Genome Sequence and Analysis of a Stress-Tolerant, Wild-Derived Strain of Saccharomyces cerevisiae Used in Biofuels Research

Sean J. McIlwain,* David Peris,*†‡§ Maria Sardi,*†‡ Oleg V. Moskvin,* Fujie Zhan,*†
Kevin S. Myers,*†ǁ Nicholas M. Riley,‡ Alyssa Buzzell,ǁ Lucas S. Parreiras,* Irene M. Ong,*
Robert Landick,*ǁǁ Joshua J. Coon,*ǁǁ Audrey P. Gasch,*ǁǁ Trey K. Sato,*
and Chris Todd Hittinger*†ǁǁ§§

*Department of Energy (DOE) Great Lakes Bioenergy Research Center, †Laboratory of Genetics, ‡Genome Center of Wisconsin, §Wisconsin Energy Institute, J. F. Crow Institute for the Study of Evolution, **Microbiology Doctoral Training Program, ‖Department of Computer Sciences, ‖‖Department of Chemistry, *ǁǁDepartment of Biochemistry, and ‖‖‖Department of Biomolecular Chemistry, University of Wisconsin-Madison, Wisconsin 53706 and §§Medical College of Wisconsin, Milwaukee, Wisconsin 53226

ORCID ID: 0000-0002-3820-8400 (S.J.M.)

ABSTRACT The genome sequences of more than 100 strains of the yeast Saccharomyces cerevisiae have been published. Unfortunately, most of these genome assemblies contain dozens to hundreds of gaps at repetitive sequences, including transposable elements, tRNAs, and subtelomeric regions, which is where novel genes generally reside. Relatively few strains have been chosen for genome sequencing based on their biofuel production potential, leaving an additional knowledge gap. Here, we describe the nearly complete genome sequence of GLBRCY22-3 (Y22-3), a strain of S. cerevisiae derived from the stress-tolerant wild strain NRRL YB-210 and subsequently engineered for xylose metabolism. After benchmarking several genome assembly approaches, we developed a pipeline to integrate PacifiC Biosciences (PacBio) and Illumina sequencing data and achieved one of the highest quality genome assemblies for any S. cerevisiae strain. Specifically, the contig N50 is 693 kbp, and the sequences of most chromosomes, the mitochondrial genome, and the 2-micron plasmid are complete. Our annotation predicts 92 genes that are not present in the reference genome of the laboratory strain S288c, over 70% of which were expressed. We predicted functions for 43 of these genes, 28 of which were previously uncharacterized and unnamed. Remarkably, many of these genes are predicted to be involved in stress tolerance and carbon metabolism and are shared with a Brazilian bioethanol production strain, even though the strains differ dramatically at most genetic loci. The Y22-3 genome sequence provides an exceptionally high-quality resource for basic and applied research in bioenergy and genetics.

KEYWORDS lignocellulosic hydrolysates Pacific Biosciences (PacBio) genome assembly genome annotation novel genes

Cellulosic bioethanol is a promising sustainable and renewable liquid transportation fuel (U.S. DOE 2006). Bioethanol is also a model fuel that is helping researchers understand the roadblocks involved in forcing cellular carbon flux away from biomass into toxic end-products, a challenge shared with advanced biofuels, including isobutanol and farnesene (Hong and Nielsen 2012; Buijs et al. 2013; U.S. DOE 2015). Although the yeast Saccharomyces cerevisiae has long been employed to convert starch sugars into ethanol, fermentation of sugars derived from the lignocellulose that makes up the cell wall of plants is more challenging. Due to its recalcitrant nature, lignocellulose-rich plant biomass, such as corn stover, must first be chemically, thermally, and/or mechanically pretreated to allow enzymes to efficiently hydrolyze cellulose and hemicellulose polymers into fermentable sugars. Although pretreatment methods can be effective at decreasing the hydrolysis time and increasing sugar yield, these methods often introduce toxic
byproducts, including weak acids, amides, and aromatic compounds derived from the lignin itself; many of these compounds have potent negative effects on microbial fermentation (Piotrowski et al. 2014). In an attempt to mitigate the impacts of these and other stresses caused by fermentation, industrial S. cerevisiae strains have been selected for their robust tolerance phenotypes and further developed for lignocellulosic ethanol production, including the strains PE-2 (Pereira et al. 2014) and Ethanol Red (Demeke et al. 2013a). The genome sequences of several bioethanol production strains, including the PE-2-derivative JAY291 (Argueso et al. 2009), have been published, but the identities of the genes and variants that confer stress tolerance and other industrially desirable properties have generally remained unclear (Baibrzad et al. 2012; Zheng et al. 2012; Sahara et al. 2014; Ulaganathan et al. 2015; Sravanthi Goud and Ulaganathan 2015).

In addition to the challenge of growth inhibitors from lignocellulosic hydrolysates, native S. cerevisiae is unable to ferment hemicellulosic pentose sugars, such as xylose, which constitute the second largest fraction of sugars in corn stover and most other plant biomass (Pauly and Keegstra 2008). Several groups have partially overcome these challenges by using strategies that combine rational engineering (e.g., over-expressing genes encoding enzymes required for xylose fermentation) and directed evolution (e.g., selecting for improved growth on xylose). These genetically modified strains of S. cerevisiae have a range of abilities to ferment the xylose present in lignocellulosic hydrolysates (van Maris et al. 2007; Koppram et al. 2012; Demeke et al. 2013a,b; Wei et al. 2013; Parreiras et al. 2014; Smith et al. 2014). Nonetheless, for evolved strains, it has often been unclear which mutations are responsible for the improved xylose fermentation.

The GLBRCY22-3 (Y22-3) yeast strain was developed to better understand the fermentation of xylose in lignocellulosic hydrolysates. Y22-3 is a monosporic derivative of NRRL YB-210 (YB-210), a wild strain of S. cerevisiae isolated from Costa Rican bananas (Mortimer and Johnston 1986). The YB-210 strain background was chosen for its unusual ability to tolerate high concentrations of ethanol (Wohlbach et al. 2014), elevated temperature, and the inhibitory compounds found in lignocellulosic hydrolysates made by two different types of alkaline pretreatment (Jin et al. 2013; Parreiras et al. 2014; Sato et al. 2014). In contrast, the standard S288c lab strain fares poorly under these stressful conditions. Although YB-210 does not utilize appreciable xylose natively, it was genetically engineered to express several heterologous enzymes required for efficient xylose metabolism; Y22-3 is one such haploid clone (Parreiras et al. 2014).

Through the directed evolution of Y22-3 on xylose as the sole sugar source, a haploid clone, Y128, was isolated that could anaerobically ferment both glucose and xylose in Ammonia Fiber Expansion (AFEX) (Balan et al. 2009) pretreated corn stover hydrolysate (ACSH) (Parreiras et al. 2014).

Strains of S. cerevisiae and other species of Saccharomyces frequently contain genes not present in the S288c reference genome, especially in their subtelomeric regions (Liti and Louis 2005; Liti et al. 2009, 2013; Novo et al. 2009; Scannell et al. 2011; Borneman et al. 2011; Hittinger 2013; Bergström et al. 2014; Borneman and Pretorius 2015; Strope et al. 2015; Baker et al. 2015). These regions of yeast genomes are frequently laboratories of innovation where gene families expand, translocate, and evolve new functions (Carlson and Botstein 1983; Liti and Louis 2005; Hittinger 2013). Occasionally, genes are also added to these regions from other species by horizontal gene transfer (Novo et al. 2009; Dunn et al. 2012; Hittinger et al. 2015). Unfortunately, most whole genome shotgun sequencing strategies perform poorly on subtelomeric regions of the genome due to the widespread presence of selfish elements and polymorphic gene families with nearly identical sequences, leaving a blind (or at least blurry) spot in many genome assemblies where many of the most interesting and dynamic genes reside (Liti et al. 2009, 2013; Scannell et al. 2011; Borneman et al. 2011; Bergström et al. 2014; Strope et al. 2015; Baker et al. 2015). These genes can be responsible for novel traits (Borneman and Pretorius 2015), but investigation of these targets requires de novo genome sequencing strategies capable of obtaining refined genome assemblies with few gaps. Even for parts of the genome conserved with an essentially complete reference genome, such as S288c, the reliability of inferences from routine resequencing applications, such as RNA sequencing (RNA-Seq), copy-number variant (CNV) detection, and mutation inference, can be improved by mapping reads against a high-quality de novo assembly of the strain or line being studied (Pool 2015). Thus, a high-quality de novo assembly for Y22-3 is required to understand whether any novel genes have undergone mutations or changed their expression during its directed evolution into its more industrially relevant derivatives, such as the anaerobic xylose-fermenting strain Y128.

To enable functional genomic investigations of this emerging biofuel strain, we have assembled a high-quality reference genome for Y22-3. We benchmarked several genome assembly approaches, developed a genome assembly pipeline that integrated Pacific Biosciences (PacBio) sequencing reads with Illumina sequencing reads, and produced a fully annotated genome sequence. With few gaps in the nuclear genome, a

Figure 1 Scaffold N50 values obtained from various de novo assemblers with PacBio and paired-end Illumina reads. Note that, for the PacBio (Pacific Biosciences) assemblies, contig N50 values are equivalent to the scaffold N50 values.
complete mitochondrial genome, and a complete 2-micron plasmid sequence, the genome sequence of Y22-3 is among the highest quality \textit{S. cerevisiae} genome sequences published. The Y22-3 genome has 92 nonrepetitive genes that \textit{S288c} lacks, many of which are predicted to encode proteins whose functions are related to carbon metabolism or stress tolerance, including several that may be relevant to the strain’s tolerance to ACSH. Interestingly, although Y22-3 and the Brazilian bioethanol strain JAY291 are not closely related across most of their genomes, they share many genes that are rarely present in other strains. The Y22-3 genome sequence will provide an important foundation for basic and applied research.

### MATERIALS AND METHODS

Complete details are available in Supplemental Material, File S1. Briefly, a single colony of Y22-3 genetically engineered for xylose metabolism (Parreiras et al. 2014) was grown in 10 g/L yeast extract, 20 g/L peptone and 20 g/L dextrose (YPD), and its genomic DNA was isolated and purified. The DNA sample was sequenced using the PacBio RS II technology with a C2 chemistry sequencing kit (Pacific Biosciences) to 155 \times depth of coverage with an extracted subread length of 2881 ± 2177 bp and maximum read length of 35,845 bp (using \textit{--minReadScore 0.75, --minLength 500} for pbh5tools, Pacific Biosciences, Menlo Park, CA), and by using the Illumina HiSeq technology with 100 bp pair-end reads with a raw depth of coverage of 1038 \times. An optimal assembly method was found by testing a variety of assembly methods that utilize either or both of the PacBio and Illumina data sets. Methods tested included the \textit{de novo} assembly programs Sprai v. 0.9.9 (Imai 2013; Kamada et al. 2014), HGAP3 smart-analysis package v. 2.2.0.133377 (Chin et al. 2013), PBcR wgs-8.2beta (Koren et al. 2013), Velvet v. 1.2.10 (Zerbino and Birney 2008), and PBJelly (English et al. 2012), as well as the read preprocessing programs Trimmomatic (Bolger et al. 2014), BLESS (Heo et al. 2014), and RACER (Ilie and Molnar 2013). Subsampling the paired-end reads down to 7% of the total number of trimmed reads was also examined. After testing the assembly methods, we assembled the nuclear genome and the 2-micron plasmid using Sprai v. 0.9.9 (Imai 2013; Kamada et al. 2014) and the mitochondrial genome using Spades v 3.5.0 (Nurk et al. 2013). We corrected single nucleotide polymorphisms (SNPs) and indels with Quiver (Chin et al. 2013) using the PacBio reads and with GATK v 3.1-1 (Van der Auwera et al. 2013) using the Illumina reads. We then annotated the nuclear and 2-micron assemblies by comparing, contrasting, and combining the predicted results from YGAP (Proux-Wéra et al. 2012) and Liftover (Kuhn et al. 2013). The mitochondrial assembly was first annotated using Liftover, followed by manual annotation using GENEIOUS v. R6 (Kearse et al. 2012).

