Submitted to *Metabolic Engineering, 1-8-2020*

A genetic toolbox for metabolic engineering of *Issatchenkia orientalis*

Mingfeng Cao[a,1], Zia Fatma[a,1], Xiaofei Song[a,b], Ping-Hung Hsieh[d], Vinh G. Tran[a], William L. Lyon[a], Maryam Sayadi[e], Zengyi Shao[f], Yasuo Yoshikuni[d,g,h], Huimin Zhao[a,c,*]

[a]Department of Chemical and Biomolecular Engineering, U.S. Department of Energy Center for Bioenergy and Bioproducts Innovation (CABBI), Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, United States

[b]Department of Microbiology, Nankai University, Tianjin, China

[c]Departments of Chemistry, Biochemistry, and Bioengineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, United States

[d]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, United States

[e]Genome Informatics Facility, Office of Biotechnology, Iowa State University, Ames, IA, 50011, United States

[f]Department of Chemical and Biological Engineering, Iowa State University, Ames, IA, 50011, United States

[g]U.S. Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, United States

[h]Biological Systems Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, United States

Running title: Development of genetic tools for *Issatchenkia orientalis*

[1]M.C. and Z.F. contributed equally to this work.

[#]To whom correspondence should be addressed. Phone: (217) 333-2631. Fax: (217) 333-5052.
E-mail: zhao5@illinois.edu

HIGHLIGHTS

- A 0.8 kb centromere-like sequence was identified from the *I. orientalis* genome and shown to improve plasmid stability.
- A set of constitutive promoters and terminators was discovered and characterized under different culture conditions.
- An efficient *in vivo* DNA assembly method was developed for plasmid and pathway assembly in *I. orientalis*.
-

Abstract

The nonconventional yeast *Issatchenkia orientalis* can grow under highly acidic conditions and has been explored for production of various organic acids. However, its broader application is hampered by the lack of efficient genetic tools to enable sophisticated metabolic manipulations. We recently constructed an episomal plasmid based on the autonomously replicating sequence (ARS) from *Saccharomyces cerevisiae* (ScARS) in *I. orientalis* and developed a CRISPR/Cas9 system for multiplex gene deletions. Here we report three additional genetic tools including: (1) identification of a 0.8 kb centromere-like (CEN-L) sequence from the *I. orientalis* genome by using bioinformatics and functional screening; (2) discovery and characterization of a set of constitutive promoters and terminators under different culture conditions by using RNA-Seq analysis and a fluorescent reporter; and (3) development of a rapid and efficient *in vivo* DNA assembly method in *I. orientalis*, which exhibited ~100% fidelity when assembling a 7 kb-plasmid from seven DNA fragments ranging from 0.7 kb to 1.7 kb. As proof of concept, we used these genetic tools to rapidly construct a functional xylose utilization pathway in *I. orientalis*.

Keywords: *Issatchenkia orientalis*, centromere-like sequence, promoters, terminators, *in vivo* DNA assembly, metabolic engineering

1. Introduction

Recent advances in synthetic biology and metabolic engineering have revolutionized our ability to engineer platform organisms to produce a wide variety of value-added compounds from renewable raw materials (Choi et al., 2019; Du et al., 2011). *Saccharomyces cerevisiae* has been regarded as a preferred workhorse due to its well-characterized physiology and availability of powerful genetic modification tools (Jensen and Keasling, 2015; Nielsen, 2019). However, *S. cerevisiae* is far from being the only yeast of economic importance, and many nonconventional yeasts have emerged as attractive production hosts due to their highly unusual metabolic, biosynthetic, physiological, and fermentative capacities (Löbs et al., 2017; Markham and Alper, 2018; Riley et al., 2016).

Issatchenkia orientalis (also named *Pichia kudriavzevii* or *Candida krusei*) (Douglass et al., 2018), renowned for its high tolerance to multiple stresses (including low pH), has demonstrated its potential as a robust organism for organic acid production. It was previously used for ethanol fermentation at pH 2 (Isono et al., 2012) and engineered to produce D-xylonate (Toivari et al., 2013), succinic acid (Xiao et al., 2014), and D-lactic acid (Park et al., 2018). However, the ability to perform extensive and sophisticated genetic manipulations in *I. orientalis* has been hampered by a lack of genetic tools such as stable episomal plasmids, strong constitutive promoters and terminators, or efficient genome editing systems. We recently discovered that the autonomously replicating sequence (ARS) from *S. cerevisiae* (ScARS) was functional for plasmid replication in *I. orientalis* and the resultant plasmid enabled efficient multiplexed gene disruptions by the CRISPR/Cas9 system in *I. orientalis* (Tran et al., 2019). Nevertheless, the plasmid was not very stable due to the lack of a functional centromere (CEN). CENs are the specialized DNA sequences on each chromosome that promote the formation of the kinetochore, the large multiprotein complex that links the sister chromatids to the spindle microtubules to ensure faithful chromosome segregation during cell division (Cleveland et al., 2003; Malik and Henikoff, 2009; Verdaasdonk and Bloom, 2011). To our knowledge, for the majority of yeast species (e.g., *S. cerevisiae* and *Kluyveromyces lactis*), point CENs contain ~125 bp of DNA and three protein binding motifs (CDEI, CDEII and CDEIII) (Coughlan and Wolfe, 2019; Dujon et al., 2004; Meraldi et al., 2006), while regional CENs possess a large array of binding sites for centromeric proteins, forming multiple CenH3 (CEN-specific histone 3) nucleosomes attached to microtubules within a specific region of the chromosome (Malik and Henikoff, 2009; Steiner and Henikoff, 2015). In *S. cerevisiae* and nonconventional yeast *Scheffersomyces stipitis*, the CEN-ARS endowed plasmids with much higher stability than just ARS sequence alone (Cao et al., 2017a). Therefore, it is desirable to isolate a functional CEN sequence from the *I. orientalis* genome since a CEN is essential to direct precise plasmid segregation.

Promoters and terminators are also important for metabolic engineering endeavors. They are the two essential distinct elements of expression systems and can be rationally designed to achieve the desired regulation or gene expression levels (Redden et al., 2015). A small selection of promoters and terminators, such as *TEF1*, *PDC1*, and *PGK1* has been utilized for gene expression in *I. orientalis* (Park et al., 2018; Tran et al., 2019; Xiao et al., 2014). However, a toolset of well characterized constitutive promoters remains necessary to explore the full potential of this strain for more extensive metabolic engineering endeavors. Particularly, since pathway optimization for chemical production requires fine-tuning of the expression levels of the genes, it is desirable to explore a collection of promoters with varying transcriptional strengths. Similarly, terminators play an important role in controlling the level of gene expression by stabilizing mRNA level. Studies involving the characterization of terminators from *S. cerevisiae* (Curran et al., 2013) and other yeasts like *S. stipitis (Gao et al., 2017)* have demonstrated that the terminator sequence affects the half-life of the transcript which later influences the level of protein expression. Therefore, it is of great importance to discover and characterize novel terminators in *I. orientalis*.

Furthermore, in metabolic pathway engineering, complete biosynthetic pathways are often required to be heterologously expressed to obtain products of interest at high yields. The conventional sequential-cloning methods, including restriction enzyme based T4-ligation, Gibson assembly (Gibson et al., 2009), and Golden Gate assembly (Engler et al., 2008), may involve multiple steps and be inefficient, or may rely on unique restriction sites that become limited for assembly of large-size plasmids harboring multiple genes in one-step fashion (Ma et al., 2019; Shao et al., 2009). We previously developed the *in vivo* assembly method, named 'DNA assembler' to enable rapid construction of large biochemical pathways in an one-step fashion based on the homologous recombination (HR) mechanism in *S. cerevisiae* (Shao et al., 2009). Since *I. orientalis* also employs a HR mechanism for double stranded DNA repair, it is desirable to extend the DNA assembler method to *I. orientalis* for fast and reliable pathway construction.

In this study, we isolated one centromere-like (CEN-L) sequence from the *I. orientalis* genome and confirmed its function in improving plasmid-based gene expression. We discovered and characterized a series of constitutive promoters and terminators with various strengths. In addition, we performed *in vivo* DNA assembly in *I. orientalis*, which exhibited high fidelity to assemble a plasmid from multiple DNA fragments. Finally, to demonstrate the utility of these genetic tools, we rapidly constructed a xylose utilization pathway in *I. orientalis*.

2. Materials and methods

2.1 Strains, media, and chemicals

All strains used in this study are listed in Table 1. *E. coli* DH5α was used to maintain and amplify plasmids. *I. orientalis* SD108 and *S. cerevisiae* YSG50 were propagated in YPAD medium consisting of 1% yeast extract, 2% peptone, 0.01% adenine hemisulphate, and 2% glucose. Recombinant *I. orientalis* strains were grown in Synthetic Complete (SC) dropout medium lacking uracil (SC-URA). LB broth, bacteriological grade agar, yeast extract, peptone, yeast nitrogen base (w/o amino acid and ammonium sulfate), ammonium sulfate, and D-xylose were obtained from Difco (BD, Sparks, MD), while complete synthetic medium was purchased from MP Biomedicals (Solon, OH). All restriction endonucleases, Q5 DNA polymerase and Phusion polymerase were purchased from New England Biolabs (Ipswich, MA). cDNA synthesis kit and SYBR Green PCR master mix were purchased from Bio-Rad (Hercules, CA). The QIAprep spin mini-prep kit and RNA isolation mini kit were purchased from Qiagen (Valencia, CA), whereas Zymoclean Gel DNA Recovery Kit and Zymoprep Yeast Plasmid Miniprep Kits were purchased from Zymo Research (Irvine, CA). All other chemicals and consumables were purchased from Sigma (St. Louis, MO), VWR (Radnor, PA), and Fisher Scientific (Pittsburgh, PA). Oligonucleotides including gBlocks and primers were all synthesized by Integrated DNA Technologies (IDT, Coralville, IA). DNA sequencing was performed by ACGT, Inc. (Wheeling, IL). Plasmid mapping and sequencing alignments were carried out using SnapGene software (GSL Biotech, available at snapgene.com).

