UNIVERSITY OF CALIFORNIA

Los Angeles

Large Insert Vectors for Synthetic Long Read Technologies

A thesis submitted in partial satisfaction

of the requirements for the degree Master of Science

in Bioengineering

by

Michael Anthony Corrado

2016

© Copyright by

Michael Anthony Corrado

2016

ABSTRACT OF THE THESIS

Large Insert Vectors for

Synthetic Long Read Technologies

by

Michael Anthony Corrado

Master of Science in Bioengineering

University of California, Los Angeles, 2016

Professor Gerard Chee Lai Wong, Chair

We are working on the development of a methodology that can increase our ability to read longer DNA fragments accurately at scale from the Illumina next-generation sequencing platform. This method takes advantage of existing technologies in DNA library preparation techniques, in combination with a barcoding strategy that can uniquely identify DNA fragments of up to 40kb. Currently, the method is being tuned on small 2 to 3 kb pUC19 vectors, and existing plasmids such as these are not suited for our proposed methodology. The goal of my project is to create a custom cosmid containing yeast genomic DNA fragments of up to ~30 kb in length, which are barcoded for deconvoluting distinct large DNA sequences. These custom cosmids will be used to validate the sequencing method for larger DNA sequences.

The thesis of Michael Anthony Corrado is approved.

Ren Sun

Dino Di Carlo

Sriram Kosuri

Gerard Chee Lai Wong, Committee Chair

University of California, Los Angeles

2016

Thank you Mother

TABLE OF CONTENTS

Introduction

Next-generation sequencing (NGS) has become a formidable tool for use in biotechnology (1, 2,). This development in sequencing technology followed the success of the Human Genome Project (3, 4). Companies such as Illumina, Life Technologies, Complete Genomics, Roche, Oxford Nanopore Technology, and Pacific Biosciences (5) have advanced the DNA sequencing space allowing for a broad range of biological studies to be accessible through NGS (6). This has been made possible by the multitude of sequencing platforms available (5,6).

Unfortunately, these platforms tend to suffer from a tradeoff between accuracy and read length (6). In applications that sequence and map large repetitive DNA regions such as haplotyping and de novo gene assemblies, a growing need for accurate longer reads at an increased read depth is needed (7). Commercially, Oxford Nanopore Technology and Pacific Biosciences both offer single molecule reading platforms that produce long reads (10kb) but the accuracy of these reads tends to be unfeasible for use (8, 9, 10). Whereas Illumina uses short "shotgun" reads to piece the genome of interest together (9). With Illumina's short DNA reading platform, synthetic long reads (SLRs) have been created using molecular barcoding and existing library preparation techniques to assemble short reads computationally.

This methodology was first leveraged by Hiatt et al. (2010). Known as subassembly, a barcoded adapter is ligated on DNA of interest and amplified. The PCR product is then broken and a second barcoded adapter is ligated. This is then used to create a sequencing library that pieces together 1kb fragments using 36 bp paired end reads (11). Notably, Voskoboynik et al. from Stephen Quake's lab (now the company Moleculo) used dilute pools to barcode and amplify 384 wells of DNA which later was used to piece together 60kb fragments for haplotyping experiments (12, 13). While this is the most used SLR, it is vulnerable to PCR amplification errors, does not uniquely barcode each molecule for upstream referencing, and has a tedious preparation of all 384 wells. Recently, Stapleton et al. (2016) ligated two uniquely different

barcoded adapters to DNA of interest. The DNA was then broken, recircularized and prepared for NGS. These original barcodes allowed for referencing when stitching the data in silico together. Stapleton was only able to generate reads up to 10kb because of PCR length limitations (14).

Currently long-read sequencing techniques are held back by PCR limitations and not uniquely barcoding all molecules downstream. We believe our method known as Barcode-directed Assembly of Genomes Enabled by LoxP (BAGEL-seq) identifies and works around these limitations. The method uniquely barcodes all unknown DNA and is not limited by PCR, but only by library and cloned in fragment size. An adapter which is placed next to a cloning site contains a linearization sequence, a barcode, a NGS primer, and a LoxP site. The DNA of interest will then be cloned in a cosmid and amplified via cloning. This methodology works as follows (Figure 1). 1) The vector is linearized using PI-PspI linearization sequence, 2) highly active Tn5 transposase (15, 16) integrates a second LoxP site randomly throughout multiple copies of the same DNA fragment from step 1, 3) A Cre-recombinase deletion reaction is used to place the LoxP sites together, bringing the barcode and primer to the unknown DNA sequence next to the second LoxP. Assuming even distribution of Tn5 LoxP in multiple clones, this method effectively primer-walks the DNA of interest. This product is then used to create a library for NGS and reads using the primer and barcode are then assembled computationally.

At the moment the method is still in its infancy and being tuned on pUC19 plasmids with small ~1kb DNA fragments cloned in. Since existing plasmids like these are not suitable for creating a protocol for SLRs, I created a custom cosmid pMAC1. With the adapter located next to a blunt end cloning site I cloned in 1kb to 30kb *Saccharomyces cerevisiae* DNA fragments. These custom vectors of up to ~40kb will later be used to validate this sequencing method for larger DNA fragments.

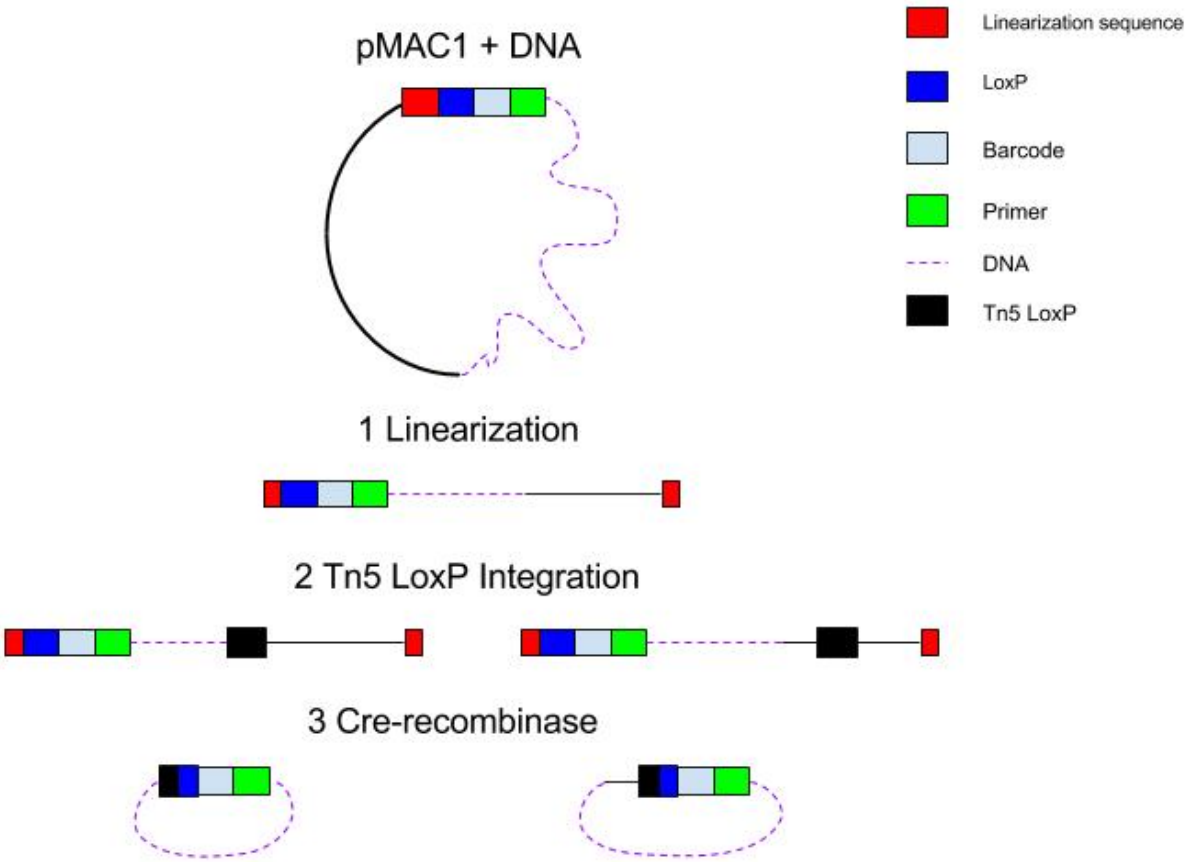

Figure 1. BAGEL-Seq steps: 1) The vector is linearized using PI-PspI linearization sequence, 2) highly active Tn5 transposase integrates in a second LoxP site randomly throughout multiple copies of the same DNA of interest, 3) A Cre-recombinase deletion reaction is used to place the LoxP sites together, bringing the barcode and primer to the unknown DNA sequence next to the second LoxP. Assuming even distribution of Tn5 LoxP in multiple clones, this method effectively primer-walks the DNA of interest.

