G3 Genes | Genomes | Genetics

GENOME REPORT

Insight into the Recent Genome Duplication of the Halophilic Yeast *Hortaea werneckii*: Combining an Improved Genome with Gene Expression and Chromatin Structure

Sunita Sinha,* Stephane Flibotte,* Mauricio Neira,* Sean Formby,* Ana Plemenitaš,† Nina Gunde Cimerman,‡ Metka Lenassi,† Cene Gostinčar,‡ Jason E. Stajich,§,1 and Corey Nislow*,1

*Faculty of Pharmaceutical Sciences, University of British Columbia, Vancouver V6T 1Z3, Canada, †Institute of Biochemistry, Faculty of Medicine, and ‡Department of Biology, Biotechnical Faculty, University of Ljubljana, 1000, Slovenia, and §Department of Plant Pathology and Microbiology, Institute for Integrative Genome Biology, University of California, Riverside, California 92521

ORCID IDs: 0000-0002-7591-0020 (J.E.S.); 0000-0002-4016-8874 (C.N.)

ABSTRACT Extremophilic organisms demonstrate the flexibility and adaptability of basic biological processes by highlighting how cell physiology adapts to environmental extremes. Few eukaryotic extremophiles have been well studied and only a small number are amenable to laboratory cultivation and manipulation. A detailed characterization of the genome architecture of such organisms is important to illuminate how they adapt to environmental stresses. One excellent example of a fungal extremophile is the halophile *Hortaea werneckii* (Pezizomycotina, Dothideomycetes, Capnodiales), a yeast-like fungus able to thrive at near-saturating concentrations of sodium chloride and which is also tolerant to both UV irradiation and desiccation. Given its unique lifestyle and its remarkably recent whole genome duplication, *H. werneckii* provides opportunities for testing the role of genome duplications and adaptability to extreme environments. We previously assembled the genome of *H. werneckii* using short-read sequencing technology and found a remarkable degree of gene duplication. Technology limitations, however, precluded high-confidence annotation of the entire genome. We therefore revisited the *H. wernickii* genome using long-read, single-molecule sequencing and provide an improved genome assembly which, combined with transcriptome and nucleosome analysis, provides a useful resource for fungal halophile genomics. Remarkably, the ~50 Mb *H. wernickii* genome contains 15,974 genes of which 95% (7608) are duplicates formed by a recent whole genome duplication (WGD), with an average of 5% protein sequence divergence between them. We found that the WGD is extraordinarily recent, and compared to *Saccharomyces cerevisiae*, the majority of the genome's ohnologs have not diverged at the level of gene expression of chromatin structure.

KEYWORDS
extremophilic yeast
gene duplication
halophile
salt tolerance
Genome Report

The study of prokaryotes able to thrive in extreme environments has led to fundamental discoveries like the archaea (Woese and Fox 1977) and the development of breakthrough technologies (polymerase chain reaction) (Henry and Debarbieux 2012; Jia *et al.* 2013; Oren 2014). In contrast, eukaryotic extremophiles in general, and fungal extremophiles in particular, remain largely unexplored. A detailed characterization of such extremophiles is important because it can advance our understanding of how organisms with a fundamentally different cellular organization and evolutionary background compared to prokaryotic extremophiles adapt to environmental stresses. Furthermore, such insights have the potential to expand the stress tolerance of industrially relevant organisms (Gostinčar *et al.* 2011). Halotolerance is a phenotype of particular interest since it can inform modifications of fungi used in biotechnology and bioremediation, and of plant crops whose growth is compromised by the widespread salinization of agricultural land (Oren 2010; Lentzen and Schwarz 2006; Shabala *et al.* 2016).

One eukaryotic extremophile to direct attention to is the ascomycetous yeast *Hortaea werneckii* (Pezizomycotina, Dothideomycetes, Capnodiales), which is exceptionally adaptable to osmotic stress and can tolerate extracellular salt concentrations that either kill or inhibit

the growth of most microorganisms (Gostinčar *et al.* 2011; Plemenitaš *et al.* 2014). It is commonly found in brines formed by the evaporation of sea water, where it thrives despite fluctuating salt concentrations (ranging from no salt to near-saturating concentrations), low water activity, high temperature, high UV radiation, varying nutrient availability, and near-alkaline pH (Gostinčar *et al.* 2011).

Several studies have identified individual mechanisms that contribute to *H. werneckii*'s extreme halotolerance (reviewed in Gostinčar *et al.* 2011 and Plemenitaš *et al.* 2014). These studies found that general strategies including cell wall melanization, changes in membrane composition, and the accumulation of glycerol (and other compatible solutes) promote resistance to salt stress. Specific adaptations include rapid changes in the expression of genes involved in salt sensing (Vaupotic and Plemenitaš 2007), and resistance to high temperature and oxidative stress (reviewed in Gostinčar *et al.* 2011 and Plemenitaš *et al.* 2014). Finally, as seen in other organisms (Gerstein *et al.* 2006; Schoustra *et al.* 2007; Dhar *et al.* 2011), *H. werneckii* has expanded its genome and contains multiple copies of halotolerance genes (Lenassi *et al.* 2013).

Interestingly, in *H. werneckii*, the mating-type locus also exists in two copies and contains the idiomorph MAT1-1. The *HwMAT1-1-1* gene sequences are 88.7% identical and have homologous 5′ and 3′ flanking regions (Lenassi *et al.* 2013). A compatible idiomorph MAT1-2 was not found in the sequenced *H. werneckii* genome, indicating that this species is heterothallic and would require a partner strain encoding *MAT1-2-1* for sexual reproduction. *MAT1-2-1* strains have not yet been identified and neither has the existence of a *H. werneckii* teleomorph. The sexual cycle of *H. werneckii* is also unknown.

