Whole-Genome Sequence and Variant Analysis of W303, a Widely-Used Strain of Saccharomyces cerevisiae

Kinnari Matheson,* 1 Lance Parsons,† 1 and Alison Gammie*, 1, 2
*Department of Molecular Biology and †Lewis-Sigler Institute for Integrative Genomics, Princeton University, New Jersey 08544-1014
ORCID IDs: 0000-0002-7934-1937 (K.M.); 0000-0002-8521-714X (L.P.); 0000-0002-4282-1056 (A.G.)

ABSTRACT The yeast Saccharomyces cerevisiae has emerged as a superior model organism. Selection of distinct laboratory strains of S. cerevisiae with unique phenotypic properties, such as superior mating or sporulation efficiencies, has facilitated advancements in research. W303 is one such laboratory strain that is closely related to the first completely sequenced yeast strain, S288C. In this work, we provide a high-quality, annotated genome sequence for W303 for utilization in comparative analyses and genome-wide studies. Approximately 9500 variations exist between S288C and W303, affecting the protein sequences of ~700 genes. A listing of the polymorphisms and divergent genes is provided for researchers interested in identifying the genetic basis for phenotypic differences between W303 and S288C. Several divergent functional gene families were identified, including flocculation and sporulation genes, likely representing selection for desirable laboratory phenotypes. Interestingly, remnants of ancestor wine strains were found on several chromosomes. Finally, as a test of the utility of the high-quality reference genome, variant mapping revealed more accurate identification of accumulated mutations in passaged mismatch repair-defective strains.

Saccharomyces cerevisiae is a genetically tractable model organism that is used to study a multitude of biological and disease processes (Botstein et al. 1997). There are many examples of the utility of yeast in uncovering fundamental biological pathways important for human health. For example, the elucidation of the conservation between yeast and human DNA mismatch repair contributed to the discovery that mismatch repair dysfunction was the causative agent in a common hereditary cancer syndrome (Fishel et al. 1993; Strand et al. 1993; Clark et al. 1999).

As yeast emerged as an important model organism, many laboratory strains were selected to express important characteristics such as the ability to mate, sporulate, and be transformed with high efficiency. Additionally, when manipulating yeast, researchers chose progeny lacking certain phenotypes such as agar invasion, clumping, and rapid sedimentation (Louis 2016). For example, S288C, a widely used laboratory strain (Goffeau et al. 1996), possesses a nonsense mutation in the FLO8 gene, which prevents clumping and invasive growth into agar, thereby allowing cells to be fully suspended in solution (Liu et al. 1996). W303, a descendant of S288C, was selected to retain the desirable characteristics of S288C, to sporulate well, and to be transformed with high efficiency (R. Rothstein, personal communication).

Differences among laboratory strains have been well documented; for example, analyses of the proteomes of several laboratory strains reveal differentially expressed proteins across various laboratory strains (Rogowska-Wrzesinska et al. 2001). Additionally, certain alleles of the SWI-SNF global transcription activator complex contribute to slow growth in the W303 background, but are lethal in S288C (Cairns et al. 1998). Given these differences, an understanding of the precise variations at the nucleotide level between strains is an important step in elucidating the underlying causes of phenotypic differences.

Since its origin, W303 has been widely used for genetic analyses of DNA repair and other biological mechanisms (Thomas and Rothstein 1997).
Table 1 Publicly available W303 sequencing data

| Reference            | Platform                   | Coverage | Accession Number |
|----------------------|----------------------------|----------|------------------|
| Ralser et al. (2012) | Illumina and Roche-454     | 376×     | GB: ALAV01000000 |
| Song et al. (2015)   | Illumina                   | 301×     | GB: JRIU01000000 |
| Lang et al. (2013)   | Illumina                   | 300×     | SRA: SRX315098   |
| Goodwin et al. (2015)| Oxford nanopore            |          | GB: JSAC01000000 |
| This work            | PacBio                     |          | GB: LYZE00000000 |

GB, GenBank; SRA, NCBI Sequence Read Archive; PacBio, Pacific Biosciences.