Results

The plasmid came linear and dephosphorylated at the blunt PmlI cloning site. The plasmid was phosphorylated and ligated back together. The circular plasmid was then cut at the PmlI restriction site and ran on an 1% agarose gel to check length and quality of blunt end ligation (Figure 2A). The plasmid was then amplified using primers Forward Primer EpiBio and Reverse Primer EpiBio. These primers added the restriction sites AatII and XhoI while deleting the unwanted LoxP site that came on the plasmid (Figure S1). The PCR product was then ran on a 1% agarose gel and the 8kb band was extracted (Figure 2B). The plasmid and adaptor were then separately digested with AatII and XhoI restriction enzymes. The plasmid and adapter were ligated at a 1:3 ratio. For barcode integration the plasmid was digested with NheI. The barcode was digested with XbaI and SpeI. The backbone was then dephosphorylated. The plasmid and barcode were ligated at a 1:3 ratio. The plasmid was also self-ligated as a control. 100uL of the transformations were then diluted, plated and grown overnight to get an estimate of the library diversity, and the remaining 900uL library was grown in 100 mL of liquid culture for 20 hours at 37^{o}C. The estimated library size is 700,000 unique barcodes with 3000 self-ligated control plasmids. The full plasmid was then Sanger sequenced. The PmlI blunt cloning site was then digested and a linearized pMAC1 library was analyzed on a 1% agarose gel for blunt end cloning (Figure 2C).

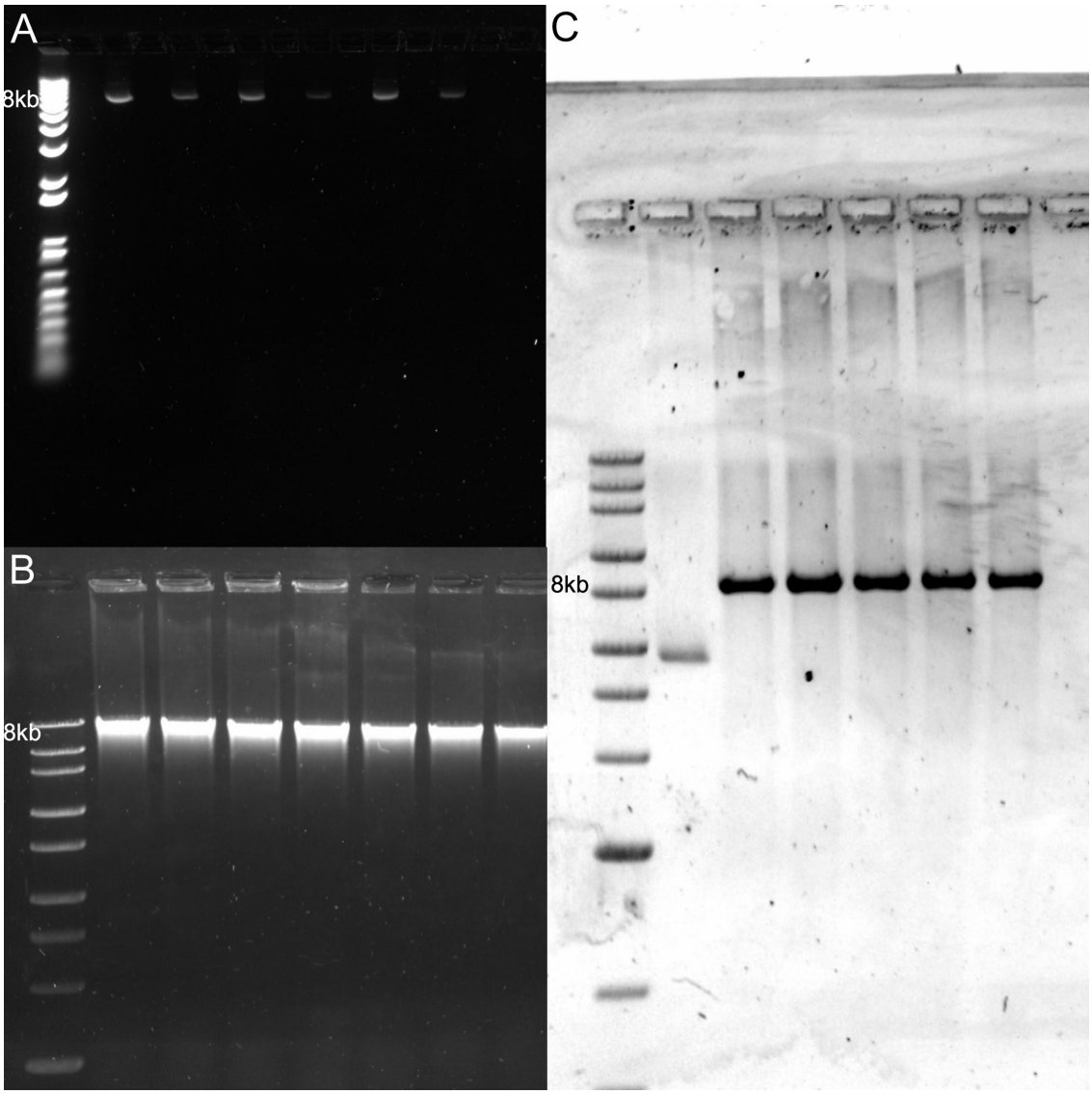

Figure 2. A) After blunt end ligation, Epicenter pCC2FOS was digested at the PmlI site and ran on a 1% agarose gel to check length and cloning site quality. B) PCR of circularized Epicenter pCC2FOS to remove unwanted LoxP site with addition of restriction sites AatII and XhoI ran on 1% agarose gel. 8kb fragments extracted for adapter integration. C) After adapter integration, PmlI cloning site checked in pMAC1 library. First well corresponds to undigested supercoiled pMAC1, and the remaining 5 wells correspond to pMAC1 digested at PmlI site.

For the integration of the yeast fragments to create the BACs, pMAC1 was digested with PmlI and dephosphorylated. pMAC1 was then ligated to the *Saccharomyces cerevisiae* genomic fragments at 1:1 ratio. The plasmid was also self-ligated as a control. The 1mL transformed products were then plated as dilutions, 700 uL, 250 uL, and 50 uL. Only 4 control colonies were obtained. The BACs were then sent to Genewiz for Sanger sequencing using 2 primers that flanked the clone site. The first Primer Skpp-1-F, annealing to the adaptor, reads the barcode and the start of the clone site in a clockwise direction. The second primer BAC Reverse Clone Site, annealing to the backbone, reads the clone site in a counter-clockwise direction. Both primers obtain ~1kb reads into the unknown DNA. Sequence alignment resulting from Sanger sequencing was performed using BLAST against the genome of *Saccharomyces cerevisiae* YJM987. Using the Sanger sequencing data and analysis of the BACs by DNA gel electrophoresis, fragments were found to range from 1kb to 30kb (Figures 4A and 4B).