2.2 Plasmid construction

Most of the plasmids were generated by the *in vivo* DNA assembly method in *I. orientalis*, while the rest were carried out either by the DNA assembler method in *S. cerevisiae* (Shao et al., 2009) or Gibson assembly (Gibson et al., 2009) in *E. coli*. The experimental design and protocols of *in vivo* DNA assembly in *I. orientalis* were very similar to DNA assembler in *S. cerevisiae*. Briefly, 50~100 ng of PCR-amplified fragments and restriction enzyme digested backbone were cotransformed into *I. orientalis* SD108 via a lithium acetate-mediated method (Gietz et al., 1995). The colonies formed on SC-URA plates were randomly picked for functional characterization, and the confirmed target cells were then used to extract plasmids for *E. coli* transformation to enrich the plasmids. The plasmids were verified by restriction digestion or DNA sequencing. If needed, the correctly assembled plasmids will be retransformed into *I. orientalis* SD108 for further characterization. The constructed plasmids were shown in Table 1, and the designed primers were listed in Table S1.

2.3 Centromere-like sequence prediction and isolation

The centromere regions were predicted using a previously developed method named *in silico* GC_3 analysis (Cao et al., 2017a; Cao et al., 2017b; Lynch et al., 2010). In brief, the whole genome sequence of *I. orientalis* was downloaded from NCBI (https://www.ncbi.nlm.nih.gov/) along with its accompanying annotations. The coding sequences (CDS) were then extracted from the genome using BEDTools (v2.20.1) (Quinlan, 2014). CodonW (v1.4.4) (http://codonw.sourceforge.net/) was used to calculate the GC_3 percentage for each CDS sequence and a line graph was generated with a moving average of 15 genes corresponding to each chromosome. The longest intergenic regions from each chromosome, which may locate the centromere sequences were chosen for alignment to achieve the conserved fragment for functional characterization. The conserved sequence (CEN-0.8 kb) was PCR-amplified from *I. orientalis* genomic DNA, and ligated with *Kas*I and *Apa*I digested ScARS (pIo-UG) plasmid backbone, resulting in ScARS/CEN-0.8kb. After verification by restriction digestion, the ScARS/CEN-0.8kb plasmid was transformed to *I. orientalis* SD108 through heat-shock and screened on SC-URA solid medium for 2 days. Next, 10 colonies were randomly picked for GFP measurement from 24 h to 120 h by flow cytometry, and the one exhibiting higher cell ratio of GFP expression than those from the ScARS-plasmid was chosen for characterization.

2.4 Centromere-like sequence characterization

The function of centromere-like (CEN-L) sequence in improving plasmid stability was characterized by evaluating *ade2* knockout efficiency and D-lactic acid production. The ScARS/CEN-L-Cas9-ade2 plasmid was constructed by integrating CEN-L to pScARS-Cas9-ade2, which was assembled by cotransforming 100 ng of Cas9 expression cassette (PCR-amplified from pVT15b-epi), single guide RNA targeting *ade2* (Table S2), and digested pScARS backbone (*Xba*I and *Not*I). After transformation, the *ade2* knockout efficiency was calculated by the ratio between pink colonies and total colonies (Tran et al., 2019). The pink colonies were also picked for further confirmation by DNA sequencing. To construct D-lactic acid producing strain, the D-lactate dehydrogenase gene (*ldhD*) from *Leuconostoc mesenteroides* was amplified from pUG6-TDH3-Lm.ldhA-CYC1 (Baek et al., 2017) and cotransformed to *I. orientalis* together with *TDH3* promoter, *TEF1* terminator, and digested ScARS and ScARS/CEN-L (Figure S1) backbone (*Bsu*36I+*Not*I). Three colonies were picked and cultivated in 2 mL SC-URA medium as seed cultures for 2 days and then transferred to new SC-URA medium with the same initial OD. The samples were collected at various time points, and the supernatants were analysed for lactic acid production by HPLC (Agilent Technologies 1200 Series, Santa Clara, CA). The HPLC was equipped with a Rezex™ ROA-Organic Acid H^+ (8%) column (Phenomenex Inc., Torrance, CA) and a refractive index detector (RID). The column was eluted with 0.005 N H_2SO_4 at a flow rate of 0.6 mL/min at 50°C (Liu et al., 2019).

Plasmid copy numbers were quantified by a previously developed method (Cao et al., 2017a; Moriya et al., 2006). Briefly, two sets of primers specific to the GFP gene in plasmids and to the TRP1 reference gene in the *I. orientalis* genomic DNA were designed (Table S1), and a 16-fold serial dilution was applied to construct the standard curves for both GFP and TRP1. qPCR was performed on a QuantStudio 7 Flex Real-Time PCR System (Applied Biosystems, Foster City, CA) using a two-step cycling reaction program. Total DNA (genomic DNA and plasmid DNA) was firstly extracted from *I. orientalis* cells by Zymolase plus freeze-thaw lysis method, and then the cell lysates were centrifuged and the supernatants were diluted appropriately for qPCR. The copy number was determined as the ratio between the calculated molar amounts of *gfp* and *trp1* genes in the total DNA extracts, according to the two standard curves. The sizes of 10.8 Mbp for the *I. orientalis* genome and 10 kb for plasmids were used in the calculation.

2.5 Promoter characterization

For promoter characterization, a single, mixed and high-complexity RNA library made of RNA samples from the following four growth conditions was used for the RNA-Seq analysis performed in the U.S. Department of Energy's Joint Genomics Institute (JGI) central facility. *I. orientalis* was first grown in YPD broth overnight under 30 °C and 200 rpm on a platform shaker. The overnight culture of *I. orientalis* was pelleted and inoculated into the following four media at an initial OD at 600 nm (OD_{600}) of 10, and then grown for 16 h in: 1) YNB medium with glucose in the aerobic condition; 2) YNB medium with glucose and lignocellulosic biomass inhibitors (i.e. 1 g/L furfural, 3 g/L hydroxymethylfuran, 10 g/L NaCl, and 3 g/L acetic acid) in the aerobic condition; 3) YNB medium with glucose in the anaerobic condition; 4) YNB medium with glucose and lignocellulosic biomass inhibitors in the anaerobic condition. The aerobic cultures were grown at 200 rpm on a platform shaker while the anaerobic cultures were grown with stir bar rotating at 400 rpm. Total RNA was extracted individually from the cells by the QIAGEN RNeasy Kit and then treated with Ambion TURBO DNase. The DNA-free RNA samples were quantified by Qubit RNA BR Assay Kit. 750 ng RNA samples of each condition were used to make a total 3000 ng mixed RNA sample for JGI library preparation and sequencing. To validate the expression of selected gene in the RNA-Seq data, qPCR was performed. *I. orientalis* cells were inoculated in YPD medium, and culture was grown at 30 °C with constant shaking at 250 rpm for overnight. The cells were then inoculated into fresh YNB medium with 2% glucose with the initial OD at 600 nm (OD_{600}) of 0.1 and grown until the OD reached to 1. Cells were collected from 1 mL of culture, and total RNA was extracted using the RNeasy mini kit from Qiagen. DNase treatment of RNA was performed in the column during the preparation of RNA using the RNase-Free DNase Set from Qiagen. cDNA synthesis was carried out using the iScript™ Reverse Transcription Supermix and iTaq Universal SYBR Green Supermix from Biorad was used for qPCR. Primers for qPCR were designed using the IDT online tool (Primer Quest). For primer design, the amplicon length was restricted to be around 140 bp and the melting temperature (T_m) was set at 58 °C. For qPCR reactions, the manufacturer's protocol was followed: 10 μL of 2× SYBR Green supermix, 300 nM of forward and reverse primers, and 1 μL of cDNA. MicroAmp Optical 384 well plates from Applied Biosystems were used for the qPCR reactions which were performed on the Applied Biosystems machine using the following program: 2 min at 50 °C and 5 min at 95 °C for one cycle followed by 15 s at 95 °C, 30 s at 60 °C, and 30 s at 72 °C for 40 cycles, with a final cycle of 5 min at 72 °C. The endogenous gene *alg9*, encoding a mannosyltransferase, involved in *N*-linked glycosylation, was used as the internal control. Expression of the selected gene for promoter characterization was normalized by the *alg9* expression level. Raw data was analyzed using QuantStudio^{TM} Real-time PCR software from Applied Biosystems.

For the cloning of promoters, either the intergenic region or the 600 bp upstream of genes were chosen for characterization. Promoter sequences are shown in the Table S2. Putative promoters were cloned with the GFP reporter gene using the *in vivo* DNA assembly method and later confirmed through restriction digestion with *Hind*III and *Sal*I. Pairs of primers used to amplify the promoter region and other genetic elements including the GFP gene, terminator elements, *E. coli* part (Col1 region and ampicillin cassette), *ura3* gene (auxotrophic marker), promoter and terminator for *ura3* gene expression, and *ura3* gene from *S. cerevisiae* along with the promoter and terminator are shown in Table S1. The resultant plasmid is a *E. coli/S. cerevisiae/I. orientalis* shuttle vector (Table 1).

2.6 Terminator characterization

A total of 14 terminators was selected, mostly of 300 bp or shorter, were selected and amplified from *I. orientalis* genomic DNA and cloned between the GFP and mCherry genes by using the *in vivo* DNA assembly method (6 fragment assembly). Primers and DNA sequences of genetic elements and structural genes used in this study are listed in Tables S1 and S2, respectively. The plasmid backbone fragment was PCR-amplified from the p247_GFP plasmid and the mCherry gene was PCR-amplified from plasmid-64324 (Addgene). A random sequence used as a negative control was PCR-amplified from a non-functional region of *I. orientalis* genomic DNA which does not code for any promoter or terminator and does not contain a stretch of polyT with more than four T's. As a control, another plasmid was constructed without any sequence between the GFP gene and the mCherry gene. The resultant plasmid was verified by restriction digestion using *Hind*III and *Xho*I.

Recombinant *I. orientalis* strains harboring control plasmids or selected terminators were evaluated using qPCR as described in section 2.4. Relative amounts of GFP and mCherry transcripts were determined using the *alg9* gene as a control followed by the calculation of the ratio of mCherry to GFP transcripts for evaluating the strength of the terminators. Experiments were performed in biological triplicates.

