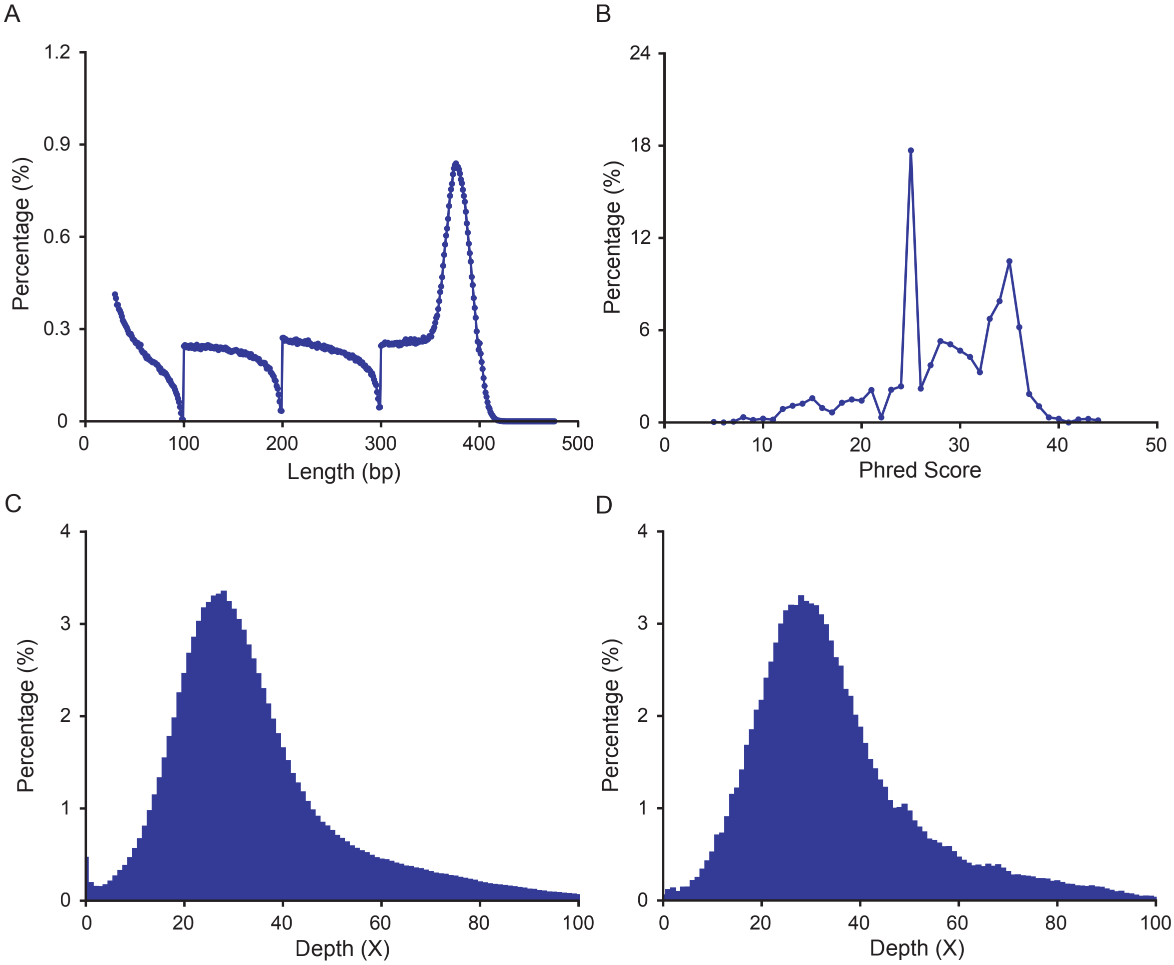


Figure 2.9 Statistics of *SynII* Ion PGM™ sequencing reads

A. Length distribution of filtered read generated from Ion PGM™ Sequencing platform. B. Quality distribution of single base. C. Mapping depth distribution to whole genome. D. Mapping depth distribution to *SynII*.

Compared with the designed sequence, 61 variations of 4 types were observed by deep sequencing: 50 single nucleotide variations (SNVs), 5 missing loxPsym sites, 4 deletions and 2 structural variations (SVs) (Table 2.7). 28 of the 50 SNVs were found to correspond exactly to the genotype of the native chromosome sequence at these positions, suggesting that these represent residual homologous recombination "patchworks" as seen in *SynIII* and other synthetic chromosomes[16,17,19-22]. When such patchwork regions are short (i.e. lying entirely between two sets of PCRTags), they can be missed by the PCRTag analysis. For the remaining 22 SNVs, 3 were found to pre-exist in synthetic minichunk DNA indicating that these mutations were introduced during synthesis. The remaining SNVs map to overlapping regions between minichunks or megachunks, suggesting that these SNVs were likely introduced during minichunk assembly or megachunk integration. Since none of these SNVs were in coding regions or noticeably altered phenotype, they were not corrected.

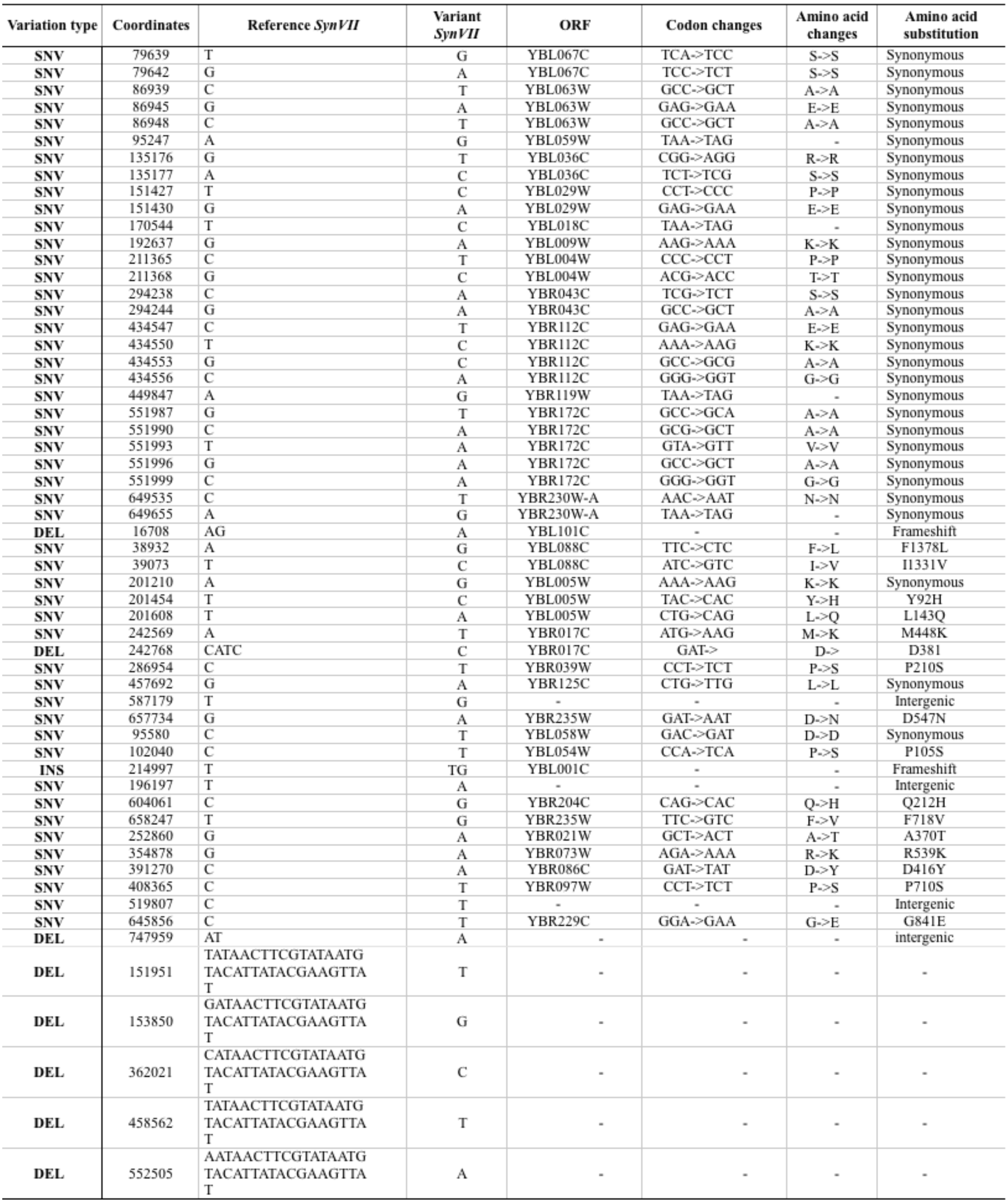
As previously seen in other larger synthetic chromosomes (*SynV*, *SynX*) [19,21], two complex structural variants (SVs) were observed: one was a ~15 kb tandem-duplication in megachunk *L* with chunk *L2-L4* duplicated and a loxPsym site located between the duplications (Figure 2.10A,B); the other SV, identified in megachunk *T*, was fully characterized by PGM sequencing (Life Technologies), and is a >30 kb complex DNA sequence including multiple copies of chunks *T4*, *T5* and a partial chunk backbone plasmid *pSBGAK* (Figure2.10C,D). We hypothesize that the 34 bp loxPsym sequence can serve as a homologous region during homologous recombination-mediated integration, albeit at a very low frequency, which led to the formation of the first SV. The mechanism of formation of the second SV remains unknown.

2.3.2 Correction of structural variants in *SynII*

We designed a straightforward strategy to repair the two structural variations, again by applying the I-*Sce*I system[86]. The 18bp I-*Sce*I recognition sequence was designed to carry a selective marker (*URA3*) and overlap with both ends of the tandem repeat junction as a donor fragment (Figure 2.11A). Upon induction of I-*Sce*I digestion, the *SynII* chromosome was broken and thereafter repaired through homology-directed recombination between the repeat sequences, and as a consequence duplicated region was effectively looped out.

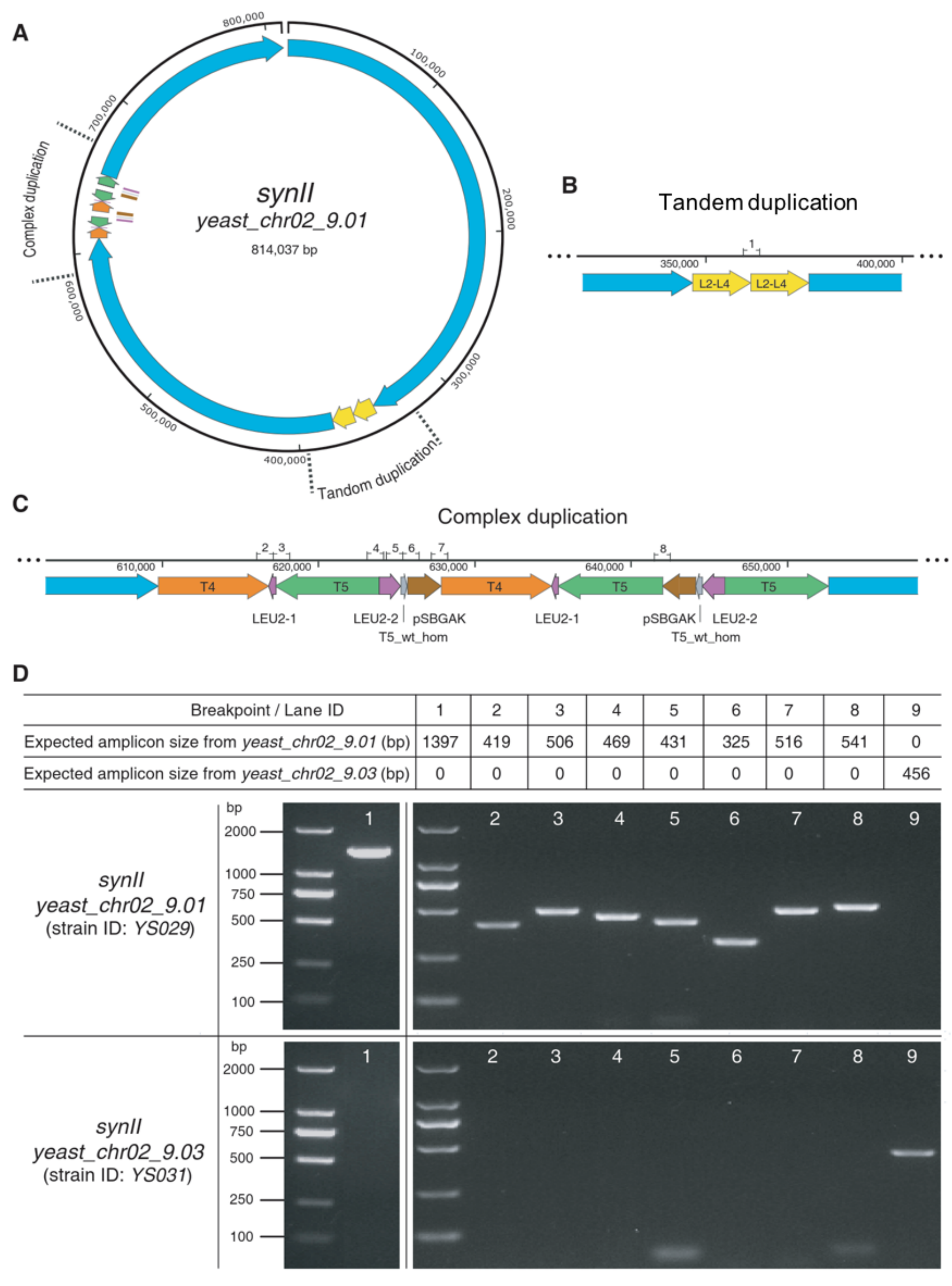
Using this strategy, the two SVs in synthetic chromosome *II* were sequentially repaired, yielding strains yeast_*chr02_9.02* and yeast_*chr02_9.03* respectively (Figure 2.11B). PCR analysis (Figure 2.10D) and deep sequencing (Figure 2.11C) validated the successful repair of the duplicated regions in *SynII.* Duplications were also observed in other synthetic chromosomes (*SynV, SynX, SynXII*). Thus the I-*Sce*I method could serve as an efficient strategy for repair of large duplications during synthetic chromosome construction.

In addition to the SVs, pulsed field gel analysis (PFGE) revealed an abnormal karyotype of native *chrXIII* and *chrXVI* (Figure 2.12). This was further confirmed by DNA sequencing and Hi-C analysis[87]. As similar instances have also been found during the construction of *SynIII*[17], we surmise that this represents a spontaneous low-frequency event occurred at some point during assembly. PFGE shows that the *chrXIII* and *chrXVI* karyotype of the *SynII* strain with SVs (yeast_*chr02_9.01*) is correct. We repaired this chromosomal crossover by first switching the mating type of the repaired *SynII* (yeast_*chr02_9.03*) and then back crossing to *SynII strain* (yeast_*chr02_9.01*) to avoid crossover between synthetic chromosome *II* and native chromosome *II.* After tetrad dissection, the correct karyotype was verified by PFGE (Figure 2.12).



Variation type	Coordinates	Reference *SynVII*	Variant *SynVII*	ORF	Codon changes	Amino acid changes	Amino acid substitution
SNV	79639	T	G	YBL067C	TCA->TCC	S->S	Synonymous
SNV	79642	G	A	YBL067C	TCC->TCT	S->S	Synonymous
SNV	86939	C	T	YBL063W	GCC->GCT	A->A	Synonymous
SNV	86945	G	A	YBL063W	GAG->GAA	E->E	Synonymous
SNV	86948	C	T	YBL063W	GCC->GCT	A->A	Synonymous
SNV	95247	A	G	YBL059W	TAA->TAG	-	Synonymous
SNV	135176	G	T	YBL036C	CGG->AGG	R->R	Synonymous
SNV	135177	A	C	YBL036C	TCT->TCG	S->S	Synonymous
SNV	151427	T	C	YBL029W	CCT->CCC	P->P	Synonymous
SNV	151430	G	A	YBL029W	GAG->GAA	E->E	Synonymous
SNV	170544	T	C	YBL018C	TAA->TAG	-	Synonymous
SNV	192637	G	A	YBL009W	AAG->AAA	K->K	Synonymous
SNV	211365	C	T	YBL004W	CCC->CCT	P->P	Synonymous
SNV	211368	G	C	YBL004W	ACG->ACC	T->T	Synonymous
SNV	294238	C	A	YBR043C	TCG->TCT	S->S	Synonymous
SNV	294244	G	A	YBR043C	GCC->GCT	A->A	Synonymous
SNV	434547	C	T	YBR112C	GAG->GAA	E->E	Synonymous
SNV	434550	T	C	YBR112C	AAA->AAG	K->K	Synonymous
SNV	434553	G	C	YBR112C	GCC->GCG	A->A	Synonymous
SNV	434556	C	A	YBR112C	GGG->GGT	G->G	Synonymous
SNV	449847	A	G	YBR119W	TAA->TAG	-	Synonymous
SNV	551987	G	T	YBR172C	GCC->GCA	A->A	Synonymous
SNV	551990	C	A	YBR172C	GCG->GCT	A->A	Synonymous
SNV	551993	T	A	YBR172C	GTA->GTT	V->V	Synonymous
SNV	551996	G	A	YBR172C	GCC->GCT	A->A	Synonymous
SNV	551999	C	A	YBR172C	GGG->GGT	G->G	Synonymous
SNV	649535	C	T	YBR230W-A	AAC->AAT	N->N	Synonymous
SNV	649655	A	G	YBR230W-A	TAA->TAG	-	Synonymous
DEL	16708	AG	A	YBL101C	-	-	Frameshift
SNV	38932	A	G	YBL088C	TTC->CTC	F->L	F1378L
SNV	39073	T	C	YBL088C	ATC->GTC	I->V	I1331V
SNV	201210	A	G	YBL005W	AAA->AAG	K->K	Synonymous
SNV	201454	T	C	YBL005W	TAC->CAC	Y->H	Y92H
SNV	201608	T	A	YBL005W	CTG->CAG	L->Q	L143Q
SNV	242569	A	T	YBR017C	ATG->AAG	M->K	M448K
DEL	242768	CATC	C	YBR017C	GAT->	D->	D381
SNV	286954	C	T	YBR039W	CCT->TCT	P->S	P210S
SNV	457692	G	A	YBR125C	CTG->TTG	L->L	Synonymous
SNV	587179	T	G	-	-	-	Intergenic
SNV	657734	G	A	YBR235W	GAT->AAT	D->N	D547N
SNV	95580	C	T	YBL058W	GAC->GAT	D->D	Synonymous
SNV	102040	C	T	YBL054W	CCA->TCA	P->S	P105S
INS	214997	T	TG	YBL001C	-	-	Frameshift
SNV	196197	T	A	-	-	-	Intergenic
SNV	604061	C	G	YBR204C	CAG->CAC	Q->H	Q212H
SNV	658247	T	G	YBR235W	TTC->GTC	F->V	F718V
SNV	252860	G	A	YBR021W	GCT->ACT	A->T	A370T
SNV	354878	G	A	YBR073W	AGA->AAA	R->K	R539K
SNV	391270	C	A	YBR086C	GAT->TAT	D->Y	D416Y
SNV	408365	C	T	YBR097W	CCT->TCT	P->S	P710S
SNV	519807	C	T	-	-	-	Intergenic
SNV	645856	C	T	YBR229C	GGA->GAA	G->E	G841E
DEL	747959	AT	A	-	-	-	intergenic
DEL	151951	TATAACTTCGTATAATGTACATTATACGAAGTTAT	T	-	-	-	-
DEL	153850	GATAACTTCGTATAATGTACATTATACGAAGTTAT	G	-	-	-	-
DEL	362021	CATAACTTCGTATAATGTACATTATACGAAGTTAT	C	-	-	-	-
DEL	458562	TATAACTTCGTATAATGTACATTATACGAAGTTAT	T	-	-	-	-
DEL	552505	AATAACTTCGTATAATGTACATTATACGAAGTTAT	A	-	-	-	-

Table 2.7 Variants in *SynII* strain revealed by sequencing data



Breakpoint / Lane ID	1	2	3	4	5	6	7	8	9
Expected amplicon size from *yeast_chr02_9.01* (bp)	1397	419	506	469	431	325	516	541	0
Expected amplicon size from *yeast_chr02_9.03* (bp)	0	0	0	0	0	0	0	0	456

Figure 2.10 *SynII* structure variation identification and repair verification

A. Full sequence map of *SynII*. B. Putative tandem duplication at megachunk *L*. Duplicated regions are depicted in yellow. C. Complex duplication at megachunk *T*. *pSBGAK* is the clone vector of chunk *T5*; *T5*_wt_hom are the wild-type homologous sequence on chunk *T5*. *LEU2*-1 (398 bp) and *LEU2*-2 (1399 bp) are upstream and downstream sequences of *LEU2*, respectively. Numbers are the breakpoints identified by sequencing. D. PCR at each identified breakpoint in the *SynII* strain was performed before and after structure variation repairing to identify the exclusion of duplicated regions. Sizes of amplicon products at each breakpoint are shown.

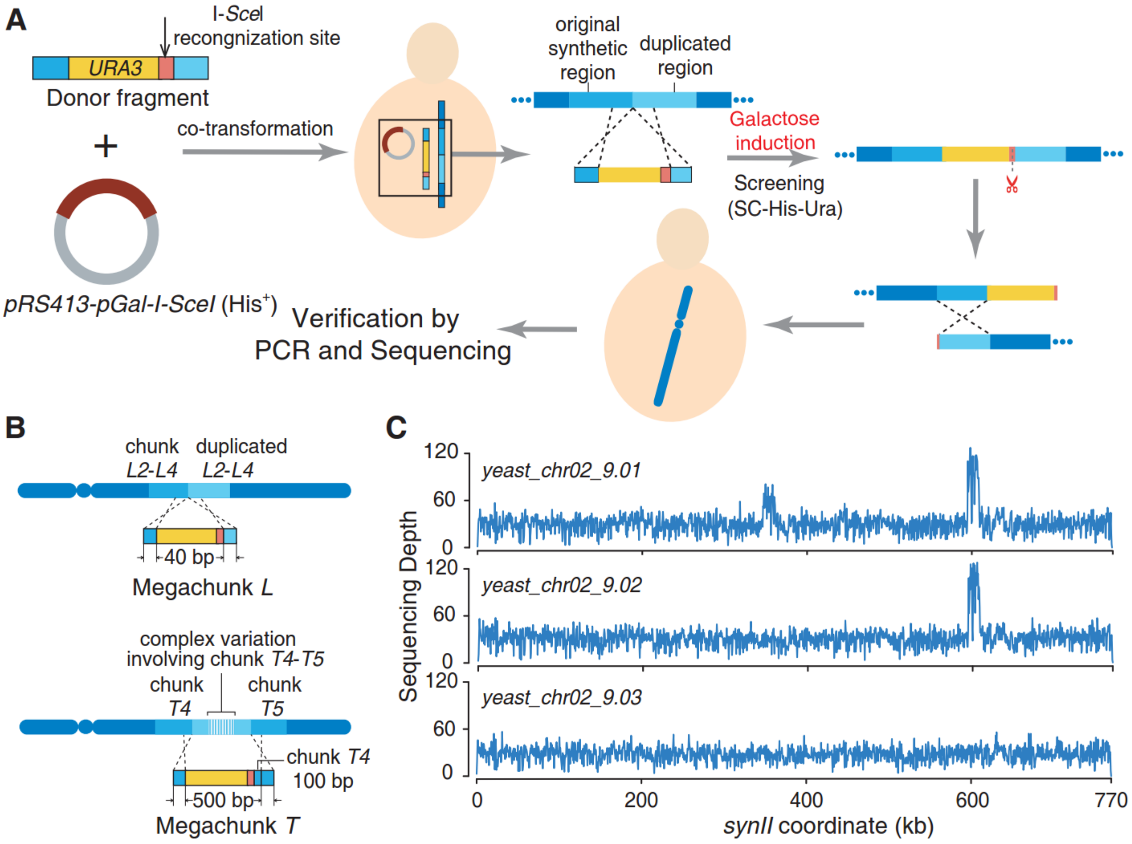


Figure 2.11 Structure variation repair through chromosome breakage

A. I-*Sce*I mediated repair strategy for *SynII* structure variations. The donor fragment was designed to carry a *URA3* cassette (yellow) and I-*Sce*I recognition site (organge), with both ends overlapping the structural variation sequences observed in *SynII* (yeast_*chr02_9.01*). The donor fragment was integrated into *SynII* between the two tandem repeats through homologous recombination, then an episomal plasmid *pRS413-pGal1-I-SceI* was transformed into the cell. A double strand break at the I-*Sce*I site was induced in galactose medium, and the homologous recombination of two partial chromosomes of *SynII* eliminated the duplication. B. Structure variations in megachunks *L* and *T* and their corresponding donor sequence design. In megachunk *L*, a copy of chunk *L2-L4* was observed following the original *L2-L4* sequence that generated a tandem-duplication. The donor fragment was inserted directly between two duplications. In megachunk *T*, the donor fragment was inserted to remove the complex variation with multiple copies of chunks *T4*, *T5* and a partial chunk backbone plasmid *pSBGAK* sequence. C. Deep sequencing read depth analysis revealed the successful sequential removal of duplications. The starting synthetic chromosome *SynII* (yeast_*chr02_9.01*) has both duplication regions, and after the first round of repair at megachunk *L*, the resulting chromosome *SynII* (yeast_*chr02_9.02*) shows only one duplication. The finished chromosome *SynII* (yeast_*chr02_9.03*) was obtained after the second repair at the megachunk *T* region.

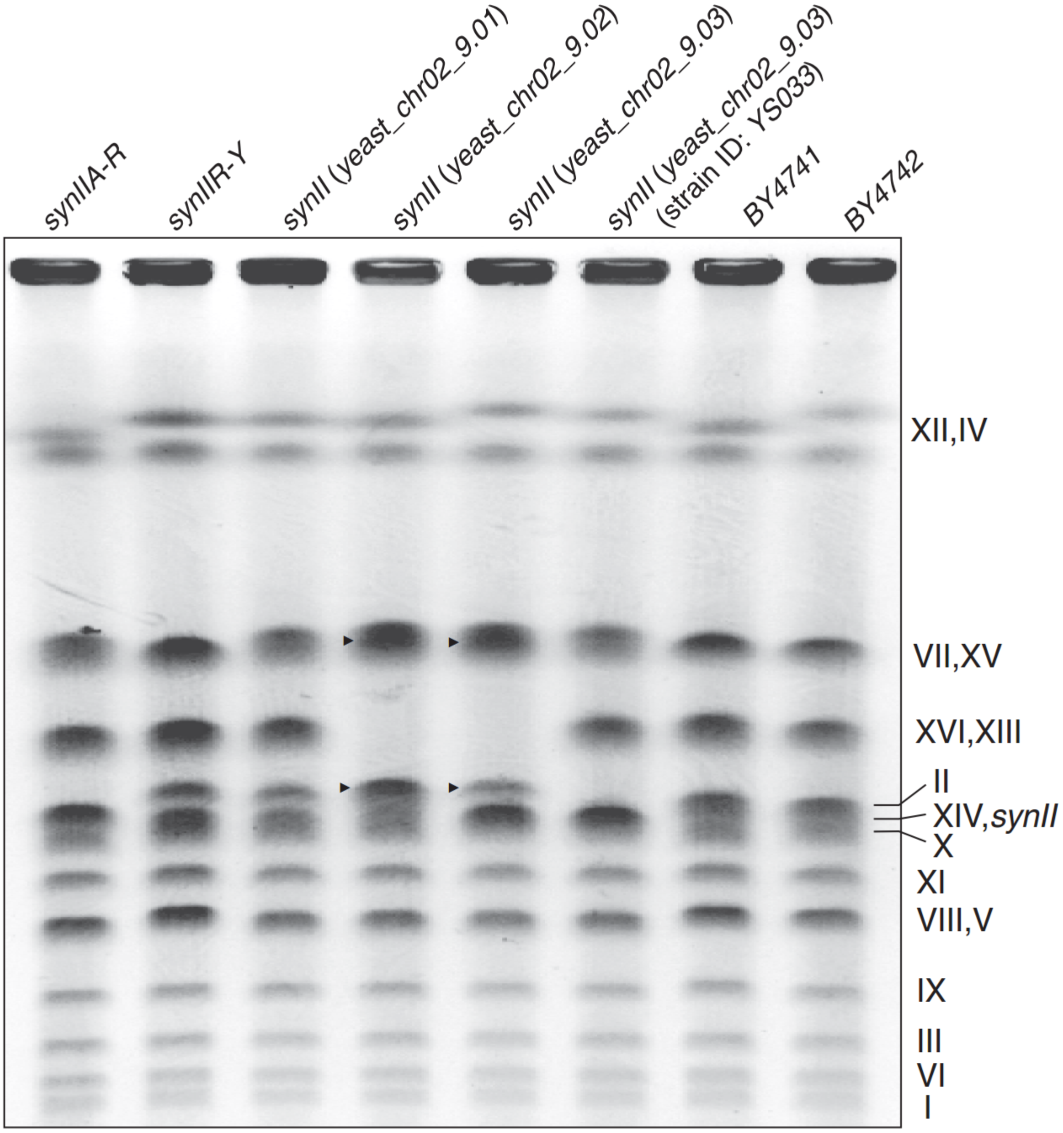


Figure 2.12 Karyotype analysis of *SynIIA-R*, *SynIIR-Y*, *SynII* with duplications, repaired *SynII* strains by PFGE

SynIIA-R (strain ID: *YS018*) and *SynIIR-Y* (strain ID: *YS026*) are the two semi-synthetic *SynII* strains used for constructing the full synthetic *SynII*, named *SynII* (yeast*_chr02_9.01*, strain ID: *YS029*). *SynII* with one of the two duplications repaired is named as *SynII* (yeast*_chr02_9.02*, strain ID: *YS030*), followed by the final repaired *SynII* (yeast*_chr02_9.03*, strain ID: *YS031*) and strain with complete *SynII* and rearrangement between chromosome 13 and 16 repaired *SynII* (yeast*_chr02_9.03,* strain ID: *YS033*). Chromosome numbers are labeled on the right side. The final *SynII* strain shows no karyotype abnormality compared to the *BY4741* and *BY4742* strains. Triangle in the figure shows abnormal karyotype comparing to native strains.

At this point, the full construction of *SynII* was completed and its genotype was carefully verified.

2.4 Building 1015-kilobase *SynVII*

Among the total 16 chromosomes, chromosome *VII* is one of the four chromosomes larger than 1Mb. It has 572 predicted ORFs, of which 341 are uncharacterized[88]. Of the ORFs, 17% show high similarity to human genes, compared to the 31% similarity to humans on whole genome scale[71]. Almost half of the ORFs could be classified into functional categories, while the number on the whole genome scale is 43%[71,89]. The number of tRNAs located on chromosome *VII* is 36, which is the highest among all chromosomes and the interplay between the synthetic chromosome *VII* and the tRNA neochromosome (part of Sc2.0 project) will be very interesting to study as well.

According to the three design principles (1. do no harm, 2. maintain genomic stability and 3. increase genetic flexibility), several changes have been made on chromosome *VII*: Telomeres are replaced by artificial telomeres of the simple sequence repeat ($(TG_{1\text{-}3})n$). 14 introns and 60 retrotransposons are deleted. All 36 tDNAs on chromosome *VII* are removed and will be placed on the “neochromosome”. 706 pairs of PCR tags are added in synthetic chromosome *VII* to distinguish it from natural chromosome. 316 loxPsym sites are added in the 3’UTR region of each non-essential ORF in chromosome *VII* for SCRaMbLE. For the total of 127 TAG stop codons on chromosome *VII*, 126 are swapped to TAA. One TAG was retained as the change might affect an overlapping ORF. The final synthetic version of chromosome *VII* will be 1015kb, 76kb shorter than the natural version (Figure 2.13).

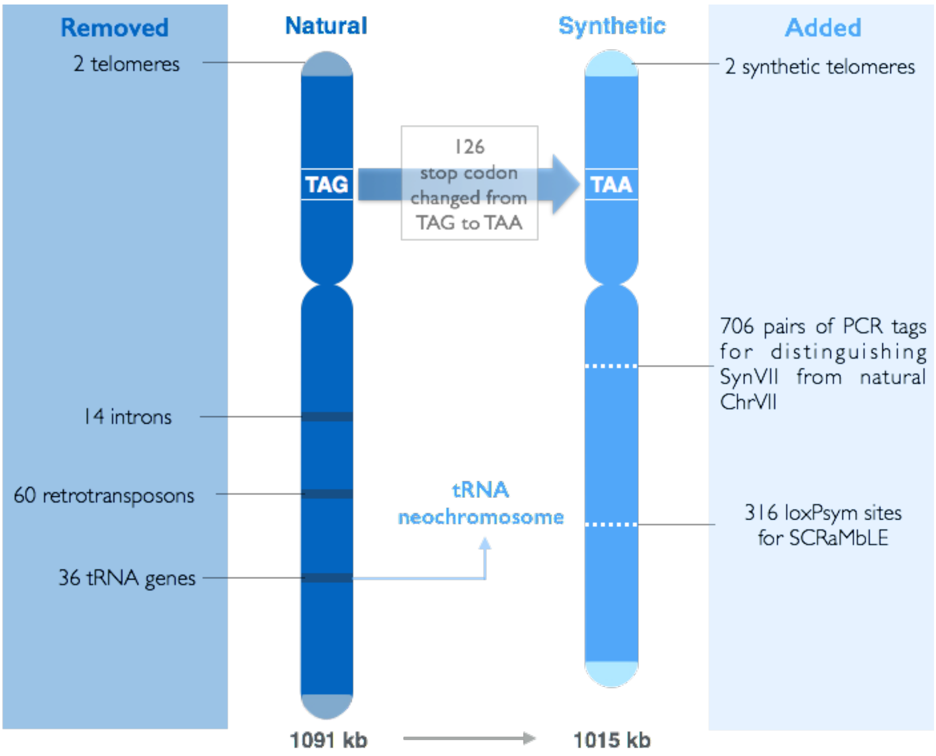


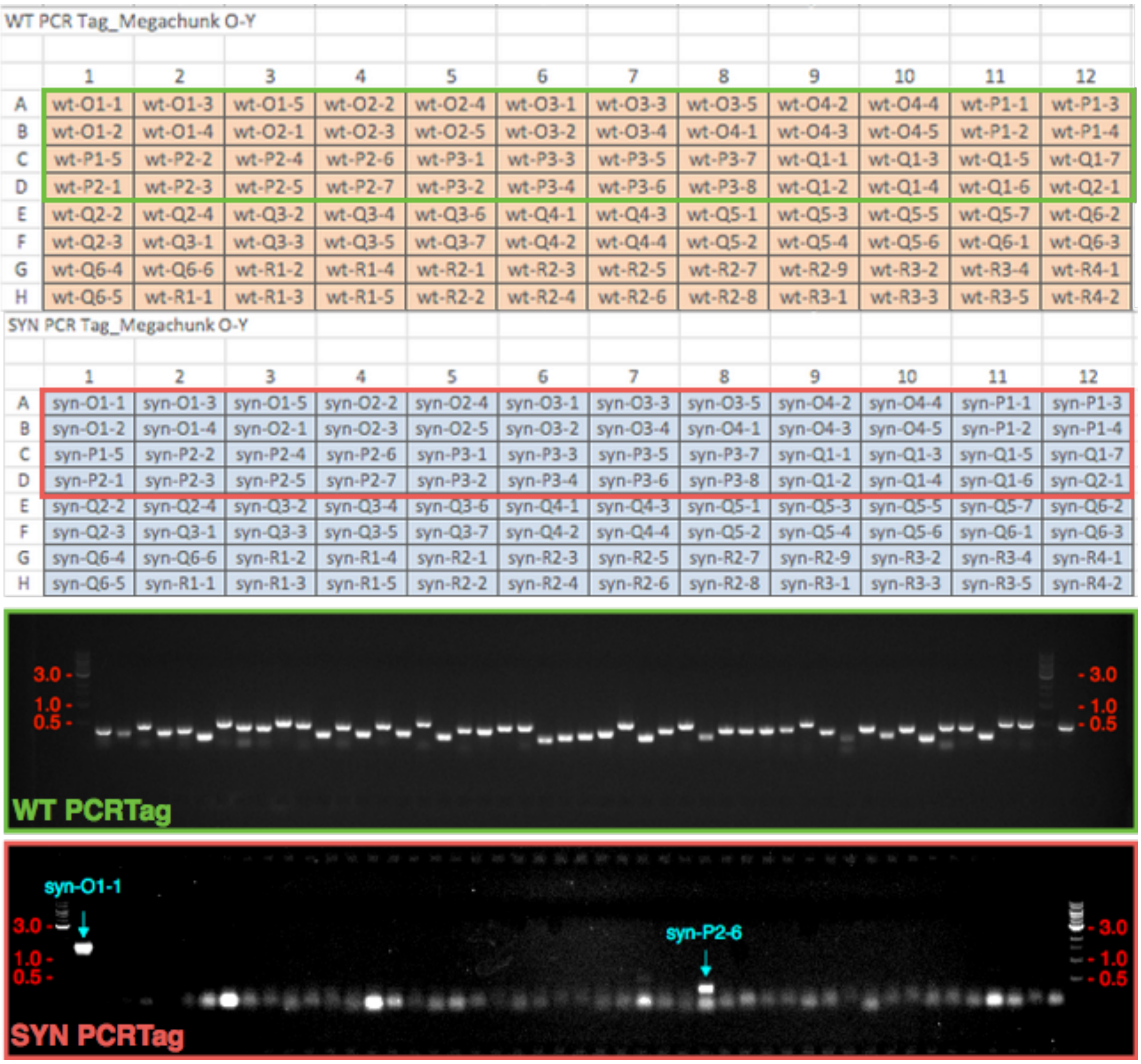
Figure 2.13 *SynVII* design

Native telomere sequence is replaced by synthetic telomere. 14 introns and 60 retrotransposons are excluded. 36 tRNA genes are relocated to tRNA neochromosome. 126 stop codons are swapped from TAG to TAA. 706 pairs of PCR tags are added into SynVII for distinguishing from natural chromosome *VII*. 316 loxPsym sites are designed evenly distributed in *SynVII* for SCRaMbLE.