As a first step toward a comprehensive understanding of halotolerance in *H. werneckii*, we previously sequenced the genome of strain EXF-2000 isolated from marine solar salterns in Slovenia (Lenassi *et al.* 2013). The short length of next generation sequencing (NGS) reads available at the time, combined with the unexpectedly large number of duplicated sequences, complicated the production of a high-quality genome assembly. Nonetheless, we confirmed that fungal ion transporters were present in multiple copies, and a close inspection of the sequence data led us to hypothesize that *H. werneckii* had recently duplicated its entire genome (Lenassi *et al.* 2013). If confirmed, this genome duplication offers a "natural experiment" in which the early stages of functional gene diversification (*e.g.*, neofunctionalization) can be interrogated.

Here we leveraged improvements in NGS to resequence the genome of EXF-2000 using long-read, single-molecule technology. Our new genome assembly reduced the number of contigs 20-fold from 12,620 to 651. Analysis and annotation of our improved genome (Hw 2.0) confirmed our previous hypothesis of a recent whole genome duplication (WGD) and further demonstrated a high degree of sequence conservation between each ohnolog pair. This greatly improved genome assembly allowed us to unambiguously compare gene pairs and closely examine the evidence for duplicate gene divergence at the level of primary sequence, gene expression, and chromatin architecture. We then compared the results from *H. werneckii* to those obtained for *Saccharomyces cerevisiae*, a species with a relatively ancient genome duplication, to compare the spectrum of genome reduction following duplication. In addition to providing a baseline for exploring genomic adaptation to extreme environmental challenge, our data set provides a framework in which to address fundamental questions of gene functionalization and divergence at the single gene, gene pair, and whole genome level. By way of example, we combined RNA-seq and nucleosome profiling to demonstrate the high degree of similarity between each ohnolog gene pair.

MATERIALS AND METHODS

Strain information and growth conditions

H. werneckii strain EXF-2000 was isolated from marine solar salterns on the Adriatic coast in Slovenia. It is archived in the Ex Culture Collection of the Department of Biology, Biotechnical Faculty, University of Ljubljana (Infrastructural Centre Mycosmo, MRIC UL). Cells were grown at 28° in synthetic defined yeast nitrogen base (YNB) liquid medium (Sigma Aldrich), supplemented with 1.7 M NaCl and adjusted to pH 7.0.

Long-read sequencing

Genomic DNA (gDNA) for PacBio sequencing was extracted using the Gentra Puregene Yeast/Bact. Kit (Qiagen) from two independent 1 ml overnight cultures of *H. werneckii* grown in YNB containing 1.7 M NaCl, taking care to avoid manipulations that could shear the gDNA. DNA samples were quantified, pooled, and sequenced using five SMRT cells with P4/C2 chemistry at the University of Washington PacBio Sequencing Facility.

RNA-seq library preparation and sequencing

H. werneckii cells were grown to midexponential phase, and 50 ml of culture was pelleted, flash frozen in liquid nitrogen, and stored at −80° until RNA isolation. Pelleted cells were ground in a mortar and pestle maintained in a liquid nitrogen bath and the grindate used for total RNA isolation using TRIzol Reagent according to the manufacturer's instructions (Thermo Fisher Scientific). Two independent RNA samples were converted to sequencing libraries using Illumina's TruSeq RNA version 2 Library Preparation Kit, and sequenced on a HiSequation 2500 or MiSeq (Illumina) to generate paired-end 100 base reads.

MNase-seq library preparation and sequencing

H. werneckii cells were grown to midexponential phase, and 200 ml of culture was used for nucleosome-bound gDNA preparation, following the protocol of Tsui *et al.* (2012) but omitting the gel extraction step as described by Henikoff *et al.* (2011). DNA was digested with titrations (12.5–50 U) of micrococcal nuclease (Fermentas), selecting for library preparation samples that showed >80% mononucleosomes based on agarose gel analysis. Two independent DNA samples were converted to sequencing libraries using NEBNext Ultra DNA Library Prep Kit for Illumina (New England Biolabs), selecting for 150 bp inserts with Agencourt AMPure XP Beads (Beckman Coulter). Libraries were sequenced on a HiSequation 2500 (Illumina), generating paired 100 base reads.

doi: https://doi.org/10.1534/g3.117.040691
Manuscript received February 19, 2017; accepted for publication May 8, 2017; published Early Online May 12, 2017.

Supplemental material is available online at www.g3journal.org/lookup/suppl/doi:10.1534/g3.117.040691/-/DC1.
[1]Corresponding authors: Department of Plant Pathology and Microbiology, Institute for Integrative Genome Biology, University of California, Riverside, 1207K Genomics Building, 900 University Ave., Riverside, CA 92521. E-mail: jason.stajich@ucr.edu; and Department of Pharmaceutical Sciences, University of British Columbia, 2405 Wesbrook Mall, Room 6619, Vancouver BC V6T 1Z3, Canada. E-mail: corey.nislow@ubc.ca

Genome assembly and annotation

Sequencing reads were *de novo* assembled using the SMRT Analysis version 2.3 suite and RS_HGAP_assembly.3 pipeline provided by Pacific Biosciences, using default parameters. Genome assembly quality was assessed with CEGMA (Parra *et al.* 2007). Protein-coding and tRNA genes were annotated with MAKER (version 2.31.8) (Holt and Yandell 2011), using as evidence for gene predictions a set of Ascomycete fungal proteomes, the complete UniProt SwissProt database, and the Genome Guided (GG) Trinity transcripts obtained from our RNA-seq data. Gene prediction training of the SNAP (Korf 2004) and Augustus (Stanke *et al.* 2006) gene callers was performed initially on protein-to-genome alignments of a conserved set of proteins in CEGMA (Parra *et al.* 2007). Paired RNA-seq reads were quality trimmed with sickle (Joshi and Fass 2011) and assembled into a consensus of transcripts with Trinity. Trinity was run with GG mode, which required the RNA-seq reads to be aligned to the genome. We used GSNAP (Wu and Nacu 2010) to align the reads and produce a BAM file, followed by running Trinity in GG mode with jaccard clipping enabled to improve fungal transcript calling and a maximum intron length of 1500 nt. The assembled consensus transcripts were used as transcript evidence for MAKER genome annotation. MAKER annotation was further processed with funannotate: Fungal genome annotation scripts (Palmer 2017) (https://github.com/nextgenusfs/funannotate) for cleanup of gene models and passed to Genome Annotation Generator (GAG) (Hall *et al.* 2014) (https://genomeannotation.github.io) for creation of Genbank submission files.