We validated the predicted protein coding genes of Y22-3 using: 1) single-end RNA-Seq data collected from four growth phases of Y22-3 grown on YP media containing 60 g/L dextrose and 30 g/L xylose (YPDX, equivalent sugar concentrations that mimic ACSH made with 6% glucan loading), 2) an optimized (Figure S1) \textit{de novo} transcriptome assembled by Trinity (Grabherr et al. 2011) using pair-end RNA-Seq data from clones derived from Y22-3 that were grown aerobically or anaerobically from four to six growth phases on YPDX and ACSH, and 3) proteomic data collected similarly to previous nanoflow liquid
chromatography tandem mass spectrometry (nLC-MS/MS) approaches (Hebert et al. 2014) from Y22-3 cells grown aerobically in YPD. We compared the potentially novel genes of Y22-3 to other representative strains of S. cerevisiae using BLAST (Altschul et al. 1997) and developed a Novelty Metric to quantify how distinct non-S288c genes were from their nonsyntenic homologs in S288c. We examined the relationship between Y22-3 and other S. cerevisiae strains by generating a maximum likelihood phylogeny using RAxML v 8.1.20 (Stamatakis 2014) on an orthologous nucleotide dataset built from protein-coding sequences conserved across all strains.

Data availability
This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession LBMA00000000. The version described in this paper is version LBMA01000000. All DNA and RNA sequencing reads have been deposited in the NCBI SRA under BioProject PRJNA279877. Raw files for mass spectrometry data from these experiments are available on Chorus (https://chorusproject.org/pages/index.html, Project ID 999). Strains are available upon request. The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article.

RESULTS AND DISCUSSION

De novo genome assembly
To optimize assembly methods, we compared each strategy by their respective scaffold N50 values and found a wide range of performances (Figure 1). Strategies using only paired-end Illumina reads performed poorly. PBJelly, an algorithm that uses PacBio reads to scaffold Illumina-based assemblies, offered modest improvement in scaffolding. Error correction of PacBio using Illumina reads proved both computationally intensive and was outperformed by PBJelly on our dataset. Genome assemblies that were produced using exclusively self-corrected PacBio reads, including Sprai, HGAP, and PBcR, performed considerably better. Since Sprai achieved the best scaffold N50 and had the highest putative accuracy (Table S1), we continued to develop our pipeline with Sprai (Figure S2 and File S2).

We made several corrections to the Sprai assembly, including polishing with Quiver (Chin et al. 2013); three iterations of corrections using a custom GATK pipeline (Van der Auwera et al. 2013); ultra-scaffolding by homology to S288c (Figure S3); and special treatment of several regions, including to recover complete assemblies for the 2-micron plasmid and mitochondrial genome (Nurk et al. 2013; Baker et al. 2015). The N50 of the final ultra-scaffolded assembly of the nuclear genome was 908 kbp, and the contig (and scaffold) N50 was 693 kbp. Only nine gaps and 15 unplaced contigs remained, most of which contained fragments of Ty elements, whose full-length size of ~6 kbp exceeded our average PacBio read length of 2.88 kbp. More than half of the chromosomes lacked any gaps, while chromosome XII contained the most gaps, including the one created by the rDNA repeats (Table S2).

Genome annotation summary
To maximize the transfer of annotations from S. cerevisiae and related species of yeasts, we compared, contrasted, integrated, and improved on the results of two annotation pipelines: Liftover (Kuhn et al. 2013),
which uses genome-wide alignment to a related genome, and the Yeast Genome Annotation Pipeline (YGAP), which features a de novo gene prediction algorithm and uses synteny and sequence similarity to infer homology (Proux-Wéra et al. 2012) (Figure S4). Using Liftover, we were able to transfer 6369 coding annotations from the S288c reference genome to the Y22-3 genome, of which 6004 were predicted to encode complete proteins. YGAP annotated 5820 genes, of which 5352 were predicted to encode complete proteins. We developed and applied an algorithm that corrected 123 of 365 Liftover annotations and 250 of 468 YGAP annotations, mainly by extending or shortening open reading frames (ORFs). After combining the annotations from Liftover and YGAP, manually correcting a few annotations, and manually correcting the mitochondrial annotations, we obtained 6319 valid coding gene annotations, 242 pseudogene annotations, and 297 tRNA annotations. Complete information for each gene, including the rationale for the proposed standard names, can be found in Table S6.

Validation of predicted genes using transcriptomic and proteomic data

To determine the impact that a nearly complete reference genome had on downstream functional genomic analyses, we compared the number of RNA-Seq reads mapped using the new Y22-3 reference genome, instead of the S288c reference genome. We observed a substantial increase in the fraction of RNA-Seq readsthat could be mapped uniquely (83% vs. 78%), as well as a decrease in the number of reads that could not be mapped at all (Table 1). These results strongly suggest that the inclusion of novel genes and divergent alleles from Y22-3 is important for genomic applications based on read mapping.

To validate the expression of predicted genes, we used transcriptomic and proteomic data to perform three different types of analyses (Figure 2, Table S3, and Table S4). First, we generated a de novo transcriptome assembly using 51 RNA-Seq experiments (Transcriptome Method). We considered a protein-coding gene validated if it had at least a 60% overlap with a predicted transcript that uniquely mapped to its locus. Second, we analyzed gene expression levels using...
eight RNA-Seq experiments (two replicates, four growth phases in YPD medium containing xylose, YPDX) (FPKM Method). We considered a protein-coding gene validated if it had an RNA expression value greater than 1 “Fragments Per Kilobase of transcript per Million mapped reads” (FPKM) in at least one experimental condition. Finally, a predicted protein was considered validated if one or more unique and unambiguously mapped peptides were detected by nLC-MS/MS (FDR < 0.01) (Protein Method).

Validations with individual methods ranged from 98.1% (6198/6319) for the FPKM Method to 51.1% (3228/6319) for the Protein Method (Figure 2). ORFs that were annotated as dubious by SGD (Cherry et al. 1998) were, perhaps not surprisingly, validated at considerably lower frequencies than ORFs that were not annotated as dubious [e.g., for the Protein Method, 54.5% (3225/5920) vs. 0.5% (3/641), P < 10^-195, Fisher’s Exact Test]. Even for genes not present in the S288c reference genome [excluding transposons, helicases, and other subtelomeric repeats detected using RepeatMasker (Smit et al. 2013)], we were able to validate 73.9% (68/92) by at least one method. Some of the genes not validated are the products of recent gene duplication events that cannot reliably be distinguished from one another.

The Y22-3 genome lacks several genes relative to S288c

The annotated Y22-3 genome lacks 296 protein-encoding genes that are present in S288c (Table S5). Of the 296 missing genes, 139 are in subtelomeric regions in S288c (defined as within 50 kbp of the end of the assembled chromosome), and 156 are annotated as dubious ORFs. All five missing essential genes correspond to ORFs annotated as dubious, and prior experimental work in S288c suggests that deletion of these five dubious ORFs is lethal in S288c due to their effects on neighboring genes, rather than their intrinsic protein-coding potential (Engel et al. 2014). The assembly gap at the rDNA locus is responsible for 14 missing ORFs, while an assembly gap in the subtelomeric region of Chromosome II could explain the absence of four S288c ORFs, including two helicases, a dubious ORF, and a gene with no known function. Thus, we conclude that the missing genes are generally not assembly artifacts, but rather reflect differences in gene content. At least 22 of the missing genes have homologous genes on different chromosomes, suggesting that some of their functions may be performed by these non-syntenic homologs. For example, Y22-3 appears to be missing SOR1, a gene encoding a sorbitol (and xylitol) dehydrogenase in S288c (Sarthy et al. 1994; Toivari et al. 2004; Wenger et al. 2010), but it retains the nearly identical paralog SOR2.

The Y22-3 genome encodes several genes previously characterized in non-S288c strains

Several genes of interest for xylose metabolism (Wenger et al. 2010), stress tolerance, or other functions have been experimentally characterized in strains of S. cerevisiae other than S288c (Borneman and Pretorius 2015). Many of these genes have homologs in the Y22-3 genome, the S288c genome, or both. To quantify how distinct non-S288c genes are from their closest homolog in S288c, we developed a Novelty Metric to compare the strength of the best TBLASTN hit to the Y22-3 genome to the best TBLASTN hit to the S288c genome. Briefly, for each query gene, we subtracted the bit score generated against the S288c genome from the bit score generated against the Y22-3 genome (or any genome). We then normalized this value against the highest bit score generated against any S. cerevisiae genome in the dataset (see File S1, Equation 1). Thus, if a genome has a sequence that is closely related to a previously characterized non-S288c gene, it scores highly, while that genome scores poorly if it only has genes that are closely related to S288c homologs. Importantly, our Novelty Metric can recover homologs that are not annotated in the target genomes and quantifies how similar these sequences are to the non-S288c genes.

Using our Novelty Metric, we found that Y22-3 contains several previously characterized genes that S288c lacks, many of which have roles in stress tolerance or metabolism that may be relevant to biofuel production (Figure 3A). These genes include BIO1 and BIO6, two genes involved in biotin synthesis (Hall and Dietrich 2007); RTMI, which confers resistance to the toxicity of molasses, a substrate often used for industrial yeast biomass and ethanol production (Ness and Aigle 1995); KHR1, which encodes a heat-resistant killer toxin (Wei et al. 2007); MPR1 or its close paralog, MPR2, which encodes a L-azetidine-2-carboxylic acid acetyl-transferase that can confer resistance to ethanol and freezing (Takagi et al. 2000); and YJM-GNAT, which encodes another N-acetyl-transferase (Wei et al. 2007). Critically, the Y22-3 genome does not encode XDH1, a gene encoding a xylitol dehydrogenase (Wenger et al. 2010) that could have interfered with the engineered xylose fermentation pathway. Although most have not been functionally characterized, many non-S288c ORFs have been predicted in other strains of S. cerevisiae (Argueso et al. 2009; Dowell et al. 2010;
Several novel genes and gene clusters are predicted to encode functions related to stress tolerance and carbon metabolism

To further explore the genetic basis of the unusual stress tolerance and carbon metabolism properties of Y22-3, we closely examined 43 genes present in Y22-3 but not in S288c (Figure 4 and Table S6). For clarity, we did not consider repeat sequences that represent selfish elements (e.g., Ty elements) and genes with no known functions [e.g., PAU (Seripauperin) and COS genes] in the main manuscript (see Table S7 for full documentation). Expression of each of these 43 genes was detected in at least one condition by the FPKM Method (Table S6). Many are nonsyntenic homologs that are similar to well-characterized genes, whereas others are more divergent, and their putative functional assignments are more tentative.