**Materials and Methods**

Genomic DNA preparation and library construction

A 500 ml culture of wild-type W303 (MY7521) MATa his3-11,15 trp1-1, ura3-1 (derived from strains generously provided by Rodney Rothstein, Columbia University) was grown in rich medium for ~24 hr. Genomic DNA extraction and purification was carried out according to Burke et al. (2000). Standard sequencing library construction was performed with Pacific Biosciences (PacBio) DNA Template Prep Kit 2.0 for one SMRT cell. The final library was sent to the University of California at Irvine for > 7 kb size selection shearing and sequencing with P6C4 chemistry.

Comparison of sequence alignments and annotations were visualized with Geneious version 9.0.3 (Kearse et al. 2012).

**Comparative analysis**

MAUVE (Darling et al. 2004) whole-genome and chromosome alignments were used to analyze single nucleotide polymorphisms (SNPs) and rearrangements between W303 and S288C. MAUVE was utilized in order to identify the position of each polymorphic site in the reference and alternative genome sequence. MAUVE alignments and polymorphisms were visualized with genoPlotR (Guy et al. 2010) and Microsoft Excel.
Variants identified from the MAUVE (Darling et al. 2010) genome alignment of S288C and W303 were characterized with CooVar version 0.07 (Vergara et al. 2012) with respect to the position and coding sequences of S288C. MUSCLE (Edgar 2004; Li et al. 2015) alignments were analyzed to identify the conservation of repeat regions in Flo1 with the S288C ortholog of the protein (S288C: YAR050W). Divergent W303 orthologs (those with nonsynonymous substitutions) were analyzed with GO Slim Mapper (yeastgenome.org/cgi-bin/GO/goSlimMapper.pl) to determine whether variants mapped onto certain root biological processes. Genes that map onto the Saccharomyces Genome Database GO slim are listed in Supplemental Material, File S1. For analysis of sequence variations from S288C, megaBLAST (Zhang et al. 2000) alignments of each chromosome against the nucleotide collection were classified. The aligned sequences of the best hits (max

Figure 1 Highly similar genome structure between W303 and its ancestor, S288C. Chromosome alignments of W303 (top) and S288C (bottom) are shown. The color blocks do not signify the degree of sequence similarity, instead they represent stretches of homology without gaps or rearrangements. Scale bars are shown for reference below each alignment.
Variant analysis of mismatch repair-deficient strains

Mapping of accumulated mutations in msh2 null (SRX315139) as well as msh2 missense variants—R542L (SRX315174), G688D (SRX315176), and A618V (SRX315175)—was carried out according to previous work (Lang et al. 2013) with more stringent quality filtering. Alignments with BWA (Li and Durbin 2009) mapping quality ≥ 80 were ignored for variant detection purposes. Variants were detected using Freebayes (Garrison and Marth 2012) and filtered to include loci with depth of coverage > 10 and variant quality > 20, with the highest genotype quality of 5000.

Data availability

Strains are available upon request. The W303 sequences from this work are available at GenBank, accession number LYZE00000000. File S1 contains the characterized substitutions based on genome alignment of S288C and W303. File S2 contains the variant calling analysis with the improved W303 reference genome.

RESULTS AND DISCUSSION

Alignment of S288C and W303 shows high similarity between the strains

The high-quality, chromosomally organized, annotated genome of the yeast strain W303 presented in this work was created by: (1) assembling long, lower fidelity reads (PacBio) into 47 contigs; (2) generating chromosome/episome length sequences using publicly available W303 data and S288C as scaffolds; and (3) error-correcting the assembled genome using short, high-fidelity reads. The complete W303 genome statistics are shown in Table 2. The genome is made up of 18 scaffolds that represent the 16 chromosomes, the mitochondrial genome, and the 2 µm plasmid.