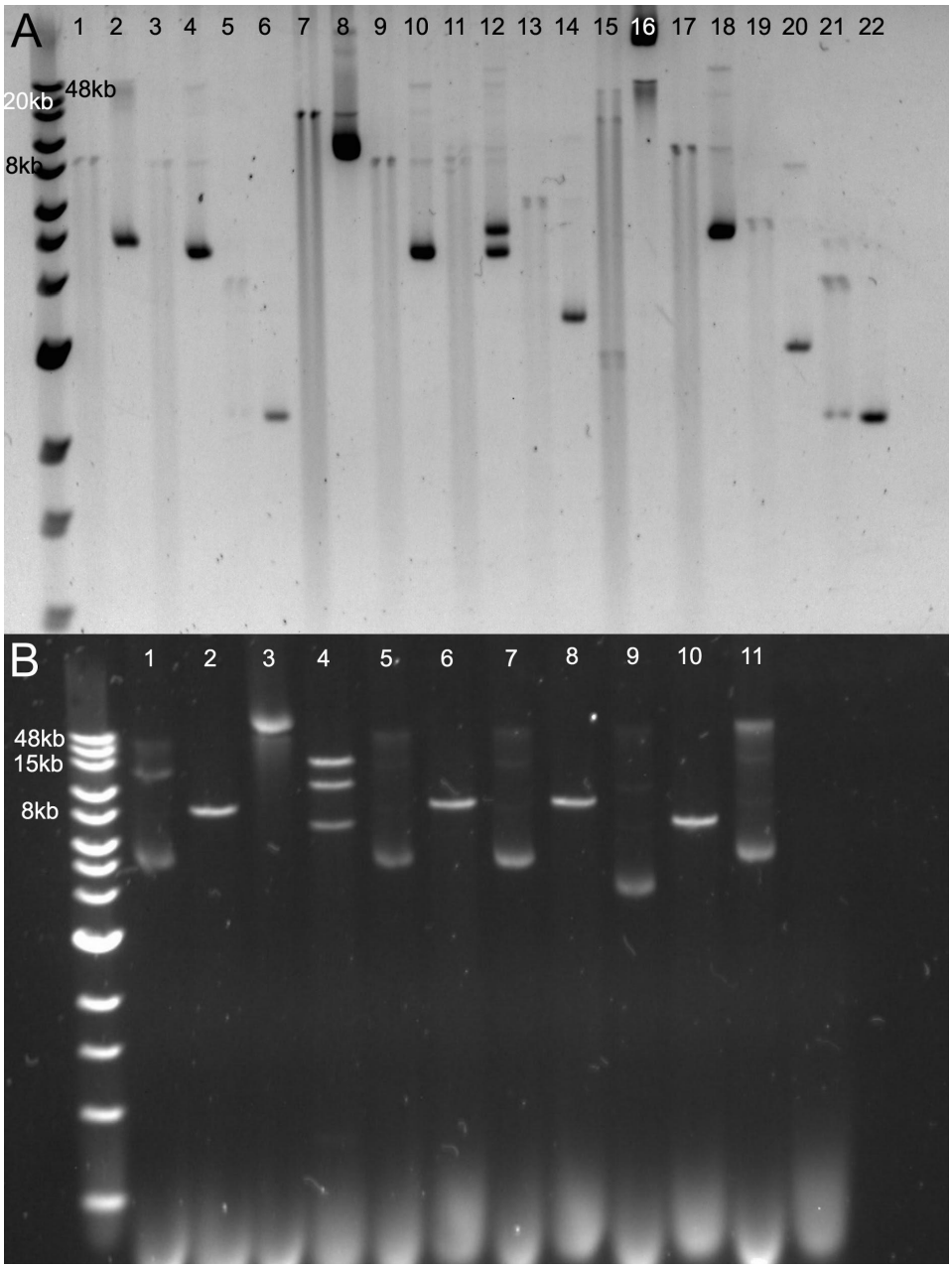

Figure 3. BACs were analyzed on a 0.6% agarose gel. All BACs digested with AfeI with control pMAC1 in lanes 1 and 2 ran as digested versus undigested. A) Lanes 7 and 8: 5kb inserts; Lanes 15 and 16 : 25kb inserts; all other lanes : ~1kb inserts. B) Inserts are predominantly ~1kb. Lanes 3 and 4: 29kb inserts.

Conclusion

With current long read platforms having a high error rate, researchers have moved towards SLRs based on accurate short reading platforms. Current SLR techniques have limitations in molecular barcoding and amplification with the additional downside of tedious library preparation. We are developing BAGEL-seq, to combat these limitations, which can potentially be extended to *in vitro* applications once established. To advance BAGEL-seq to longer DNA fragments than ~1kb, I created a custom BAC plasmid pMAC1 , and *Saccharomyces cerevisiae* genomic DNA of up to 30kb were introduced into the plasmid and subsequently characterized. These BACs are applicable for use with Tn5-LoxP inserts, Cre recombinase, and next-generation sequencing.

Methods

Adapter

The adapter was purchased as a gene block from IDT. The adapter includes the following components from 5' to 3': 4 base pair overhang, XhoI site, PI-PspI linearization sequence, skpp-1-F primer, Lox66 mutant site, NheI site, skpp-2-R primer, AatII site, 4 base-pair overhang. The gBlock was inserted into a standard TOPO plasmid pCR™-Blunt II-TOPO® by TOPO cloning. *pCR™-Blunt II-TOPO®* with the gBlock was then transformed into NEB 5-alpha Competent *E. coli*. Cells were plated and grown overnight at 37^{o}C. 12 colonies were picked and inoculated each in 5 mL liquid culture. Liquid cultures were grown overnight at 37^{o}C and miniprepped. The miniprepped plasmids were then sent to Genewiz for Sanger sequencing. The perfect LoxP sequence was picked and saved as backup supply. The gBlock was then amplified with PCR using primers Forward Tag PCR and Reverse Tag PCR. The PCR product was then checked on a 4% E-gel for length and quality (figure 4). The adapter was then purified and eluted into water.

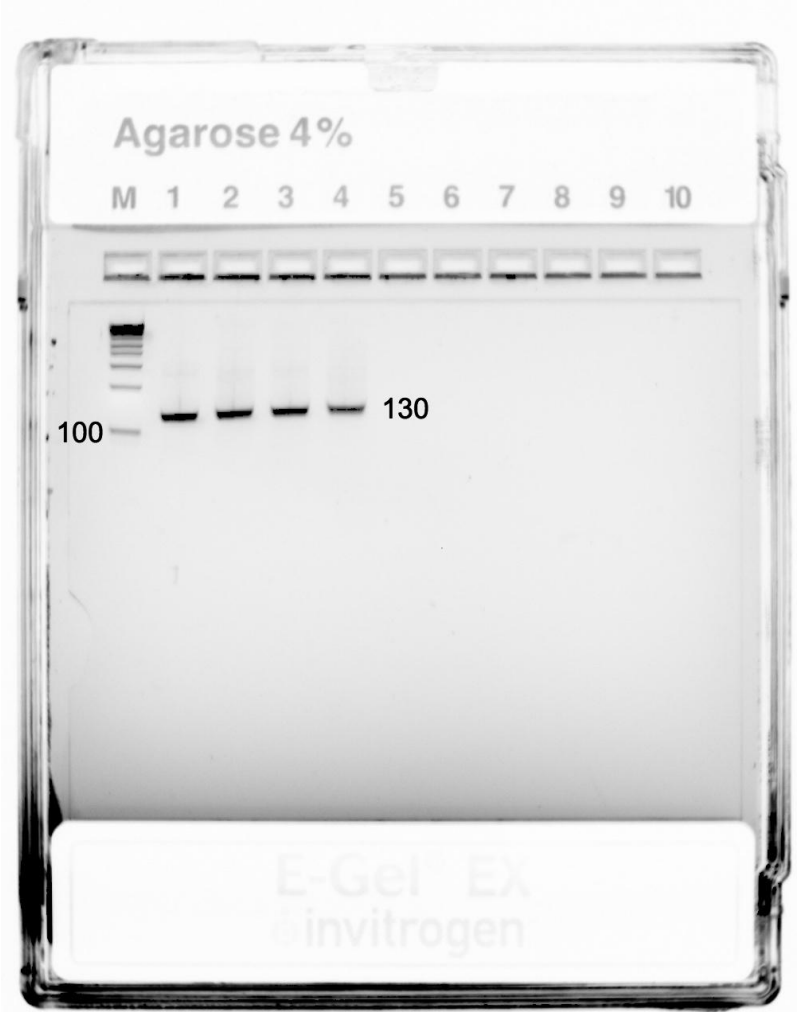

Figure 4. 4% agarose E-gel used to check the length of the adapter after PCR. Bottom of the ladder corresponds 100bp, adapter length is ~130bp.