Bioinformatics analysis of genome duplication

To globally investigate the duplication of genomic sequences in our new assembly Hw 2.0, the genome was aligned to itself with the PROmer algorithm, as implemented in MUMmer 3.23, and plotted with the mummerplot utility (Kurtz *et al.* 2004). The pairs of most similar contigs were divided into two lists, resulting in two nonoverlapping sets of contigs, each set roughly representing half of the duplicated genome (a haploid genome). To compare protein paralogs in the genome, duplicated genes were identified from the published Hw 1.0 and the improved Hw 2.0 genomes as in Lenassi *et al.* (2013). An all-against-all protein sequence similarity search of *H. werneckii* proteins to a *H. werneckii* protein database was performed by BLASTP, part of BLAST 2.2.28+ (Altschul *et al.* 1997) using an *e*-value cutoff of $1e^{-50}$. Additionally, predicted proteins were aligned back to the genomic sequence with Exonerate version 2.2.0 using the protein2genome model (Slater and Birney 2005), and the number of loci to which each protein aligned with at least 75% of the maximal score was counted. To assess sequence conservation, duplicated proteins were identified by BLASTP using a cutoff of $1e^{-50}$. Each protein was allowed to be in only one pair (the one with the lowest *e*-value), which resulted in 7608 protein pairs. The two proteins in each pair were aligned to each other with MAFFT software in the "–auto" mode (Katoh and Toh 2008a,b). The alignment was split into 20 sectors of equal length, and the number of identical and mismatched amino acids for each was counted. The total number of amino acid substitutions within each protein pair was calculated by summing the substitutions in all 20 alignment sectors. For analysis of upstream regions, protein pairs were determined by BLASTP and a cutoff of $1e^{-90}$. For each duplicated protein pair, regions extending from 1 kb upstream and the first 100 bp of each CDS were aligned with MAFFT software in the "–auto" mode (Katoh and Toh 2008a,b). All pairs without a start codon at the expected location, or with <75% of identical nucleotides in the region between duplicated proteins were discarded. This resulted in 4441 pairs of genes, for which the conserved nucleotide positions were counted within the pair, summed, and expressed as the proportion of conserved nucleotides per position within the whole data set. Synonymous site divergence was calculated from these pairs by aligning the protein sequences with Muscle (Edgar 2004), followed by a back translation to coding sequences using bp_mrtrans (Stajich *et al.* 2002), and substitutions calculated with YN00 (Yang and Nielsen 2000) reimplemented for multi-sequence processing in the package subopt-kaks (https://github.com/hyphaltip/subopt-kaks). Substitutions per synonymous site (dS) values above dS = 2 were removed as they are unlikely to be useful estimates and represent the limit of estimation.

RNA-seq analysis

After alignment to the Hw 2.0 assembly, the number of aligned RNA-seq reads for each gene was tabulated directly with the alignment program STAR (Dobin *et al.* 2013). The same procedure was also applied to RNA-seq reads for *S. cerevisiae* available at the NCBI Short Read Archive (accession numbers SRR488142 and SRR488143) using the reference genome S288C version R64 [*Saccharomyces* Genome Database (SDG); http://www.yeastgenome.org].

MNAse-seq analysis

H. werneckii sequencing reads were aligned to the Hw 2.0 genome using BWA version 0.7.12 (Li and Durbin 2009) and the resulting files sorted with the SAMtools suite version 1.2 (Li *et al.* 2009). DANPOS2 (version 2.2.2) was then used on the alignments to extract nucleosome positioning and occupancy (Chen *et al.* 2013). The same analysis procedure was applied to publicly available reads for *S. cerevisiae* using the reference genome S288C version R64 (SDG; http://www.yeastgenome.org).

Analysis of divergence in gene regulation

The difference in expression level between paralog pairs was calculated with custom Perl and R scripts. Briefly, for each sample, gene paralog pairs were kept for further analysis if both members were covered by at least 50 aligned reads. An expression ratio in the $\log_2$ scale was then calculated for each pair, taking into account the coding size of the individual genes. Those ratios were then averaged between the two samples for each species and finally the absolute value of the resulting fold changes was used in the calculation of the distribution for each species. In addition, for each gene the output from DANPOS2 provided an average nucleosome occupancy value calculated every 10 bases for a 1 kb window centered on the transcription start site. Those nucleosome occupancy profiles were then normalized to an average of one. In other words, only the shape of the profile and not the absolute occupancy level was kept for the downstream analysis. The similarity between paralog pairs was then assessed by calculating the root mean square between the individual normalized profiles (analogous to a SD or the goodness of fit in a χ^2 minimization).

Data availability

Sequencing reads have been deposited in Genbank under the BioProject PRJNA356640. Scripts and data are available from https://github.com/stajichlab/Hortaea_werneckii.

RESULTS

The Hortaea werneckii genome assembly version 2.0

Our previously published genome sequence (Hw 1.0) used short 75 bp Illumina reads, which limited our assembly and ability to analyze *H. werneckii*'s genomic content (Lenassi *et al.* 2013). Although the data

provided a case for the highly duplicated nature of the genome, the large number of contigs (>12,000) prevented us from unambiguously distinguishing *bona fide* gene duplications from potential mis-assemblies. In this study, we took advantage of long-read, single-molecule Pacific Biosciences RS (PacBio) sequencing technology, which offers considerably longer read lengths. gDNA was extracted from *H. werneckii* strain EXF-2000 and sequenced using PacBio P4/C2 chemistry. Reads (average length of 5458 bases) were *de novo* assembled to generate genome sequence Hw 2.0 (Table 1) using the SMRT Analysis version 2.3 suite and RS_HGAP_assembly.3 protocol with default parameters (Pacific Biosciences). In generating this new assembly, we considered a hybrid short-/long-read assembly approach (*e.g.*, Youssef *et al.* 2013) but found that there was little improvement based on summary statistics and empirical gene content inspection to the PacBio alone assembly (data not shown).