To analyze the divergence of W303 from its parent strain, S288C, the genomes were aligned using MAUVE (Darling et al. 2010). Figure 1 shows the collinear blocks of homology between the strains. In Figure 1, each colored segment represents a distinct region of DNA that shares homology without gaps or rearrangement. Despite some telomeric rearrangements, S288C and W303 are highly similar in genomic structure and sequence identity. The alignment of chromosome IX exhibits high synteny with S288C and shows only one breakpoint between the collinear homologous regions of the chromosome (Figure 1).

In contrast, chromosome XVI shows rearrangement near a terminal region of the chromosome. A transposable element and a Y′-encoded ATP helicase flank the junction of this region. This finding is not surprising because both transposable elements (Mieczkowski et al. 2006) and Y′-helicases, thought to have originated as mobile elements, are associated with chromosomal rearrangement and recombination (Louis and Haber 1992; Schmidt and Kolodner 2006).

The divergence in telomeric regions includes changes beyond the large rearrangement discussed above. A comparison of the gene content between S288C and the annotation of W303 shows expansion of Y′ element ATP-dependent helicase protein throughout the genome, including the acquisition of Y′ regions on chromosomes without these subtelomeric elements in S288C. These differences were identified on the right arm of chromosomes III and XIV (Louis et al. 1994; Louis 1995). Previous work demonstrated that subtelomeric elements undergo recombination and expansion in telomerase-deficient strains in order to restore telomeres (Lundblad and Blackburn 1993), and that the presence of these Y′ helicases varies between related strains of S. cerevisiae on homologous chromosomes (Chan and Tye 1983).

Although the chromosome structure (Figure 1) and lengths (Figure 2A) of W303 and S288C are similar, there are 9500 single nucleotide variations (Figure 2A and File S1). Figure 2B shows that at the nucleotide level, some chromosomes are more homologous to S288C (blue), while others are more divergent (gray). Overall, chromosome XI is the most distinct from its respective chromosome in S288C (Figure 2B). This observation prompted further analysis of the divergence or ancestry at the chromosome level.
While S288C and W303 are highly similar, each chromosome was analyzed to identify regions that may be divergent. After performing megablast BLASTn (Zhang et al. 2000) alignments against the nucleotide collection, W303 chromosomes III and XI were found to share significant sequence identity to the respective chromosomes in strains: EC1118, max score = 2.018e + 05, E value = 0.0 (Novo et al. 2009) and YJM244, max score = 2.461e + 05, E value = 0.0 (Strope et al. 2015). Interestingly, both are wine fermentation strains with European ancestry. These regions of similarity with the wine strains include continuous regions of chromosome III and XI. In contrast, when these same regions in W303 were aligned with S288C, the output showed shorter segments of homology with multiple gaps.

Phylogenetic analysis of several S. cerevisiae strains from various populations confirms the close relationship between S288C and W303 genome-wide (Figure 3A). Interestingly, S288C and W303 branch in a clade with the commercial wine strain EC1118 mentioned above (Figure 3A). This sequence similarity may reflect a shared wine strain ancestry among these three strains. To determine whether W303 has distinct wine strain ancestry, chromosome alignments of the S. cerevisiae strains described above were conducted to identify polymorphisms among the strains. The analysis revealed identical polymorphic sites shared among the wine strains EC1118 and YJM244, and the laboratory strains S288C and W303, on chromosomes III and XI (Figure 3B, COMMON). Interestingly, on W303 chromosome XI, which is the

**Figure 3** Phylogenetic analyses reveal potential remnants of wine strain ancestry. (A) Phylogeny of whole genomes of S. cerevisiae strains from various populations. A key of the populations associated with each strain are given in the upper right rectangle. (B) Identified polymorphisms across yeast species. The polymorphisms were identified using MAUVE chromosomal alignments with the strains shown in (A). Sites with common nucleotides only in the EC1118 and YJM244 wine strains, W303, and S288C are shown as points in orange (COMMON), while blue points represent potential sites of divergence from S288C where sites are only identical across the EC1118 and YJM244 wine strains and W303 (DIVERGENT).