Barcoded insert

To create the barcode, a 61 bp oligonucleotide was synthesized with 15 randomized bases (IDT). This oligo contained from 5' to 3': primer Skpp-8-F, restriction site XbaI, 15 randomized bases, and primer Skpp-8-R. The barcode was then PCRed using primers Skpp-8-F and Barcode Reverse hang off. The Barcode Reverse Hang Off primer contained from 5' to 3': primer Skpp-8-R, primer Skpp-9-R, and restriction site SpeI with a 4 base overhang. To contain

the complexity of the barcode, only 15 PCR cycles were done and the quality of the DNA was checked on a 4% E-Gel (figure 5). The PCR product was purified and eluted in 10uL water. The barcode insert was purposely created to be 91 bp long to increase recovery yield after restriction enzyme digestion.

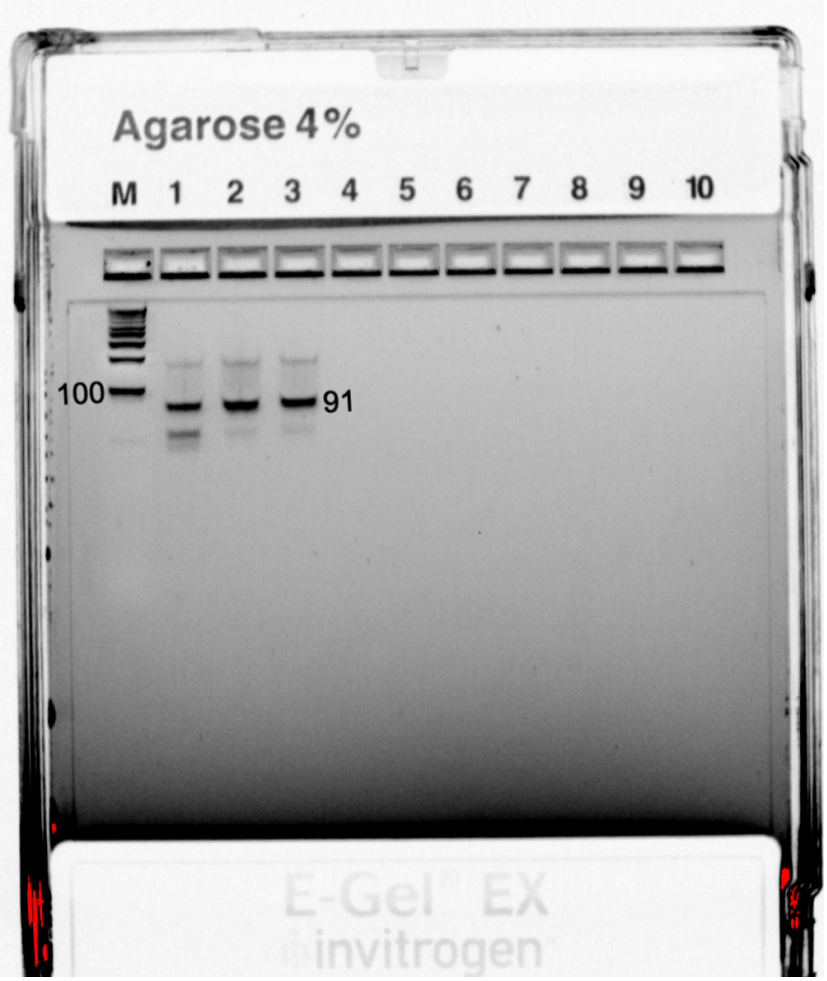

Figure 5. PCR product of barcode oligos ran on 4% agarose E-Gel. Primer-dimers seen in all lanes. Bottom of ladder bar corresponds to 100bp, barcode length: ~91bp.

Yeast Genomic DNA Prep

Saccharomyces cerevisiae was plate-streaked and grown overnight at 30°C. The plate culture was then used to inoculate 50 mL of liquid culture. Liquid cultures were grown at 30°C for 16-20 hours. To extract the yeast genomic DNA the yeast cells were digested using Zymo Yeastar kit. The DNA was then end repaired using Invitrogen DNA End Repair Mix. The end repair reaction was then analyzed on a 0.6% agarose gel at 15 volts for 24 hours (Figure 6).

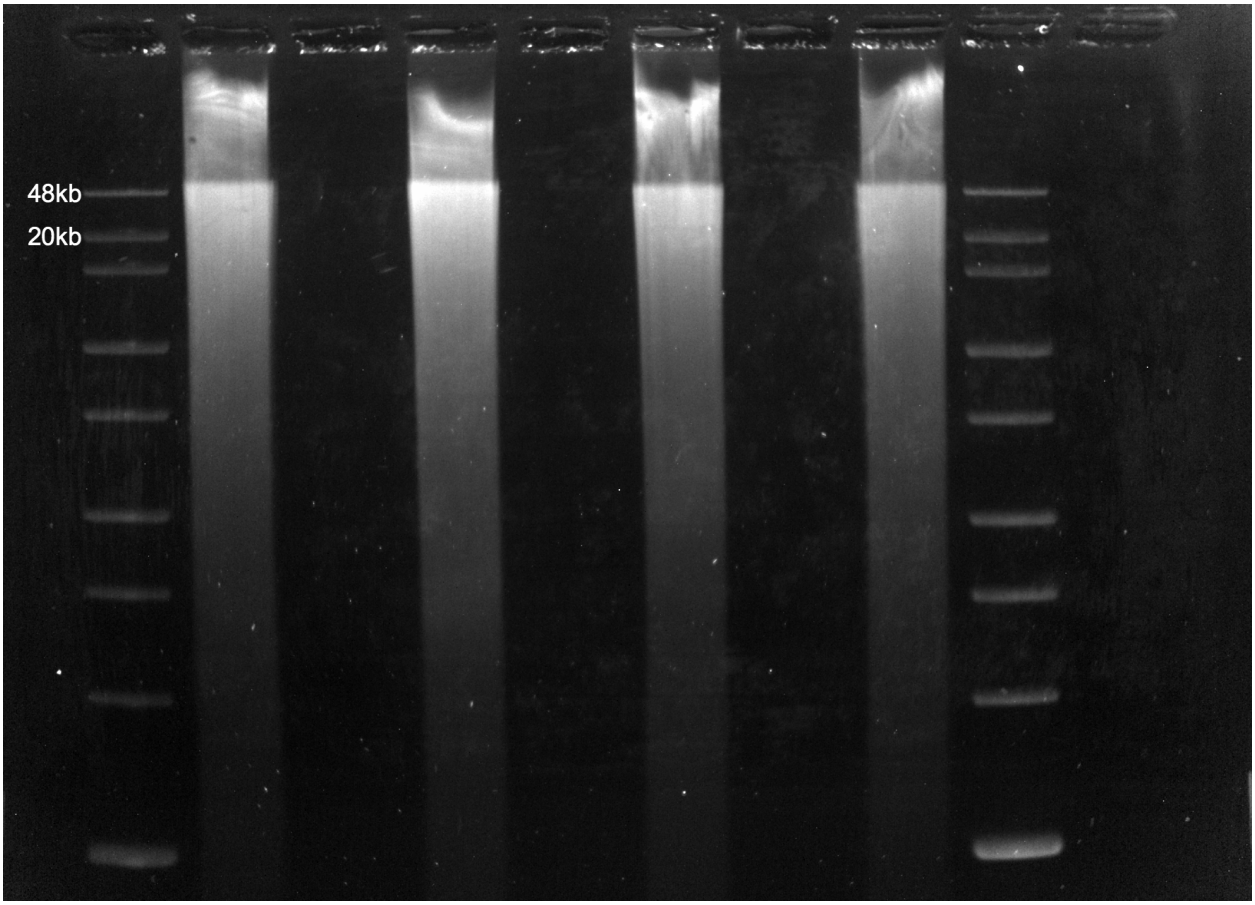

Figure 6. End repaired yeast fragments ran on a 0.6% agarose gel for 24 hours at 15V. The top ladder band of 48kb and the second band at 20kb were used as the extraction point makers for excision of the gel slice.

