Comprehensive annotation of the transcriptome of the human fungal pathogen *Candida albicans* using RNA-seq

Vincent M. Bruno,[1] Zhong Wang,[2] Sadie L. Marjani,[3] Ghia M. Euskirchen,[4] Jeffrey Martin,[2] Gavin Sherlock,[4,5] and Michael Snyder[1,4,5]

[1]*Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, Connecticut 06520, USA;* [2]*DOE Joint Genome Institute (JGI), Walnut Creek, California 94598, USA;* [3]*Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06520, USA;* [4]*Department of Genetics, Stanford University Medical School, Stanford, California 94305-5120, USA*

Candida albicans is the major invasive fungal pathogen of humans, causing diseases ranging from superficial mucosal infections to disseminated, systemic infections that are often lifethreatening. We have used massively parallel high-throughput sequencing of cDNA (RNA-seq) to generate a high-resolution map of the *C. albicans* transcriptome under several different environmental conditions. We have quantitatively determined all of the regions that are transcribed under these different conditions, and have identified 602 novel transcriptionally active regions (TARs) and numerous novel introns that are not represented in the current genome annotation. Interestingly, the expression of many of these TARs is regulated in a condition-specific manner. This comprehensive transcriptome analysis significantly enhances the current genome annotation of *C. albicans*, a necessary framework for a complete understanding of the molecular mechanisms of pathogenesis for this important eukaryotic pathogen.

[Supplemental material is available online at http://www.genome.org. The sequencing data from this study have been submitted to the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi) under accession no. SRA020929.]

Candida albicans, the major invasive fungal pathogen of humans, asymptomatically inhabits the mucosal surfaces of most healthy individuals and causes disease upon immune debilitation or disruption of the host's microbial flora. It is the etiological agent of mucosal infections such as oral and vaginal thrush and can also disseminate through the bloodstream to establish infection at several different anatomical sites (Klepser 2006). Hematogenously disseminated candidiasis has a 47% mortality rate despite the rapid administration of antifungal therapy (Gudlaugsson et al. 2003). A steady increase in the number of AIDS cases and of patients undergoing chemotherapy has led to an increase in the number of people suffering from *C. albicans* infections (Pfaller and Diekema 2004). This increase in prevalence, as well as an increasing resistance to existing antifungal therapies, provides a strong impetus to understand the molecular mechanisms of pathogenesis and the acquisition of drug resistance, with the hopes of identifying novel therapeutic targets. In order to obtain a comprehensive understanding of these mechanisms, it is necessary to have a complete description of the transcriptome of *C. albicans*.

The ability of *C. albicans* to cause disease largely depends on the ability to alter its transcriptome in response to different environmental stimuli and stresses to ensure survival in different host niches. A complex transcriptional circuitry ensures that morphogenesis between ovoid yeast cells and elongated filamentous cells, a process tightly linked to virulence, takes place in response to specific stimuli (Biswas et al. 2007). Changes in the transcriptional network also ensure that *C. albicans* can grow in tissues with vastly different pH values and survive stresses believed to be inflicted upon them by cells of the innate immune system (Lorenz and Fink 2001; Rubin-Bejerano et al. 2003; Bensen et al. 2004; Hromatka et al. 2005; Enjalbert et al. 2006; Chiranand et al. 2008).

RNA-seq (deep-sequencing of cDNA) provides a largely unbiased method to define comprehensively and systematically the transcriptome (the complete set of transcribed regions in a genome) of an organism in a manner that is significantly more sensitive than microarray hybridization approaches (Wang et al. 2009). This approach has been used to identify novel transcribed regions in *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, bacteria, humans, and plants (Emrich et al. 2007; Weber et al. 2007; Mi et al. 2008; Nagalakshmi et al. 2008; Wilhelm et al. 2008). The data obtained using RNA-seq have also been used to identify new splicing events and to quantify gene expression from cells grown under different experimental conditions or grown as different cell types (Marioni et al. 2008; Mortazavi et al. 2008; Sultan et al. 2008; Trapnell et al. 2009; Wu et al. 2010).

Here, we report a comprehensive transcriptome annotation of *C. albicans* using RNA-seq data. We generated data for *C. albicans* (strain SC5314) grown under a total of nine different in vitro conditions. Using our data set of about 177 million mapped reads, we have significantly refined the primary annotations of the *C. albicans* genome by determining the position of transcripts in the genomes, including 602 newly identified transcripts for strain SC5314, and by identifying 41 new introns. Furthermore, we have determined the expression levels of each of the transcripts for all nine conditions and have uncovered several examples of condition-specific

[5]**Corresponding authors.**
E-mail mpsnyder@stanford.edu.
E-mail sherlock@genome.stanford.edu.
E-mail vmb25@email.med.yale.edu.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.109553.110.

expression of novel transcripts as well as examples of condition-specific expression of previously annotated genes that were missed by traditional microarray hybridization experiments.

Results

Deep RNA sequencing of *C. albicans* isolates under diverse growth conditions

In order to perform a comprehensive analysis of the *C. albicans* transcriptome, we performed high-throughput sequencing of cDNA (RNA-seq) made from poly(A) purified RNA isolated from *C. albicans* grown under a variety of in vitro conditions. In total, nine different growth conditions were analyzed; six different experiments were performed (in duplicate) using matched media conditions (control and experimental cells were grown under equivalent conditions). A wide variety of growth conditions that are known to be relevant to pathogenesis were investigated. The conditions tested were as follows: hyphae-inducing conditions (YPD plus 10% fetal calf serum [FCS] vs. YPD alone), tissue culture medium buffered to different pH's (M199 media buffered to pH 4 vs. pH 8), high oxidative stress (YPD + 5 mM H_2O_2 vs. YPD alone), low oxidative stress (YPD + 0.5 mM H_2O_2 vs. YPD alone), nitrosative stress (YPD + 1 mM DPTA-NONOate vs. YPD alone), and cell wall damage–inducing conditions (YPD + 100 μg/mL Congo Red vs. YPD alone). From all of the experiments combined, we obtained a total of 177 million reads that mapped to the SC5314 genome (Table 1). By aligning the mapped reads against the reference genomes, we observed that our background signal was very low as evidenced by low read counts in intergenic regions and across the centromeres. Sample signal tracks of two different genomic regions are depicted in Figure 1. Biological replicates showed a very high level of correlation ($r > 0.92$) (see Supplemental Fig. S1).

