Verification of genetic engineering in yeasts with nanopore whole genome sequencing

Joseph H. Collins[1], Kevin W. Keating[1], Trent R. Jones[1], Shravani Balaji[1], Celeste B. Marsan[1], Marina Çomo[1], Zachary J. Newlon[1], Tom Mitchell[2], Bryan Bartley[2], Aaron Adler[2], Nicholas Roehner[2], and Eric M. Young[1,*]

[1]Department of Chemical Engineering, Worcester Polytechnic Institute, Worcester, MA, USA
[2]Synthetic Biology, Raytheon BBN Technologies, Cambridge, MA, USA
[*]emyoung@wpi.edu

ABSTRACT

Yeast genomes can be assembled from sequencing data, but genome integrations and episomal plasmids often fail to be resolved with accuracy, completeness, and contiguity. Resolution of these features is critical for many synthetic biology applications, including strain quality control and identifying engineering in unknown samples. Here, we report an integrated workflow, named Prymetime, that uses sequencing reads from inexpensive NGS platforms, assembly and error correction software, and a list of synthetic biology parts to achieve accurate whole genome sequences of yeasts with engineering annotated. To build the workflow, we first determined which sequencing methods and software packages returned an accurate, complete, and contiguous genome of an engineered *S. cerevisiae* strain with two similar plasmids and an integrated pathway. We then developed a sequence feature annotation step that labels synthetic biology parts from a standard list of yeast engineering sequences or from a custom sequence list. We validated the workflow by sequencing a collection of 15 engineered yeasts built from different parent *S. cerevisiae* and nonconventional yeast strains. We show that each integrated pathway and episomal plasmid can be correctly assembled and annotated, even in strains that have part repeats and multiple similar plasmids. Interestingly, Prymetime was able to identify deletions and unintended integrations that were subsequently confirmed by other methods. Furthermore, the whole genomes are accurate, complete, and contiguous. To illustrate this clearly, we used a publicly available *S. cerevisiae* CEN.PK113 reference genome and the accompanying reads to show that a Prymetime genome assembly is equivalent to the reference using several standard metrics. Finally, we used Prymetime to resequence the nonconventional yeasts *Y. lipolytica* Po1f and *K. phaffii* CBS 7435, producing an improved genome assembly for each strain. Thus, our workflow can achieve accurate, complete, and contiguous whole genome sequences of yeast strains before and after engineering. Therefore, Prymetime enables NGS-based strain quality control through assembly and identification of engineering features.

Introduction

Whole genome sequencing (WGS) is an attractive method for evaluating genetic engineering because it does not depend on specific sequence features and it captures unintended editing. Yet, engineered organisms are rarely evaluated with WGS, even though the few engineered genomes reported to date show unpredictable features that can only be detected with WGS. These include detection of multiple insertion events[1], gene loss and chromosomal rearrangement[2], unexpected mutations affecting phenotype[3], unpredictable off-target mutations from Cas9 editing[4,5], insertion of DNA from a plasmid used for cloning[6], and even insertion of genomic DNA from the cloning host[7]. This does not include a large number of unpublished accounts of WGS revealing unexpected sequences and genome structures in engineered industrial strains. This evidence challenges the assumption that an observed phenotype is the direct result of intended engineering, illuminating a possible explanation for variation between replicates and irreproducible findings - a common problem for biology-related disciplines[8]. Clearly, WGS must be used more broadly to detect and validate genetic engineering.

WGS is particularly needed for engineered yeast strains which can have complex genome features like multiple deletions[9], multiple plasmids[10], many insertions[11], and SCRaMbLEd chromosomes[12,13]. Furthermore, yeast are a crucial testbed for genome-scale design[14,15], and accurate WGS will be necessary for validating written eukaryotic genomes. Finally, engineered yeast have significant economic value as promising cell factories for the manufacture of medicines[16,17], fuels[18,19], materials[20,21], and chemicals[22,23]. Given the economic importance and increasing use of engineered yeast cell factories, it is crucial that WGS methods are developed that can efficiently validate the presence of intended engineering and confirm the absence of unintended variation. Without practical WGS workflows, the majority of strains are currently validated with inferior methods like PCR and targeted sequencing.

Yet, applying WGS is a challenge because of the diversity of genetic backgrounds, the variety of engineering features, and the current scale of yeast strain engineering. Myriad laboratory strains of the baker's yeast *Saccharomyces cerevisiae*[9,24,25] and nonconventional yeasts like *Yarrowia lipolytica*[26–28] and *Komagataella phaffii* (formerly *Pichia pastoris*)[29,30] are used to create yeast cell factories, so there are many potential genetic backgrounds. Methods of yeast engineering leave myriad sequence features behind, including standard plasmid sets with standard expression parts[31–34], high efficiency transformation[35–37], homologous recombination[10,38–40], gene knockouts using the Cre recombinase system[41], and genome editing using RNA-guided endonucleases[7,11,42,44–46]. Furthermore, the scale of yeast engineering is increasing both in the fraction of a genome that may be rewritten[12,47,48], and in the numbers of engineered strains created through adaptive laboratory evolution[49–51] and combinatorial pathway engineering[1,52,54–56]. Each of these factors make accurate, complete, and contiguous genomes difficult to attain without significant allocation of resources.

A WGS workflow involves five steps - DNA isolation, sequencing library preparation, sequencing, assembly, and annotation. First, genomic DNA is isolated from all other cellular components using one of a variety of methods, including phenol-choloroform, bead beating, or enzymatic lysis[57]. Second, the sequencing library is prepared by attaching adapters and barcodes. This can be done via ligation, which involves shearing the DNA to create free ends for DNA ligase to attach adapters, or tagmentation, which randomly inserts adapter attachment points without shearing[58]. Third, the library is sequenced with a next-generation sequencing (NGS) platform that either obtains short reads (150-300 base pairs long) with high accuracy[58] or long reads (1.5 kilobases to megabases long) with lower accuracy[59]. The average read length and the number of reads (genome coverage) output by the NGS platform is dependent on sequencing technology and the preceding DNA isolation and adapter attachment steps[60]. Fourth, the reads are computationally assembled into a final genome sequence with software that uses either an overlap-layout-consensus (OLC) or De Bruijn graph (DBG) algorithm[61]. OLC algorithms piece together reads based on overlapping sequences to construct progressively larger contiguous sequences (contigs). These algorithms use an All-versus-All consensus step[62] that may discard highly identical sequences in order to reach consensus. In contrast, DBG algorithms split reads into shorter k-mers followed by a Eulerian walk approach to construct contigs, thus DBG may be less prone to discarding highly identical sequences[62]. These algorithms assemble a genome sequence *de novo* or, when available, a reference genome may be used to aid assembly[63]. Fifth, an annotation is performed. Eukaryotic annotation involves first predicting genes in the genome sequence, followed by functional annotation[64]. However, engineering features like synthetic biology parts are not annotated. The quality of the assembled, annotated genome sequence is dependent on the read depth, the average read length, and the read accuracy.

Genome assembly quality also depends on genome assembly software. A variety of software based on OLC and DBG algorithms have been made publicly available. These fall into three categories: short read only, hybrid, and long read with error correction. Short read only software uses short read data exclusively, thus it achieves high sequence accuracy but requires high genome coverage to resolve structures like chromosomes as single contiguous sequences (contigs). These include the OLC assembler Edena[26] and the DBG assemblers ABySS[25] and Velvet[27]. Hybrid assembly software uses both short read and long read data. These assemble short reads first, then stitch contigs together with long reads. Popular hybrid assemblers include the OLC assembler Masurca[68] and the DBG assemblers HybridSPAdes[69] and Unicycler[28]. Long read with error correction software also uses both short read and long read data. These assemble only long reads with software like the OLC assemblers MiniASM[20], Canu[22], and SMARTdenovo[23] or the DBG assembler Flye[24]. Due to the error rate of long reads, the resulting contigs are error prone[59]. Therefore, the initial long read contigs only provide a "skeleton" for mapping additional reads[75–84]. The read mapping is performed by error correction software such as Medaka[85], which uses long read data, and Racon[29] or Pilon[31], which use short read data. Both hybrid and long read with error correction assembly approaches currently hold the most promise to achieve accurate genome sequence and structure at low read depths.

Ideally, one would be able to quickly obtain a genome assembly with WGS that resolves engineering with both accurate sequence and structure (*e.g.* the correct number of chromosomes and plasmids). Yet, unique challenges arise when applying WGS to engineered yeasts. First, engineered yeasts contain many engineering features that are important to annotate, yet there is currently no way to do this. Second, the high sequence identity in many engineered constructs, such as common plasmid elements or parts derived from the host genome, can cause identical sequences to be omitted[88,89]. In particular, OLC assemblers struggle to reproduce the expected representation and resolution of repeats[20,22,90]. Third, genome assembly software constructs either linear or circular sequences, not both. This is insufficient for yeasts that have a hybrid structure consisting of both linear chromosomes and circular plasmids. Fourth, the scale of yeast strain engineering limits the broad application of WGS due to cost. Currently, iterative design cycles and biological replicates result in many more strains than can be reasonably sequenced. For example, one combinatorial library for itaconic acid production consisted of 1,152 unique strains, including replicates[1]. Another library for penicillin synthesis consisted of 120 unique strains, only a subset of which were validated with targeted sequencing[56]. Therefore, it is necessary for WGS to not only achieve accurate resolution of all engineering signatures, but do so with the minimal amount of resources used per genome.