Using a size standard, the DNA was gel-extracted between the 48kb to 20kb markers. The gel slice was then placed into Thermo Scientific™ SnakeSkin™ Dialysis Tubing, 10K MWCO (16 mm) and electro-eluted. After electro-elution, the extracted gDNA fragments were EtOH precipitated. The quality, quantity, and length of the DNA was then verified on a 0.6% agarose gel (Figure 7).

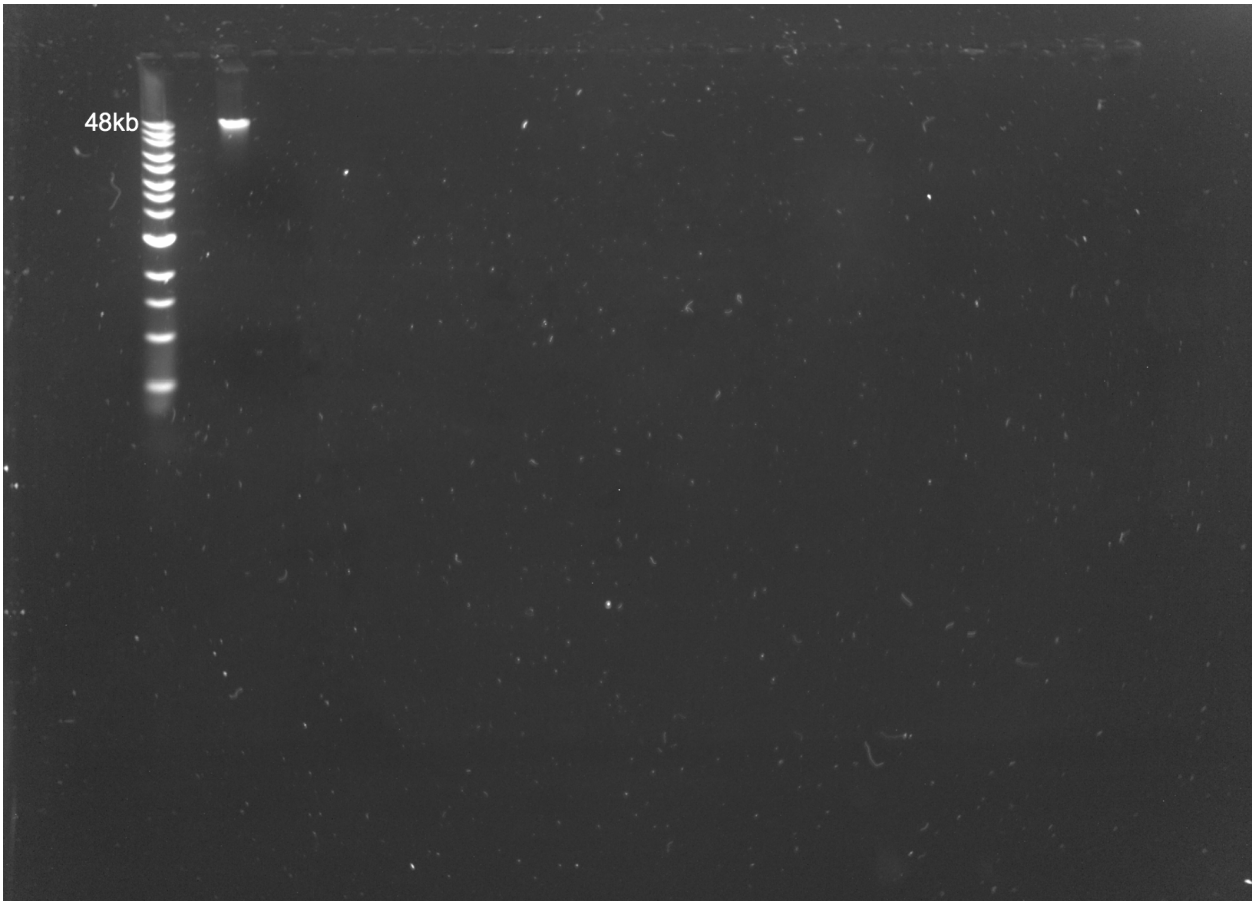

Figure 7. After electroelution of the yeast fragment, the length, quality, and quantity of the DNA was analyzed on a 0.6% agarose gel (5 hours,40V). The top ladder band corresponds to 48kb and 30ng.

pMAC1

The original 8kb plasmid (pCC2FOS) was acquired from Epicentre. The plasmid came linear and dephosphorylated at the blunt PmlI cloning site. The plasmid was phosphorylated using NEB T4 Polynucleotide Kinase and ligated back together. The ligation was purified using

Qiagen MinElute PCR Purification Kit and transformed into NEB 5-alpha Competent *E. coli.* The cells were then plated on blue/white screening plates, as the circular plasmid contained a completed LacZ gene, and grown overnight at 37^{o}C. 12 blue colonies were picked and used to inoculate 12 5 mL liquid culture tubes. Liquid cultures were grown overnight and mini-prepped. The circular plasmid was then cut at the PmlI restriction site and analyzed on an 1% agarose gel to check its length and quality (Figure 9).