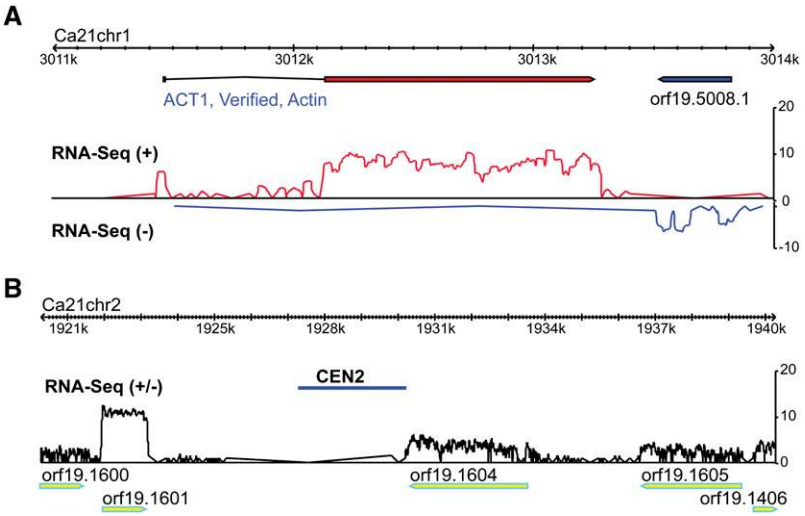

Figure 1. Representative signal tracks of RNA-seq data. Data are from *C. albicans* strain SC5314 grown in YPD. (*A*) Strand-specific representation of expression in the genomic region surrounding the *ACT1* locus. Red or blue boxes *above* the signal tracks represent *Candida* Genome Database (CGD) annotated features. The red line represents transcription from the positive strand. The blue line represents transcription from the negative strand. (*B*) Strand nonspecific representation of transcription in the genomic region immediately surrounding the centromere on chromosome 2 (CEN2).

Table 1. Sequence read mapping statistics

Growth condition	Unique mappable reads
Neg. control (CWD) #1	5,600,388
Neg. control (CWD) #2	4,940,838
CWD #1	6,119,027
CWD #2	4,952,319
High XS #1	5,618,920
High XS #2	5,817,324
Low XS #1	4,803,218
Low XS #2	5,259,905
Neg. control (NS) #1	6,042,559
Neg. control (NS) #2	5,620,768
NS #1	4,522,265
NS #2	5,049,030
Neg. control (XS) #1	4,903,719
Neg. control (XS) #2	5,412,937
pH4 #1	4,789,805
pH4 #2	5,595,842
pH8 #1	4,385,364
pH8 #2	4,396,487
YPD + serum #1	8,763,631
YPD + serum #2	12,249,820
YPD #1	12,213,306
YPD #2	11,045,798
YPD #1-SS	9,690,207
YPD #2-SS	9,689,915
YPD + serum #1-SS	10,073,614
YPD + serum #2-SS	9,601,796
Total	177,158,802

Detection of annotated genes

Using our RNA-seq data set, which includes RNA assayed from cells grown under several different conditions, we were able to successfully detect at least some expression for 6006 (97%) of the 6197 previously annotated open reading frames (ORFs) in the genome. Of the 191 ORFs that we failed to detect the expression of, 148 are annotated as "dubious" and 18 are annotated as "pseudogenes," "blocked_open_reading_frames," or "transposable element genes." Gene ontology (GO) analysis of the remaining 25 undetected ORFs revealed enrichment for in the GO Term "response to hydroperoxide," with a *P*-value of 0.0017. Failure to detect these genes could stem from them not being real genes, in the case for the "dubious ORFs" and the "pseudogenes," or from the genes simply not being expressed at a detectable level under any of the conditions that we tested.

We also analyzed our data to determine 5′ and 3′ UTR lengths (Supplemental Table S1). Strikingly, those ORFs with long 5′ UTRs (> 500 bp, of which there were 286 ORFs) were significantly enriched for genes annotated to the GO terms filamentous growth (corrected *P*-value = 2.17×10^{-6}) and regulation of biological process (6.73×10^{-6}). In contrast, those with short 5′ UTRs (<200 bp) were enriched for RNA metabolic process (5.43×10^{-6}).

Intron discovery

We set out to analyze the intron annotation, by specifically searching for confirmation of existing annotations as well as discovery of new introns. We first compiled a list of 499 existing

intron annotations based on the gene annotations from the *Candida* Genome Database (CGD) (http://www.candidagenome.org/; Skrzypek et al. 2010). We next took two complementary approaches to identifying a comprehensive list of introns in the genome using the deep-sequencing data. First, we used the TopHat software (Trapnell et al. 2009), which identifies reads that match to regions on either side of a canonical splice site, to identify both known and novel introns. Using this approach, we validated 385 of the 499 (77%) known introns (data not shown). Among the 114 introns that we were unable to detect using RNA-seq, 67 were introns in tRNA genes, nine were introns in mitochondrial genes, one was a 5′ UTR intron in a nuclear encoded gene, and 37 were in nuclear encoded ORFs. Since our analysis was performed in poly(A) purified mRNA, we did not expect to detect introns in tRNA genes and mitochondrial genes as these are not polyadenylated. We examined the sequence data for the 38 nuclear encoded genes whose introns we failed to detect, and determined that there was not enough sequence coverage in our data to be able to identify the introns in 24 of those genes. A GO term analysis revealed that these genes are significantly enriched for the annotations "meiosis" and "sexual sporulation" (transferred based on orthology to *S. cerevisiae*) (Arnaud et al. 2009), which are conditions that were not tested; 13 of the genes contain meiotic/sporulation annotations. For the remaining 11 genes, for several we observed that either one of the exons showed no expression (four cases), or there were sequence reads that mapped throughout the presumed intron (seven cases), though no ORF was clearly present in those regions. It is possible that these are unspliced products, whose efficiency of splicing is low under our growth conditions.