We conducted RNA-seq to aid genome annotation and to compare the expression of each ohnolog gene pair. Sequencing reads were *de novo* assembled using Trinity, an analysis recommended for *de novo* transcriptome assembly from RNA-seq data in nonmodel organisms (Haas *et al.* 2013). One caveat is that the gene models used have not been hand curated, therefore some inaccuracies are to be expected, especially for genes with multiple exons. In order to estimate the accuracy of our gene models, we compared the alignments of RNA-seq reads performed with both the STAR (Dobin *et al.* 2013) and HISAT2 (Kim *et al.* 2015) aligners. Visual inspection of the alignments using the IGV viewer (Robinson *et al.* 2011; Thorvaldsdottir *et al.* 2013) suggested that ~30% of the genes with multiple exons have imperfect gene models. A more global inspection at the transcriptome level using StringTie (Pertea *et al.* 2015) and GffCompare (https://github.com/gpertea/gffcompare) agreed with this estimate.

Our genome resequencing reduced the number of contigs from 12,620 in Hw 1.0 to 651 in Hw 2.0, yet the genome composition (*e.g.*, GC content) remained largely unchanged (Table 1). The greatest improvement in the genome assembly and annotation addressed the key questions of (1) total gene number, and (2) the number of duplicated gene pairs. The number of predicted protein-coding genes was reduced by 30% to 15,974 genes (Table 1). This decrease likely reflects the fact that the large number of contigs in Hw 1.0 comprised a substantial number of fragmented or misassembled sequences, resulting in an overestimate of the number of unique protein-coding genes.

At 49.9 Mb, the *Hortaea* genome is larger than many other species in the order Capnodiales (16.9–74.1 Mb, average size 35.8 Mb), which also includes the "whiskey fungus" *Baudoinia compniacensis* (21.88 Mb) and the wheat pathogen *Zymoseptoria tritici* (previously *Mycosphaerella graminicola*; 39.69 Mb) (Goodwin *et al.* 2011; Ohm *et al.* 2012; Stukenbrock *et al.* 2012). In this order, differences in genome size tend to reflect differences in the amount of repeat sequences; large genomes have large amounts of repeated DNA content and vice versa (de Wit *et al.* 2012). In contrast, our new genome assembly shows that Hw 2.0 contains a small amount of repetitive DNA (3.2% of total genome sequence; Table 1) and that the large genome can be attributed to a greater-than-average number of predicted protein-coding genes (~30% higher than other sequenced Capnodiales species, average of ~11,300). We previously hypothesized that this large number of genes results from a WGD event (Lenassi *et al.* 2013) and as described below, the data from Hw 2.0 supports this conclusion. In addition, the *H. werneckii* genome is more gene dense than that of its relatives: gene, protein, and exon lengths are slightly longer, while the intergenic distance is smaller (Table 1) (Ohm *et al.* 2012). As a consequence, over two thirds of the genome (69.9%) codes for proteins (Table 1). While genome size varies among the Capnodiales, the proportion of genes with introns and the length of those introns are similar among species in the order.

Table 1 Assembly statistics of the Hw 2.0 genome

	Value
Assembly size (Mb)	49.9
Number of contigs	651
Contig N50 (bp)	153,735
Contig max (bp)	787,827
GC content (%)	53.5
Repeat content (%)	3.2
Number of predicted protein-coding genes	15,974
Median mRNA length (bp)	1,833
Number of coding exons	38,282
Median coding exon length (bp)	342
Number of introns	22,308
Number of genes with introns	11,528
Median intron length (bp)	62
mRNA length/total length (%)	69.9
Number of tRNAs	148

Statistics were obtained using the MAKER genome annotation.

WGD

We next used our improved genome assembly to investigate the duplication status of all proteins and protein pairs to evaluate the recent genome duplication event in greater detail. One simple approach to address this question is to divide the genome into two "haploid" single gene copies and compare the paralogs. When contigs were randomly divided into two nonoverlapping subsets, each roughly representing a haploid complement of the genome (see *Materials and Methods*), the subsets could be aligned to each other with few gaps or rearrangements (Supplemental Material, Figure S1; a representative alignment using the 10 largest contigs is shown in Figure 1A). When aligning the subsets with an 85% minimum nucleotide similarity threshold and counting only the best one-to-one alignments for each locus, 69.46% of the genome was covered by alignable duplicated regions. Alignment of each subset to its pair also showed that there were no additional large-scale duplications beyond the global WGD in either of the subgenomes (Figure S2). This confirms that *H. werneckii* contains two highly similar copies of the genome, and that the duplicated proteins arose from a large-scale genome duplication and not through duplications of individual chromosome segments.

To comprehensively examine the predicted protein sequences (in terms of amino acid sequence similarity) for each duplicate protein pair, we used BLASTP to compare the predicted proteomes as a matrix. Over 90% of proteins were found in more than one copy (71% present as duplicates) at an *e*-value cutoff of $1e^{-50}$ (Figure 1B, left panel). In a different analysis, we aligned the predicted proteins within the proteome back to the genome using Exonerate (with a cutoff of 75% similarity) and corroborated that ~90% of protein sequences aligned to two separate genomic locations (Figure 1B, right panel). This provides unambiguous confirmation that most protein-coding genes have been duplicated. In fact, both comparisons showed that the proportion of duplicated genes is even higher than originally estimated in Hw 1.0 (Figure 1B).