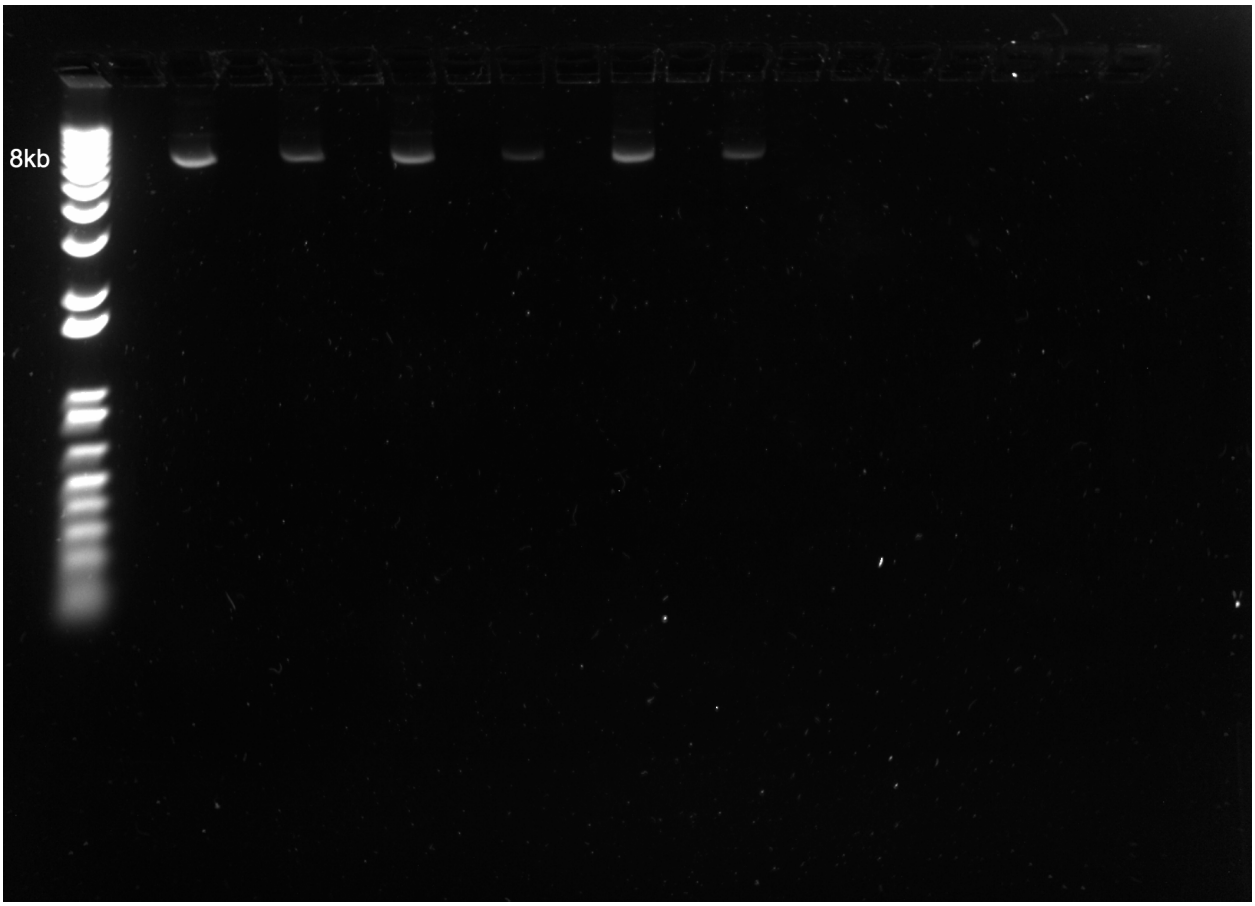

Figure 8. After self-ligation of Epicenter pCC2FOS at the PmlI site, the plasmid was digested and ran on a 1% agarose gel. Checking for cloning site quality, and plasmid quantity.

The plasmid was then amplified using primers Forward Primer EpiBio and Reverse Primer EpiBio. These primers added the restriction sites AatII and XhoI while removing the unwanted LoxP site on the plasmid. The PCR product was then analyzed on a 1% agarose gel and the 8kb band was excised from the gel(Figure 9).

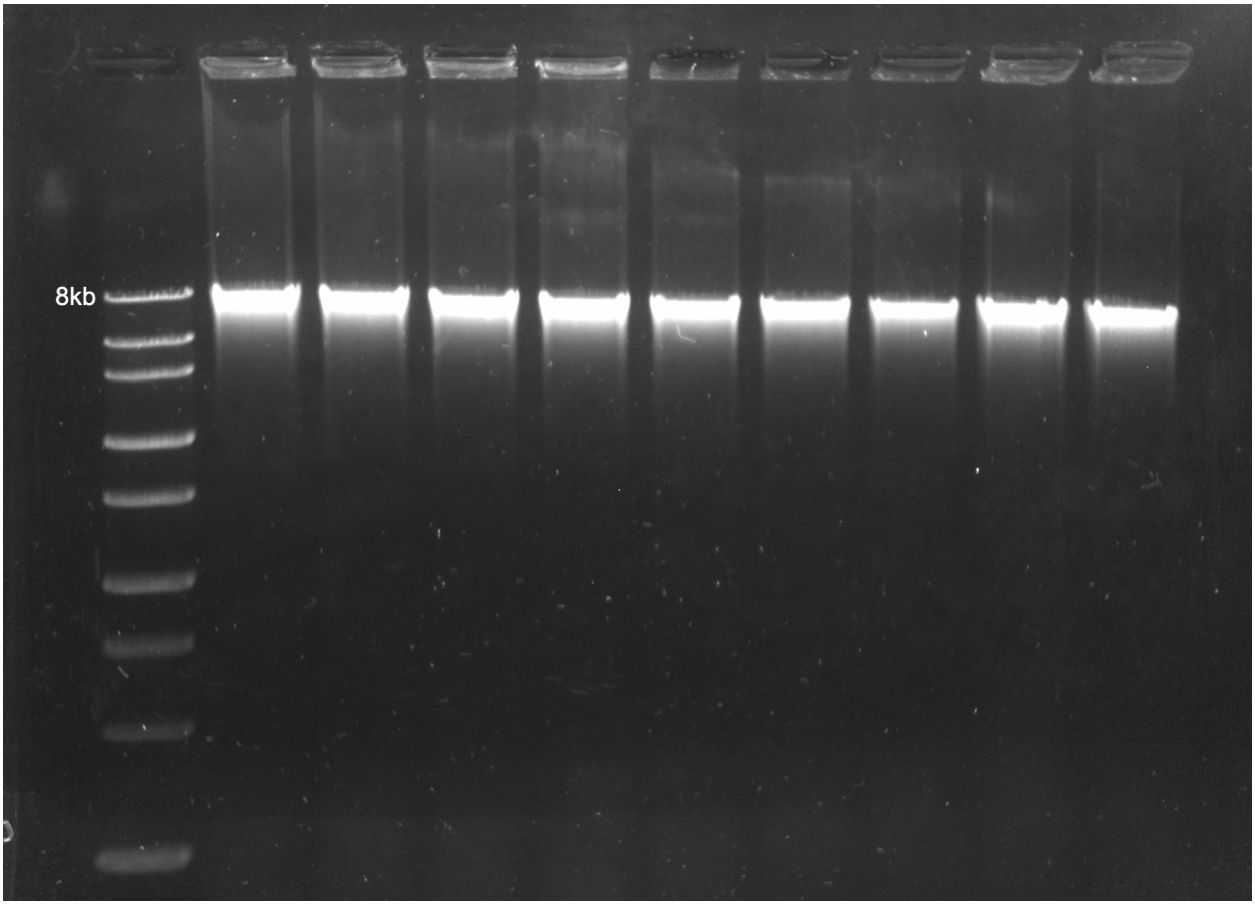

Figure 9. Plasmid ran on 1% agarose gel after Epicenter pCC2FOS PCR deletion of unwanted LoxP site and addition of restriction sites AatII and XhoI. Plasmid extracted from gel for further preparation.

The DNA was recovered from the gel slice using Qiagen Gel Extraction kit. The plasmid and adapter were then separately digested with AatII and XhoI. The backbone was then dephosphorylated using NEB Shrimp Alkaline Phosphatase (rSAP). The digestions were then cleaned with Qiagen MinElute PCR Purification Kit, and the plasmid and adapter were ligated

together at a 3:1 ratio. The digested plasmid was also self-ligated without insert as a control. Both ligations were cleaned with Qiagen MinElute PCR Purification Kit and transformed using NEB 5-alpha Competent *E. coli.* Cells were plated and grown overnight at 37^{o}C. 12 colonies were picked and used to inoculate 5 mL liquid culture. Liquid cultures were grown overnight at 37^{o}C and miniprepped. 1 mL of each culture was turned into frozen glycerol stock. The miniprepped plasmids were then sent to Genewiz for sequencing using primer Tag Sequence Check. The perfect LoxP site was picked.