To quantify how novel these genes are, we again used our Novelty Metric to search for these genes in a panel of diverse strains with published genome sequences, as well as two other wild stress-tolerant strains (Birren et al. 2005; Wei et al. 2007; Argueso et al. 2009; Novo et al. 2009; Dowell et al. 2010; Borneman et al. 2011, 2012; Roncoroni et al. 2011; Akao et al. 2011; Zheng et al. 2012; Wohlbach et al. 2014; Fay et al. 2014a, b; Song et al. 2015). Most of these genes were found in a minority of the strains examined (Figure 4), suggesting that they could be at least partly responsible for some of the Y22-3 traits relevant to biofuel production. Interestingly, many of these genes are shared with another biofuel strain, JAY291 (Argueso et al. 2009), despite the fact that these strains are not phylogenetically closely related across most of their genome (Figure 5). This remarkable overlap advances the shared novel genes as particularly promising candidates for future studies investigating shared industrially relevant traits.

Several novel genes are predicted to encode functions related to stress tolerance, carbon metabolism, aldehyde or alcohol detoxification, and biofuel synthesis. A total of 28 genes with functional annotations were not syntenic and lacked reciprocal best-BLAST hits with S288c, and we have proposed standard names for them (Figure 4 and Table S6). For example, a homolog of ADH6, which encodes a cinnamyl alcohol dehydrogenase (Larroy et al. 2002), was especially divergent in sequence (48% maximum protein sequence identity), and we propose ADH8 as its standard name. Since ferulic acid, p-coumaric acid, and related aromatic lignin degradation products are among the most toxic fermentation inhibitors in ACSH and many other lignocellulosic hydrolysates (Piotrowski et al. 2014), genes that reduce aromatic aldehydes into their less toxic alcohols may be beneficial. We also found two nonsyntenic homologs of DDI2 and DDI3, which were recently shown to encode identical cyanamide hydratases in S288c (Li et al. 2013). If their activity is broader or the divergent (88% identical) homolog (DDI72) present in Y22-3 has novel activities, these genes might also metabolize other amides present in ACSH, such as acetamide, feruloyl amide, and p-coumaroyl amide (Chundawat et al. 2010).

The Y22-3 genome encodes several nonsyntenic homologs of genes involved in vitamin B1 (thiamine) and vitamin B6 metabolism. The novel gene THI75 is distantly related (39% identical) to known thiamine transporters, while several additional genes are involved in the synthesis of pyridoxal 5′-phosphate, which is the active form of vitamin B6 (the novel genes SNOS and SNZ5, as well as additional nonsyntenic homologs of each). Previous studies on sugarcane bioethanol strains have found that increased copy numbers of SNO and SNZ genes improve growth in high sugar media lacking pyridoxine (vitamin B6) (Stambuk et al. 2009). Pyridoxal 5′-phosphate is a precursor for thiamine biosynthesis, and thiamine pyrophosphate is an obligate cofactor for many enzymes required for fermentation and the pentose phosphate pathway, including pyruvate decarboxylase and transketolase. The presence of additional copies of these genes in the Y22-3 genome
suggests that similar constraints on vitamin B1 and B6 metabolism may also be important for lignocellulosic biofuel production.

As is common in *S. cerevisiae* (Liti and Louis 2005; Liti et al. 2009; Strope et al. 2015), most (37/43) of these novel genes and nonsyntenic homologs mapped to subtelomeric regions, including an invertase (SUC72r), an α-galactosidase (MEL11), several floculins (FLO59, FLO95, and FLO70), and three Zn(II)Cys₆ transcription factors (ZTF2, ZTF3, and ZTF4). As is typically seen in *S. cerevisiae* (Liti and Louis 2005; Liti et al. 2009; Strope et al. 2015), most of the novel genes and nonsyntenic homologs are present in clusters (Figure 4 and Table S6). Ten clusters of two or more of these genes were found in Y22-3 (Figure 6, Figure S6, Figure S7, Figure S8, Figure S9, Figure S10, Figure S11, Figure S12, and Figure S13), but several clusters deserve special mention. One of the few nonsubtelomeric clusters is located in the interior of the right arm of chromosome IV and encodes a fungal transcription factor (FTF1), a floculin (ZbaiFLO11), a nicotinic acid permease (ZbaiTNA1), an oxoprolinase (ZbaiOXP1), and a Zn(II)Cys₆ transcription factor (ZTF1) (Wohlbach et al. 2014; Parreiras et al. 2014) (Figure S7). These genes were also horizontally transferred from *Zygosaccharomyces bailii* into several wine strains (Novo et al. 2009) and apparently into Y22-3. The revised genome assembly presented here both completes and firmly places this and other clusters onto Y22-3 chromosomes, whereas the previous assembly often left such clusters incomplete and unplaced (Wohlbach et al. 2014). Many clusters include genes whose functions are likely related, such as the subtelomeric region of the right arm of chromosome VII, which includes a second complete maltose utilization cluster embedded within the MAL1 cluster present in S288c; this novel cluster encodes a divergent isomaltase (MAL72), maltose transporter (MAL71), and activating Zn(II)Cys₆ transcription factor (MAL73) (Figure S10). The interior of the right arm of chromosome VIII contains at least six copies of CUP1 with a spacing consistent with the recently described Type 3 (Zhao et al. 2014) configuration (Figure S11); the locus could also contain additional copies because no PacBio reads fully spanned the repeats. Most strikingly, both the left and right subtelomeric regions of chromosome X contain clusters of genes related to thiamine metabolism and encoding amide hydrolases (Figure 6). The nonsyntenic homologs of the genes present in the right subtelomeric region of chromosome X are relatively closely related to genes present on the left subtelomeric region of chromosome VI in S288c, while those in the left subtelomeric region of chromosome X appear to be highly divergent in Y22-3 and are often shared only with the bioethanol strain JAY291 (Figure 4).

Conclusions

Here, we have developed a genome assembly pipeline that integrates PacBio and deep Illumina paired-end sequencing coverage. The Y22-3 genome sequence assembled is one of the highest quality *S. cerevisiae* genome sequences published. Most nuclear chromosomes are complete, including several challenging regions, such as subtelomeric regions. The mitochondrial genome and 2-micron plasmid sequences are complete. Careful annotation revealed several novel genes and gene clusters, many of which have predicted roles in stress tolerance or fermentation. Genes involved in thiamine metabolism, involved in carbon metabolism, encoding enzymes that act on aromatic lignin degradation products, and encoding amidases, are likely to be particularly relevant for biofuel production by Y22-3 in ACHS and other lignocellulosic hydrolysates. Strikingly, many closely related genes are also found in the genome of the Brazilian bioethanol strain JAY291, suggesting that there may be a common genetic basis for some of their industrially relevant properties. The complete genome sequence of Y22-3 will enable ongoing and future investigations into its novel properties, including approaches using molecular genetics, functional genomics, and directed evolution.

ACKNOWLEDGMENTS

We thank Brian Haas for his assistance in optimizing the Trinity pipeline’s parameters, Mary Tremaine for technical support, Yaoping Zhang for running the fermentations with Y22-3, David Benton for management and critical reading of the manuscript, Dana Wohlbach for providing the assembled genome sequences and novel gene predictions for NNRRL Y-2209 (= LEP) and NNRRL Y-389 (= MUSH), and Gavin Sherlock for providing the XDH1 sequence from *S. cerevisiae* strain EC1118. We also thank the University of Wisconsin Biotechnology Center DNA Sequencing Facility and Gene Expression Center for performing Illumina DNA and RNA sequencing, respectively, and the United States Department of Energy (DOE) Joint Genome Institute (JGI), a DOE Office of Science User Facility, for providing Illumina RNA sequencing. The work conducted by the DOE JGI is supported by the Office of Science of the United States DOE under contract number DE-AC02-05CH11231. This material is based upon work supported by the National Science Foundation (NSF) Graduate Research Fellowship Program under grant number DGE-1256259 to M.S. and N.M.R. This work was funded in part by the DOE Great Lakes Bioenergy Research Center (DOE Office of Science BER-FC02-07ER64494 to R.L., J.C., A.P.G., T.K.S., and C.T.H.) and the NSF (grant number DEB-1253634 to C.T.H.). C.T.H. is a Pew Scholar in the Biomedical Sciences and an Alfred Toepfer Faculty Fellow, supported by the Pew Charitable Trusts and the Alexander von Humboldt Foundation, respectively.

LITERATURE CITED

Akao, T., I. Yashiro, A. Hosoyama, H. Kitagaki, H. Horikawa et al., 2011 Whole-Genome Sequencing of Sake Yeast *Saccharomyces cerevisiae* Kyokai no. 7. DNA Res. 18: 423–434.


Borneman, A. R., B. A. Desany, D. Riches, J. P. Affourtit, A. H. Forgan et al., 2011 Whole-genome comparison reveals novel genetic elements that...
characterize the genome of industrial strains of Saccharomyces cerevisiae. PLoS Genet. 7: e1001287.


Li, J., M. Biss, Y. Fu, X. Xu, S. A. Moore et al., 2015 Two duplicated genes DDI2 and DDI3 in budding yeast encode a cyanamide hydratase and are induced by cyanamide, J. Biol. Chem. 290: 12664–12675.


Liti, G., A. N. Nguyen Ba, M. Blythe, C. A. Müller, A. Bergström et al., 2013 High quality de novo sequencing and assembly of the Saccharomyces arboricus genome. BMC Genomics 14: 69.


Sravanthi Goud, B., and K. Ulaganathan, 2015  Draft genome sequence of Saccharomyces cerevisiae strain NCIM3186 used in the production of bioethanol from sweet sorghum. Genome Announc. 3: e00813–e00815.


Ulaganathan, K., B. Sravanthi Goud, M. M. Reddy, V. Praveen Kumar, S. Radhakrishna et al., 2015  Genome sequence of Saccharomyces cerevisiae NCIM3107, used in bioethanol production. Genome Announc. 3: e01557-14.


Zhao, Y., P. K. Strope, S. G. Kozmin, J. H. McCusker, F. S. Dietrich et al., 2014  Structures of naturally evolved CUP1 tandem arrays in yeast indicate that these arrays are generated by unequal nonhomologous recombination. G3 (Bethesda) 4: 2259–2269.