In a second approach to discover novel introns, we used a method developed in house (Nagalakshmi et al. 2008) to identify reads that need to be split into two segments in order to be properly aligned to the genome (junction reads). Combined, these methods revealed the presence of 41 previously unannotated introns under the growth conditions examined (for sample signal track, see Fig. 2B; Supplemental Table S2). We subsequently verified the existence of all 41 novel introns by RT-PCR, the data for 14 of which are shown in Figure 2C. In six cases, these novel introns extended currently annotated ORFs at their 5′ ends, resulting in additional predicted N-terminal protein sequence, which often aligned well with the predicted ortholog sequence from *Candida dubliniensis*. For example, for orf19.6013, we discovered two additional introns upstream of the currently annotated ORF. These two introns extend the protein coding sequence at the N terminus, such that

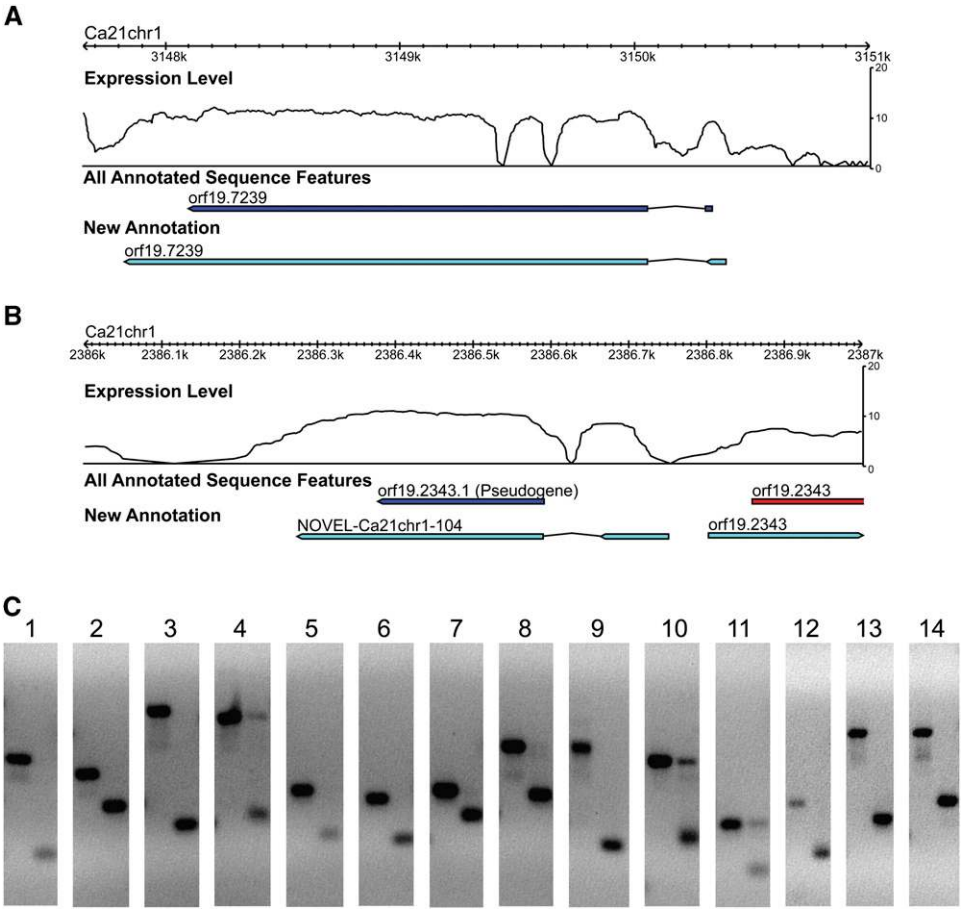

Figure 2. Intron identification and validation. (*A*, *B*) Signal tracks representing an example of a known, previously annotated, intron (*A*) and a novel intron (*B*) identified using our RNA-seq data. Red and dark blue bars represent annotated ORFs according to the *Candida* Genome Database (CGD). Light blue bars represent RNA-seq–driven annotations, which include ORF plus 5′ and 3′ untranslated regions (UTRs). Introns are depicted as two colored bars (exons) connected by a bent gray line. (*C*) Validation of introns by RT-PCR. Each panel represents a different novel intron being assayed by a different pair of oligonucleotides flanking the intron. The *left* lane in each panel is a PCR product derived from genomic DNA. The *right* lane in each panel is a PCR product derived from cDNA.

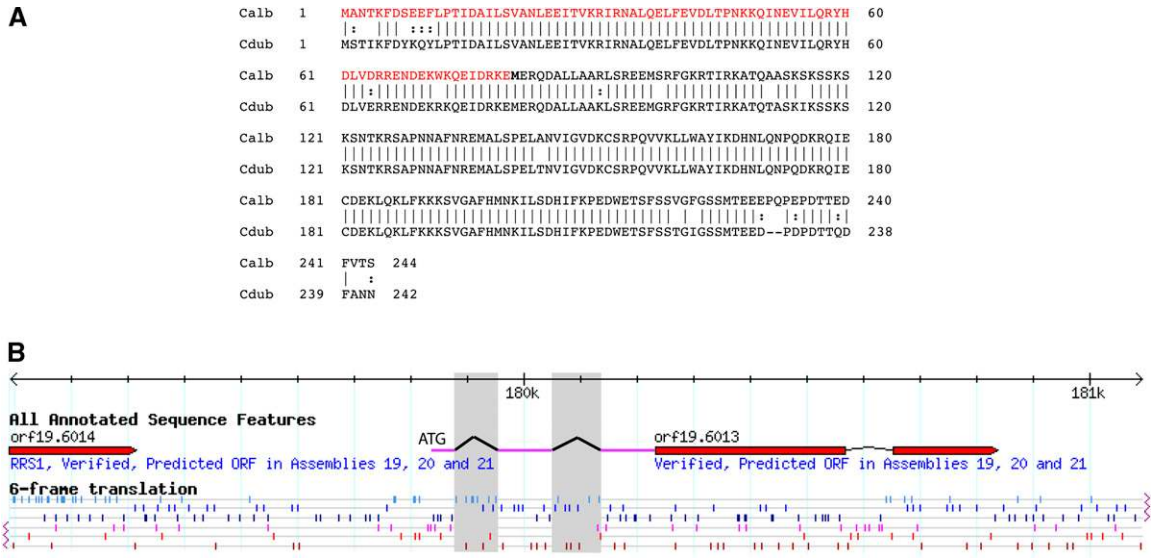

Figure 3. Re-annotation of orf19.6013 due to discovery of two novel introns. (*A*) Alignment of the new amino acid sequence (which results from the discovery of new introns) for orf19.6013 to its homolog in *C. dubliniensis*. Red letters represent new translation that results from considering the new introns. (*B*) Schematic representation of re-annotation set against the *Candida* Genome Database (CGD) GBrowse. Shaded areas represent novel introns. Red bars represent annotated CGD ORFs. Pink lines represent new coding sequence resulting from the novel introns.