**Figure 4** Divergent coding sequences in W303 when compared to S288C. Divergent regions of the protein sequence of Flo1 are highlighted in the alignment with other variants of the protein residues (S288C: YAR050W). Protein domains are shown above the alignment, PA14: pink, flocculin: yellow. Expansions of the flocculin repeats in W303 are shown.
most divergent from S288C (Figure 2B), there are many polymorphic sites that are distinct from S288C, but identical to ones in the wine strains EC1118 and YJM244 (Figure 3B, DIVERGENT).

W303 ancestors include D311-3A and D190-9C, strains with unknown ancestry (R. Rothstein, personal communication and Rogowska-Wrzesinska et al. 2001). The data presented in this paper suggest that these strains might also have European wine ancestry. Further sequencing of laboratory strains in the pedigree of W303 would allow for the characterization of the source of the divergence of W303 from S288C.

Divergent coding sequences of W303 compared to S288C

To analyze potential functional consequences of the differences between W303 and S288C, synonymous and as well as the conservative and nonconservative nonsynonymous substitutions were characterized using Coovar version 0.07 (Vergara et al. 2012). The analysis was based on the variants identified from MAUVE (Darling et al. 2010) genome alignment between S288C and W303. The results are provided as a comprehensive listing of the genomic variation between the strains that may be a useful tool for researchers interested in understanding the genetic basis of phenotypic differences (File S1).

Because nonsynonymous substitutions have the capacity to have biological consequences, the complete list of highly divergent genes with the number of conservative and nonconservative nonsynonymous substitutions is supplied in File S1. The variants with nonsynonymous changes were mapped to Gene Ontology (GO) terms. There was not a significant enrichment in any category for the entire group of ~700 genes with nonsynonymous differences or with the ~220 genes with nonconservative substitutions (File S1).

Although there was not enrichment in a specific functional category, certain genes were strikingly divergent; for example, YHL008C, an uncharacterized open reading frame, sustained substantially more nonsynonymous substitutions than the any other gene. YHL008C has 83 nonsynonymous substitutions (11 of which are nonconservative) over the 1884 nt open reading frame. Little is known about the function; however, deletion of this open reading frame decreases chloride accumulation (Jennings and Cui 2008). Early yeast transformation procedures often employed calcium chloride to increase transformation efficiency. As mentioned previously, W303 was selected to have superior transformation efficiency over S288C. Variations in YHL008C and the other 42 coding sequences involved in ion transport (GO:0006811, File S1) might be associated with the selection of spores with high transformation efficiency during crosses that gave rise to W303.

The second gene with the most nonconservative, nonsynonymous substitutions is AAD4, an aryl alcohol dehydrogenase (AAD). The W303 AAD4 gene has 48 variants (nine conservative and nine nonconservative substitutions) in the 990 nt open reading frame. Variability in AADs has been associated with wine and other fermentation strains (Borneman et al. 2011). The AAD enzymes convert aldehydes and ketones into their corresponding aromatic alcohols. As such, the variability of AAD genes in different fermentation yeast strains is thought to influence the volatile aromas produced during wine fermentation, and aroma characteristics are an important component of wine quality (Li et al. 2014).

With an understanding of the history of W303, we examined certain other processes that had been selected for during the crosses to create the strain. As mentioned above, W303 was selected to have a higher sporulation efficiency than S288C (Gerke et al. 2006) (R. Rothstein, personal communication). Interestingly, differences in 19 of 176 sporulation genes (GO:0043934) (Hong et al. 2008) were identified.