Communicating editor: M. J. Cherry
Isolation of Y22-3 genomic DNA: The genetic background and genetic engineering of GLBRCY22-3 (Y22-3) have been described elsewhere (Parreiras et al. 2014). Briefly, Y22-3 is a haploid MATa derivative of YB-210, which has been engineered for xylose utilization. A single colony of Y22-3 strain was cultured in 500 mL 10 g/L yeast extract, 20 g/L peptone, and 20 g/L dextrose (YPD) media at 30 °C overnight. Cells were centrifuged at 3,000 relative centrifugal force (RCF) for 15 min at 4 °C. The resulting cell pellet was washed with 10 mM Tris, 0.5 mM EDTA, pH 8.0 and centrifuged. The washed cell pellet was then resuspended in 30 mL of 50 mM EDTA and incubated with 1,000 U Zymolyase at 30 °C for 1 hr. The resulting spheroplasts were centrifuged at 3,000 RCF at 4 °C for 15 min; completely resuspended in 15 mL sterile 0.2 M Tris, pH 8.5, 0.25 M NaCl, 25 mM EDTA, 0.5% sodium dodecylsulfate, 3 mg RNase A, 8 mg Proteinase K; and incubated at 50 °C for 30 min. Cell lysate was then clarified by centrifugation; extracted with 10 mL Tris-equilibrated phenol, pH 8.0; and incubated at room temperature for 20 min. The phenol-lysate mixture was then centrifuged at 3,000 RCF for 15 min, and the upper aqueous phase was transferred to a sterile 50 mL conical tube. The aqueous phase was subsequently extracted twice with 10 mL chloroform, as done with the phenol. To the final phenol/chloroform-extracted aqueous phase, 1/10th of the total volume of 3 M sodium acetate and two volumes of isopropanol were added and mixed by inversion. DNA was allowed to precipitate by incubation at -80 °C for 20 min and pelleted by centrifugation at 5,000 RCF for 30 min at 4 °C. The DNA pellet was washed with ice-cold 70% ethanol, air dried for 15 min, and then dissolved in 10 mM Tris, 0.5 mM EDTA, pH 8.0 (TE buffer). Purified genomic DNA
(gDNA) was then further column-purified (Genomic Tip-500, Qiagen) to a final concentration of 60 ng/μL in 200 μL TE buffer.

**Pacific Biosciences library preparation and sequencing:** The isolated gDNA was submitted to the Medical College of Wisconsin for PacBio sequencing. Initial quantification was performed using a Quant-iT™ PicoGreen® dsDNA Assay Kit (Life Technologies, Carlsbad, CA, USA) and a NanoDrop 2000 (Thermo Fisher Scientific Inc., Waltham, MA, USA). gDNA was purified and concentrated with a 0.45x AMPure PB bead wash (Pacific Biosciences, Menlo Park, CA, USA). About 5 μg of concentrated gDNA was sheared to 10 kbp using a Covaris gTube (Covaris Inc, Woburn, MA, USA). A Pacific Biosciences (PacBio) sequencing library, or SMRTbell™, was constructed using the SMRTbell™ Template Prep Kit 1.0 and the 10 kbp Template Preparation and Sequencing with Low-Input DNA procedure (Pacific Biosciences). P4 Polymerase was coupled with the resulting SMRTbell™ library. For sequencing, the library was bound to MagBeads by incubating for 1 hr at 4 °C. The final library was run over 8 V2 SMRT cells on the PacBio RSII using a C2 chemistry sequencing kit (Pacific Biosciences). Each SMRT cell was visualized using a 1x180 min movie. The PacBio subreads were extracted using pbh5tools (PacificBiosciences, Menlo Park, CA, USA) with --minReadScore 0.75 and --minLength 500.

**Illumina library preparation and sequencing:** The isolated gDNA was submitted to the University of Wisconsin Biotechnology Center (UWBC) DNA Sequencing Facility for Illumina sequencing. DNA concentration and sizing were verified using the Qubit® dsDNA HS Assay Kit (Life Technologies, Grand Island, NY) and Agilent DNAHS chip (Agilent Technologies, Inc., Santa Clara, CA), respectively.
Samples were prepared according the TruSeq® Genomic DNA Sample Preparation kit v2 (Illumina Inc., San Diego, California, USA) with minor modifications. Samples were sheared using either a Diagenode Bioruptor (Diagenode USA, Denville, NJ) or a Covaris M220 Ultrasonicator (Covaris Inc, Woburn, MA, USA). Libraries were size-selected for an average insert size of 275 bp using a 2% Invitrogen E-Gel (Life Technologies). Quality and quantity of the finished libraries were assessed using an Agilent DNA1000 chip and Qubit® dsDNA HS Assay Kit, respectively. Cluster generation was performed using standard Cluster Kits and the Illumina Cluster Station. Paired-end, 100-bp sequencing was performed using standard SBS chemistry on an Illumina HiSeq 2000 sequencer. Images were analyzed using the standard Illumina Pipeline, version 1.8.2.

**Optimization of the de novo genome assembly:** Many de novo genome assembly pipelines can utilize long but error-prone PacBio reads and/or short Illumina reads. Some pipelines can leverage the two, either by using the Illumina data to correct the sequencing errors in the PacBio data or by scaffolding prior de novo assemblies using the PacBio data. Methods also exist to self-correct Illumina data prior to assembly (Ilie and Molnar 2013; Heo et al. 2014). To determine the optimal genome assembly method, we tested several de novo genome assembly and read preprocessing algorithms.

Using only PacBio reads, we ran Sprai v. 0.9.9 (Imai 2013; Kamada et al. 2014), HGAP3 smart-analysis package v. 2.2.0.133377 (Chin et al. 2013), and PBcR wgs-8.2beta (Koren et al. 2013) using their default or provided parameters. Using only
paired-end Illumina reads, we created assemblies using Velvet v. 1.2.10 (Zerbino and Birney 2008) with a series of different \textit{kmer} sizes \{45, 47, \ldots 99\}. In addition to using raw reads, we also tried preprocessing the reads using Trimmomatic (Bolger \textit{et al.} 2014), Bless (Heo \textit{et al.} 2014), and RACER (Ilie and Molnar 2013). We also tried subsampling the paired-end reads down to 7\% of the total number of trimmed reads (6,121,566/87,466,528 or about 44X/625X trimmed coverage). Finally, we ran RACER with and without prior trimming with Trimmomatic (Bolger \textit{et al.} 2014). PBcR (Koren \textit{et al.} 2013) allows for the use of Illumina reads to correct the PacBio reads prior to assembly using Celera (Myers \textit{et al.} 2000). Unfortunately, computational limitations prevented us from using all of the Illumina data, so we again reduced the number of reads to 7\% (PBcR-44X). We also used PBJelly (English \textit{et al.} 2012) to scaffold PBcR (PBJelly-PBcR) and Velvet (PBJelly) assemblies using the PacBio reads. The accuracy of the Sprai, HGAP, and PBcR assemblies was assessed using QUAST v3.2 build 26.11.2015 (Gurevich \textit{et al.} 2013).

Although the contig (and scaffold) N50 of the Sprai assembly was of exceptionally high continuity at 671 kbp, polishing our assembly with Quiver (Chin \textit{et al.} 2013) detected and corrected 1359 SNPs and 1073 indels (247 deletions, and 826 insertions). We then removed redundant contigs with the check\_redundancy.pl script included in the Sprai package and ordered the remaining contigs by aligning homologous regions to the S288c genome. After the putative order was determined, adjacent contigs were checked for possible overlaps using Staden (Bonfield and Whitwham 2010), a process that replaced eight gaps with a consensus sequence. Contigs that could not be joined were then ultra-scaffolded by homology to S288c using
a gap with an arbitrary size of 5 kbp. A handful of small inversions were detected using dot plots built with Gepard (Krumsieck et al. 2007), but no major non-subtelomeric translocations were detected relative to S288c (Figure S3); we cannot definitively exclude the possibility that translocations exist at the few gaps remaining in our assembly. We trimmed and reoriented the complete 2-micron plasmid to match the orientation of S288c. Sprai performed poorly on the mitochondrial genome, so we assembled the Sprai-corrected PacBio reads and paired-end Illumina reads into the complete mitochondrial genome with Spades v 3.5.0 (Bankevich et al. 2012), which was recently shown to yield complete Saccharomyces mitochondrial assemblies from paired-end Illumina reads alone (Baker et al. 2015). We iteratively deployed GATK v 3.1-1 (Van der Auwera et al. 2013) to correct an additional 9 SNPs and 217 indels with the Illumina paired-end data, reaching convergence after three iterations.

After selecting Sprai as the optimal de novo assembler, additional single nucleotide polymorphisms (SNPs) and indels were found with GATK v 3.1-1 (Van der Auwera et al. 2013) using the Illumina paired-end reads as previously described (Wohlbach et al. 2014; Clowers et al. 2015). Briefly, paired-end reads were aligned to the assembly using Bowtie2 (Langmead and Salzberg 2012) with default settings with the exception of $-N=1$ to allow for one mismatch. SNPs and indels were identified with GATK using base quality score recalibration, indel realignment, duplicate removal, and depth of coverage analysis (McKenna et al. 2010; Van der Auwera et al. 2013) using the default parameters with the exception for $-mbq 30$ for the UnifiedGenotyper program. Variants were filtered using the following suggested GATK criteria: $QD < 2$, $FS > 60$, and $MQ < 40$ for SNPs and $QD < 2$, and $FS > 200$ for indels. The predicted SNPs
and indels were visually verified using IGV (Robinson et al. 2011; Thorvaldsdóttir et al. 2013), and the corrections were incorporated into the final assembly.

**Preliminary genome annotation:** The final assembly was annotated using YGAP (Proux-Wéra et al. 2012) and Liftover using S288C-R64 as the reference with \textit{minIdentity} 98\% and \textit{minScore} 100 (Kuhn et al. 2013). Initial annotations and quality checks for these methods proceeded separately until the annotations were integrated as noted below. We used BLASTN to find the \textit{ScTAL1}, \textit{CpXylA}, \textit{PsXYL3}, and \textit{Kan}^R genes from the xylose utilization cassette engineered into Y22-3 (Parreiras et al. 2014). Although we annotated these genes and examined their expression, we excluded them from all analyses of novel genes and non-syntenic homologs because they were not in the parent strain NRRL YB-210. All annotations with Ns within their sequences were removed. For protein-coding annotations, we determined whether the predicted open reading frame (ORF) would produce a valid protein by noting whether each ORF included:

1) A correct start codon
2) A correct stop codon
3) A sequence length divisible by 3
4) No internal stop codons

For both YGAP and Liftover annotations, we found that some predicted ORFs failed these tests. We considered any ORF annotations invalid if they did not pass all four criteria. We attempted to fix the annotations that lacked introns in the following ways:
Invalid start codons: For sequences containing only an invalid start codon, we searched upstream in frame until we encountered a valid start codon. If a stop codon was first encountered while searching upstream, we then searched downstream in frame for the next valid start codon.