The sequence of synthetic chromosome *VII* is computationally segregated by BGI customized software "Segman" to 25 50kb-megachunks, then to 129 10kb-chunks and to final 485 3kb-minichunks the synthesis of which was directly outsourced to a DNA synthesis company. In addition, chromosome *VII* was finished with efforts from both UK and China, under the collaboration between UoE and BGI. *SynVIIL* (carrying Megachunk A-N) was constructed by team of BGI, while *SynVIIR* (carrying Megachunk O-Y) was assembled in Edinburgh. Then *SynVIIL* and *SynVIIR* carrying I-*Sce*I sites were combined to generate the fully synthetic *SynVII*.

PCRTag assay

The flexibility to analyze and quantify the synthetic content of the genome is important throughout the expanded genome synthesis process. For SynVIIR, 288 WT PCRTag primer pairs and 288 SYN PCRTag primer pairs are used for PCRTag assay during synthetic megachunks integration. PCRTag assays of wild-type yeast (BY4741) gDNA showed the presence of only wild-type PCRTags, demonstrating the specificity of all wild-type PCRTag primers (Figure 2.14). However, 4 synthetic PCRTags gave a positive signal of wild type gDNA. Sequence of these primers were aligned to BY4741 genome reference with BLASTN, with 3' end mapping (>10bp and only 3bp mismatches at 3' end) been considered as possible binding sites. The alignment showed that these primers cannot align to other locations on the genome, indicating the design should be correct. Therefore, these primers were re-synthesized to exclude the possibility that amplification might be caused by mutations on these primers. Analysis confirms that the presence of these synthetic PCRTags is caused by mutations on primers introduced during synthesis (Figure 2.15A). For wild-type PCRTags, only one wild-type PCRTag primer pair, namely WT_chr07_T1_3_F and WT_T1_3_R, did not yield any product under the PCR conditions used for PCRTag assay. Re-design of this primer was performed by reducing the length from 28bp to 23bp without changing the binding site. Subsequent analysis shows that with new designed primers, WT_chr07_T1_3 PCRTag is present and can be used to rapidly verify the introduction of synthetic sequence and removal of native wild-type yeast chromosome *VII* sequence by PCR (Figure 2.15B).



WT PCR Tag_Megachunk O-Y

	1	2	3	4	5	6	7	8	9	10	11	12
A	wt-O1-1	wt-O1-3	wt-O1-5	wt-O2-2	wt-O2-4	wt-O3-1	wt-O3-3	wt-O3-5	wt-O4-2	wt-O4-4	wt-P1-1	wt-P1-3
B	wt-O1-2	wt-O1-4	wt-O2-1	wt-O2-3	wt-O2-5	wt-O3-2	wt-O3-4	wt-O4-1	wt-O4-3	wt-O4-5	wt-P1-2	wt-P1-4
C	wt-P1-5	wt-P2-2	wt-P2-4	wt-P2-6	wt-P3-1	wt-P3-3	wt-P3-5	wt-P3-7	wt-Q1-1	wt-Q1-3	wt-Q1-5	wt-Q1-7
D	wt-P2-1	wt-P2-3	wt-P2-5	wt-P2-7	wt-P3-2	wt-P3-4	wt-P3-6	wt-P3-8	wt-Q1-2	wt-Q1-4	wt-Q1-6	wt-Q2-1
E	wt-Q2-2	wt-Q2-4	wt-Q3-2	wt-Q3-4	wt-Q3-6	wt-Q4-1	wt-Q4-3	wt-Q5-1	wt-Q5-3	wt-Q5-5	wt-Q5-7	wt-Q6-2
F	wt-Q2-3	wt-Q3-1	wt-Q3-3	wt-Q3-5	wt-Q3-7	wt-Q4-2	wt-Q4-4	wt-Q5-2	wt-Q5-4	wt-Q5-6	wt-Q6-1	wt-Q6-3
G	wt-Q6-4	wt-Q6-6	wt-R1-2	wt-R1-4	wt-R2-1	wt-R2-3	wt-R2-5	wt-R2-7	wt-R2-9	wt-R3-2	wt-R3-4	wt-R4-1
H	wt-Q6-5	wt-R1-1	wt-R1-3	wt-R1-5	wt-R2-2	wt-R2-4	wt-R2-6	wt-R2-8	wt-R3-1	wt-R3-3	wt-R3-5	wt-R4-2

SYN PCR Tag_Megachunk O-Y

	1	2	3	4	5	6	7	8	9	10	11	12
A	syn-O1-1	syn-O1-3	syn-O1-5	syn-O2-2	syn-O2-4	syn-O3-1	syn-O3-3	syn-O3-5	syn-O4-2	syn-O4-4	syn-P1-1	syn-P1-3
B	syn-O1-2	syn-O1-4	syn-O2-1	syn-O2-3	syn-O2-5	syn-O3-2	syn-O3-4	syn-O4-1	syn-O4-3	syn-O4-5	syn-P1-2	syn-P1-4
C	syn-P1-5	syn-P2-2	syn-P2-4	syn-P2-6	syn-P3-1	syn-P3-3	syn-P3-5	syn-P3-7	syn-Q1-1	syn-Q1-3	syn-Q1-5	syn-Q1-7
D	syn-P2-1	syn-P2-3	syn-P2-5	syn-P2-7	syn-P3-2	syn-P3-4	syn-P3-6	syn-P3-8	syn-Q1-2	syn-Q1-4	syn-Q1-6	syn-Q2-1
E	syn-Q2-2	syn-Q2-4	syn-Q3-2	syn-Q3-4	syn-Q3-6	syn-Q4-1	syn-Q4-3	syn-Q5-1	syn-Q5-3	syn-Q5-5	syn-Q5-7	syn-Q6-2
F	syn-Q2-3	syn-Q3-1	syn-Q3-3	syn-Q3-5	syn-Q3-7	syn-Q4-2	syn-Q4-4	syn-Q5-2	syn-Q5-4	syn-Q5-6	syn-Q6-1	syn-Q6-3
G	syn-Q6-4	syn-Q6-6	syn-R1-2	syn-R1-4	syn-R2-1	syn-R2-3	syn-R2-5	syn-R2-7	syn-R2-9	syn-R3-2	syn-R3-4	syn-R4-1
H	syn-Q6-5	syn-R1-1	syn-R1-3	syn-R1-5	syn-R2-2	syn-R2-4	syn-R2-6	syn-R2-8	syn-R3-1	syn-R3-3	syn-R3-5	syn-R4-2

Figure 2.14 PCRTag assay of wild-type gDNA (BY4741)

Partial PCRTag assay of wild-type yeast (BY4741) gDNA showed the presence of WT PCRTags and absence of SYN PCRTags. Two SYN PCRTags, syn-O1-1 and syn-P2-6 are present, demonstrating it is either caused by the errors created by the gene synthesis company and lead to unspecific amplification.

SynVII synthesis and assembly

The full *SynVII* sequence is segregated into ~3kb minichunks or 10kb chunks for *de novo* DNA synthesis. The nomenclature of minichunk uses the following naming convention: minichunk is named with Chunk ID followed by “HoM”, “Ho” or “Fx”. “HoM” refers to the homologous region between megachunks; “Ho” means the homologous region between Chunks and assembly accepting vector, while “Fx” indicates the x fragment. Each minichunk is designed with a 40bp overlapping region on both sides to enable efficient chunk assembly via in vitro Gibson assembly strategy[5]. For chunk assembly, the strategy used for *SynIII* was to ligate each chunk first and then transform them into yeast for integration via homologous recombination[17]. However, for chromosome *VII*, an alternative strategy was applied: each chunk is designed to have an 800bp-1200bp homologous region, and the ligation step was skipped by directly transforming all 5-6 chunks (equal to 1 megachunk) into yeast for integration.

For megachunk O, Q, S, U, W, X and Y, chunk assembly from corresponding ~3kb minichunks is required. The old strategy includes three steps: first, digesting out the minichunk fragment from corresponding carrying vector; second, Gibson assembly and bacterial transformation; and third, verifying by miniprep and digestion. However, efficiency of each Gibson assembly varies for several reasons, such as the number of DNA fragments, the length and potential secondary structure of overhangs. And normally for a chunk assembly, miniprep and digestion of 10 to 20 colonies is required in a single run in order to find the correct construct. Here, the workflow is refined by adding one additional step of colony PCR to reduce the cost and unnecessary workload (Figure 2.15). Two sets of primers are designed for the accepting vector and chunk sequence. A successfully assembled chunk will have both amplicons with sizes of ~300-800bp that cover the junctions of accepting vector and chunk sequence. And usually 80% of assembled products showing both bands are found to be correct following verification by digestion. By doing so, more colonies can be screened for each chunk assembly (~24 colonies on average), while much less work of miniprep and digestion is needed (2-4 on average). This single added step increases the chance to find correct chunk assembly product and significantly reduces the cost of reagents and consumables. By applying the refined workflow, over 90% of chunks were successfully assembled with only one attempt.

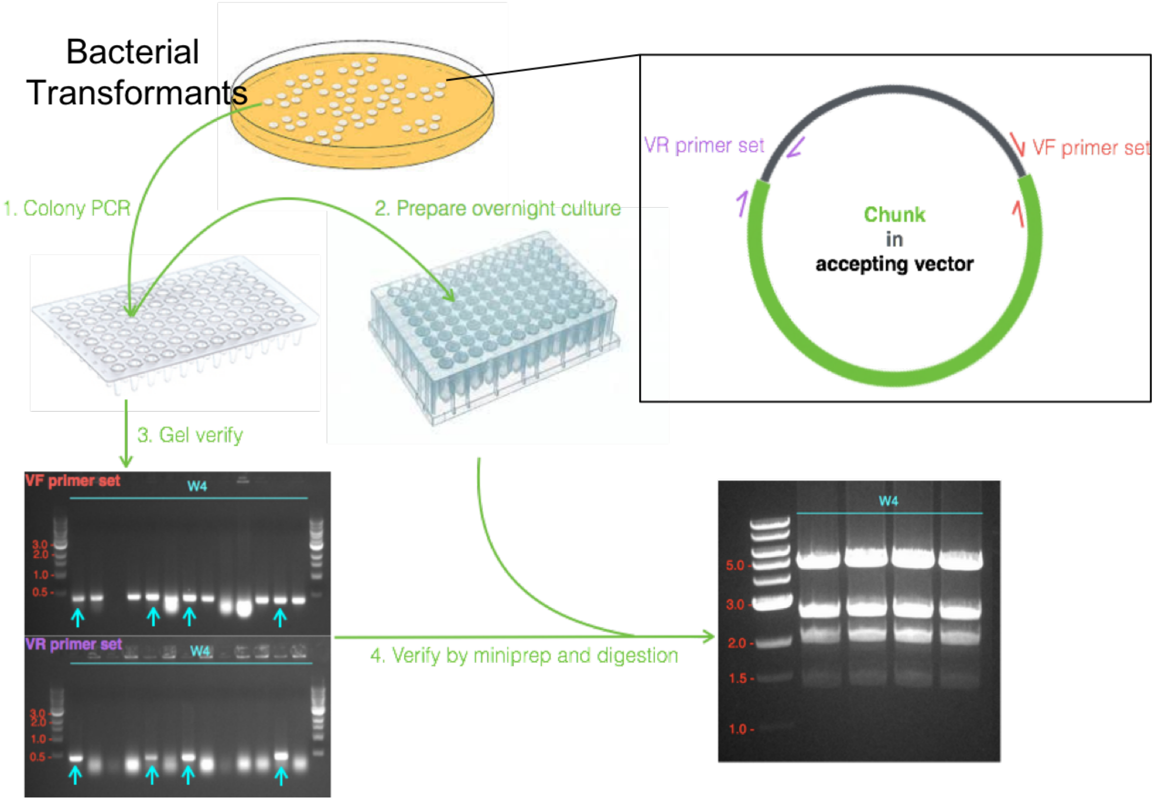


Figure 2.15 Refining chunk assembly workflow by adding a colony PCR step

Two sets of primer pairs, VF primer set and VR primer set, are designed to amplify junctions between accepting vector and chunk sequence and used for colony PCR. Colony is first dipped into PCR master mix, followed by dipping into corresponding well of a 96-deep well plate for overnight culture. After verification by gel electrophoresis, colonies with both positive bands are further verified by digestion.

Minichunk and Chunk fingerprinting

For minichunks/chunks that were outsourced to a DNA synthesis company (Life Tech), a major issue discovered so far was unwanted mixing of different minichunk samples. Take Y4Ho minichunk as an example, repeated failure of Y4 chunk Gibson assembly was experienced without a reasonable explanation. Once sent for sequencing, it was realized that the sequence was aligned to a region in megachunk A. After further investigation, the minichunk used for chunk Y4 Gibson assembly is actually A4Ho, which was from the same batch sent for synthesis with Y4Ho. Therefore the only plausible explanation was that the DNA synthesis company mixed up the samples in the same batch. And failure of discovering this issue earlier, in part, was due to the similar size after digesting out these minichunks from the accepting vector. Therefore, a QC process called “Plasmid Fingerprinting” is included to verify all minichunks/chunks outsourced for synthesis. Instead of using restriction enzymes designed to digest out minichunks/chunks, 1-2 restriction enzymes with multiple sites present in each minichunk/chunk plasmid were used to cut the plasmid into 3-8 pieces such that most, if not all, fragments are small enough to be accurately sized by again gel electrophoresis, generating a very unique pattern that will distinguish each minichunk/chunk plasmid.

A computer program was developed to automatically analyze the multiple cutter sites for each sample, which generates a spreadsheet containing information such as list of multiple cutter enzymes, number of fragments and corresponding sizes after digestion. In total 132 minichunks and 19 chunks, only two minichunks, namely Y2F3 and W2F3, failed to pass the fingerprinting QC. Sequencing of W2F3 minichunk again confirms that this sample is mixed up with other minichunks. Further investigation of the Y2F3 minichunk is discussed in the following "Problematic Minichunk troubleshooting" part. The DNA synthesis company has been contacted to further verify these two samples and requested to send correct products. In summary, "Plasmid Fingerprinting" is highly recommended as an efficient QC method, especially for high throughput DNA synthesis platform.

Problematic minichunk trougbleshooting

During the chunk Y2 assembly, an aberrant digestion result was noticed during minichunk digestion. According to gel simulation, the size of target Y2F3 minichunk should be 2639bp. However, the result from gel electrophoresis showed a band over 3kb (Figure 2.16A). A repeat digestion was performed and further confirmed that this was not due to experimental error. To further investigate, successive Primer walking was performed to verify the sequence of Y2F3 minichunk. Two primers that matched the beginning and the end of target Y2F3 minichunk sequence, namely Y2F3_sequencing_F and Y2F3_sequencing_R, were designed for first run of sequencing. Then each end of the sequenced strand was used as a primer (Y2F3_sequencing2_F and Y2F3_sequencing2_R) for the next run of sequencing. That way, the short part of Y2F3 minichunk that was sequenced kept "walking" along the target sequence. After 2 runs of primer walking, a 777bp sequence was confirmed to be inserted into Y2F3 minichunk, which was not consistent with the sequence design. The insert was further analysed by aligning its sequence to that of the plasmid and found it could be partially aligned to the plasmid backbone at bacterial ColE1 origin sequence (Figure 2.16B). These data showed that the Y2F3 minichunk construct received from DNA synthesis company was incorrect, therefore, resynthesis of this minichunk was performed.

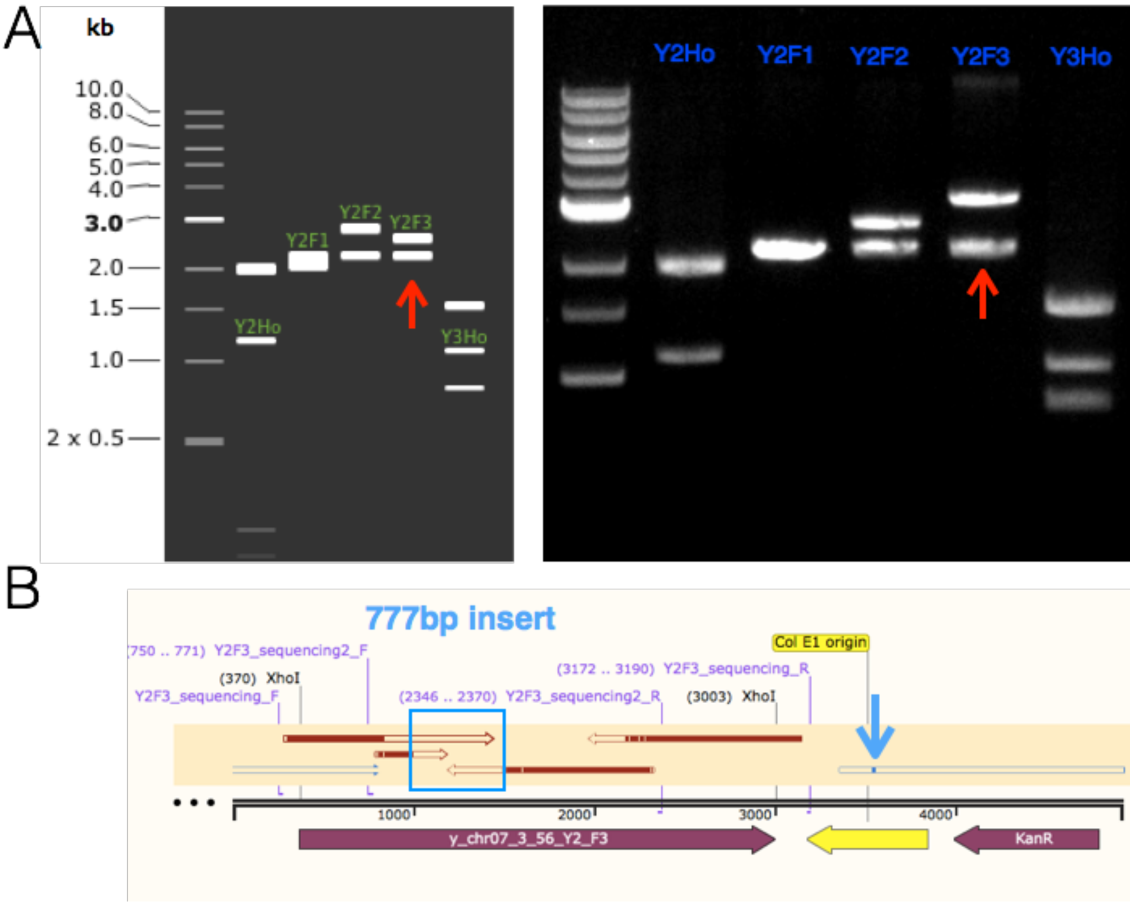


Figure 2.16 Digestion and sequencing verify of Y2F3 minichunk

A. Digestion result of the Y2F3 minichunk shows a band ~1kb larger than predicted (Arrow pointed). B. Two sequential runs of sequencing confirmed a 777bp insert (blue box) within target Y2F3 minichunk sequence, which can be partially aligned to the ColE1 origin sequence in the plasmid backbone.

The synthesis of minichunk Y1F1 (3433bp in size) failed repeatedly by the outsourced company. The best they could provide were two constructs carrying the full length of Y1F1 fragment, but each of them harbored a non-synonymous point mutation (T mutated to C at 775bp, Ser to Ile; G mutated to T at 1234bp, Met to Thr, repectively). The reference sequence contained in Saccharomyces Genome Database (SGD) shows that there are three ORFs in this minichunk: YGR275W, YGR276C and YGR277C (Table 2.8). Both point mutations were found in YGR276C coding region, indicating that this gene might be toxic to *E. coli* during cloning, which may explain why it is hard to obtain a colony carrying 100% correct Y1F1 fragment sequence. Therefore, an alternative strategy was designed: Y1F1 was split into two pieces Y1F1-A & Y1F1-B, each of them contained a ~500bp overlap region. Y1F1-A was assembled with upstream minichunks to generate Y1-A intermediate chunk, and Y1F1-B was assembled with downstream minichunks to generate Y1-B intermediate chunk (Figure 2.17). Then Y1-A and Y1-B were combined with other chunks (Y2, Y3, Y4 and Y5) for the megachunk integration through homologous recombination.

Construct ID	Location of Mutation	Codon Change	Amino Acid Change
#1	1234	AGC > ATC	Ser > Ile
#2	775	ATG > ACG	Met > Thr

Table 2.8 Mutation information of two Y1F1 constructs

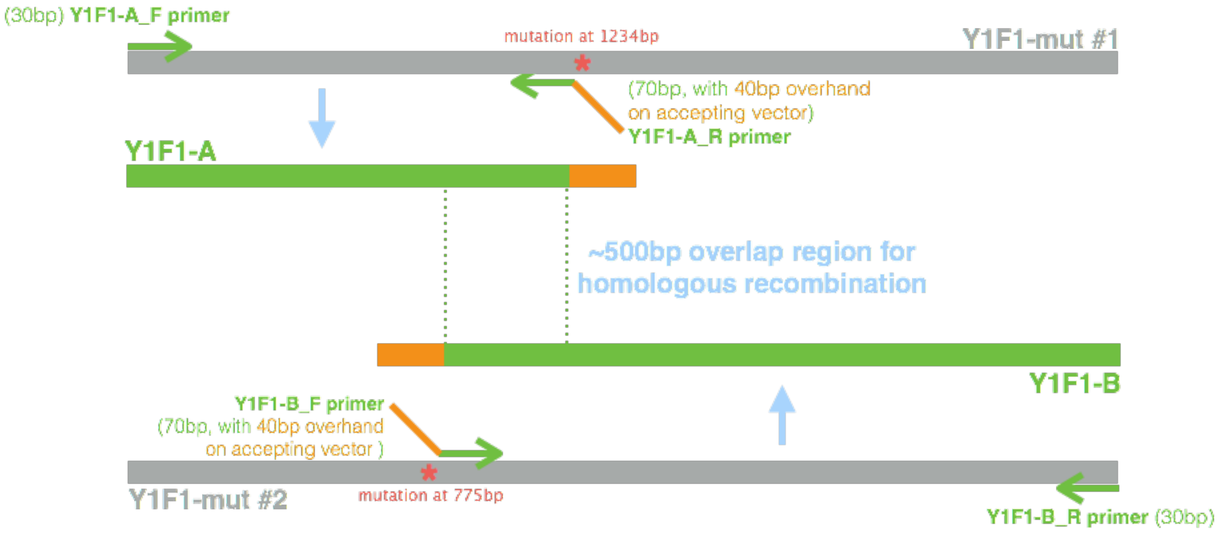

(30bp) Y1F1-A_F primer
mutation at 1234bp
Y1F1-mut #1
(70bp, with 40bp overhand on accepting vector)
Y1F1-A_R primer
Y1F1-A
~500bp overlap region for homologous recombination
Y1F1-B
Y1F1-B_F primer
(70bp, with 40bp overhand on accepting vector)
Y1F1-mut #2
mutation at 775bp
Y1F1-B_R primer (30bp)

Figure 2.17 Design of Y1F1-A and Y1F1-B construction

Primers are designed to amplify Y1F1-A and Y1F1-B separately. A 40bp overlapping sequence on the primers is included to facilitate the assembly of Y1F1-A and Y1F1-B separately with other minichunks. For Y1F1-A, the 35bp forward primer is designed to fully align to the template, while the 82bp reverse primer has 36bp sequence that can align to the template and 46bp tail that overlaps with the 5' end of linear vector; for Y1F1-B, the 86bp forward primer has 40bp sequence that can align to the template and 46bp tail that overlaps with the 3' end of linear vector, while the 42bp reverse primer can fully align to the template. Y1F1-A and Y1F1-B are expected to be obtained by PCR, with size at 1179bp and 2689bp respectively.

Three attempts were tried to construct Y1F1-A and Y1F1-B fragments with TOPO cloning and pJET cloning with several different pairs of primers. For Y1F1-A, results show that Y1F1-A could not be obtained from cloning and bacterial transformation. Sequencing results show that all constructs obtained were either partial sequence or rearranged sequence. Similar results were also observed during Y1F1-B construction. In conclusion, generating Y1F1-A & Y1F1-B by PCR and TOPO/pJET cloning in Dh5α competent cell or 5-alpha F'I^{q} Competent *E. coli* competent cell was not successful, likely due to the toxicity to bacteria used.

Since Y1F1 minichunk sequence is originally from BY4741 yeast, it is possible that repaired fragment could be obtained by yeast transformation. Therefore, Multichange ISOthermal (MISO) mutagenesis[90] was applied to fix the single nucleotide mutation in Y1F1 construct. The pRS413 plasmid was used as the accepting vector and template for PCR. Two pairs of primers carrying the corrected base at the mutation site were used to amplify two fragments from the Y1F1 mutant #1, a 40bp overhang was added on both sides of the two fragments (Figure 2.19).

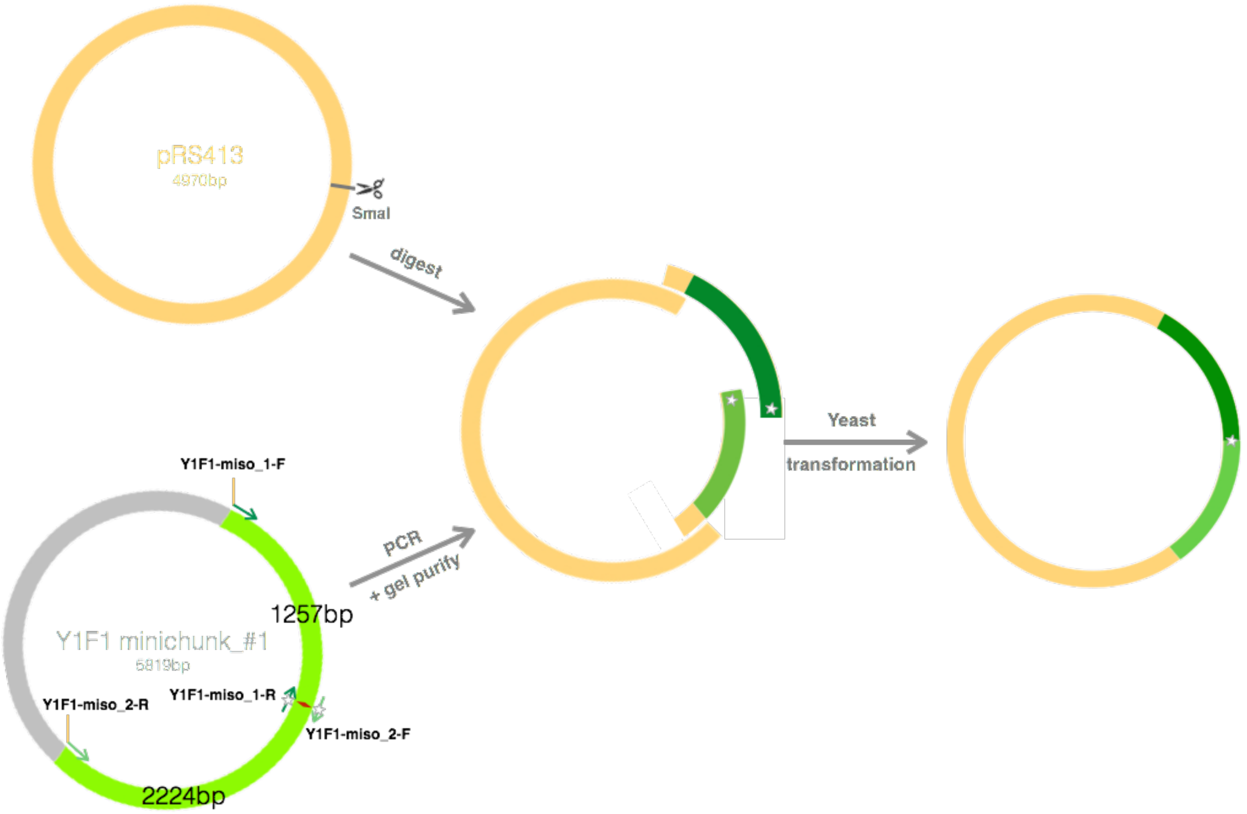


Figure 2.18 Design of Y1F1 MISO

Two pairs of primers carrying the correct base at the mutation site were used to amplify two fragments (one with the size of 1257bp and the other with the size of 2224bp) with 40bp overlapping regions on both sides from the Y1F1 mutant #1. SmaI digestion was performed to linearize pRS413 plasmid. Then the three fragments were assembled through yeast transformation.

Instead of Gibson assembly and bacterial transformation, yeast transformation was applied to assemble the three pieces of DNA, with pRS413 plasmid as the positive control and ddH$_2$O as the negative control for yeast transformation (Figure 2.19A). Then the two sets of PCR primers used to amplify the two fragments of Y1F1 were used again for colony PCR verification. The result gel shows that Y1F1 was successfully assembled in pRS413 plasmid (Figure 19B). In addition, pRS413-Y1F1 construct was isolated from yeast and re-transformed into bacteria to further verify whether Y1F1 minichunk is toxic to bacteria. After re-transform back to *E. coli* (Dh5α), colonies were only observed in positive control plate, confirming that the Y1F1 minichunk was toxic to *E. coli* (Figure

2.20). Therefore, an alternative strategy was applied to integrate the synthetic megachunk Y in BY4742 initial strain to replace the wild-type Megahunk Y. Considering the high efficient homologous recombination mechanism in yeast, here I propose that minichunks can be used directly for integration as each minichunk carries 40bp overhangs (Figure 2.22A). These results show that one colony was obtained with all synthetic fragment integrated at the designated site (Figure 2.21B).

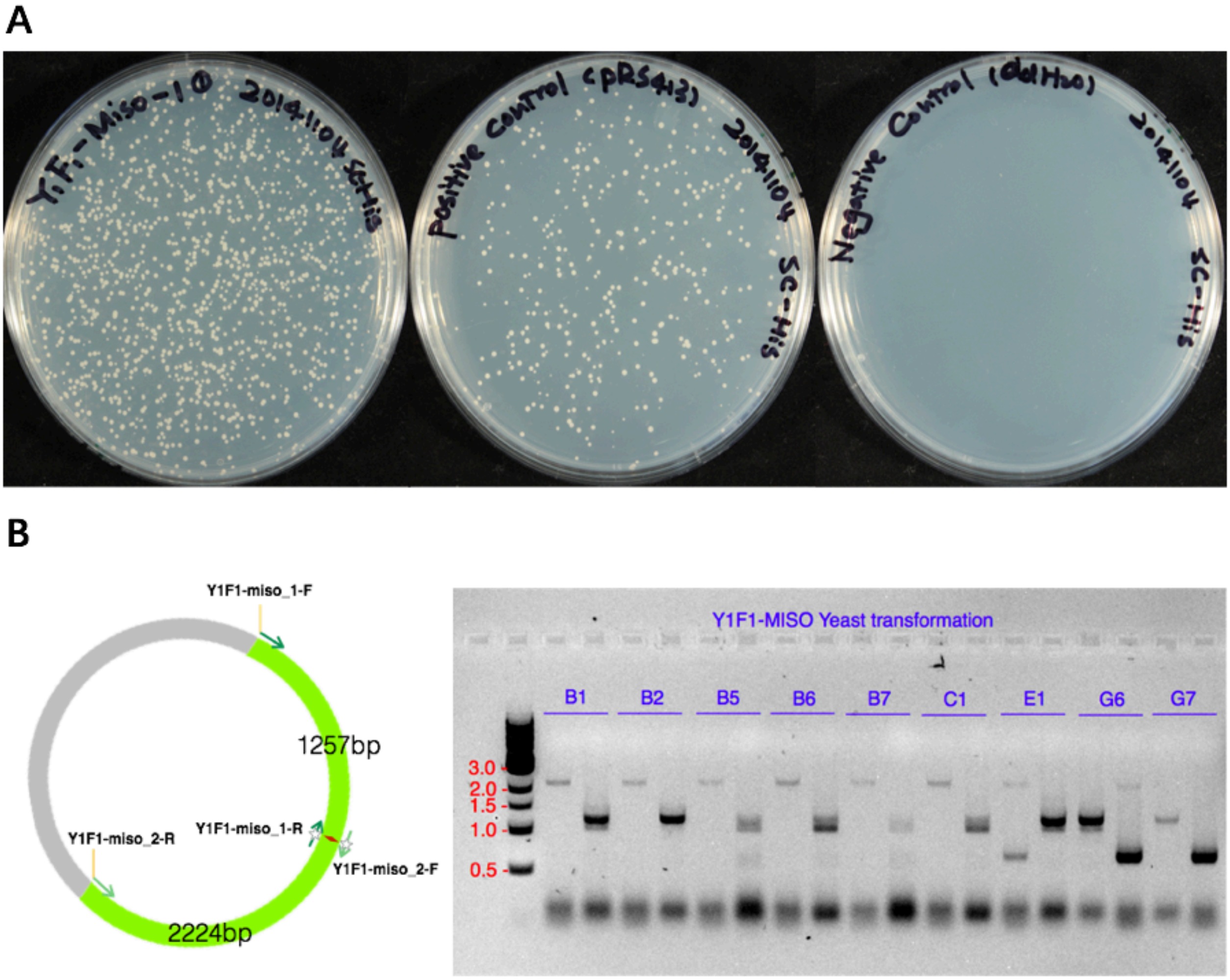

A
Y1F1-Miso-1① 20141104 SC-His
Positive Control (pRS413) 20141104 SC-His
Negative Control (ddH2O) 20141104 SC-His
B
Y1F1-miso_1-F
1257bp
Y1F1-miso_1-R
Y1F1-miso_2-R
Y1F1-miso_2-F
2224bp
Y1F1-MISO Yeast transformation
B1
B2
B5
B6
B7
C1
E1
G6
G7
3.0
2.0
1.5
1.0
0.5

Figure 2.19 Yeast transformation and verification of Y1F1 MISO

A. Yeast transformation of Y1F1 MISO. From left to right, yeast transformation of Y1F1-MISO product, yeast transformation of pRS413 as positive control, yeast transformation of ddH_2O as negative control. B. Colony PCR verification of Y1F1 MISO constructions. From left to right, primer design and colony PCR gel image.

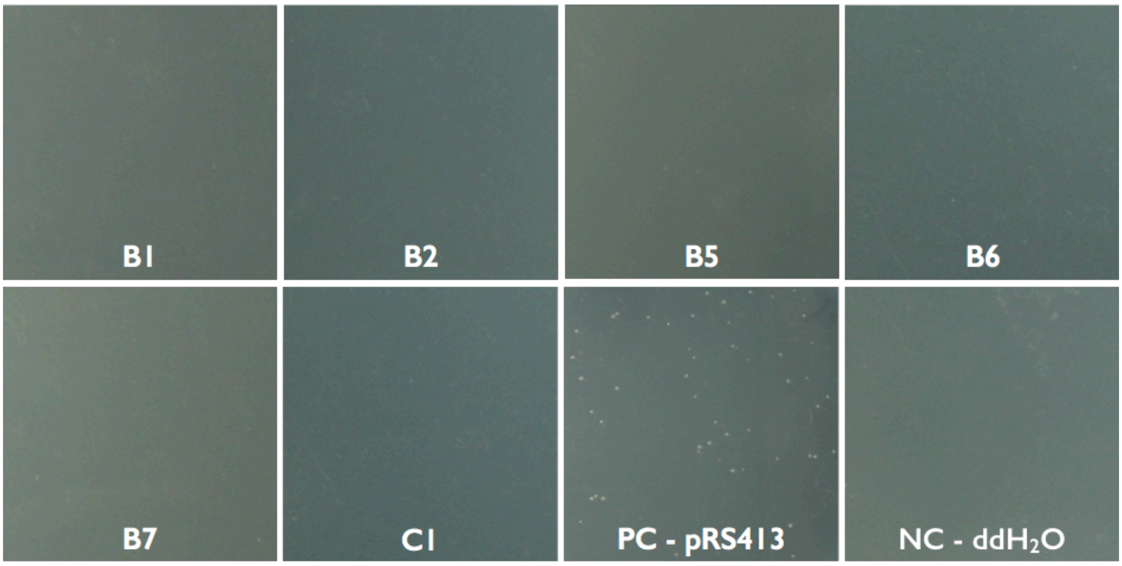


Figure 2.20 Bacterial transformation of pRS413-Y1F1 plasmid

Upper, from left to right, bacterial transformation of colony PCR verified pRS413-Y1F1 construct B1, B2, B5, and B6. Lower, from left to right, bacterial transformation of colony PCR verified pRS413-Y1F1 construct B7, C1, with pRS413 as positive control and ddH_2O as negative control.