To investigate the degree of sequence divergence between each pair of duplicated proteins, we aligned protein pairs generated by BLASTP to each other using MAFFT. In general, the sequence identity between pairs was very high, with most proteins differing by <5% (Figure 1C). This high sequence identity was also apparent when analyzing the dS for

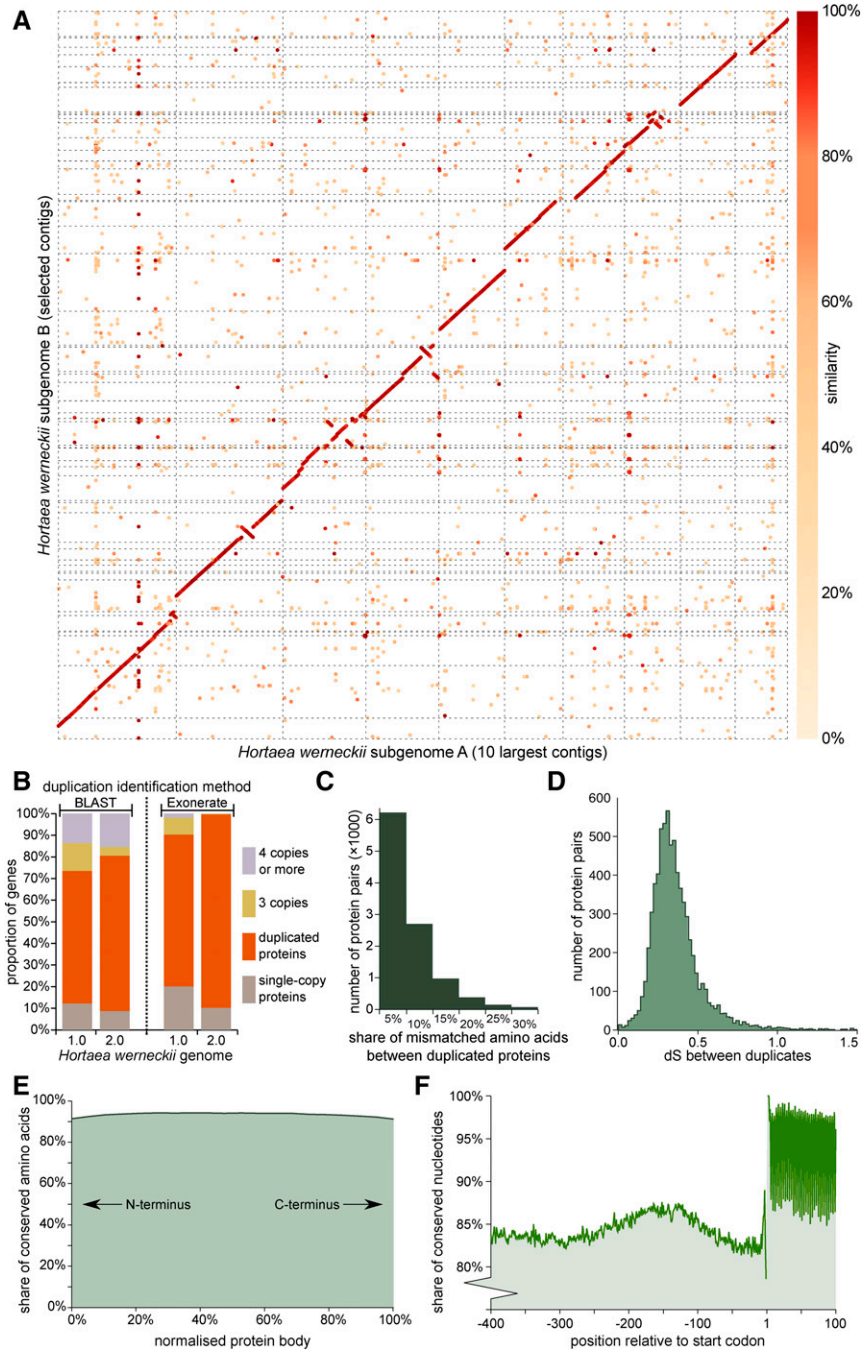

Figure 1 Evidence for whole genome duplication in *H. werneckii*. (A) Partial alignment of two genome copies of the duplicated *H. werneckii* genome. The 10 largest contigs of one genome copy were aligned against the corresponding contigs of the other genome copy, and the alignment was visualized with mummerplot. The gridlines separate individual contigs. (B) The share of duplicated predicted proteins in Hw 1.0 and Hw 2.0 genomic sequences. Duplications were determined by an all-against-all proteome BLASTP (left two bars; similarity threshold $1e^{-50}$) and by aligning the proteins back to the genomic sequence using Exonerate algorithm (right two bars; similarity threshold 75% of maximal score). (C) Assessment of sequence conservation. The number of duplicated proteins with a certain share of amino acid substitutions between the proteins in the duplicated pair. (D) Distribution of synonymous site substitution rates for paralogous genes. A histogram of dS for all paralog pairs. (E) Sequence conservation along the amino acid sequence. The average share of amino acids that remain identical between the duplicated proteins after the WGD is shown along the relative length of the protein sequence. (F) The share of conserved positions in the upstream and partial CDS regions of duplicated genes. The average share of nucleic acids that remain identical between the duplicated genes after the WGD is shown relative to the start codon.

each pair of duplicated sequences, which showed a discrete peak at a low dS (Figure 1D). This profile is consistent with a model of evolution where most of the duplicates were created at the same evolutionary time, followed by gradual accumulation of changes; multiple WGDs would show multiple peaks while duplicates created at random over evolutionary time would instead have a flat distribution across dS values. This further confirms *H. werneckii*'s WGD as seen in similar approaches to uncover plant WGD (Blanc and Wolfe 2004). The age of the WGD can be inferred to be relatively recent based on a median dS of 0.33 (mean 0.36) and the relatively tight peak. Not unexpectedly, the conservation between duplicated proteins extended along the entire length of the protein sequence (Figure 1E). A parallel analysis restricted to the CDS of these pairs also showed high sequence identity (data not shown).