Similarly, selection against flocculation in ancestral laboratory strains likely gave rise to lessened selective pressure of these genes. As mentioned above, S288C harbors an inactivating point mutation in FLO8, whose gene product is a transcription factor required for flocculation and invasive growth (Liu et al. 1996). In W303, there are 13 aa differences in the 2277 nt open reading frame for Mss11, a protein that
coregulates cell wall genes with Flo8 (Bester et al. 2012). Additionally, another flocculation gene, Flo9, harbors 12 nonsynonymous substitutions (two nonconservative) in an open reading frame of 3969 nt. Finally, the W303 Flo1 gene maintains expansions in the flocculin repeat region (Figure 4) that directly correlate with adhesion phenotypes (Verstrepen et al. 2005). SPSC01, a constitutively flocculent strain, contains an expansion of these domains in Flo1 (He et al. 2012). The variation in this region may be due to instability at these repetitive regions, or reflect a more flocculent ancestor of W303. Taken together, these divergences might be a consequence of changes that occur in the continuous laboratory selection against the flocculation function.

Although only a few observations are cited above, the analysis of the polymorphisms identified from alignment of S288C and W303 should serve as a tool to begin to understand the mechanisms underlying phenotypic variations between the strains.

**Improvement of mapping of mutation accumulation in a mismatch repair-defective strain**

The assembled genome sequence of W303 described in this work was employed to validate the efficacy in accurate mutation calling. Previously, we conducted a mutation accumulation analysis with a lower quality S288C SNP-adjusted draft genome that required the manual verification of all called single base substitutions and indels at repetitive elements (Lang et al. 2013). By manual verification, we refer to final steps in the SNP calling pipeline to eliminate false positives. The process includes filtering out commonly called false positives and then visualizing the aligned sequencing reads of the passed strains along with the ancestors using genome viewing software to verify the fixed mutations in the passed mutator strains (Lang et al. 2013). In the previous analysis, the identification of insertion/deletion mutations required less stringent SNP calling parameters; however, while capturing the mutations, the less stringent SNP calling output resulted in a large number of false positives. We reasoned that high-throughput mapping of mutations, particularly insertion/deletions, should be more accurate and require less manual verification with a higher quality reference genome. To test this, DNA reads from serially passaged mismatch repair-defective strains (Lang et al. 2013) were mapped onto the S288C SNP-adjusted W303 draft genome (Lang et al. 2013) and to the high-quality W303 genome presented in this work. The SNP calling parameters were similar to those used previously with minor modifications (described in the Materials and Methods).

As anticipated, an improvement in the number of calls was observed with the high-quality W303 genome in contrast to the SNP-adjusted S288C draft genome (Figure S5). For example, with the null msh2 passaged strain, mapping onto the current high-quality W303 assembly increased insertion or deletion calls from 422 to 248 and the number of SNP calls decreased from 138 to 44. The in all cases, the number of SNP variants called with the high-quality genome was closer to the actual number of mutations verified manually (Figure 5 and File S2). Importantly, the mutations identified in mismatch repair-defective strains using the high-quality W303 assembly recapitulate the increased identification of insertions and deletions, without creating a large number of false positives (Figure 5). In conclusion, these data represent an improvement on the S288C SNP-adjusted draft W303 genome and can be employed for analysis of the ancestry, variant detection, and other genome-wide studies.

**ACKNOWLEDGMENTS**

The authors would like to acknowledge Wei Wang, Jessica Wiggins, and the Princeton Sequencing Core Facility for their support of this project. We would also like to acknowledge Mark Rose for his support of this work. The research that is the basis of this manuscript was conducted at Princeton University. This research was supported by a New Jersey Commission on Cancer Research Seed grant (10-1064-CCR-E0) and a Cancer Institute (NCI) of New Jersey NCI Cancer Center Support grant (P30 CA072720). K.M. was partially supported by a T32 GM07388 training grant. Some equipment and supplies were purchased through R01 GM037739. The Princeton Sequencing Core Facility was supported in part by the National Institute of General Medical Sciences (P50 GM071508). The authors declare that they have no competing interests.

Author contributions: K.M. and A.G. conceived the study. L.P. assembled the genome sequence. K.M. carried out all subsequent analysis. K.M. and A.G. wrote the manuscript. All three authors read and approved the final manuscript.