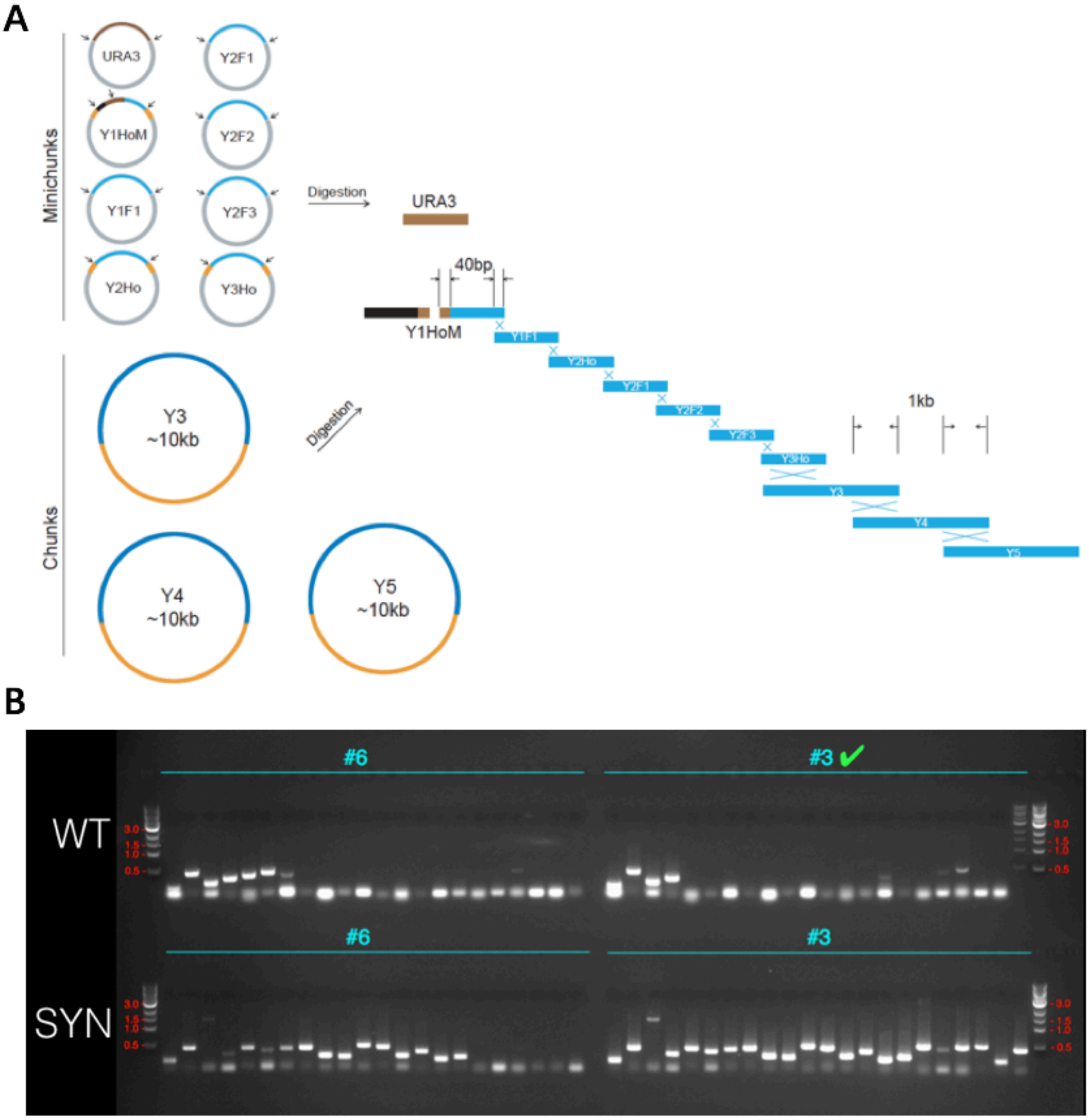


Figure 2.21 Integration of Megachunk Y

A. Minichunks of chunk Y1 and Y2, together with chunk Y3, Y4 and Y5 are digested out and mixed with chunks directly for integration. B. PCRTag verification of integration products. Presence of synthetic PCRTags and absence of wild-type PCRTags indicate the successful substitution of synthetic megachunk Y.

Another problem during integration was also experienced: substitution of Megachunk W was very challenging due to repeated failures. After several runs of attempts, the best construction obtained was one strain carrying a partial wild-type sequence: a ~10kb region junction chunk W4, W5 and

W6 (Figure 2.22). This implies some Sc2.0 design elements within the wild-type region could be responsible for the integration failure. Considering the troubleshooting could be a very time-consuming process, the strain carrying partial wild-type sequence in Megachunk W was used for further construction, while troubleshooting was performed in parallel (details explained in Chapter Three). Once the problem is identified, repair can be performed on the final strain to obtain the final construct.

As a result, the constructed *SynVIIL* and *SynVIIR* were combined through the same strategy used for *SynII*, generating the full *SynVII* chromosome.

2.4.1 Sequencing and correction of variants in *SynVII*

Compared with the designed sequence, Sequencing of the *SynVII* strain revealed 88 SNVs, 2 insertions, 6 deletions, and the partial wild-type region in Megachunk W (Figure 2.23, Table 2.9). 20 of the 88 SNVs were found to correspond exactly to the genotype of the native chromosome sequence at these positions, in agreement with the observation in *SynII* and other chromosomes. 29 SNVs found in intergenic regions were likely introduced during assembly and integration. One base insertion was found in YGR076C gene, which encodes large subunit of the mitochondrial ribosome[91]. A frameshift mutation in this gene might cause respiratory deficiency, which is also consistent with the growth defect observed. Therefore, this particular variation needed to be repaired, and details are explained in chapter three.

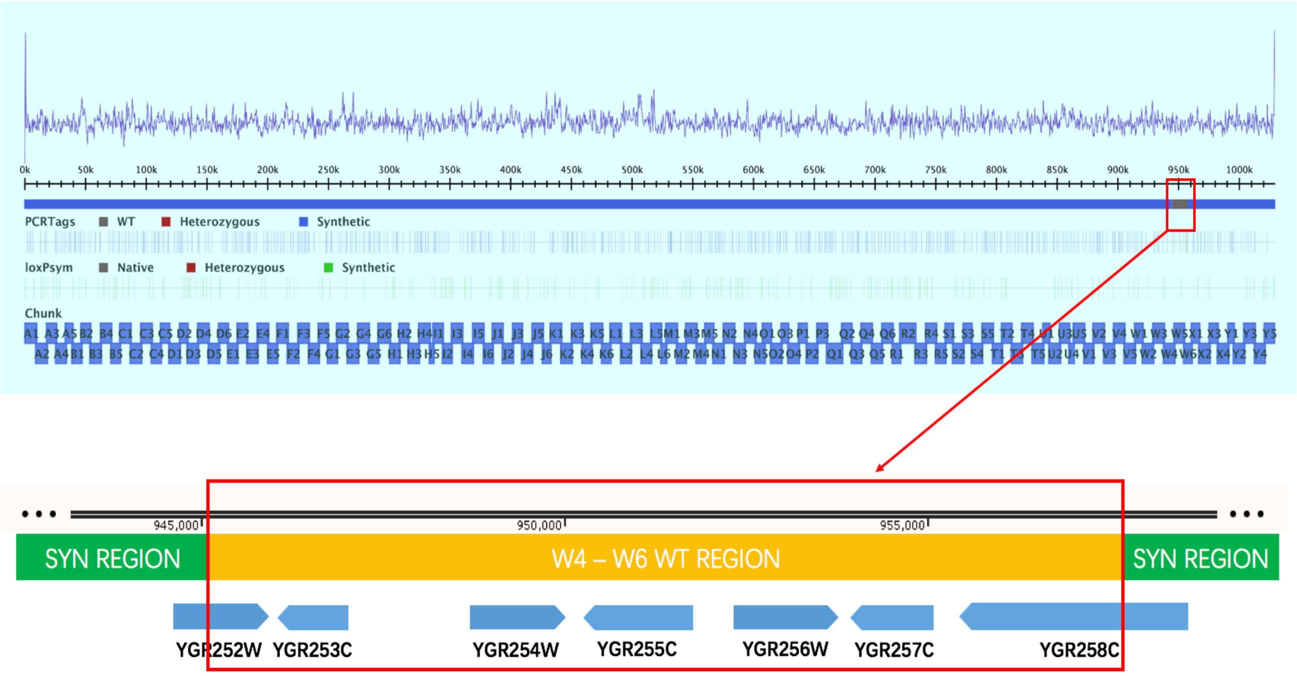


Figure 2.22 Sequencing revealed a missing loxP site and a wild-type region in Megachunk W of *SynVII*

A wild-type region in megachunk W was identified after sequencing verification, The ~10kb region contains the partial coding region of *YGR252W* and *YGR258C*, and ORFs of *YGR253C*, *YGR254W*, *YGR255C*, *YGR256W* and *YGR257C*.

Variation type	Coordinates	Reference *SynVII*	Variant *SynVII*	ORF	Codon changes	Amino acid	Amino acid substitution
SNV	35647	G	A	YGL243W	AGA->AAA	R->K	R375K
SNV	45050	G	A	-	-	-	Intergenic
SNV	45052	A	G	-	-	-	Intergenic
SNV	63955	T	C	YGL227W	TTA->TCA	L->S	L929S
SNV	118823	G	A	YGL197W	GAA->AAA	E->K	E1276K
SNV	211372	A	T	YGL150C	CTA->CAA	L->Q	L1278Q
SNV	228617	G	A	YGL141W	TGC->TAC	C->Y	C155Y
SNV	228865	G	A	YGL141W	GGC->AGC	G->S	G238S
SNV	228978	A	G	YGL141W	ATA->ATG	I->M	I275M
SNV	280790	A	T	YGL116W	GAA->GTA	E->V	E326V
SNV	290392	C	T	YGL111W	CTT->TTT	L->F	L126F
SNV	294107	G	A	YGL108C	TCA->TTA	S->L	S13L
SNV	302039	C	T	YGL102C	GAA->AAA	E->K	E24K
SNV	319501	G	A	YGL094C	CCT->CTT	P->L	P861L
SNV	319502	G	A	YGL094C	CCT->TCT	P->S	P861S
SNV	353788	T	C	YGL076C	AAT->GAT	N->D	N39D
SNV	354196	C	T	-	-	-	Intergenic
SNV	372848	A	G	-	-	-	Intergenic
INS	383112	G	GA	-	-	-	Intergenic
SNV	394086	T	C	YGL049C	AAT->GAT	N->D	N901D
SNV	408323	A	G	-	-	-	Intergenic
SNV	452503	A	G	YGL015C	TTT->TCT	F->S	F52S
SNV	470259	G	A	-	-	-	Intergenic
SNV	495623	C	T	-	-	-	Intergenic
SNV	510745	G	T	-	-	-	Intergenic
SNV	513232	A	G	YGR019W	AAG->GAG	K->E	K109E
SNV	528855	G	A	-	-	-	Intergenic
SNV	576521	T	A	YGR059W	TTC->ATC	F->I	F23I
SNV	578187	A	G	-	-	-	Intergenic
SNV	578499	C	T	-	-	-	Intergenic
DEL	579163	AT	A	-	-	-	Intergenic
SNV	579255	C	T	-	-	-	Intergenic
SNV	580950	A	C	YGR061C	TTG->GTG	L->V	L1326V
SNV	581229	A	G	YGR061C	TCT->CCT	S->P	S1233P
SNV	582191	T	C	YGR061C	GAA->GGA	E->G	E912G
SNV	586977	A	T	YGR064W	TAT->TTT	Y->F	Y110F
SNV	591572	T	C	YGR067C	CAA->CGA	Q->R	Q782R
SNV	601536	T	C	YGR071C	AAA->AGA	K->R	K587R
SNV	602470	G	A	YGR071C	CTT->TTT	L->F	L276F
INS	606435	A	AT	YGR076C	-	-	Frameshift
SNV	606442	C	T	YGR076C	GAA->AAA	E->K	E146K
SNV	638403	T	A	YGR092W	CTC->CAC	L->H	L216H
SNV	638759	A	G	YGR092W	AAG->GAG	K->E	K335E
SNV	639493	T	C	-	-	-	Intergenic
DEL	639616	CAT	C	-	-	-	Intergenic
SNV	639768	A	C	-	-	-	Intergenic
SNV	640573	A	G	YGR093W	AAG->AGG	K->R	K195R
SNV	640791	A	G	YGR093W	AGT->GGT	S->G	S268G
SNV	641509	A	C	YGR093W	AAC->ACC	N->T	N507T
SNV	641809	T	C	-	-	-	Intergenic
SNV	642293	T	G	YGR094W	TAT->GAT	Y->D	Y158D
SNV	662375	A	T	YGR100W	GAA->GAT	E->D	E809D
SNV	664209	T	C	-	-	-	Intergenic
SNV	669668	A	T	-	-	-	Intergenic
SNV	669836	T	C	-	-	-	Intergenic
DEL	670661	TA	T	-	-	-	Intergenic
SNV	671269	T	C	-	-	-	Intergenic
DEL	671909	ATT	A	YGR107W	-	-	Frameshift
DEL	690516	AT	A	YGR118W	-	-	Frameshift
SNV	692026	C	A	YGR119C	CAG->CAT	Q->H	Q291H
SNV	693273	A	G	YGR120C	ATA->ACA	I->T	I259T
SNV	694871	T	C	YGR121C	GAC->GGC	D->G	D415G
SNV	696304	T	C	-	-	-	Intergenic
SNV	696891	A	G	-	-	-	Intergenic
SNV	697061	T	C	-	-	-	Intergenic
SNV	697078	A	G	-	-	-	Intergenic
SNV	697270	T	C	YGR122W	TCA->CCA	S->P	S39P
SNV	697348	A	G	YGR122W	ACC->GCC	T->A	T65A
DEL	728595	GA	G	YGR139W	-	-	Frameshift
SNV	735252	T	C	-	-	-	Intergenic
SNV	737005	C	T	-	-	-	Intergenic
SNV	747229	C	T	YGR146C-A	GAG->AAG	E->K	E6K
SNV	773455	C	G	YGR162W	GCC->GGC	A->G	A193G
SNV	787845	G	T	YGR170W	GCT->TCT	A->S	A585S
SNV	992161	A	G	YGR276C	ATG->ACG	M->T	M500T
SNV	1018180	G	T	-	-	-	Intergenic

Table 2.9 Variants in *SynVII* strain revealed by sequencing data

Synonymous SNVs are not shown in this table.

The repair of the remaining wild-type region of megachunk W showed significant phenotypic defect (details discussed in the next chapter), suggesting potential design flaws that need to be eliminated from *SynVII*. Apart from this, the full construction of *SynVII* at this point can be considered accomplished.

2.5 Conclusion and discussion

In the first part of this chapter, we demonstrated that nanoliter PCR and DNA assemblies were designed and successfully tested using the Echo robotic platform. As the continual efforts are being conducted by the Sc2.0 consortium for other to-be-finished chromosomes, this method can significantly help to save the cost as well as reduce the labor. In the world of laboratory automation, efficiency and robustness are as important as cost-effectiveness. With this in mind, we overlaid a number of correct assemblies (efficiency) with standard deviation (robustness) in the same plot with the cost of reactions for the Gibson assembly (Figure 2.4D) and the Golden Gate assembly (Figure 2.5D). 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 the use of the acoustic dispenser in synthetic biology, and we envision that this technology will be instrumental in lab automation, in particular in the era of DNA foundries.

In the second part of this chapter, the design and construction efforts were explained in detail. *SynII* and *SynVII* were constructed from both ends. Other innovative approaches with the goal of parallelized construction are possible. *SynXII* was constructed in 6 parallel production strains (a method termed Meiotic Recombination-mediated Assembly), which can further promote the construction efficiency[22]. Future efforts combining both approaches for developing a more efficient construction method can be envisioned. Compared to *SynII*, the construction of *SynVII* seems to be more challenging. The failure with the Minichunk assembly of Y1F1 and the substitution of Megachunk W indicate that some previously unrealized biological properties could be the main "limiting steps" towards the genome-wide design.

2.6 Materials and methods

Echo PCR

The GoTaq Green Master Mix (Promega, Madison, WI) was used in the PCR. 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, followed by 7 min at 72 °C, and hold at 4° C. GoTaq Green Master Mix (35 µL) and double-distilled water (ddH2O; 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 study was MicroAmp EnduraPlate (Life Technologies, Carlsbad, CA).

Echo Gibson assembly

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[5]. 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.

Echo Golden Gate assembly

The Golden Gate master mix contained 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. 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 for Echo assembly reactions

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 additional 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.

qPCRTag assay

For qPCRTag reaction setup, 500nL of mastermix, 5nL of gDNA and 10nL of PCRTag primers were dispensed into a 1536 multiwell plate using Echo550. Then the plate was sealed and centrifuged. The LightCycler 1536 was used for the qPCR, with a two-step amplification protocol described previously [110].

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 ddH2O to a final volume of 5 μL whenever 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.

SynII and *SynVII* Segmentation

Once *SynII* and *SynVII* design were completed, segmentation of synthetic chromosome sequence into megachunks (~30 kb) and chunks (~10 kb) was performed using *BioStudio*[70]. By applying an in-house developed program called SegMan, each chunk was segmented into minichunks (~3 kb). Minichunks were designed to be assembled into chunks by Gibson assembly[5], and therefore 40 bp overlaps between each adjacent minichunks were added by SegMan. Terminal restriction endonuclease sites were also added on both ends of each minichunk to allow excision from the plasmid to ensure only 5′ sticky or blunt ends could be generated by digestion. The 40 bp overlap regions were chosen based on the following criteria: minimal free energy > –3 kcal/mol, and melting temperature (Tm) of 68±4 °C. Higher free energy of overlap sequence leads to reduced probability of self-folding of single-stranded DNA and results in higher efficiency for overlap-based in vitro assembly methods. Here the RNAfold program of the ViennaRNA Package 2.0[92] was used to calculate the minimal free energy of the DNA sequence. A simplified formula derived from Rychlik, W. et al.[93] was applied to estimate the melting temperature: $Tm = 945*\Delta H/(\Delta S + R*\log(0.0001)) - 273.15$, in which ΔH is enthalpy (kJ/mol), ΔS is entropy, R is the molar gas constant (1.9872 cal/mol-K). In addition, to reduce synthesis cost, *LEU2* and *URA3* markers were designed separately as independent cassettes and used repeatedly in minichunk assembly.

Minichunk Assembly

Synthesis of ~3 kb minichunks was outsourced to Invitrogen, Genscript, BGI Tech and Life Tech. *pSBGAA* or *pSBGAK* (sequence information can be found *at www.syntheticyeast.org*) was chosen as the accepting vector for chunks. Minichunks were excised from the plasmids using the terminal restriction sites. *Bam*HI was used to linearize the chunk-accepting vectors *pSBGAA* (*Amp*^{R}) and *pSBGAK* (*Kan*^{R}). The chunks were assembled using a modified Gibson assembly strategy[5]. In brief, 1 μL of Taq ligase (New England Biolabs) was added in the final volume of 20 μL for each reaction. The molar ratio of accepting vector to minichunk DNA was 1:5. After thorough mixing, a one-hour incubation at 50°C was performed. Then 10 μL of the reaction mixture was transformed into 50 μl *E.coli* DH5α competent cells (TAKARA). For assembly verification, single colonies were picked and subjected to overnight culture at 37°C. After miniprep, restriction digestion of the terminal restriction sites was performed to check the assembly outcomes.

Replacement of WT Yeast Chromosome with Synthetic Chunks

BY4741 (*MAT***a** *his3Δ1 leu2Δ0 LYS2 met15Δ0 ura3Δ0*) containing the *KanMX* marker (strain ID: *YS000-L*) and *BY4742* (*MATα*) containing the *URA3* marker (strain ID: *YS000-R*) were used as the initial strains for synthetic megachunk replacement from the left arm and right arm, respectively. The *URA3* and *LEU2* selectable markers were used iteratively for replacing the native sequences of chromosome *II/VII* with synthetic chunks. The chunks constituting each megachunk were co-transformed using the LiOAc transformation protocol[17] with 300 ng of each chunk DNA and 200 ng of each junction DNA added. The transformation products were re-suspended in 100 μL 5 mM $CaCl_2$ and plated on an appropriate selectable media (SC–Ura or SC–Leu) with serial dilutions where appropriate. After several successive rounds (from both left and right arms) of replacements, two semi-synthetic chromosomes were successfully constructed. The selectable marker *URA3* in one of the semi-synthetic chromosome was removed through yeast transformation with the corresponding markerless synthetic fragment and screened with 5-FOA[94].

Integration of parallel constructed semi-synthetic chromosomes

An I-*Sce*I-mediated method was developed to combine the two semi-synthetic chromosomes, *synIIA-R* and *synIIR-Y*. For *synIIA-R*, a fragment containing the I-*Sce*I recognition sequence and *URA3* (I-*Sce*I-*URA3*) was designed to have a 40 bp overlapping sequence both upstream and downstream of the *LEU2* marker on one of the semi-synthetic chromosome. A previously described method[95] was used to integrate the I-*Sce*I-*URA3* cassette into one of the semi-synthetic chromosomes. In addition, a *LEU2* marker was inserted into *the* other semi-synthetic chromosome upstream of the centromeric region. Mating of the two semi-synthetic chromosomes was performed by overnight co-culturing in 3mL YPD medium at 30°C. The I-*Sce*I expression vector *pRS413-pGAL-I-SceI* was transformed into the diploid cells followed by plating on SC–His plate. Single colonies were picked and subjected to overnight culture in 3 mL SC–His/glucose medium. The overnight culture was then added to 20 mL SC–His/raffinose (2% raffinose and 0.1% glucose) to reach an initial OD_{600} of 0.1 and incubated until OD_{600} reached 0.4 (~4-hour). Cells were harvested by centrifugation at 6,200 rcf for 5 min and re-suspended in 20 mL SC–His/galactose (2% galactose). After a 2-hour I-*Sce*I induction in galactose-containing media, 20 μL cells were inoculated into 3 mL YPD medium, followed by overnight culture at 30°C. Overnight culture (1 mL) was harvested by centrifugation at 13,800 rcf for 1 min. The pellets were re-suspended in 200 μL ddH_2O, patched on a SPOR plate and incubated at room temperature for 1 day. Then 5-7 days incubation at 30°C was performed until a significant number of tetrads was observed.

Cells were re-suspended in 25 μL Zymolyase-20T (25 mg/mL in 1 M sorbitol), then incubated at 37°C, with shaking at 800 rpm for 60 min. Then 500 μL ddH_2O was added and mixed thoroughly, followed by plating on 5-FOA plates and incubating at 30°C for 3 days. Then replica plating on SC–Leu and YPD plates was performed, followed by 24 hours incubation at 30°C. FOA^R Leu^- colonies were selected from YPD plates, inoculated into 3 mL YPD medium, and cultured overnight. To verify the integration of semi-synthetic chromosomes, one pair of synthetic PCRTag and wild-type PCR tag were chosen from each megachunk (in total 25 pairs with one from each Megachunk) to perform PCRTag analysis. Then sequencing was performed for further verification.

PCRTags Analysis Reaction setting

PCRTag amplification was performed using rTaq Polymerase (TAKARA). Forward and reverse PCRTag primer pairs (400 nM each, detailed information of PCRTag primers can be found at www.syntheticyeast.org), 1 μL 20Mm NaOH treated cell culture and 2.25 μL ddH_2O were added to a final reaction volume of 12.5 μL. The PCR thermal-cycler program setting was used as follows: 94°C/5 min, 30 cycles of (94°C/30 sec, 55°C/30 sec, 72°C/30 sec), and a final extension of 72°C/5 min. Detection of PCRTags was carried out by gel electrophoresis. PCR product (3 μL) was loaded onto a 2% agarose gel, and electrophoresis was performed at 180V for 20 min.

Yeast Genomic DNA Preparation for DNA Sequencing

Yeast cells were grown in 5 mL YPD medium using a 14 mL round bottom tube for 2 days till saturation. Pellets were collected by centrifugation at 13,800 rcf for 1 min. Breaking buffer (400 μL) was added to re-suspend the pellet. Glass beads (0.2 g, 0.5 mm of diameter) and 400 μL PCI (Phenol:Chloroform:Isoamyl alcohol=25:24:1) were added, the re-suspension was vortexed at maximum speed for 3 min, and then centrifuged at 13,800 rcf for 10 min. 400 μL of the aqueous layer was transferred to a new 1.5 mL tube. The genomic DNA was precipitated by adding 400 μL of isopropyl alcohol and kept at room temperature for 5 min. Then genomic DNA was pelleted by centrifugation at 13,800 rcf for 10 min. The pellet was washed with 500 µL 70% ethanol, followed by 5 min drying at 37°C. The genomic DNA was re-suspended in 50 µL TE buffer (10 mM Tris-HCl pH8.0, 1mM EDTA) with RNase (25 µg/mL) and incubated at 37°C for 30 min.

Nucleotide Sequence Analysis of *SynII* with Hiseq2500 sequencing platform

Library preparation and whole genome sequencing

Paired-end whole genome sequencing was performed for the *SynII* and *SynVII* on the HiSeq2500 platform. A 500-bp library was prepared according to standard Illumina DNA preparation protocols.

Sequencing quality control and mapping to genome sequence

Before mapping of reads, quality control of sequencing reads was performed. Reads with adapters or shorter than 90 bp were removed. Reads with more than one base having a Phred-score below 10 or with more than one unknown base were removed, leaving 673 Mbp cleaned paired-end reads, with 55.6-fold sequencing depth of the genome. Cleaned reads of each sample were mapped to yeast reference sequences (the original sequence of chromosome *II/VII* being replaced by synthetic chromosome *II/VII* sequence) using BWA 0.5.6[96] with standard settings. For each alignment result, local realignment was performed with GATK 2.7-2 RealignerTargetCreator and IndelRealigner tools[85] to clean up mapping artifacts caused during reads mapping on the edges of indels. The resulting files in BAM format were then prepared for initial SNV/indel calling.

Identification of SNVs and indels

Both Samtools[97] and GATK 2.7[85] pipelines were used to identify the SNVs and indels using default parameters. The variants were filtered by the criteria of QUAL < 50 or PV4 < 0.1,0.1,0.1,0.1 or MQ <10 or DP < 10 or DP4 < 0,0,3,3 for Samtools results and QUAL < 50 or FS >3 or BaseQRankSum >3 or MQRankSum>3 or ReadPosRankSum>3 or MQ <10 or DP < 10 for GATK results. The variants identified by either tool were merged with CombineVariants implemented in GATK. The merged variants in the synthetic chromosome were checked manually using Tablet[98] to exclude false-positive results caused by sequencing or mapping errors. Annotation was performed for observed variants, with the following types: synonymous type, non-synonymous type, frameshifts, and variant outside coding region.

Identification of loxPsym sites

In *SynIII*, *loxPsym* sites can be absent from expected locations[17]. To check whether all loxPsym sites are present in *SynII/SynVII*, sequencing reads containing loxPsym sites were extracted from the whole read library for identifying the presence of all expected *loxPsym* sites. For each loxPsym site, the read mapped span the loxPsym site of upstream and downstream flanking sequence >10 nucleotides and the sequenced loxPsym site with bases of >=24 matching or mismatching loxPsym sequence was recognized as a loxPsym read and bases of >=15 deleted was counted as a "*loxPsym*_lost" read. The loxPsym site with p-value of < 0.001, which estimated by Poisson mode, and supporting read number >=5 was identified as a loss of *loxPsym* site.

Structural variant detection with the PGM sequencing platform

Library preparation and whole genome sequencing

A 400-bp DNA library of a *SynII* was prepared for single-end whole genome sequencing according to the Life Tech standard preparation protocol using the Ion Xpress™ Barcode Adapter 1-96 Kit (Cat.no.4474517) and sequenced on the Ion PGM™ platform. Quality control of sequencing reads was performed. Reads shorter than 30bp or duplicated were removed. Reads with more than 1% of bases having a Phred-based quality score <10 or with unknown bases were trimmed first to meet the filtering criteria and removed if the first trimming step failed, leaving 488 Mbp cleaned reads, with 32.6-fold sequencing depth of the genome and 248 bp average length. Cleaned reads of each sample were mapped to synthetic yeast genome sequences using bowtie2-2.0.0[84] with standard settings.

Structural variant reconstruction

To identify whether structural variations exist in *SynII*, reads that did not map to the reference genome were split to pairwise ends (split reads, at least 30 bp) by scanning over all intermediate positions at least 30bp from the ends of the read. Then these pairwise ends were then aligned to the reference using Bowtie2[84] by single-end mapping with parameter –k 100. The pairwise ends that matched parental sequence were analyzed for breakpoints to provide direct evidence for structural variants. One read was assigned as the most probable mapping type based on five priorities: no recombination events (top priority), intra-chromosome recombination, inter-chromosome recombination within wild-type chromosomes and external chromosome recombination between synthetic and wild-type chromosomes, and single end mapping. For each identified breakpoint, a 5 bp error range was allowed. These reads at a breakpoint site with at least 3 supporting reads were locally assembled using PolyPhred[99], and then the assembly results were annotated by Blastn against the reference containing the selection marker and acceptor vector and wild-type chromosome *II* sequence to identify the potential extrinsic sequence. Combining the split-read mapping, sequencing depth and local assembly contigs, the structural variations were reconstructed.

SynII Structural Variant Correction

An I-*Sce*I mediated transformation strategy was designed to correct the two structural variations observed in Megachunks *L* and *T* of *SynII*. For the duplication in Megachunk *L*, a donor fragment containing the *URA3* sequence and an I-*Sce*I site (TAGGGATAACAGGGTAAT) was designed to carry 40 bp overlapping regions at the end of the original chunk *L4* and the beginning of duplicated chunk *L2*. The donor fragment was produced by PCR using primers including the I-*Sce*I site and overlap sequence. I-*Sce*I expression vector *pRS413-pGAL-I-SceI* and the donor fragment were co-transformed into the *synII* strain. After I-*Sce*I induction on galactose media, cells were plated directly on 5-FOA plates directly and incubated at 30°C for 3 days. Then FOA^R colonies were selected and cultured overnight in YPD medium at 30°C. Confirmation of repair was performed by PCR using a pair of primers designed to amplify across the *L4-L2* breakpoint (primer sequence information can be found *at www.syntheticyeast.org*).

For the complex variation in Megachunk *T*, a donor sequence was designed to contain a *URA3* cassette and the I-*Sce*I site, with the left end overlapping with the right end of chunk *T4* by 500 bp, and the right end overlapping with the left end of chunk *T5* by 500 bp (Fig. 1B). Both 500 bp overlapping sequences and the I-*Sce*I-*URA3* cassette were produced by PCR with 40 bp overlaps with each other and assembled together through the Gibson assembly method[5]to form the donor fragment. An additional 100 bp of chunk *T4* was added between the I-*Sce*I-*URA3* cassette and 500 bp overlapping sequence of *T5* to facilitate the homologous recombination. The donor fragment and I-*Sce*I expression vector *pRS413-pGal-I-SceI* were then co-transformed into the *SynII* in which the structural variation in Megachunk *L* had been repaired. I-*Sce*I induction was performed on galactose-containing media and colonies were screened on 5-FOA-containing plates. PCR verification was then performed with 8 pairs of verification primers designed across each breakpoint observed in this structural variation. The *SynII* was sequence-verified again using the PGM sequencing platform, with library preparation and sequencing method as described in "structural variant detection with the PGM sequencing platform".

Mating Type Switch of Yeast

pJD138 (*pGAL-HO*) was transferred into the *SynII* strain using the LiOAc transformation protocol[17], followed by overnight culture in 5ml SC–Ura/galactose (2% galactose) liquid at 30°C. The mating type was then determined by crossing with *MAT***a** and *MATα* tester strains and confirmed by microscopy.

Mating and Sporulation

Mating of a *SynII* strain and a second *SynII* strain that had its mating type switched to alpha (strain ID: YS032) were recovered on YPD plates. Colonies were picked and patched in the same section (mating section) on a new YPD plate with an inoculating loop, followed by overnight culture at 30^{o}C. Then diploid colonies were patched on SPOR plates [10 g potassium acetate (Sigma), 1.25 g yeast extract (Oxoid), 1 g glucose (Sangon Biotech), 20 g agar (Sangon Biotech), ddH$_2$O up to 1 L] and left at room temperature for one day, followed by 8 days incubation at 30°C. Then cells were picked from SPOR plates and re-suspended in 25 μL zymolyase solution (MP Biomedicals, 1 mg/ml Zymolyase in 1M sorbitol), followed by 25 min incubation at 37°C. Cells were checked under the microscope to determine zymolyase efficiency. After tetrads were appropriately digested, 500 μL ddH$_2$O was added and the suspension mixed vigorously by vortexing to break up the tetrads. Then a series of 10-fold dilutions were made before plating on YPD plates. The plates were incubated at 30°C for 2-3 days. Single colonies were picked and streaked on fresh plates, followed by overnight incubation. Mating type of random spores was confirmed by crossing with *MAT***a** and *MATα* tester strains and confirmed by microscopy. Spore-derived colonies were selected for PCR verification. The nine pairs of primers used for identifying breakpoints in *SynII* were used here to verify the correct *SynII* sequence (primer sequence information can be found *at www.syntheticyeast.org*). Pulsed-field gel electrophoresis (PFGE) was performed for karyotype analysis. Finally, candidates were further verified using the PGM sequencing platform, with library preparation and sequencing method as described in "structural variant detection with the PGM sequencing platform" above.

Pulsed-field Gel Electrophoresis

Samples were prepared for pulsed-field gel electrophoresis as previously described by Herschler J, et al[100]. Identification of chromosomes was inferred from the karyotype of WT controls (*BY4741*, *BY4742*) on the same gel. Samples were analyzed on a 1.0% agarose gel in 1 × TAE (pH 8.0) for 22 hrs at 14°C on a CHEF apparatus. The voltage was set to 5 V/cm, at an angle of 120°. Switch time was set to 60 – 90 sec ramped over 22 hrs.

BigDye Reaction setting

BigDye sequencing PCR reaction includes 0.32ul sequencing primer(10uM), 2ul DNA (80-120ng), 2ul of 5X Sequencing buffer, 2ul 1:4 diluted Big Dye and 3.68ul ddH$_2$O. Reaction setting is as follows:32 cycles of (95°C/30sec, 50°C/20sec, 60°C/4min), hold at 4°C.

TOPO Cloning

Reaction of TOPO Cloning includes 1ul TOPO® vector, 0.5-4ul PCR product, 1ul salt solution and ddH_2O to a final volume of 6ul. Mix the reaction mixture gently and incubate for 5-30mins at RT.

pJET Cloning

For cloning of blunt-end PCR product, ligation reaction includes 1ul pJET1.2/blunt cloning vector (50ng/ul), 1ul purified PCR product, 10ul 2X reaction buffer, 1ul T4 ligase and ddH_2O to a final volume of 20ul. Mix the reaction mixture gently and incubate for 5-30mins at RT.

T4 ligation

T4 ligation reaction includes 2ul DNA, 0.5ul T4 ligase, 1ul T4 ligation buffer and ddH_2O to a final volume of 10ul. Mix the reaction mixture gently and incubate overnight at 16°C.

Bacterial Transformation

For standard bacterial transformation, a maximum 10ul of transformation DNA is added into 100ul Dh5α competent cells, followed by 20 mins incubation on ice; after heatshock at 42°C for 45 seconds the cells were stored on ice for additional 10 mins; then 750ul of S.O.C medium was added and the cells were recovered in 37°C for 1 hour; finally 200ul of transformation product was plated on selection plates.

For bacterial transformation with MAX Efficiency® Stbl2™ competent cell, 5uL of DNA was added to 100ul competent cells followed by incubation on ice for 30 mins; heat shock for 25 sec; further incubation on ice for 2 mins; addition of 900ul S.O.C medium; shaking at 200rpm, 30°C for 90 mins; spreading 200ul on selection plates; and incubation overnight at 30°C.

Plasmid Isolation from Yeast

Cells were harvested from 3 ml overnight culture in YPD and resuspended in 250 μl of buffer P1 (QIAprep miniprep kit). After addition of 100 μl of glass beads, cell walls were broken by vortexing for 5 minutes. Following this treatment, plasmids were isolated using the standard alkaline lysis and a Qiagen miniprep spin column to isolate the DNA. Aliquots of the shuttle plasmid were used to transform *E. coli*.

Chapter Three

Part 2: Characterization - the fitness assays for *SynII* and *SynVII*

3.1 Introduction

As described in Chapter Two, thousands of design elements were introduced to the synthetic chromosomes following the design principles. The extensive changes made to the chromosome naturally raise the question whether there are any phenotypic effects caused by these changes. Here in this chapter, in-depth characterizations of both *SynII* and *SynVII* were described. Both synthetic chromosomes showed varying levels of defects, therefore, debugging methods were also developed to quickly identify the defect origins.

3.2 *SynII*

During the construction of *SynII*, a significant fitness defect was observed after Megachunk *E* integration (Figure 3.1, Figure 3.4). Further investigation revealed that the defect was introduced by the insertion of the selective marker *URA3* (for integration selection) adjacent to the *NCL1* gene, which encodes a tRNA:m5C-methyltransferase. Loss of function of *NCL1* gene was previously reported to be involved in slow growth and temperature sensitivity[101,102], which could explain the observed phenotypic defect. This auxotrophic marker interference phenotype was corrected in the next step of SwAP-IN, in which the *URA3* marker was replaced by an intact *NCL1* gene. Similar instances of such reversible fitness defects were also found after integration of Megachunk *B*, *U*, and *W* (Figure 3.1).