Here, we present a sample and data processing workflow that is capable of resolving all chromosomes and plasmids within

complete, contiguous genomes of engineered yeasts with synthetic biology parts annotated. First, we optimize sequencing library preparation to increase nanopore read length and the number of reads from plasmids. Then, we test different assembly algorithms for their ability to achieve correct, contiguous sequences of the engineering features. We then develop an annotation strategy to label common yeast engineering sequences. We integrate all software steps into a single package, called Prymetime. Then, we validate Prymetime on a panel of 15 engineered yeasts, resolving all engineering sequences in different genetic backgrounds. Further analysis of two strains reveals unintended recombination and insertion events, demonstrating the utility of Prymetime as a quality control tool. We further show that the whole genomes produced by Prymetime using 40X read depth of both nanopore and Illumina reads are equivalent to or better than reference genome assemblies of *S. cerevisiae* CEN.PK113, *Y. lipolytica* Po1f, and *K. phaffii* CBS 7435. Thus, Prymetime allows resolution and annotation of intended engineering signatures, identification of unintended changes, and assembly of quality parent strain genomes.

Results

Optimizing Nanopore Sequencing Library Preparation for Engineered Yeasts

From the beginning, we set a standard that our genome assembly workflow must be able to resolve chromosomal integrations and multiple plasmids used in yeast engineering. Therefore, we built a *S. cerevisiae* CEN.PK113 strain containing an integrated carotenoid pathway, the native 2μ plasmid, a dCas9 plasmid, and a gRNA plasmid, shown in Figure 1a. We named this strain "FEY_2," and a picture of several colonies of this strain are shown in 1b. Initially, we prepared sequencing libraries of FEY_2 with a nanopore ligation kit. Sequencing these initial libraries had low average read length that varied from run to run, possibly because of differential DNA shearing during isolation. To limit this, we developed a gentle genomic DNA isolation protocol which increased average nanopore read length and reduced variance (see Supplementary Methods). However, the sequencing results contained very few reads from plasmids, as determined by comparing the average normalized mapped reads of the plasmid antibiotic selection markers to those of the ACT1 genomic locus using Minimap2. We could isolate plasmids from FEY_2 using a yeast miniprep kit, so we reasoned that the sequencing library preparation step was so gentle that it was not linearizing circular plasmids for adapter ligation. Thus, we turned to a tagmentation library preparation method. The improvement in average normalized mapped plasmid reads is shown in Figure 1c for the low copy plasmid and Figure 1d for the high copy plasmid. Interestingly, the 2:1 and 20:1 marker to ACT1 ratios for each plasmid are equivalent to the approximate plasmid copy number in yeast for each origin[25,35]. Furthermore, tagmentation also increased the representation of other circular elements like the native 2μ plasmid and mitochondrial DNA. These results indicate that tagmentation is key to achieving long average read lengths while also generating linear molecules from small circular DNA so that they can pass through the nanopore flow cell. Thus, with gentle isolation and tagmentation, nanopore sequencing of FEY_2 resulted in adequate representation of plasmid reads.

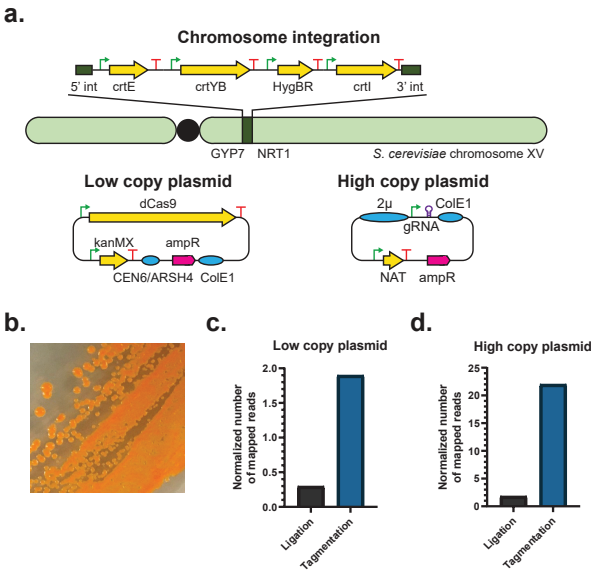

Figure 1. *FEY_2 strain design and nanopore library preparation methods affecting FEY_2 read representation.* **a.** Illustration of the engineering signatures comprising FEY_2, which included a carotenoid pathway chromosomal integration, a low copy plasmid expressing dCas9, and a high copy plasmid expressing gRNA. **b.** Photograph of FEY_2 streaked onto an agar plate, showing the carotenoid pathway is functional. **c.** Normalized number of mapped reads from libraries prepared by ligation or tagmentation for the low copy plasmid in FEY_2. **d.** Normalized number of mapped reads from libraries prepared by ligation or tagmentation for the high copy plasmid in FEY_2.

Developing a *de novo* Genome Assembly Workflow for Complete, Contiguous Plasmids and Integrations

Once we achieved appropriate read representation, we investigated which assembly algorithm would correctly assemble the reads into contiguous sequences. This requirement is stringent, particularly for the three plasmids because they each have significant sequence identity between each other and the genome. We evaluated *de novo* assemblers of the following types: short-read only, hybrid, and long read with error correction.

To do this, we used our optimized library preparation to obtain long reads at 60X genome coverage from the ONT MinION and short reads at 125X genome coverage from the Illumina iSeq 100. This common set of reads was then fed to each assembler, and the resulting genome assembly was analyzed using BLASTN for the presence of the integrated pathway, both plasmids, and the native 2μ plasmid. A visual representation of the BLASTN results is shown in Figure 2a, with a key in Figure 2b describing the glyphs used to represent the assembly features. The engineering features were rarely complete or assembled into one contiguous sequence. The short-read only *de novo* assemblers ABySS, Edena, and Velvet returned a fragmented, incomplete pathway and plasmids. The hybrid assemblers SPAdes and Masurca produced more complete sequences than the short read only assemblers, but the genome integration was fragmented, and Masurca also omitted portions of the three plasmids. The long read *de novo* assemblers MiniASM, Canu, Flye, and SMARTdenovo each returned a single contiguous sequence for the genome integration, yet, MiniASM, Canu, and SMARTdenovo omitted sections of the three plasmids. Only Flye was able to return the genome integration and each plasmid correctly in contiguous sequences.

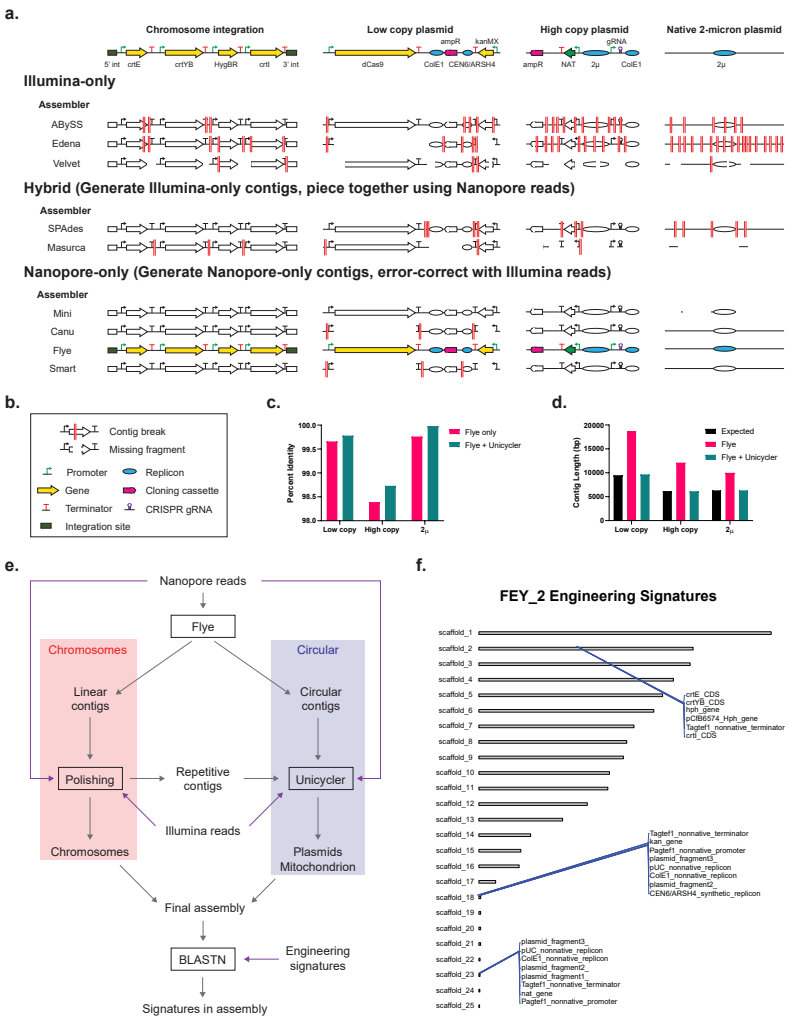

Figure 2. *Detection of engineering signatures in S. cerevisiae FEY_2 using various genome assembly strategies.* **a.** BLASTN results from querying known engineering signatures against assemblies. **b.** Key describing assembly failure modes and synthetic biology part glyphs. The colored pathways and plasmids represent assemblies where all engineering signatures were found in contiguous sequences. **c.** Plasmid percent identity, found using BLASTN, in the FEY_2 strain both before and after reassembly with Unicycler. **d.** Plasmid contig length in the FEY_2 strain both before after reassembly with Unicycler, compared to the expected plasmid length. **e.** Overview of Prymetime genome assembly pipeline. **f.** Output from Prymetime, showing the FEY_2 genome assembly scaffolds and the engineering signatures found using BLASTN.