there are an additional 80 amino acid residues, which align with high sequence identity to the *C. dubliniensis* sequence (Fig. 3). A second such example is orf19.5569, where the discovery of a new intron extends the protein sequence at the N terminus by 63 residues, which align with almost 100% identity to the *C. dubliniensis* ortholog (Supplemental Fig. S2). Thus, our RNA-Seq data greatly improve the intron annotation of the *C. albicans* genome.

Identification of novel expressed regions

The current gene annotation is almost exclusively based on gene prediction programs that detect protein coding regions. It does not contain genes that code for small proteins (<100 codons), and it contains few genes (other than tRNAs) with no coding potential (noncoding genes). We used our data set to search for novel transcripts. To do this, all of the strand nonspecific reads from all of the growth conditions were combined and used to identify regions of the genome with a minimum coverage of at least four reads/nucleotide for a stretch of at least 250 nucleotides (nt). We required that these regions were at least 250 nt away from an annotated gene, and allowed a gap in coverage of no greater than 50 nt to allow for introns and regions of low mapability. Using these criteria, we identified 602 regions of the genome that are expressed as polyadenylated RNA. We used an additional 39 million reads that were generated from a strand-specific protocol, from cells grown in YPD and in YPD + 10% serum, to confidently assign the expressed strand to 522 of these novel transcripts (see Supplemental Table S3). Signal tracks of two novel transcripts are displayed in Figure 4. These novel expressed regions have a median size of 434 bp. Even though most are not expected to encode long ORFs (425 have ORFs ≤ 50 codons), we did discover a few (13) that have ORFs with the potential to encode proteins with greater than 100 amino acids, which may be bona fide novel protein coding genes.

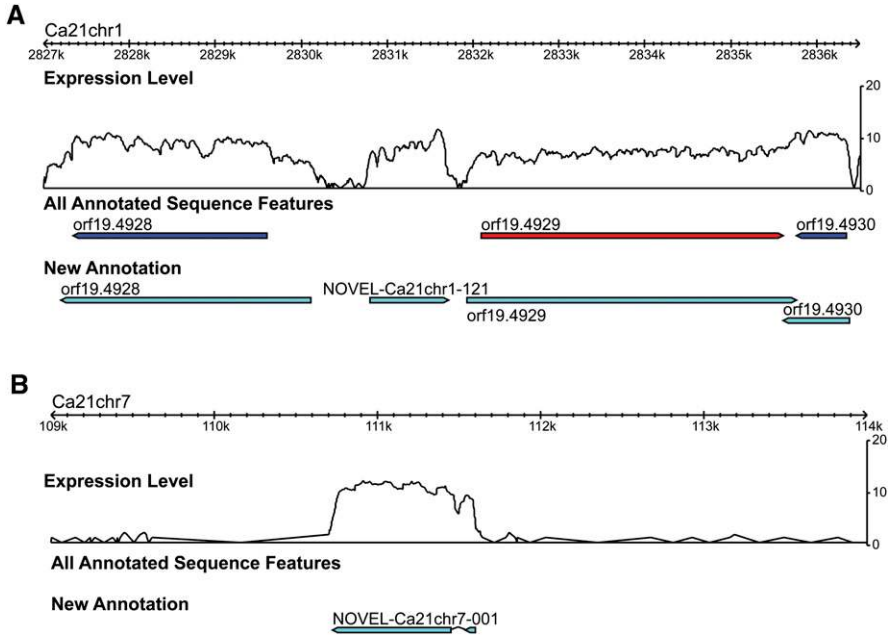

Figure 4. Identification of novel transcriptionally active regions (TARs). Signal tracks for examples of novel transcripts that are present in a region of chromosome 1 that has several annotated features (*A*) and a region on chromosome 7 that does not have any annotated features (*B*). Red and dark blue bars represent annotated ORFs according to the *Candida* Genome Database (CGD). Light blue bars represent RNA-seq–driven annotations, which include ORF plus 5′ and 3′ untranslated regions (UTRs). Introns are depicted as two colored bars (exons) connected by a bent gray line.

However, the majority of the novel transcribed regions are not expected to encode proteins, based on the absence of long recognizable ORFs capable of encoding proteins with significant similarity (*P*-value < 1×10^{-10}) to those in existing databases, with only 23 of the novel transcripts meeting these criteria, several of which were transposon related.

Analysis of gene expression reveals condition-specific expression for many *C. albicans* genes

RNA-seq provides a platform to measure differences in gene expression in a manner that is more sensitive than traditional microarray hybridization experiments (Wilhelm and Landry 2009). We used our RNA-seq data to analyze the expression of all of the previously annotated genes as well as the set of novel transcripts that we have uncovered in this study. RPKM (reads per kilobase per million mapped reads) values (Mortazavi et al. 2008) were determined for all genes in each of the conditions tested, and the resulting data were transformed by first dividing each value for a gene under a particular condition by that gene's mean RPKM value across all conditions and then taking the log (base 2) of the resulting values. Effectively, this transformed the data into the familiar log ratio values typically used for gene expression analyses. We did this so that we could determine the similarity in relative change for each transcript across the set of conditions and how those changes were similar and differed between transcripts. These data were then subjected to hierarchical clustering using the Pearson correlation as the distance metric (Fig. 5A). To identify subclusters with functional enrichment, we determined a significant Pearson correlation through permutation analysis as done previously (Brown et al. 2006). We then cut the tree at this correlation, and resulting subclusters were refined by visual inspection and then analyzed for GO term enrichment using GO::Termfinder (Boyle et al. 2004). Example subclusters are shown in Figure 5, B–D. We also clustered the RPKM data themselves, which are a representation of absolute abundance for the transcripts, and noted that some clusters also showed functional enrichment, suggesting that many transcripts that contribute to a process are maintained at similar levels to one another across conditions (e.g., Supplemental Fig. S3).