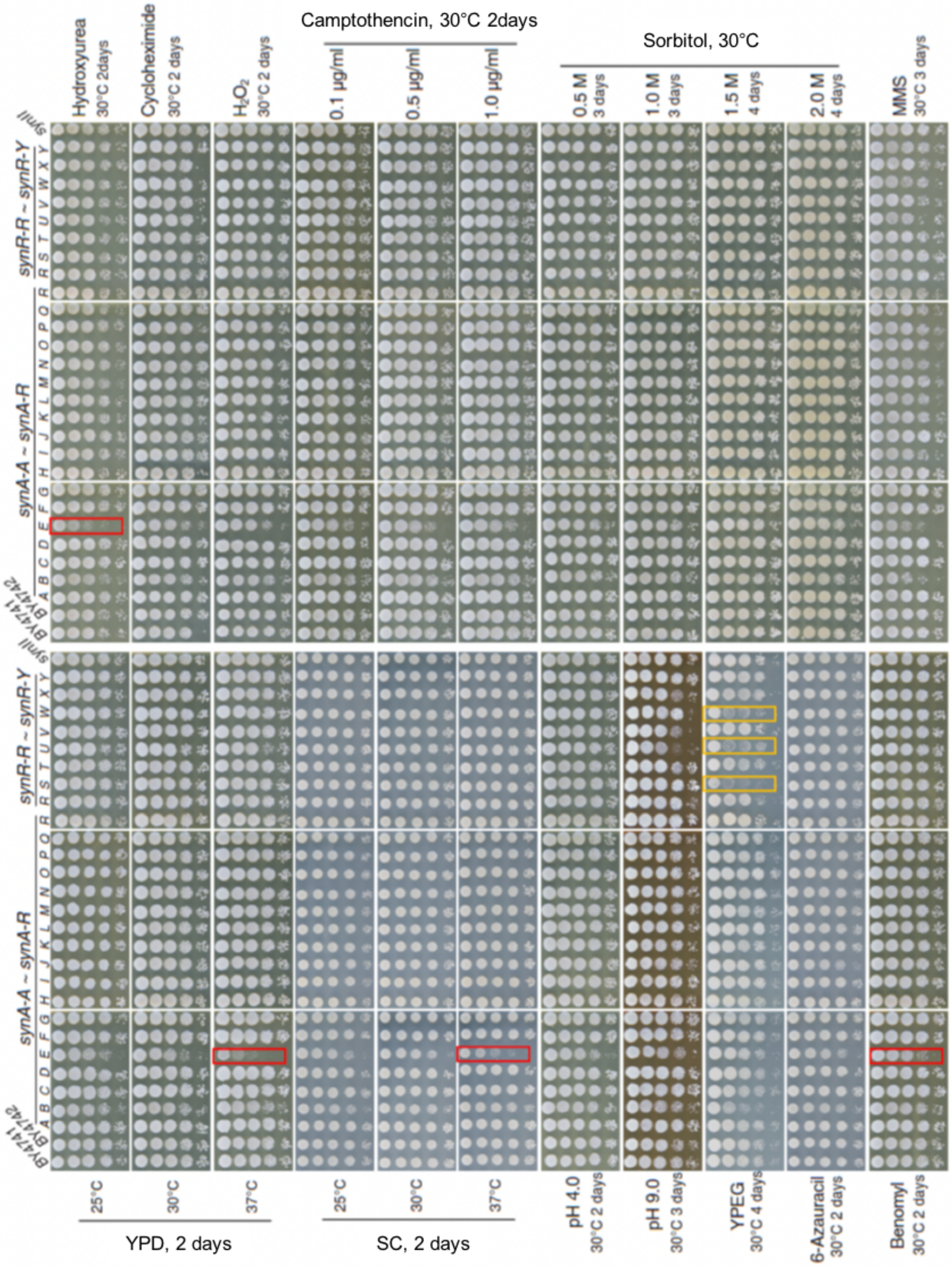


Figure 3.1 Phenotypic assay of all intermediates and the final *SynII* strain under different conditions

Phenotypic assays under 22 different conditions were performed for all intermediate strains to monitor the fitness defect(s). 10-fold serial dilutions of overnight cultures were used for plating. A major phenotype defect was observed in the *SynIIA-E* intermediate strain when culturing under high temperature and in the presence of Benomyl and Hydroxyurea (Red highlighted). Another major defect on the YPEG plate was also revealed in *SynIIR-S*, *SynIIR-U* and *SynIIR-W* intermediate strains (Orange highlighted).

3.1.1 Phenotypic and morphological assays of *SynII*

Growth curves, phenotype and morphology under various culture and stress conditions were evaluated to check the fitness of *SynII strain* compared to the wild-type counterparts (BY4741 and BY4742) (Figure 3.2, Figure 3.3). Results show that the final *SynII* strains are largely on par with wild-type strains.

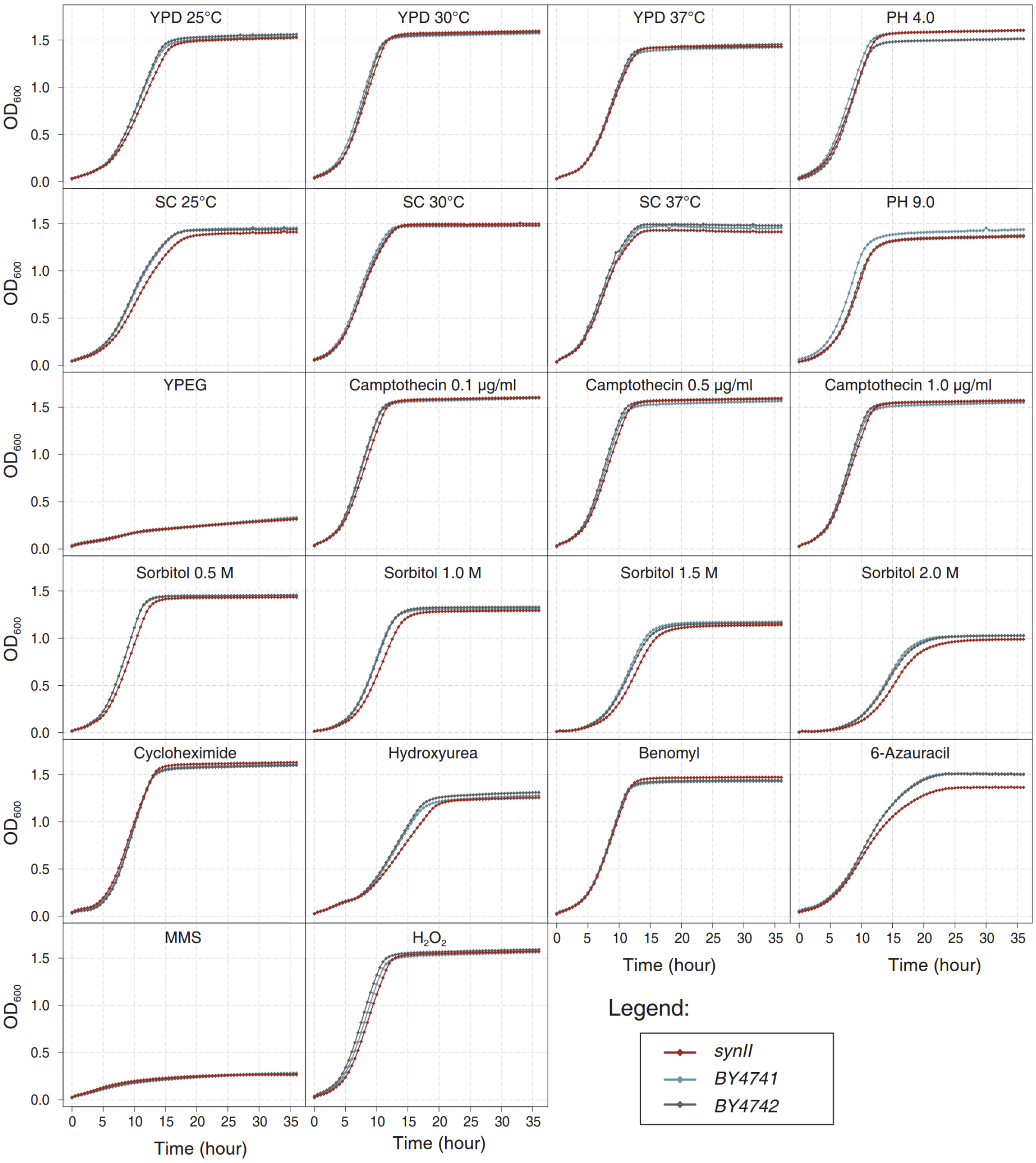


Figure 3.2 Growth curves of *SynII* strain under different conditions

SynII cells were cultured in 22 different conditions and OD_{600} were measured at indicated time to check for growth defects (BY4741 and BY4742 strain as reference). Except for a minor growth defect in media with 6-Azauracil (6AU), no major growth defect was observed.

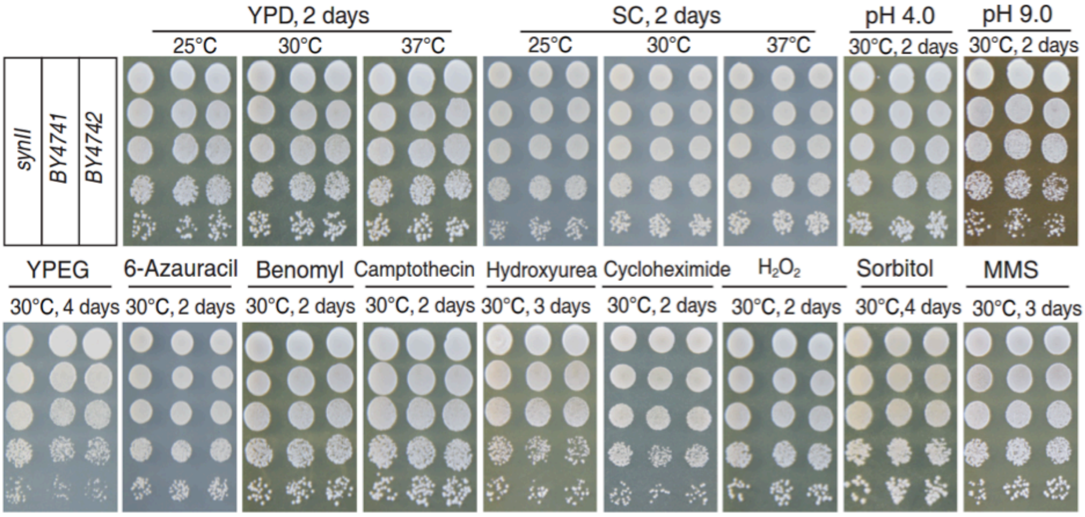


Figure 3.3 Phenotypic profiling of *SynII* on different media

10-fold serial dilutions of overnight cultures of *synII* and wild-type (*BY4741* and *BY4742*) strains were used for plating. From left to right: YPD at 25°C, 30°C, and 37°C; SC at 25°C, 30°C, and 37°C; low pH YPD (pH 4.0) and high pH YPD (pH 9.0); YPEG; SC+6-Azauracil; YPD+Benomyl; YPD+Camptothecin; YPD+Hydroxyurea; YPD+Cycloheximide (10 μg/ml, 2 hrs pretreatment); YPD+H_2O_2 (1 mM, 2 hrs pretreatment); YPD+Sorbitol; YPD+MMS, (YPD, yeast extract peptone dextrose; YPEG, yeast extract peptone glycerol ethanol; MMS, methyl methane sulfone; SC, synthetic complete).

Fitness defect rescue of *SynII* by complementation assay

A defect at high temperature (37°C) in medium with glycerol as the carbon source was observed in *SynII*_cells: colony size is much smaller than wild-type. By crossing *SynII* with wild type strain (BY4742), it was confirmed that the slow-growth phenotype is recessive (Figure 3.4A). To ascertain the origin of this defect, the phenotype of each intermediate strain under this particular growth condition was examined. The results show that the defect originated from Megachunk X (Figure 3.4B). After checking all genes carrying design features in Megachunk X, one essential gene *YBR265W,* which encodes 3-ketosphinganine reductase (Tsc10p) that catalyzes the second step in the pathway for sphingolipid synthesis in *S. cerevisiae*[103], was found modified carrying a synthetic PCRTag. As sphingolipids are reportedly involved in the regulation of the yeast high-osmolarity glycerol (HOG) response pathway[104], we reasoned that the defect might be caused by the modification in *YBR265W*. Using complementation assay by replacing the synthetic PCRTag in *YGR265W* by its corresponding wild-type sequence, we proved that the defect is caused by PCRTag recoding in *YBR265W* (Figure 3.5). This highlights the effectiveness of the genome debugging mechanism provided by the modular chromosome construction strategy in the Sc2.0 project.

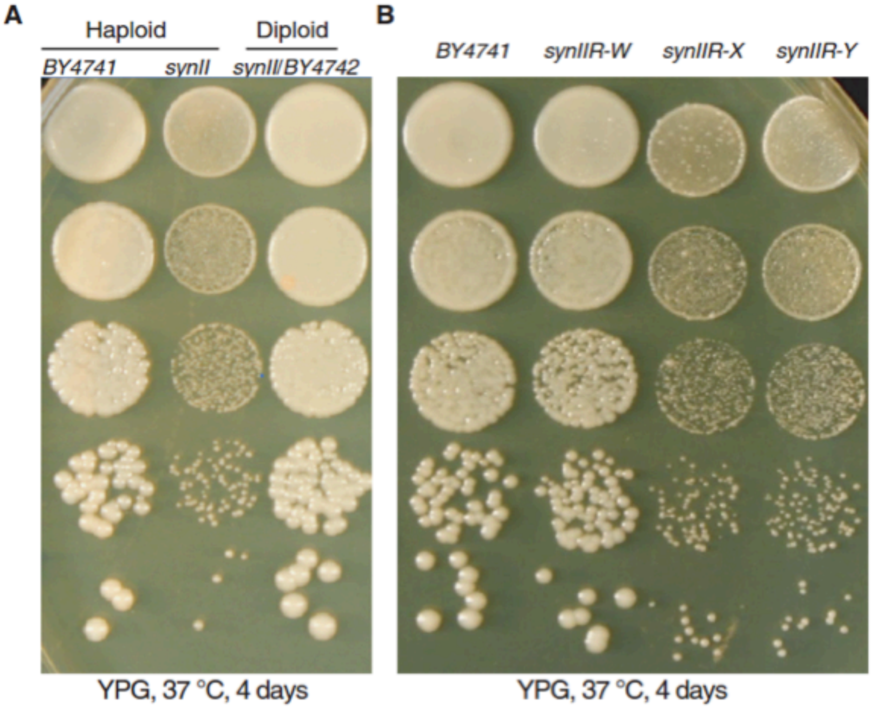


Figure 3.4 *SynII defect* at 37°C on YPG media

A. *SynII* was mated to BY4742 and the resulting diploid cells were spotted on YPG media and incubated for 4 days at 37°C for checking whether the mutation responsible for the defect was recessive. B. Spotting on YPG plate with intermediate strains *SynIIR-W*, *SynIIR-X* and *SynIIR-Y*, with BY4741 as control. The defect started appearing with the *SynIIR-X* strain, indicating the defect originates within megachunk *X* region.

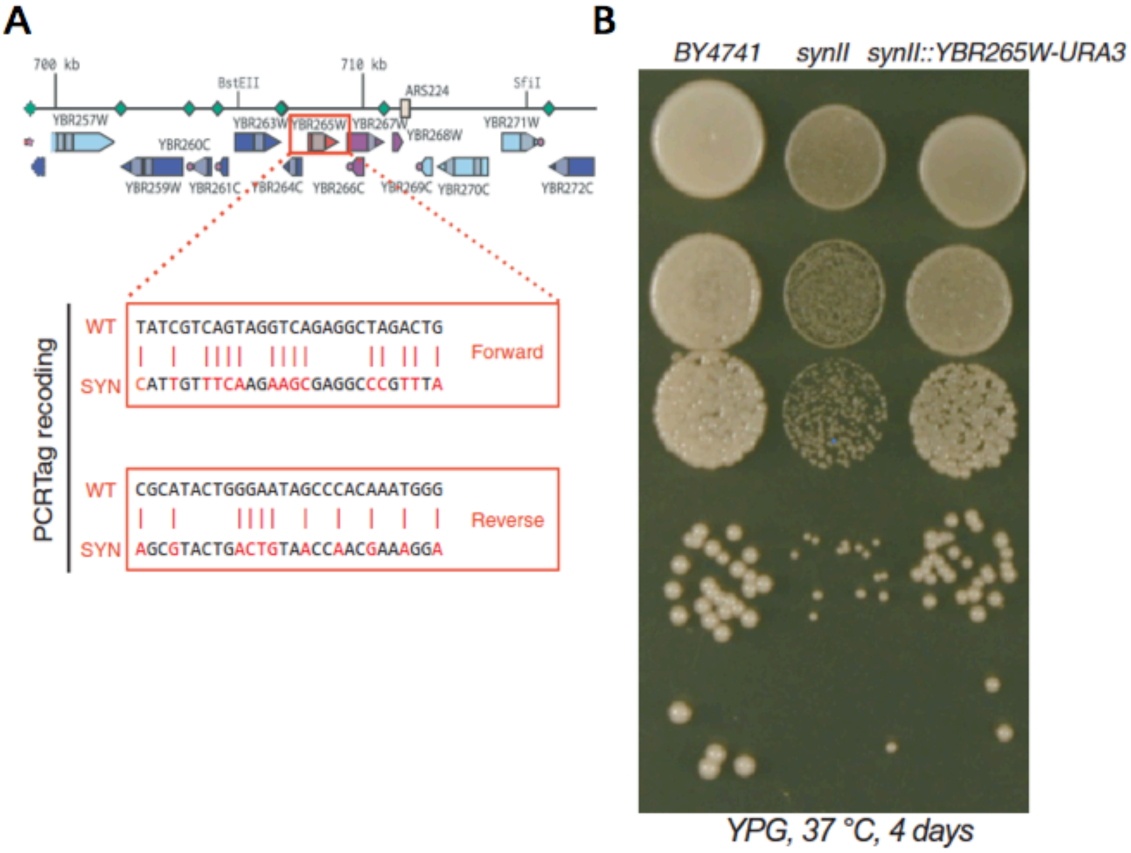


Figure 3.5 *SynII defect* at 37°C on YPG media is rescued by complementing with wild-type *YBR265W* gene

A. PCRTag recoding of *YBR265W* gene. *YBR265W* was modified in *SynII* to contain recoded one PCRTag. B.The PCRTag was recoded by replacing the modified *YBR265W* gene with the wild-type gene. *SynII::YBR265W-URA3* strain spotting on YPG medium was performed, with BY4741 and *SynII* as control.

As more and more individual synthetic yeast chromosomes are near complete, combining them into a single haploid cell will be the next step. To this end, we successfully merged *SynII* and *SynIII* using "endoreduplication intercross" method established previously[20]. The resulting *SynII/III* cell grows just like WT cells (Figure 3.6), indicating that *SynII* can be successfully incorporated into another synthetic background without introducing growth defects.

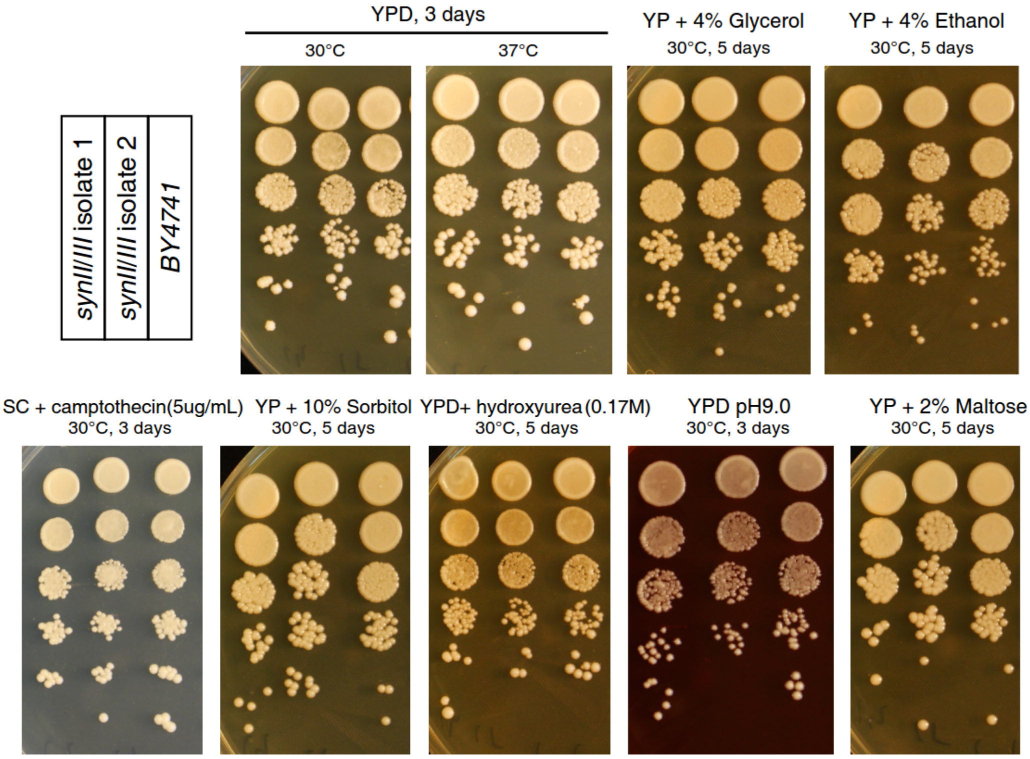


Figure 3.6 Phenotypic assay of two *SynII/III* isolates under different conditions

10-fold serial dilutions of overnight cultures of *SynII/III* isolates and wild-type strain (BY4741) were used for plating. From left to right: YPD at 30°C, and 37°C; YP + 4% Glycerol at 30°C; YP + 4% Ethanol at 30°C; SC + Camptothecin (5 μg/mL) at 30°C; YP + 10% Sorbitol at 30°C; YPD+ Hydroxyurea (0.17 M) at 30°C; high pH YPD (pH 9.0) and YP + 2% Maltose at 30°C (SC, synthetic complete).

3.1.2 Segregation and replication of *SynII*

Once *SynII* was constructed and phenotypically profiled, a logical next question is to ask how this synthetic chromosome replicates and segregates. Therefore, to further investigate whether the modifications introduced in *SynII* compared to WT chromosome II might have introduced subtle chromosome defects not detected in large-scale phenotypic assays, we examined and compared DNA replication and segregation processes of *SynII* and WT (BY4741) strains by tagging *SynII* with an array of *tet* operators and using the previously described TetR-GFP method[105]. The pRS306*tetO14* plasmid (AMp327) carries 224 copies of *tetO* sequence on the pRS306 backbone. p128tetR-GFP (AMp326) is another vector in which GFP is fused with *tet-R* on the pRS305 backbone. A 950kb region to the 3' of *CEN2* was cloned into pRS306*tetO14* and integrated into *SynII* and BY4741 strains both expressing tetR-GFP, resulting in the generation of a GFP label 15 kb to the right of *CEN2*. Before measurement, alpha factor was first added in the medium to synchronize the *CEN2-GFP* strains in G1. Next, alpha factor was removed by filtration and washing with fresh medium to release *SynII* and wild-type cells from the G1 block into the cell cycle. Sample was collected every 15 mins for the next three hours and fixed for later analysis. The separation of sister chromatids at *CEN2* was scored by counting the percentage of cells in which two GFP dots were visible (Figure 3.7A). We found that *SynII* and BY4741 strains were comparable. The progression of cells from metaphase to anaphase was identified by changes in spindle morphology (Figure 3.7A). The overall ratio of metaphase to anaphase spindles further confirmed normal cell cycle progression in the *SynII* strain (Figure 3.7B).

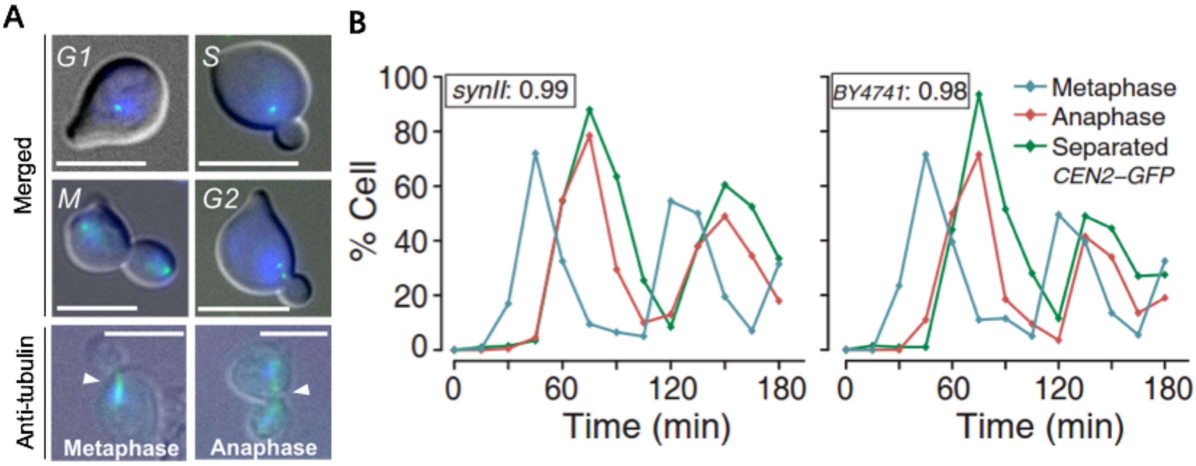


Figure 3.7 Cell cycle comparisons between *SynII* and BY4741.

A. Representative images showing cell morphology at different stages during the cell cycle after release from *G1* block. DNA staining is shown in blue; CEN2-GFP in *G1*, *S*, *G2* and *M* phase are shown in green. Spindle morphology in metaphase and anaphase was visualized by immunofluorescence (White arrow pointed). The scale bar represents 5 µm. B. Graphs showing the percentage of *SynII* cells with separated *CEN2-GFP* dots, metaphase spindles and anaphase spindles during the cell cycle. For each time point at least 200 cells were counted. The inset numbers indicate the overall ratio of metaphase to anaphase cells throughout the time course for *SynII* and *BY4741* strains.

Early activating replication origins are frequently found adjacent to tRNAs and transposable elements containing LTRs[106]. In *SynII*, two early origins are no longer linked with tRNAs and LTRs, as they were removed by design. Therefore, *SynII* offers an opportunity to test the functional requirement of linkages between these two early replication origins and their corresponding tRNAs and LTRs. The replication dynamics analysis for *SynII* done by Conrad A. Nieduszynski's group demonstrated that the synthetic chromosome showed identical replication dynamics to the wild-type chromosome *II* (Figure 3.8). This finding is consistent with the linkage between early activating origins, tRNAs and LTRs not being a functional requirement for early origin activation.

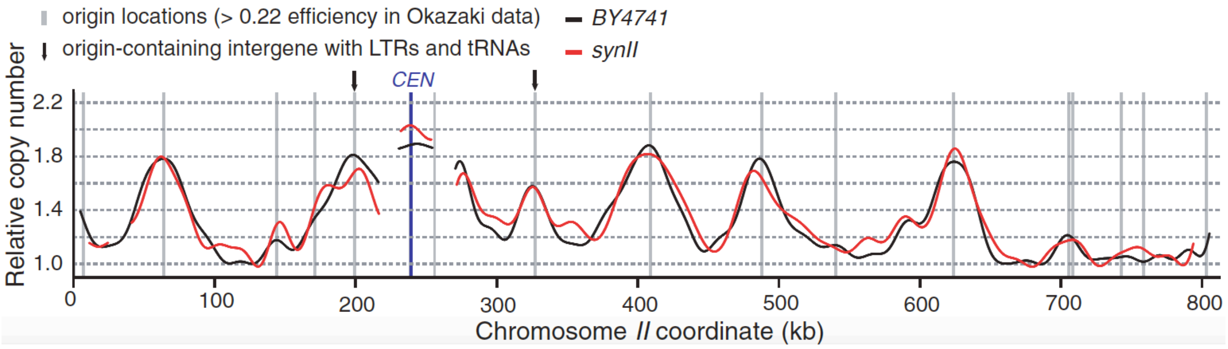


Figure 3.8 Replication profiles of *SynII* (red) and BY4741 (black)

Replication time is presented as relative copy number by deep sequencing.

Our results show that modifications in *SynII* show no gross negative effect on key cell cycle transitions; including chromosome replication (*S* phase) and segregation of sister chromatids (anaphase). In additional, Hi-C analysis of *SynII* performed by Romain Koszul's group also revealed no substantial changes between the synthetic and native chromatin, suggesting that the designed sequence has little to no negative effect on the average global folding of the chromosome[87].

3.1.3 Trans-Omics analysis of *SynII*

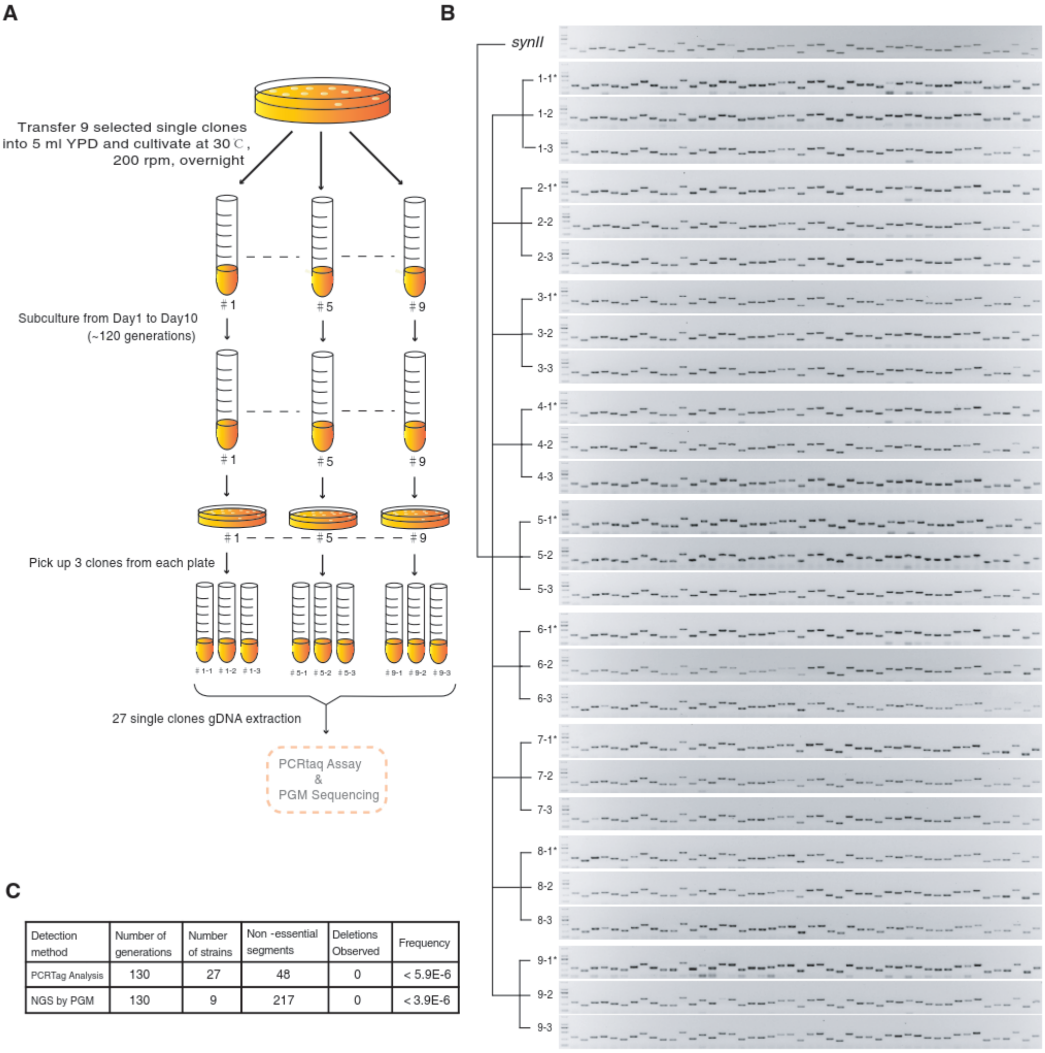
A systems biology approach ("Trans-Omics" analysis) was used here to evaluate the impact of *SynII* on the genomics, transcriptomics, proteomics and metabolomics of *S. cerevisiae*.

Genomic stability test of *SynII*

The insertion of 267 loxPsym sites in *SynII* could affect genomic stability. Therefore, we evaluated genome integrity and the frequency of chromosome segment loss in the absence of Cre expression. In total, 27 independent single colonies of *SynII* strain were selected after successive subculture of ~130 generations for PCRTag analysis. PCRTag analysis showed that no deletions were observed in all 27 independent isolates derived from over 130 mitotic generations from 9 independent lineages (Figure 3.9A,B). The overall loss rate was estimated lower than 5.9 x 10^{-6}. And in addition, 9 of the 27 single colonies were sequenced by the PGM sequencing platform. The deep sequencing analysis of 9 derived single strains showed that no mutation or genome rearrangement was observed, indicating that genome stability was faithfully maintained even after 100 generations of nonselective growth (Figure 3.9C).

Omics profiling of *SynII*

Transcriptome profiling identified only 18 out of 6,561 genes to have differential mRNA expression in comparison to the wild-type, with 7 up-regulated and 11 down-regulated (FDR < 0.01 and p-value < 7.62E-06, Figure 3.10A, Table 3.1). For the 18 identified differentially expressed genes, 6 are on *SynII*. However, KEGG and GO analysis did not find further evidence of these genes significantly enriched in particular functional pathways.

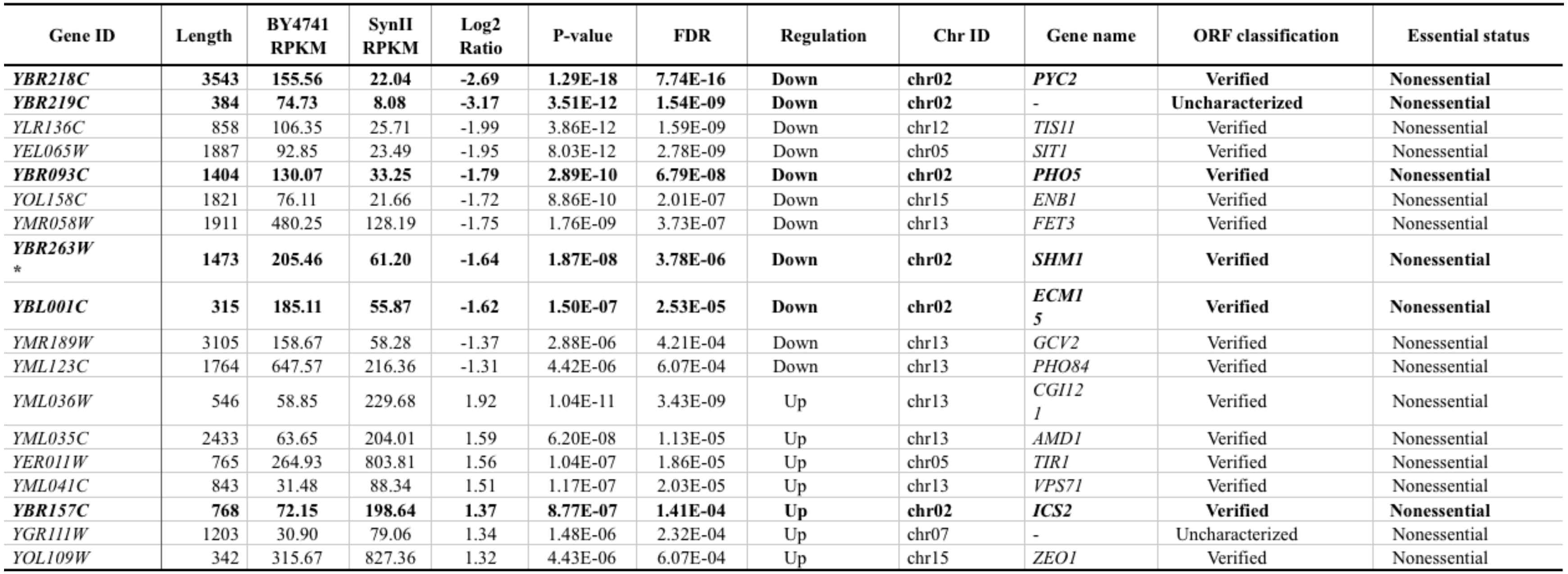


A
B
Transfer 9 selected single clones into 5 ml YPD and cultivate at 30 ℃, 200 rpm, overnight
#1
#5
#9
Subculture from Day1 to Day10 (~120 generations)
Pick up 3 clones from each plate
#1-1 #1-2 #1-3
#5-1 #5-2 #5-3
#9-1 #9-2 #9-3
27 single clones gDNA extraction
PCRtaq Assay & PGM Sequencing
synII
1-1*
1-2
1-3
2-1*
2-2
2-3
3-1*
3-2
3-3
4-1*
4-2
4-3
5-1*
5-2
5-3
6-1*
6-2
6-3
7-1*
7-2
7-3
8-1*
8-2
8-3
9-1*
9-2
9-3
C

Detection method	Number of generations	Number of strains	Non -essential segments	Deletions Observed	Frequency
PCRTag Analysis	130	27	48	0	< 5.9E-6
NGS by PGM	130	9	217	0	< 3.9E-6

Figure 3.9 Genome stability of the *SynII* strain

A. In total, 27 independent single colonies of *SynII* strain were selected after ~130 generations for PCRTag analysis, and in addition, 9 of the 27 single colonies were sequenced by the PGM sequencing platform. B. In total, 48 PCRTags (2 PCRTags located in nonessential genes from each megachunk) were chosen for PCRTag analysis to check the loss of different segments in the absence of SCRaMbLE. C. PCRTag analysis and PGM sequencing result show that without SCRaMbLE no losses were observed. Here frequency refers to the estimated maximum loss frequency per generation.