Invalid stop codons or ORFs not divisible by 3: For sequences containing only an invalid stop codon, we searched downstream in frame until a valid stop codon was found. If the sequence had a start codon but was not divisible by 3, we removed the remainder (i.e. one or two bp) from the end of the annotation and attempted to find a downstream stop codon in the frame of the start codon.

Invalid start and stop codons: For sequences containing both an invalid start and stop codon, we first tried to fix the start codon as described above and, if successful, then attempted to fix the stop codon.

Invalid annotations due to internal stop codons were not fixed, nor were annotations that failed for other reasons not listed above. After attempting to fix the ORF annotations, we annotated all remaining invalid ORF annotations as putative pseudogenes.

Integrating YGAP and Liftover annotations: To integrate YGAP and Liftover gene annotations, we first compared the genomic coordinates. We categorized annotations as 1) agreeing in both the genomic coordinates and the annotation name, 2) agreeing in coordinates but disagreeing in name, 3) conflicting based upon their coordinates, or 4) conflicting and overlapping with multiple annotations. We also compared the coding sequences between introns and observed some that agreed,
others that conflicted, and others that conflicted with multiple annotations. We resolved conflicts using the following rules:

1) Keep the annotation that encodes a valid protein.

2) If both encode valid proteins, keep the annotation from Liftover.

3) If multiple Liftover annotations overlap (e.g. S288c pseudogenes or split ORFs), keep the annotation that matches a valid YGAP annotation.

4) For cases where multiple annotations conflict, keep only the annotations where YGAP and Liftover agree.

We also manually fixed some annotations as follows: The YJM-GNAT (Wei et al. 2007) gene was found in the sequence using BLASTN. ORFs YMR084W and YMR085W were merged into a complete opening reading frame, which is known to exist in some strains of S. cerevisiae (Kellis et al. 2003); since it is a paralog of GFA1, we called this merged annotation GFA2 in Y22-3. Similarly, YAR073W and YMR075W comprise the complete ORF for IMD1 (Escobar-Henriques and Daignan-Fornier 2001) in Y22-3. A paralog of VTH1 was found by extending YCR101C, and the overlapping YCR100C and YCR099C ORFs were deleted. Finally, annotations to other non-S288c genes previously described (MPR1, RTM1, BIO1, BIO6, and KHR1) were renamed to match the literature (Borneman and Pretorius 2015).

We included other annotations found by Liftover (e.g. rRNAs, repeat_regions) within the final annotation. For tRNAs, we used both the annotations found by Liftover and YGAP, keeping the Liftover annotation in cases of disagreements regarding the genomic coordinates. Some centromeres were identified manually using BLAST. After manual inspection and circularization to match the orientation of S288c, the 2-micron
plasmid was annotated using Liftover (S288C reference, minIdentity 85%, minScore 30). The mitochondrial genome sequence was initially annotated using Liftover and finished manually using Geneious v. R6 (Kearse et al. 2012). Repeats were found using RepeatMasker v4.0.5, database 20140131 (Smit et al. 2013) with parameters -species 'Saccharomyces cerevisiae' -s -no_is -cutoff 255 -frag 20000.

**Y22-3 RNA-Sequencing for gene expression analysis (UWBC dataset):** Y22-3 was cultured as described previously (Parreiras et al. 2014). In brief, Y22-3 was inoculated to an optical cell density at 600 nm (OD600) of 0.1 in 2 L Applikon bioreactors containing 1.8 L 10 g/L yeast extract, 20 g/L peptone, 60 g/L dextrose, 30 g/L xylose, 50 mM potassium phosphate, pH 5.0 (YPDX) and sparged with air. Cell samples were harvested during four stages of growth: minimal amount of glucose left (Residual Glucose Phase), during the transition from glucose to xylose consumption (Transition Phase), 1-10 hours after the transition (Early Xylose Phase), and ~20 hours after the transition (Xylose Phase) (Parreiras et al. 2014). For each harvest time point, 40 mL of culture were removed from the bioreactors; added to 5 mL ice-cold 95:5 ethanol:water-saturated phenol, pH 6.6; and mixed by inversion. Fixed cells were then pelleted by centrifugation at 10,000 RCF for 5 min at 4 °C. The supernatant was decanted and the remaining pellet flash frozen in a dry ice-ethanol bath. Total RNA was extracted from the Y22-3 cell pellet by hot phenol extraction (Gasch 2002). RNA was further purified using RNeasy Plus columns (Qiagen) according to manufacturer's protocol.

Total RNA was submitted to the UWBC Gene Expression Center, which verified purity and integrity using a NanoDrop2000 Spectrophotometer and an Agilent 2100 BioAnalyzer, respectively. Samples that met the Illumina sample input guidelines were
prepared according the TruSeq® Stranded Total RNA Sample Preparation Guide (Rev. E) using the Illumina® TruSeq® Stranded Total RNA Sample Preparation kits (Illumina Inc., San Diego, California, USA) with minor modifications. For each library preparation, 2 μg of total RNA was subjected to ribosomal RNA depletion using the EpiCentre RiboZero™ Gold Ribosomal RNA Removal (Yeast) kit (EpiCentre Inc., Madison, WI, USA) as directed. Ribosomal RNA-depleted RNA samples were purified using paramagnetic beads (Agencourt RNA Clean XP beads, Beckman Coulter, Indianapolis IN, USA). Subsequently, each rRNA-depleted sample was fragmented using divalent cations under elevated temperature. The fragmented RNA was synthesized into double-stranded cDNA using SuperScript II Reverse Transcriptase (Invitrogen, Carlsbad, California, USA) and random primers for first-strand cDNA synthesis, followed by second-strand synthesis using DNA Polymerase I, and RNase H to degrade the mRNA. Double-stranded cDNA was purified using paramagnetic beads (Agencourt AMPure XP beads, Beckman Coulter). The cDNA products were incubated with Klenow DNA Polymerase to add an ‘A’ nucleotide (deoxyadenosine monophosphate) to the 3’ end of the blunt DNA fragments. DNA fragments were ligated to forked Illumina adapters, which have a single ‘T’ nucleotide (deoxythymidine monophosphate) overhang at their 3’ end. The adapter-ligated DNA products were purified using paramagnetic beads. Adapter-ligated DNA was amplified in a Linker Mediated PCR reaction (LM-PCR) for 11 cycles using Phusion™ DNA Polymerase and Illumina’s PE genomic DNA primer set, followed by purification with paramagnetic beads. Quality and quantity of the finished libraries were assessed using an Agilent DNA1000 chip (Agilent Technologies, Inc., Santa Clara, CA, USA) and Qubit® dsDNA HS Assay Kit (Invitrogen, Carlsbad,
California, USA), respectively. Cluster generation was performed using standard Cluster Kits (v3) and the Illumina Cluster Station. Single-end, 100-bp sequencing was performed using standard SBS chemistry (v3) on an Illumina HiSeq 2500 sequencer. Images were analyzed using the standard Illumina Pipeline, version 1.8.2.

**RNA-Sequencing for de novo transcriptome assembly (JGI dataset):** Clones derived from Y22-3 were inoculated to an optical cell density at 600 nm (OD$_{600}$) of 0.1 in 2 L Applikon bioreactors containing 1.8 L YPDX or AFEX-pretreated corn stover hydrolysate and sparged with air (aerobic) or nitrogen (anaerobic). Cell samples were harvested during six stages of growth: during glucose consumption (Late Glucose Phase), minimal amount of glucose left (Residual Glucose Phase), during the transition from glucose to xylose consumption (Transition Phase), 2-6 hours after the transition (Early Xylose Phase), 8 hours after transition to xylose (Intermediate Xylose Phase) and ~20 hours after the transition (Xylose Phase) (Parreiras *et al.* 2014). RNA was purified as described above and submitted to JGI. Plate-based RNA sample preparation was performed on the PerkinElmer Sciclone NGS robotic liquid handling system using Illumina’s TruSeq Stranded mRNA HT sample prep kit ([http://support.illumina.com/sequencing/sequencing_kits/truseq_stranded_mrna_ht_sample_prep_kit.html](http://support.illumina.com/sequencing/sequencing_kits/truseq_stranded_mrna_ht_sample_prep_kit.html)) following the protocol outlined by Illumina in their user guide and with the following conditions: total RNA starting material was 1 μg per sample, and 10 cycles of PCR were used for library amplification.

The library was quantified using KAPA Biosystem’s next-generation sequencing library qPCR kit and run on a Roche LightCycler 480 real-time PCR instrument. The quantified library was then multiplexed with other libraries for a pool size of six, and the
pool was then prepared for sequencing on the Illumina HiSeq sequencing platform utilizing a TruSeq paired-end cluster kit v3 and Illumina’s cBot instrument to generate a clustered flowcell for sequencing. Sequencing of the flowcell was performed on the Illumina HiSeq 2000 sequencer using a TruSeq SBS sequencing kit 200 cycles v3 following a 2x150 indexed run recipe.

**Proteomic experiments and validation (Protein Method):** Proteomic experiments were performed similarly to our previous approach (Hebert et al. 2014). For protein extraction and digestion, yeast cell pellets were lysed by glass bead milling (Retsch GmbH, Germany). Lysate protein concentration was measured via bicinchoninic acid protein assay (Thermo Pierce, Rockford, IL), and yeast proteins were reduced through incubation in 5 mM dithiothreitol for 45 min at 58 °C. Free cysteines were alkylated in 15 mM iodoacetamide in the dark for 30 min. The alkylation was stopped with 5 mM DTT. A 1 mg protein aliquot was digested overnight at room temperature in 1.5 M Urea with trypsin (Promega, Madison, WI) added at a 1:50 (w/w) enzyme to protein ratio. Digestions were quenched by the addition of trifluoroacetic acid and were then desalted over tC18 Sep-Pak cartridges (Waters, Milford, MA).

For online nanoflow liquid chromatography tandem mass spectrometry (nLC-MS/MS), reversed phase columns were packed in-house using 75 µm ID, 360 µm OD bare fused silica capillary. A nanoelectrospray tip was laser pulled (Sutter Instrument Company, Novato, CA) and packed with 1.7 µm diameter, 130 Å pore size Ethylene Bridged Hybrid C18 particles (Waters) to a length of 30-35 cm. Buffer A consisted of 0.2% formic acid and 5% DMSO in water, and Buffer B consisted of 0.2% formic acid in acetonitrile. Two µg of peptides were loaded onto the column in 95% buffer A for 12 min
Gradient elution was performed at 300 nL min\(^{-1}\), and gradients increased linearly from 5 to 35% buffer B over 190 min, followed by an increase to 70% B at 215 min and a wash at 70% B for 5 min. The column was then re-equilibrated at 5% B for 20 min. Eluting peptides were ionized with electrospray ionization at +2 kV, and the inlet capillary temperature was held at 300 °C on an ion-trap Orbitrap hybrid mass spectrometer (Orbitrap Elite, Thermo Fisher Scientific, San Jose, CA). Survey scans of peptide precursors were collected over the 300-1500 Thompson range in the Orbitrap with an automatic gain control target value of 1,000,000 (50 ms maximum injection time), followed by data-dependent ion trap MS/MS scans using collisional activation dissociation (CAD) of the 20 most intense peaks (AGC target value of 5,000 and maximum injection times of 100 ms). Precursors with charge states equal to one or unassigned were rejected.