For barcode integration the plasmid was digested with NheI. The barcode was then digested with XbaI and SpeI. The backbone was then dephosphorylated using NEB Shrimp Alkaline Phosphatase (rSAP). The digestions were cleaned using Qiagen MinElute PCR Purification Kit. The cleaned products were then ligated together at a 3:1 ratio. The plasmid was also self-ligated as a control. The ligations were then purified using Qiagen MinElute PCR Purification Kit and transformed into competent *E. coli* cells. 100 uL of the transformations were then diluted, plated and grown overnight to get a library size estimate, and the remaining 900 uL library was grown in 100 mL of liquid culture for 20 hours at 37^{o}C. The estimated library size is 700,000 unique barcodes (with 3000 self ligated control plasmids). The liquid culture library was then maxiprepped. 12 colonies were picked from the dilution plates and grown in 5mL liquid culture overnight at 37^{o}C. The picked colonies and the library were sent to Genewiz for sequencing using the primer Tag Sequence Check. The full plasmid was then sequenced using primers: Skpp-1-F, 9_R, Chl 5'_F, Chl 3'_F, redF_F, redF-prime-OriV_F, OriV_F, OriII-prime-repE_F, repE 5'_F, repE 3'_F, ParA_F, ParA mid_F, ParA 3'_F, ParB 5'_F, ParB 3'_F, ParC 5'_F. The PmlI blunt cloning site was then cut and analyzed on a 1% agarose gel (Figure 10)

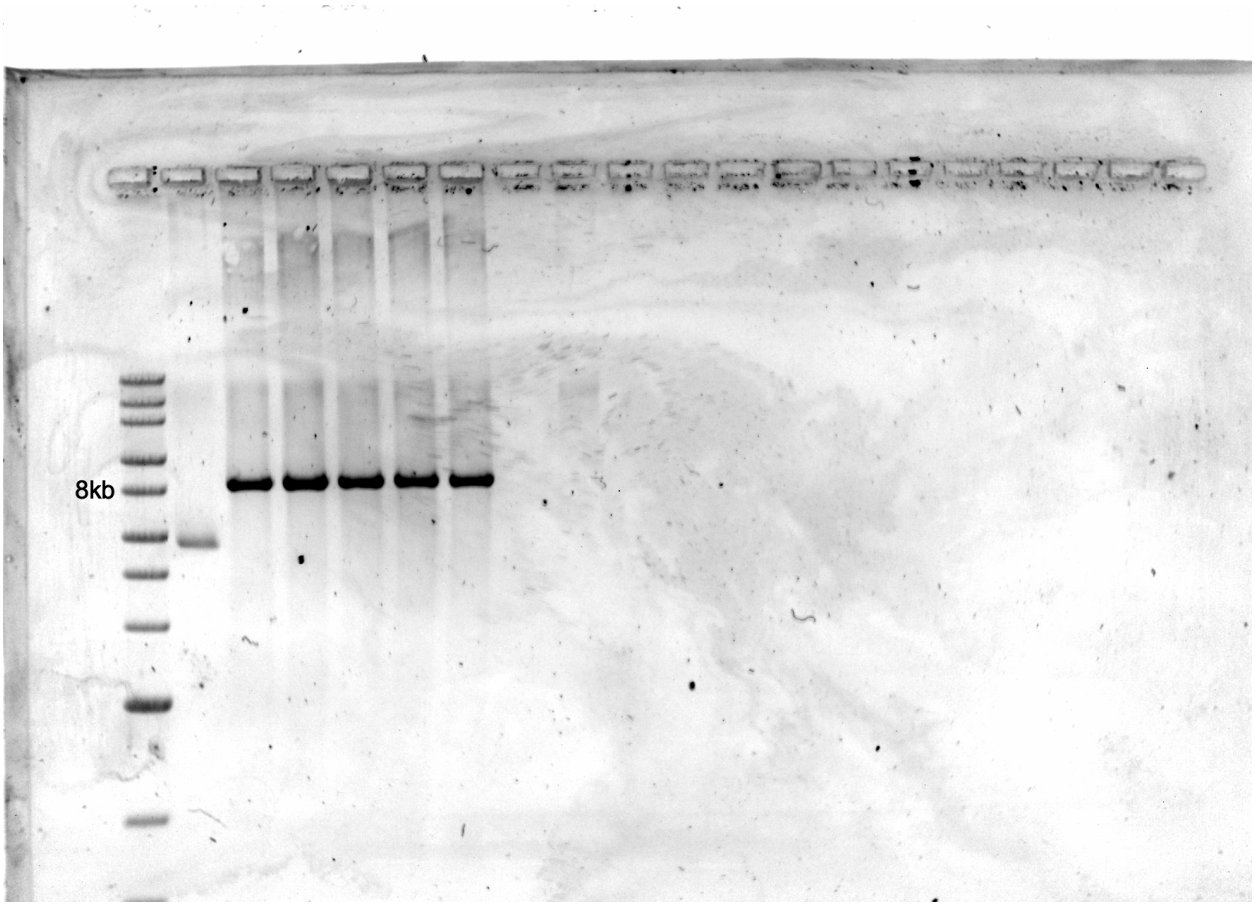

Figure 10. First lane is undigested pMAC1, and the next lanes are PmlI digests of the pMAC1 library. 1% Agarose gel.

Yeast BAC

To create the BAC, pMAC1 was digested with PmlI restriction enzymes and dephosphorylated using NEB Shrimp Alkaline Phosphatase (rSAP). The digestion was cleaned with Qiagen MinElute PCR Purification Kit. pMAC1 was then ligated to the yeast fragments at a 1:1 ratio. The ligation was then cleaned using drop dialysis on a Type-VS Millipore membrane and transformed. The plasmid was also ligated to itself as a control. The 1mL transformed products were then plated as dilutions: 700 uL, 250 uL, 50 uL. The plates were grown overnight at 37^{o}C. Only 4 control colonies were obtained. 12 colonies were picked and used to inoculate 12 5 mL liquid culture tubes. Liquid cultures were grown overnight and miniprepped. 1 mL of each culture

was turned into frozen glycerol stock. BACs were then sent to Genewiz for Sanger sequencing. 2 primers that flanked the clone site were used: Skpp-1-F, found on the adapter, which reads the barcode and the start of the clone site, and the primer BAC Reverse Clone Site. NIH BLAST was used against the sequencing results with the yeast species *Saccharomyces cerevisiae* YJM987. Using the sequencing data and analyzing the BACs on a gel, the fragments were found to range from 1kb to 30kb (Figure 4).

Supplementary Information

Table S1. Primers

Forward Tag PCR	ATGCCTCGAGTGGCAAACAGCTAT
Reverse Tag PCR	TACGGACGTCTGGTAGTAATAAGGGC
Skpp-1-F	ATATAGATGCCGTCCTAGCG
Skpp-2-R	TGGTAGTAATAAGGGCGACC
Skpp-21-F	TCGGATTAGAGTTCTCTCGC
Skpp-21-R	CTGAATGTTCGGGATTTCCC
Skpp-8-F	TAACTACCAGTATCCACCGC
Skpp-8-R	ATGATCCTATGCGTCTGTGT
Skpp-9-R	AATATCGAACAGCTGTGCAC
Forward Primer EpiBio	ATGCGACGTCTGGCTTAACTATGCGGCATC
Reverse Primer EpiBio	ATGCCTCGAGATTAGCGATGAGCTCGGACT
Tag Sequence Check	CAGGAAACAGCCTAGGAAC
9_R	CAGGAAACAGCCTAGGAAC
Chl 5'_F	CATTAAGCATTCTGCCGACATG
redF_F	GTATACCGCTGAAAGTTCTG
Chl 3'_F	GAACGGTCTGGTTATAGGTAC
redF-prime-OriV_F	CCTTTTACAGCCAGTAGTGC
OriV_F	GCACCTATTGACATTTGAGG
Orill-prime-repE_F	TTATCCATGCTGGTTCTAG
repE 5'_F	GTATCTGCGAGATCCATGTTG
repE 3'_F	CACTTCCATGACGACAGGAT
ParA_F	GGATGTTCAGAATGAAACTC
ParA mid_F	GGACGATGTCACTTATGC
ParA 3'_F	CGGATGAAGTTGGTAAAGG
ParB 5'_F	CTGGTGTCATAGAAATTGC
ParB 3'_F	GTTCCAACTGAGTGTATAGAG
ParC 5'_F	CGTCGGTCTGATTATTAGTC

