RESEARCH ARTICLE

RNA Enrichment Method for Quantitative Transcriptional Analysis of Pathogens *In Vivo* Applied to the Fungus *Candida albicans*

Sara Amorim-Vaz,[a] Van Du T. Tran,[b] Sylvain Pradervand,[b,c] Marco Pagni,[b] Alix T. Coste,[a] Dominique Sanglard[a]

Institute of Microbiology, University Hospital Lausanne, and University Hospital Center, Lausanne, Switzerland[a]; Vital-IT Group, SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland[b]; Genomic Technologies Facility, University of Lausanne, Lausanne, Switzerland[c]

A.T.C. and D.S. contributed equally to this work.

ABSTRACT ***In vivo*** **transcriptional analyses of microbial pathogens are often hampered by low proportions of pathogen biomass in host organs, hindering the coverage of full pathogen transcriptome. We aimed to address the transcriptome profiles of *Candida albicans*, the most prevalent fungal pathogen in systemically infected immunocompromised patients, during systemic infection in different hosts. We developed a strategy for high-resolution quantitative analysis of the *C. albicans* transcriptome directly from early and late stages of systemic infection in two different host models, mouse and the insect *Galleria mellonella*. Our results show that transcriptome sequencing (RNA-seq) libraries were enriched for fungal transcripts up to 1,600-fold using biotinylated bait probes to capture *C. albicans* sequences. This enrichment biased the read counts of only ~3% of the genes, which can be identified and removed based on *a priori* criteria. This allowed an unprecedented resolution of *C. albicans* transcriptome *in vivo*, with detection of over 86% of its genes. The transcriptional response of the fungus was surprisingly similar during infection of the two hosts and at the two time points, although some host- and time point-specific genes could be identified. Genes that were highly induced during infection were involved, for instance, in stress response, adhesion, iron acquisition, and biofilm formation. Of the *in vivo*-regulated genes, 10% are still of unknown function, and their future study will be of great interest. The fungal RNA enrichment procedure used here will help a better characterization of the *C. albicans* response in infected hosts and may be applied to other microbial pathogens.**

IMPORTANCE **Understanding the mechanisms utilized by pathogens to infect and cause disease in their hosts is crucial for rational drug development. Transcriptomic studies may help investigations of these mechanisms by determining which genes are expressed specifically during infection. This task has been difficult so far, since the proportion of microbial biomass in infected tissues is often extremely low, thus limiting the depth of sequencing and comprehensive transcriptome analysis. Here, we adapted a technology to capture and enrich *C. albicans* RNA, which was next used for deep RNA sequencing directly from infected tissues from two different host organisms. The high-resolution transcriptome revealed a large number of genes that were so far unknown to participate in infection, which will likely constitute a focus of study in the future. More importantly, this method may be adapted to perform transcript profiling of any other microbes during host infection or colonization.**

Received 12 June 2015 **Accepted** 19 August 2015 **Published** 22 September 2015

Citation Amorim-Vaz S, Tran VDT, Pradervand S, Pagni M, Coste AT, Sanglard D. 2015. RNA enrichment method for quantitative transcriptional analysis of pathogens *in vivo* applied to the fungus *Candida albicans*. mBio 6(5):e00942-15. doi:10.1128/mBio.00942-15.

Editor Joseph Heitman, Duke University

Address correspondence to Dominique Sanglard, Dominique.Sanglard@chuv.ch.

Candida albicans is the most prevalent fungal pathogen. It is able to live and proliferate in a wide range of human body sites, including skin and mucosa, but also in the bloodstream and virtually all internal organs. Systemic infections in immunocompromised patients result in mortality rates of around 50% (1, 2). Therefore, available antifungal therapies need to be complemented by new drugs, preferentially directed against new targets to avoid cross-resistance. To identify potential novel targets in *C. albicans*, a deep understanding of the host-microbe interactions is needed. In particular, the mechanisms that allow adaptation of the pathogen to the different environments in the infected host are of great interest.

Transcriptional regulation is crucial for these adaptive processes, and thus transcriptomic studies during the course of infection, or under conditions that mimic this process, can provide valuable insights. Previous studies attempted to characterize the transcriptional response of *C. albicans* during the infection process. Most of these studies were performed *in vitro*, under conditions that mimic some of the stresses encountered by the fungus within its host (3–14), or using mammalian cells (15–18) and tissue cultures (19–21). Recent studies suggest that regulatory circuits take different trajectories *in vitro* and *in vivo* (22–24). For example, a recent study showed that the transcriptional profile of *C. albicans* transcription factor mutants (for example, *RIM101*, *EFG1*, and *ZAP1* mutants) are dramatically different during infection and during growth *in vitro* (24). These findings further emphasize the importance of performing transcriptional analysis directly during infection. Some transcriptional studies have been

TABLE 1 Enrichment of *C. albicans* RNA using the SureSelect procedure

	Galleria, 2 h p.i. (G6)		*Galleria*, 24 h p.i. (G12)		Mouse, 16 h p.i. (K23)		Mouse, 48 h p.i. (K29)	
Method	Nonenriched	Enriched	Nonenriched	Enriched	Nonenriched	Enriched	Nonenriched	Enriched
Total reads	189M[a]	24M	172M	191M	183M	38M	169M	31M
No. aligned to *C. albicans*	66K[a]	14M	93K	121M	72K	15M	169K	21M
% aligned reads	0.03	58	0.05	63	0.04	39	0.1	69
Fold enrichment		1670		1172		1003		677

[a] M, million; K, thousand.

performed on mouse kidneys systemically infected with *C. albicans* (25), on liver (26), on biofilms grown on devices placed in the bloodstream (27), or on feces of gastrointestinally infected mice (28). However, all these attempts have turned out to be a great challenge for investigators, since fungal RNA levels in recovered infected organs were very low compared to host RNA. Until now, different strategies were developed to overcome this problem, including isolation of fungal cells prior to RNA extraction (25), or specific fungal RNA amplification post-RNA extraction (26). In the first case, cells were necessarily exposed to environmental changes due to the enrichment procedure before transcription and RNA degradation could be mitigated, thus modifying the observed transcriptional response (25). In the second case, the RNA population was biased due to a nonlinear amplification of fungal RNA because of the presence of large amounts of host RNA (26). Alternative animal models have also been used, such as the rabbit (29) or the zebrafish (30), in order to recover higher fungal biomass and to be able to perform direct transcript profiling analyses on fungal RNA. Most of the studies mentioned so far used microarrays to analyze *C. albicans* transcription regulation, a method with relatively low sensitivity in quantifying the absolute expression values and in the detection of low abundance genes (31). A different technology (NanoString) was recently adapted to *C. albicans* and allows fungal transcription profiling on samples containing less than 0.1% *C. albicans* RNA (22). Even though NanoString overcomes the problem of low fungus/host RNA ratio and has a high multiplex capability, it is still limited to a restricted number of target genes and thus cannot yield a comprehensive transcriptional profile (32).