2.7 Assembly of a xylose utilization pathway

Plasmid ScARS/CEN-L was digested with *Apa*I and *Not*I to obtain the backbone, and it was also used as a PCR template to obtain the URA3 expression cassette. *XR*, *XDH*, and *XKS* were PCR-amplified from pRS416Xyl-Zea_A_EVA (Shao et al., 2009). Promoters and terminators were PCR-amplified from the genomic DNA of *I. orientalis* (Table S2). All overlaps were designed to have 70-80 bp to facilitate *in vivo* homologous recombination, except for the overlaps between the fragments and the backbone (~40 bp). Approximately 100 ng of each fragment was transformed into *I. orientalis*, and the resultant transformants were spread onto SC-URA plates and incubated at 30 °C. Yeast colonies were collected for plasmid extraction, and the resultant plasmids were transformed to *E. coli* for enrichment. For assembly of a helper plasmid harboring the individual *XR*/*XDH*/*XKS* cassette, plasmids were extracted from randomly picked *E. coli* colonies and were verified by restriction digestion and DNA sequencing. Afterwards, the individual cassettes, *TDH3p-XR-MDH1t*, *HSP12p-XDH-PDC1t*, and *INO1p-XKS-PFK1t* were PCR-amplified from the helper plasmids (primers listed in Table S1), and mixed with ScARS/CEN-L backbone (digested by *Apa*I and *Not*I) and the URA3 expression cassette. *I. orientalis* was transformed with 100 ng of each fragment, spread on a SC-URA plate, and incubated at 30 °C. Plasmids were then extracted from *I. orientalis* and transformed to *E. coli*. Plasmids were extracted from three different *E. coli* colonies and were confirmed by restriction digestion and DNA sequencing.

The recombinant *I. orientalis* strain carrying the xylose utilization pathway was analyzed by monitoring the cell growth in SC-URA liquid medium supplemented with 2% xylose (SC-URA+XYL) as the sole carbon source (Shao et al., 2009). Colonies were picked into 2 mL SC-URA liquid medium supplemented with 2% glucose and grown for 2 days. Cells were spun down and washed with SC-URA+XYL medium twice to remove the remaining glucose and finally resuspended in fresh SC-URA+XYL medium with an initial OD_{600} of 0.2. Then, the cells were grown at 30 °C for 144 hours and OD_{600} was measured. The residual xylose was measured through HPLC after diluting the samples by 10-fold. The HPLC setup protocol was the same as the lactic acid measurement described in Materials and Methods 2.4 (Liu et al., 2019). Meanwhile, the sub-cultured cells in SC-URA medium were collected before growing to mid-log phase for qPCR analysis. The RNA extraction, cDNA synthesis, and qPCR were performed as described above in Section 2.4.

2.8 Flow cytometry

The GFP expression was measured by flow cytometry as described elsewhere (Cao et al., 2017a; Tran et al., 2019). In brief, the transformed *I. orientalis* cells were cultured in SC-URA medium for 24 h to 120 h and then centrifuged for 2 min at 2,000 x *g* to remove the supernatant. The cell pellets were resuspended in 10 mM phosphate-buffered saline (PBS, pH 7.4) and then analyzed by flow cytometry at 488 nm on a BD LSR II flow cytometer analyzer (BD Biosciences, San Jose, CA).

Similarly, for promoter characterization, constructs were transformed into *I. orientalis* and single colonies were picked from SC-URA plates and inoculated in the SC-URA medium and grown for 24 h. Cells were then inoculated in YNB medium with 2% glucose and YNB with glucose and lignocellulosic hydrolysate (1 g/L furfural, 3 g/L HMF, 3 g/L acetate and 10 g/L NaCl) and cultured under aerobic and anaerobic conditions. Samples were taken after 48 h for GFP fluorescence measurement. For terminator characterization, flow cytometer BD LSR FORTESSA with HTS was used to determine the fluorescence intensities of mCherry at 610 nm (Piatkevich and Verkhusha, 2011) and GFP at 488 nm.

3. Results and discussion

3.1 A centromere-like sequence improves gene expression on a plasmid

We experimentally confirmed that ScARS was functional for plasmid replication in *I. orientalis*, and the percentage of the cells carrying the ScARS-GFP containing plasmid was 55% of the entire population based on the flow cytometry analysis of the GFP expression at 5 days (Tran et al., 2019). Considering that in the benchmark system represented by *S. cerevisiae*, expressing GFP by the commercial vector pRS416 containing the native centromere resulted in a symmetric GFP peak representing >80% of the entire population (Cao et al., 2017a), isolating a functional CEN sequence from *I. orientalis* genome is essential for stable plasmid segregation.

Douglass *et al*. sequenced 32 *I. orientalis* genomes and predicted that each of the 5 centromeres is a 35-kb gene desert containing a large inverted repeat (Douglass et al., 2018). In parallel, we performed *in silico* GC_3 analysis of the genome of *I. orientalis* SD108, which was previously confirmed as an efficient prediction method to locate centromere sequences for some yeast species (Cao et al., 2017b; Lynch et al., 2010). Five long intergenic regions with sizes of 38.3-46.2 kb (covering the 31.7-37.8 kb centromeres predicted by Douglass et al.) were identified to contain potential centromeres (Table 2).

Due to the large sizes of these predicted sequences, integrating them to the plasmid for functional characterization was undesirable. We proposed that the conserved sequences shared by the five long predicted sequences may possess the functionality of a centromere. Therefore, the five centromere sequences were aligned, and an 811-bp conserved fragment (~2% of the original size) was obtained (Figure 1A). The 811-bp fragment (CEN-0.8kb) was amplified and integrated to ScARS-plasmid (previously reported as pIo-UG, (Tran et al., 2019)) with GFP as a reporter (Figure 1B), and transformed into *I. orientalis* SD108 strain for functional characterization. It was shown that among the 10 randomly picked colonies, only CEN-0.8kb-2 could express GFP at ratios of 81% and 67% at 24 h and 120 h, respectively (Figure 1C), while the other nine colonies were associated with similar peaks (Figure S2) to the cells harboring ScARS-plasmid (60% and 53% at 24 h and 120 h, respectively, Figure 1C). After DNA sequencing and aligning the different CEN-0.8kb fragments, we found there were a few nucleotide variants among them (Figure S3), which may be essential for the function of CEN-0.8kb. We also observed that the spacing sequence between ScARS and CEN-0.8kb-2 affected the CEN-0.8kb-2 function. The currently used spacing sequence of ScLeu2 cassette with a size of 2.2-kb could guarantee a GFP^+ population of > 80% at 24 h. However, when ScARS and

CEN-0.8kb-2 were rearranged in tandem, the percentage of the GFP$^+$ population decreased to 60%.

Collectively, these observations provided valuable information regarding CEN epigeneticity. In many eukaryotes, it is generally thought that CENs are epigenetically specified by their specialized chromatin structure and no conserved sequences or common features were found to predict CENs across species (Coughlan and Wolfe, 2019; Dalal, 2009; McKinley and Cheeseman, 2016). The CenH3 has been proposed to be the epigenetic mark of CENs, and its post-translational modifications (e.g., phosphorylation, methylation, acetylation, and ubiquitylation) contribute to CEN function (Bernad et al., 2009; Srivastava and Foltz, 2018). In this study, only one of the 0.8-kb sequence (CEN-0.8kb-2) demonstrated the benefit to plasmid stability. Whether the inefficiency of the other 0.8-kb sequences in the context of the plasmid was caused by the nucleotide variations or the different contexts rendered by the plasmid and the genome remains intriguing and will be investigated in the future.

The function of CEN-0.8kb-2 was further investigated by evaluating the *ade2* knockout efficiency via CRISPR/Cas9 and D-lactic acid production via overexpression of D-lactate dehydrogenase gene (*ldhD*) from *Leuconostoc mesenteroides* using plasmids harboring ScARS and ScARS/CEN-0.8kb-2 (Figure S1). As shown in Figure 1D and Figure S4, the *ade2* knockout efficiency was 95% using pScARS/CEN-0.8kb-2, while it was only 80% for ScARS plasmid. Meanwhile, the D-lactic acid produced by an *I. orientalis* strain overexpressing *ldhD* by ScARS/CEN-0.8kb-2 could reach 1.46 g/L in test tube, which was around 1.8-fold higher than the level achieved with the corresponding ScARS vector. To elucidate if the increased gene expression was originated from improved plasmid stability, we assayed the copy number of the two GFP expressing vectors (i.e., ScARS and ScARS/CEN-0.8kb-2) by qPCR. As shown in Figure 1E, the copy number of ScARS/CEN-0.8kb-2 plasmid was ~2.2 at 24 h, slightly higher than that of the ScARS plasmid (~1.9), indicating that CEN-0.8kb-2 improved the plasmid stability and resulted in a higher gene expression level. However, the copy numbers of both plasmids decreased over time, suggesting that they were still not as stable as the reported CEN-containing plasmids in *S. cerevisiae* and *S. stipitis* (Cao et al., 2017a). We hypothesized that the insufficient function of CEN-0.8kb-2 was caused by the incompleteness of the CEN sequence, which suggests the possibility that the surrounding sequences next to CEN-0.8kb-2 in the context of the plasmid might play an important role in plasmid segregation. Nevertheless, CEN-0.8kb-2 is beneficial for the gene expression system by improving the CRISPR gene knockout efficiency and the production of valuable chemicals in *I. orientalis*. Future studies are necessary to explore what nucleotide variants of CEN and what spacing sequences affect the CEN function, what sequences serve as centromeric protein binding sites, and whether CEN-0.8kb-2 functionality was conferred from the long terminal repeat (LTR) of centromere-associated retrotransposon (Coughlan and Wolfe, 2019; Douglass et al., 2018). To distinguish it from a fully functional CEN, CEN-0.8kb-2 was renamed as centromere-like sequence, i.e., CEN-L hereafter.