Because the sequence of *H. werneckii*'s ohnologs is highly conserved, we next asked if their transcription, and therefore the resulting protein levels, differs between gene pairs. We first looked at gene structure on the premise that the diversity of upstream noncoding sequences could serve as a potential indicator of regulatory divergence. We examined the degree of sequence variation around the transcriptional start site, focusing on 4441 high-confidence gene pairs (generated using a BLASTP cutoff of $1e^{-90}$). In the 400 bp upstream of the ATG codon, ~85% of nucleotides were conserved between duplicates, with the highest conservation observed ~150 bp upstream of the start codon (Figure 1F). As expected, sequence identity was highest in the first 100 bp of the CDS (Figure 1F); throughout this region, the third base of each codon was much less conserved than the first two bases, and its conservation

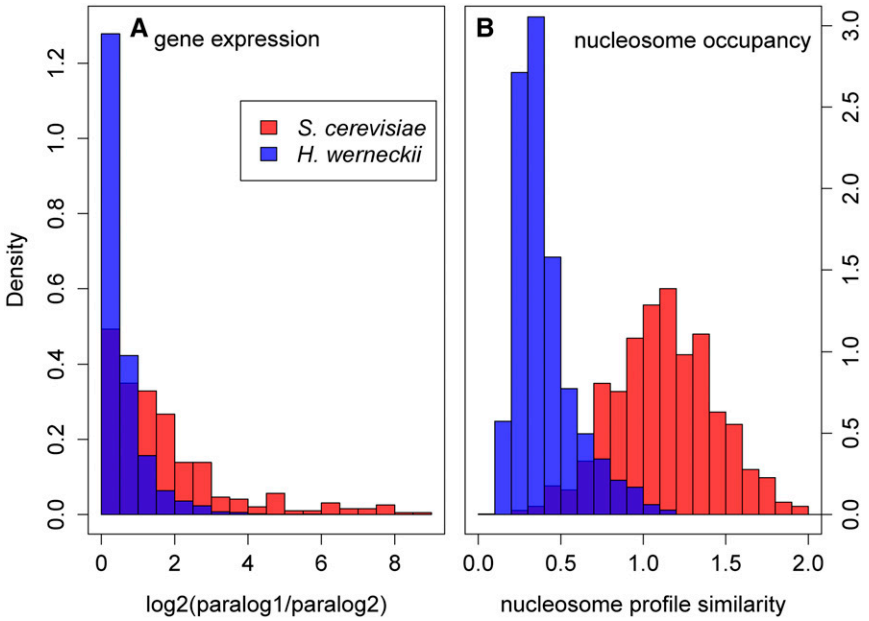

Figure 2 Similarity of gene expression level (A) and nucleosome occupancy profile (B) between paralog pairs. (A) Distribution of the absolute value of gene expression ratio between paralog pairs in $\log_2$ scale for *H. werneckii* (blue) and *S. cerevisiae* (red). (B) Distribution of nucleosome profile similarity between paralog pairs in *H. werneckii* (blue) and paralogs in *S. cerevisiae* (red) (see *Materials and Methods* for details).

was close to the region upstream of CDS. The higher divergence of sequences upstream of the CDS, where most transcriptional factors would bind, suggests that gene regulation, for example by binding of transcription factors, may have already diverged between ohnolog pairs.

To further explore the similarities between each ohnolog pair, we used two approaches to examine potential differences between the 4441 high-confidence *H. werneckii* gene pairs. To directly measure gene expression divergence, we compared transcription levels of ohnolog pairs. Since eukaryotic gene regulation is also affected by chromatin architecture (*e.g.*, nucleosome occupancy and positioning) (Iyer 2012), we also compared nucleosome occupancy between ohnolog pairs. RNA and nucleosome-bound gDNA were extracted, converted to libraries, and sequenced to generate short paired-end reads. To put these observations in context, we performed the same comparisons using comparable data available for paralog pairs of *S. cerevisiae*. The rationale behind this comparison is that while *H. werneckii* has undergone a recent large-scale gene duplication, several lines of evidence support the idea that genome duplication of *S. cerevisiae* is considerably older and the limited number of gene pairs which have persisted have undergone significant functional divergence (Musso *et al.* 2008; Li *et al.* 2010). Consistent with this premise, our analysis of the gene expression and nucleosome occupancy showed much greater conservation between ohnolog pairs in *H. werneckii* than *S. cerevisiae* (Figure 2). This suggests that not only the primary sequence, but also the function of each ohnolog is highly conserved in *H. werneckii*.

DISCUSSION

We present an improved genome assembly for *H. werneckii*, which is a foundational resource for future global studies of this unique extremophile. The associated Web sites (https://github.com/stajichlab/Hortaea_werneckii) incorporate the sequencing data, genome annotation, and associated methods in a single archive. The Gbrowse genome browser site (http://gb2.fungalgenomes.org/gb2/gbrowse/hortaea) provides additional interactive views of the genome and annotation.

A careful assessment of sequence duplication and conservation enabled us to confirm our previously hypothesized WGD. This duplicated genome content likely allows *H. werneckii* to possess the genetic redundancy associated with diploidy, thereby avoiding a haploid phase, which normally predominates in the life cycle of ascomycetes. The fact that the genome is homozygous for the MAT1-1-1 gene (Lenassi *et al.* 2013) suggests two possibilities: (1) the fungus is asexual, which might enable it to maintain the genome configuration that is well-adapted to extreme salinity (similar to observations in some pathogens; Ene and Bennett 2014); or (2) it can mate with haploid or diploid strains of the opposite mating type, resulting in triploid or tetraploid progeny, respectively. WGDs and polyploidization events are common in plants, and tend to result in increased fitness and stress resistance (Vanneste *et al.* 2014; del Pozo and Ramirez-Parra 2015). However, they still appear to be relatively rare in other eukaryotic lineages. Instead, transient ploidy changes are described in *S. cerevisiae* and others in laboratory experimental evolution experiments. In these studies, genomic content increases rapidly in response to environmental stresses (Gerstein *et al.* 2006; Dhar *et al.* 2011; Dujon 2015). *H. werneckii* therefore appears unusual among fungi, as a very recent WGD has only been described in some zygomyceteous Mucoromycotina fungi, the human pathogen *Rhizopus oryzae* (Ma *et al.* 2009), and the dung fungus *Phycomyces blakesleeanus* (Corrochano *et al.* 2016). The *R. oryzae* and *P. blakesleeanus* WGDs have been suggested to contribute to pathogenicity and the expansion of signal transduction and light sensing, while for *H. werneckii* it may confer a selective advantage in hypersaline environments. However, the relative age of the WGD in these Mucoromycotina fungi is older than that of *H. werneckii* as the duplicate pairs are much more divergent and there are many fewer sets of gene pairs. Many lineages await genomic investigations so the ever decreasing cost of sequencing and improvements in long-read and high-repeat content sequencing may help uncover additional recent WGD events in other eukaryotes.