A total of three replicates were performed, two of which were biological (i.e. independent cultures) and one of which was technical (i.e. a second pellet was collected for one of the biological replicates). These three pellets were collected separately and also handled separately at each step of the proteomics analysis (including cell lysis, proteolytic digestion, and LC-MS/MS analysis). Peptide mixtures from each of these pellets were injected/analyzed once, giving a total of three replicates. Since peptide quantification was highly correlated for both the biological and technical replicates (\(R^2 > 0.96\)), we processed all data in batch but used the match-between-runs feature of MaxQuant (Version 1.4.1.2) (Cox and Mann 2008), thus allowing validation of the expression of predicted proteins by any of the three replicates.
Raw data was processed using MaxQuant (Version 1.4.1.2) (Cox and Mann 2008), and tandem mass spectra were searched with the Andromeda search algorithm (Cox et al. 2011). Oxidation of methionine was specified as a variable modification, while carbamidomethylation of cysteine was set as a fixed modification. A precursor search tolerance of 20 ppm and a product mass tolerance of 0.35 Da were used for searches, and three missed cleavages were allowed for full trypsin specificity. Peptide spectral matches (PSMs) were made against a target-decoy custom database of the Y22-3 protein-coding annotations (created August 25, 2015), which was concatenated with a reversed sequence version of the forward database. Peptides were filtered to a 1% false discovery rate (FDR), and a 1% protein FDR was applied according to the target-decoy method. Proteins were identified using at least one peptide (razor + unique), where the razor peptide is defined as a non-unique peptide assigned to the protein group with the most other peptides (Occam’s razor principle). Using these methods, we identified 29,227 unique peptides (Table S4). For the final analysis, we only considered predicted proteins to be validated we could detect at least one unique peptide that could be unambiguously assigned to it.

Validating genes by RNA-Sequencing (FPKM Method): For validation of predicted protein-coding genes by the FPKM (Fragments Per Kilobase of transcript per Million mapped reads) Method, we first trimmed the aligned the UWBC single-end Illumina data using Trimmomatic (Bolger et al. 2014) with the following rules: 1) remove any number of bp from 3’ end that have the average quality score < 26 in a 3-bp sliding window, and 2) keep the trimmed read if 25 or more bp are left. The trimmed reads were then aligned using Tophat v2.1.0 (Trapnell et al. 2012). We then removed non-
uniquely mapped reads. Finally, we extracted counts and calculated the FPKM for each gene annotation using Cufflinks v2.2.1 (Trapnell et al. 2012).

De novo transcriptome assembly (Transcriptome Method): To describe transcriptional activity of Y22-3 over a wide range of conditions, nearly 1.5 billion (1,433,309,474) paired-end Illumina RNA-Seq reads from the JGI dataset were assembled de novo to yield a generalized transcriptome model (Parreiras et al. 2014). To remove both low-quality and nucleotide composition-biased parts of the sequencing reads, the Trimmomatic software (Bolger et al. 2014) was applied to pre-process the reads with the following rules: 1) remove the first 12 bp from 5' end, 2) remove any number of bp from 3' end that have the average quality score < 30 in a 3-bp sliding window, and 3) keep the trimmed read if 36 or more bp are left.

Transcriptome assembly was performed using the Trinity pipeline (Grabherr et al. 2011). The pool of reads was normalized to a target coverage of 50 using Trinity's in silico normalization routine. Transcriptome assembly with default parameters produced numerous artificial fusion transcripts. To optimize the assembly parameters for our particular case, extensive parameter scanning and optimization were performed by generating 270 de novo assemblies that combined 10 levels of minimal k-mer coverage (Inchworm stage of the Trinity), three levels of minimal glue, three levels of minimal iso ratio, and three levels of glue factor (Chrysalis stage). Optimization of the Butterfly-stage parameters was not recommended by the Trinity developers.

Selection of the best assembly was performed with the aid of the DETONATE package (Li et al. 2014), using both the RSEM-EVAL and REF-EVAL pipelines. Our
preliminary applications of DETONATE to smaller-size assemblies, combined with visual assessment of the assembly quality after subsequent mapping to the genome sequence, revealed that 3 out of 47 output statistics generated by the package - *Transcript Length Distribution Related Factors* (when maximized), *Unweighted K-mer KL_A_to_M* (when minimized), and *Unweighted_Pair_F1* (when maximized) - are better representatives of the overall assembly quality for our transcriptome. We selected the three candidate assemblies with top scores according to each of the three statistics. The final assembly was selected by minimizing the sum of ranks; this assembly belonged to the top 5% of assemblies for all three ranked lists. The following advanced Trinity parameters were used to generate the optimized transcriptome assembly: 

- *min_kmer_cov 32* (Inchworm stage), *min_glue 4*, *min_iso_ratio 0.01*, and *glue_factor 0.01* (Chrysalis stage). The transcripts from the optimized transcriptome assembly were mapped onto the Y22-3 genome sequence via the first stage of the PASA pipeline (Haas *et al.* 2003) with blat and gmap aligners and the following options: --

- *MAX_INTRON_LENGTH 2000* and --*transcribed_is_aligned_orient*. Alignment stringency was set to minimum of 95% identity across at least 90% of the transcript sequence. Visual comparison of the mapping results derived from the optimized and default-parameter assemblies (*Figure S1*) revealed that cases of artificial both-strand coverage by predicted transcripts, which were abundant in the default assembly, were essentially eliminated in the optimized assembly, without sacrificing sensitivity (seen as coverage of the genomic features predicted at DNA level).
Comparing strains by presence of genes using TBLASTN: To compare the genome content of several diverse S. cerevisiae genomes, we developed a Novelty Metric to integrate TBLASTN results. We used a TBLASTN word size of two. From the bits for the best TBLASTN hit of the query gene against a given subject strain, we subtracted the number of bits for the best hit against S288c. We then divided that value by the number of bits for the best hit from all of the strains searched for that particular gene. Negative values were set to zero, and the result was scaled from 0-100 (Equation 1).

\[
\text{MatchScore}(\text{gene}_i, \text{strain}_j) = \max[0, \frac{\text{Bits}(\text{gene}_i, \text{strain}_j) - \text{Bits}(\text{gene}_i, \text{strain}_S288c)}{\max_{k \in \text{strain}}(\text{Bits}(\text{gene}_i, \text{strain}_k) - \text{Bits}(\text{gene}_i, \text{strain}_S288c))}] \times 100
\]

This Novelty Metric is capable of distinguishing novel genes, even when there is a relatively close homolog in S288c. We generated heat maps of previously characterized non-S288c genes (Goto et al. 1991; Borneman and Pretorius 2015; Kowalec et al. 2015) and for genes predicted but not characterized in other strains (Figure 3). We excluded \textit{IMI1} from strain W303 because the difference in gene content is caused by the absence of a stop codon relative to S288c, which fuses two adjacent genes and their intergenic region (Kowalec et al. 2015), rendering the Novelty Metric uninformative. We recorded the value as "not applicable" for any query gene where the Novelty Metric for the all strains, including the source strain, was zero (e.g. the S288c hit was equally good, or the novel ORF predicted by other authors did not match the published genome sequence).
Examination of non-S288c genes with predicted functions: Of the 92 genes predicted in Y22-3 but not in S288c we carefully examined those genes where we could assign a putative function, leading to the 43 genes in Figure 4 and Table S6. These 43 genes exclude repeat sequences missed by RepeatMasker (e.g. some Y' helicases) and genes homologous to genes with no known functions (e.g. PAU (which encode seripauperins) and COS genes) (Table S7). KHR1 (Goto et al. 1991) was retained for these analyses, even though it was found within a LTR region. For each of these genes, we used the WU-BLASTP web tool on SGD (with default settings, except with no filters) to recover and record the best High-scoring Segment Pair (HSP) in S288c. We manually examined other close matches and gene annotations in SGD. We used the BLASTP web tool on NCBI (with default settings, except with no filters) to search the "nr" GenBank database for annotated proteins in other organisms, including other strains of S. cerevisiae. We manually examined entries for other close matches to determine whether putative functional inferences could be made or whether similar genes had already been named. For each of the 28 novel genes named, the best hit in S288c was already annotated as orthologous in the Y22-3 genome (i.e. it was not a reciprocal best BLAST hit). This criterion excluded 3 clusters of subtelomeric genes that may represent translocations. Furthermore, except for systematic names, each of these genes did not closely match a non-S288c gene that had already been named in another strain of S. cerevisiae.

Phylogenetic comparison to other Saccharomyces strains: To study the relationship between Y22-3 and Saccharomyces strains, we generated a maximum likelihood (ML) phylogeny using RAxML v. 8.1.20 (Stamatakis 2014) on a nucleotide
dataset containing orthologous protein-coding sequences conserved across all strains, including the outgroup *S. paradoxus* (Scannell *et al.* 2011). We performed multiple sequence alignments using DIALIGN2 (Morgenstern 1999) and removed indels, ambiguous nucleotides, and heterozygous nucleotides by codon as previously described (Hittinger *et al.* 2010). We reconstructed the ML phylogeny using 30 BFGS searches and the GTR+GAMMA (Yang 1993) model of nucleotide evolution. Bootstrap support values were determined for the best ML tree using 100 pseudoreplicates.

**R Software Used:** In house scripts for data analysis and plotting were written in R (R Core Team 2015) using the packages *Biostrings* (Pages *et al.*), *VennDiagram* (Chen and Boutros 2011), *gplots* (Warnes *et al.* 2015), *ggtree* (Yu *et al.*), and *ggplot2* (Wickham 2009).

**LITERATURE CITED**


Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin *et al.*, 2012 SPAdes: A
New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing.


Kowalec, P., M. Grynberg, B. Pająk, A. Socha, K. Winiarska et al., 2015 Newly identified protein Imi1 affects mitochondrial integrity and glutathione homeostasis in *Saccharomyces cerevisiae*. FEMS Yeast Res 15: fov048.


Parreiras, L. S., R. J. Breuer, R. Avanasi Narasimhan, A. J. Higbee, A. La Reau et al.,


R Core Team, 2015 R: A Language and Environment for Statistical Computing.


 Warnes, G. R., B. Bolker, L. Bonebakker, R. Gentleman, W. H. A. Liaw et al., 2015
gplots: Various R programming tools for plotting data.

Wei, W., J. H. McCusker, R. W. Hyman, T. Jones, Y. Ning et al., 2007 Genome
sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM789.


Wohlbach, D. J., N. Rovinskiy, J. a. Lewis, M. Sardi, W. S. Schackwitz et al., 2014
Comparative genomics of *Saccharomyces cerevisiae* natural isolates for bioenergy

Yang, Z., 1993 Maximum-likelihood estimation of phylogeny from DNA sequences when

visualization and annotation of phylogenetic tree with different types of meta-data.