3.2 Systematic characterization of constitutive promoters

Previously, a certain number of promoters such as *TDH3p*, *PGK1p*, *TEF1p*, and *FBA1p* were used to create an *I. orientalis* strain capable of producing 11.63 g/L succinic acid (Xiao et al., 2014). However, so far, no comparative and systematic approach has been adopted for the characterization of a panel of constitutive promoters in *I. orientalis*. Therefore, we sought to identify a panel of strong, moderate, and weak constitutive promoters based on the RNA-sequencing data. Note that we used the RNA-seq data only as a primary screening tool to identify some constitutive promoter candidates for further investigation. A total number of 5141 genes were detected, and they were ranked from the most highly expressed to the least expressed based on their Reads Per Kilobase of transcript, per Million mapped reads (RPKM) values (Wagner et al., 2012). Functional annotation of the genes was performed based on the homology with the *S. cerevisiae* proteins. We selected the 1% most highly expressed genes based on RPKM values and narrowed down the collection to 52 genes. Out of the 52, only 35 genes were mapped to the *Saccharomyces* database as listed in Table 3. RNA-Seq data has revealed that the topmost expressed transcript is about ten-fold higher than most of the expressed genes, as shown in Figure 2A.

To quantify the strengths of the promoters, we measured the intensity of GFP fluorescence of the corresponding reporter strains at 48 h for aerobic conditions and 72 h for anaerobic conditions using flow cytometry. Additionally, we have included controls with no promoter for GFP expression and set the gate in a way to subtract any auto fluorescence value from the cells (Figure S5). Cells carrying the constructs were grown in four equivalent conditions as used for cultivation of the cells for RNA-Seq analysis. Results of GFP fluorescence for YNB minimal medium are mostly consistent with the qPCR results. In comparison to the positive control (*g527, belongs to FBA1p*), seven promoters (*g247*, *g5025*, *g853*, *g917*, *g3376*, *g2204*, and *g3540*) had led to strong expression (Figure 2B) and the analysis very closely correlated with the qPCR results (Figure S6). Some promoters showed quite similar fluorescence values to *g527p* and are included in the list of moderate promoters (*g3824*, *g43*, *g3767*, *g172*, *g973*, and *g4288*), whereas others are included in the list of weak promoters. Surprisingly, we did not detect the activity of a few promoters such as *g2880p*, *g2120p*, and *g2815p*, which correlates with the qPCR data (Figure S6). This reflects that either these promoters are not functional at all in the minimal medium or may require a different inducer. These inducers could be a stress induced by an anaerobic condition, inhibitors present in lignocellulosic biomass or both. To test this hypothesis, we measured the fluorescence in YNB medium in anaerobic condition (Figure S7A) or in YNB medium supplemented with inhibitory compounds present in lignocellulosic hydrolysate such as furfural, HMF, NaCl, and acetic acid, and grown in aerobic as well as in anaerobic condition. These molecules have shown to hamper the growth and fermentation ability of *S. cerevisiae* (Palmqvist and Hahn-Hägerdal, 2000a; Palmqvist and Hahn-Hägerdal, 2000b). Comparing the GFP expression driven by the *g2880p*, *g529p*, and *g2815p* did not show any noticeable difference when compared in the aerobic and anaerobic conditions in YNB medium (Figure S7A), or with lignocellulosic hydrolysate inhibitors under aerobic and anaerobic conditions (Figure S7B and S7C). Interestingly, the identified strong promoters listed in Table S3 were concluded to be constitutive promoters because they were expressed at similar levels in all the culture conditions. Moreover, comparing the promoter strength in YNB and stress-inducing medium has led to the identification of a few different promoters such as *g5025p* and *g3767p* in aerobic condition, and *g5025p, g3767p, g697*, and *g4194p* in anaerobic condition (Table S3). By comparative analysis, we have identified a few strong, medium, and weak constitutive promoters which can be used to express long biosynthetic pathways in *I. orientalis*.

3.3 Identification and characterization of strong terminators

The corresponding putative terminators of the 16 above-identified strong promoters were selected for characterization (Table S3). Furthermore, we sought to demonstrate the strength of these terminators at both transcriptional and translation levels. Out of the 16 targets, only 14 terminators were included for this study, since the terminators of the *pdc6* and *tdh3* genes have been used previously for the expression of the succinic acid pathway (Xiao et al., 2014). These terminators regions were amplified from either the intergenic sequences or the 300-bp sequences downstream of the target genes following a similar approach described previously (Gao et al., 2017), and then cloned between the two reporter genes, *gfp*, and *mCherry* (Figure 3A). Notably, we found that the 300 bp sequence of the *TEF1* terminator also included the promoter region and therefore we also selected the first 150 bp of this terminator for further study (*g2204t**).

As shown in Figure 3A, the two reporter genes (*gfp* and *mCherry*) share a single promoter (*TDH3p*, *g247*) and the terminator of the *pgk1* gene is placed after the *mCherry* gene, whereas the target terminators are placed between the two reporter genes. Note that the same design was used to discover new terminators (Gao et al., 2017). Additionally, 2 controls were included, one with no terminator sequence inserted between the reporter genes (Control 1) and the other where a random sequence of 300 bp that does not correspond to any promoter or terminator region inserted between the reporter genes (Control 2). In both the cases, the transcriptional ratio of mCherry and GFP was calculated to be approximately 0.64-0.62 (Figure 3B). Interestingly, except for the terminator of the *g73* gene that had a transcriptional ratio of 0.23, the rest of the terminators had a transcriptional ratio ranging from 0.03 to 0, and therefore were concluded to be strong terminators.

To further investigate the effect of the selected terminators on gene expression efficiency, their corresponding GFP fluorescence intensities were measured by flow cytometry, which have shown that changing the terminator has changed the expression level of GFP. Interestingly, terminators from all selected promoters except for *g1414* have shown similar fluorescence intensities (Figure 3C). This demonstrates that corresponding terminators of strong promoters can be used for the expression of biosynthetic pathways responsible for production of chemicals and fuels.

3.4 *In vivo* DNA assembly in *I. orientalis*

Rapid plasmid construction is critical in metabolic engineering, especially for large biochemical pathway assembly in one-step. Since *I. orientalis* employs the homologous recombination mechanism for double-stranded DNA repair (Douglass et al., 2018; Tran et al., 2019; Xiao et al., 2014), we sought to develop an *in vivo* DNA assembly method in *I. orientalis* for fast and reliable pathway construction. Here we intended to skip the usage of the helper elements corresponding to *S. cerevisiae*, which would save at least 3 days in generating a construct. As proof of concept, we performed the assembly of a shortened version of the ScARS plasmid (S-ScARS, 6.4 kb, Figure S1) containing IoURA3, ScARS and GFP cassettes, by co-transforming the linearized ScARS plasmid backbone (digested by *Ppu*MI+*Apa*I, ~6 kb) lacking ScARS and the amplified ~0.4 kb ScARS with 40 bp overlaps at two sides into *I. orientalis* (Figure 4A). Only the successfully assembled plasmid containing ScARS could grow on SC-URA, and three randomly picked colonies were chosen for GFP fluorescence analysis by flow cytometry and plasmid digestion by *Ppu*MI+*Kpn*I. The results showed that the GFP expression profile from S-ScARS was the same as that from the ScARS plasmid, with ~55% cells expressing GFP at 24 h (Figure 4B), and two bands (5.9 kb, 0.5 kb) were observed on the agarose DNA gel for the digested S-ScARS plasmid (Figure 4C), indicating 100% fidelity for two-fragment assembly.

We then performed *in vivo* assembly of a modified plasmid ScARS (M-ScARS, Sed1 promoter for GFP expression, ~7.4 kb, Figure S1) using multiple fragments. 2-7 fragments (2F-

7F) were PCR-amplified from the previously constructed M-ScARS backbone (Figure 4D-E) and co-transformed to *I. orientalis*. Plasmid digestion showed that all three randomly picked colonies from the 2, 3, 4, 6 and 7-fragment assembly groups were correctly assembled (3/3, 100%), while 5-fragment (5F) assembly showed 67% efficiency (2/3) (Figure 4F). Notably, 12-fragment assembly of M-ScARS was also successful with 100% fidelity (3/3) (data not shown), providing the foundation for assembling large biochemical pathways in *I. orientalis*.

Next, we attempted to expand the *in vivo* assembly and the aforementioned tools to a longer pathway, the xylose utilization pathway. This pathway includes three genes, *XR*, *XDH*, and *XKS*, which encode for xylose reductase, xylitol dehydrogenase, and xylulokinase, respectively. First, we constructed three helper plasmids by assembling the ScARS/CEN-L backbone (digested by *Apa*I and *Not*I) with the URA3 expression cassette, *XR*, *XDH*, and *XKS* genes, and the constitutive promoters and terminators characterized above (Figure 5A). After obtaining the helper plasmids, we then constructed the plasmid containing the xylose utilization pathway (ScARS/CEN-L-Xylose, Figure S1) by assembling the backbone, the URA3 cassette, and the three individual gene expression cassettes, *TDH3p-XR-MDH1t*, *HSP12p-XDH-PDC1t*, and *INO1p-XKS-PFK1t*. For *in vivo* assembly, 100 ng of each fragment, with 70-80 bp overlaps (40 bp overlap with backbone) were co-transformed to *I. orientalis* and the resultant plasmids were confirmed by restriction digestion and DNA sequencing. As shown in Figure 5B, the correct clones of XR helper plasmid exhibited three bands with sizes (bp) of 6127, 2561 and 1217, while XDH helper plasmid exhibited four bands with sizes of 4044, 2561, 1861 and 1217; XKS helper plasmid, exhibited three bands with sizes of 7224, 2561 and 1217; and the combined XR-XDH-XKS xylose pathway plasmid (ScARS/CEN-L-Xylose) exhibited four bands with sizes of 7016, 3736, 2561 and 1217. The results showed 100% fidelity was achieved for the assembly of the 6.5

kb xylose utilization pathway with an 8 kb plasmid backbone. The function of the assembled xylose utilization pathway was analyzed by growing the recombinant *I. orientalis* strain containing xylose utilization pathway in SC-URA medium supplemented with xylose instead of glucose. The recombinant *I. orientalis* strain carrying the whole xylose utilization pathway grew faster than the control strain containing the ScARS/CEN-L plasmid in xylose medium (Figure 5C), and the residual xylose at 144 h were 16.1 g/L and 17.6 g/L in the media of engineered and control strains (Figure 5D), respectively, indicating that the assembled xylose utilization pathway was functional. qPCR was also used to verify the expression levels of the three pathway genes, and the results showed that *XR* was poorly expressed while *XDH* and *XKS* were expressed well (Figure S8), which could explain the slow growth rate of the recombinant *I. orientalis* strain.