Postduplication events normally include large-scale gene deletions, reductions, and genome reshuffling. Indeed, only 550 pairs of ohnologs remain in *S. cerevisiae* and similar numbers are seen in relatives (Dujon *et al.* 2004; Byrne and Wolfe 2005; Cliften *et al.* 2005; Scannell *et al.* 2007). This does not seem to be the case in the evolution of *H. werneckii* as the WGD, for now, appears stable. Indeed, we have shown that even after long-term growth in the presence of salt stress, the genome is not reduced (C. Gostinčar, A. Kejžar, J. Zajc, C. Nislow, S. Sinha *et al.*,

unpublished data). A WGD could result from either an endoreduplication of the genome, or from hybridization between two strains (intraspecific) or two species (interspecific). While further research is required to distinguish between these possibilities, the current genomic data does not appear to support the hybridization hypothesis. Hybridization would likely occur between strains of opposite mating types, resulting in a hybrid containing both MAT1-1-1 and MAT1-2-1 genes *vs.* two copies of the MAT1-1-1 gene as seen in *H. werneckii*. While this could have arisen secondarily through replacement of a MAT1-2-1 by recombination from the other idiomorph, this seems unlikely considering that the two MAT1-1-1 copies have diverged and are only 88.7% identical.

Considering the endoreduplication hypothesis, the stability of the WGD is worth remarking on further. Usually, after gene duplication one of the copies is expected to accumulate deleterious mutations and be lost through nonfunctionalization (Lynch and Conery 2000, 2003). The number of observed substitutions between *H. werneckii* duplicates suggests that since the WGD event there has not been a substantial loss of one of the duplicate copies from the genome, nor have there been major changes in the gene expression and chromatin structure between the ohnologs. Further comparison and reconstruction of the WGD will be important future research steps, taking into account comparisons to the closest members of the genus.

We suggest that the improved genome presented in this work, combined with the relative resistance of the genome to environmental stresses make *H. werneckii* an attractive "emerging" model organism both for understanding genetic redundancy and for developing strains for demanding biotechnological conditions.

ACKNOWLEDGMENTS

We thank the University of British Columbia Sequencing and Bioinformatics Consortium for sample preparation, sequencing, and analysis. This work was partially supported by the United States Department of Agriculture, Agricultural Experimental Station at the University of California (UC), Riverside; and the National Institute of Food and Agriculture Hatch Project CA-R-PPA-5062-H to J.E.S. Computational analyses were performed on high performance computing resources in the Institute for Integrative Genome Biology at UC Riverside supported by National Science Foundation DBI-1429826 and National Institutes of Health S10-OD016290.

LITERATURE CITED

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Blanc, G., and K. H. Wolfe, 2004 Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16: 1667–1678.

Byrne, K. P., and K. H. Wolfe, 2005 The yeast gene order browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res. 15: 1456–1461.

Chen, K., Y. Xi, X. Pan, Z. Li, K. Kaestner *et al.*, 2013 DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. Genome Res. 23: 341–351.

Cliften, P. F., R. S. Fulton, R. K. Wilson, and M. Johnston, 2005 After the duplication: gene loss and adaptation in Saccharomyces genomes. Genetics 172: 863–872.

Corrochano, L. M., A. Kuo, M. Marcet-Houben, S. Polaino, A. Salamov *et al.*, 2016 Expansion of signal transduction pathways in fungi by extensive genome duplication. Curr. Biol. 26: 1577–1584.

del Pozo, J. C., and E. Ramirez-Parra, 2015 Whole genome duplications in plants: an overview from *Arabidopsis*. J. Exp. Bot. 66: 6991–7003.

de Wit, P. J. G. M., A. van der Burgt, B. Ökmen, I. Stergiopoulos, K. A. Abd-Elsalam *et al.*, 2012 The genomes of the fungal plant pathogens *Cladosporium fulvum* and *Dothistroma septosporum* reveal adaptation to different hosts and lifestyles but also signatures of common ancestry. PLoS Genet. 8: e1003088.

Dhar, R., R. Sägesser, C. Weikert, J. Yuan, and A. Wagner, 2011 Adaptation of *Saccharomyces cerevisiae* to saline stress through laboratory evolution. J. Evol. Biol. 24: 1135–1153.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski *et al.*, 2013 STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21.

Dujon, B., 2015 Basic principles of yeast genomics, a personal recollection. FEMS Yeast Res. 15: fov047.

Dujon, B., D. Sherman, G. Fischer, P. Durrens, S. Casaregola *et al.*, 2004 Genome evolution in yeasts. Nature 430: 35–44.

Edgar, R. C., 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32: 1792–1797.

Ene, I. V., and R. J. Bennett, 2014 The cryptic sexual strategies of human fungal pathogens. Nat. Rev. Microbiol. 12: 239–251.

Gerstein, A. C., H.-J. E. Chun, A. Grant, and S. P. Otto, 2006 Genomic convergence toward diploidy in Saccharomyces cerevisiae. PLoS Genet. 2: e145.

Goodwin, S. B., S. Ben M'barek, B. Dhillon, A. H. J. Wittenberg, C. F. Crane *et al.*, 2011 Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. PLoS Genet. 7: e1002070.

Gostinčar, C., M. Lenassi, N. Gunde-Cimerman, and A. Plemenitaš, 2011 Fungal adaptation to extremely high salt concentrations. Adv. Appl. Microbiol. 77: 71–96.

Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood *et al.*, 2013 *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8: 1494–1512.

Hall, B., T. De Rego, and S. Geib, 2014 GAG: the Genome Annotation Generator (Version 1.0) [Software] http://genomeannotation.github.io/GAG.

Henikoff, J. G., J. A. Belsky, K. Krassovsky, D. M. MacAlpine, and S. Henikoff, 2011 Epigenome characterization at single base-pair resolution. Proc. Natl. Acad. Sci. USA 108: 18318–18323.

Henry, M., and L. Debarbieux, 2012 Tools from viruses: bacteriophage successes and beyond. Virology 434: 151–161.