A powerful technology consisting of direct RNA sequencing (RNA-seq) has increased detection sensitivity and is capable of absolute quantification of gene expression (33). It has lately been used to study genome-wide *C. albicans* transcriptional profiles *in vitro* (34–36) or in *C. albicans*-infected mammalian cells (37, 38). So far, to our knowledge, only two studies have attempted the analysis of the *C. albicans* transcriptome by RNA-seq directly from host infections (38, 39). However, these studies were also confronted with the recurrent problem of the low fungal transcript proportion in the total RNA extracted, resulting in extremely low percentages of fungal transcripts. Such samples do not permit an acceptable sequencing depth, resulting in the detection of a small number of highly expressed genes only.

Here, we developed a strategy to extract RNA from infected hosts at early and late stages of infection and to enrich the total RNA extract in *C. albicans* mRNA, using an mRNA capture-based technology. This technology was initially developed to analyze the human exome or a small targeted panel of genes by RNA-seq approaches. In this study, we adapted this tool to *C. albicans*. For these analyses, we chose two different hosts, the mouse model of systemic infection, which is the animal model most commonly used to study *C. albicans* virulence, and the larvae of *Galleria mellonella*. This insect model is increasingly used as an alternative to mammals. Several studies showed good correlations between the results obtained with this model and the gold standard mouse model when *C. albicans* virulence factors were investigated (reviewed in reference 40). First, we demonstrate here that our selected RNA enrichment method does not introduce a bias for ~97% of the genes in the relative *C. albicans* RNA population when enriched and nonenriched samples were compared. Second, our results highlight the idea that an unprecedented resolution of the *C. albicans* transcriptome can be achieved under *in vivo* conditions, since we were able to attribute RNA-seq reads to over 6,000 *C. albicans* genes. Among these, we found a core of about 1,200 genes that are commonly regulated in the two infection models at early and late stages of infection, relative to *in vitro* grown cells. Among the *in vivo*-regulated genes, we confirmed the involvement of genes previously known to participate in infection, but we also found a large group of genes with yet-uncharacterized functions which constitute interesting future avenues of investigation.

RESULTS

Development of the RNA enrichment method to analyze the transcriptional profile of a microorganism within its host. (i) Adaptation of the SureSelect method to *C. albicans*. After extraction of total RNA from *C. albicans* infected mouse kidneys or *G. mellonella* (see Fig. S1 in the supplemental material), we estimated the percentage of *C. albicans* RNA in samples. Infecting animals with 5×10^5 *C. albicans* cells resulted in RNA samples with a range of 0.1 to 1% pathogen RNA at early time points (mouse kidneys, 16 h postinfection [p.i.]; *G. mellonella*, 2 h p.i.). Because the amounts of pathogen RNA in our samples were so small, standard RNA-seq was not feasible, since it would result in low transcriptome coverage and prevent the analysis of genes expressed at low levels. Therefore, an approach for fungal RNA enrichment based on the SureSelect capture system (Agilent Technologies) was applied. This system uses biotinylated probes that we designed to specifically match the *C. albicans* ORFome (see Materials and Methods). When this system was applied, the percentages of reads aligned to the pathogen of a given *in vivo* sample were increased up to 1,670-fold (Table 1; also, see Table S1 in the supplemental material). However, even if this technology was earlier applied and validated for human exome sequencing, it was still necessary to verify that the enrichment process did not bias the results, i.e., that the relative *C. albicans* RNA population in enriched samples remained the same as in the original samples.

(ii) Validation of *C. albicans* RNA enrichment. Our aims were to (i) establish for which *C. albicans* genes the SureSelect protocol yielded quantitatively reliable counts, (ii) rationally discard outlier genes (those unlikely to produce reliable counts), and