Pathway engineering is an important strategy for producing value-added bioproducts with high yield and productivity especially for long biosynthetic pathways. HR-based DNA assembler has been proved to be efficient for assembling large biochemical pathways in *S. cerevisiae* (Du et al., 2012; Shao et al., 2009; Shi et al., 2016). However, limited attempts were reported to apply HR-based assembly in other yeast hosts for rapid pathway engineering, even though some yeast species exhibit much more attractive capabilities, such as the high acid tolerant *I. orientalis*. Here, we performed *in vivo* DNA assembly in *I. orientalis* and achieved a very high fidelity for the assembly of a 14.5 kb-plasmid carrying a xylose utilization pathway from 5 fragments of different sizes. Although the pathway did not function well, which may be due to codon bias or imbalance of the promoter/terminator strengths for the *XR* gene, it still demonstrated that the DNA assembly could be adopted for efficient construction of biochemical pathways in *I. orientalis*.

4. Conclusion

In this work, we have developed a set of genetic tools for *I. orientalis*. Specifically, a CEN-L sequence was identified to be able to stabilize the ScARS-harboring vector and improve the expression of the target genes. More than 10 different strengths of constitutive promoters and terminators were screened and characterized, which can be used to tune gene expression in pathway engineering. In addition, HR-based *in vivo* assembly was shown to be a rapid and efficient method for plasmid construction and pathway assembly in *I. orientalis*. This toolbox has laid the foundation for facile metabolic engineering of *I. orientalis* as a platform organism for synthesis of chemicals and fuels.

Supporting information

Supplementary data to this article can be found online at xxx.

Conflict of interest

No conflict of interest was declared for this study.

Author contributions

M.C., Z.F., and H.Z. conceived the study and wrote the manuscript. M.C. and Z.F. performed the experiments and analyzed the data. V.T., X.S., and W.L. contributed to the plasmid construction, HPLC, and GFP measurements. P.H. and Y.Y. provided the RNA-Seq data. M.S. and Z.S. performed the *in silico* GC_3 analysis.

Acknowledgments

This work was supported by the U.S. Department of Energy award DE-SC0018420. RNA-Seq was conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract No. DE-AC02-05CH1123. X.S. was supported by a fellowship from the China Scholarship Council. M.S. and Z.S. were supported by the National Science Foundation Grant MCB 1716837. We thank Dr. Teresa Martin and Carl Schultz for proofreading the manuscript.

References

Baek, S. H., Kwon, E. Y., Bae, S. J., Cho, B. R., Kim, S. Y., Hahn, J. S., 2017. Improvement of D-lactic acid production in *Saccharomyces cerevisiae* under acidic conditions by evolutionary and rational metabolic engineering. Biotechnol. J. 12, 1700015

Bernad, R., Sánchez, P., Losada, A., 2009. Epigenetic specification of centromeres by CENP-A. Experimental Cell Res. 315**,** 3233-3241.

Cao, M., Gao, M., Lopez-Garcia, C. L., Wu, Y., Seetharam, A. S., Severin, A. J., Shao, Z., 2017a. Centromeric DNA facilitates nonconventional yeast genetic engineering. ACS Synth. Biol. 6**,** 1545-1553.

Cao, M., Seetharam, A. S., Severin, A. J., Shao, Z., 2017b. Rapid isolation of centromeres from *Scheffersomyces stipitis*. ACS Synth. Biol. 6**,** 2028-2034.

Choi, K. R., Jang, W. D., Yang, D., Cho, J. S., Park, D., Lee, S. Y., 2019. Systems metabolic engineering strategies: Integrating systems and synthetic biology with metabolic engineering. Trends Biotechnol. 37**,** 817-837.

Cleveland, D. W., Mao, Y., Sullivan, K. F., 2003. Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling. Cell 112**,** 407-21.

Coughlan, A. Y., Wolfe, K. H., 2019. The reported point centromeres of *Scheffersomyces stipitis* are retrotransposon long terminal repeats. Yeast 36**,** 275-283.

Curran, K. A., Karim, A. S., Gupta, A., Alper, H. S., 2013. Use of expression-enhancing terminators in *Saccharomyces cerevisiae* to increase mRNA half-life and improve gene expression control for metabolic engineering applications. Metab. Eng. 19**,** 88-97.

Dalal, Y., 2009. Epigenetic specification of centromeres. Biochemistry and Cell Biology. 87**,** 273-282.

Douglass, A. P., Offei, B., Braun-Galleani, S., Coughlan, A. Y., Martos, A. A. R., Ortiz-Merino, R. A., Byrne, K. P., Wolfe, K. H., 2018. Population genomics shows no distinction between pathogenic *Candida krusei* and environmental *Pichia kudriavzevii*: One species, four names. PLoS Pathog. 14**,** e1007138.

Du, J., Shao, Z., Zhao, H., 2011. Engineering microbial factories for synthesis of value-added products. J. Ind. Microbiol. & Biotechnol. 38**,** 873-890.

Du, J., Yuan, Y., Si, T., Lian, J., Zhao, H., 2012. Customized optimization of metabolic pathways by combinatorial transcriptional engineering. Nucleic Acids Res. 40**,** e142-e142.

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuveglise, C., Talla, E., Goffard, N., Frangeul, L., Aigle, M., Anthouard, V., Babour, A., Barbe, V., Barnay, S., Blanchin, S., Beckerich, J. M., Beyne, E., Bleykasten, C., Boisrame, A., Boyer, J., Cattolico, L., Confanioleri, F., De Daruvar, A., Despons, L., Fabre, E., Fairhead, C., Ferry-Dumazet, H., Groppi, A., Hantraye, F., Hennequin, C., Jauniaux, N., Joyet, P., Kachouri, R., Kerrest, A., Koszul, R., Lemaire, M., Lesur, I., Ma, L., Muller, H., Nicaud, J. M., Nikolski, M., Oztas, S., Ozier-Kalogeropoulos, O., Pellenz, S., Potier, S., Richard, G. F., Straub, M. L., Suleau, A., Swennen, D., Tekaia, F., Wesolowski-Louvel, M., Westhof, E., Wirth, B., Zeniou-Meyer, M., Zivanovic, I., Bolotin-Fukuhara, M., Thierry, A., Bouchier, C., Caudron, B., Scarpelli, C., Gaillardin, C., Weissenbach, J., Wincker, P., Souciet, J. L., 2004. Genome evolution in yeasts. Nature. 430**,** 35-44.

Engler, C., Kandzia, R., Marillonnet, S., 2008. A one pot, one step, precision cloning method with high throughput capability. PloS One 3**,** e3647.

Gao, M., Cao, M., Suastegui, M., Walker, J., Rodriguez Quiroz, N., Wu, Y., Tribby, D., Okerlund, A., Stanley, L., Shanks, J. V., Shao, Z., 2017. Innovating a nonconventional yeast platform for producing shikimate as the building block of high-value aromatics. ACS Synth. Biol. 6**,** 29-38.

Gibson, D. G., Young, L., Chuang, R. Y., Venter, J. C., Hutchison, C. A., 3rd, Smith, H. O., 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods 6**,** 343-5.

Gietz, R. D., Schiestl, R. H., Willems, A. R., Woods, R. A., 1995. Studies on the transformation of intact yeast cells by the LiAc/SS-DNA/PEG procedure. Yeast 11**,** 355-360.

Isono, N., Hayakawa, H., Usami, A., Mishima, T., Hisamatsu, M., 2012. A comparative study of ethanol production by *Issatchenkia orientalis* strains under stress conditions. J. Biosci. Bioeng. 113**,** 76-8.

Jensen, M. K., Keasling, J. D., 2015. Recent applications of synthetic biology tools for yeast metabolic engineering. FEMS Yeast Res. 15**,** 1-10.

Liu, J.-J., Zhang, G.-C., Kwak, S., Oh, E. J., Yun, E. J., Chomvong, K., Cate, J. H. D., Jin, Y.-S., 2019. Overcoming the thermodynamic equilibrium of an isomerization reaction through oxidoreductive reactions for biotransformation. Nat. Commun. 10**,** 1356.

Löbs, A.-K., Schwartz, C., Wheeldon, I., 2017. Genome and metabolic engineering in non-conventional yeasts: Current advances and applications. Synth. Syst. Biotechnol. 2**,** 198-207.

Lynch, D. B., Logue, M. E., Butler, G., Wolfe, K. H., 2010. Chromosomal G + C content evolution in yeasts: systematic interspecies differences, and GC-poor troughs at centromeres. Genome Biol. Evol. 2**,** 572-83.

Ma, X., Liang, H., Cui, X., Liu, Y., Lu, H., Ning, W., Poon, N. Y., Ho, B., Zhou, K., 2019. A standard for near-scarless plasmid construction using reusable DNA parts. Nat. Commun. 10**,** 3294.

Malik, H. S., Henikoff, S., 2009. Major evolutionary transitions in centromere complexity. Cell. 138**,** 1067-82.

Markham, K. A., Alper, H. S., 2018. Synthetic Biology Expands the Industrial Potential of *Yarrowia lipolytica*. Trends Biotechnol. 36**,** 1085-1095.

McKinley, K. L., Cheeseman, I. M., 2016. The molecular basis for centromere identity and function. Nature reviews. Mol. Cell Biol. 17**,** 16-29.

Meraldi, P., McAinsh, A. D., Rheinbay, E., Sorger, P. K., 2006. Phylogenetic and structural analysis of centromeric DNA and kinetochore proteins. Genome Biol. 7**,** R23.

Moriya, H., Shimizu-Yoshida, Y., Kitano, H., 2006. In Vivo Robustness Analysis of Cell Division Cycle Genes in *Saccharomyces cerevisiae*. PLOS Genet. 2**,** e111.

Nielsen, J., 2019. Yeast systems biology: model organism and cell factory. Biotechnol. J. 14**,** 1800421.

Palmqvist, E., Hahn-Hägerdal, B., 2000a. Fermentation of lignocellulosic hydrolysates. I: inhibition and detoxification. Bioresour. Technol. 74**,** 17-24.

Palmqvist, E., Hahn-Hägerdal, B., 2000b. Fermentation of lignocellulosic hydrolysates. II: inhibitors and mechanisms of inhibition. Bioresour. Technol. 74**,** 25-33.