Of the assemblies missing large portions of at least one of the three plasmids, almost all were generated with an OLC assembler. To investigate this further, we used BLASTN at each step in the OLC-based Canu pipeline to determine when sequences were omitted. We noted that the complete low-copy plasmid was initially present before the consensus step, and was then lost in the final assembly. It seems that Canu discarded the plasmid at a certain threshold during the consensus step, likely because of high sequence identity with the other plasmids. In contrast, the DBG assemblers Flye, ABySS, and SPAdes did not omit sections of plasmids. While the plasmids were fragmented among different contigs with ABySS and SPAdes, Flye assembled all plasmids into single contiguous sequences. This indicates that, along with sequencing library preparation, genome assembly quality is also dependent on the *de novo* assembler.

While the plasmid contigs from Flye were complete, they were longer than expected. Further inspection revealed that the contigs consisted of several repeats of the expected plasmid sequence. This is a common problem for long-read assemblers, as they use linear logic to merge contigs[22,91]. To obtain structurally representative plasmid contigs, we wrote a script to identify putative circular sequences and send them to be re-assembled with Unicycler, software that was built to assemble circular contigs from bacterial sequencing data[28]. Reassembly of plasmids with Unicycler improved the accuracy as measured by BLASTN (Figure 2c), and length (Figure 2d) of the contigs for the three plasmids in FEY_2.

To simplify the search for engineering features with BLASTN, we added a step that annotates features in the genome from synthetic biology part sequences. We developed a list of many yeast engineering features, including promoters, terminators, selection markers, fluorescent reporters, common coding sequences, origins, and conserved plasmid fragments. This is the default part annotation list for Prymetime, but the user may easily use a different list. A sequence in the genome assembly with high identity to a part on the list are annotated and then plotted using the genome plotter karyoploteR[36], shown in Figure 2f for the FEY_2 genome assembly. This feature allows the user to quickly visualize all engineering signatures in the entire genome, particularly highlighting plasmids and integrations. Additionally, this feature permits identification of known engineering sequences in an unknown sample.

We coded the final workflow into a single dockerized software package called Prymetime: "Pipeline for Recombinant Yeast genoMEs That Identifies Markers of Engineering." To our knowledge, this is the first workflow able to annotate engineering features in yeast genomes and assemble both linear and circular contigs. An overview of the workflow is shown in Figure 2e, with a more detailed diagram in Figure S1. In our automated workflow, Flye first assembles nanopore reads and classifies contigs as linear or circular. Then, linear contigs are error-corrected using the polishing software Medaka, Racon, and Pilon, while the circular contigs are re-assembled with long read and short read data using Unicycler. Finally, in an optional step, engineering features are annotated and visualized with karyoploteR. As Figure 2f shows, this workflow assembled chromosomes, plasmids, and successfully annotated engineering features in the FEY_2 genome.

Resolving Engineering Signatures in a Collection of Engineered Yeasts

We next validated Prymetime on a collection of engineered laboratory and nonconventional yeast. We constructed 15 strains from *S. cerevisiae* S288C, CEN.PK113-7D, W303-α, BY4741, BY4742, and *K. phaffii* ATCC 76273 (CBS 7435)[93,94] and *Y. lipolytica* ATCC MYA-2613 (Po1f)[95]. A description of each strain is shown in Figure 3, with more detailed descriptions of each strain in Table S1. Engineering signatures were inserted into the genome or maintained on episomal plasmids. *S. cerevisiae* integrations were targeted to the HO locus[32] or between NRT1 and GYP7 in chromosome XV[1,45]. *S. cerevisiae* plasmids consisted of custom TypeIIS-compatible yeast shuttle vectors with either *S. cerevisiae* replicon (2μ or CEN6/ARSH4). Engineering was broadly categorized into biosynthetic pathways, gene editing components, deletions, and synthetic biology elements. Biosynthetic pathways included propane[96], β-carotene[6], prespatane[10], carnosic acid[13], and limonene[100,101]. Genome editing associated tools included SpCas9[8], dCas9[7], LbCpf1[45], FnCpf1[44], and Cre recombinase[41]. Deletions included the synthetic auxotrophies already present in *S. cerevisiae* W303-α, BY4741, BY4742, and *Y. lipolytica* Po1f. Synthetic biology elements included fluorescent proteins[9,16] and the 2A sequence[17]. The engineered *Y. lipolytica* strain "FEY_74" contained a CRISPR-Cas9 expression plasmid that contained a codon-optimized version of the Cas9 protein, along with a gRNA expression cassette[46]. The engineered *K. phaffii* strain "FEY_75" contained a chromosomally-integrated red fluorescent protein (RFP) cassette. This strain was transformed using a two-step recombinase based system with integrative plasmids[34].

Strain	Species	Parent	Purpose	Integration	Deletion	Low copy	High copy
FEY_1	*S. cerevisiae*	S288C	Pathway (Propane)	ChrXV			
FEY_2	*S. cerevisiae*	CEN.PK113-7D	Pathway (β-carotene) + Editing (dCas9)	ChrXV		CEN6/ARSH4	2μ
FEY_5	*S. cerevisiae*	CEN.PK113-7D	Common elements (Fluorescent protein)	ChrXV			
FEY_15	*S. cerevisiae*	CEN.PK113-7D	Editing (dCas9)	HO			2μ
FEY_18	*S. cerevisiae*	W303-α	Editing (Cas9)		X	CEN6/ARSH4	2μ
FEY_27	*S. cerevisiae*	S288C	Editing (Cpf1)			CEN6/ARSH4	
FEY_29	*S. cerevisiae*	S288C	Editing (Cpf1)				2μ
FEY_30	*S. cerevisiae*	S288C	Editing (Cre)			CEN6/ARSH4	
FEY_37	*S. cerevisiae*	S288C	Pathway (Prespatane)				2μ
FEY_43	*S. cerevisiae*	S288C	Pathway (Carnosic acid)	HO			
FEY_45	*S. cerevisiae*	S288C	Pathway (Limonene)	HO			
FEY_48	*S. cerevisiae*	BY4742	Common elements (Fluorescent proteins)	HO	X	CEN6/ARSH4	
FEY_55	*S. cerevisiae*	BY4741	Common elements (Fluorescent protein)		X	CEN6/ARSH4	
FEY_74	*Y. lipolytica*	Po1f	Editing (Cas9)		X	CEN	
FEY_75	*K. phaffii*	CBS 7435	Common elements (Fluorescent protein)	trp2			

Figure 3. *Panel of diverse engineered yeast strains constructed.* Presence of an integration, deletion, low copy plasmid, or high copy plasmid icon indicates a strain has the respective engineering signature, while the absence of an icon indicates it does not have the engineering signature.

We sequenced this collection with the ONT MinION and the Illumina iSeq 100. From the sequencing data, Prymetime produced genome assemblies that captured each engineering signature in each *S. cerevisiae* genetic background as measured by BLASTN of the reference sequence against the assembly. Shown in Figure 4a, the genome assemblies resolved seven different genome integrations in two genome loci and eleven different plasmids. Metrics for the BLASTN results for all engineered *S. cerevisiae* strains are in Table S2, while the annotated karyoploteR visualizations can be found in Figure S2. Figure S3 shows that including Unicycler for plasmid assembly improves length and accuracy in every strain, not just FEY_2. Furthermore, neither the type of gene (metabolic, selective, editing, or reporter), nor repetitive parts (Ptef1, Pgal10), nor plasmid copy number affected the accuracy or structural completeness of the assembly.