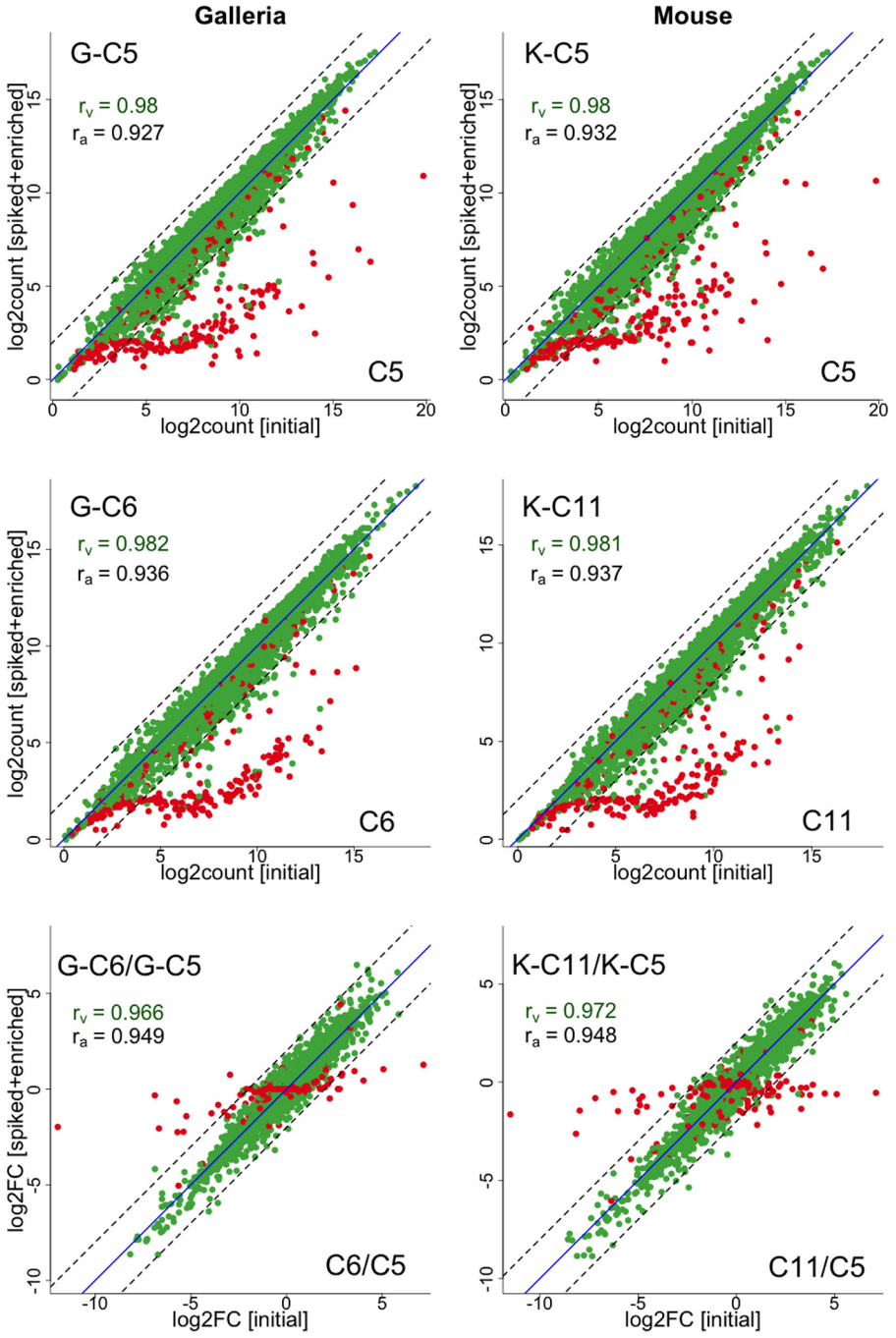

FIG 1 Correlation between enriched and nonenriched mRNA read counts of *C. albicans* genes. The enriched reads were recovered from a sample of host mRNA spiked with 1% of the same *C. albicans* mRNA. Scatter plots of $\log_2$ normalized RNA-seq counts and scatter plots of $\log_2$ fold change (log2FC) are shown. Green indicates valid genes and red indicates genes rejected via our classification with selected features. r_a, Pearson correlation for all genes; r_v, Pearson correlation for valid genes.

(iii) issue recommendations for optimizing the bait probe design. For this purpose, we investigated the correlation between the read counts obtained from the sequencing of an RNA sample from *in vitro* grown *C. albicans* cells, without enrichment, and the read counts obtained using the SureSelect enrichment protocol from a sample of host RNA spiked with 1% of the same *C. albicans* RNA (samples designated with "C"). We performed two replicate experiments using *G. mellonella* (samples designated with "G") as the host RNA (C5 versus G-C5; C6 versus G-C6) and two other experiments using mouse kidney RNA (samples designated with "K") as the host RNA (C5 versus K-C5; C11 versus K-C11). These samples are referred to here as nonenriched (C5, C6, and C11) and enriched (G-C5, K-C5, G-C6, and K-C11) samples. The experimental read counts are plotted in Fig. 1 and in Fig. S2 in the supplemental material.

We exploited a machine learning approach to identify a com-

TABLE 2 Binary features[a]

Property	Threshold values	No. of features
No. of probes per gene	>1, 2, 3, 4, 5	5
% GC[b]	>5, 10, 15, . . ., 95	57
% low-complexity sequence[b]	>5, 10, 15, . . ., 95	57
RNA-folding energy[b]	>−40, −35,−30, . . ., −5	24
Redundancy[b]	>1, 2, 3, 4, 5	15

[a] The 158 binary features associated with every gene and that can be computed from the bait probe sequences and locations.
[b] Minimum, maximum, or average per gene.

bination of gene features that can be used to determine which genes are suitable for the enrichment protocol. This combination must be made of *a priori* criteria, like the sequence and the location of the bait probes, and not of the experimental observation of gene expression. We considered five different properties that can be computed on the probes and that are likely to affect in some way the nucleic acid hybridization on which the SureSelect protocol relies. These properties are the following: (i) number of probes per gene; (ii) GC content (percent) of the probes, (iii) presence of low-complexity regions, (iv) RNA-folding energy, and (v) presence of highly similar probe sequences between different genes (redundancy). One hundred fifty-eight binary features were computed for these properties by introducing numerical thresholds, for example a required minimal GC content, and by taking the minimum, maximum, or average values over the probes that cover every single gene. Table 2 summarizes the definitions of these 158 binary features.

To select a subset of discriminative and nonredundant features, we conducted a machine learning-based feature selection process on an initial approximate classification of the genes. The genes with <4-fold changes between spiked enriched and nonenriched samples in all four experiments (5,615 genes) were tagged as valid; those with >4-fold changes in all four experiments were tagged as rejected (144 genes); the remaining 124 unclassified genes were left out of this selection step (see Fig. S2 in the supplemental material). Table 3 shows the different optimal subsets of features returned by the five different algorithms investigated. As each feature selection technique has its own advantages and disadvantages (41), we eventually retained the consensus of the features obtained from these five methods, expressed as follows: number of probes of >1, average GC content of >5%, and maximum GC content of >25%. That is, the valid genes must be covered by more than one bait probe, on average the probe GC content must be greater than 5%, and there should be at least one probe with a GC content greater than 25% (*C. albicans* average GC content is 33.5% [42]). Interestingly, the low complexity and redundancy features were never selected, possibly because the probe design protocol form SureSelect (possibly) already includes such filters.

We then trained a support vector machine (SVM) (43) model on the 5,615 valid and 144 rejected genes with the three selected features. The obtained model showed that the three features had equivalent importance in the SVM classifier. Table 4 shows all combinations of the features and their frequency of occurrence for the *C. albicans* genes and for the particular set of bait probes investigated here. Figure 1 shows the corresponding final classification of the genes. The correlations of the gene counts (Fig. 1) were improved from 0.93-0.94 to 0.98 by rejecting 662 genes out of 6,468 (according to SC5314 genome version A21-s02-m09-r07). The standard deviations of $\log_2$ fold change were reduced from 0.72-0.76 to 0.44-0.47. Figure 1 also shows the correlations of the $\log_2$ fold change between the initial nonenriched and the enriched samples. The rejected genes (see File S1, "Gene_Filter_Accepted_Rejected" in the supplemental material) clearly appear as outliers in these two comparisons, which indirectly confirms the effectiveness of our gene selection procedure.

Out of the 662 rejected genes, 411 were open reading frames (ORFs), of which only 287 are expressed (see File S1, "Rejected_genes_expressed" in the supplemental material). These corresponded mostly to genes with mitochondrial functions or ribosomal genes (see File S1, "GO_term_rejected_expressed"). In addition, baits were designed and constructed for only 8 of these ORFs, the remaining being smaller than our size threshold of 370 bp (the first 250 bp of each gene not included in the design plus 120 bp for the first bait). Taken together, the results of this analysis show that the capture process was efficient and that the excluded genes were unlikely to significantly affect the biological interpretations drawn from these data.