Gene ID	Length	BY4741 RPKM	SynII RPKM	Log2 Ratio	P-value	FDR	Regulation	Chr ID	Gene name	ORF classification	Essential status
YBR218C	**3543**	**155.56**	**22.04**	**-2.69**	**1.29E-18**	**7.74E-16**	**Down**	**chr02**	***PYC2***	**Verified**	**Nonessential**
YBR219C	**384**	**74.73**	**8.08**	**-3.17**	**3.51E-12**	**1.54E-09**	**Down**	**chr02**	**-**	**Uncharacterized**	**Nonessential**
YLR136C	858	106.35	25.71	-1.99	3.86E-12	1.59E-09	Down	chr12	*TIS11*	Verified	Nonessential
YEL065W	1887	92.85	23.49	-1.95	8.03E-12	2.78E-09	Down	chr05	*SIT1*	Verified	Nonessential
YBR093C	**1404**	**130.07**	**33.25**	**-1.79**	**2.89E-10**	**6.79E-08**	**Down**	**chr02**	***PHO5***	**Verified**	**Nonessential**
YOL158C	1821	76.11	21.66	-1.72	8.86E-10	2.01E-07	Down	chr15	*ENB1*	Verified	Nonessential
YMR058W	1911	480.25	128.19	-1.75	1.76E-09	3.73E-07	Down	chr13	*FET3*	Verified	Nonessential
YBR263W *	**1473**	**205.46**	**61.20**	**-1.64**	**1.87E-08**	**3.78E-06**	**Down**	**chr02**	***SHM1***	**Verified**	**Nonessential**
YBL001C	**315**	**185.11**	**55.87**	**-1.62**	**1.50E-07**	**2.53E-05**	**Down**	**chr02**	***ECM1 5***	**Verified**	**Nonessential**
YMR189W	3105	158.67	58.28	-1.37	2.88E-06	4.21E-04	Down	chr13	*GCV2*	Verified	Nonessential
YML123C	1764	647.57	216.36	-1.31	4.42E-06	6.07E-04	Down	chr13	*PHO84*	Verified	Nonessential
YML036W	546	58.85	229.68	1.92	1.04E-11	3.43E-09	Up	chr13	*CGI12 1*	Verified	Nonessential
YML035C	2433	63.65	204.01	1.59	6.20E-08	1.13E-05	Up	chr13	*AMD1*	Verified	Nonessential
YER011W	765	264.93	803.81	1.56	1.04E-07	1.86E-05	Up	chr05	*TIR1*	Verified	Nonessential
YML041C	843	31.48	88.34	1.51	1.17E-07	2.03E-05	Up	chr13	*VPS71*	Verified	Nonessential
YBR157C	**768**	**72.15**	**198.64**	**1.37**	**8.77E-07**	**1.41E-04**	**Up**	**chr02**	***ICS2***	**Verified**	**Nonessential**
YGR111W	1203	30.90	79.06	1.34	1.48E-06	2.32E-04	Up	chr07	-	Uncharacterized	Nonessential
YOL109W	342	315.67	827.36	1.32	4.43E-06	6.07E-04	Up	chr15	*ZEO1*	Verified	Nonessential

Table 3.1 List of genes with differential expression levels in *SynII* compared to BY4741

Genes on *SynII* are shown in bold. *Gene adjacent to telomere.

In the proteomics analysis, mass spectrometry was performed on *SynII* strain, with BY4741 strain as control. Mass spectrometry provided abundance data for 3,965 out of 6,682 protein-coding genes, and only 6 proteins showed substantial differential abundance in *SynII* strain compared with wild-type strain (Figure 3.10B, Table 3.2). 4 of the 6 proteins are expressed from genes on *SynII*: *YBR296C* protein expression level was doubled than that of wild type, while *YBL101C*, *YBR218C* and *YBR056W* protein expression level were down-regulated. Finally, metabolomics profiling was performed by metabolic LC-MS analysis. In total 4,941 and 6,417 mass spectra peaks with CV <30% were identified in positive and negative mode respectively, mapping to 1,032 unique metabolites, Fewer than 0.78% of the metabolites were differentially represented (Figure 3.10C,D). Potential differentially regulated metabolites included sphingolipids, glycerophospholipids, steroids and steroid derivatives (Table 3.3). Interestingly these are all molecules associated with the membranes.

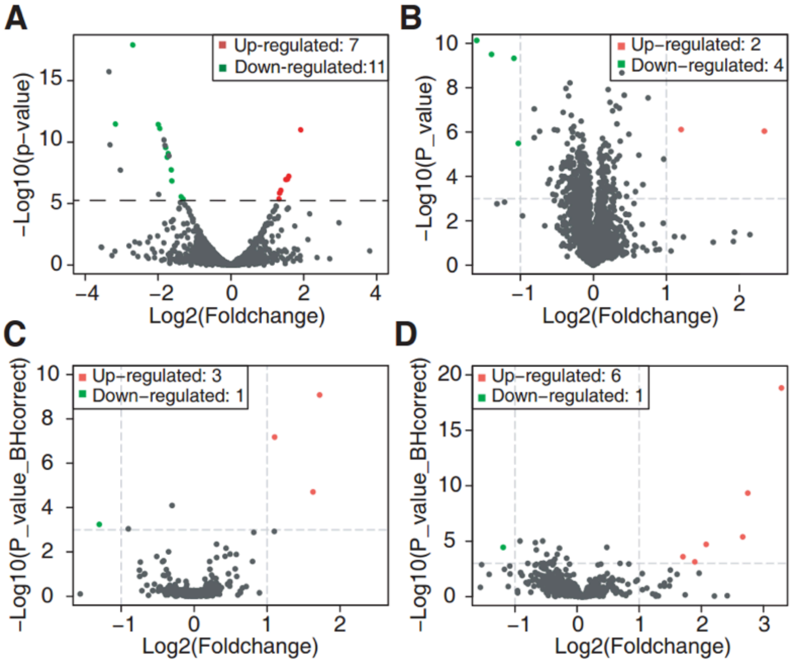
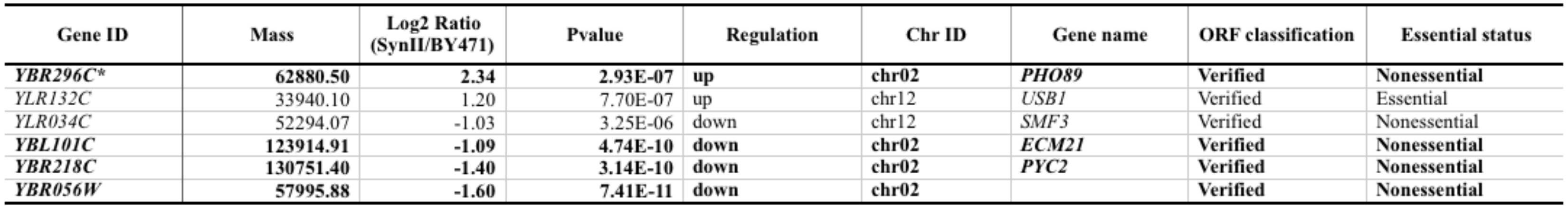


Figure 3.10 *SynII* strain Omics profile (BY4741 as reference)

A-D Identified dysregulated genetic features at (A) transcriptome level, (B) proteome level and (C and D) metabolome level (Metabolic and lipid profiling in LC-MS positive mode respectively) of *SynII* cells, compared to BY4741 cells. The total number of molecules with differential abundance expressed (p-value < 0.001) in transcriptome, proteome and metabolome are also presented as well. Up-regulated and down-regulated molecules are labeled in red and green respectively.



Gene ID	Mass	Log2 Ratio (SynII/BY471)	Pvalue	Regulation	Chr ID	Gene name	ORF classification	Essential status
YBR296C*	**62880.50**	**2.34**	**2.93E-07**	**up**	**chr02**	***PHO89***	**Verified**	**Nonessential**
YLR132C	33940.10	1.20	7.70E-07	up	chr12	*USB1*	Verified	Essential
YLR034C	52294.07	-1.03	3.25E-06	down	chr12	*SMF3*	Verified	Nonessential
YBL101C	**123914.91**	**-1.09**	**4.74E-10**	**down**	**chr02**	***ECM21***	**Verified**	**Nonessential**
YBR218C	**130751.40**	**-1.40**	**3.14E-10**	**down**	**chr02**	***PYC2***	**Verified**	**Nonessential**
YBR056W	**57995.88**	**-1.60**	**7.41E-11**	**down**	**chr02**		**Verified**	**Nonessential**

Table 3.2 List of genes identified from proteomics data with differential protein expression levels in *SynII* compared to BY4741

*Gene adjacent to telomere.

Data resource	Description	Mode l	Minimum CV (%)	Average abundance			Log2 Ratio	Formula
				QC	BY4742	*SynII*		
Metabolomics data	Indoleacetaldehyde	+	6.5	65747.3	33935.9	111779.1	1.72	C10H9NO
	PI(16:0/0:0)	+	19.2	29372.5	19574.7	60544.5	1.63	C25H49O12P
	4-Hydroxyphenylacetic acid	+	13.7	37866.2	26082.8	56030.9	1.07	C8H8O3
	Methyl 2-hydroxy benzoate	+	13.7	37866.2	26082.8	56030.9	1.07	C8H8O3
	Vanillin	+	13.7	37866.2	26082.8	56030.9	1.07	C8H8O3
	Anisic acid	+	13.7	37866.2	26082.8	56030.9	1.07	C8H8O3
	Ethyl-2-butenoate	+	10.4	51016.9	57177.6	23222.8	-1.32	C6H9O2-
	PI(16:0/0:0)	-	24.4	44249.4	49788.4	185260.6	1.81	C25H49O12P
	Engeletin	-	9.3	108556.1	74900.1	250574.7	1.72	C21H22O10
	PI(16:0/0:0)	-	7.4	718486.9	675483.1	1847842.0	1.43	C25H49O12P
	PI(18:0/0:0)	-	21.1	63194.7	101861.9	274244.0	1.43	C27H53O12P
Lipidomics data	3-Dehydrosphinganine	+	4.0	123286.8	17828.9	164754.6	3.29	C18H37NO2
	Sphingosine	+	4.0	123286.8	17828.9	164754.6	3.29	C18H37NO2
	3-Dehydrosphinganine	+	22.1	8283.2	1992.9	12877.8	2.74	C18H37NO2
	Sphingosine	+	22.1	8283.2	1992.9	12877.8	2.74	C18H37NO2
	palmitoleic acid	+	5.5	12136.7	809.4	15058.0	2.66	C16H30O2
	(2S,3S,4R)-2-aminohexadecane-1,3,4-triol	+	5.5	12136.7	809.4	15058.0	2.66	C16H35NO3
	ergosterol	+	7.3	34915.2	10362.1	40174.7	2.07	C28H44O
	3-dehydro-4-methylzymosterol	+	7.3	34915.2	10362.1	40174.7	2.07	C28H44O
	ergosta-5,7,24(28)-trien-3beta-ol	+	7.3	34915.2	10362.1	40174.7	2.07	C28H44O
	4beta-methylzymosterol-4alpha-carboxylic acid	+	17.5	20782.0	5883.2	20922.5	1.89	C29H46O3
	ergosterol	+	11.6	13925.6	4393.0	14396.8	1.72	C28H44O
	3-dehydro-4-methylzymosterol	+	11.6	13925.6	4393.0	14396.8	1.72	C28H44O
	ergosta-5,7,24(28)-trien-3beta-ol	+	11.6	13925.6	4393.0	14396.8	1.72	C28H44O
	PE(12:0/16:0)	+	8.5	54118.6	48559.0	20465.7	-1.32	C33H66NO8P
	PA(14:1(9Z)/16:0)	+	8.5	54118.6	48559.0	20465.7	-1.32	C33H63O8P
	PA(14:0/16:1(9Z))	+	8.5	54118.6	48559.0	20465.7	-1.32	C33H63O8P
	Uracil	-	12.4	6196.3	10482.1	4056.1	1.26	C4H4N2O2
	PI(18:0/0:0)	-	7.6	17987.7	6065.6	20421.4	-1.74	C27H53O12P

Table 3.3 List of metabolites with different levels in *SynII* compared to BY4741 from metabolomics data and Lipidomics data

Moreover, we also observed cellular defects that potentially arise from the Sc2.0 overall design principles. Compared to the WT strain, a subtle but potentially biologically significant up-regulation of genes with the GO terms “ribosome” and “cytoplasmic translation” was observed in the *synII* strain at both the transcriptome and proteome level (Figure 3.11, Figure 3.12A, B, C, D & E). Similar up-regulation was also observed in the *synV* and *synX* transcriptomes (Figure 3.13). It is known that deletion of multi-copy tRNA genes can lead to increased output of the translation machinery[107], thus we reasoned that the up-regulation of translational activity might be caused by the deletion of the 13 tRNA genes in *SynII,* which are all from multi-copy tRNA gene families. To test this we introduced an array containing all tRNA gene from chromosome II into *SynII* and found that the introduction of tRNA array to *SynII* greatly diminished the increase in translational functions (Figure 3.14).

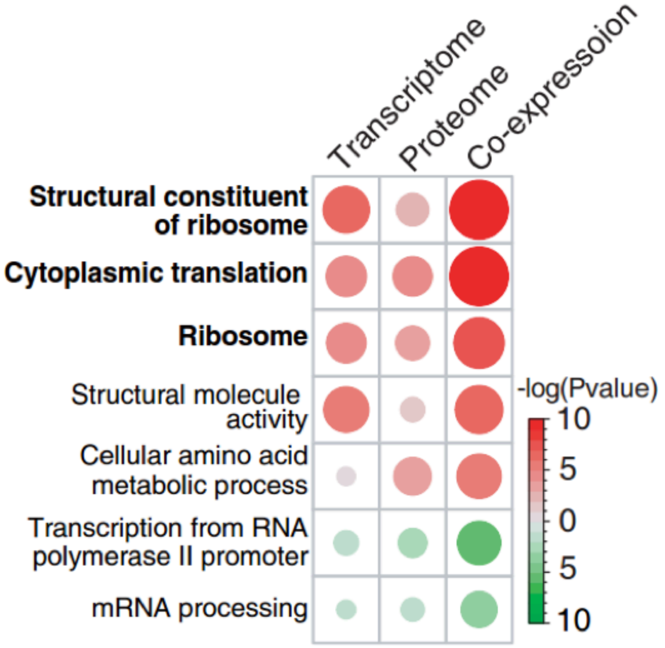


Figure 3.11 GO terms of the enriched pathways and the co-expression profile revealed by transcriptome and proteome analyses

Up-regulated features are labeled in red and down-regulated features are labeled in green circles. Significance level is indicated by color intensities and sizes of the circles.

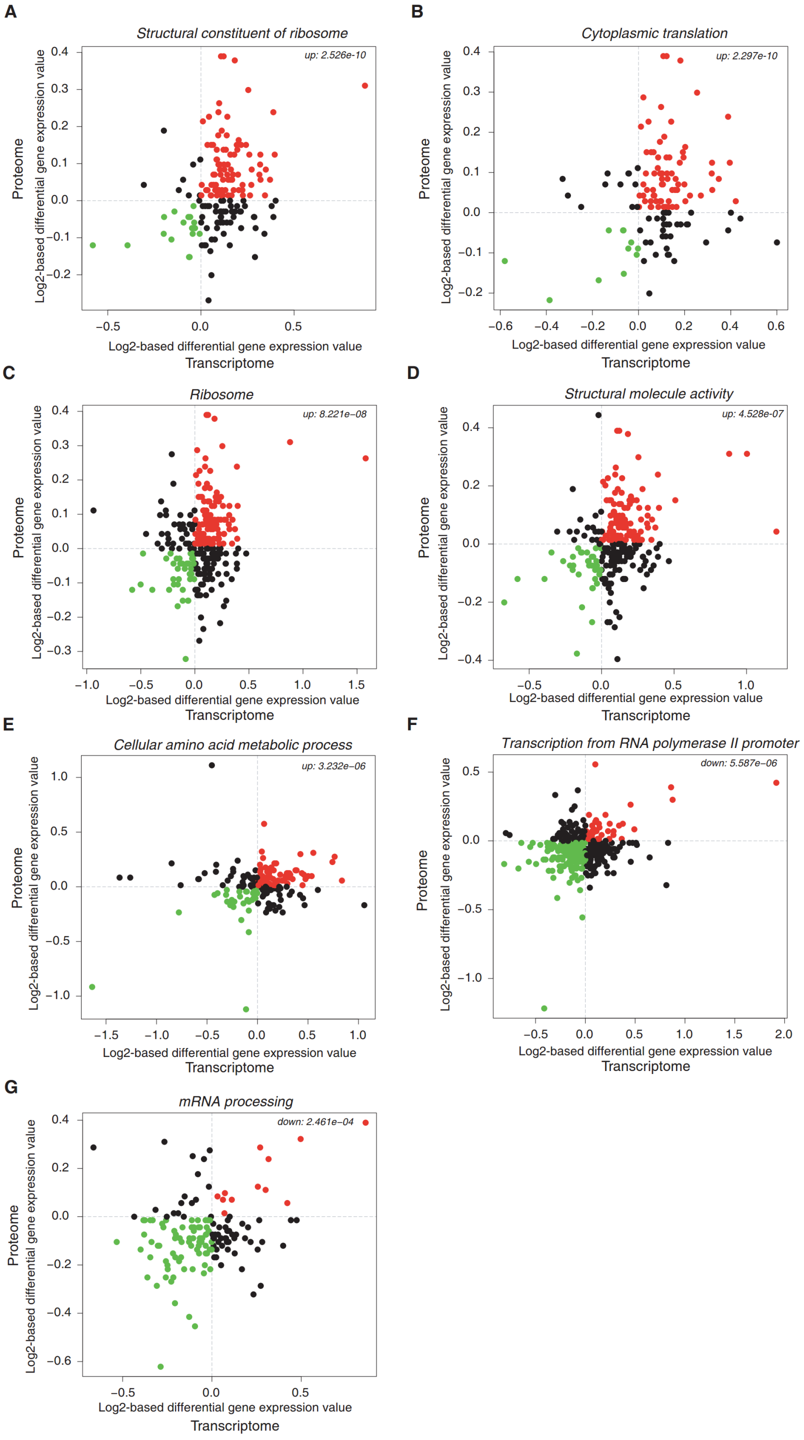


Figure 3.12 Correlation between the differentially expressed genes with different GO terms identified by transcriptomics and proteiomics analyses

X-axes and Y-axes are log2-based gene expression values of transcript level and protein level respectively. Within “Structural constituent of ribosome” (A), “Cytoplasmic translation” (B), “Ribosome” (C), “Structural molecule activity” (D) pathways, “Cellular amino acid metabolic process” (E) “Transcription from RNA polymerase II promoter” (F), and (G) “mRNA Processing”, genes that are significantly up-regulated (A-E) or down-regulated (F and G) are labeled in red and green respectively.

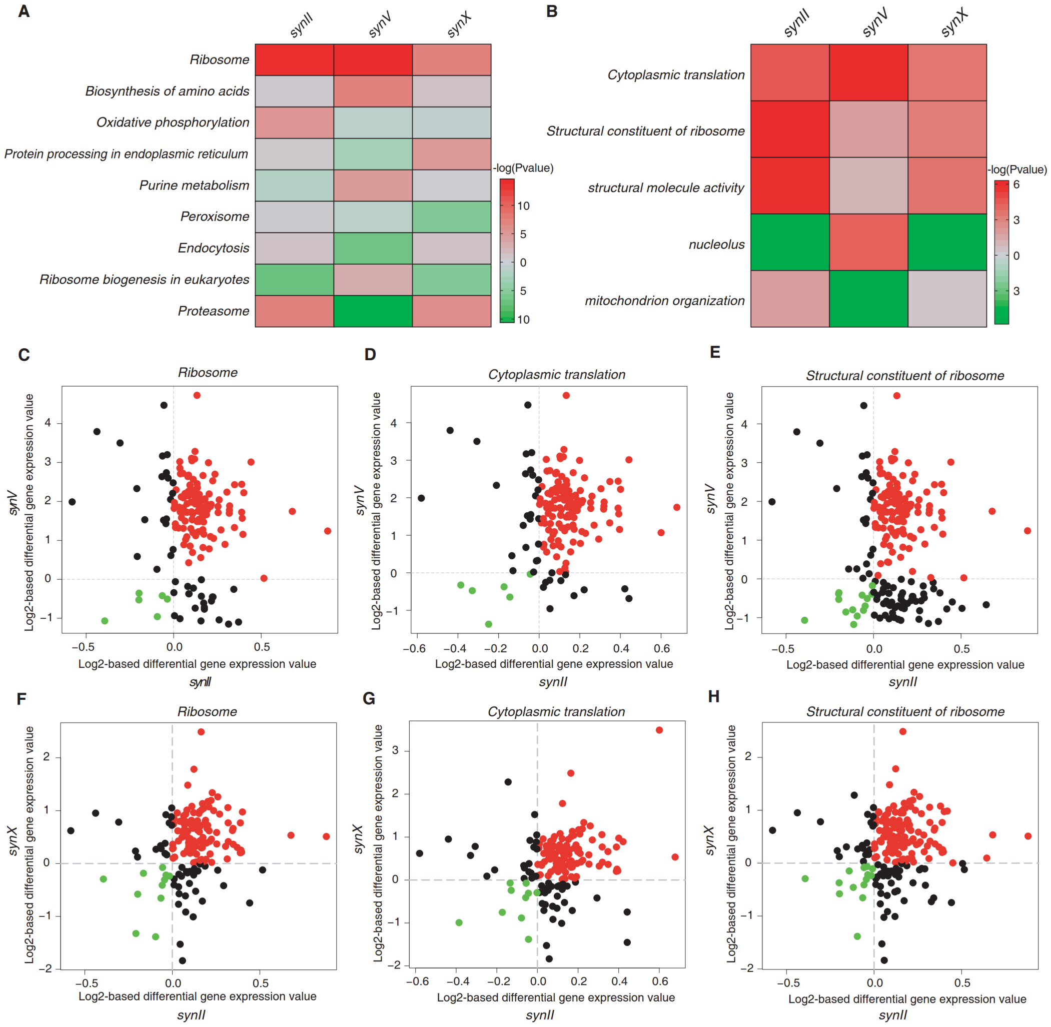


Figure 3.13 Enriched yeast KEGG pathways (A) and Go terms (B) with significant up-regulated and down-regulated genes of *SynII, SynV and SynX* strains compared to BY4741 control

Significance level is indicated by color intensity. (C-H) correlation analyses of differentially expressed genes categorized by GO terms between *SynII and SynV, or SynII and SynX* strains. The scatter plots indicate log2-based differential gene expression values of *SynII*, *SynV* and SynX strains. Within "Ribosome [sce03010]" (C, F), "Cytoplasmic translation [GO:0002181]" (D, G) and "Structural constituent of ribosome [GO:0003735]" (E, H), genes that are consistently up-regulated/down-regulated are labeled in red/green respectively.

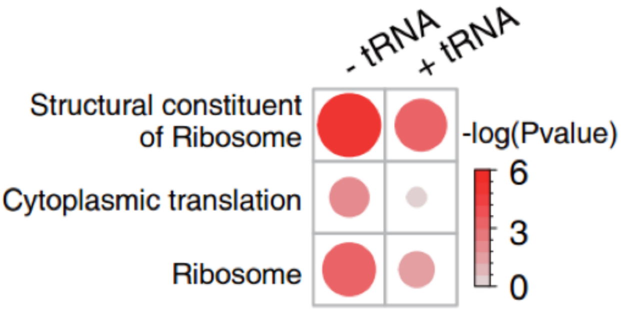


Figure 3.14 RNAseq analysis of *SynII* with and without tRNA array

By adding back the tRNA array of *SynII*, the increase in the translation is greatly reduced. Significance level is indicated by color intensities and sizes of the red circles.

In summary, despite some subtle differences, the Trans-Omics analysis provides clear evidence that the biological processes within the *SynII* strain are highly consistent with the wild-type strain. Therefore, the yeast genome displays a great degree of plasticity, and can readily cope with the large degree of editing encoded into *SynII*.

3.3 *SynVII*

3.2.1 Fitness defect rescue of *SynVII* by complementation assay

In contrast to *SynII*, the construction of *SynVII* proved to be more challenging. Several defects have been identified in *SynVII*, which are most likely introduced by the design features.

Defect mapping of *SynVIIS-Y* intermediate strain by SGA

The semi-synthetic right arm of *SynVII*, *SynVIIR*, was fully constructed after 11 runs of step-wise replacement with synthetic fragments. After a routine phenotypic check by growth curve analysis, *SynVIIR* was found to have a minor growth defect is that its lag phase is longer than wild-type cells (Figure 3.15). This defect was further confirmed by spotting assays in solid media (Figure 3.16).

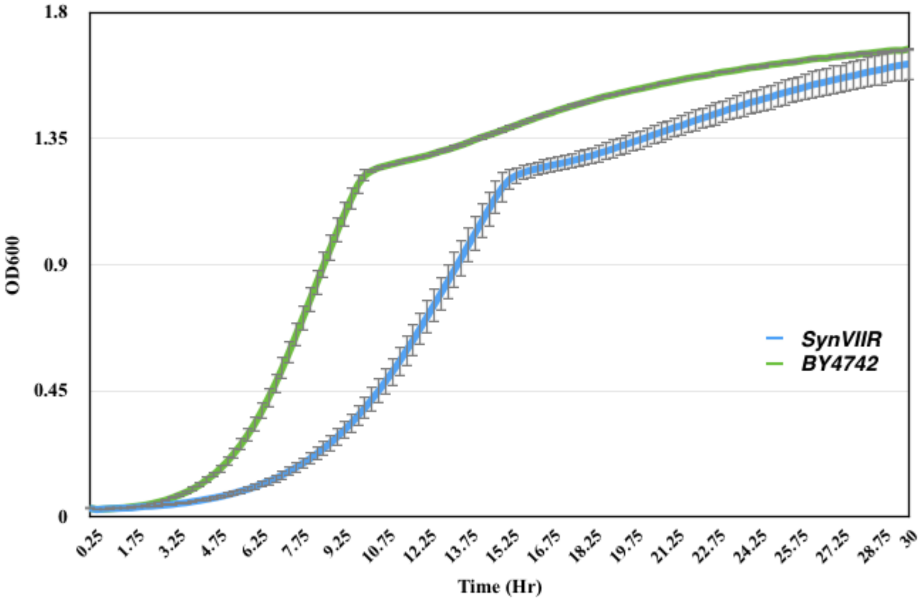


Figure 3.15 Growth curve analysis of *SynVIIR*

A 30-hours growth curve analysis in YPD medium was performed at 30℃ for *SynVIIR*, with wild-type strain (BY4742) as control. A longer lag phase was found in *SynVIIR* comparing to BY4742.

In order to pinpoint the defect origin, all 11 intermediate strains were recovered and spotted side-by-side and incubated at 30℃ for 4 days. The results revealed that the defect origin was within Megachunk S, as the defect starts to show in all intermediate strains after the Megachunk S integration (Figure 3.16). In total 23 genes were modified to incorporate different design features in Megachunk S (Table 3.4). Neither prior studies nor any supporting evidence was found that linked any of these genes to the observed defect. Therefore, the complementation assay that has been effective to rescue the fitness defect of *SynII* is not feasible here due to the fact that in total 49 design features, including PCRTag recoding, loxP site insertion and stop codon swap, are potentially involved (Table 3.4).

As the defect was found in the most common yeast culturing condition, indicating the potential responsible modification(s) altered genetic interactions which failed to maintain normal cellular functions[108]. This led to a hypothesis that if a single mutant (here refers to the defect causing gene) is responsible for the cell fitness defect, then the double mutant of the gene should also show the same phenotype.

Gene ID	Essential_status	ORF_classification	Design features
YGR154C	nonessential	verified	PCRTag, loxP
YGR155W	fast_growth	verified	PCRTag, loxP
YGR156W	essential	verified	PCRTag
YGR157W	nonessential	verified	PCRTag, loxP
YGR158C	essential	verified	PCRTag, Stop codon swap
YGR159C	fast_growth	verified	PCRTag, loxP
YGR160W	fast_growth	dubious	loxP
YGR161C	nonessential	uncharacterized	PCRTag, loxP, Stop codon swap
YGR161W-C	nonessential	uncharacterized	loxP
YGR162W	fast_growth	verified	PCRTag, loxP
YGR163W	nonessential	verified	PCRTag
YGR164W	nonessential	dubious	loxP
YGR165W	fast_growth	verified	PCRTag, loxP
YGR166W	fast_growth	verified	PCRTag, loxP
YGR167W	fast_growth	verified	PCRTag
YGR168C	nonessential	uncharacterized	loxP
YGR169C	nonessential	verified	PCRTag
YGR169C-A	nonessential	uncharacterized	loxP
YGR170W	nonessential	verified	PCRTag
YGR171C	fast_growth	verified	PCRTag, loxP
YGR172C	essential	verified	PCRTag
YGR173W	nonessential	verified	PCRTag, loxP, Stop codon swap
YGR174C	nonessential	verified	PCRTag, loxP, Stop codon swap

Table 3.4 Design features in *SynVII* Megachunk S

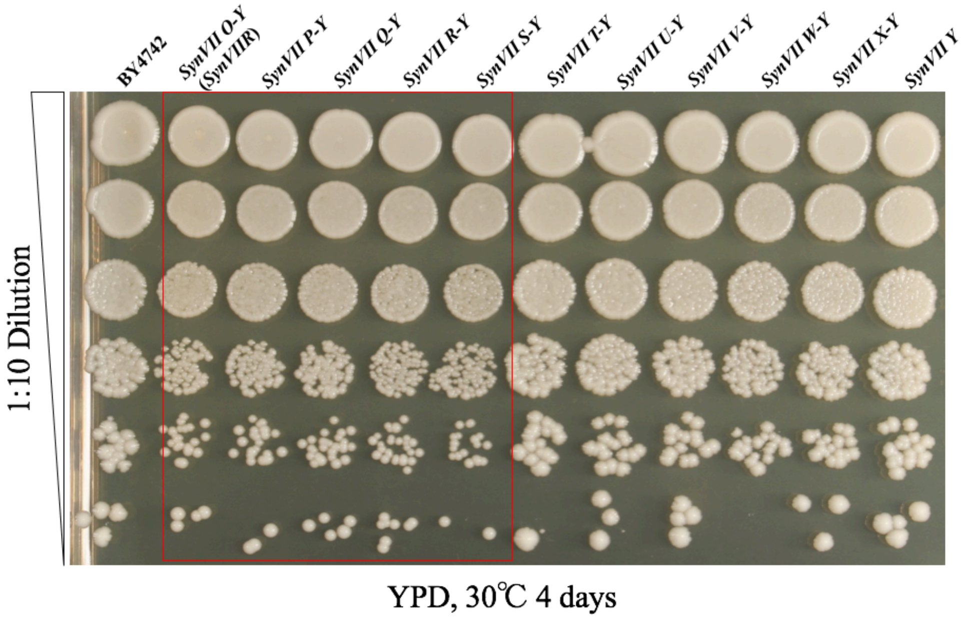
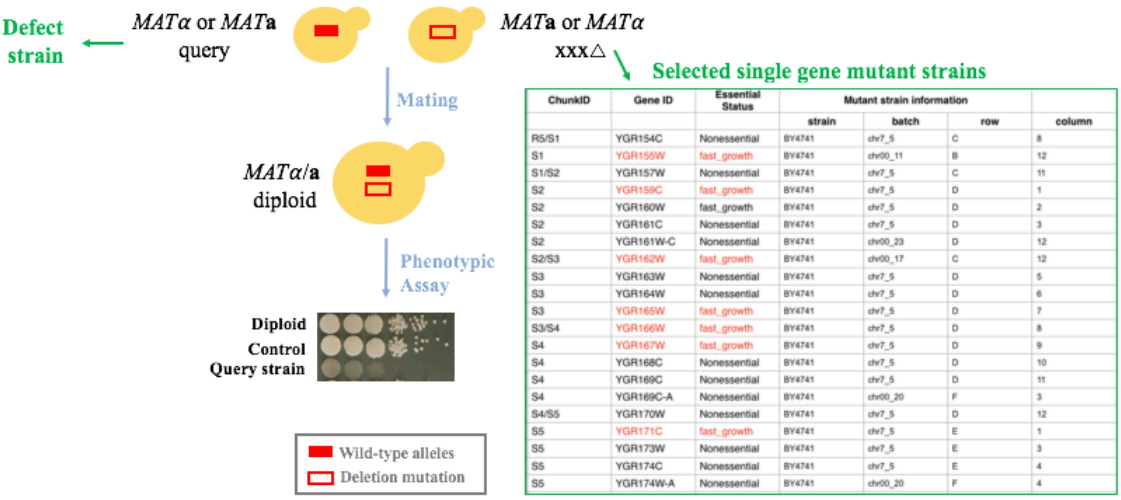


Figure3.16 Growth defect found in *SynVIIR* intermediate strains

With BY4742 strain as control, all intermediate *SynVIIR* strains were spotted on YPD and incubated at 30℃ for 4 days. A growth defect is detected in all strains containing synthetic Megachunk S.

A well-established method termed "synthetic genetic array" (SGA) analysis, in which a query mutation is crossed with a pre-designed yeast gene-deletion mutant library to identify of functional relationships between genes in a high-throughput manner[109,110]. Our hypothesis is that the synthetic design features such as PCRTag recoding, loxp site insertion and stop codon swap might lead to the functional dysregulation of genes containing synthetic design features. Therefore, the deletion mutant library offers a great opportunity to potentially identify the nature of the defect in synthetic Megachunk S . In Megachunk S, 20 of the 23 genes in Megachunk S have corresponding gene-deletions in the SGA library (the remaining 3 are essential genes). By mating these deletion strains individually with the defective *SynVIIS-Y* strain, the resulting diploid strain with one copy gene been deleted and another copy gene that cannot maintain proper function might also show the observed growth defect, and spontaneously lead us to the defect origin (Figure 3.17). On the other hand, even if the defect is rescued after mating, then it means our target(s) can be narrowed down to the 3 remaining essential genes.

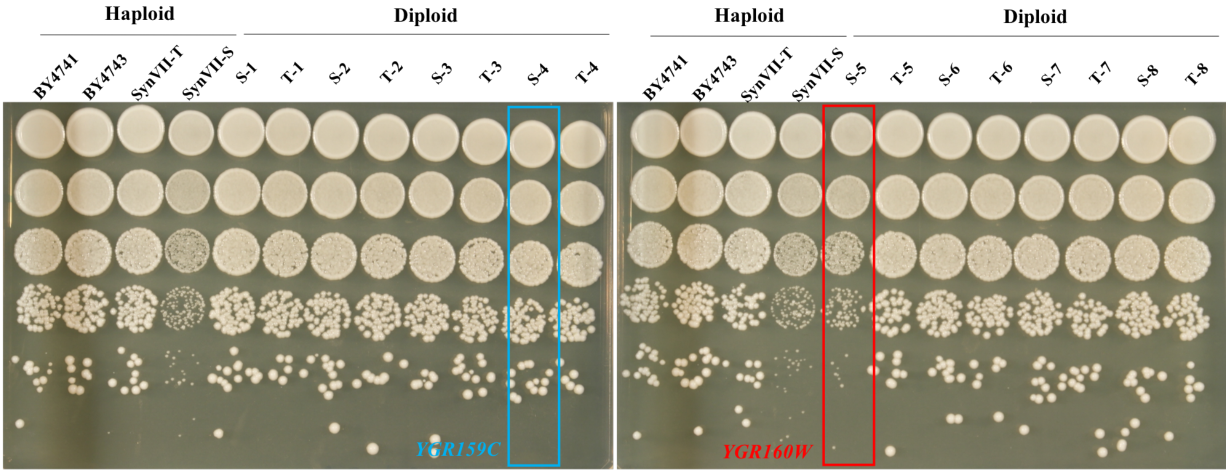


ChunkID	Gene ID	Essential Status	Mutant strain information			
			strain	batch	row	column
R5/S1	YGR154C	Nonessential	BY4741	chr7_5	C	8
S1	YGR155W	fast_growth	BY4741	chr00_11	B	12
S1/S2	YGR157W	Nonessential	BY4741	chr7_5	C	11
S2	YGR159C	fast_growth	BY4741	chr7_5	D	1
S2	YGR160W	fast_growth	BY4741	chr7_5	D	2
S2	YGR161C	Nonessential	BY4741	chr7_5	D	3
S2	YGR161W-C	Nonessential	BY4741	chr00_23	D	12
S2/S3	YGR162W	fast_growth	BY4741	chr00_17	C	12
S3	YGR163W	Nonessential	BY4741	chr7_5	D	5
S3	YGR164W	Nonessential	BY4741	chr7_5	D	6
S3	YGR165W	fast_growth	BY4741	chr7_5	D	7
S3/S4	YGR166W	fast_growth	BY4741	chr7_5	D	8
S4	YGR167W	fast_growth	BY4741	chr7_5	D	9
S4	YGR168C	Nonessential	BY4741	chr7_5	D	10
S4	YGR169C	Nonessential	BY4741	chr7_5	D	11
S4	YGR169C-A	Nonessential	BY4741	chr00_20	F	3
S4/S5	YGR170W	Nonessential	BY4741	chr7_5	D	12
S5	YGR171C	fast_growth	BY4741	chr7_5	E	1
S5	YGR173W	Nonessential	BY4741	chr7_5	E	3
S5	YGR174C	Nonessential	BY4741	chr7_5	E	4
S5	YGR174W-A	Nonessential	BY4741	chr00_20	F	4

Figure 3.17 Illustration of SGA-mediated debugging strategy

Defect carrying strain is served as the query strain and mated individually with single mutant strain library selected with gene of interest. The resulting diploid query strains are cultured on solid media for phenotypic assays.