The genome assemblies from the two engineered nonconventional yeasts - *Y. lipolytica* strain FEY_74 and *K. phaffii* strain FEY_75 - revealed unintentional engineering. FEY_74 was intended to contain the pCRISPR-yl[46] plasmid, shown in Figure 4c. However, the plasmid contig from the genome assembly was missing the entire Cas9 transcription unit and a portion of the *E. coli* origin of replication. Inspection of the raw reads failed to identify a single read with the missing ylCas9 sequence. We performed a genomic DNA isolation and a yeast plasmid miniprep on FEY_74 and transformed the resulting DNA back into *E. coli*, yet did not observe any colonies. This indicates that the disrupted origin of replication in the assembly reflects an actual unintended loss rather than an assembly error. This was further confirmed by PCR of DNA isolated from FEY_74 with primers spanning the missing region of the plasmid. The length of the PCR product indicated that the Cas9 transcription unit was indeed missing (Figure S4). Similarly, FEY_75 was designed to have an RFP transcription unit integrated into chromosome II (Figure 4d). However, PCR of the integration site in chromosome II failed to confirm integration, even though the strain was nourseothricin resistant and RFP positive. Prymetime was able to annotate the entire pathway in the FEY_75 genome, but the pathway was integrated into chromosome IV. These results indicate that Prymetime can be used to find and validate engineering sequences, which is useful for both strain quality control and identification of engineering in unknown samples.

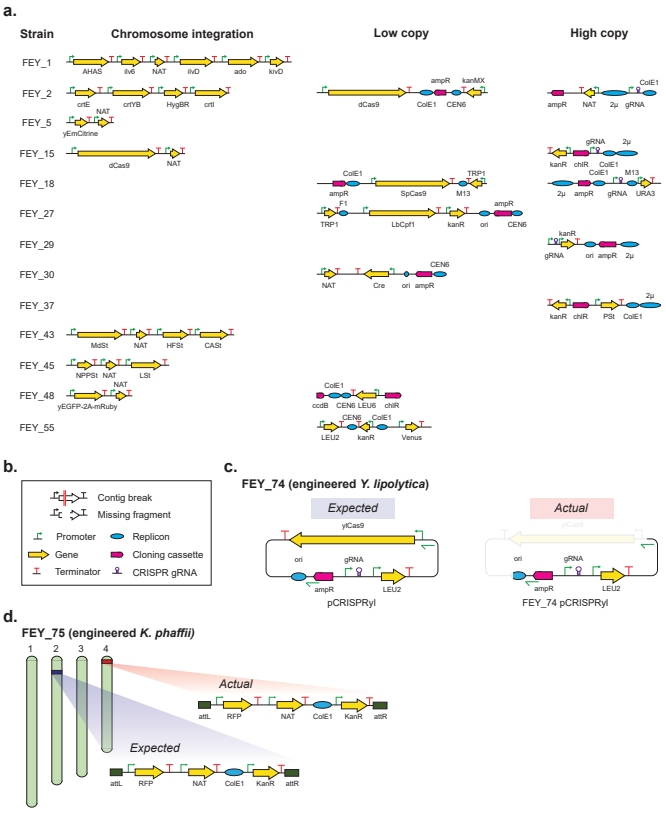

Figure 4. *Resolving signatures of engineering from the panel of engineered yeast strains.* **a.** Visual representation of the BLASTN results from querying known engineering signatures against Prymetime-assembled genome assemblies of the 15 engineered *S. cerevisiae* strains. **b.** Key describing assembly failure modes and synthetic biology part glyphs. The colored pathways and plasmids represent assemblies where all engineering signatures were found in contiguous sequences. **c.** The expected CRISPR-Cas9 expression vector for FEY_74, an engineered *Y. lipolytica* strain, and the actual plasmid from the Prymetime genome assembly (see Figure S3). **d.** Illustration showing the expected location of the RFP integration cassette into chromosome II of FEY_75, an engineered *K. phaffii* strain, and the actual location of the cassette into chromosome IV.

Whole Genome Assembly Quality

We then analyzed the whole genome quality of the Prymetime assemblies. Each engineered *S. cerevisiae*, *Y. lipolyitica*, and *K. phaffii* genome had high contiguity, sequence accuracy, and genome completeness. However, to clearly show that Prymetime can assemble quality whole genomes we benchmarked it against the best publicly available *S. cerevisiae* CEN.PK113-7D genome assembly for which the raw reads were also available[105,106]. We used these same raw reads to generate a genome assembly with Prymetime. As shown in Figure 5a, the reference assembly used raw nanopore, Illumina, and PacBio reads[106], while Prymetime assembled random subsets of the raw nanopore and Illumina reads at different read depths. These subsets were passed through each step in Prymetime (Figure 5b). We compared the assembly quality at each step to the reference genome.

First, we evaluated the structure of the initial contigs from Flye using metrics from QUAST[32], specifically, N50, the number of contigs, and the length of the largest contig (Figure 5c). The standard deviation was calculated from the variation in the metrics resulting from three different random read subsets. The Flye step achieves N50 equivalent to the *S. cerevisiae* CEN.PK113-7D reference at a long read genome coverage of 40X and above. Similarly, 40X genome coverage and above produced the expected 18 contigs - sixteen chromosomes, the native 2μ plasmid, and mitrochondrial DNA. Further, the longest contig from 40X genome coverage and above is equivalent to the reference genome. To be thorough, we also tested nanopore assemblers other than Flye to see if these produced improvements, depicted in Table S3. Flye remained the best assembler. These results indicate that the Flye assembly step can generate reference quality genome structure with a minimum long read sequencing depth of 40X.

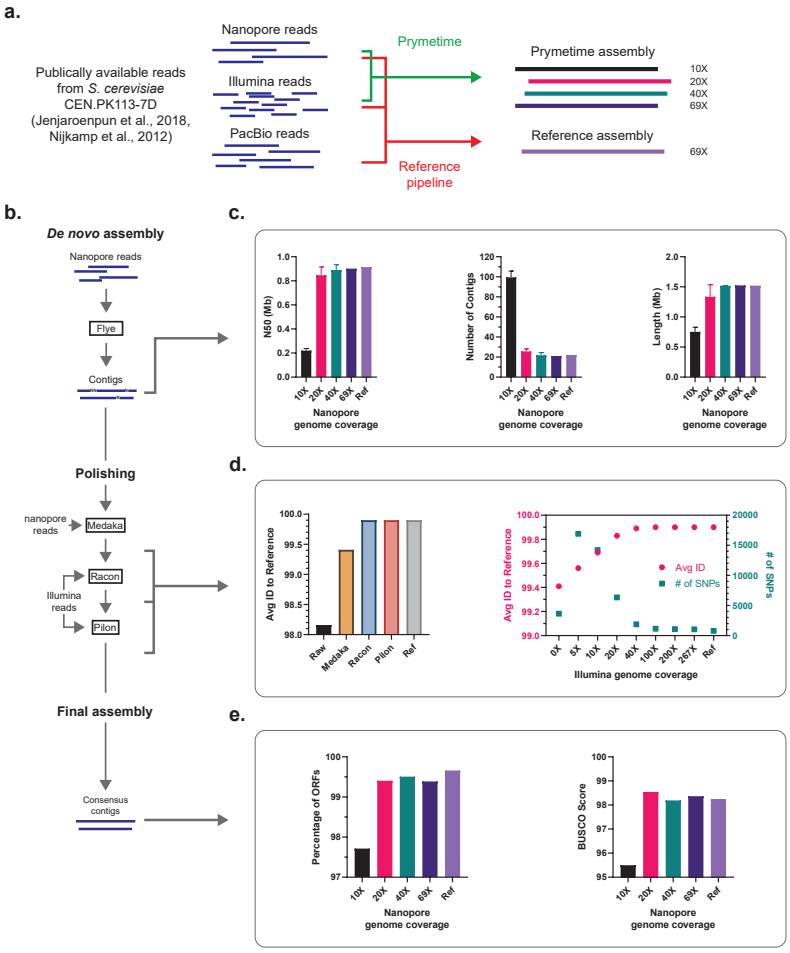

Figure 5. *Genome assembly quality at varying genome coverage depths using publicly available reads from a* S. cerevisiae *CEN.PK113-7D reference genome.* **a.** Prymetime and reference assembly workflows and read depths. **b.** Assembly and polishing workflow. Assemblies were evaluated at different points in this workflow, indicated by the arrows. **c.** Genome structure comparison, using QUAST. **d.** Genome accuracy with different polishing steps and read depths. **e.** Genome completeness quantified by the percentage of open reading frames (ORFs) from *S. cerevisiae* S288C in each genome assembly and by the BUSCO score.