Zerbino, D. R., and E. Birney, 2008 Velvet: algorithms for de novo short read assembly
FILE S2: GENOME ASSEMBLY SCRIPTS

**Sprai+Quiver**

ezez_vx1.pl ec.spec pbasm.spec
pbalign.py --nproc 8 --forQuiver --tmpDir /tmp all.fofn $assembly $b.cmp.h5
quiver -j 8 $b.cmp.h5 -r $assembly -o consensus.fa -o variants.gff

ec.spec

#### common ####

# input_for_database: filtered subreads in fasta or fastq format
input_for_database PacBio/Y22-3.pacbio.filtered_subreads.fastq

# min_len_for_query: the subreads longer than or equal to this value will be corrected
min_len_for_query 500

#if you don't know the estimated genome size, give a large number
estimated_genome_size 12156677
#if you don't know the estimated depth of coverage, give 0
estimated_depth 100

#
ca_path Programs/wgs-8.0/Linux-amd64/bin/

# the number of processes used by all vs. all alignment
# = 'partition' (in single node mode)
# = 'pre_partition' * 'partition' (in many node mode)
pre_partition 2
partition 12

# sprai prefer full paths
# if you use ezez4qsub*.pl. you MUST specify blast_path & sprai_path
# blast_path: where blastn and makeblastdb exist in
blast_path /usr/bin/blastn
# sprai_path: where binaries of sprai (bfmt72s, nss2v_v3 and so on) exist in
sprai_path /home/smcilwain/bin/

#### many node mode (advanced) ####

# sge: options for all the SGE jobs
# sge -soft -l ljob,lmem,sjob
# queue_req: additional options for all the SGE jobs
# queue_req -l s_vmem=4G -l mem_req=4
# longestXx_queue_req: if valid, displaces queue_req
# longestXx_queue_req -l s_vmem=64G -l mem_req=64
#BLAST_RREQ: additional options for SGE jobs of all vs. all alignment
#BLAST_RREQ -pe def_slot 4

##### common (advanced) #####

# used by blastn
word_size 18
evalue 1e-50
num_threads 1

#valid_voters 11

#trim: both ends of each alignment by blastn will be trimmed 'trim' bases to detect chimeric reads
trim 42

pbasm.spec
# Copyright (c) 2011-2013, Pacific Biosciences of California, Inc.
#
# All rights reserved.
#
# Redistribution and use in source and binary forms, with or without
# modification, are permitted (subject to the limitations in the
# disclaimer below) provided that the following conditions are met:
#
#  * Redistributions of source code must retain the above copyright
#    notice, this list of conditions and the following disclaimer.
#  *
#  * Redistributions in binary form must reproduce the above
#    copyright notice, this list of conditions and the following
#    disclaimer in the documentation and/or other materials provided
#    with the distribution.
#  *
#  * Neither the name of Pacific Biosciences nor the names of its
#    contributors may be used to endorse or promote products derived
#    from this software without specific prior written permission.
#
# NO EXPRESS OR IMPLIED LICENSES TO ANY PARTY'S PATENT RIGHTS ARE
# GRANTED BY THIS LICENSE. THIS SOFTWARE IS PROVIDED BY PACIFIC
# BIOSCIENCES AND ITS CONTRIBUTORS "AS IS" AND ANY EXPRESS OR
# IMPLIED
# WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES
# OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE
# DISCLAIMED. IN NO EVENT SHALL PACIFIC BIOSCIENCES OR ITS
# CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL,
# SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT
unitigger = bogart
#utgErrorRate = 0.015
#utgErrorLimit = 4.5

cnsErrorRate = 0.25
cgwErrorRate = 0.25
ovlErrorRate = 0.015

frgMinLen = 1000
ovlMinLen = 40

merSize=14

merylMemory = 16384
merylThreads = 8

ovlStoreMemory = 16384

# grid info
useGrid = 0
scriptOnGrid = 0
frgCorrOnGrid = 0
ovlCorrOnGrid = 0

sge = -S /bin/bash -V -q all.q
#sge = -S /bin/bash -sync y -V -q all.q
sgeScript = -pe threads 1
sgeConsensus = -pe threads 1
sgeOverlap = -pe threads 4
sgeFragmentCorrection = -pe threads 4
sgeOverlapCorrection = -pe threads 1

#ovlHashBits = 22
#ovlHashBlockLength = 46871347
#ovlRefBlockSize = 537
ovlHashBits = 25
ovlThreads = 4
ovlHashBlockLength = 50000000
ovlRefBlockSize = 100000000

ovlConcurrency = 6
frgCorrThreads = 4
frgCorrBatchSize = 100000
ovlCorrBatchSize = 100000

cnsMinFrgs = 7500
cnsConcurrency = 24

# change sgeName every time if you do not want to wait for the jobs not necessary to wait
sgeName = iroha

**PBCR**

PBCR -length 500 -partitions 200 -l Y22-3.pacbio -s pacbio.spec -fastq $pacbio_fastq
genomeSize=12156677

pacbio.spec
merSize=14

PBcR + Illumina

pacbio_fastq="Y22-3.pacbio.filtered_subreads.fastq"

pbcrl_length=500
pbcrl_partitions=200
genome_size=12156677

illumina_fastq_R1=sample_R1.fastq
illumina_fastq_R2=sample_R2.fastq

illumina_insert_size=1000
illumina_insert_size_d=50

if [ ! -e illumina.frg ]; then
  echo "creating illumina.frg"
  cmd="fastqToCA -libraryname illumina -technology illumina -type sanger \\
    -insertsize $illumina_insert_size $illumina_insert_size_d \\
    -innie -mates $illumina_fastq_R1,$illumina_fastq_R2"
  $cmd"
echo $cmd
$cmd > illumina.frg.tmp
mv illumina.frg.tmp illumina.frg
echo "illumina.frg created"
fi

# Run PBcR
PBcR -length 500 -partitions 200 -l Y22-3.pacbio.illumina -s pacbio.illumina.spec -fastq $pacbio_fastq genomeSize=12156677 illumina.frg

HGAP
if [ ! -e hgap3 ]; then
  mkdir hgap3
fi

TMP=/home/GLBRCORG/smcilwain/tmp
SHARED_DIR=/home/GLBRCORG/smcilwain/shared_dir
echo "Running smrtpipe.py"

smrtpipe.py -DTMP=$TMP -DSHARED_DIR=$SHARED_DIR --params hgap3.xml --output=hgap3 xml:my_inputs.xml >& hgap3.log

hgap3.xml:

<?xml version="1.0"?>
<smrtpipeSettings>
  <global>
    <param name="version">
      <value>3</value>
    </param>
    <param name="fetch">
      <value>common/protocols/preprocessing/Fetch.1.xml</value>
    </param>
    <param name="state">
      <value>active</value>
    </param>
    <param name="Control Filtering">
      <value>common/protocols/control/KeepControlReads.1.xml</value>
    </param>
    <param name="assembly">
      <value>common/protocols/assembler/PreAssemblerHGAP.3.xml</value>
    </param>
    <param name="consensus">
      <value>common/protocols/consensus/AssemblyPolishing.1.xml</value>
    </param>
  </global>
</smrtpipeSettings>
(BETA) HGAP version 3. PacBio de novo assembler optimized for speed.

common/protocols/filtering/PreAssemblerSFilter.1.xml

common/protocols/mapping/BLASR.1.xml

reference

common/protocols/referenceuploader/ReferenceUploaderUnitig.1.xml

true

RS_HGAP_Assembly_3

Sets up inputs

Filter reads for use in the pre-assembly step of HGAP, the hierarchical genome assembly process.

100
<param name="readScore">
  <value>0.80</value>
</param>
</module>

<module name="P_FilterReports">
  <param name="moduleName">
    <value>P_FilterReports</value>
  </param>
</module>

<module name="P_PreAssemblerDagcon">
  <param name="minLongReadLength">
    <value>1000</value>
  </param>
  <param name="totalBestn">
    <value>24</value>
  </param>
  <param name="minCorCov">
    <value>6</value>
  </param>
  <param name="moduleName">
    <value>P_PreAssemblerDagcon</value>
  </param>
  <param name="splitBestn">
    <value>10</value>
  </param>
  <param name="targetChunks">
    <value>6</value>
  </param>
  <param name="computeLengthCutoff">
    <value>True</value>
  </param>
  <param name="blasrOpts">
    <value>-noSplitSubreads -minReadLength 200 -maxScore -1000 -maxLCPLength 16</value>
  </param>
  <param name="title">
    <value>Using DAG-based consensus algorithm, pre-assemble long reads as the first step of the Hierarchical Genome Assembly process (HGAP). Version 2 is a stepping stone for scaling to much larger genomes.</value>
  </param>
</module>

<module name="P_AssembleUnitig">
  <param name="merSize">
    <value>14</value>
  </param>
</module>
<param name="defaultFrgMinLen">
  <value>500</value>
</param>

<param name="moduleName">
  <value>P_AssembleUnitig</value>
</param>

<param name="ovlMinLen">
  <value>40</value>
</param>

<param name="libraryName">
  <value>pacbioReads</value>
</param>

<param name="specTmpl">
  <value>analysis/etc/celeraAssembler/unitig.spec</value>
</param>

<param name="description">
  <value>This module runs Celera Assembler v8.1 to the unitig step, then finishes with our custom unitig consensus caller</value>
</param>

<param name="genomeSize">
  <value>12156677</value>
</param>

<param name="ovlErrorRate">
  <value>0.06</value>
</param>

<param name="xCoverage">
  <value>25</value>
</param>

<param name="maxSlotPerc">
  <value>1</value>
</param>

</module>

<module name="P_ReferenceUploader">
  <param name="moduleName">
    <value>P_ReferenceUploader</value>
  </param>

  <param name="runUploaderUnitig">
    <value>True</value>
  </param>

  <param name="runUploaderHgap">
    <value>False</value>
  </param>

  <param name="sawriter">
    <value>sawriter -blt 8 -welter</value>
  </param>

</module>

<param name="runUploader">
<value>False</value>
</param>
<param name="samIdx">
  <value>samtools faidx</value>
</param>
<param name="name">
  <value>reference</value>
</param>
</module>
<module name="P_Mapping">
  <param name="maxHits">
    <value>10</value>
  </param>
  <param name="maxDivergence">
    <value>30</value>
  </param>
  <param name="placeRepeatsRandomly">
    <value>True</value>
  </param>
  <param name="pulseMetrics">
    <value>DeletionQV,IPD,InsertionQV,PulseWidth,QualityValue,MergeQV,SubstitutionQV,DeletionTag</value>
  </param>
  <param name="pbalign_opts">
    <value>>--seed=1 --minAccuracy=0.75 --minLength=50 --algorithmOptions="-useQuality"</value>
  </param>
  <param name="samBam">
    <value>True</value>
  </param>
  <param name="moduleName">
    <value>P_Mapping</value>
  </param>
  <param name="minAnchorSize">
    <value>12</value>
  </param>
  <param name="description">
    <value>BLASR maps reads to genomes by finding the highest scoring local alignment or set of local alignments between the read and the genome. The first set of alignments is found by querying an index of the reference genome, and then refining until only high scoring alignments are retained. Additional pulse metrics are loaded into the resulting cmp.h5 file to enable downstream use of the Quiver algorithm.</value>
  </param>
  <param name="loadPulsesOpts">
    <value>bymetric</value>
  </param>
</module>
<param>
<param name="gff2Bed">
<value>True</value>
</param>
</module>
<module name="P_MappingReports">
<param name="moduleName">
<value>P_MappingReports</value>
</param>
</module>
<module name="P_AssemblyPolishing">
<param name="moduleName">
<value>P_AssemblyPolishing</value>
</param>
<param name="description">
<value>Polish a pure-PacBio assembly for maximum accuracy using the Quiver algorithm.</value>
</param>
<param name="enableMapQVFilter">
<value>True</value>
</param>
</module>
</smrtpipeSettings>