To test the hypothesis, 20 gene-deletion mutants from the ~5000 yeast gene-deletion mutant library were picked and recovered. These mutant strains were mated with the *SynVIIS-Y* strain as well as a control *SynVIIT-Y* strain. The spotting assay revealed that the *SynVIIS-Y* /*YGR160W*-deletion diploid strain showed exactly the same phenotype comparing to *SynVIIS-Y* haploid strain (Figure 3.18). Therefore, modification of the *YGR160W* gene, which a loxp site was inserted in its 3'UTR region, is most likely the cause for the fitness defect observed.



YPD, 30°C 3 days

Figure 3.18 *SynVIIS-Y* and *SynVIIT-Y* diploid spotting

SynVIIS-Y and *SynVIIT-Y* were mated with selected BY4742 derived single gene mutants and spotted for phenotype verification, with wild-type haploid and diploid strains, and *SynVIIS-Y* and *SynVIIT-Y* haploid strains spotted on the side as phenotypic control. The *SynVIIS-Y/YGR159C*-deletion diploid strain and *SynVIIS-Y /YGR160W*-deletion diploid strain were highlighted in red and blue respectively. "S" labeled diploid strains are *SynVIIS-Y* mating strains with single gene mutant strains, "T" labeled diploid strains are *SynVIIT-Y* mating strains with single gene mutant strains. Number "1 – 8" are *YGR154C*, *YGR155W*, *YGR157W*, *YGR159C*, *YGR160W*, *YGR161C*, *YGR161W-C* and *YGR162W* single gene mutant strains respectively.

Interestingly, the YGR160W gene largely overlaps with the YGR159C coding region. Assuming the gene-deletion strategy used is replacing each of the ORFs in the yeast genome with a *KanMX4* cassette (details can be found from the *Saccharomyces* Genome Deletion Project website: http://www-sequence.stanford.edu/group/yeast_deletion_project/deletions3.html), the *YGR159C* double mutant is expected to show the same fitness defect. However, the *YGR159C* double mutant maintained a normal phenotype (Figure 3.18, blue highlighted). No further details were found regarding whether an alternative deletion strategy was used for genes with overlapping ORFs, therefore, sequencing of the *YGR159C* and *YGR160W* gene-deletion strains was performed. Results showed that the whole ORF of YGR160W was disrupted as expected. However, the whole ORF of YGR159C gene-deletion mutant remained intact, indicating that the general deletion strategy for the *Saccharomyces* Genome Deletion Project is not the same for overlapping ORFs. This explained why the *SynVIIS-Y/YGR159C*-deletion diploid strain did not show a fitness defect.

Both RNAseq analysis (Table 3.5) and rtPCR analysis show that the level of *YGR159C* mRNA in *SynVIIS-Y* strain is significantly higher than that in wild type strain (blue bar in Figure 3.19B), while YGR159C mRNA abundance in *synVIIT-Y* strain is similar to wild type (yellow bar in Figure 3.19B). Upregulated YGR159C gene expression suggests that modifications introduced by design is responsible for the phenotypic defect.

Gene ID	Gene ID	Fold change	P-value	Regulation
SynVIIS-Y/BY4741	NSR1	5.015113	3.55E-05	Up
SynVIIT-Y/BY4741	NSR1	0.695991	0.124162	-

Table 3.5 RNAseq analysis of YGR159C gene in *SynVIIS-Y* and *SynVIIT-Y* strains

YGR159C gene (gene ID: NSR1) with differential expression levels in *SynVIIS-Y* and *SynVIIT-Y* compared to BY4741 (p < 0.05).

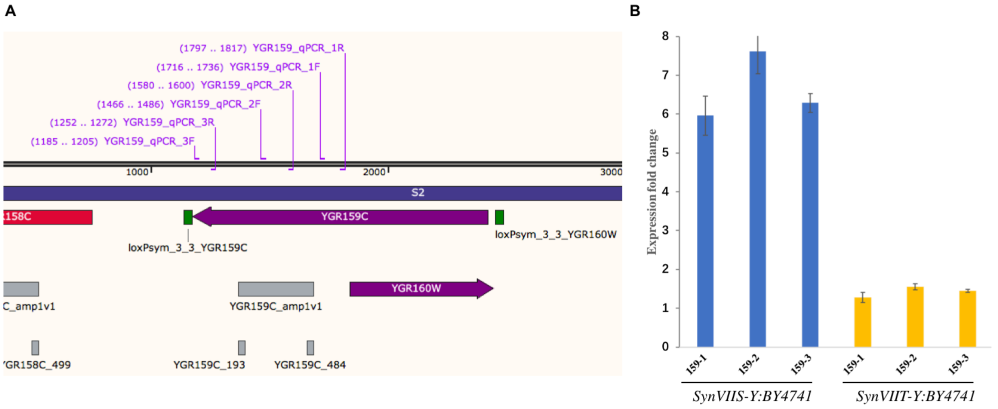


Figure 3.19 rtPCR analysis of YGR159C gene in *SynVIIS-Y* and *SynVIIT-Y* strains

A. Three sets of primers were designed for the rtPCR analysis of YGR159C gene. B. Relative expression levels of mRNA in *SynVIIS-Y* and *SynVIIT-Y* strains in comparing with native strain BY4741.

Although the role of each modification to the *YGR159C* gene in contributing to the defect is not entirely clear, SGA analysis and sequencing have narrowed down the potential defect targets from 49 to 4, and facilitate further defect troubleshooting by complementation assays.

Other defect mapping in the final *SynVII* strain

Another severe defect was found when a ~10 kb wild-type region in megachunk W was replaced with synthetic DNA in the final *SynVII* strain. Three yeast isolates were obtained that either partially or completely replaced the wild-type region with synthetic megachunk W (Figure 3.20A). However, growth curve analysis shows that all three isolates had severe growth defects (Figure 3.20B). 7 genes with a total of 23 design features are potentially responsible for this defect (Figure 3.21A). In order to find the cause of the defect, complementation assays with each gene was performed. Surprisingly, modifications of multiple genes were found to be responsible for this severe defect (Figure 3.21B), suggesting the modifications in this region interfere with function of multiple genes.

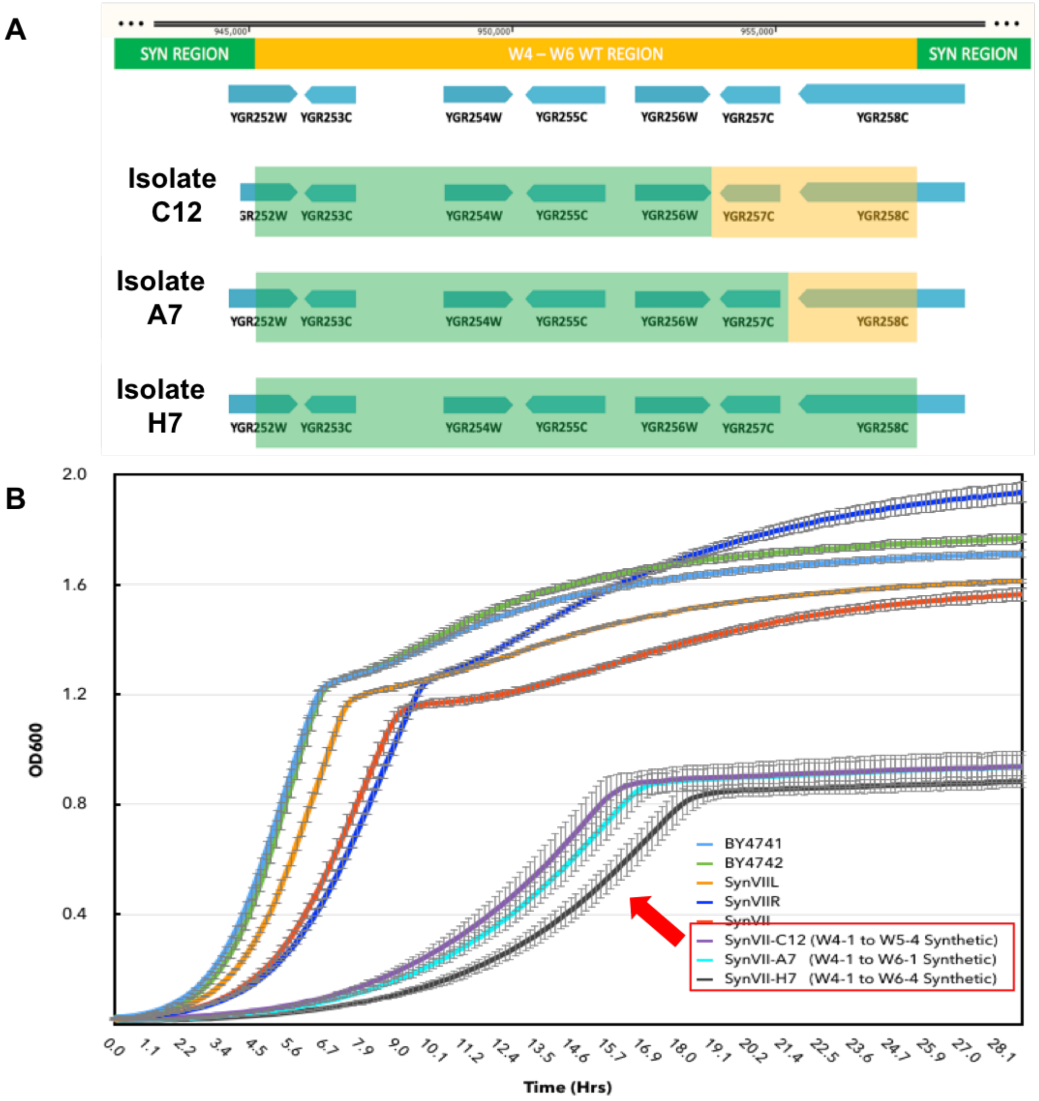


Figure 3.20 Growth defect of final *SynVII* strain caused by design features in megachunk W

A. The sequence map constructed by sequencing data of three *SynVII* isolates with partial or complete synthetic sequence in megachunk W. Green: synthetic region, Orange: wild type region, Blue: involving genes in this region. B. Growth curve analysis of the three isolates shows severe defects compared to wild-type strains, semi-synthetic SynVII strains or complete SynVII strain with the 10kb wild-type sequence in megachunk W.

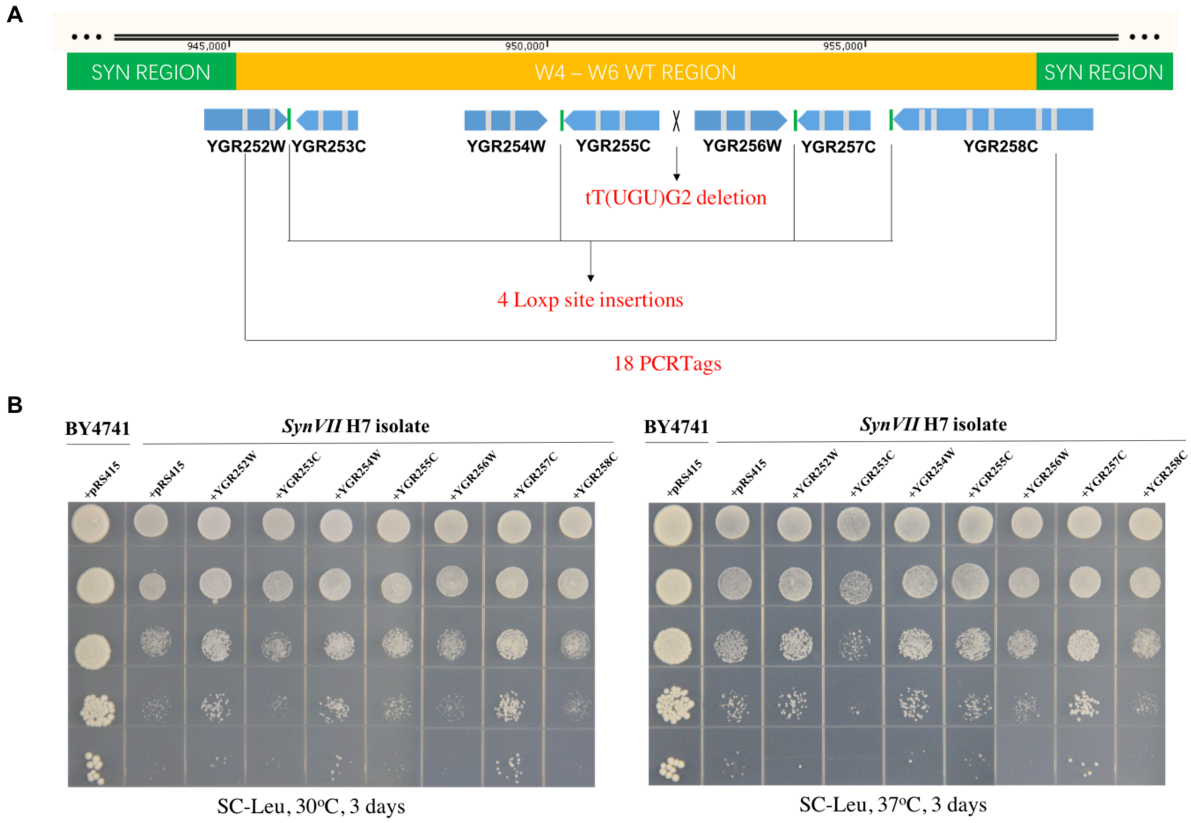


Figure 3.21 Design feature summary of defect causing region in megachunk W and corresponding complementation assay

A. Three types of design features were included in the defect causing region: tRNA deletion labeled as "X", PCRTag recoding (grey bars) and loxP site insertion (green bars). Wild type region are in orange. B. Complementation assay of each involved genes in the defect causing region. Each involved gene was amplified out from wild type strain and cloned into pRS415 vector and transformed into *SynVII* H7 strains containing full synthetic megachunk W for phenotypic assays.

3.4 Conclusion and discussion

To summarize, extensive Trans-Omics tests were conducted, including phenomics, transcriptomics, proteomics, chromosome segregation and replication analyses, which demonstrated that the *SynII* strain, despite significant sequence alterations, maintains a near wild type phenotype. However, our analyses reveal that 10% scale up-regulation of many components of the translational machinery (for example ribosomal proteins at both the RNA and protein levels) is a typical feature characteristic of synthetic chromosome replacement strains, and the main cause for this was shown to be the reduction in the tRNA gene copy number. Our comprehensive phenotypic assays of *SynII* strains also revealed the plasticity of the yeast genome. *SynVII* seems to be more sensitive to perturbations as much more severe growth defects were found related to certain design features introduced.

For each designer chromosome in Sc2.0 projects, thousands of changes have been introduced. It may be premature to suggest that the Sc2.0 design is too conservative, since strains hosting multiple synthetic chromosomes may have more profound phenotypic differences from the wild type due to the trans-interactions among designer features in different chromosomes. Growth rates determined under standard conditions would not be enough to reveal subtle changes. The Trans-Omics approach represents a powerful way to capture subtle changes at different levels. Therefore, as shown in this chapter, more in-depth analyses were introduced and established, e.g. monitoring cell replication and segregation, examining DNA, RNA, protein and metabolic profiles. For defect mapping, the modular construction approach implemented in Sc2.0 projects provides an important mechanism to systematically discover and repair phenotypic defects during construction and has greatly reduced the workload of debugging by narrowing down the target into one 30-50 kb megachunk region. The defect generated by PCRTag recoding in *YBR165W* gene of *SynII* was identified in such way. However, this approach loses its advantage when there are multiple potential defect origins involved, such as the severe growth defect found in *SynVII,* indicating genetic interactions occur when mutant alleles of two or more genes collaborate to generate an unusual composite phenotype. In this case, SGA-mediated defect mapping method was developed to address the challenge, as it enables systematical screening of potential targets in parallel. Interestingly, most defects were found to be caused by the PCRTag recoding (*SynII, SynV* and *SynXII*). Although the recoding was performed synonymously, it may have altered protein expression levels due to altered codon usage. Therefore, for future genome redesign it is imperative to take the codon usage bias into account.

In conclusion, we have demonstrated that the *SynII* strain segregates, replicates and functions (at the Trans-Omics level) in a very similar way to its wild-type counterpart which has naturally evolved over millions of years. On-going further investigations have been arranged for *SynVII* to eliminate all potential defects that might interfere with future studies.

3.5 Materials and methods

Growth Curve Assay

The growth curve analysis of BY4741, BY4742, and *SynII* was carried out using a Bioscreen C system (Oy Growth Curves Ab Ltd). Overnight cultures were sub-cultured into 300 μL of the different media to a final OD_{600} of ~0.1 in the 96 well micro-plates. The assay was performed in 3 replicates for each sample. Medium alone was used as the negative control.

Serial dilution assay on different medium plates

BY4741, BY4742, all intermediate and final *SynII* strains were inoculated in YPD medium at 30°C overnight. Then 1:10 serial dilutions were performed using these samples and plated onto different medium plates, which including: YPD, YPD with MMS (testing for DNA damage repair), YPD with benomyl (a microtubule inhibitor), YPD with Camptothecin (a topoisomerase inhibitor), YPD with Hydroxyurea (testing for defective DNA replication), YPD with various final concentrations of Sorbitol (0.5 M, 1 M, 1.5 M, 2 M) to cause osmotic stress, synthetic complete (SC) medium, SC with 6-Azauracil (testing for defective transcription elongation), YPEG (respiratory defects) containing 2% Glycercol and 2% Ethanol (testing for respiratory defects), and YPD adjusted to pH 4.0 and pH 9.0 with HCl and NaOH (testing for vacuole formation defects). Two specific drugs, Hydrogen peroxide (testing for oxidative stress) and Cycloheximide (testing for defective protein synthesis), were used to treat the cells for two hours by directly adding them to overnight cultures. Yeast cells were collected by centrifugation and resuspended in water before serial dilution and plating. Plates were incubated at 25°C /30°C /37°C for 2/3/4 days.

Cell Morphology

Cells were grown to log phase in YPD at 30°C. DIC Images were collected using a Nikon microscope Ti-E (100X) with an Andor Zyla 5.5 camera.

CEN2-GFP Strain Construction and Culture

To create the *CEN2-GFP* strain, a 950 bp region to the right of *CEN2* was cloned into *AMp327*[105] and integrated into the *SynII* strain and BY4741 carrying *tetR-GFP*, generating a GFP label 15 kb to the right of *CEN2*. Cell cycle synchronization of *CEN2-GFP* strains was performed by adding alpha factor to YEP medium. Alpha factor was removed by filtration and washing with fresh YEP medium to release cells from G1 block into the cell cycle.

Microscopy

Methods of fixing cells for GFP-labeled chromosome visualization and indirect immunofluorescence were previously described by Fernius et al.[105]. α-tubulin was visualized using a rat anti-α-tubulin antibody at a dilution of 1:50 in PBS/BSA and an anti-rat FITC antibody at a dilution of 1:16.67 in PBS/BSA. Microscopy was performed on a Zeiss Axioplan 2 microscope and images were captured using a Hamamatsu camera operated through Axiovision software. To score GFP dots and spindles in mitosis, cells were co-stained with DAPI to visualize DNA morphology. For each sample obtained at each 15-min time point, 200 cells were counted in the field to score GFP dots and spindles. EC Plan NeoFluar objective lens (x63, Oil, numerical aperture of 1.25) was used. Camera pixel size is 6.3 pixels/μm.

SynII Replication Timing Profiles Analysis

Replication timing profiles were generated from asynchronous cultures using fluorescence-activated cell sorting technique been established previously[111]. Asynchronous cells were fixed with 70% ethanol, treated with RNaseA and Proteinase K and stained with 10 × SYTOX® green nucleic acid stain (Invitrogen). At least 30 million cells were sorted from a particular cell cycle stage using a MoFlo Sorter (Coulter Beckman). Sorted cells were treated with Zymolase, RNaseA and Proteinase K. DNA was purified using phenol-chloroform extraction and ethanol precipitation. Single end Illumina HiSeq2500 sequencing yielded a minimum of 29.6 million uniquely mapped reads per sample. The ratios between uniquely mapped reads from the non-replicating and the replicating samples were calculated for every 1000 bp window. The ratios were normalized to control with a baseline of one for differences in the number of reads between the samples.

Genome Stability Analysis

SynII was streaked on YPD plate and incubated at 30°C for 2 days. Nine single colonies were selected for successive subculture in YPD medium for ~130 generations, followed by plating on YPD plates overnight. Three single colonies of each initial isolate were selected. Genomic DNA preparation was performed for all 27 isolated single colonies. 48 pairs of PCRTags were chosen to perform PCRTag assay for all 27 isolated single colonies. Sequencing on the PGM platform was conducted for 9 of the 27 isolated single colonies (one from each initial isolate), with library preparation and sequencing method as described in “structural variant detection with PGM sequencing platform” in Materials and Methods of Chapter Two. Sequence data was analyzed according to the method described in “*Nucleotide Sequence Analysis of SynII with Hiseq2500 sequencing platform*” in Materials and Methods of Chapter Two.

Yeast Total RNA Isolation for RNA Sequencing

3 biological replicates of both *SynII* and *BY4741* strains were cultured overnight in 3 mL YPD medium at 30°C. The cultures were added to 10 mL fresh YPD medium and incubated until the OD_{600} reached ~0.8. The cells were harvested by centrifugation at 900 rcf for 5 min. Total RNA was isolated using the RiboPure-yeast kit (Ambion) according to the manufacturer’s instructions.

RNA-seq Analysis of *SynII*

The 330 bp cDNA libraries were prepared according to the manufacturer’s instructions (Illumina Inc.) and were paired-end sequenced using the Illumina Hiseq2500. Raw reads were filtered using the following criteria: no N bases, no adaptor sequences, minimum read length 100, bases of low quality (<10) was no more than 1% in a read. An average of 7.0 M and 8.6 M clean reads were obtained from BY4741 and *SynII* samples respectively, with 95.7% and 95.2% mapped to the corresponding genome reference. Clean reads were mapped to genomes by tophat v2.0.10[112], with the parameter –r (--mate-inner-dist) of 130. After reads counting, RPKM was calculated for each gene. Differential gene expression was analyzed by DEseq v1.20.0[113], with the no replicates scenario (parameters were: --method blind, --sharingMode fit-only, --fittype, local). For each gene, a raw p-value and adjusted p-value using the Benjamini–Hochberg procedure were obtained. Genes were assessed for statistical significance by rejecting the null hypothesis if the adjusted p-value (FDR) was <0.01 and if the raw p-value fell below the threshold of the 5% Family Wise Error Rate (FWER) after Bonferroni correction (threshold=7.62E-06). False positive results were inferred by the following rules and removed: dubious genes, transposable genes, genes with low coverage (<60%) for native and synthetic strains.

Proteome and Metabolome Analysis of *SynII*

3 biological replicates of both BY4741 and *SynII* strains were cultured overnight in 3 mL YPD medium at 30°C and re-inoculated in 10 mL fresh YPD medium for further incubation until the OD_{600} reached ~0.8. The cells were collected for both proteome and metabolome analysis. Proteins from the yeast cell were extracted with Urea, reduced, alkylated, digested with trypsin and iTRAQ labeled-

BY4741 (113, 115, 117, 121) and *SynII* (114, 116, 118 and 119) (AB SCIEX, Framingham, MA, USA). After labeling, the peptides were fractionation with the SCX method and analyzed by an Orbitrap Q Exactive mass spectrometer (Thermo Fisher Scientific, San Jose, CA) coupled with an online HPLC. Mascot and IQuant[114] software was used for protein identification and quantification. For the global metabolomics analysis, metabolites were extracted with buffer (50% methanol and 50% water), separated with a BEH C18 column. For the lipidomics, metabolites were extracted with buffer (75% dichloromethane and 25% methanol), separated with a CSH C18 column. The metabolites were detected by a XEVO-G2XS QTOF mass spectrometer (Waters, Manchester, UK) and the raw spectrums were processed by Progenesis QI 2.0 software (Nonlinear Dynamics, Newcastle, UK) for peak picking, alignment, normalization and identification. Further statistical analysis was performed on the resulting normalized peak intensities using in-house developed software metaX. Pathway analysis was conducted using the MetaboAnalyst pathway tool[115].

Yeast KEGG pathway and GO enrichment

To identify the differences in biological processes between *SynII* and BY4741 strains, gene enrichment and co-expression enrichment analyses were performed using yeast KEGG pathways and yeast Gene Ontology (GO) annotations using transcriptomics, proteomics and metabolomics data respectively. Genes and metabolites with differential expression of log2 (fold-change) >0 and down-regulation of log2 (fold-change) < 0 were considered as up-regulated and down-regulated for enrichment analysis. The significance of each KEGG pathway and GO term in genes, metabolites, as well as their co-expressions was individually identified using the hyper-geometric test and Chi-squared test with false discovery rate (FDR) correction and the threshold P-value < 0.001. The KEGG category and interaction between genes and metabolites were classified based on the Yeast Metabolome Database[116] (http://www.ymdb.ca/system/downloads/current/ymdb.json.zip).

Mating type verification by yeast colony PCR

Three primers were designed that can specifically amplify the corresponding mating type sequence (primers YCp3517, YCp3518 and YCp3519), generating a 404bp amplicon for *MATα* type strain and a 544 bp amplicon for *MAT***a** type strain. Cell culture is pre-treated with 40nM NaOH at 95°C for 15 mins. Then the supernatant was used as template for PCR. PCR program setting is as follows: 95°C/3 min, 32 cycles of (95°C/30sec, 55°C/1min 30sec, 72°C/30sec), and a final extension of 72°C/7 min. Visualization of PCR products was carried out by gel electrophoresis. 12 uL of each PCR product was loaded on 2% agarose gel, running at 180V for 20 min.

Chapter Four

Part 3: SCRaMbLE - to explore the potential of *SynII* and *SynVII* by applying SCRaMbLE

4.1 Introduction

As one of the key features of Sc2.0 project, the SCRaMbLE system, which is based on a chemically inducible Cre recombinase[16,117], was incorporated to permit global genome rearrangements. The loxP site introduced in synthetic yeast genome is symmetric (hereinafter referred to as loxPsym site), thereby the combination of each two loxPsym sites can generate a diverse population of cells with a selectable phenotypic diversity. Theoretically, the “build-in” tool SCRaMbLE holds the potential to allow accelerated evolution of the synthetic yeast strains. It can be used for basic research to gain knowledge about structure and influence of chromosomal architecture for the phenotype as well to generate miniaturized yeast strains. On the other hand, it can be used to optimize yeast strains for biotechnological applications in a faster, more complex and directed way compared to standard random mutagenesis experiments which rely mostly on small deletions and point mutations. In this chapter, the workflow of genome sequence reconstruction of SCRaMbLEd *SynIXR* strains using deep sequencing data will be described first followed by the discussion of using SCRaMbLE for fitness defect debugging of *SynII* and aneuploid *SynVII*.

4.2 Genome sequence reconstruction of SCRaMbLEd strains with deep sequencing data

While initial applications of SCRaMbLE have been promising[117], many potential challenges exist. First, introducing multiple loxPsym sites may induce genome instability even in the absence of Cre recombinase; after Cre induction has been shut off, leaky Cre expression or continuing Cre activity may lead to instability. When Cre is active, recombination between loxPsym sites and off-target sites, albeit at extremely low frequency, may occur[118]. For desired recombination at loxPsym sites, random pairing is desirable to obtain maximum diversity. However, the 82 bp minimum distance required for loxP recombination[119] and emergent recombination hotpots may further reduce diversity of SCRaMbLEd products. Detailed characterization of the genomes resulting from SCRaMbLE is required to answer these questions, however, the genome rearrangements generated by SCRaMbLE may not be amenable to standard genome sequencing and assembly methods. Therefore, it is important to develop alternative method(s) that allow accurate analysis of SCRaMbLEd genomes. To this end, 63 previously isolated circular-*SynIXR* SCRaMbLEd strains from Dymond, et al. [16]were used to establish the genome reconstruction workflow as well as to detect any unexpected ectopic recombination event. The 63 *SynIXR* SCRaMbLEd strains were generated by Cre induction and then selected for auxotrophies arising from loss of function of *LYS1* or *MET28* encoded in the synthetic region. 33 lys$^-$, 20 met$^-$ and 10 lys$^-$ met$^-$ strains were isolated.

Together with one wild-type strain BY4741 and two non-SCRaMbLE parental strains, all 63 SynIXR SCRaMbLEd strains were sequenced. For each strain, short-insert libraries with average insert size of 500 bp were generated without PCR amplification to avoid creating artefacts through crossover PCR at loxPsym sites. For ten strains whose rearrangements were too complex to be resolved with short inserts (see details in section 4.1.3 "Full sequence reconstruction"), long-insert libraries with an average 10 kb insert size were prepared; long-insert library generation was less efficient and required a standard PCR amplification step[120]. Paired-end sequencing for all libraries was performed on the Illumina HiSeq 2000 platform. After stringent quality filtering, short-insert sequences with 37-fold sequence coverage on average were obtained ready for analysis.

4.2.1 Nomenclature of segments and junctions

SCRaMbLE is designed to generate diversity by combinatorial rearrangement of segments flanked by designed recombination sites. The original segments are represented as consecutive integers, 1 through 43 for *SynIXR* (Figure 4.1). The loxPsym junctions are denoted by the unique left ("L") and right ("R") ends connecting the two segments. For example, the loxPsym site at the junction between segments 1 and 2 involves half-sites 1R and 2L. After the SCRaMbLE process, the rearranged chromosome is represented using standard gene order conventions as a list of the segments in their new order, with junctions connecting adjacent segments, and an additional junction between the final and initial segment in the circular-*SynIXR* chromosome. Deletions and duplications change the number of times that a segment appears, and inversions reverse the order of the affected region.

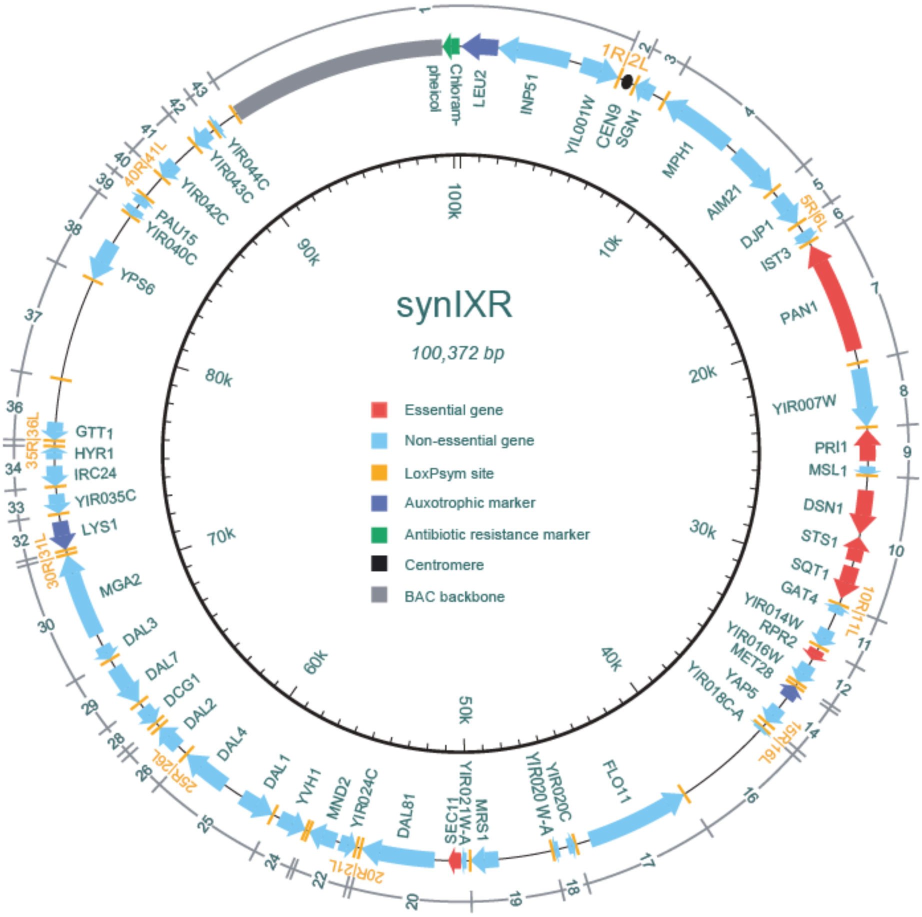


Figure 4.1 The circular SynIXR synthetic chromosome

SynIXR chromosome comprises 43 segments, numbered consecutively as shown, with loxPSym sites (orange bars) serving as junctions between adjacent segments. Arrows indicate genes, with colors denoting essential genes (red), auxotrophic markers (purple), other non-essential genes (light blue), and the chloramphenicol resistance marker (green). The centromere, in segment 2, is shown as a black circle. The left and right ends of each segment are denoted as "L" and "R", respectively. And loxPsym sites are denoted by the unique half-sites of the segments they join.

With junction and segment defined, an in-house software pipeline was developed to efficiently identify recombination breakpoints and other sequence variation relative to the parental strain (Figure

4.2). Reads with loxPsym sequence were analyzed separately for direct evidence of recombinations and classified as "parental junctions", "novel junctions" involving a recombination between pairs of designed loxPsym sites, or "off-target junctions" involving the recombination of a designed loxPsym site with any off-target sequence. Recombinations not involving loxPsym sites were termed "ectopic rearrangements". Novel junctions and copy number variation were then used to classify underlying events as deletions, inversions, insertions, tandem and non-tandem duplications, and more complex rearrangements.

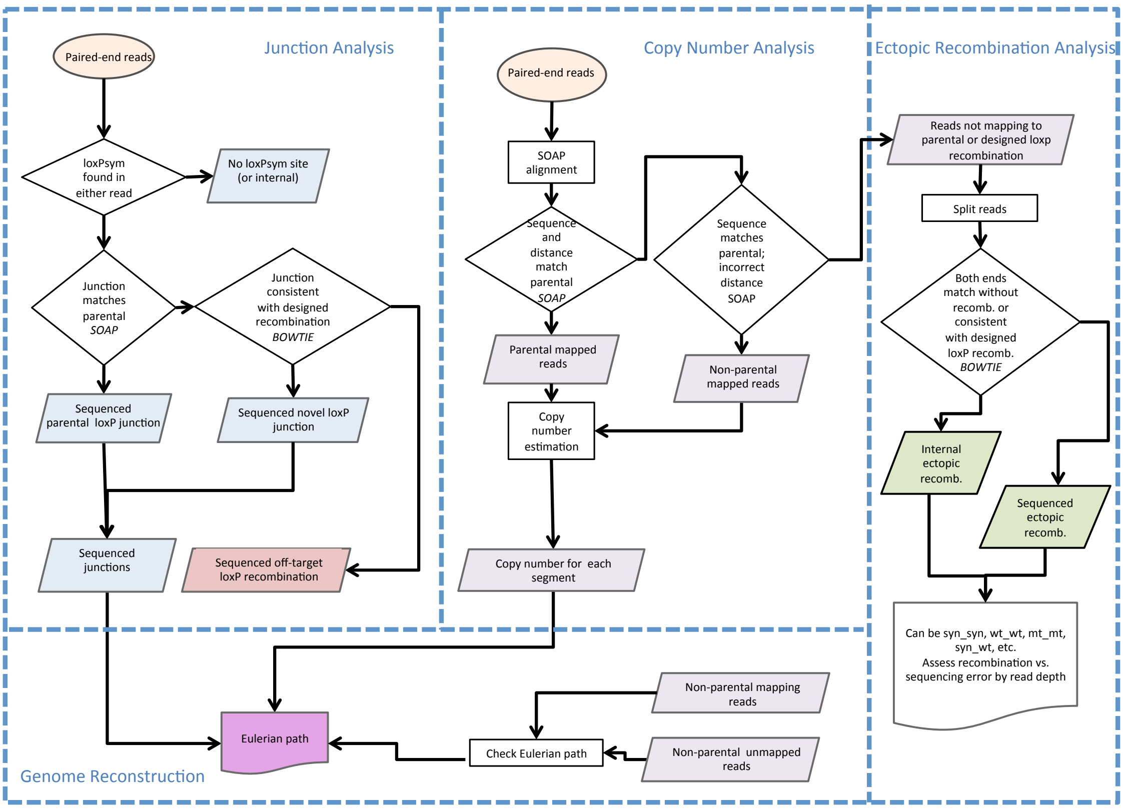

Junction Analysis
Paired-end reads
loxPsym found in either read
No loxPsym site (or internal)
Junction matches parental SOAP
Junction consistent with designed recombination BOWTIE
Sequenced parental loxP junction
Sequenced novel loxP junction
Sequenced junctions
Sequenced off-target loxP recombination
Copy Number Analysis
Paired-end reads
SOAP alignment
Sequence and distance match parental SOAP
Sequence matches parental; incorrect distance SOAP
Parental mapped reads
Non-parental mapped reads
Copy number estimation
Copy number for each segment
Ectopic Recombination Analysis
Reads not mapping to parental or designed loxp recombination
Split reads
Both ends match without recomb. or consistent with designed loxP recomb. BOWTIE
Internal ectopic recomb.
Sequenced ectopic recomb.
Can be syn_syn, wt_wt, mt_mt, syn_wt, etc.
Assess recombination vs. sequencing error by read depth
Genome Reconstruction
Eulerian path
Check Eulerian path
Non-parental mapping reads
Non-parental unmapped reads

Figure 4.2 SCRaMbLE reconstruction pipeline

Paired-end reads were aligned to identify both parental and novel junctions, copy number of each segment and ectopic recombination events. By combining the analysis of junction and copy number, Eulerian path was applied for the genome reconstruction.