We used qRT-PCR to validate the differences in gene expression determined by RPKM analysis of the RNA-seq data. The analysis was performed on a total of 41 genes. Twenty-six of these were novel transcripts (including three that were tested under two different sets of conditions), and 15 were previously annotated genes that were not previously known to be regulated in conditions

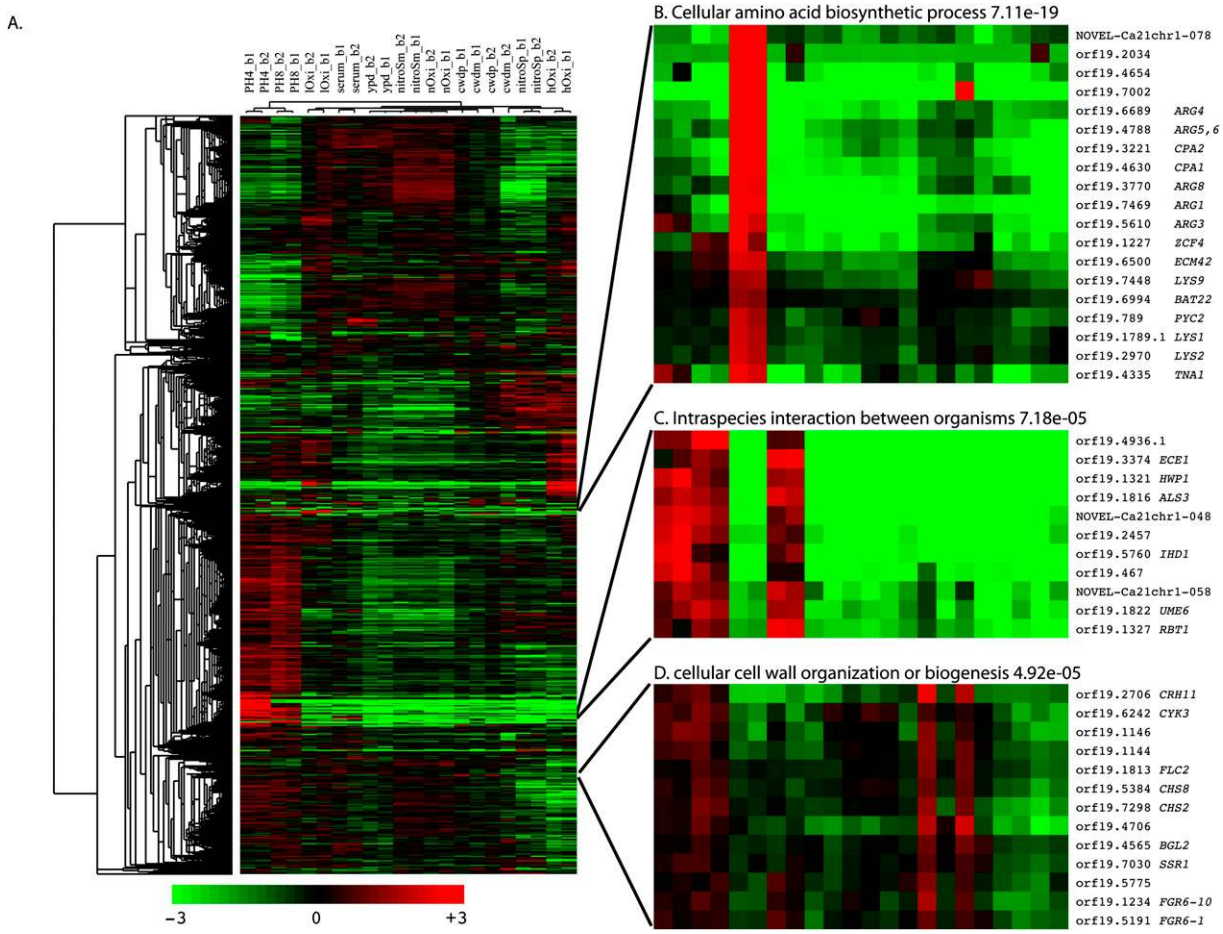

Figure 5. Cluster analysis of gene expression based on log ratio RPKM data. (*A*) Heat map depicting the results of cluster analysis of the log ratio RPKM data. (*B*) A cluster that is enriched in genes involved in "cellular amino acid biosynthetic process." (*C*) A cluster that is enriched in genes involved in "intraspecies interaction between organisms." (*D*) A cluster that is enriched in genes involved in "cellular cell wall organization or biogenesis." Green represents lower expression, red represents high expression, column represent individual experiments, and rows represent transcriptional units.

that our RPKM analysis indicated. We observed a high correlation between the gene expression changes that were calculated by the two different methods (Pearson correlation = 0.940) (see Fig. 6). Thus, we have uncovered and validated several gene expression changes of known genes that were not uncovered by traditional microarray analysis as well as condition-specific expression for many of our novel transcripts.

Inspection of the expression data revealed three classes of findings that further our understanding of *C. albicans* biology. First, we identified many examples of novel transcripts whose expression is regulated by the growth conditions. For example, novel transcript NOVEL-Ca21chr1-048 is expressed at 299-fold higher levels during growth in the hyphal form (YPD + 10% serum) than grown in the yeast form (YPD alone) (see Supplemental Table S4). It is found in a cluster (Fig. 5C) along with hyphal-specific cell wall proteins, some of which (*ALS3* and *HWP1*) have demonstrated roles in virulence via mediating interactions with host cells (Staab et al. 1999; Tsuchimori et al. 2000; Phan et al. 2007).

Second, we found several examples of completely uncharacterized annotated genes whose expression is regulated by growth conditions. These examples will help to begin characterizing gene functions for the many genes for which absolutely no functional data exists. For example, the expression of orf19.4936.1 is higher in the presence of serum (267-fold) than in the control condition (YPD) (see Supplemental Table S4). These results represent the first functional data points for orf19.4936.1 and suggest a possible role in the responses to serum response.