PBJelly
#!/bin/bash -el

fasta="contigs.fa"
source $HOME/Programs/PBSuite_14.9.9/setup.sh

if [ ! -e reference ]; then
  mkdir reference
fi

sed s/\./_/ $fasta > ./reference/reference.fasta

echo "setup"
Jelly.py setup Protocol.xml

echo "mapping"
Jelly.py mapping Protocol.xml

echo "support"
Jelly.py support Protocol.xml
```
Protocol.xml
<jellyProtocol>
  <reference>./reference/reference.fasta</reference>
  <outputDir>pbjelly_out/</outputDir>
  <blasr>-minMatch 8 -minPctIdentity 70 -bestn 1 -nCandidates 20 -maxScore -500 -nproc 4 -noSplitSubreads</blasr>
  <input baseDir="GLBRC-Data/Yeast/Y22-3/PacBio/">
    <job>Y22-3.pacbio.filtered_subreads.fastq</job>
  </input>
</jellyProtocol>

Velvet
#!/bin/bash -el

R1=GLBRCY22-3.R1.trimmed.fastq
R2=GLBRCY22-3.R2.trimmed.fastq

velveth=velveth
velvetg=velvetg

hash_list="45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75 77 79 81 83 85 87 89 91 93 95 97 99"

if [ ! -e results ]; then
  mkdir results
fi

for hash in $hash_list; do
  out_dir="results.$hash/
  if [ ! -e velveth.$hash.run ]; then
    cmd="$velveth $out_dir $hash -fastq -shortPaired -separate $R1 $R2"
    echo $cmd
    $cmd >& velveth.$hash.log
    touch velveth.$hash.run
```
contig="/results/contigs.h$hash.fa"
if [ ! -e $contig ]; then
  cmd="$velvetg $out_dir -cov_cutoff auto -exp_cov auto"
  echo $cmd
  $cmd >& velvetg.h$hash.log
  cp $out_dir/contigs.fa $contig
fi
done
echo "Done!"

**Velvet-Bless**

#Need untrimmed reads
R1=GLBRCY22-3_R1.fastq
R2=GLBRCY22-3_R2.fastq

mkdir k31
bless -read1 $R1 -read2 $R2 -prefix k31/Y22-3.bless -kmerlength 31

#Run velvet as above using
R1=./k31/Y22-3.bless.1.corrected.fastq
R2=./k31/Y22-3.bless.2.corrected.fastq

**Velvet-Racer**

racer=./RACER

#Can be either trimmed or raw reads
R1=GLBRCY22-3.R1.trimmed.fastq
R2=GLBRCY22-3.R2.trimmed.fastq

genome_size=12156677

$racer $R1 RACER.R1.fastq $genome_size
$racer $R2 RACER.R2.fastq $genome_size
#Run velvet as above using
R1=RACER.R1.fastq
R2=RACER.R2.fastq
Mitochondrial genome sequence and structure: The size of the Y22-3 mitochondrial genome is 82897 bp, which is 2.8 kbp smaller than the S288c mitochondrial genome. Y22-3 has 24 mitochondrial tRNA annotations and a similar gene order to S288c, but it differs in the number of introns in COX1, the positions of GC clusters, and the presence of several free-standing homing endonucleases and maturases (Figure S5). Specifically, the first four exons of COX1 are combined into a unique exon in Y22-3 such that the COX1 gene contains four introns, instead of the seven found in S288c. Although not previously noted, the unusual exon configuration of Y22-3 COX1 is also present in two clinical S. cerevisiae strains, YJM1549 and YJM1332 (Wolters et al. 2015). Intron 2 of COX1 encodes a maturase, ai4β, which is similar to a maturase found in the clinical S. cerevisiae isolate YJM789 (Wei et al. 2007) and 19 Mosaic-A, four wine/European, and 15 Mosaic-B S. cerevisiae strains (Wolters et al. 2015). Although this intron and its maturase exist in only 38 out of 103 S. cerevisiae strains, ai4β also exists in Saccharomyces pastorianus W34/70 (Nakao et al. 2009).

Y22-3 contains two putative homing endonucleases not found in S288c, homologs of which have been described previously in other S. cerevisiae strains, such as CBS 2354 (= NCYC 74) and D273-10B, as RF2 and RF3, respectively (Michel 1984; Séraphin et al. 1985, 1987). Although not annotated in the recent analysis of 103 S. cerevisiae mitochondrial genomes (Wolters et al. 2015), we found RF2 in NCYC3594 and YJM1439, and we found RF3 in YJM789, YJM1078, and YJM1273, suggesting that their presence in Y22-3 is rare but not unique. Along with ORF1, which is present in
S288c, these homing endonucleases are members of the LAGLIDAG family. Y22-3
ORF1 has no premature stop codons or GC clusters shifting ORF1 frame, so it may be
active, as has been suggested for other Saccharomyces strains (Peris et al. 2015).

A gene sequence comparison, using the coding sequences of COX1, COX2,
COX3, ATP6, ATP8, ATP9, COB, and VAR1 showed that the mitochondrial genomes of
Y22-3 and S288c were surprisingly divergent with 1.23% nucleotide differences
between them, which is only slightly less than the divergences between Y22-3 and the
reference genome of the sister species, Saccharomyces paradoxus CBS 432 (2.22%),
or S. pastorianus W34/70 (6.35%). High diversity (1.07%) among 103 S. cerevisiae
mitochondrial coding sequences has recently been noted (Wolters et al. 2015). For Y22-
3 and for S. cerevisiae more broadly, the high nucleotide diversity is mainly driven by
COX2 and VAR1. High nucleotide diversity at COX2 has been suggested to be the
result of reticulate evolution (e.g. recombination, introgression, horizontal gene transfer)
mediated by ORF1, GC clusters, and AT tandem repeat regions (Peris et al. 2015).

LITERATURE CITED

Michel, F., 1984 A maturase-like coding sequence downstream of the OXI2 gene of
yeast mitochondrial DNA is interrupted by two GC clusters and a putative end-of-

Nakao, Y., T. Kanamori, T. Itoh, Y. Kodama, S. Rainieri et al., 2009 Genome Sequence
of the Lager Brewing Yeast, an Interspecies Hybrid. DNA Res 16: 115–129.


**Figure S1.** A sample region from chromosome XIV demonstrating the predicted genes (blue track), the alignment of the predicted transcripts generated with default-parameter Trinity assembly (grey zone), and the alignment of the predicted transcripts generated with the optimized Trinity assembly (green zone). Blocks, zones of both-strand coverage by predicted transcripts; arrows, zone of both-strand coverage in the optimized assembly that are too narrow (several nucleotide pairs) for block highlighting. For both Trinity assemblies, the genome sequence was not used to assist in the transcript assembly, so the de novo transcriptome provides an independent validation of the predicted protein-coding genes.
Figure S2. Flowchart of the steps taken to generate the final de novo assembly of Y22-3.
Figure S3. Dot plots for each chromosome using S228c as the reference on the x-axis and the Y22-3 assembly on the y-axis. Note that there are a handful of small inversions but no detectable translocations.
Genome Assembly

YGAP

Liftover (S288c-R64)

Corrected YGAP CDS Annotations

Check and Correct CDS

Corrected Liftover CDS Annotations

Check and Correct CDS

Compare Annotations

Integrated Annotations

Integrate

Figure S4. Y22-3 genome annotation pipeline.
Figure S5. Schematic representation of A) Y22-3 and B) S288c mitochondrial genome annotations. Mitochondrial protein-coding genes, including their introns; coding sequences (CDS); rRNAs; tRNAs; and non-coding RNAs are represented in green, yellow, red, pink, and brown, respectively. Genes with asterisks are additional coding sequences in each mitochondria. The triangle in the COX1 gene represents the fusion of exons 1 to 4 into an exon in Y22-3 that S288c lacks.
Figure S6. GenePalette depiction of novel genes and non-syntenic homologs found in the subtelomeric region of the left arm of chromosome II. \( \Psi \), pseudogene. Features syntenic with S288c are in blue, and novel genes and non-syntenic homologs with valid coding regions are in green, and pseudogenes are in red. The scale bar represents 1000 bp.
Figure S7. GenePalette depiction of novel genes and non-syntenic homologs found on chromosome IV. Ψ, pseudogene. Features syntenic with S288c are in blue, and novel genes and non-syntenic homologs with valid coding regions are in green, and pseudogenes are in red. The scale bar represents 1000 bp.
Figure S8. GenePalette depiction of novel genes and non-syntenic homologs found in the subtelomeric region of the left arm of chromosome VI. Ψ, pseudogene. Features syntenic with S288c are in blue, and novel genes and non-syntenic homologs with valid coding regions are in green, and pseudogenes are in red. The scale bar represents 1000 bp.
Figure S9. GenePalette depiction of novel genes and non-syntenic homologs found in the subtelomeric region of the right arm of chromosome VI. Ψ, pseudogene. Features syntenic with S288c are in blue, and novel genes and non-syntenic homologs with valid coding regions are in green, and pseudogenes are in red. The scale bar represents 1000 bp.
Figure S10. GenePalette depiction of novel genes and non-syntenic homologs found in the subtelomeric region of the right arm of chromosome VII. Ψ, pseudogene. Features syntenic with S288c are in blue, and novel genes and non-syntenic homologs with valid coding regions are in green, and pseudogenes are in red. The scale bar represents 1000 bp.
Figure S11. GenePalette depiction of the *CUP1* locus found on chromosome VIII. Features syntenic with the S288c reference genome are in blue, while extra copies of *CUP1* are depicted in green. The scale bar represents 1000 bp.
Figure S12. GenePalette depiction of novel genes and non-syntenic homologs found in the subtelomeric region of the right arm of chromosome VIII. Ψ, pseudogene. Features syntenic with S288c are in blue, and novel genes and non-syntenic homologs with valid coding regions are in green, and pseudogenes are in red. The scale bar represents 1000 bp.
**Figure S13.** GenePalette depiction of novel genes and non-syntenic homologs found in the subtelomeric region of the left arm of chromosome IX. Ψ, pseudogene. Features syntenic with S288c are in blue, and novel genes and non-syntenic homologs with valid coding regions are in green, and pseudogenes are in red. The scale bar represents 1000 bp.