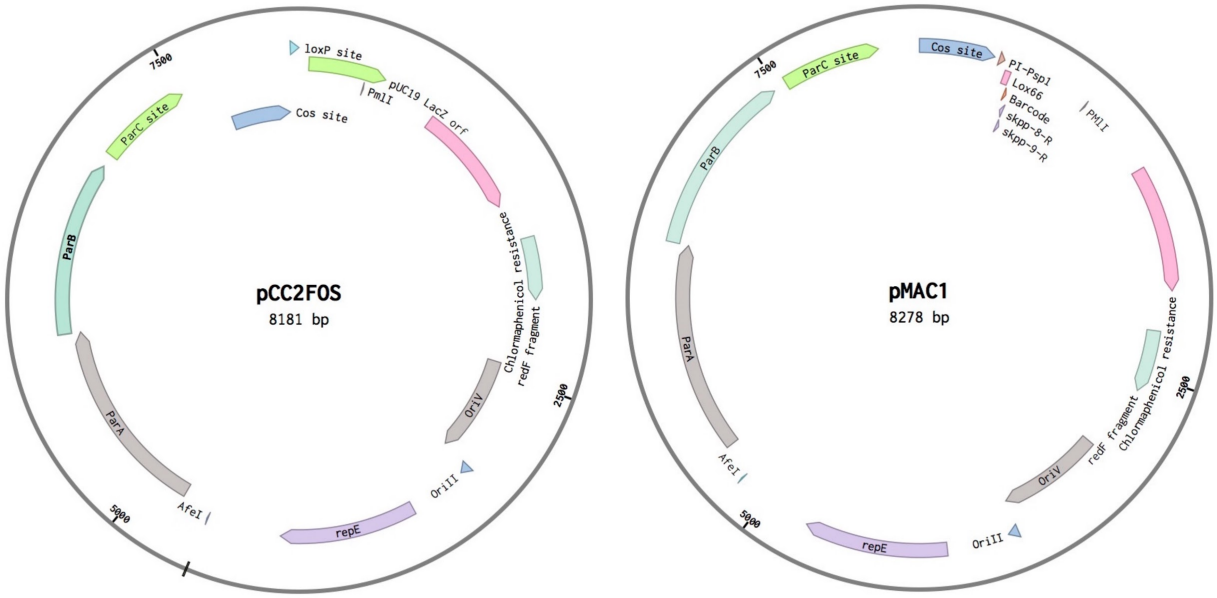

Figure S1. Epicentre pCC2FOS comparison to custom pMAC1. Deletion of LoxP and pUC19 LacZ orf. Addition of adaptor next to blunt PmlI cloning site, which contains: PI-PspI linearization sequence, mutant Lox66, 15bp randomized barcode, and two primers used for NGS and Sanger sequencing.

Protocols

Gel clean up

Qiagen
Standard protocol

Electroelution
See below for protocol

PCR clean up

Qiagen MinElute PCR Purification Kit

Digestion clean up

Qiagen MinElute PCR Purification Kit
Standard protocol

Transformation

NEB 5-alpha Competent *E. coli.*
Standard protocol

Ligation
NEB T4 DNA Ligase
Standard protocol with ratio changes as noted

Digestion
NEB restriction enzymes
Standard protocol

Electroelution
Cut two fingers length of dialysis tubing and place the tubing bag in a petri-dish of the same buffer as the gel box.
Seal one end with a dialysis clip after folding the end over multiple times.
Insert the gel slice into the tubing and add 200-1000 uL of gel box buffer into the bag.
Seal the other end of the tubing with a second dialysis clip after folding over the end enough to create a strong seal and encapsulate the gel slice in the buffer.
Place the sealed bag in an gel box so that the DNA band is parallel to the electrodes.
Run for 3 to 4 hours forward at 100 volts.
Reverse the bag and run for 5 to 10 minutes at 100 volts.
Remove the buffer from the tubing and place into a 1.5-5 mL microfuge tube.
Ethanol precipitate the DNA.

8kb PCR
KAPA HiFi HotStart ReadyMix PCR Kit
Protocol on desk

EtOH precipitation
Starting volume 30uL
Add 3M NaOAC (3.2uL) to final [] of 0.3M
Add 87.5ul 100% EtOH to to final [] of 70%
Place tube in -80 for 20 minutes (although you can go over night or even without this step)
Spin at max (15000) for 30 *minutes* at RT
Remove supernatant
Wash once with 500uL of 70% EtOH, making sure to flick/rock the tube to allow the ethanol to penetrate the pellet.
Spin at max (15000) for 5 minutes at RT
Remove supernatant and let the pellet air dry for 5 minutes, typically the pellet will turn clear once dry.
Re-suspend in 4.5uL-20uL of H2O *flick the pellet to allow it to dissolve easier.
*Scale up the reaction volumes as needed.

rSAP
NEB Shrimp Alkaline Phosphatase (rSAP)

Standard protocol

Kinase
NEB T4 Polynucleotide Kinase
Standard protocol

Drop Dialysis
Type-VS Millipore membrane
Standard NEB protocol

TOPO Clone
Invitrogen's Zero Blunt TOPO Clone kit.
Standard protocol

Miniprep
Qiagen Miniprep Kit
Standard protocol used

Maxiprep
Qiagen Maxiprep Kit
Standard protocol used

References

1 Mardis, Elaine R. "The impact of next-generation sequencing technology on genetics." *Trends in genetics* 24.3 (2008): 133-141.

2 Metzker, Michael L. "Sequencing technologies—the next generation." *Nature reviews genetics* 11.1 (2010): 31-46.

3 Shendure, Jay, et al. "Advanced sequencing technologies: methods and goals." *Nature Reviews Genetics* 5.5 (2004): 335-344.

4 Lander, Eric S., et al. "Initial sequencing and analysis of the human genome."*Nature* 409.6822 (2001): 860-921.

5 Shendure, Jay, and Erez Lieberman Aiden. "The expanding scope of DNA sequencing." *Nature biotechnology* 30.11 (2012): 1084-1094.

6 Shendure, Jay, and Hanlee Ji. "Next-generation DNA sequencing." *Nature biotechnology* 26.10 (2008): 1135-1145.

7 Treangen, Todd J., and Steven L. Salzberg. "Repetitive DNA and next-generation sequencing: computational challenges and solutions." *Nature Reviews Genetics* 13.1 (2012): 36-46.

8 Eid, John, et al. "Real-time DNA sequencing from single polymerase molecules." *Science* 323.5910 (2009): 133-138.

9 Quail, Michael A., et al. "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers." *BMC genomics* 13.1 (2012):1.

10 Jain, Miten, et al. "Improved data analysis for the MinION nanopore sequencer." *Nature methods* 12.4 (2015): 351-356.

11 Hiatt, Joseph B., et al. "Parallel, tag-directed assembly of locally derived short sequence reads." *Nature methods* 7.2 (2010): 119-122.

12 Voskoboynik, Ayelet, et al. "The genome sequence of the colonial chordate, Botryllus schlosseri." *Elife* 2 (2013): e00569.

13 Kuleshov, Volodymyr, et al. "Whole-genome haplotyping using long reads and statistical methods." *Nature biotechnology* 32.3 (2014): 261.

14 Stapleton, James A., et al. "Haplotype-phased synthetic long reads from short-read sequencing." *PloS one* 11.1 (2016): e0147229.

15 Reznikoff, William S. "Tn5 as a model for understanding DNA transposition."*Molecular microbiology* 47.5 (2003): 1199-1206.

16 Reznikoff, WILLIAM S. "The Tn5 transposon." *Annual Reviews in Microbiology* 47.1 (1993): 945-964.

17 Araki, Kimi, et al. "Comparative analysis of right element mutant lox sites on recombination efficiency in embryonic stem cells." *BMC biotechnology* 10.1 (2010): 1.

18 Altschul, Stephen F., et al. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic acids research*25.17 (1997): 3389-3402.