Analysis of *C. albicans* transcriptome during infection following enrichment procedure. After validating the enrichment method, we proceeded to the analysis of *C. albicans* transcriptome. For this analysis, samples from two infected hosts were compared at two different time points to an *in vitro* reference spiked with noninfected host material. All of these samples were subjected to the same enrichment procedure (see Fig. S1 and Table S2 in the supplemental material).

Normalized and Voom-transformed gene counts (see Materials and Methods) from enriched RNA-seq samples were subjected to hierarchical clustering and principal component analysis (PCA) (Fig. 2; also, see File S1, "voom_normalized_gene_expression" in the supplemental material). As shown in Fig. 2a, samples clustered according to their origin either from mouse, *G. mellonella*, or *in vitro* samples. The *in vivo* samples tended to cluster together, distinct from *in vitro* samples. Biological replicates fell into closely related groups. This is also visible in the PCA (Fig. 2b). Taken together, these results suggest that the data produced were of high quality, even after separate enrichment procedures.

Next, a linear model was built using the five different conditions, *in vitro* plus early and late time points from mouse kidneys

TABLE 3 Feature combinations selected by the different feature selection methods

Method	Selected features
Best first	#probe > 1, #probe > 3, avg(%GC) > 5, avg(%GC) > 10, max(%GC) > 25, min(%GC) > 10
Greedy stepwise	#probe > 1, avg(%GC) > 5, max(%GC) > 25
Linear forward selection	#probe > 1, #probe > 3, avg(%GC) > 5, avg(%GC) > 15, max(%GC) > 25, min(%GC) > 10
Scatter search	#probe > 1, avg(%GC) > 5, max(%GC) > 25
Subset size forward selection	#probe > 1, avg(%GC) > 5, max(%GC) > 25

TABLE 4 Gene classification based on the three selected features[a]

No. of genes	No. of probes > 1	Avg % GC > 5	Max % GC > 25	Class
5806	1	1	1	Valid
8	1	1	0	Rejected
289	0	1	1	Rejected
18	0	1	0	Rejected
347	0	0	0	Rejected

[a] No gene had the combination 101, 100, or 001 (impossible combination), which would have been classed as "acceptable."

(16 h and 48 h) and *G. mellonella* (2 h and 24 h), as factors to identify *C. albicans* transcriptome relationships between the different host models and time points. Differential expression of *C. albicans* genes upon host and time of infection was calculated relative to *in vitro* growth conditions (see File S1, "FC_mouse_16h", "FC_mouse_48h", "FC_Galleria_2h", and "FC_Galleria_24h" in the supplemental material). The resulting fold changes in gene expression were compared between hosts and time points in a pairwise manner (Fig. 3). As expected, the best correlation coefficients were obtained within the same host (Fig. 3a and b) ($r = 0.752$ between *G. mellonella* at 2 and 24 h; $r = 0.724$ between mouse kidneys at 16 and 48 h). Interhost comparisons resulted in lower correlation coefficients; however, the coefficient exhibited higher values when transcriptomes were compared at late time points (Fig. 3c and d) ($r = 0.707$ between *G. mellonella* at 24 h and mouse kidneys at 48 h; $r = 0.611$ between *G. mellonella* at 2 h and mouse kidneys at 16 h). Even if the *C. albicans* transcriptomes were derived from two distant hosts with expected differences in response patterns, the correlations obtained highlighted common transcriptomic responses for *C. albicans* in the two infection models.

After processing of the RNA-seq data, we first evaluated differential expression of the *C. albicans* genes during infection compared to *in vitro* growth conditions. Table 5 gives an overview of the 20 most upregulated genes in *G. mellonella* and mice at different time points, compared to *in vitro* conditions. The most upregulated genes at any time point included genes known to be important for cell host adhesion, invasion, and dissemination, including genes for a number of GPI-anchored proteins, such as *HWP1*, *RBT5*, *SOD5*, *ALS3*, and the hypha-specific gene *ECE1* (reviewed in reference 44). Two other genes involved in the regulation of hyphal morphogenesis were present (*HGC1* and *UME6*) (45, 46). Finally, genes relevant to iron acquisition were also observed (*CFL2* and *CFL5*) (47). These data demonstrated that the RNA-seq enrichment approach yielded data consistent with published studies on genes involved in pathogenesis of *C. albicans*. In addition, it demonstrated that some of the most important variations in gene expression occurred, unexpectedly, in the same way in two distant hosts, mice and *G. mellonella*.

As an additional independent verification, we validated the RNA-seq results by analyzing the expression levels of a selection of genes by qPCR on cDNA obtained directly from the original RNA samples, in which no RNA enrichment was carried out. These included genes up- or downregulated in both mouse and *G. mellonella* and genes inversely regulated between the two animal models (see Fig. S3a in the supplemental material). Fold change expression in *in vivo* samples compared to *in vitro* inocula were highly concordant between the two techniques (Spearman's correlation [r] = 0.92; $P < 0.0001$) (see Fig. S3b in the supplemental material).