Third, we found examples where our analysis suggests a novel role for genes and pathways that are relatively well established and annotated. These examples could uncover novel aspects of *C. albicans* biology. For example, several genes involved in arginine biosynthesis (*ARG1*, *ARG3*, *ARG4*, *ARG5/6*, *ARG8*, *CPA1*, *CPA2*, and *ECM42*) and lysine biosynthesis (*LYS1*, *LYS2*, *LYS9*) are induced specifically in our experiments in which the cells were subject to mild oxidative stress (0.5 mM hydrogen peroxide) (for expression data, see Fig. 5C; Supplemental Table S4). None of these genes have been previously reported to be induced in response to oxidative stress in *C. albicans* or *S. cerevisiae*. Our results suggest a novel linkage between arginine and lysine biosynthesis and the oxidative stress response.

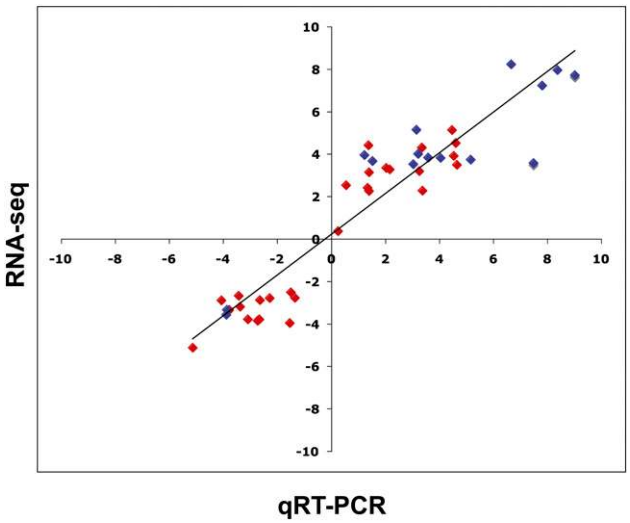

Figure 6. Verification of gene expression analysis by quantitative real-time PCR (qRT-PCR). Individual gene expression ratios (treated/untreated) were calculated using RPKM data generated by RNA-seq and plotted against calculations done for the same gene using qRT-PCR. Red diamonds represent novel transcripts. Blue diamonds represent annotated genes that were previously not known to be regulated under that particular condition. The Pearson correlation is 0.929.

Discussion

The task of defining the complete set of transcripts that an organism expresses is complicated by the fact that transcriptomes are dynamic entities that change in response to the extracellular environment. Thus, not all of the genes are expressed under any given condition or developmental state, and many genes even when they are expressed, will only be expressed at low levels. We reasoned that performing RNA-seq on cells grown under several different conditions and different developmental states would generate a more complete transcriptome map than simply assaying one growth condition. To this end, we generated RNA-seq data from cells grown under nine different in vitro conditions. The specific conditions were chosen because they approximate many of the different environments and stresses that *C. albicans* is thought to encounter during the course of an infection and as a commensal organism.

Previous genomic studies of gene expression in *C. albicans* used microarrays based on the existing genome annotations. The fact that the genome annotation is constantly being updated and even the current annotation is incomplete means that these earlier studies were missing many genes (more than 602 by our estimates), many of which could potentially play important roles in the pathogenesis of the organism. It is only with a complete annotation of the transcriptome that we can fully understand how an organism responds and reacts to different environments and causes disease.

Despite the considerable depth of sequencing coverage that we obtained, we did not detect the expression of all of the previously annotated genes in the genome. It should be noted that all of the RNA that we analyzed in this study was put through two sequential rounds of poly(A) mRNA selection, so we do not expect to detect transcripts which are not polyadenylated. Long-lived transcripts with short poly(A) tails and short transcripts, which may have been size-selected against in our isolation protocol, will also be underrepresented in our data set.

We have identified 602 transcripts that do not correspond to known annotated features in the CGD. Our analysis of gene expression of all the novel transcripts as well as the previously annotated genes has revealed a number of interesting features of gene expression. First, the expression of many of these transcripts is regulated in a condition-specific manner. Although the functions of these transcripts are currently unknown, cluster analysis of the RPKM data combined with GO analysis can provide clues about their function. For example, the possibility exists that NOVEL-Ca21chr1-048, whose expression is regulated in a manner similar to *HWP1* and *ALS3* (Fig. 5C), is involved in mediating the interaction between *C. albicans* and host cells and, by extension, virulence. A more detailed functional analysis is required, and is currently underway, to determine if this novel transcript is indeed involved in virulence. Second, we found several instances of the condition-specific gene expression of previously annotated genes that have never been reported in the literature. These cases will serve as starting points to functionally characterize the large number of genes in the genome for which no functional data exist as well as make some novel connections between well-characterized pathways and biological phenomena.

The arginine biosynthetic genes are known to be induced upon phagocytosis by neutrophils (Rubin-Bejerano et al. 2003) and macrophages (Lorenz et al. 2004). Exactly why these genes are induced in macrophages is not known, but the induction upon phagocytosis into neutrophils is thought to be a response to amino acid starvation conditions inside the phagosome and not due to oxidative stress, a crucial part of the host defense (Babior et al. 1973). This notion is supported by the inability to detect the induction of *ARG* genes upon exposure to hydrogen peroxide in a microarray experiment (Rubin-Bejerano et al. 2003). Our results suggest that *C. albicans* cells do indeed respond to mild oxidative stress by inducing the expression of arginine biosynthetic genes, and may explain why these genes are induced upon phagocytosis into cells of the innate immune system. One intriguing possibility is that *C. albicans* overproduces arginine and lysine as a mechanism to deal with cellular damage resulting from free radicals in the cells. Additional experiments are required to understand exactly why and how *C. albicans* would respond to oxidative stress in this manner, and may uncover an exciting interplay between this pathogen and the members of the innate immune system and thus further our understanding of *C. albicans* biology.

We used our data set to identify introns that are not present in the existing CGD annotation, and discovered 41 such cases. Six of the novel introns result in extensions to existing ORFs, and another six other appear to be new 5′ UTR introns. The remaining ones either are introns in novel transcripts or generate alternative splices of existing transcripts. This study dramatically improves the intron annotation and provides a more accurate view of the organism's protein coding potential.

The ribosomal protein L30, encoded by *RPL30*, is the only gene in the *C. albicans* genome that is known to undergo alternative splicing (Mitrovich et al. 2007). We were able to detect both of the alternatively spliced junctions of *RPL30* by using BLAT on the aggregate of all the reads for strain SC5314 (data not shown). Further mining of our data set is expected to be useful to identify other potential examples of alternative splicing as well as examples of genes whose splicing is regulated in a condition-dependent manner.