Next, we evaluated the sequence accuracy of Prymetime. Average identity to the reference genome was calculated with MUMmer[33] at different points in the Prymetime polishing workflow, shown in Figure 5d. The unpolished long read assembly from Flye only matches 98.1% of the reference genome, while successive Medaka (long read polishing), and Racon and Pilon (short read polishing) steps improve the assembly, eventually matching the reference. Then, we assessed the read depth of short reads needed to optimize Racon and Pilon polishing. To do this, we calculated average identity to reference and the number of single nucleotide polymorphisms (SNPs) for different short read genome coverages, again using MUMmer. The results indicate that short reads at 40X genome coverage and above are sufficient to match the reference genome. Additionally, the accuracy of polishing assemblies from each nanopore assembler other than Flye is presented in Table S4 (adding polishing steps) and Table S5 (changing short read depth). These results indicate that successive polishing with Illumina short reads at a minimum read depth of 40X is critical for sequence accuracy.

Finally, the polished genome assemblies were evaluated for completeness. Two completeness metrics were used - the percentage of *S. cerevisiae* S288C open reading frames (ORFs) contained in an assembly, calculated by BLASTN, and the Benchmarking Universal Single-Copy Orthologs (BUSCO) score, calculated using the Saccharomycetales dataset[34]. The results are shown in Figure 5e. In terms of percentage of *S. cerevisiae* S288C ORFs, Prymetime genome assemblies using a long read depth of 40X most closely matched the *S. cerevisiae* CEN.PK113-7D reference. In terms of the BUSCO score, all assemblies with long read depths above 20X were equivalent to the reference. Additionally, these metrics were calculated for polished assemblies using other nanopore assemblers, shown in Table S5 (BUSCO score) and Table S6 (percent of S288C ORFs). Taken together, these results indicate that the genome assemblies generated by Prymetime are structurally correct, accurate, and complete using only 40X read depth for both long and short reads.

Re-Sequencing Nonconventional Yeasts

To demonstrate that the entire laboratory and computational workflow can obtain high quality reference genomes, we re-sequenced the two nonconventional yeasts in this study with the ONT MinION and Illumina iSeq 100. The resulting *de novo* genome assemblies output from Prymetime were compared to the publicly available reference assembly for *Komatagaella phaffii* CBS 7435[110] and *Yarrowia lipolytica* Po1f[28]. Comparing the whole genomes with Mauve[111] qualitatively confirmed the completeness of the Prymetime assemblies (Figure 6a depicts *K. phaffii* and Figure 6b depicts *Y. lipolytica*). Interestingly, Prymetime resolved a large region in the third contig of *Y. lipolytica* Po1f that is not in the reference (purple shading in Figure 6b). Quantitatively, the Prymetime genome assemblies had high genome contiguity as measured by the number of contigs (Figure 6c), had no assembly gaps (Figure 6d), and improved genome completeness measured by BUSCO score (Figure 6e). The higher BUSCO scores are because there are 6 more essential genes in the *K. phaffii* Prymetime assembly and 13 more essential genes in the *Y. lipolytica* Prymetime assembly. This is, to our knowledge, the first assembly of these essential genes in these nonconventional yeasts. Overall, these results indicate that Prymetime can be used to generate high quality *de novo* reference genomes of nonconventional yeasts.

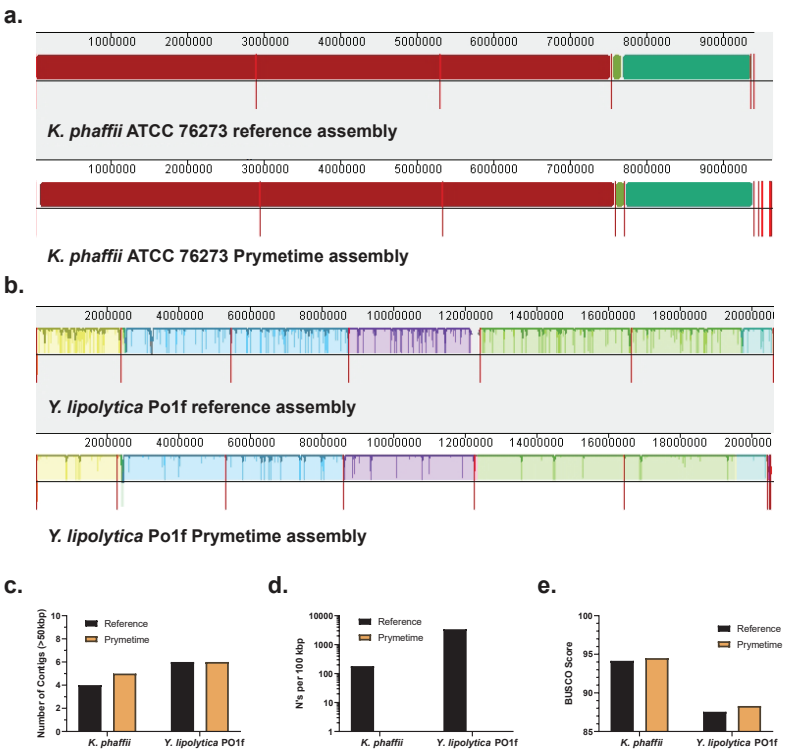

Figure 6. *Re-sequencing of two nonconventional yeast strains* **a.** Mauve visualization of a *K. phaffii* whole genome comparison between the publicly available reference assembly and the Prymetime assembly. **b.** Mauve visualization of a *Y. lipolytica* whole genome comparison between the publicly available reference assembly and the Prymetime assembly. The red lines on each alignment indicate a new contig. **c.** Number of contigs in the reference genome assemblies versus the Prymetime genome assemblies. **d.** Gaps in the reference genome assemblies versus the Prymetime genome assemblies. The number of gaps were represented by the number of N's per 100 kbp. **e.** Genome BUSCO score.

Discussion

This work describes development of a novel workflow for WGS of engineered yeasts. The validated workflow consists of gentle gDNA isolation, tagmentation, long and short read NGS, accurate *de novo* assembly of both linear and circular elements, and synthetic biology part annotation. Using this workflow, diverse engineering signatures can be resolved in complete, contiguous sequences even with multiple similar plasmids in one strain. The resulting whole genome quality is comparable to high-quality reference assemblies, therefore, it is possible to generate accurate genome assemblies both before and after engineering. This permits verification of genetic engineering in yeasts with WGS.

Interestingly, only Flye was suitable for the first assembly step, fully satisfying the strict requirements of engineering signature accuracy, completeness, and contiguity. The other *de novo* genome assemblers we tested omitted engineering signatures at the read depths tested - even with appropriate representation of plasmid reads. This is because the strain used to benchmark the assemblers had multiple plasmids with high sequence identity between them. We observed that signature omission most commonly occurred with assemblers built around OLC algorithms. Assemblers built around DBG algorithms were consistently better at resolving all signatures. Yet, Flye remained the only assembler able to resolve all of the signatures in contiguous sequences. These observations highlight the difficulty of applying otherwise effective genome assembly software to engineered yeasts, which have complex and repetitive sequence elements, and the need to continually improve assembly algorithms to better handle complex features. Based on our results, we recommend benchmarking future assemblers on the performance of Flye.

Whole genome sequencing is rarely used in strain engineering cycles due to the barriers of NGS cost, time, and required bioinformatics expertise. The WGS workflow we developed with the inexpensive ONT MinION and Illumina iSeq 100 platforms and the integrated, dockerized Prymetime software package overcomes these barriers. With Prymetime, we were able to achieve high-quality genomes at relatively low read depth, finding that 40X for both long and short reads was sufficient for accuracy, completeness, and contiguity of whole yeast genomes and the engineering features within them. With 40X read depth, up to 30 *S. cerevisiae* genomes can be sequenced on one MinION flow cell and up to 4 genomes can be sequenced on one Illumina iSeq flow cell. This is because 0.5 Gb is needed for 40X read depth of the 12.1 Mb *S. cerevisiae* genome and our typical yield is approximately 15 Gb from the MinION and 2.4 Gb from the iSeq 100. Not accounting for labor, this level of multiplexing would cost around $200 per genome. The entire workflow is fast - it takes under a week to start from a single colony and acquire a genome assembly, requiring only 15 hours of hands-on time. Finally, our workflow requires only a few coding steps - future users can simply load NGS reads and run the Prymetime script (see the GitHub repository at https://github.com/emyounglab/prymetime for more details).

The integrated workflow described here permits rapid, on-site acquisition of reference quality yeast genome sequences and annotation of genetic parts. Thus, it detects and validates genetic engineering in yeasts. The utility of this workflow for verification of engineering and resequencing of nonconventional yeasts was demonstrated, indicating that it may also be applied to sequence novel yeast isolates and identify engineering in unknown samples. We envision that this novel approach is broadly applicable to any effort that involves engineered yeasts.

Data Availability

The reference genomes for *S. cerevisiae*, *Y. lipolytica*, and *K. phaffii* can be accessed on NCBI. All engineered genome sequences are available on NCBI. The raw reads have been deposited on NCBI.

Code Availability

Prymetime can be accessed as a Docker image on GitHub at https://github.com/emyounglab/prymetime.