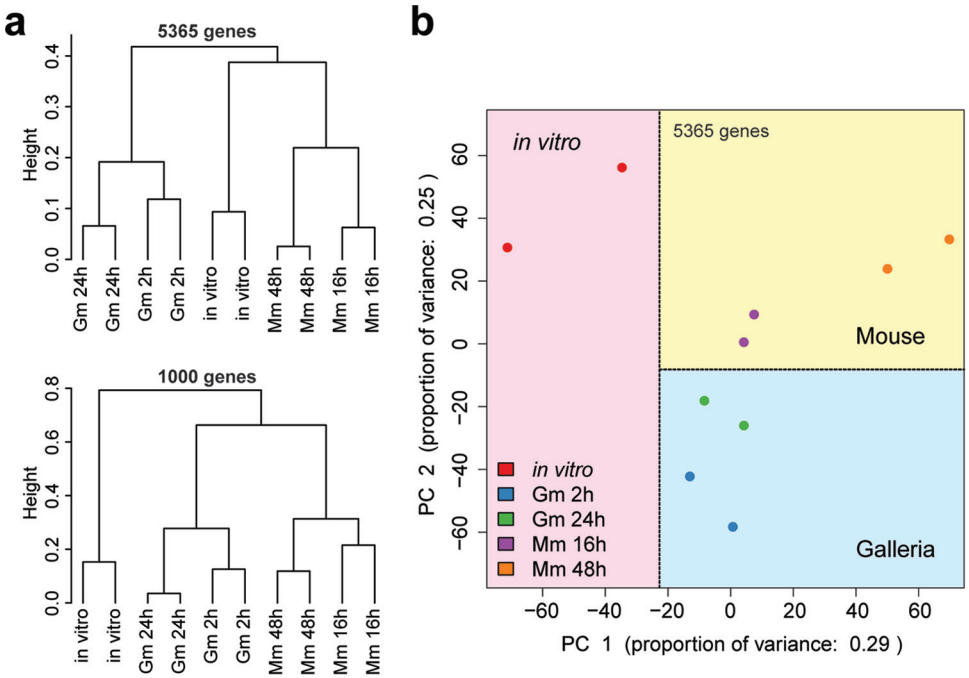

FIG 2 Hierarchical clustering (a) and principal component analysis (b) of *in vivo* and *in vitro* samples used to characterize *C. albicans* responses in the two host models. Clustering and PCA were performed using Voom-transformed and normalized gene counts. The 5,365 *C. albicans* genes with at least 1 count per million in at least one sample were used for clustering and PCA. Additionally, clustering was also performed on the 1,000 genes with the highest variance across all 10 samples. Genes that did not meet the enrichment quality criteria were excluded. Gm, *G. mellonella*; Mm, *Mus musculus*.

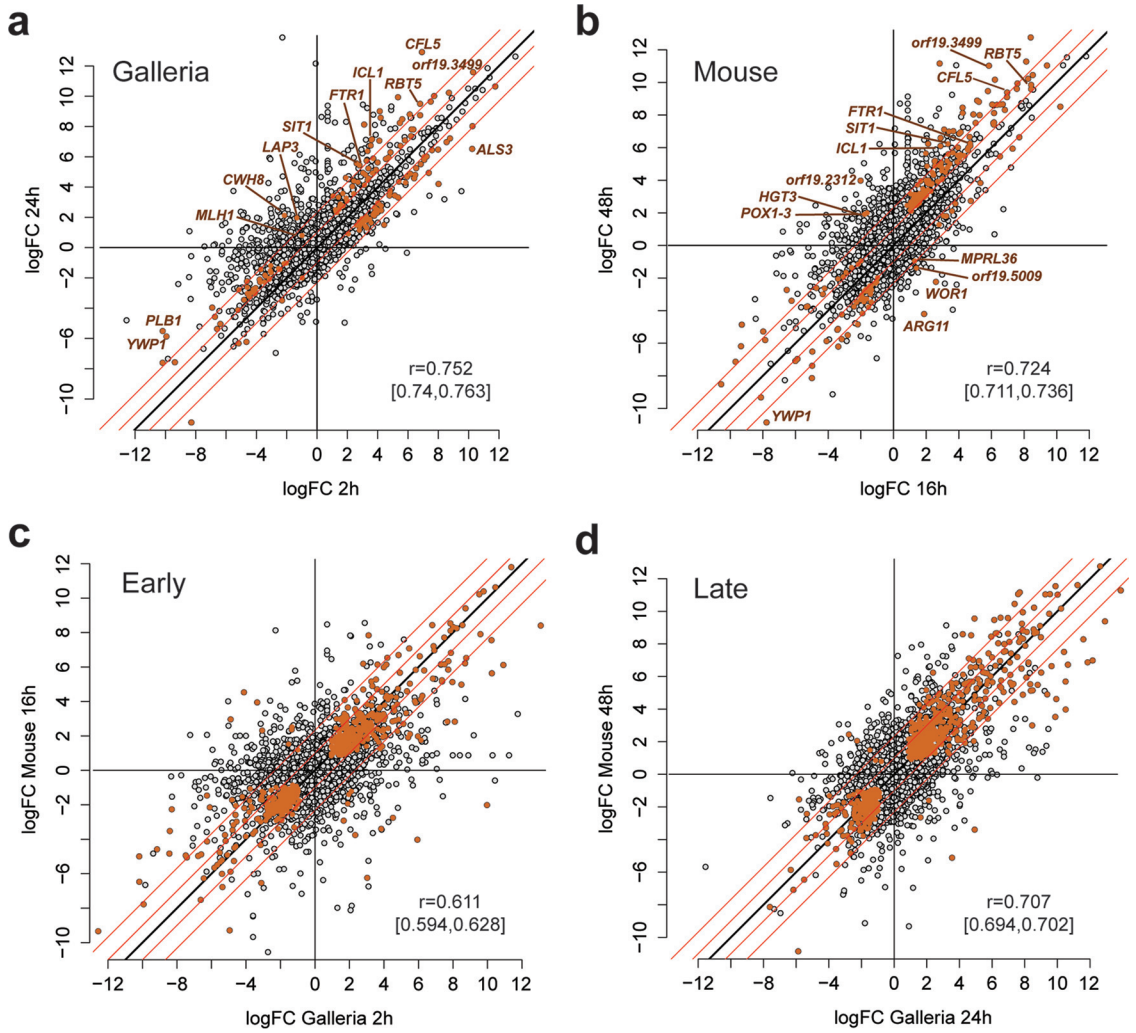

FIG 3 Correlations between $\log_2$ fold changes (logFC). For each *in vivo* condition, its log fold change versus the *in vitro* condition was computed. (a and b) Early versus late responses in *G. mellonella* and mouse. Brown dots indicate significant genes (false discovery rate [FDR] < 5%) with a difference between early and late fold changes larger than 2 (see Materials and Methods for statistical analysis). (c and d) *G. mellonella* versus mouse responses for early and late time points. Brown dots are genes significant (FDR < 5%) in both hosts. For all plots, the identity line is indicated in black. Red lines show logFC differences of −5-, −2-, 2- and 5-fold. *r*, Pearson correlation coefficients with confidence interval.