In summary, our data set of 177 million uniquely mapped reads will serve to significantly improve the current genome annotations of *C. albicans* through the discovery of novel transcripts and identification of novel introns. Furthermore, our approach of examining several different growth conditions has allowed us to obtain a more complete view of the transcriptome as well as the ability to uncover condition-specific regulation of annotated genes that were missed by traditional microarray hybridization experiments.

The summary data, as well as signal tracks and the novel annotations have been submitted to the CGD website.

Methods

Media and growth conditions

C. albicans strain SC5314 was routinely passaged in YPD (2% dextrose, 2% Bacto Peptone, 1% yeast extract) at 30°C.

For the serum-induction experiments, we followed a protocol established by Kadosh and Johnson (2005). Briefly, a saturated overnight culture of strain SC5314 was diluted into 100 mL of YPD medium and allowed to grow at 30°C overnight until cells reached an OD_{600} of ~13. Twenty-five-milliliter aliquots from this culture were diluted into 250 mL of fresh, prewarmed YPD medium in the presence or absence of 10% FCS (GIBCO) and grown at 30°C (absence of serum) or 37°C (presence of serum). One hour after dilution into the 250-mL cultures, cells were harvested by centrifuging and immediately stored at −80°C.

For pH4 vs. pH8 experiments, we followed a protocol established by Bensen et al. (2004). Strain SC5314 was grown overnight in YPD at 30°C. The following day, cells were pelleted, washed with M199 medium at either pH 4 or pH 8, and diluted into fresh M199 pH 4 or pH 8 medium prewarmed to 37°C to an OD_{600} of 0.5. Cells were then incubated for 4 h at 37°C with shaking and then harvested by centrifuging and immediately stored at −80°C.

For the oxidative stress experiments, a saturated overnight culture of strain SC5314 was diluted into 800 mL of YPD medium to an OD_{600} of 0.1 and allowed to grow at 30°C until the culture reached an OD_{600} of ~1.0. At this point, the culture was split into three 250-mL cultures in separate 1-L flasks. To these cultures, hydrogen peroxide (H_2O_2) was added to a final concentration of 5 mM (high oxidative stress) or 0.5 mM (low oxidative stress) or was omitted (untreated control). After 15 min of treatment at 30°C, the cells were harvested by vacuum filtration and immediately stored at −80°C.

For the cell wall damage experiments, a saturated overnight culture of strain SC5314 was diluted into 600 mL of YPD medium to an OD_{600} of 0.1 and allowed to grow at 30°C until the culture reached an OD_{600} of ~1.0. At this point, the culture was split into two 250-mL cultures in separate 1-L flasks. To one of the cultures, a stock solution of Congo Red (Sigma Aldrich), dissolved in water, was added to a final concentration of 100 μg/μL. An equal volume of water was added to the other 250-mL culture as an untreated control. After 2 h of treatment at 30°C, the cells were harvested by centrifuging and immediately stored at −80°C for later RNA isolation.

For the nitrosative stress experiments, a saturated overnight culture of strain SC5314 was diluted into 600 mL of YPD + 80 mM HEPES (pH 7.5) to an OD_{600} of ~0.1 and allowed to grow at 30°C until an OD_{600} of ~1.0. At this point, the culture was split into two 250-mL cultures in separate 1-L flasks. To one of the cultures, a 750 mM stock solution of DPTA-NONOate (Cayman chemicals; dissolved in 10 mM NaOH) was added to a final concentration of 1mM. An equal volume of 10 mM NaOH was added to the other 250-mL culture as an untreated control. After 15 min of treatment at 30°C, the cells were harvested by vacuum filtration and immediately stored at −80°C for later RNA isolation.

Total RNA isolation and mRNA purification

Total RNA was extracted from frozen pellets using the Ribopure Yeast Kit (Ambion), and poly(A) mRNA was purified from total RNA using the Micro Poly(A) Purist Kit (Ambion) according to the manufacturer's instructions.

Preparation of libraries for Illumina deep-sequencing

mRNA, from cells grown under several conditions, was fragmented into 150- to 300-bp fragments by incubation in RNA Fragmentation Reagent (Ambion) for 5 min at 70°C. The fragmented mRNA was then purified away from the fragmentation buffer by Agencourt RNAClean beads (Beckman Coulter) following the manufacturer's instructions. The purified, fragmented mRNA was then converted into double-stranded cDNA using the SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen) by priming with random hexamers. Strand nonspecific cDNA libraries were prepared for Illumina deep-sequencing according to the method previously described by Nagalakshmi et al. (2008). Strand-specific cDNA libraries were prepared as described by Parkhomchuk et al. (2009). About 30 nt of sequence was determined from one end of each cDNA fragment using high-throughput DNA sequencing (Bentley et al. 2008).

Intron discovery

To identify reads that spanned potential junctions, TopHat 1.0.12 was used (Trapnell et al. 2009), requiring a minimum intron length of 10 bp and a maximum intron length of 1500 bp. Reads were compared to Assembly 21 of *C. albicans*, available from the CGD. After identification of novel introns, the aligned reads were visually inspected using GenomeView (http://genomeview.sourceforge.net/) to further confirm them, before experimental validation was performed.

Intron validation

A pool of poly(A)-enriched mRNA was generated by combining 500 ng of each mRNA sample, which was used for deep-sequencing, into a single microcentrifuge tube. cDNA was then prepared from this pool using the SuperScript First-Strand Kit (Invitrogen). The cDNA was then used as a template for PCR using primers that flank each of the intron junctions tested. As a control, each primer pair was also used to prime PCR off of *C. albicans* genomic DNA. All PCR products are subject to agarose gel electrophoresis and visualized by staining with ethidium bromide. The oligonucleotides were systematically designed using an in-house–developed script that utilizes Primer3 to pick primers that span the splicing junctions (for a list of oligonucleotides used for validation, see Supplemental Table S5).

Acknowledgments

The part of work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231. S.L.M. is supported by a Ruth L. Kirschstein National Research Service Award (F32GM087109) from the National Institute of General Medical Sciences; G.S. is supported by R01AI077737 from the NIAID at the NIH. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute Of General Medical Sciences or the National Institutes of Health.