4.2.2 Analysis of off-target and ectopic recombination

The first question addressed was: Do Cre generated semi-ectopic recombinations between native loxPsym sites and cryptic loxP sequences occur in the yeast genome? Such events have been reported to occur albeit at an extremely low frequency[118]. No evidence for such events was detected in the 63 SCRaMbLEd isolates. Across the 63 SCRaMbLE isolates, all 612 distinct novel junctions involved designed loxPsym sites. The average read depth of 21.5 for these novel junctions was not appreciably different from the average of 21.2 for the 2384 sequenced parental junctions. In contrast, the maximum read depth for any ectopic loxPsym recombination was only 2, suggesting these reads represent ligation artefacts rather than real recombination events.

Next, fully ectopic recombinations not involving loxPsym or loxP relevant sequences was examined. The average read depth for non-synthetic nuclear chromosomes was 36.1, greater than the read depth for *SynIXR* and possibly reflects greater recovery of linear vs. circular chromosomes. Previous studies using comparative genome hybridization and quantitative PCR also found reduced recovery of *SynIXR*[16]. The maximum read depth for any putative ectopic recombination was essentially the same for the parental strain (maximum read depth 8) and the SCRaMbLEd strains (maximum 10 across all 64 strains), and far less than the average read depth. Furthermore, 80.8% of the reads supporting putative ectopic recombinations involved mitochondrial genome sequence, which, due to mitochondrial copy number, has a much higher average read depth of 562. In summary, no strong evidence for off-target or ectopic recombination caused by SCRaMbLE was found from our sequencing data.

4.2.3 Full sequence reconstruction

“SCRaMbLEgrams” was used to depict the segment order and orientation for each strain (Figure 4.3). Each strain had a unique structure; the diverse recombinations and resulting genomes support the use of SCRaMbLE to generate combinatorial diversity through random recombinations between loxPsym sites. Simple deletions ranged in length from 1 to 16 segments, or 135 to 41,999 bp. The largest simple inversion was in JS611, involving 20 segments and extending over 38,925 bp. The strain JS613, with the greatest number of simple recombinations that could be unambiguously mapped, had 8 simple deletions and 2 simple inversions. Compared to uninduced *SynIXR*, 6 hour induction of *SynIXR* led to 10-fold fewer viable colonies; 12 hour induction to about 100-fold fewer colonies; and longer induction periods to 1000-fold fewer colonies[16], providing indirect but strong evidence for more events with longer induction periods.

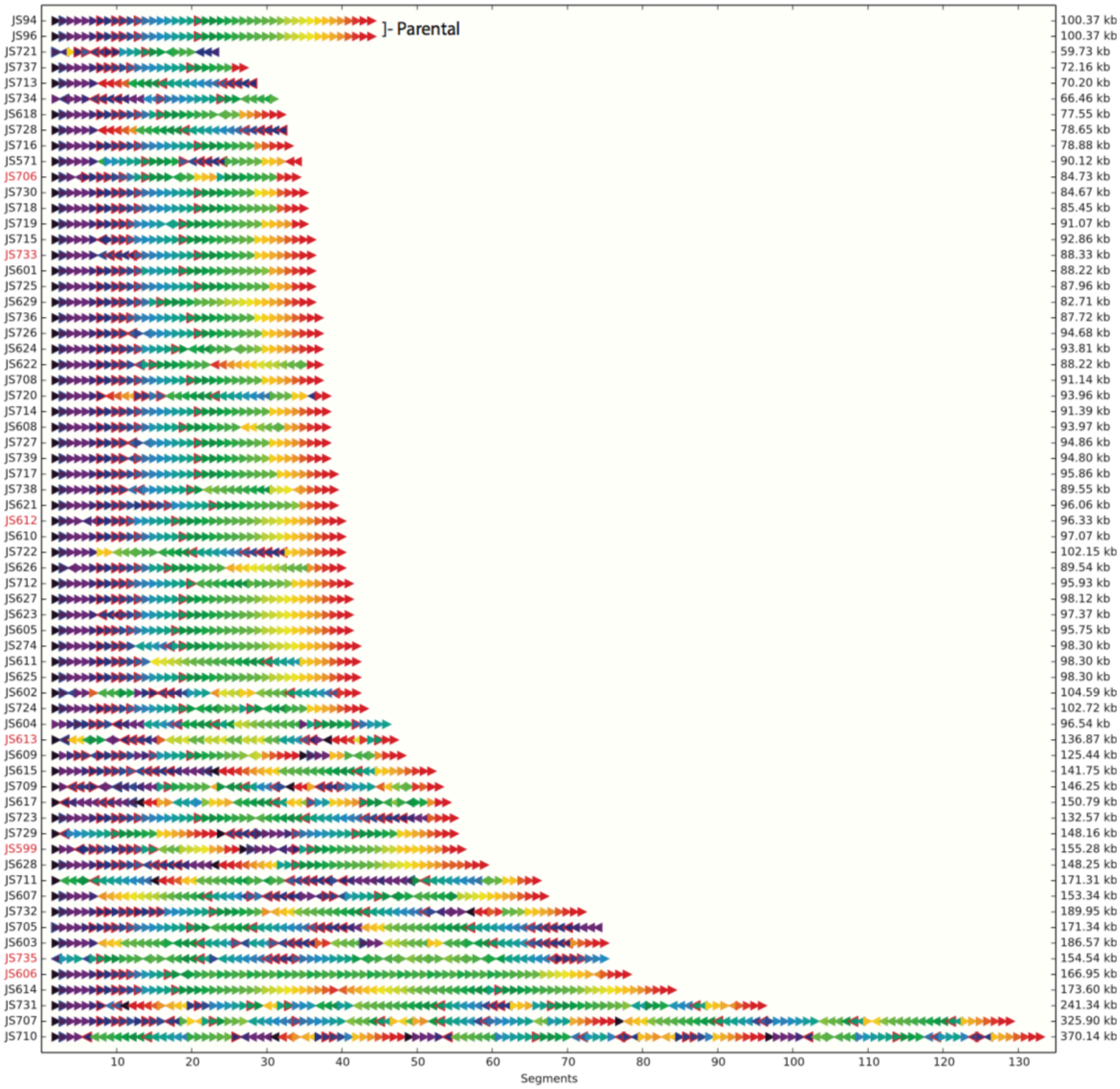


Figure 4.3 Rearrangements observed in *SynIXR* SCRaMbLE strains

Each SCRaMbLE strain is represented as a sequence of arrows. The colour of each arrow indicates the segment number in the parental chromosome, and the direction of the arrow represents the orientation (SCRaMbLEgram visualization). A red border denotes a segment containing an essential gene. The names of slow growth strains are indicated in red text.

The observed *SynIXR* junctions were then used as inputs to Euclidean path algorithms to reconstruct the rearranged chromosomes (Figure 4.2). These algorithms require only linear time to check the existence of a sequence consistent with the observed loxPsym junctions, and then near-linear time to reconstruct feasible solutions. For 39 of the scrambled *SynIXR* strains, junctions in the 500 bp library helped derive a unique reconstruction (Table 4.1). The remaining strains had more than one possible reconstruction because the insert size was insufficient to resolve rearrangements involving large duplications and higher amplifications. Strain JS735, for example, has 41 parental junctions, 33 novel junctions, and 23 duplicated segments, and 1,732,332 possible solutions based on junctions from short inserts.

Strain ID	Total # of recombination s	# of segments observed at copy number 1	# of segments observed at copy number 2 or more	# of segments deleted	# of reconstruction s	# of 10kb reads	Final # of possible unique reconstruction
JS94	0	43	0	0	1	-	1
JS96	0	43	0	0	1	-	1
JS274	3	41	0	2	1	-	1
JS571	8	33	0	10	1	-	1
JS601	3	33	1	9	1	-	1
JS605	2	40	0	3	1	-	1
JS606	19	34	4	5	1	-	1
JS608	5	37	0	6	1	-	1
JS610	3	39	0	4	1	-	1
JS611	2	41	0	2	1	-	1
JS612	5	39	0	4	1	-	1
JS618	4	31	0	12	1	-	1
JS621	3	30	4	9	1	-	1
JS622	6	36	0	7	1	-	1
JS623	4	40	0	3	1	-	1
JS624	6	36	0	7	1	-	1
JS625	1	41	0	2	1	-	1
JS626	5	39	0	4	1	-	1
JS627	2	40	0	3	1	-	1
JS629	2	35	0	8	1	-	1
JS708	2	36	0	7	1	-	1
JS710	44	1	35	7	-	2667118	1
JS713	4	27	0	16	1	-	1
JS714	1	37	0	6	1	-	1
JS715	6	35	0	8	1	-	1
JS716	1	32	0	11	1	-	1
JS717	1	38	0	5	1	-	1
JS718	1	34	0	9	1	-	1
JS719	6	34	0	9	1	-	1
JS722	8	35	2	6	1	-	1
JS724	8	24	9	10	1	-	1
JS725	2	35	0	8	1	-	1
JS726	7	36	0	7	1	-	1
JS727	6	37	0	6	1	-	1
JS728	3	31	0	12	1	-	1
JS729	7	10	22	11	1	-	1
JS730	2	34	0	9	1	-	1
JS733	3	35	0	8	1	-	1
JS736	2	36	0	7	1	-	1
JS737	3	26	0	17	1	-	1
JS738	7	38	0	5	1	-	1
JS739	4	37	0	6	1	-	1
JS602	20	33	4	6	2	-	2
JS603	21	10	28	5	6592	2651354	2
JS613	28	16	12	15	132	5213592	2
JS617	23	10	18	15	15040	2908656	2
JS709	32	18	17	8	390	5862856	3
JS712	6	36	2	5	2	-	2
JS720	9	35	1	7	2	-	2
JS721	16	15	3	25	2	-	2
JS723	3	10	22	11	2	-	2
JS607	18	14	26	3	312	2377080	4
JS614	3	17	23	3	4	-	4
JS615	4	15	18	10	4	-	4
JS732	16	8	31	4	272	5222530	4
JS734	9	20	5	18	4	-	4
JS604	13	16	14	13	8	-	8
JS711	22	5	29	9	222	2586462	3
JS628	5	20	18	5	12	-	12
JS706	9	19	7	17	12	-	12
JS609	18	23	12	8	18	-	18
JS705	6	9	24	10	36	-	36
JS599	19	13	21	9	160	-	160
JS735	33	3	23	17	1732332	-	1732332
JS707	41	6	32	5	-	4712378	48
JS731	27	10	29	4	171072	4966128	1
BY4741	-	-	-	-	-	-	-

Table 4.1 SCRaMbLE analysis summary of *SynIXR* and BY4741 strains

Therefore, long-insert libraries were used to constrain and reduce the number of feasible solutions and found to be powerful. Of the 25 strains with multiple solutions, 10 were chosen for long-insert sequencing. Each of these strains had at least 100 feasible solutions with the short-insert library, and two had over 100,000 solutions. Long-insert reads reduced the number of solutions to 4 or fewer for each sequenced strain, except for JS707, whose over 1,000,000 feasible solutions from short inserts were reduced to 48 feasible solutions.

Recombination events were classified as deletions, inversions, tandem and inverted duplications, and more complex rearrangements with respect to the parental *SynIXR* sequence (Figure 4.4). Most frequently retained segments were mostly a single copy (Figure 4.5A), and, as expected, essential genes and the centromere are retained in all strains. Segments 19 and 24, which lack essential genes but contain genes required for fast growth, are also retained in all strains. The *MRS1* gene in segment 19 is essential for mitochondrial gene expression[121]. Segment 19 is also immediately upstream of essential gene *SEC11* on segment 20, and each strain also retains the 19R-20L junction; segment 19 might therefore contain regulatory elements required for proper *SEC11* expression. Segment 24 contains *YVH1*, and *YVH1* knockouts confer slow growth[121]. Auxotrophy screening previously performed on the strains chosen for study generated deletion peaks close to segment 14 (containing *MET28*) and segment 32 (*LYS1*) as expected (Figure 4.5A). The segment containing *MET28* is close to essential genes in segments 12 and 20, accordingly, deletions of *MET28* do not extend beyond these limits. Deletions of *LYS1* can be much larger, with the closest essential segments in the circular chromosome being 20 and 2.

While the SCRaMbLE system was designed primarily to generate deletions and inversions, duplications were also found to be surprisingly frequent and widespread. At least one duplication occurred in 31 of the 64 *SynIXR* strains, and each of the 43 segments is amplified in at least one strain (Figure 4.5B). Prolific amplifications, sometimes spanning most of the surviving regions of *SynIXR*, suggest that duplications of genes on *SynIXR* do not incur a strong fitness defect, even for essential genes (p-value: 0.329, two-sided t-test for the number of strains with essential vs. non-essential segments amplified). Amplifications may arise from the double rolling circle mechanism, which can amplify micron plasmids[122,123] and has recently been used in engineered systems as a model for gene amplifications in cancer[124]. Double rolling circle amplification may occur when loxPsym sites recombine across a replication fork within a DNA replication bubble, creating a topology in which replication forks travel in the same direction around the bubble until one is reversed by a second recombination. Unequal crossing over between two loxP sites during the late *S* or *G2* phases is another mechanism that can generate duplicated regions.

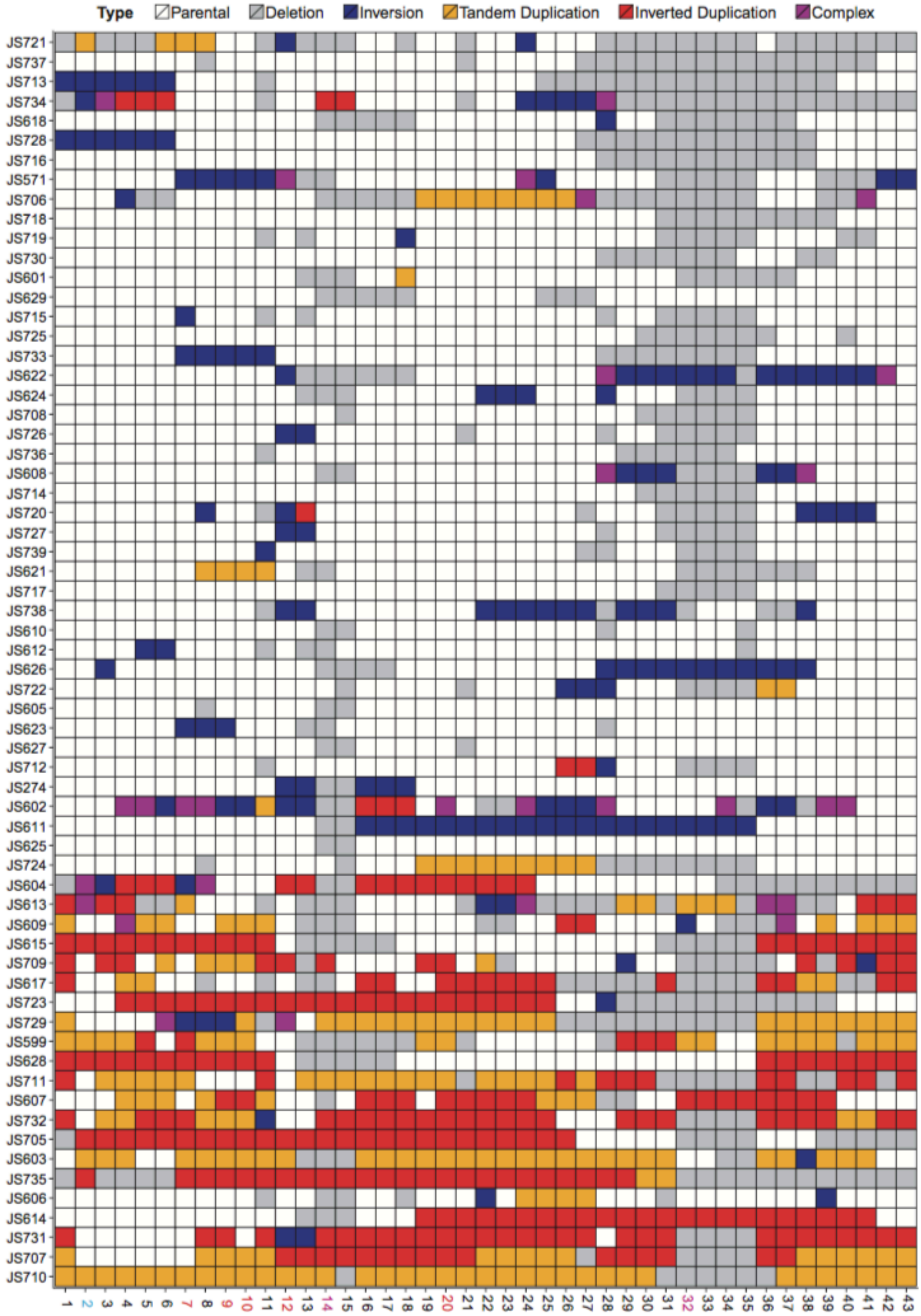


Figure 4.4 Recombination events of *SynIXR* SCRaMbLEd strains

The fate of each segment in each strain in indicated as deleted (gray), inverted (blue), tandem duplication (orange), inverted duplication (red), complex (purple), or unaffected by any SCRaMbLE event (white).

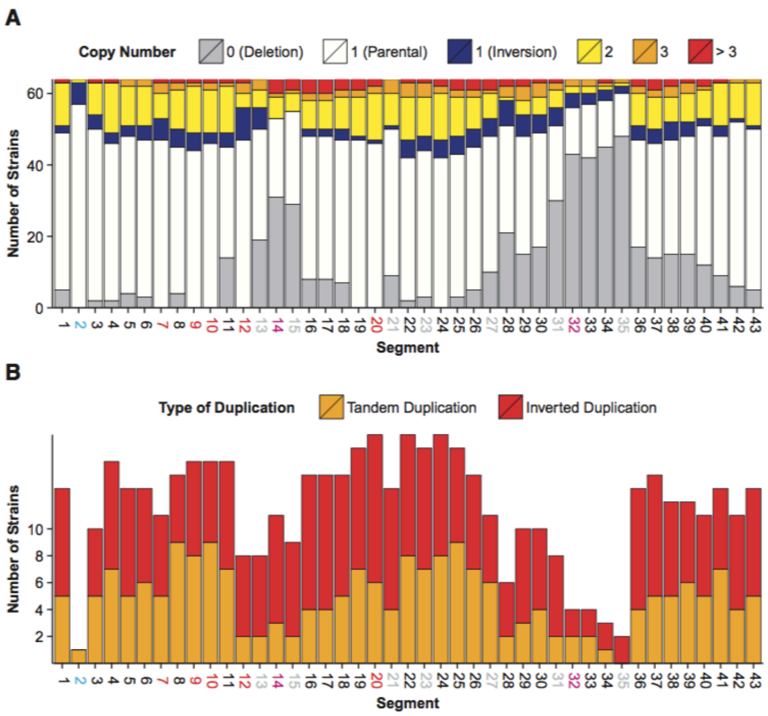


Figure 4.5 Copy number distribution of segments in *SynIXR*

A. The distribution of copy number for each segment is shown across strains. B. Each segment was involved in at least one duplication event.

While the mean number of events per strain was 6.2 ± 4.9 (Figure 4.6), the distribution of events per strain has a long tail, with 18 events observed in one strain (Figure 4.7A). The loxPsym site was designed to allow recombination in both directions, which, regardless of fitness effects and subtle differences in DNA bending required for deletion versus inversion, should result in equal numbers of these types of events. While 156 total deletions were observed, 63 of these were required by phenotypic selection. The remaining 93 deletion events were indistinguishable from the 89 inversion events (P = 0.83, Poisson test for rates). Similarly, 50 tandem duplications and 44 inverted duplications were observed, at roughly equal frequencies (P = 0.66, Poisson test for rates). Recombination event frequencies negatively correlate with the distance between loxPsym sites prior to an event (Figure 4.7B). Deletions are further limited in length by the requirement to retain essential genes. Inversion frequencies also decay with length, consistent with statistical probabilities of DNA looping[125,126], and more slowly than the length decay of deletions.

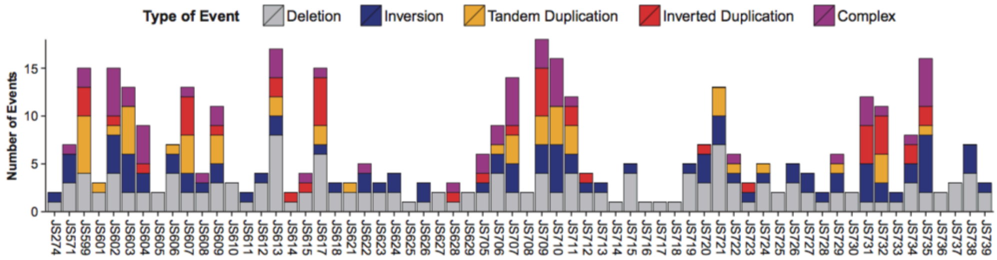


Figure 4.6 SCRaMbLE events of *SynIXR* strains

The number of SCRaMbLE events of each type is depicted for each strain, with the number of events per strain is 6.2 ± 4.9.

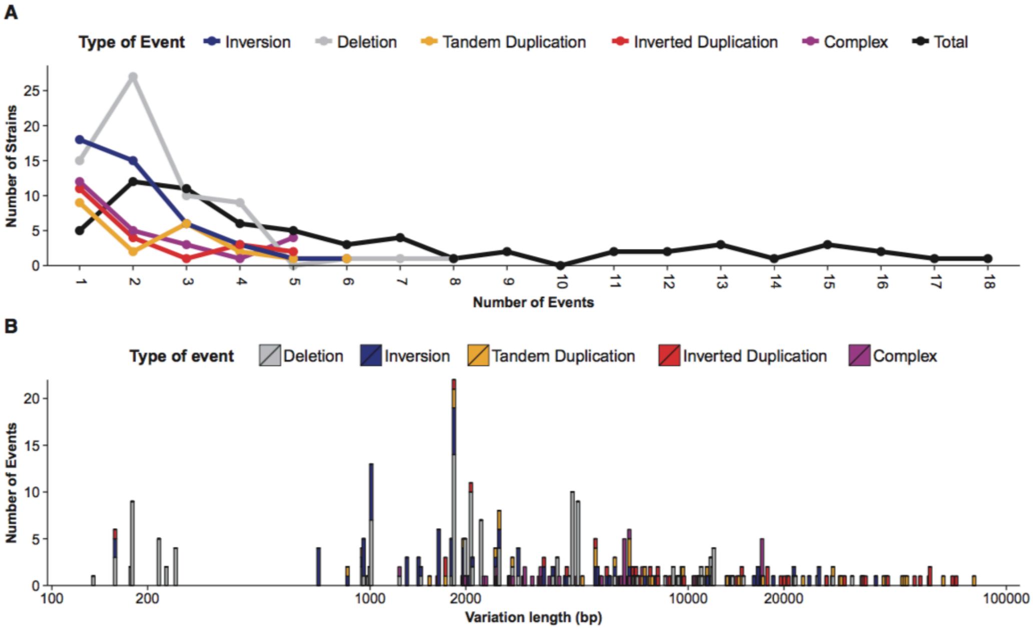


Figure 4.7 Number and variation length of different types of events in SCRaMbLEd *SynIXR*

A. The number of events per strain has a long tail, with strains observed having 17 and 18 distinct recombination events. B. Events are more likely at short distances but continue to be observed for long separations between recombination sites. There is a preponderance of deletions over inversions at the short lengths.

To summarize, the analysis of 63 SynIXR SCRaMbLEd strains relative to the parental *SynIXR* strain using the established analysis pipeline shows that SCRaMbLE combines designed recombination sites with controlled expression of recombinase for efficient, robust generation of genomic diversity. Deep sequencing of yeast strains before and after SCRaMbLE demonstrates that the parental genome is stable, with no evidence for recombination outside the designed sites. Observed variations include deletions, inversions and duplications. Deletion frequencies readily identify genes required for high fitness. The double rolling circle mechanism likely contributes to the amplification of large regions, which is predicted to produce useful "gain-of-function" phenotypes.

4.3 Fitness defect rescue of *SynII* by SCRaMbLE

Genome reconstruction analysis of circular *SynIXR* shows that SCRaMbLE system will be valuable in combinatorial exploration of genomic diversity for gene-gene and gene-environment interactions and phenotype-based selections. During the construction of Sc2.0 chromosomes, given that some defects might involve more than one bug, the complementation method used for construction of *SynII* may not be applicable (Details can be found in Chapter Two). In this case, SCRaMbLE may provide a powerful alternative approach for correcting any defect that arise.

Here, the feasibility of this approach has been demonstrated by quickly evolving a *SynII* deriviative to identify and overcome a design problem. By turning on the SCRaMbLE system in *synII* cells and directly cultivating the SCRaMbLEd population in conditions that appear the defect, a few wild type-like large colonies were identified (Figure 4.8A). From two large colonies isolated for sequencing, no rearrangement was observed within or near the *YBR265W* gene, in which the synonymously recoded PCRTag altered *YBR265W* protein level and led to the defect. However, all genomic rearrangements on *SynII* involve genes coding for regulatory proteins that have either genetic or physical interactions with high osmolarity glycerol (HOG) regulatory proteins (Figure 4.8B). It is also interesting to see that a phenotypic defect can be rescued through a number of different routes within the complex cellular interactome network.

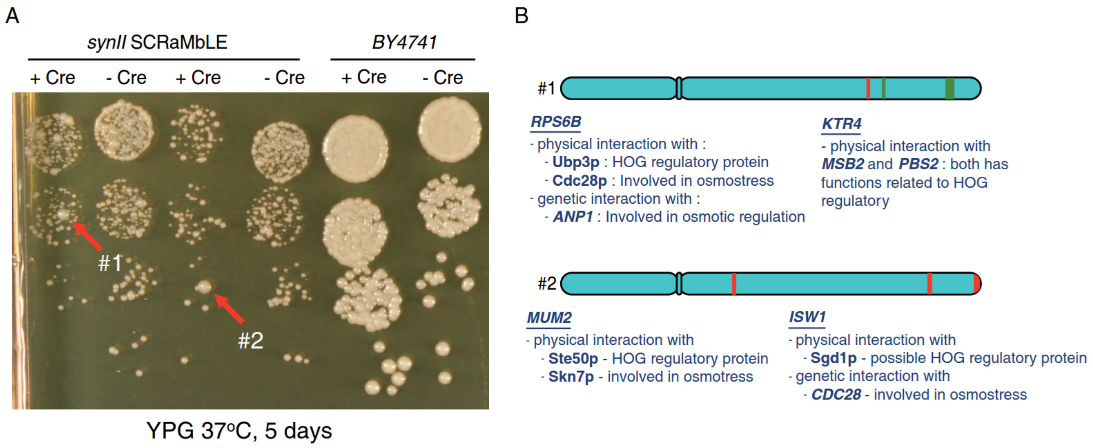

A
synII SCRaMbLE
BY4741
+ Cre
- Cre
+ Cre
- Cre
+ Cre
- Cre
#1
#2
YPG 37°C, 5 days
B
#1
RPS6B
- physical interaction with :
- Ubp3p : HOG regulatory protein
- Cdc28p : Involved in osmostress
- genetic interaction with :
- ANP1 : Involved in osmotic regulation
KTR4
- physical interaction with
MSB2 and PBS2 : both has functions related to HOG regulatory
#2
MUM2
- physical interaction with
- Ste50p - HOG regulatory protein
- Skn7p - involved in osmotress
ISW1
- physical interaction with
- Sgd1p - possible HOG regulatory protein
- genetic interaction with
- CDC28 - involved in osmostress

Figure 4.8 *SynII* defect at 37°C on YPG media can be rescued by SCRaMbLE

A. Screening the SCRaMbLEd *SynII* colonies with recovered phenotype. Two large colonies were isolated for sequencing verification (red arrows). B. Genome rearrangements identified on *SynII* in growth recovery strains. Inversions and deletions are highlighted in green and red respectively. Genes involving the HOG response regulation are listed. *RPS6B, MUM2* and *ISW1* were deleted (Red) and *KTR4* generated a convergent ORF (CDS-CDS) in an inversion (Green).

4.4 Conclusion and discussion

This chapter started with the introduction of a genome analysis pipeline developed for detailed analysis of SCRaMbLEd Sc2.0 strains. And from the analysis of 64 SCRaMbLEd SynIXR strains, a few conclusions of the SCRaMbLE system can be addressed:

First, analysis of 63 synIXR SCRaMbLE strains confirms that the synthetic system functions as designed. In the absence of Cre induction, the parental chromosome remained stable. For induced strains, no off-target recombination events were observed as all novel junctions were derived from designed loxPsym sites, with no ectopic recombinations or other rearrangements observed. Studies of *SynIII* strain also support this conclusion[17]. However, instead of integrating Cre-EBD in the genome, it is better to provide Cre-EBD on a plasmid as leaky expression from integrated Cre-EBD could lead to continued recombinations in the absence of estradiol induction.

Second, short-insert libraries were sufficient to characterize the recombination junctions and copy number variations for every strain and provided unambiguous genome structures for most SCRaMbLEd strains. For strains with multiple solutions yielded by short read sequencing, additional sequencing with long-insert libraries helped to narrow down the number of solutions. The result of *SynIXR* SCRaMbLEd strains demonstrates that a combination of short-insert and long-insert sequencing can be sufficient to determine genome structure even for highly rearranged and duplicated SCRaMbLEd genomes.

Third, each strain had a unique genome, demonstrating the ability of SCRaMbLE to generate genomic diversity. The spectrum of loxPsym recombinations was consistent with retention of essential genes and selection for loss of phenotypic markers. Recombination hotspots and cold-spots in general may depend on selection pressure. Deletion patterns from SCRaMbLE provided ample information to identify all annotated essential and slow-growth genes. The ability of SCRaMbLE to generate gene amplifications may permit evolution of new functions.

Several applications of SCRaMbLE with synthetic yeast strains were presented in this chapter. First, we found *SynII* strain has fitness defect under a specific growth condition (glycerol as carbon source and growth temperature at 37°C). The strength of SCRaMbLE was demonstrated by generating *SynII* strains with restored tolerance to this stress condition. In the previous chapter, *YBR265W* gene was found carrying a recoded PCRTag that is responsible for the defect. However, SCRaMbLE analysis revealed that ideletions and inversions obtained in the two SCRaMbLEd *SynII* strains involve genes (*RPS6B*, *MUM2*, *ISW1* and *KTR4*) which are all involved or interact with the HOG response regulatory pathway.

Applications of SCRaMbLE have also been attempted by other studies. For example, SCRaMbLE has been used to generate strains with increased tolerance to several stress conditions, such as ethanol, heat and acetic acid[127]. In addition, SCRaMbLE has been applied for exogenous pathway optimization and chassis engineering[128]. It is encouraging to see that this system could be used in combination with directed evolution to further increase permutations and potentially lead to more desired new strains. The power of SCRaMbLE is not only to identify strains, but also to dissect the potential underlying mechanisms through the combination of other genetics tools, phenotypic assays and sequencing efforts. Therefore, the system holds great potential for fundamental biology studies and industrial applications. With the merging of multiple synthetic chromosomes and the complete synthesis of the whole yeast genome in near future, more applications for the SCRaMbLE system will emerge.

4.5 Materials and methods

Sequencing Library preparation

The sequencing libraries were made without PCR amplification to avoid creating junction artifacts through crossover PCR driven by hybridization of loxPsym sites. In addition, in order to accurately estimate copy number of target regions, amplification-free sequencing decreased the likelihood that an appreciable proportion of these sequences would be duplicates and preserved a more even distribution of read coverage across the targeted sequencing regions[129]. Cultures of reference and SCRaMbLE strains were grown in 30 mL YPD at 30°C until saturated. DNA was extracted by the glass beads method[130], and the amplicon-free libraries were prepared following the Illumina protocol. Ten micrograms of DNA were sheared using a Covaris S2 to an average length of 500 bp, end-repaired, and ligated to Illumina paired-end adapters. Ligated fragments were selected on agarose gels and purified to yield the corresponding libraries. A large amplicon library of average length 10 kb was also prepared for 10 strains by published methods[120]. All constructed libraries were sequenced on the Illumina HiSeq 2000 platform.

Mapping to the parental genome

Illumina HiSeq 2000 paired-end short reads were trimmed to remove adapter sequence. Then strict filtering was used to remove <100 bp or duplicated reads. Reads with unknown bases or with bases having a Phred-score below seven were removed. Filtered reads were mapped to the *SynIXR* reference sequences JN020955 using the SOAP short-read aligner (http://soap.genomics.org. cn/) for standard data processing of reads mapping to the parental genome with no rearrangements[96]. For the nonsynthetic chromosomes, BY4741 served as the reference. Paired-end reads that could not be mapped directly were analyzed using Bowtie[84] using standard settings.

Splitting reads for junction identification

Reads containing loxPsym sites that did not map to the parental genome, with at least 15 bp of sequence flanking the loxPsym site, were then trisected into a loxPsym site and two single ends, which remain associated with the paired sequence from the other end of the fragment. The two single ends were then aligned to the reference using Bowtie 2[131] by single-end mapping to determine a novel loxPsym junction, with at least three reads required for support.

Ectopic and parental structural variant detection

Reads without loxPsym sites that did not map to the parental genome were directly split into two ends. In order to precisely locate a possible breakpoint, the split site was scanned over all intermediate positions at least 15 bp from the ends of the read. Then the two ends were aligned to the reference using Bowtie 2 by single-end mapping with parameter –k 100 for breakpoint detection, which could provide direct evidence of an ectopic recombination. For each identified breakpoint, a 5-bp error range was allowed. For a pre-existing parental variation relative to the BY4741_v2 reference, a 15-bp error range was allowed and at least two reads per sample were required for identifying pre-existing breakpoints.

To detect recombination events outside of the designed target loxPsym site, we also applied CREST (Clipping REveals STructure), an algorithm that detects genomic structural variations at single base resolution[132]. A sequence read that spans a breakpoint will have partial alignment to both sides of the junction. Therefore, BWA[96] was used first to perform local alignment to all generated reads after quality control. The unaligned portion of reads were masked by a process termed "soft-clipping" because the unaligned subsequence is retained but not trimmed even though it does not map to the current genomic location. Then soft-clipped reads were extracted from the binary alignment/map (BAM) file. First, soft-clipped reads at a putative breakpoint were assembled into a contig; the contig was then mapped against the reference genome to identify candidate partner breakpoints; then all possible soft-clipped reads were identified and assembled into a contig, followed by an alignment of the contig derived from the partner back to the reference genome. Finally, a match to the initial breakpoint was performed to check whether the breakpoint prediction was correct. All breakpoints with read depth close to the average depth involved recombinations between pairs of designed loxPsym sites.

Copy number variation

Copy numbers were estimated to detect deletions, duplications, and higher amplifications. The average sequencing depth of *SynIXR* was 29.02. Then every segment with read depth of less than five was marked as deleted. After proceeding with the copy number estimation algorithm below, the average read depth for deleted segments was 0.27 and the maximum was 4.98. Segments 1 and 38 exhibit a read depth close to five when deleted because reads from homologous regions in the genome are erroneously mapped. To account for possible systematic bias in sequencing depth due to confounding factors from library preparation and mapping, an iterative algorithm was used to refine the copy number estimation and the potential bias.

Structural variation and genome reconstruction

Each rearranged genome, denoted T, was represented by an undirected graph G = (V,E). The vertices V represented the segments of T, integer values with magnitude one through the total number of segments L (in this case 43) and sign indicating direction relative to the reference. The edges E represented the junctional-reads for adjacent segments. For each tandem duplication of segment i with copy number Ni, Ni − 1 vertices were added to the graph and connected as a ring. Given this graph G, the rearranged genome sequence was reconstructed by finding a cycle that visits each edge in G once, termed a Eulerian cycle over G. Algorithms based on Eulerian cycles have become standard in DNA sequence assembly. First the existence of a Eulerian cycle was checked by verifying that each vertex has an even degree and all belong to one and only one connected component; these operations can be performed in linear time. While it is possible that a mixture of two distinct genomes may lead to no cycles being found, in each case we found at least one Eulerian cycle. Multiple Eulerian cycle solutions were found in some cases. These result from rearrangements that extend farther than the length of the paired-end library and introduce ambiguities similar to genome assembly with long repeats. The multiple solutions have equal length and identical composition of segments and segment junctions.

Long-insert libraries

Paired-end reads from long-insert libraries with fragment sizes of ~10 kb were used to constrain the solutions for the 10 strains with complex rearrangements. Quality thresholds for sequencing reads from the 10-kb library were the same as for the 500- bp library. The reads of ~90 bp were aligned to the reference (Ref Strain: BY4741, the original right arm sequence of Chromosome 9 was replaced with the synthetic sequence) using SOAPaligner 2.21[96], available at http://soap.genomics.org.cn/

soapaligner.html, with parameters -m 7500 -x 12500 -r 1 -v 4 –R.