(i) Time-dependent genes in infection models. As mentioned above, the 2h and 16h time points in *G. mellonella* and mouse experiments were considered early time points, while the 24h and 48h time points in both hosts were taken as late time points. When comparing the results obtained at two time points (Fig. 3a and b), we observed that the majority of genes remained in the same category at both time points, either up- or downregulated compared to the *in vitro* condition. Nevertheless, when a cutoff of a 2-fold change between the two time points was applied, some early-upregulated genes saw a further increase of their expression at late time points, while for others, the upregulation was less pronounced at the later time point (Fig. 3a and b, colored data points). The same was valid for downregulated genes. Large differences (more than 5.5-fold between time points) were observed for several of these genes. For example, *CFL5*, encoding a ferric reductase induced under low-iron conditions, was 5.3-fold more expressed at the late time point than at the early time point in mice, while this difference was 50-fold in *G. mellonella*. Other genes participating to iron homeostasis and hemoglobin utilization, including *SIT1*, *FTR1*, and *RBT5* (48) were 6- to 7-fold more expressed at late time points in *G. mellonella* than at early time points. The same genes followed similar trends in mice; however, they were only 2.5- to 7-fold more expressed at the late time point. These data suggest that *C. albicans* needs to increase its iron capture capacity during the course of infection. Other genes, such as *ICL1*, a gene which encodes isocitrate lyase and which is critical for the glyoxylate cycle, were 11- and 9-fold more expressed at late time points than early time points in both infection models, thus revealing increased adaptation to alternative carbon sources in *C. albicans* during these infection phases (49). The gene encoding orf19.3499, a secreted protein that is Hap43 repressed, is upregulated in both *G. mellonella* and mice at early time points (50- to 200-fold compared to expression *in vitro*). However, it is further upregulated 1,000- to 2,000-fold in both hosts at later time points as compared to *in vitro*. *ALS3*, which is important for adhesion to host cells (50), was 250-fold upregulated at early time point but

TABLE 5 *C. albicans* genes most upregulated during systemic infection

orf19 name	Gene name	Description	Fold change compared to *in vitro* growth (log_2 values)[a]			
			Gm, 2 h	Gm, 24 h	Mm, 16 h	Mm, 48 h
orf19.7455	orf19.7455	Ortholog of *C. dubliniensis* CD36: Cd36_86630	11.38	11.24	11.80	11.60
orf19.1321	*HWP1*	Hyphal cell wall protein	10.46	9.89	10.63	11.54
orf19.6028	*HGC1*	Hypha-specific G1 cyclin-related protein involved in regulation of morphogenesis	9.78	9.55	10.39	11.05
orf19.3374	*ECE1*	Hypha-specific protein	8.73	9.26	9.40	11.04
orf19.1363	orf19.1363	Putative protein of unknown function	7.85	7.38	8.53	10.45
orf19.5636	*RBT5*	GPI-linked cell wall protein	6.80	9.50	8.43	9.85
orf19.2060	*SOD5*	Cu and Zn-containing superoxide dismutase	8.55	6.97	8.43	9.59
orf19.7094	*HGT12*	Glucose, fructose, mannose transporter	13.09	12.62	8.41	12.76
orf19.1816	*ALS3*	Cell wall adhesin	8.02	4.20	8.31	8.59
orf19.5585	*SAP5*	Secreted aspartyl proteinase	8.22	8.57	8.25	10.22
orf19.1822	*UME6*	Zn(II)2Cys6 transcription factor	7.86	7.86	8.11	8.87
orf19.2457	orf19.2457	Protein of unknown function	9.59	8.99	7.91	8.41
orf19.5952	orf19.5952	Protein of unknown function	3.11	8.13	7.84	8.73
orf19.2062	*SOD4*	Cu and Zn-containing superoxide dismutase	7.50	9.64	7.71	6.64
orf19.2061	orf19.2061	Ortholog of *C. albicans* WO-1: CAWG_03846	10.23	6.52	7.56	9.61
orf19.4599	*PHO89*	Putative phosphate permease	6.87	7.54	7.21	7.20
orf19.1264	*CFL2*	Oxidoreductase; iron utilization	5.46	8.23	7.08	6.86
orf19.1930	*CFL5*	Ferric reductase	6.92	12.92	6.96	9.38
orf19.113	*CIP1*	Possible oxidoreductase	7.88	7.90	6.82	8.31
orf19.6148	orf19.6148	Homolog of nuclear distribution factor NudE, NUDEL	4.84	6.29	6.69	8.66

[a] Mm, *M. musculus*; Gm, *G. mellonella*.

only 18-fold upregulated at later infection times in *G. mellonella*. In mice, the same gene did not vary its expression pattern during the course of infection (approximately 350-fold upregulated compared to the *in vitro* condition at the two time points; *ALS3* is not labeled in Fig. 3b). With regard to downregulated genes, *PLB1*, encoding a phospholipase required for host cell penetration (51), was 25-fold less expressed at 2 h than at 24 h in *G. mellonella*. This gene was downregulated in mice compared to the *in vitro* condition, however, about 88-fold at both time points. *YWP1* (encoding a yeast wall protein) was 16-fold less expressed at early time points than at later time points in *G. mellonella*. This gene followed a similar trajectory in mice at the early time point but was further downregulated (8-fold) at the later time point.

Very few genes were inversely regulated between early and late time points. Namely, *WOR1*, *ARG11*, *MPRL36* and orf19.5009 were upregulated at the early time point but downregulated at the late time point in mice (Fig. 3b). *HGT3*, *POX1-3*, and orf19.2312 (*CFL11*) in mice (Fig. 3b), plus *CWH8*, *LAP3*, and *MLH1* in *G. mellonella* (Fig. 3a), were downregulated at early time points but upregulated at late time points. In spite of the mentioned differences between time points for some of the genes, the majority of the transcriptome remained stable across time, at least at the time points investigated here. This conclusion can also be drawn from Fig. 3c and d, in which most of >2-fold-regulated genes between conditions remained in the same cluster of colored data points. A more detailed time course experiment might be needed to observe a comprehensive kinetic of gene expression during infection.

(ii) *C. albicans* transcriptomic signature of host infection. The correlations between *C. albicans* expression profiles under the different host infection conditions prompted us to identify a common gene expression signature reflecting a state of infection. A two-step meta-analytical approach was used to identify genes affected *in vivo* regardless of the host model and the time point (Fig. 4a). First, *z* scores computed from *P* values were combined meta-analytically for each host separately. Then, a unique *z* score per gene was derived from the two host-specific *z* scores. Positive and negative *z* scores reflect the fact that genes are either up- or downregulated, respectively, in both hosts and at any time of infection compared to *in vitro* cultures. This approach allowed the identification of 1,169 *C. albicans* genes significantly affected during the host infection process (Bonferroni *P* value, ≤0.05) (Fig. 4b and c; also, see File S2, "meta_analysis," in the supplemental material). Only 35 genes in mouse and 49 in *G. mellonella* showed a clear host-specific response (see File S2, "Genes_Galleria_ specific" and "Genes_Mouse_specific"). Among these genes, *TLO1* stands out as a gene with an inverse regulation in the two hosts. *TLO1* from a closely related species (*Candida dubliniensis*) was shown to be critical for expression of genes relevant to virulence (adhesins, iron acquisition genes, amino acid catabolic genes, etc.), thus suggesting that *TLO1* in *C. albicans* may have similar functions (52). The difference of expression of *TLO1* between the two investigated hosts may suggest that this gene has a niche-specific role.