Acknowledgements

The authors thank James Kingsley at WPI for his help implementing Prymetime on WPI's server. This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) under Finding Engineering Linked Indicators (FELIX) program contract #N66001-18-C-4507. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. This work is also supported by Worcester Polytechnic Institute startup funds.

Author Contributions Statement

JHC, AA, NR, and EMY conceived of the study. JHC conducted all sequencing runs and bioinformatics analysis. JHC, KWK, TRJ, SB, CBM, MC, and ZJN built the collection of engineered yeasts for sequencing. JHC, TM, and EMY wrote PRYMETIME scripts and created the Docker image.

Competing Interests Statement

The authors declare no competing interests.

References

1. Anton, B. P., Fomenkov, A., Raleigh, E. A. & Berkmen, M. Complete genome sequence of the engineered escherichia coli shuffle strains and their wild-type parents. *Genome Announc.* **4**, e00230–16, DOI: 10.1128/genomeA.00230-16 (2016).

2. Solis-Escalante, D. *et al.* The genome sequence of the popular hexose-transport-deficient saccharomyces cerevisiae strain eby.vw4000 reveals loxp/cre-induced translocations and gene loss. *FEMS Yeast Res* **15**, DOI: 10.1093/femsyr/fou004 (2015).

3. Gallegos, J. E., Hayrynen, S., Adames, N. & Peccoud, J. Challenges and opportunities for strain verification by whole-genome sequencing. *bioRxiv* DOI: 10.1101/515338 (2019). https://www.biorxiv.org/content/early/2019/01/09/515338.full.pdf.

4. Li, J. *et al.* Whole genome sequencing reveals rare off-target mutations and considerable inherent genetic or/and somaclonal variations in crispr/cas9-edited cotton plants. *Plant Biotechnol. J.* **17**, 858–868, DOI: 10.1111/pbi.13020 (2019). https://onlinelibrary.wiley.com/doi/pdf/10.1111/pbi.13020.

5. Veres, A. *et al.* Low incidence of off-target mutations in individual crispr-cas9 and talen targeted human stem cell clones detected by whole-genome sequencing. *Cell Stem Cell* **15**, 27 – 30, DOI: https://doi.org/10.1016/j.stem.2014.04.020 (2014).

6. Young, A. E. *et al.* Genomic and phenotypic analyses of six offspring of a genome-edited hornless bull. *Nat. Biotechnol.* DOI: 10.1038/s41587-019-0266-0 (2019).

7. Schwarzhans, J.-P. *et al.* Non-canonical integration events in pichia pastoris encountered during standard transformation analysed with genome sequencing. *Sci. Reports* **6**, 38952 EP – (2016). Article.

8. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454, DOI: 10.1038/533452a (2016).

9. Brachmann, C. B. *et al.* Designer deletion strains derived from saccharomyces cerevisiae s288c: A useful set of strains and plasmids for pcr-mediated gene disruption and other applications. *Yeast* **14**, 115–132, DOI: 10.1002/(SICI)1097-0061(19980130)14:2<115::AID-YEA204>3.0.CO;2-2 (1998).

10. DiCarlo, J. E. *et al.* Yeast oligo-mediated genome engineering (yoge). *ACS Synth. Biol.* **2**, 741–749, DOI: 10.1021/sb400117c (2013).

11. Ronda, C. *et al.* CrEdit: CRISPR mediated multi-loci gene integration in saccharomyces cerevisiae. **14**, 97, DOI: 10.1186/s12934-015-0288-3.

12. Blount, B. A. *et al.* Rapid host strain improvement by in vivo rearrangement of a synthetic yeast chromosome. **9**, 1932, DOI: 10.1038/s41467-018-03143-w (2018).

13. Gowers, G.-O. F. *et al.* Improved betulinic acid biosynthesis using synthetic yeast chromosome recombination and semi-automated rapid LC-MS screening. **11**, 868, DOI: 10.1038/s41467-020-14708-z (2020).

14. Ostrov, N. *et al.* Technological challenges and milestones for writing genomes. **366**, 310–312, DOI: 10.1126/science.aay0339.

15. Bartley, B. A., Beal, J., Karr, J. R. & Strychalski, E. A. Organizing genome engineering for the gigabase scale. **11**, 689, DOI: 10.1038/s41467-020-14314-z.

16. Collins, J. H. & Young, E. M. Genetic engineering of host organisms for pharmaceutical synthesis. *Curr Opin Biotechnol* **53**, 191–200, DOI: 10.1016/j.copbio.2018.02.001 (2018).

17. Paddon, C. J. & Keasling, J. D. Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development. *Nat. Rev. Microbiol.* **12**, 355–367, DOI: 10.1038/nrmicro3240 (2014).

18. Zhou, Y. J., Kerkhoven, E. J. & Nielsen, J. Barriers and opportunities in bio-based production of hydrocarbons. *Nat. Energy* **3**, 925–935, DOI: 10.1038/s41560-018-0197-x (2018).

19. Peralta-Yahya, P. P., Zhang, F., del Cardayre, S. B. & Keasling, J. D. Microbial engineering for the production of advanced biofuels. *Nature* **488**, 320–328, DOI: 10.1038/nature11478 (2012).

20. Werten, M. W. T., Eggink, G., Cohen Stuart, M. A. & de Wolf, F. A. Production of protein-based polymers in pichia pastoris. *Biotechnol. advances* **37**, 642–666, DOI: 10.1016/j.biotechadv.2019.03.012 (2019). 30902728[pmid].

21. Keating, K. W. & Young, E. M. Synthetic biology for bio-derived structural materials. *Curr. Opin. Chem. Eng.* **24**, 107 – 114, DOI: https://doi.org/10.1016/j.coche.2019.03.002 (2019). Materials engineering: bio-derived/bio-inspired materials.

22. Borodina, I. & Nielsen, J. Advances in metabolic engineering of yeast saccharomyces cerevisiae for production of chemicals. *Biotechnol. J.* **9**, 609–620, DOI: 10.1002/biot.201300445 (2014). https://onlinelibrary.wiley.com/doi/pdf/10.1002/biot.201300445.

23. Ekas, H., Deaner, M. & Alper, H. S. Recent advancements in fungal-derived fuel and chemical production and commercialization. *Curr. Opin. Biotechnol.* **57**, 1 – 9, DOI: https://doi.org/10.1016/j.copbio.2018.08.014 (2019). Energy Biotechnology • Environmental Biotechnology.

24. Thomas, B. J. & Rothstein, R. Elevated recombination rates in transcriptionally active dna. *Cell* **56**, 619 – 630, DOI: https://doi.org/10.1016/0092-8674(89)90584-9 (1989).

25. Sikorski, R. S. & Hieter, P. A system of shuttle vectors and yeast host strains designed for efficient manipulation of dna in saccharomyces cerevisiae. *Genetics* **122**, 19–27 (1989).

26. Markham, K. A. & Alper, H. S. Synthetic biology expands the industrial potential of yarrowia lipolytica. *Trends Biotechnol.* **36**, 1085 – 1095, DOI: https://doi.org/10.1016/j.tibtech.2018.05.004 (2018).

27. Abdel-Mawgoud, A. M. *et al.* Metabolic engineering in the host yarrowia lipolytica. *Metab. Eng.* **50**, 192 – 208, DOI: https://doi.org/10.1016/j.ymben.2018.07.016 (2018). Metabolic Engineering Host Organism Special Issue.

28. Madzak, C., Treton, B. & Blanchin-Roland, S. Strong hybrid promoters and integrative expression/secretion vectors for quasi-constitutive expression of heterologous proteins in the yeast Yarrowia lipolytica. *J. Mol. Microbiol. Biotechnol.* **2**, 207–216 (2000).

29. Peña, D. A., Gasser, B., Zanghellini, J., Steiger, M. G. & Mattanovich, D. Metabolic engineering of pichia pastoris. *Metab. Eng.* **50**, 2 – 15, DOI: https://doi.org/10.1016/j.ymben.2018.04.017 (2018). Metabolic Engineering Host Organism Special Issue.

30. Gasser, B. & Mattanovich, D. A yeast for all seasons – Is Pichia pastoris a suitable chassis organism for future bioproduction? *FEMS Microbiol. Lett.* **365**, DOI: 10.1093/femsle/fny181 (2018). Fny181, http://oup.prod.sis.lan/femsle/article-pdf/365/17/fny181/25431392/fny181.pdf.

31. Mumberg, D., Müller, R. & Funk, M. Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds. *Gene* **156**, 119 – 122, DOI: https://doi.org/10.1016/0378-1119(95)00037-7 (1995).

32. Voth, W. P., Richards, J. D., Shaw, J. M. & Stillman, D. J. Yeast vectors for integration at the HO locus. *Nucleic Acids Res.* **29**, e59–e59, DOI: 10.1093/nar/29.12.e59 (2001). http://oup.prod.sis.lan/nar/article-pdf/29/12/e59/9905922/2900e59.pdf.