Park, H. J., Bae, J.-H., Ko, H.-J., Lee, S.-H., Sung, B. H., Han, J.-I., Sohn, J.-H., 2018. Low-pH production of d-lactic acid using newly isolated acid tolerant yeast *Pichia kudriavzevii* NG7. Biotechnol. Bioeng. 115**,** 2232-2242.

Piatkevich, K. D., Verkhusha, V. V., 2011. Guide to red fluorescent proteins and biosensors for flow cytometry. Methods Cell Biol. 102**,** 431-61.

Quinlan, A. R., 2014. BEDTools: The swiss-army tool for genome feature analysis. Curr. Protoc. Bioinformatics 47**,** 11.12.1-34.

Redden, H., Morse, N., Alper, H. S., 2015. The synthetic biology toolbox for tuning gene expression in yeast. FEMS Yeast Res. 15**,** 1-10.

Riley, R., Haridas, S., Wolfe, K. H., Lopes, M. R., Hittinger, C. T., Göker, M., Salamov, A. A., Wisecaver, J. H., Long, T. M., Calvey, C. H., Aerts, A. L., Barry, K. W., Choi, C., Clum, A., Coughlan, A. Y., Deshpande, S., Douglass, A. P., Hanson, S. J., Klenk, H.-P., LaButti, K. M., Lapidus, A., Lindquist, E. A., Lipzen, A. M., Meier-Kolthoff, J. P., Ohm, R. A., Otillar, R. P., Pangilinan, J. L., Peng, Y., Rokas, A., Rosa, C. A., Scheuner, C., Sibirny, A. A., Slot, J. C., Stielow, J. B., Sun, H., Kurtzman, C. P., Blackwell, M., Grigoriev, I. V., Jeffries, T. W., 2016. Comparative genomics of biotechnologically important yeasts. Proc. Natl. Acad. Sci. 113**,** 9882.

Shao, Z., Zhao, H., Zhao, H., 2009. DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways. Nucleic Acids Res. 37**,** e16-e16.

Shi, S., Liang, Y., Zhang, M. M., Ang, E. L., Zhao, H., 2016. A highly efficient single-step, markerless strategy for multi-copy chromosomal integration of large biochemical pathways in *Saccharomyces cerevisiae*. Metab. Eng. 33**,** 19-27.

Srivastava, S., Foltz, D. R., 2018. Posttranslational modifications of CENP-A: marks of distinction. Chromosoma. 127**,** 279-290.

Steiner, F. A., Henikoff, S., 2015. Diversity in the organization of centromeric chromatin. Curr. Opin. Genet. Dev. 31**,** 28-35.

Toivari, M., Vehkomäki, M.-L., Nygård, Y., Penttilä, M., Ruohonen, L., Wiebe, M. G., 2013. Low pH d-xylonate production with *Pichia kudriavzevii*. Bioresour. Technol. 133**,** 555-562.

Tran, V. G., Cao, M., Fatma, Z., Song, X., Zhao, H., 2019. Development of a CRISPR/Cas9-Based Tool for Gene Deletion in *Issatchenkia orientalis*. mSphere. 4**,** e00345-19.

Verdaasdonk, J. S., Bloom, K., 2011. Centromeres: unique chromatin structures that drive chromosome segregation. Nature reviews. Mol. Cell Biol. 12**,** 320-332.

Wagner, G. P., Kin, K., Lynch, V. J., 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. 131**,** 281-285.

Xiao, H., Shao, Z., Jiang, Y., Dole, S., Zhao, H., 2014. Exploiting *Issatchenkia orientalis* SD108 for succinic acid production. Microb. Cell Fact. 13**,** 121.

Table 1. Strains and plasmids used in this study

Strains/Plasmids	Features	Sources
Strains		
E. coli DH5α	Cloning host	NEB
I. orientalis SD108	*ura3*Δ, host for plasmids in this study	(Xiao et al., 2014)
S. cerevisiae YSG50	*ade2-1, ade3Δ22, ura3-1, his3-11,15, trp1-1, leu2-3,112, can1-100*, used for plasmid assembly	(Shao et al., 2009)
Plasmids		
pScARS	Also reported as pIo-UG, derived from pRS415, containing *E. coli* elements, ScARS, ScLEU2, *IoURA3* and GFP cassette	(Tran et al., 2019)
pVT15b-epi	CRISPR/Cas9 plasmid, containing ScARS, *IoURA3*, iCas9, *RPR1* promoter, and sgRNA scaffold. Used for PCR of iCas9 and sgRNA cassettes	(Tran et al., 2019)
pScARS/CEN-0.8kb	Derived from pScARS by integrating the conserved 0.8 kb sequence from predicted CEN1~5	This study
pScARS/CEN-L	Also mentioned as pScARS-CEN-0.8kb-2, the screened centromere-like sequence with improved pScARS stability	This study
pScARS-Cas9-ade2	Derived from pScARS by changing GFP cassette to Cas9 cassette, also containing sgRNA targeting *ade2*	This study
pScARS/CEN-L-Cas9-ade2	Derived from pScARS/CEN-L by changing GFP cassette to Cas9 cassette, also containing sgRNA targeting *ade2*	This study
pUG6-TDH3-Lm.ldhA-CYC1	Used for amplifying *ldhD* gene	(Baek et al., 2017)
pScARS-LDH	Derived from pScARS by changing GFP cassette to LDH cassette	This study
pScARS/CEN-L-LDH	Derived from pScARS/CEN-L by changing GFP cassette to LDH cassette	This study
pS-ScARS	The shortened version of pScARS by removing ScLeu2 element	This study
pM-ScARS	The modified version of pScARS by replacing GFP promoter from *TDH3p* to *SED1p_g5025*	This study
pRS416Xyl-Zea_A	Used for amplifying xylose utilization pathway genes, *XR*,	(Shao et al.,

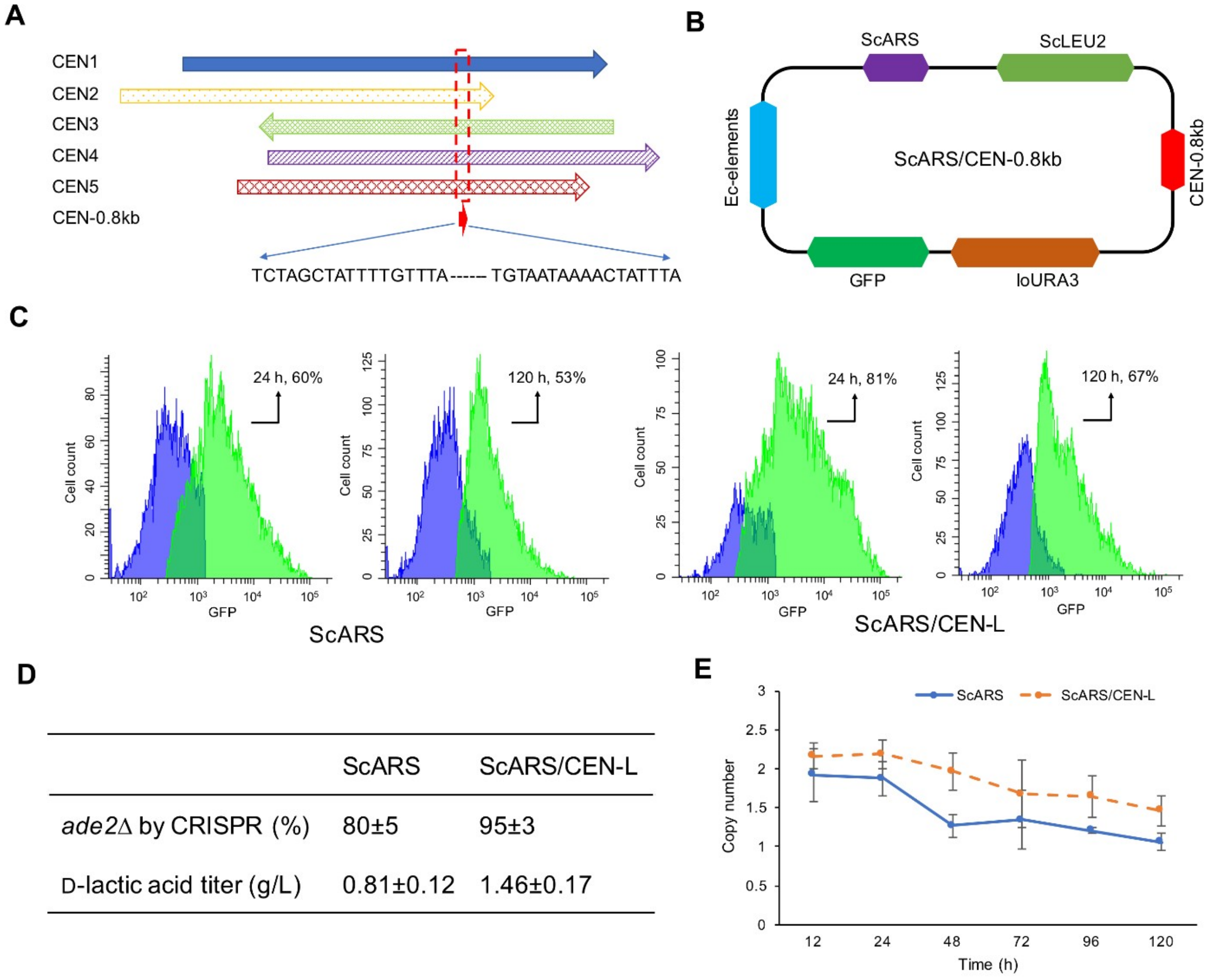
_EVA	*XDH*, and *XKS*	2009)
pScARS/CEN-L-Xylose	Derived from pScARS/CEN-L, containing xylose utilization pathway genes, *XR*, *XDH*, and *XKS*	This study
Plasmid-64324	pU6-(*Bbs*I) CBh-Cas9-T2A-mCherry, for mCherry amplification	Addgene
p247_GFP	Modified version of pScARS by replacing GFP promoter with *g247* (*TDH3*) promoter	This study
pWP_GFP	Modified version of pScARS by removing GFP promoter; negative control	This Study
pX_GFP	Modified version of pScARS by replacing GFP promoter with *X* promoter, and *X* represents *g853* (*GPM1*), *g917*, *g3540*, *g3376*, *g5025*, *g527*, *g2204*, *g1414*, *g4288*, *g3767*, *g5125*, *g73*, *g4282*, *g697*, *g4194*, and other tested promoters	This study
p247_mCherry	The modified version of p247_GFP by replacing GFP with *mCherry* gene and *ENO2t* terminator with PGK1t	This study
p247_GFP_mCherry	The modified version of p247_GFP, where *mCherry* added after *ENO2t* terminator, and *PGK1t* after *mCherry*	This study
pControl1	The modified version of p247_GFP_mCherry where *mCherry* are cloned in continuity of GFP, removed *ENO2t* terminator	This study
pControl2	The modified version of p247_GFP_mCherry, where *ENO2t* terminator sequence were replaced by random 300 bp sequence	This study
pZF_ter	The modified version of p247_GFP_mCherry, where *ENO2t* terminator sequence were replaced by different putative terminator sequence	This study