Classifying recombination events

Duplication events were identified as repeating regions of the reconstruction genome. Repeating regions in the same direction were further classified as tandem duplications and repeat regions in the opposite orientation were classified as inverted duplications. Subregions with the opposite direction within repeat and non-repeat regions were identified as inversion events. Deletion events were identified as segments with copy number of zero. A junction between two repeat regions or between a repeat and a non-repeat region was classified as a duplication junction. A junction between an inversion and a non-inversion region was an inversion junction. A deletion junction was a junction between two segments that are adjacent to the left end and the right end of a deleted region in the parental genome. Non-parental junctions not associated with the endpoints of deletion, duplication, or inversion events were classified as complex junctions. Contiguous regions of complex junctions were defined as a single complex event.

Cre SCRaMbLE of *SynII*

BY4741 and *SynII* with *pSCW11-CreEBD-His3*[133] plasmid were cultured in SC–His liquid media overnight at 30°C, with duplicates for each strain. Cells were inoculated into SC–His liquid medium with 1 μM EST and cultured for 24 hrs at 30°C. Then cells were washed, serial diluted and spotted on YPG plates, followed by 4 days of incubation at 37°C.

Chapter Five

Overview and perspectives

In this thesis, I have presented three major areas that are the focus of my research and explained their contributions in the development of yeast synthetic genomics. The first part focused on the development of a highly efficient and cost-effective DNA construction strategy for megabase long chromosomes for the Sc2.0 project, which holds the potential to be applied for genome-scale construction in other organisms. The second part provided details on the use of an extensive trans-omics approach to dissect the fitness of strains carrying synthetic chromosomes, which facilitate the identification of subtle phenotypic changes at different levels. In the third part, the SCRaMbLE system is discussed in detail. Although several other debugging methods are explained in the second part, SCRaMbLE offers a new way of studying how the model organism, yeast, can adapt to challenging conditions by systemically altering its genomic content.

The Sc2.0 project is considered to be a current landmark of synthetic genomics. To date, over one-third of the *S. cerevisiae* chromosomes have been individually constructed by now. Meanwhile, a method has also been developed to build up a polysynthetic strain[20]. Through the dedicated efforts of the Sc2.0 consortium, it is expected that this will be realised in the near future the completion of Sc2.0 strain carrying all synthetic chromosomes as well as the tRNA neochromosome.

The Sc2.0 strain will be useful to answer biological questions, e.g. how the synthetic strain adapts to incorporate an extra neochromosome carrying all tRNA genes. If the final Sc2.0 strain maintains a wild type like phenotype, this would suggest that gene order might not be a major contributor to cell viability. Then one of the next design frontiers might be to focus on reorganizing the whole genome into several contiguous DNA cassettes, which represent known biological functions. The majority of the 250-300 introns in native budding yeast genome have been removed from the Sc2.0 synthetic chromosomes. Ribosomal introns are still retained as a previous study showed that deletion of introns encoded within certain ribosomal protein genes has led to fitness defects[69]. Intriguingly, exclusion of all ribosomal introns from *SynII* did not seem to affect fitness under several growth conditions examined so far (Yue Shen, et al. unpublished). One possible explanation is that all ribosomal proteins on *SynII* have duplicate copies on other chromosomes, and this might have contributed to maintain the wild-type like phenotype. It will be interesting, in future studies, to further examine the possibility of constructing a completely intron-free Sc2.0 strain.

The Sc2.0 project represents the first genome-scale synthesis effort in any eukaryotic organism. Although sequencing, analyzing, editing and synthesis techniques continue to advance at lightning speed, genome-scale synthesis is still limited to relatively small genomes. Discussions on building large, gigabase (Gb)-sized genomes, for both animals and plants, to further our understanding of genetic blueprints have been ongoing[23]. The consensus is that combining de novo genome synthesis and genome editing (here we call this combination 'genome-write') may constitute a superior approach for assembling large artificial genomes[134]. The homologous recombination efficiency in cells from advanced species may not be as high as that in yeasts, thus *in vivo* genome editing tools, such as CRISPR/Cas9, are necessary for homology-independent targeted integration of synthetic DNA fragments in both dividing and non-dividing cells in vitro and, more importantly, in vivo (for example, in neurons of postnatal mammals) [135].

Genome-write technologies will definitely improve our understanding of many fundamental biological questions.

One of the most ambitious genome-write projects is to generate a synthetic cell with a functional synthetic genome. The need to design an artificial genome comes from a fundamental question: what is gene essentiality? The essentiality of cellular pathways can be context dependent and evolvable in all kingdoms of life[136]. Essential pathways are mostly tasked with basic and fundamental cellular functions that support cell/organism viability. In order to adapt to the extrinsic environment, cellular metabolic pathways often undergo extensive evolution, resulting in the development of a series of complex secondary metabolic mechanisms. It has been a long-standing interest and effort to study the biological processes and metabolic pathways of natural cells and organisms and in this regard synthetic genomics can provide a different perspective. For example, recently Hutchison et al. reported the design and construction of a synthesized minimal *M. mycoides* genome containing only 473 genes, which is half the size of the native genome. Surprisingly, among these 473 genes, they found there are still over one hundred genes with unknown functions, but seemingly essential for cell viability[62].

Another future focus could be the rewriting of the genetic coding rules. Genome write does not have to follow the existing rules of genetic encoding and decoding, thereby holding potential to create new life forms that do not exist in nature. Although different organisms differ vastly in many aspects, their genomes are encoded by the same four nucleotides, and their proteins are assembled with the same 20 types of amino acids. Synthetic organisms, on the other hand, may be designed to incorporate additional types of nucleotides into their genomes and non-canonical amino acids into their proteins. By altering the codon rules and using modified tRNAs, new proteins containing non-canonical amino acids can be obtained. Lajoie et al. replaced all TAG stop codons with the TAA stop codons and reassigned TAG codon to a non-standard amino acid in bacteria[137]. Similar efforts have also been applied in the Sc2.0 project[70]. Very recently, Zhang et al. incorporated unnatural base pairs into DNA and non-canonical amino acids into GFP with the use of specific tRNA[138]. These studies expanded the repertoire for DNA replication, transcription and translation, thereby paving the way for the creation of a variety of artificial organisms with special features.

Last but not least, one of the major genome-write applications is therapeutics. Genome-wide association studies (GWAS) have revealed countless small or large variations that might be associated with medically important phenotypes [139]. Many studies are trying to establish causal relationships between particular mutations and specific diseases, especially for single gene disorders. Subsequently, repair of such defects could be realized by introducing synthetic DNA with the mutation eliminated by genome editing tools.

Despite the potential mentioned above, however, there remain many challenges that require continued investment and effort before full-fledged genome-write applications become possible:

First, we need to continue improving our understanding of cellular biological processes and metabolic pathways, which is a prerequisite for developing any new synthetic biology design and applications. There are more than eight million different species living on Earth. Among them, many are complex multicellular organisms that have a multitude of cell types with different compositions and functions. These cells share common fundamental metabolism, but also differ in specific metabolic pathway(s). What are the factors involved, and what are the relations between these factors? Many researches are needed to answer these questions.

Second, new technologies are needed to further reduce the cost of genome synthesis and targeted integration/delivery. At present, implementing a large-scale synthetic genomic project is still costly and time consuming. The synthetic DNA fragment is first synthesized as oligonucleotides by a chemical method, then assembled into larger pieces through multiple DNA assembly techniques. Due to the limited efficiency of chemical-based DNA synthesis, the efficiency of synthesizing oligonucleotides more than 200 bases is quite challenging. Therefore, new methods which can improve the chemistry reaction efficiency or techniques that use enzymes to synthesize long DNA are necessary. For targeted integration of synthetic DNA, the current techniques are still inefficient in precision delivery, while the size the synthetic DNA to be delivered is also limited. Customizable nuclease-based genome editing relies on the cell's intrinsic DNA double-strand break (DSB) repair mechanism. Due to the difficulty of precisely controlling the DNA repair process, homology directed repair (HDR) based error-free repair efficiency is inherently low. Thus, new genome editing tools which will allow large pieces of DNA (from 100 kb to megabase scale) to be delivered into host cells in designated site are needed.

Third, public engagement is needed for addressing bioethics and biosafety concerns. Synthetic biology has made it easy to synthesise human-animal chimeric genomes. An organism containing both human and animal genetic material can be made by incorporating large amounts of human genomic fragments into an animal genome, or vice versa. Depending on the degree of genomic mixing, it may become increasingly difficult to define the species identity of the generated cells and/or organisms. In addition, it has been reported that synthetic viruses can have the same infectivity as the natural virus[140]. These synthetic pathogens may be used for criminal activities that endanger public safety. Therefore, we must take extreme precautions to prevent synthetic biology technologies from being used to engineer toxic, infectious and other harmful organisms or biological materials.

In general, synthetic biology (synthetic genomics in particular) marks a new era in the field of biology. In contrast to physics and chemistry, there are no simple and straightforward principles to follow to develop applications in biology. Biological processes in all species are black boxes with extremely complex structures and functions. Understanding of the underlying rules is prerequisite for developing applications for every discipline. In the past century, biological studies have undergone rapid development and revolutionized our understanding of organism heredity and development. These consequently laid an important foundation for the birth of synthetic biology. In the future, with a combination of genome writing technology, genetic circuit technology, new genetic encoding and decoding technology, etc. we might be able to create new species or resuscitate extinct species in accordance with human needs.

Reference

1. Wöhler F. Ueber künstliche Bildung des Harnstoffs. *Annalen der Physik und Chemie*. 1828;88(2):253-256. doi:10.1002/andp.18280880206.

2. Sekiya T, Brown EL, Belagaje R, et al. Total synthesis of a tyrosine suppressor tRNA gene. XV. Synthesis of the promoter region. *J Biol Chem*. 1979;254(13):5781-5786.

3. Guo Y, Dong J, Zhou T, et al. YeastFab: the design and construction of standard biological parts for metabolic engineering in Saccharomyces cerevisiae. *Nucleic Acids Res*. 2015;43(13):e88-e88. doi:10.1093/nar/gkv464.

4. O'Brien EJ, Monk JM, Palsson BØ. Using Genome-scale Models to Predict Biological Capabilities. *Cell*. 2015;161(5):971-987. doi:10.1016/j.cell.2015.05.019.

5. Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA, Smith HO. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods*. 2009;6(5):343-345. doi:10.1038/nmeth.1318.

6. Engler C, Kandzia R, Marillonnet S. A one pot, one step, precision cloning method with high throughput capability. El-Shemy HA, ed. *PLoS ONE*. 2008;3(11):e3647. doi:10.1371/journal.pone.0003647.

7. Weber E, Engler C, Gruetzner R, Werner S, Marillonnet S. A modular cloning system for standardized assembly of multigene constructs. Peccoud J, ed. *PLoS ONE*. 2011;6(2):e16765. doi:10.1371/journal.pone.0016765.

8. DiCarlo JE, Conley AJ, Penttilä M, Jäntti J, Wang HH, Church GM. Yeast oligo-mediated genome engineering (YOGE). *ACS Synth Biol*. 2013;2(12):741-749. doi:10.1021/sb400117c.

9. Mitchell LA, Chuang J, Agmon N, et al. Versatile genetic assembly system (VEGAS) to assemble pathways for expression in S. cerevisiae. *Nucleic Acids Res*. 2015;43(13):6620-6630. doi:10.1093/nar/gkv466.

10. Kouprina N, Larionov V. Transformation-associated recombination (TAR) cloning for genomics studies and synthetic biology. *Chromosoma*. 2016;125(4):621-632. doi:10.1007/s00412-016-0588-3.

11. Jiang Y, Chen B, Duan C, Sun B, Yang J, Yang S. Multigene editing in the Escherichia coli genome via the CRISPR-Cas9 system. Kelly RM, ed. *Applied and Environmental Microbiology*. 2015;81(7):2506-2514. doi:10.1128/AEM.04023-14.

12. Cello J, Paul AV, Wimmer E. Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. *Science*. 2002;297(5583):1016-1018. doi:10.1126/science.1072266.

13. Smith HO, Hutchison CA, Pfannkoch C, Venter JC. Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. *Proc Natl Acad Sci USA*. 2003;100(26):15440-15445. doi:10.1073/pnas.2237126100.

14. Chan LY, Kosuri S, Endy D. Refactoring bacteriophage T7. *Molecular Systems Biology*. 2005;1(1):2005.0018–E10. doi:10.1038/msb4100025.

15. Gibson DG, Benders GA, Andrews-Pfannkoch C, et al. Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. *Science*. 2008;319(5867):1215-1220. doi:10.1126/science.1151721.

16. Dymond JS, Richardson SM, Coombes CE, et al. Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. *Nature*. 2011;477(7365):471-476. doi:10.1038/nature10403.

17. Annaluru N, Muller H, Mitchell LA, et al. Total synthesis of a functional designer eukaryotic chromosome. *Science*. 2014;344(6179):55-58. doi:10.1126/science.1249252.

18. Shen Y, Wang Y, Chen T, et al. Deep functional analysis of synII, a 770-kilobase synthetic yeast chromosome. *Science*. 2017;355(6329). doi:10.1126/science.aaf4791.

19. Xie Z-X, Li B-Z, Mitchell LA, et al. "Perfect" designer chromosome V and behavior of a ring derivative. *Science*. 2017;355(6329). doi:10.1126/science.aaf4704.

20. Mitchell LA, Wang A, Stracquadanio G, et al. Synthesis, debugging, and effects of synthetic chromosome consolidation: synVI and beyond. *Science*. 2017;355(6329):eaaf4831. doi:10.1126/science.aaf4831.

21. Wu Y, Li B-Z, Zhao M, et al. Bug mapping and fitness testing of chemically synthesized chromosome X. *Science*. 2017;355(6329):eaaf4706. doi:10.1126/science.aaf4706.

22. Zhang W, Zhao G, Luo Z, et al. Engineering the ribosomal DNA in a megabase synthetic chromosome. *Science*. 2017;355(6329). doi:10.1126/science.aaf3981.

23. Boeke JD, Church G, Hessel A, et al. The Genome Project-Write. *Science*. June 2016:1-4. doi:10.1126/science.aaf6850.

24. Durkacz B, Carr A, Nurse P. Transcription of the cdc2 cell cycle control gene of the fission yeast Schizosaccharomyces pombe. *EMBO J*. 1986;5(2):369-373.

25. Enoch T, Nurse P. Mutation of fission yeast cell cycle control genes abolishes dependence of mitosis on DNA replication. *Cell*. 1990;60(4):665-673.

26. Benner SA, Sismour AM. Synthetic biology. *Nat Rev Genet*. 2005;6(7):533-543. doi:10.1038/nrg1637.

27. Purnick PEM, Weiss R. The second wave of synthetic biology: from modules to systems. *Nat Rev Mol Cell Biol*. 2009;10(6):410-422. doi:10.1038/nrm2698.

28. Sprinzak D, Elowitz MB. Reconstruction of genetic circuits. *Nature Publishing Group*. 2005;438(7067):443-448. doi:10.1038/nature04335.

29. Keasling JD. Synthetic biology and the development of tools for metabolic engineering. *Metab Eng*. 2012;14(3):189-195. doi:10.1016/j.ymben.2012.01.004.

30. Yofe I, Zafrir Z, Blau R, et al. Accurate, model-based tuning of synthetic gene expression using introns in S. cerevisiae. Copenhaver GP, ed. *PLoS Genet*. 2014;10(6):e1004407. doi:10.1371/journal.pgen.1004407.

31. Wang Y, Ma M, Xiao X, Wang Z. Intronic splicing enhancers, cognate splicing factors

and context-dependent regulation rules. *Nat Struct Mol Biol*. 2012;19(10):1044-1052. doi:10.1038/nsmb.2377.

32. Spingola M, Grate L, Haussler D, Ares M. Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae. *RNA*. 1999;5(2):221-234.

33. Zalatan JG, Lee ME, Almeida R, et al. Engineering complex synthetic transcriptional programs with CRISPR RNA scaffolds. *Cell*. 2015;160(1-2):339-350. doi:10.1016/j.cell.2014.11.052.

34. Berens C, Groher F, Suess B. RNA aptamers as genetic control devices: the potential of riboswitches as synthetic elements for regulating gene expression. *Biotechnol J*. 2015;10(2):246-257. doi:10.1002/biot.201300498.

35. Tucker BJ, Breaker RR. Riboswitches as versatile gene control elements. *Current Opinion in Structural Biology*. 2005;15(3):342-348. doi:10.1016/j.sbi.2005.05.003.

36. Grate D, Wilson C. Inducible regulation of the S. cerevisiae cell cycle mediated by an RNA aptamer-ligand complex. *Bioorg Med Chem*. 2001;9(10):2565-2570.

37. Hanson S, Berthelot K, Fink B, McCarthy JEG, Suess B. Tetracycline-aptamer-mediated translational regulation in yeast. *Mol Microbiol*. 2003;49(6):1627-1637.

38. Futcher AB. Saccharomyces cerevisiae cell cycle: cdc28 and the G1 cyclins. *Semin Cell Biol*. 1991;2(4):205-212.

39. Becskei A, Séraphin B, Serrano L. Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *EMBO J*. 2001;20(10):2528-2535. doi:10.1093/emboj/20.10.2528.

40. Ajo-Franklin CM, Drubin DA, Eskin JA, et al. Rational design of memory in eukaryotic cells. *Genes Dev*. 2007;21(18):2271-2276. doi:10.1101/gad.1586107.

41. Williams P. Quorum sensing, communication and cross-kingdom signalling in the bacterial world. *Microbiology (Reading, Engl)*. 2007;153(Pt 12):3923-3938. doi:10.1099/mic.0.2007/012856-0.

42. Lee SC, Ni M, Li W, Shertz C, Heitman J. The evolution of sex: a perspective from the fungal kingdom. *Microbiol Mol Biol Rev*. 2010;74(2):298-340. doi:10.1128/MMBR.00005-10.

43. Hennig S, Rödel G, Ostermann K. Artificial cell-cell communication as an emerging tool in synthetic biology applications. *J Biol Eng*. 2015;9(1):13. doi:10.1186/s13036-015-0011-2.

44. Chen M-T, Weiss R. Artificial cell-cell communication in yeast Saccharomyces cerevisiae using signaling elements from Arabidopsis thaliana. *Nature Biotechnology*. 2005;23(12):1551-1555. doi:10.1038/nbt1162.

45. Khakhar A, Bolten NJ, Nemhauser J, Klavins E. Cell-Cell Communication in Yeast Using Auxin Biosynthesis and Auxin Responsive CRISPR Transcription Factors. *ACS Synth Biol*. 2016;5(4):279-286. doi:10.1021/acssynbio.5b00064.

46. Havens KA, Guseman JM, Jang SS, et al. A synthetic approach reveals extensive tunability of auxin signaling. *Plant Physiol*. 2012;160(1):135-142. doi:10.1104/pp.112.202184.

47. Benenson Y. Biomolecular computing systems: principles, progress and potential. *Nat Rev Genet*. 2012;13(7):455-468. doi:10.1038/nrg3197.

48. Anderson JC, Voigt CA, Arkin AP. Environmental signal integration by a modular AND gate. *Molecular Systems Biology*. 2007;3(1):133. doi:10.1038/msb4100173.

49. Moon TS, Lou C, Tamsir A, Stanton BC, Voigt CA. Genetic programs constructed from layered logic gates in single cells. *Nature Publishing Group*. 2012;491(7423):249-253. doi:10.1038/nature11516.

50. Friedland AE, Lu TK, Wang X, Shi D, Church G, Collins JJ. Synthetic gene networks that count. *Science*. 2009;324(5931):1199-1202. doi:10.1126/science.1172005.

51. Danino T, Mondragón-Palomino O, Tsimring L, Hasty J. A synchronized quorum of genetic clocks. *Nature Publishing Group*. 2010;463(7279):326-330. doi:10.1038/nature08753.

52. Win MN, Smolke CD. Higher-order cellular information processing with synthetic RNA devices. *Science*. 2008;322(5900):456-460. doi:10.1126/science.1160311.

53. Regot S, Macia J, Conde N, et al. Distributed biological computation with multicellular engineered networks. *Nature Publishing Group*. 2011;469(7329):207-211. doi:10.1038/nature09679.

54. Mee MT, Wang HH. Engineering ecosystems and synthetic ecologies. *Mol Biosyst*. 2012;8(10):2470-2483. doi:10.1039/c2mb25133g.

55. Biliouris K, Babson D, Schmidt-Dannert C, Kaznessis YN. Stochastic simulations of a synthetic bacteria-yeast ecosystem. *BMC Syst Biol*. 2012;6(1):58. doi:10.1186/1752-0509-6-58.

56. Shou W, Ram S, Vilar JMG. Synthetic cooperation in engineered yeast populations. *Proc Natl Acad Sci USA*. 2007;104(6):1877-1882. doi:10.1073/pnas.0610575104.

57. Gore J, Youk H, van Oudenaarden A. Snowdrift game dynamics and facultative cheating in yeast. *Nature Publishing Group*. 2009;459(7244):253-256. doi:10.1038/nature07921.

58. Youk H, Lim WA. Secreting and sensing the same molecule allows cells to achieve versatile social behaviors. *Science*. 2014;343(6171):1242782-1242782. doi:10.1126/science.1242782.

59. De Roy K, Marzorati M, Van den Abbeele P, Van de Wiele T, Boon N. Synthetic microbial ecosystems: an exciting tool to understand and apply microbial communities. *Environ Microbiol*. 2014;16(6):1472-1481. doi:10.1111/1462-2920.12343.

60. Paddon CJ, Westfall PJ, Pitera DJ, et al. High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature Publishing Group*. 2013;496(7446):528-532. doi:10.1038/nature12051.

61. Gibson DG, Glass JI, Lartigue C, et al. Creation of a bacterial cell controlled by a chemically synthesized genome. *Science*. 2010;329(5987):52-56. doi:10.1126/science.1190719.

62. Hutchison CA, Chuang R-Y, Noskov VN, et al. Design and synthesis of a minimal bacterial genome. *Science*. 2016;351(6280):aad6253-aad6253. doi:10.1126/science.aad6253.

63. DiCarlo JE, Norville JE, Mali P, Rios X, Aach J, Church GM. Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. *Nucleic Acids Res*. 2013;41(7):4336-4343. doi:10.1093/nar/gkt135.

64. Horwitz AA, Walter JM, Schubert MG, et al. Efficient Multiplexed Integration of Synergistic Alleles and Metabolic Pathways in Yeasts via CRISPR-Cas. *Cell Syst*. 2015;1(1):88-96. doi:10.1016/j.cels.2015.02.001.

65. Ronda C, Maury J, Jakočiunas T, et al. CrEdit: CRISPR mediated multi-loci gene integration in Saccharomyces cerevisiae. *Microb Cell Fact*. 2015;14(1):97. doi:10.1186/s12934-015-0288-3.

66. Ji H, Moore DP, Blomberg MA, et al. Hotspots for unselected Ty1 transposition events on yeast chromosome III are near tRNA genes and LTR sequences. *Cell*. 1993;73(5):1007-1018.

67. Admire A, Shanks L, Danzl N, et al. Cycles of chromosome instability are associated with a fragile site and are increased by defects in DNA replication and checkpoint controls in yeast. *Genes Dev*. 2006;20(2):159-173. doi:10.1101/gad.1392506.

68. Shen Y, Stracquadanio G, Wang Y, et al. SCRaMbLE generates designed combinatorial stochastic diversity in synthetic chromosomes. *Genome Res*. 2016;26(1):36-49. doi:10.1101/gr.193433.115.

69. Parenteau J, Durand M, Morin G, et al. Introns within Ribosomal Protein Genes Regulate the Production and Function of Yeast Ribosomes. *Cell*. 2011;147(2):320-331. doi:10.1016/j.cell.2011.08.044.

70. Richardson SM, Mitchell LA, Stracquadanio G, et al. Design of a synthetic yeast genome. *Science*. 2017;355(6329):1040-1044. doi:10.1126/science.aaf4557.

71. Tettelin H, Agostoni Carbone ML, Albermann K, et al. The nucleotide sequence of Saccharomyces cerevisiae chromosome VII. *Nature*. 1997;387(6632 Suppl):81-84.

72. Parenteau J, Durand M, Veronneau S, et al. Deletion of Many Yeast Introns Reveals a Minority of Genes that Require Splicing for Function. *Mol Biol Cell*. 2008;19(5):1932-1941. doi:10.1091/mbc.E07-12-1254.

73. Dunn DA, Feygin I. Challenges and solutions to ultra-high-throughput screening assay miniaturization: submicroliter fluid handling. *Drug Discov Today*. 2000;5(12 Suppl 1):84-91.

74. Szita N, Polizzi K, Jaccard N, Baganz F. Microfluidic approaches for systems and synthetic biology. *Current Opinion in Biotechnology*. 2010;21(4):517-523. doi:10.1016/j.copbio.2010.08.002.

75. Kong DS, Carr PA, Chen L, Zhang S, Jacobson JM. Parallel gene synthesis in a microfluidic device. *Nucleic Acids Res*. 2007;35(8):e61-e61. doi:10.1093/nar/gkm121.

76. Tewhey R, Warner JB, Nakano M, et al. Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nature Biotechnology*. 2009;27(11):1025-1031. doi:10.1038/nbt.1583.

77. ELLSON R, MUTZ M, BROWNING B, et al. Transfer of low nanoliter volumes between microplates using focused acoustics?automation considerations. *Journal of the Association for Laboratory Automation*. 2003;8(5):29-34. doi:10.1016/S1535-5535(03)00011-X.

78. Quan J, Tian J. Circular polymerase extension cloning for high-throughput cloning of complex and combinatorial DNA libraries. *Nat Protoc*. 2011;6(2):242-251. doi:10.1038/nprot.2010.181.

79. Zhang Y, Werling U, Edelmann W. SLiCE: a novel bacterial cell extract-based DNA cloning method. *Nucleic Acids Res*. 2012;40(8):e55-e55. doi:10.1093/nar/gkr1288.

80. Li MZ, Elledge SJ. Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. *Nat Methods*. 2007;4(3):251-256. doi:10.1038/nmeth1010.

81. Engler C, Gruetzner R, Kandzia R, Marillonnet S. Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. Peccoud J, ed. *PLoS*

ONE. 2009;4(5):e5553. doi:10.1371/journal.pone.0005553.

82. Mitchell LA, Phillips NA, Lafont A, Martin JA, Cutting R, Boeke JD. qPCRTag Analysis - A High Throughput, Real Time PCR Assay for Sc2.0 Genotyping. *JoVE*. 2015;(99):1-7. doi:10.3791/52941.

83. Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature Publishing Group*. 2004;428(6983):617-624. doi:10.1038/nature02424.

84. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3):R25. doi:10.1186/gb-2009-10-3-r25.

85. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303. doi:10.1101/gr.107524.110.

86. Parenteau J, Durand M, Véronneau S, et al. Deletion of many yeast introns reveals a minority of genes that require splicing for function. *Mol Biol Cell*. 2008;19(5):1932-1941. doi:10.1091/mbc.E07-12-1254.

87. Mercy G, Mozziconacci J, Scolari VF, et al. 3D organization of synthetic and scrambled chromosomes. *Science*. 2017;355(6329). doi:10.1126/science.aaf4597.

88. Goffeau A, Barrell BG, Bussey H, et al. Life with 6000 genes. *Science*. 1996;274(5287):546–563–7.

89. Mewes HW, Albermann K, Bähr M, et al. Overview of the yeast genome. *Nature*. 1997;387(6632 Suppl):7-65. doi:10.1038/42755.

90. Mitchell LA, Cai Y, Taylor M, et al. Multichange isothermal mutagenesis: a new strategy for multiple site-directed mutations in plasmid DNA. *ACS Synth Biol*. 2013;2(8):473-477. doi:10.1021/sb300131w.

91. Heeren G, Rinnerthaler M, Laun P, et al. The mitochondrial ribosomal protein of the large subunit, Afo1p, determines cellular longevity through mitochondrial back-

signaling via TOR1. *Aging (Albany NY)*. 2009;1(7):622-636. doi:10.18632/aging.100065.

92. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, et al. ViennaRNA Package 2.0. *Algorithms Mol Biol*. 2011;6(1):26. doi:10.1186/1748-7188-6-26.

93. Rychlik W, Spencer WJ, Rhoads RE. Optimization of the annealing temperature for DNA amplification in vitro. *Nucleic Acids Res*. 1990;18(21):6409-6412.

94. Boeke JD, LaCroute F, Fink GR. A positive selection for mutants lacking orotidine-5'-phosphate decarboxylase activity in yeast: 5-fluoro-orotic acid resistance. *Mol Gen Genet*. 1984;197(2):345-346.

95. Güldener U, Heck S, Fielder T, Beinhauer J, Hegemann JH. A new efficient gene disruption cassette for repeated use in budding yeast. *Nucleic Acids Res*. 1996;24(13):2519-2524.

96. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760. doi:10.1093/bioinformatics/btp324.

97. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352.

98. Milne I, Stephen G, Bayer M, et al. Using Tablet for visual exploration of second-generation sequencing data. *Brief Bioinformatics*. 2013;14(2):193-202. doi:10.1093/bib/bbs012.

99. Nickerson D. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res*. 1997;25(14):2745-2751. doi:10.1093/nar/25.14.2745.

100. Herschleb J, Ananiev G, Schwartz DC. Pulsed-field gel electrophoresis. *Nat Protoc*. 2007;2(3):677-684. doi:10.1038/nprot.2007.94.

101. Alexandrov A, Chernyakov I, Gu W, et al. Rapid tRNA decay can result from lack of

nonessential modifications. *Mol Cell*. 2006;21(1):87-96. doi:10.1016/j.molcel.2005.10.036.

102. Dewe JM, Whipple JM, Chernyakov I, Jaramillo LN, Phizicky EM. The yeast rapid tRNA decay pathway competes with elongation factor 1A for substrate tRNAs and acts on tRNAs lacking one or more of several modifications. *RNA*. 2012;18(10):1886-1896. doi:10.1261/rna.033654.112.

103. Kim SK, Noh YH, Koo J-R, Yun HS. Effect of expression of genes in the sphingolipid synthesis pathway on the biosynthesis of ceramide in Saccharomyces cerevisiae. *J Microbiol Biotechnol*. 2010;20(2):356-362.

104. Tanigawa M, Kihara A, Terashima M, Takahara T, Maeda T. Sphingolipids regulate the yeast high-osmolarity glycerol response pathway. *Mol Cell Biol*. 2012;32(14):2861-2870. doi:10.1128/MCB.06111-11.

105. Fernius J, Marston AL. Establishment of Cohesion at the Pericentromere by the Ctf19 Kinetochore Subcomplex and the Replication Fork-Associated Factor, Csm3. Lichten M, ed. *PLoS Genet*. 2009;5(9):e1000629–17. doi:10.1371/journal.pgen.1000629.

106. Di Rienzi SC, Collingwood D, Raghuraman MK, Brewer BJ. Fragile genomic sites are associated with origins of replication. *Genome Biol Evol*. 2009;1(0):350-363. doi:10.1093/gbe/evp034.

107. Bloom-Ackermann Z, Navon S, Gingold H, Towers R, Pilpel Y, Dahan O. A Comprehensive tRNA Deletion Library Unravels the Genetic Architecture of the tRNA Pool. Copenhaver GP, ed. *PLoS Genet*. 2014;10(1):e1004084–16. doi:10.1371/journal.pgen.1004084.

108. Costanzo M, Baryshnikova A, Myers CL, Andrews B, Boone C. Charting the genetic interaction map of a cell. *Current Opinion in Biotechnology*. 2011;22(1):66-74. doi:10.1016/j.copbio.2010.11.001.

109. Tong AH, Evangelista M, Parsons AB, et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*. 2001;294(5550):2364-2368. doi:10.1126/science.1065810.

110. Winzeler EA, Shoemaker DD, Astromoff A, et al. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science*. 1999;285(5429):901-906.

111. Müller CA, Hawkins M, Retkute R, et al. The dynamics of genome replication using deep sequencing. *Nucleic Acids Res*. 2014;42(1):e3-e3. doi:10.1093/nar/gkt878.

112. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009;25(9):1105-1111. doi:10.1093/bioinformatics/btp120.

113. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010;11(10):R106. doi:10.1186/gb-2010-11-10-r106.

114. Wen B, Zhou R, Feng Q, Wang Q, Wang J, Liu S. IQuant: an automated pipeline for quantitative proteomics based upon isobaric tags. *Proteomics*. 2014;14(20):2280-2285. doi:10.1002/pmic.201300361.

115. Xia J, Wishart DS. Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. *Nat Protoc*. 2011;6(6):743-760. doi:10.1038/nprot.2011.319.

116. Jewison T, Knox C, Neveu V, et al. YMDB: the Yeast Metabolome Database. *Nucleic Acids Res*. 2012;40(Database issue):D815-D820. doi:10.1093/nar/gkr916.

117. Dymond J, Boeke J. The Saccharomyces cerevisiae SCRaMbLE system and genome minimization. *Bioeng Bugs*. 2012;3(3):168-171. doi:10.4161/bbug.19543.

118. Sauer B. Identification of cryptic lox sites in the yeast genome by selection for Cre-mediated chromosome translocations that confer multiple drug resistance. *J Mol Biol*. 1992;223(4):911-928.

119. Hoess R, Wierzbicki A, Abremski K. Formation of small circular DNA molecules via an in vitro site-specific recombination system. *Gene*. 1985;40(2-3):325-329.

120. Asan, Geng C, Chen Y, et al. Paired-end sequencing of long-range DNA fragments for de novo assembly of large, complex Mammalian genomes by direct intra-molecule ligation. Aboobaker AA, ed. *PLoS ONE*. 2012;7(9):e46211.

doi:10.1371/journal.pone.0046211.

121. Giaever G, Chu AM, Ni L, et al. Functional profiling of the Saccharomyces cerevisiae genome. *Nature*. 2002;418(6896):387-391. doi:10.1038/nature00935.

122. Futcher AB. Copy number amplification of the 2 micron circle plasmid of Saccharomyces cerevisiae. *J Theor Biol*. 1986;119(2):197-204.

123. Futcher AB. The 2 micron circle plasmid of Saccharomyces cerevisiae. *Yeast*. 1988;4(1):27-40. doi:10.1002/yea.320040104.

124. Watanabe T, Tanabe H, Horiuchi T. Gene amplification system based on double rolling-circle replication as a model for oncogene-type amplification. *Nucleic Acids Res*. 2011;39(16):e106-e106. doi:10.1093/nar/gkr442.

125. Hagerman P. Flexibility Of Dna. *Annual Review of Biophysics and Biomolecular Structure*. 1988;17(1):265-286. doi:10.1146/annurev.biophys.17.1.265.

126. Rippe K. Making contacts on a nucleic acid polymer. *Trends in Biochemical Sciences*. 2001;26(12):733-740. doi:10.1016/S0968-0004(01)01978-8.

127. Luo Z, Wang L, Wang Y, et al. Identifying and characterizing SCRaMbLEd synthetic yeast using ReSCuES. *Nature Communications*. May 2018:1-10. doi:10.1038/s41467-017-00806-y.

128. Liu W, Luo Z, Wang Y, et al. Rapid pathway prototyping and engineering using in vitro and in vivo synthetic genome SCRaMbLE-in methods. *Nature Communications*. May 2018:1-12. doi:10.1038/s41467-018-04254-0.

129. Korbel JO, Urban AE, Affourtit JP, et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science*. 2007;318(5849):420-426. doi:10.1126/science.1149504.

130. Xiao W. *Yeast Protocols*. Vol 313. New Jersey: Humana Press; 2005. doi:10.1385/1592599583.

131. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*.

2012;9(4):357-359. doi:10.1038/nmeth.1923.

132. Wang J, Mullighan CG, Easton J, et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat Methods*. 2011;8(8):652-654. doi:10.1038/nmeth.1628.

133. Lindstrom DL, Gottschling DE. The mother enrichment program: a genetic system for facile replicative life span analysis in Saccharomyces cerevisiae. *Genetics*. 2009;183(2):413–22–1SI–13SI. doi:10.1534/genetics.109.106229.

134. Chari R, Church GM. Beyond editing to writing large genomes. *Nat Rev Genet*. 2017;18(12):749-760. doi:10.1038/nrg.2017.59.

135. Suzuki K, Tsunekawa Y, Hernandez-Benitez R, et al. In vivo genome editing via CRISPR/Cas9 mediated homology-independent targeted integration. *Nature Publishing Group*. 2016;540(7631):144-149. doi:10.1038/nature20565.

136. Rancati G, Moffat J, Typas A, Pavelka N. Emerging and evolving concepts in gene essentiality. *Nat Rev Genet*. 2018;19(1):34-49. doi:10.1038/nrg.2017.74.

137. Lajoie MJ, Rovner AJ, Goodman DB, et al. Genomically recoded organisms expand biological functions. *Science*. 2013;342(6156):357-360. doi:10.1126/science.1241459.

138. Zhang Y, Ptacin JL, Fischer EC, et al. A semi-synthetic organism that stores and retrieves increased genetic information. *Nature Publishing Group*. 2017;551(7682):644-647. doi:10.1038/nature24659.

139. The NHGRI-EBI GWAS Catalog, a curated resource of SNP-trait associations. doi:10.6019/tol.gwas-w.2017.00001.1.

140. Wimmer E, Mueller S, Tumpey TM, Taubenberger JK. Synthetic viruses: a new opportunity to understand and prevent viral disease. *Nature Publishing Group*. 2009;27(12):1163-1172. doi:10.1038/nbt.1593.