33. Lee, M. E., DeLoache, W. C., Cervantes, B. & Dueber, J. E. A highly characterized yeast toolkit for modular, multipart assembly. *ACS Synth. Biol.* **4**, 975–986 (2015).

34. Perez-Pinera, P. *et al.* Synthetic biology and microbioreactor platforms for programmable production of biologics at the point-of-care. *Nat. Commun.* **7**, 12211, DOI: 10.1038/ncomms12211 (2016).

35. Christianson, T. W., Sikorski, R. S., Dante, M., Shero, J. H. & Hieter, P. Multifunctional yeast high-copy-number shuttle vectors. *Gene* **110**, 119 – 122, DOI: https://doi.org/10.1016/0378-1119(92)90454-W (1992).

36. Gietz, R. D. & Schiestl, R. H. High-efficiency yeast transformation using the liac/ss carrier dna/peg method. *Nat. Protoc.* **2**, 31–34, DOI: 10.1038/nprot.2007.13 (2007).

37. Liu, G., Lanham, C., Buchan, J. R. & Kaplan, M. E. High-throughput transformation of saccharomyces cerevisiae using liquid handling robots. *PLOS ONE* **12**, 1–15, DOI: 10.1371/journal.pone.0174128 (2017).

38. Shao, Z., Zhao, H. & Zhao, H. DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways . *Nucleic Acids Res.* **37**, e16–e16, DOI: 10.1093/nar/gkn991 (2008). https://academic.oup.com/nar/article-pdf/37/2/e16/16756870/gkn991.pdf.

39. Si, T. *et al.* Automated multiplex genome-scale engineering in yeast. *Nat. Commun.* **8**, 15187, DOI: 10.1038/ncomms15187 (2017).

40. Luo, J., Sun, X., Cormack, B. P. & Boeke, J. D. Karyotype engineering by chromosome fusion leads to reproductive isolation in yeast. *Nature* **560**, 392–396, DOI: 10.1038/s41586-018-0374-x (2018).

41. Hegemann, J. H. & Heick, S. B. *Delete and Repeat: A Comprehensive Toolkit for Sequential Gene Knockout in the Budding Yeast Saccharomyces cerevisiae*, 189–206 (Humana Press, Totowa, NJ, 2011).

42. DiCarlo, J. E. *et al.* Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. *Nucleic Acids Res.* **41**, 4336–4343, DOI: 10.1093/nar/gkt135 (2013). http://oup.prod.sis.lan/nar/article-pdf/41/7/4336/25342046/gkt135.pdf.

43. Farzadfard, F., Perli, S. D. & Lu, T. K. Tunable and multifunctional eukaryotic transcription factors based on crispr/cas. *ACS Synth. Biol.* **2**, 604–613, DOI: 10.1021/sb400081r (2013).

44. Świat, M. A. *et al.* FnCpf1: a novel and efficient genome editing tool for Saccharomyces cerevisiae. *Nucleic Acids Res.* **45**, 12585–12598, DOI: 10.1093/nar/gkx1007 (2017). http://oup.prod.sis.lan/nar/article-pdf/45/21/12585/22146005/gkx1007.pdf.

45. Verwaal, R., Buiting-Wiessenhaan, N., Dalhuijsen, S. & Roubos, J. A. Crispr/cpf1 enables fast and simple genome editing of saccharomyces cerevisiae. *Yeast (Chichester, England)* **35**, 201–211, DOI: 10.1002/yea.3278 (2018). 28886218[pmid].

46. Schwartz, C. M., Hussain, M. S., Blenner, M. & Wheeldon, I. Synthetic RNA Polymerase III Promoters Facilitate High-Efficiency CRISPR-Cas9-Mediated Genome Editing in Yarrowia lipolytica. *ACS Synth Biol* **5**, 356–359 (2016).

47. Galanie, S., Thodey, K., Trenchard, I. J., Filsinger Interrante, M. & Smolke, C. D. Complete biosynthesis of opioids in yeast. *Science* **349**, 1095–1100, DOI: 10.1126/science.aac9373 (2015). https://science.sciencemag.org/content/349/6252/1095.full.pdf.

48. Meadows, A. L. *et al.* Rewriting yeast central carbon metabolism for industrial isoprenoid production. *Nature* **537**, 694 EP – (2016).

49. Dragosits, M. & Mattanovich, D. Adaptive laboratory evolution – principles and applications for biotechnology. **12**, 64, DOI: 10.1186/1475-2859-12-64.

50. Mans, R., Daran, J.-M. G. & Pronk, J. T. Under pressure: evolutionary engineering of yeast strains for improved performance in fuels and chemicals production. *Curr. Opin. Biotechnol.* **50**, 47 – 56, DOI: https://doi.org/10.1016/j.copbio.2017.10.011 (2018). Energy biotechnology • Environmental biotechnology.

51. Strucko, T. *et al.* Laboratory evolution reveals regulatory and metabolic trade-offs of glycerol utilization in saccharomyces cerevisiae. *Metab. Eng.* **47**, 73 – 82, DOI: https://doi.org/10.1016/j.ymben.2018.03.006 (2018).

52. Burén, S. *et al.* Formation of nitrogenase nifdk tetramers in the mitochondria of saccharomyces cerevisiae. *ACS Synth. Biol.* **6**, 1043–1055, DOI: 10.1021/acssynbio.6b00371 (2017).

53. Young, E. M. *et al.* Iterative algorithm-guided design of massive strain libraries, applied to itaconic acid production in yeast. *Metab. Eng.* **48**, 33–43, DOI: https://doi.org/10.1016/j.ymben.2018.05.002 (2018).

54. Casini, A. *et al.* A pressure test to make 10 molecules in 90 days: External evaluation of methods to engineer biology. *J. Am. Chem. Soc.* **140**, 4302–4316, DOI: 10.1021/jacs.7b13292 (2018). PMID: 29480720, https://doi.org/10.1021/jacs.7b13292.

55. Denby, C. M. *et al.* Industrial brewing yeast engineered for the production of primary flavor determinants in hopped beer. *Nat. Commun.* **9**, 965, DOI: 10.1038/s41467-018-03293-x (2018).

56. Awan, A. R. *et al.* Biosynthesis of the antibiotic nonribosomal peptide penicillin in baker's yeast. *Nat. Commun.* **8**, 15202, DOI: 10.1038/ncomms15202 (2017).

57. Ali, N., Rampazzo, R. d. C. P., Costa, A. D. T. & Krieger, M. A. Current nucleic acid extraction methods and their implications to point-of-care diagnostics. *BioMed research international* **2017**, 9306564–9306564, DOI: 10.1155/2017/9306564 (2017). 28785592[pmid].

58. van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. The third revolution in sequencing technology. *Trends Genet.* **34**, 666 – 681, DOI: https://doi.org/10.1016/j.tig.2018.05.008 (2018).

59. Watson, M. & Warr, A. Errors in long-read assemblies can critically affect protein prediction. *Nat. Biotechnol.* **37**, 124–126, DOI: 10.1038/s41587-018-0004-z (2019).

60. van Dijk, E. L., Jaszczyszyn, Y. & Thermes, C. Library preparation methods for next-generation sequencing: Tone down the bias. *Exp. Cell Res.* **322**, 12 – 20, DOI: https://doi.org/10.1016/j.yexcr.2014.01.008 (2014).

61. Wajid, B. & Serpedin, E. Review of general algorithmic features for genome assemblers for next generation sequencers. *Genomics, Proteomics & Bioinforma.* **10**, 58 – 73, DOI: https://doi.org/10.1016/j.gpb.2012.05.006 (2012).

62. Li, Z. *et al.* Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Brief Funct Genomics* **11**, 25–37, DOI: 10.1093/bfgp/elr035 (2012).

63. Lischer, H. E. L. & Shimizu, K. K. Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC bioinformatics* **18**, 474–474, DOI: 10.1186/s12859-017-1911-6 (2017). 29126390[pmid].

64. Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **13**, 329–342, DOI: 10.1038/nrg3174 (2012).

65. Hernandez, D., François, P., Farinelli, L., Østerås, M. & Schrenzel, J. De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. *Genome Res.* **18**, 802–809, DOI: 10.1101/gr.072033.107 (2008). http://genome.cshlp.org/content/18/5/802.full.pdf+html.

66. Simpson, J. T. *et al.* Abyss: a parallel assembler for short read sequence data. *Genome Res* **19**, 1117–23, DOI: 10.1101/gr.089532.108 (2009).

67. Zerbino, D. R. & Birney, E. Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Res.* **18**, 821–829, DOI: 10.1101/gr.074492.107 (2008).

68. Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677, DOI: 10.1093/bioinformatics/btt476 (2013). http://oup.prod.sis.lan/bioinformatics/article-pdf/29/21/2669/18533361/btt476.pdf.

69. Antipov, D., Korobeynikov, A., McLean, J. S. & Pevzner, P. A. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* **32**, 1009–1015, DOI: 10.1093/bioinformatics/btv688 (2015). http://oup.prod.sis.lan/bioinformatics/article-pdf/32/7/1009/19568450/btv688.pdf.

70. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Comput. Biol.* **13**, 1–22, DOI: 10.1371/journal.pcbi.1005595 (2017).

71. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–10, DOI: 10.1093/bioinformatics/btw152 (2016).

72. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**, 722–736, DOI: 10.1101/gr.215087.116 (2017).

73. Ruan, J. Ultra-fast de novo assembler using long noisy reads. *GitHub* (2018).

74. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546, DOI: 10.1038/s41587-019-0072-8 (2019).

75. Giordano, F. *et al.* De novo yeast genome assemblies from minion, pacbio and miseq platforms. *Sci Rep* **7**, 3935, DOI: 10.1038/s41598-017-03996-z (2017).

76. Salazar, A. N. *et al.* Nanopore sequencing enables near-complete de novo assembly of saccharomyces cerevisiae reference strain cen.pk113-7d. *FEMS Yeast Res* **17**, DOI: 10.1093/femsyr/fox074 (2017).

77. Jenjaroenpun, P. *et al.* Complete genomic and transcriptional landscape analysis using third-generation sequencing: a case study of saccharomyces cerevisiae cen.pk113-7d. *Nucleic Acids Res* **46**, e38, DOI: 10.1093/nar/gky014 (2018).

78. Khatri, I., Tomar, R., Ganesan, K., Prasad, G. S. & Subramanian, S. Complete genome sequence and comparative genomics of the probiotic yeast saccharomyces boulardii. *Sci Rep* **7**, 371, DOI: 10.1038/s41598-017-00414-2 (2017).

79. Fournier, T. *et al.* High-quality de novo genome assembly of the dekkera bruxellensis yeast using nanopore minion sequencing. *G3 (Bethesda)* **7**, 3243–3250, DOI: 10.1534/g3.117.300128 (2017).

80. Tondini, F., Jiranek, V., Grbin, P. R. & Onetto, C. A. Genome sequence of australian indigenous wine yeast torulaspora delbrueckii coft1 using nanopore sequencing. *Genome Announc.* **6**, DOI: 10.1128/genomeA.00321-18 (2018).

81. Bizzarri, M. *et al.* Draft genome sequences of the highly halotolerant strain zygosaccharomyces rouxii atcc 42981 and the novel allodiploid strain zygosaccharomyces sapae atb301(t) obtained using the minion platform. *Microbiol Resour Announc.* **7**, DOI: 10.1128/MRA.00874-18 (2018).

82. Love, K. R. *et al.* Comparative genomics and transcriptomics of pichia pastoris. *Bmc Genomics* **17**, DOI: ARTN55010.1186/s12864-016-2876-y (2016).

83. Olsen, R. A. *et al.* De novo assembly of dekkera bruxellensis: a multi technology approach using short and long-read sequencing and optical mapping. *Gigascience* **4**, DOI: ARTN5610.1186/s13742-015-0094-1 (2015).

84. McIlwain, S. J. *et al.* Genome sequence and analysis of a stress-tolerant, wild-derived strain of saccharomyces cerevisiae used in biofuels research. *G3-Genes Genomes Genet.* **6**, 1757–1766, DOI: 10.1534/g3.116.029389 (2016).

85. ONT. Medaka: Sequence correction provided by ont research. *GitHub* (2018).

86. Vaser, R., Sovic, I., Nagarajan, N. & Sikic, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**, 737–746, DOI: 10.1101/gr.214270.116 (2017).

87. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963, DOI: 10.1371/journal.pone.0112963 (2014).

88. de Toro, M., Garcilláon-Barcia, M. P. & De La Cruz, F. Plasmid diversity and adaptation analyzed by massive sequencing of escherichia coli plasmids. *Microbiol. Spectr.* **2** (2014).

89. Arredondo-Alonso, S., Willems, R. J., van Schaik, W. & Schürch, A. C. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb. genomics* **3**, e000128–e000128, DOI: 10.1099/mgen.0.000128 (2017). 29177087[pmid].

90. Berlin, K. *et al.* Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol* **33**, 623–30, DOI: 10.1038/nbt.3238 (2015).

91. Miller, J. R. *et al.* Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**, 2818–2824, DOI: 10.1093/bioinformatics/btn548 (2008). http://oup.prod.sis.lan/bioinformatics/article-pdf/24/24/2818/16885558/btn548.pdf.

92. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090, DOI: 10.1093/bioinformatics/btx346 (2017). https://academic.oup.com/bioinformatics/article-pdf/33/19/3088/25164703/btx346.pdf.

93. Valli, M. *et al.* Curation of the genome annotation of pichia pastoris (komagataella phaffii) cbs7435 from gene level to protein function. *FEMS Yeast Res* **16**, DOI: 10.1093/femsyr/fow051 (2016).

94. Kuberl, A. *et al.* High-quality genome sequence of pichia pastoris cbs7435. *J Biotechnol* **154**, 312–20, DOI: 10.1016/j.jbiotec.2011.04.014 (2011).

95. Liu, L. & Alper, H. S. Draft genome sequence of the oleaginous yeast yarrowia lipolytica po1f, a commonly used metabolic engineering host. *Genome Announc.* **2**, DOI: 10.1128/genomeA.00652-14 (2014).

96. Zhang, L., Liang, Y., Wu, W., Tan, X. & Lu, X. Microbial synthesis of propane by engineering valine pathway and aldehyde-deformylating oxygenase. *Biotechnol. for biofuels* **9**, 80–80, DOI: 10.1186/s13068-016-0496-z (2016). 27042209[pmid].

97. Verwaal, R. *et al.* High-level production of beta-carotene in saccharomyces cerevisiae by successive transformation with carotenogenic genes from xanthophyllomyces dendrorhous. *Appl. Environ. Microbiol.* **73**, 4342–4350, DOI: 10.1128/AEM.02759-06 (2007). https://aem.asm.org/content/73/13/4342.full.pdf.

98. Kersten, R. D. *et al.* A red algal bourbonane sesquiterpene synthase defined by microgram-scale nmr-coupled crystalline sponge x-ray diffraction analysis. *J. Am. Chem. Soc.* **139**, 16838–16844 (2017).

99. Scheler, U. *et al.* Elucidation of the biosynthesis of carnosic acid and its reconstitution in yeast. *Nat. Commun.* **7**, 12942 (2016).

100. Cao, X. *et al.* Metabolic engineering of oleaginous yeast *Yarrowia lipolytica* for limonene overproduction. *Biotechnol. for Biofuels* **9**, 214 (2016).

101. Jongedijk, E. *et al.* Capturing of the monoterpene olefin limonene produced in saccharomyces cerevisiae. *Yeast* **32**, 159–171, DOI: 10.1002/yea.3038 (2015). https://onlinelibrary.wiley.com/doi/pdf/10.1002/yea.3038.

102. Sheff, M. A. & Thorn, K. S. Optimized cassettes for fluorescent protein tagging in saccharomyces cerevisiae. *Yeast* **21**, 661–670, DOI: 10.1002/yea.1130 (2004). https://onlinelibrary.wiley.com/doi/pdf/10.1002/yea.1130.

103. Lee, S., Lim, W. A. & Thorn, K. S. Improved blue, green, and red fluorescent protein tagging vectors for s. cerevisiae. *PLOS ONE* **8**, 1–8, DOI: 10.1371/journal.pone.0067902 (2013).

104. Souza-Moreira, T. M. *et al.* Screening of 2A peptides for polycistronic gene expression in yeast. *FEMS Yeast Res.* **18**, DOI: 10.1093/femsyr/foy036 (2018). Foy036, https://academic.oup.com/femsyr/article-pdf/18/5/foy036/24968970/foy036.pdf.

105. Nijkamp, J. F. *et al.* De novo sequencing, assembly and analysis of the genome of the laboratory strain saccharomyces cerevisiae cen.pk113-7d, a model for modern industrial biotechnology. *Microb Cell Fact* **11**, 36, DOI: 10.1186/1475-2859-11-36 (2012).

106. Jenjaroenpun, P. *et al.* Complete genomic and transcriptional landscape analysis using third-generation sequencing: a case study of saccharomyces cerevisiae cen.pk113-7d. *Nucleic Acids Res* **46**, e38, DOI: 10.1093/nar/gky014 (2018).

107. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. Quast: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–5, DOI: 10.1093/bioinformatics/btt086 (2013).

108. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12, DOI: 10.1186/gb-2004-5-2-r12 (2004).

109. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. Busco: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–2, DOI: 10.1093/bioinformatics/btv351 (2015).

110. Küberl, A. *et al.* High-quality genome sequence of pichia pastoris cbs7435. *J. Biotechnol.* **154**, 312 – 320, DOI: https://doi.org/10.1016/j.jbiotec.2011.04.014 (2011).

111. Darling, A. E., Mau, B. & Perna, N. T. progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLOS ONE* **5**, 1–17, DOI: 10.1371/journal.pone.0011147 (2010).