The 1,169 genes of the host infection signature were analyzed for their biological significance by two different approaches. First, a gene set enrichment analysis (GSEA) was performed using a list of regulated genes obtained from transcriptional data compiled from the literature, including several conditions in which *C. albicans* (i) was exposed to host cells, (ii) was grown *in vitro* under conditions mimicking host environments, or (iii) was extracted from host tissues (see File S2, "CandidaL_exp_dataL2.gmt" in the supplemental material). The GSEA list also contains *C. albicans* genes with binding sites of several transcription factors. *C. albicans* genes that were classified as "*in vivo*-regulated" in our study were used as a ranked list of *z* scores in GSEA. As shown in Fig. 5, there were two major clusters of genes with positive (red) and negative (green) values. The nodes within the positively regulated genes

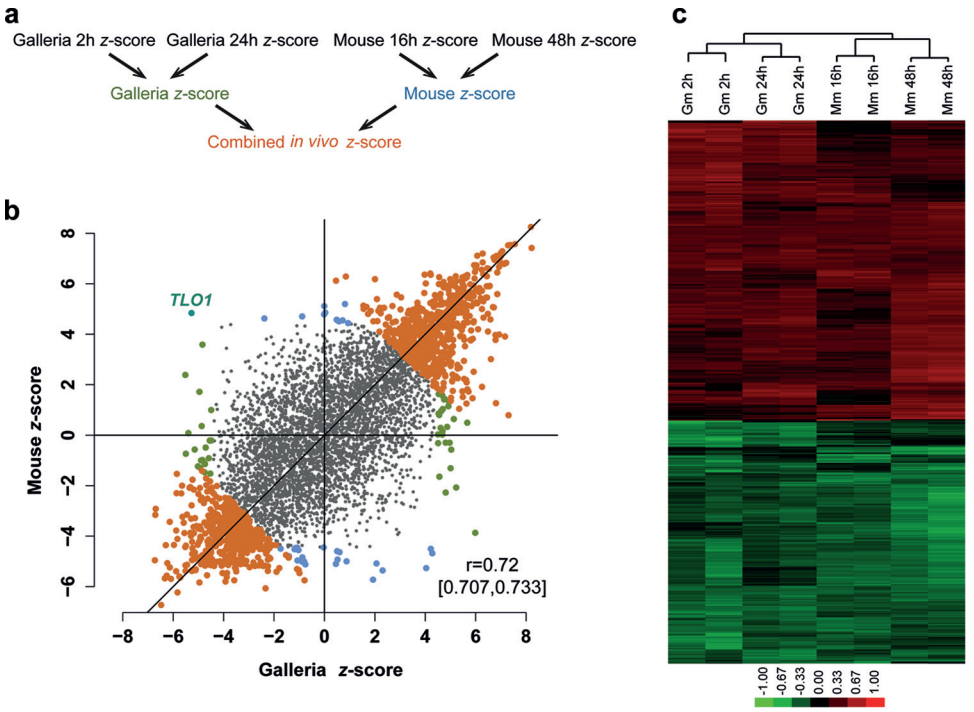

FIG 4 Identification of *C. albicans* genes differentially expressed *in vivo* versus *in vitro* using a meta-analytical approach. (a) Meta-analysis strategy. Limma contrast statistics were converted to *z* scores (see Materials and Methods). *z* scores were then combined meta-analytically as illustrated. (b) Scatter plot of mouse and *Galleria z* scores obtained meta-analytically. Mouse and *Galleria* z scores are further combined into one *in vivo z* score. The 1,169 genes for which this combined *z* score is significant (Bonferroni *P* value ≤ 0.05) are indicated in brown. Genes for which the combined *z* score is not significant are indicated in blue if the mouse *z* score is significant or in green if the *G. mellonella* z score is significant. *TLO1* is the only gene with significant and anti-correlated *z* scores in mouse and *G. mellonella*. *r*, Pearson correlation coefficient with confidence interval. (c) Heat map of the 1,169 significant genes by meta-analysis. For each sample, a log fold change versus the average *in vitro* expression was computed. The log fold change values were variance scaled and are represented on the heat map. Hierarchical clustering tree of the samples is indicated at the top. Gm, *G. mellonella*; Mm, *Mus musculus*.

cluster formed a network in which the node "BIOFILMS_UP" is central, since it exhibits the highest number of genes connected to other gene sets. About 300 *in vivo*-upregulated genes overlap genes that are also upregulated in biofilms (1,001 genes) (53). Consistently, the cluster of positively regulated genes contains nodes of transcription factor binding sites (Fig. 5, green labels for nodes BCR1, BRG1, ZAP1, and NDT80) that were reported as important for biofilm formation (54). Other gene sets, including genes induced by several hypha-inducing conditions, were also present in this cluster (Fig. 5, red labels; conditions include growth at 37°C, addition of serum, and growth in Spider medium). Hyphal formation is a key element of the *in vivo* lifestyle of *C. albicans* (55), and our transcriptional data reflect this feature. Besides these enrichment sets, others were more related to the exposure of *C. albicans* to host cells or host environment (Fig. 5, orange labels; conditions include exposure to reconstituted human epithelium, to blood, to plasma, to macrophages, and to *C. albicans* from caecum). Nodes relative to the white-opaque switch in *C. albicans* are also present and suggest that genes regulated by this process are also enriched in the *in vivo* gene cluster shown in Fig. 5. Furthermore, this cluster also contains a gene expression set from *C. albicans* infecting zebrafish, suggesting a significant overlap (52 genes over a total of 192 upregulated genes) in the response of the fungus to infection of zebrafish, mice, and *G. mellonella*. In addition, the *in vivo*-downregulated genes of the present study were clustered with gene sets containing mostly downregulated genes, thus mirroring the opposite cluster. Lastly, some nodes containing *in vivo*-expressed genes were identified by GSEA but were not associated with the 2 major clusters of Fig. 5. Among them, iron-regulated genes are present (Fig. 5, yellow labels) and also indicate the relevance of iron homeostasis in the *in vivo* lifestyle of *C. albicans*. Some nodes also indicate an inverse regulation overlapping the group of *in vivo* genes. For example, genes upregulated by hypoxia *in vitro* ("HYPOXIAB_UP" and "HYPOXIA20MIN_UP") overlap genes downregulated *in vivo*. While these results may suggest an absence of hypoxia for *C. albicans* in the two hosts tested, this inverse regulation pattern could also reflect that *in vitro* conditions mimicking host conditions may not always reflect the physiological status of *C. albicans in vivo*.