Table 2. Centromere-containing loci predicted by *in silico* GC_3 analysis

	IoCEN1	IoCEN2	IoCEN3	IoCEN4	IoCEN5
Predicted CEN loci on chromosomes	1463934-1510092	1451832-1492638	188014-226662	360477-403218	1093806-1132090
Predicted CEN sizes (bp)	46159	40807	38649	42742	38285

Table 3. Selected 36 promoters from *I. orientalis* SD108

Chr_locus	**Locus Tag (*I. orientalis*)**	**CDS_product**	**Threshold value**
1	JL09_g247	glyceraldehyde-3-phosphate dehydrogenase (TDH3)	0.04
1	JL09_g5025	SED1	0.06
3	JL09_g3824	Enolase	0.08
1	JL09_g220	PGK	0.12
3	JL09_g527	FBA1	0.16
1	JL09_g43	RTC3	0.18
5	JL09_g853	GPM1	0.19
1	JL09_g917	indolepyruvate decarboxylase 6	0.21
1	JL09_g5125	triose-phosphate isomerase TPI1	0.27
2	JL09_g3767	thioredoxin peroxidase TSA1	0.29
2	JL09_g2880	heat shock protein HSP150	0.33
1	JL09_g172	RCF2	0.35
1	JL09_g4285	pyruvate kinase CDC19	0.37
2	JL09_g3376	inositol-3-phosphate synthase INO1	0.39
5	JL09_g4565	ubiquitin	0.41
5	JL09_g697	RGI1	0.43
2	JL09_g31	peptidylprolyl isomerase CPR1	0.47
5	L09_g1318	ribosomal 60S subunit protein L10	0.51
2	JL09_g2204	translation elongation factor EF-1 alpha	0.53
2	JL09_g2120	amino acid transporter AGC1	0.56
4	JL09_g3008	pyridoxamine-phosphate oxidase PDX3	0.58
3	JL09_g529	alcohol dehydrogenase ADH3	0.62
1	JL09_g867	PBI2	0.68
1	JL09_g73	low-affinity Cu transporter	0.70
2	JL09_g2815	ribosomal 40S subunit protein S30A	0.72
2	JL09_g4565	ubiquitin-ribosomal 40S subunit protein S31 fusion protein	0.76
5	JL09_g1368	NADPH dehydrogenase	0.78
4	JL09_g4461	hexose transporter HXT6	0.80
2	JL09_g1383	cytochrome c isoform 2	0.86
2	JL09_g1414	hexose transporter HXT2	0.89
1	JL09_g3540	lipid-binding protein HSP12	0.91
4	JL09_g2950	cytochrome c oxidase subunit VII	0.93
5	JL09_g850	ubiquinol--cytochrome-c reductase subunit 8	0.95
3	JL09_g426	thioredoxin TRX1	0.97
2	JL09_g1530	amino acid starvation-responsive transcription factor GCN4	1.01



	ScARS	ScARS/CEN-L
*ade2*Δ by CRISPR (%)	80±5	95±3
D-lactic acid titer (g/L)	0.81±0.12	1.46±0.17

Figure 1. Discovery and characterization of a centromere-like (CEN-L) sequence. (A) Alignment of the centromere sequences predicted by *in silico* GC_3 analysis. (B) Plasmid map of ScARS/CEN-0.8kb containing *I. orientalis* CEN-0.8kb and URA3 selection marker, GFP expression cassette, *E. coli* elements (Ec-elements), *S. cerevisiae* ARS (ScARS), and *LEU2* selection marker (ScLEU2). (C) GFP expression profiles by ScARS or ScARS/CEN-L harboring plasmids at 24 h and 120 h measured by flow cytometry. (D) *ade2* knockout efficiencies by CRISPR/Cas9 and D-lactic acid productions using ScARS and ScARS/CEN-L plasmids. (E) Copy number assay for ScARS and ScARS/CEN-L vectors. Note: CEN-0.8kb-2 was named as CEN-L.

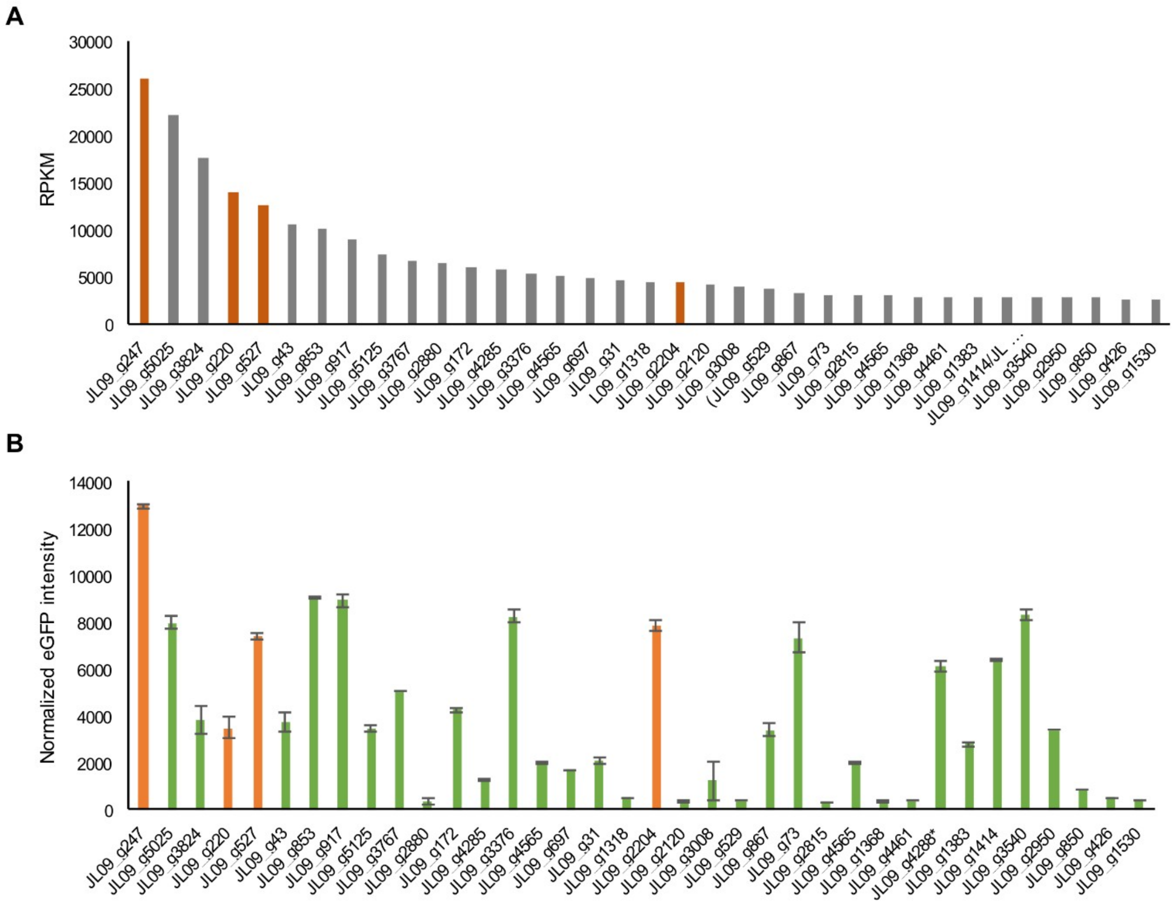


Figure 2. Promoter selection and characterization. (A) Plot showing the expression levels of the most highly expressed genes based on RNA-Seq analysis. (B) GFP expression driven by selected promoters. Promoters highlighted in orange were used previously for the assembly of the succinic acid pathway in *I. orientalis*. JL09_g527 (*fba*1) gene, highlighted in orange was used as a positive control for this study.