References

Arnaud MB, Costanzo MC, Shah P, Skrzypek MS, Sherlock G. 2009. Gene Ontology and the annotation of pathogen genomes: The case of *Candida albicans*. *Trends Microbiol* **17:** 295–303.

Babior BM, Kipnes RS, Curnutte JT. 1973. Biological defense mechanisms. The production by leukocytes of superoxide, a potential bactericidal agent. *J Clin Invest* **52:** 741–744.

Bensen ES, Martin SJ, Li M, Berman J, Davis DA. 2004. Transcriptional profiling in *Candida albicans* reveals new adaptive responses to extracellular pH and functions for Rim101p. *Mol Microbiol* **54:** 1335–1351.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456:** 53–59.

Biswas S, Van Dijck P, Datta A. 2007. Environmental sensing and signal transduction pathways regulating morphopathogenic determinants of *Candida albicans*. *Microbiol Mol Biol Rev* **71:** 348–376.

Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G. 2004. GO:TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20:** 3710–3715.

Brown JA, Sherlock G, Myers CL, Burrows NM, Deng C, Wu HI, McCann KE, Troyanskaya OG, Brown JM. 2006. Global analysis of gene function in yeast by quantitative phenotypic profiling. *Mol Syst Biol* **2:** 2006.0001.

Chiranand W, McLeod I, Zhou H, Lynn JJ, Vega LA, Myers H, Yates JR 3rd, Lorenz MC, Gustin MC. 2008. CTA4 transcription factor mediates induction of nitrosative stress response in *Candida albicans*. *Eukaryot Cell* **7:** 268–278.

Emrich SJ, Barbazuk WB, Li L, Schnable PS. 2007. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res* **17:** 69–73.

Enjalbert B, Smith DA, Cornell MJ, Alam I, Nicholls S, Brown AJ, Quinn J. 2006. Role of the Hog1 stress-activated protein kinase in the global transcriptional response to stress in the fungal pathogen *Candida albicans*. *Mol Biol Cell* **17:** 1018–1032.

Gudlaugsson O, Gillespie S, Lee K, Vande Berg J, Hu J, Messer S, Herwaldt L, Pfaller M, Diekema D. 2003. Attributable mortality of nosocomial candidemia, revisited. *Clin Infect Dis* **37:** 1172–1177.

Hromatka BS, Noble SM, Johnson AD. 2005. Transcriptional response of *Candida albicans* to nitric oxide and the role of the YHB1 gene in nitrosative stress and virulence. *Mol Biol Cell* **16:** 4814–4826.

Kadosh D, Johnson AD. 2005. Induction of the *Candida albicans* filamentous growth program by relief of transcriptional repression: A genome-wide analysis. *Mol Biol Cell* **16:** 2903–2912.

Klepser ME. 2006. *Candida* resistance and its clinical relevance. *Pharmacotherapy* **26:** 68S–75S.

Lorenz MC, Fink GR. 2001. The glyoxylate cycle is required for fungal virulence. *Nature* **412:** 83–86.

Lorenz MC, Bender JA, Fink GR. 2004. Transcriptional response of *Candida albicans* upon internalization by macrophages. *Eukaryot Cell* **3:** 1076–1087.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18:** 1509–1517.

Mi S, Cai T, Hu Y, Chen Y, Hodges E, Ni F, Wu L, Li S, Zhou H, Long C, et al. 2008. Sorting of small RNAs into *Arabidopsis argonaute* complexes is directed by the 5′ terminal nucleotide. *Cell* **133:** 116–127.

Mitrovich QM, Tuch BB, Guthrie C, Johnson AD. 2007. Computational and experimental approaches double the number of known introns in the pathogenic yeast *Candida albicans*. *Genome Res* **17:** 492–502.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5:** 621–628.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320:** 1344–1349.

Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A. 2009. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res* **37:** e123. doi: 10.1093/nar/gkp596.

Pfaller MA, Diekema DJ. 2004. Rare and emerging opportunistic fungal pathogens: Concern for resistance beyond *Candida albicans* and *Aspergillus fumigatus*. *J Clin Microbiol* **42:** 4419–4431.

Phan QT, Myers CL, Fu Y, Sheppard DC, Yeaman MR, Welch WH, Ibrahim AS, Edwards JE Jr, Filler SG. 2007. Als3 is a *Candida albicans* invasin that binds to cadherins and induces endocytosis by host cells. *PLoS Biol* **5:** e64. doi: 10.1371/journal.pbio.0050064.

Rubin-Bejerano I, Fraser I, Grisafi P, Fink GR. 2003. Phagocytosis by neutrophils induces an amino acid deprivation response in *Saccharomyces cerevisiae* and *Candida albicans*. *Proc Natl Acad Sci* **100:** 11007–11012.

Skrzypek MS, Arnaud MB, Costanzo MC, Inglis DO, Shah P, Binkley G, Miyasato SR, Sherlock G. 2010. New tools at the Candida Genome Database: Biochemical pathways and full-text literature search. *Nucleic Acids Res* **38:** D428–D432.

Staab JF, Bradway SD, Fidel PL, Sundstrom P. 1999. Adhesive and mammalian transglutaminase substrate properties of *Candida albicans* Hwp1. *Science* **283:** 1535–1538.

Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, et al. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321:** 956–960.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25:** 1105–1111.

Tsuchimori N, Sharkey LL, Fonzi WA, French SW, Edwards JE Jr, Filler SG. 2000. Reduced virulence of *HWP1*-deficient mutants of *Candida albicans* and their interactions with host cells. *Infect Immun* **68:** 1997–2002.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet* **10:** 57–63.

Weber AP, Weber KL, Carr K, Wilkerson C, Ohlrogge JB. 2007. Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. *Plant Physiol* **144:** 32–42.

Wilhelm BT, Landry JR. 2009. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* **48:** 249–257.

Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bahler J. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453:** 1239–1243.

Wu JQ, Habegger L, Noisa P, Szekely A, Qiu C, Hutchison S, Raha D, Egholm M, Lin H, Weissman S, et al. 2010. Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. *Proc Natl Acad Sci* **107:** 5254–5259.

Received April 22, 2010; accepted in revised form July 29, 2010.