Holt, C., and M. Yandell, 2011 MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12: 491.

Iyer, V., 2012 Nucleosome positioning: bringing order to the eukaryotic. Trends Cell Biol. 22: 250–6.

Jia, B., G.-W. Cheong, and S. Zhang, 2013 Multifunctional enzymes in archaea: promiscuity and moonlight. Extremophiles 17: 193–203.

Joshi, N. A., and J. N. Fass, 2011 Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at: https://github.com/najoshi/sickle.

Katoh, K., and H. Toh, 2008a Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. BMC Bioinformatics 9: 212.

Katoh, K., and H. Toh, 2008b Recent developments in the MAFFT multiple sequence alignment program. Brief. Bioinform. 9: 286–298.

Kim, D., B. Langmead, and S. L. Salzberg, 2015 HISAT: a fast spliced aligner with low memory requirements. Nat. Methods 12: 357–360.

Korf, I., 2004 Gene finding in novel genomes. BMC Bioinformatics 5: 59.

Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway *et al.*, 2004 Versatile and open software for comparing large genomes. Genome Biol. 5: R12.

Lenassi, M., C. Gostinčar, S. Jackman, M. Turk, I. Sadowski *et al.*, 2013 Whole genome duplication and enrichment of metal cation transporters revealed by *de novo* genome sequencing of extremely halotolerant black yeast *Hortaea werneckii*. PLoS One 8: e71328.

Lentzen, G., and T. Schwarz, 2006 Extremolytes: natural compounds from extremophiles for versatile applications. Appl. Microbiol. Biotechnol. 72: 623–634.
Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.
Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078–2079.
Li, J., Z. Yuan, and Z. Zhang, 2010 The cellular robustness by genetic redundancy in budding yeast. PLoS Genet. 6: e1001187.
Lynch, M., and J. S. Conery, 2000 The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155.
Lynch, M., and J. S. Conery, 2003 The evolutionary demography of duplicate genes. J. Struct. Funct. Genomics 3: 35–44.
Ma, L.-J., A. S. Ibrahim, C. Skory, M. G. Grabherr, G. Burger *et al.*, 2009 Genomic analysis of the basal lineage fungus *Rhizopus oryzae* reveals a whole-genome duplication. PLoS Genet. 5: e1000549.
Musso, G., M. Costanzo, M. Huangfu, A. M. Smith, J. Paw *et al.*, 2008 The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. Genome Res. 18: 1092–1099.
Ohm, R. A., N. Feau, B. Henrissat, C. L. Schoch, B. A. Horwitz *et al.*, 2012 Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen *Dothideomycetes* fungi. PLoS Pathog. 8: e1003037.
Oren, A., 2010 Industrial and environmental applications of halophilic microorganisms. Environ. Technol. 31: 825–834.
Oren, A., 2014 Taxonomy of halophilic Archaea: current status and future challenges. Extremophiles 18: 825–834.
Palmer, J., 2017 Funannotate: Fungal genome annotation scripts. https://github.com/nextgenusfs/funannotate
Parra, G., K. Bradnam, and I. Korf, 2007 CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23: 1061–1067.
Pertea, M., G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell *et al.*, 2015 StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33: 290–295.
Plemenitaš, A., M. Lenassi, T. Konte, A. Kejžar, J. Zajc *et al.*, 2014 Adaptation to high salt concentrations in halotolerant/halophilic fungi: a molecular perspective. Front. Microbiol. 5: 199.
Robinson, J. T., H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander *et al.*, 2011 Integrative genomics viewer. Nat. Biotechnol. 29: 24–26.
Scannell, D. R., G. Butler, and K. H. Wolfe, 2007 Yeast genome evolution—the origin of the species. Yeast 24: 929–942.
Schoustra, S. E., A. J. M. Debets, M. Slakhorst, and R. F. Hoekstra, 2007 Mitotic recombination accelerates adaptation in the fungus Aspergillus nidulans. PLoS Genet. 3: e68.
Shabala, S., J. Bose, A. T. Fuglsang, and I. Pottosin, 2016 On a quest for stress tolerance genes: membrane transporters in sensing and adapting to hostile soils. J. Exp. Bot. 67: 1015–1031.
Slater, G. S. C., and E. Birney, 2005 Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6: 31.
Stajich, J. E., D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz *et al.*, 2002 The Bioperl toolkit: perl modules for the life sciences. Genome Res. 12: 1611–1618.
Stanke, M., O. Schöffmann, B. Morgenstern, and S. Waack, 2006 Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7: 62.
Stukenbrock, E. H., F. B. Christiansen, T. T. Hansen, J. Y. Dutheil, and M. H. Schierup, 2012 Fusion of two divergent fungal individuals led to the recent emergence of a unique widespread pathogen species. Proc. Natl. Acad. Sci. USA 109: 10954–10959.
Thorvaldsdottir, H., J. T. Robinson, and J. P. Mesirov, 2013 Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. 14: 178–192.
Tsui, K., T. Durbic, M. Gebbia, and C. Nislow, 2012 Genomic approaches for determining nucleosome occupancy in yeast. Methods Mol. Biol. 833: 389–411.
Vanneste, K., S. Maere, and Y. Van de Peer, 2014 Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 369: 20130353.
Vaupotic, T., and A. Plemenitaš, 2007 Differential gene expression and Hog1 interaction with osmoresponsive genes in the extremely halotolerant black yeast *Hortaea werneckii*. BMC Genomics 8: 280.
Woese, C. R., and G. E. Fox, 1977 Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. USA 74: 5088–5090.
Wu, T. D., and S. Nacu, 2010 Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26: 873–881.
Yang, Z., and R. Nielsen, 2000 Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. 17: 32–43.
Youssef, N. H., M. B. Couger, C. G. Struchtemeyer, A. S. Liggenstoffer, R. A. Prade *et al.*, 2013 The genome of the anaerobic fungus *Orpinomyces* sp. strain C1A reveals the unique evolutionary history of a remarkable plant biomass degrader. Appl. Environ. Microbiol. 79: 4620–4634.

Communicating editor: J. Rine