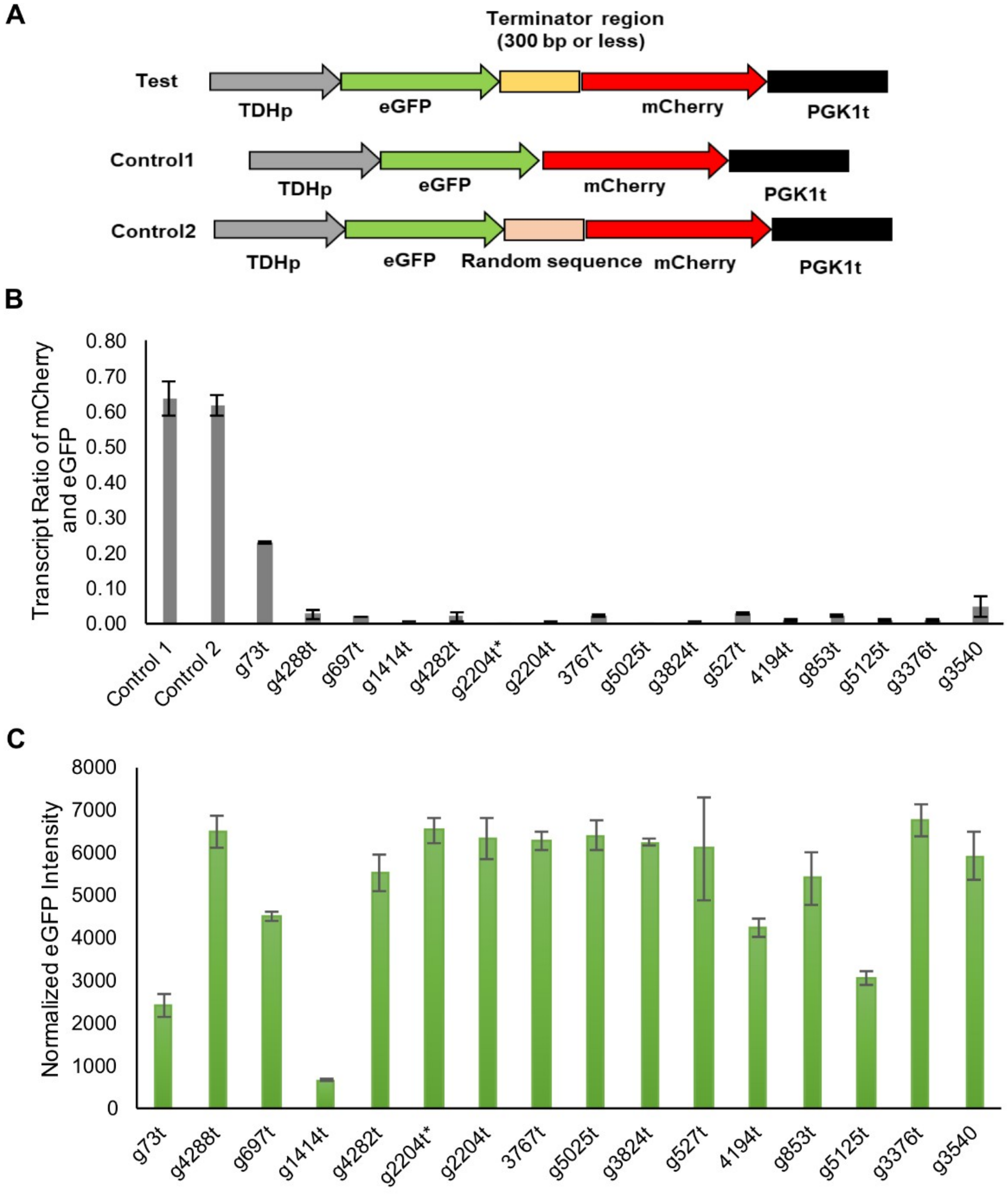
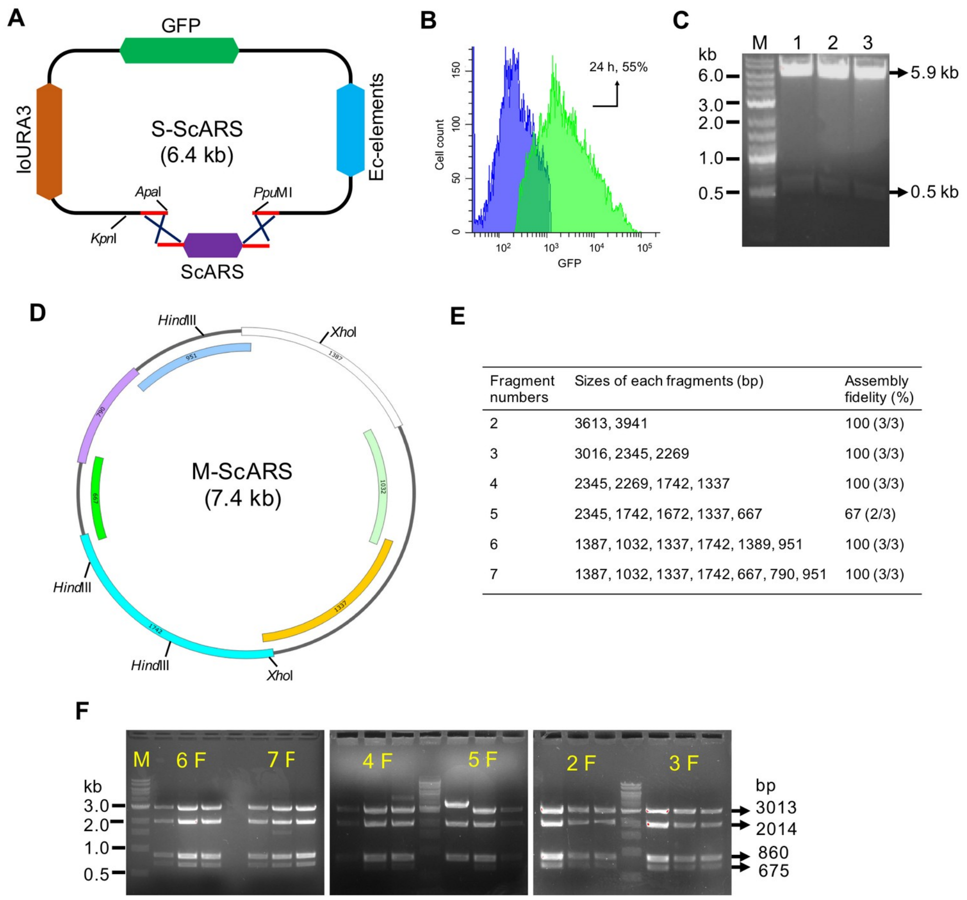


Figure 3. Characterization of the terminators from *I. orientalis*. (A) Terminators are cloned between two reporter genes, *gfp* and *mCherry* (Test) whereas either a random sequence (Control 2) or no sequence are inserted between the reporter genes (Control 1). (B) Termination efficiency of the selected terminators was calculated at the transcriptional level by determining the ratio of *mCherry* transcripts to *gfp* transcripts. (C) Terminator characterization based on GFP fluorescence intensity. Error bars represent standard deviations of two biological replicates.



Fragment numbers	Sizes of each fragments (bp)	Assembly fidelity (%)
2	3613, 3941	100 (3/3)
3	3016, 2345, 2269	100 (3/3)
4	2345, 2269, 1742, 1337	100 (3/3)
5	2345, 1742, 1672, 1337, 667	67 (2/3)
6	1387, 1032, 1337, 1742, 1389, 951	100 (3/3)
7	1387, 1032, 1337, 1742, 667, 790, 951	100 (3/3)

Figure 4. *In vivo* DNA assembly in *I. orientalis*. (A) Shortened ScARS plasmid (S-ScARS) assembled from the 6 kb backbone and 0.4 kb ScARS. (B) GFP expression profiles of randomly picked colonies containing S-ScARS at 24 h. (C) Restriction digestion analysis of the plasmids isolated from randomly picked colonies by *Ppu*MI and *Kpn*I (5.9 kb and 0.5 kb bands). M represents 1 kb plus DNA ladder from NEB. (D) The modified ScARS plasmid (M-ScARS) used for *in vivo* assembly of various numbers of fragments, and only the 7 fragments assembly was shown here. (E) Various numbers of fragments, their sizes, and assembly fidelity. (F) Restriction digestion analysis of assembled plasmids from different fragments by *Hind*III and *Xho*I, and 3013 bp, 2014 bp, 860 bp, 860 bp, and 675 bp bands were observed. Three colonies were picked for each assembly test.

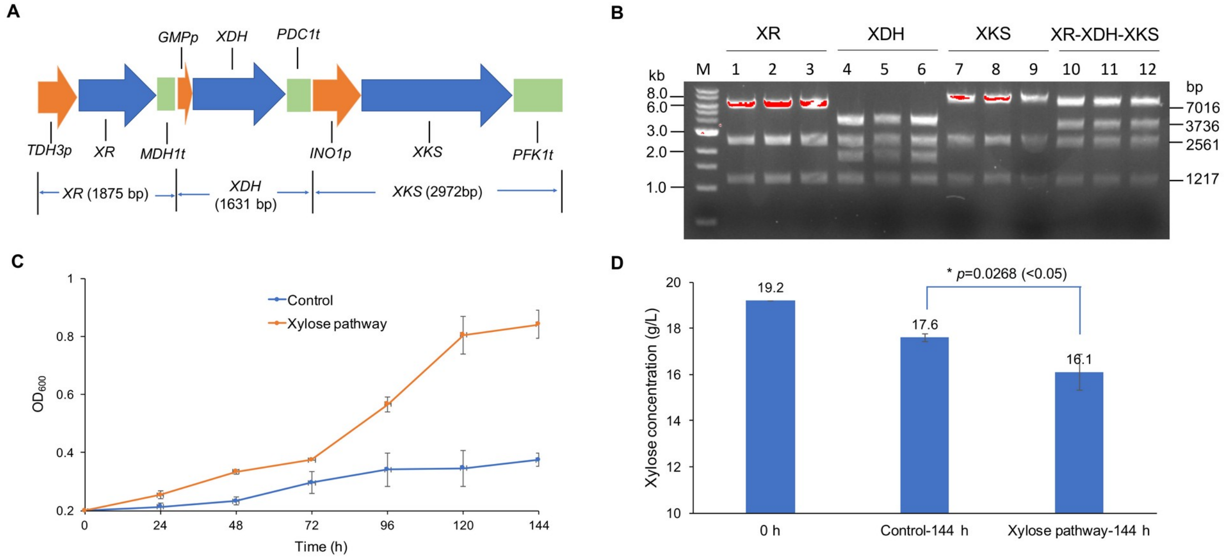


Figure 5. Construction of the xylose utilization pathway in *I. orientalis*. (A) Schematic representation of the assembled xylose utilization pathway. Each gene and its promoter/terminator were individually assembled first in *I. orientalis*. (B) Restriction digestion analysis of randomly picked colonies from assembled individual XR/XDH/XKS helper plasmids and combined XR-XDH-XKS plasmid by *Hind*III and *Eco*RI. The correct bands for digested plasmids (bp): XR: 6127, 2561 and 1217; XDH: 4044, 2561, 1861 and 1217; XKS: 7224, 2561 and 1217; and XR-XDH-XKS (ScARS/CEN-L-Xylose): 7016, 3736, 2561 and 1217. M represents 1 kb DNA ladder from NEB. (C) Functional analysis of the xylose utilization pathway by monitoring cell growth in SC-URA medium supplemented with 2% xylose. Cells carrying the ScARS/CEN-L were used as the negative control. (D) Residual xylose concentrations in liquid culture of the engineered strain containing the xylose utilization pathway and control strain. Error bars represent standard deviations for biological triplicates. The asterisk indicates statistical difference (p<0.05) using a two-tailed Student *t* test.