A second approach in the biological interpretation of our results was taken by subjecting *in vivo* genes to gene ontology (GO) enrichment analysis. The genes were separated by their positive and negative *z* scores. As shown in Table 6, positively regulated genes were enriched for several processes, including those involved in response to external stimulus, pathogenesis, biofilm/hypha formation, and iron homeostasis. All these processes are known to participate to the virulence of the pathogen. In addition, many of the listed genes belong to pathways involved in cell wall biogenesis and maintenance. Negatively regulated genes were enriched for processes such as cellular amino acid biosynthesis, glycolysis, mitochondrial electron transport, translation, and induction of host defense response. The enrichment of such processes in this category of regulated genes reflects the fact that metabolic functions (glycolysis and respiration), biosynthesis of cellular

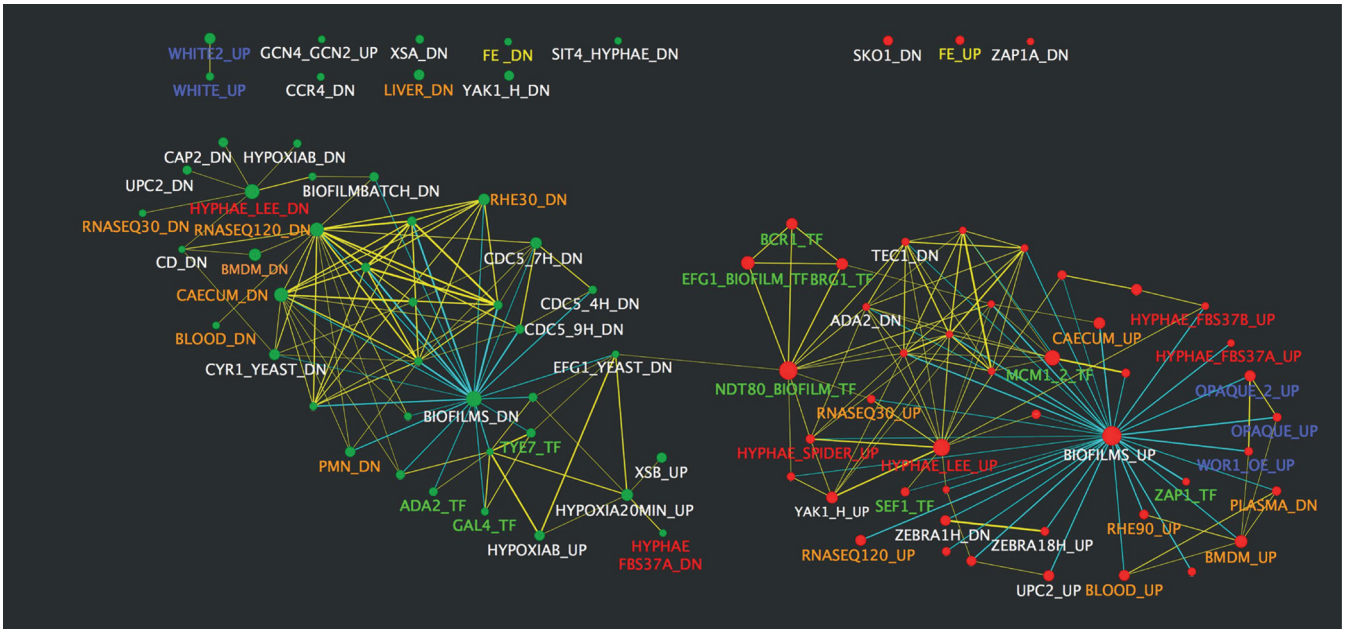

FIG 5 GSEA of *C. albicans* genes regulated *in vivo*. The gene list was produced from data in File S2 in the supplemental material ("meta-analysis", "CandidaL_exp_dataL2.gmt"), in which genes with *P* values of ≤0.05 (*in vivo*) were chosen. The genes were ranked according to their *z* scores. The list was then imported into the GSEA software. Analysis parameters were as follows: norm, meandiv; scoring_scheme, weighted; set_min, 15; nperm, 1000; set_max, 500. GSEA results were uploaded into Cytoscape 3.0 with the following parameters: *P* value cutoff, 0.01; FDR *q* value, 0.05. Red nodes represent enriched gene lists in upregulated genes from the GSEA. Green nodes represent enriched gene lists in downregulated genes from the GSEA. Nodes are connected by edges when overlaps exist between nodes. The size of nodes reflects the total number of genes that are connected by edges to neighboring nodes. Edge thickness reflects the level of confidence between nodes. Colored labels of nodes are defined in the text and indicate specific classes of genes.

components, and protein translation are less active in *in vivo*- than *in vitro*-grown cells, which is also typical of a less favorable growth environment.

DISCUSSION

Studies of the *C. albicans* transcriptome during host infections have been hampered by the low abundance of microbial transcripts within samples containing excess host RNA. In this study, we showed that by using a hybridization-based RNA enrichment procedure, one can target the pathogen transcriptome even when it is diluted in a large excess of host RNA. We demonstrated that the enrichment and amplification steps permit the quantitative recovery of RNA from most fungal genes but not from a few that can be identified based on *a priori* criteria. Essentially, insufficient bait coverage and low GC content preclude quantitative recovery of some of the genes through a hybridization-based procedure. A machine-learning approach contributed to the establishment of the relevant gene features and numerical thresholds for gene rejection criteria. In the future, these results could serve to guide the design an improved set of baits. More generally, we believe that this integration of wet-lab and *in silico* methods is not limited to the particular experimental systems investigated here but is applicable to a vast range of parasite/host experimental systems.

This approach allowed an unprecedented resolution of the *C. albicans* transcriptome during infection. For example, in the microarray-based *in vivo* study by Thewes et al. (26), only 787 *C. albicans* genes (strain SC5314) could be detected in mouse kidneys, and 476 were differentially expressed in a statistical significant manner. In comparison, our approach allows detection of 5,365 *C. albicans* genes (86% of the 6,218 ORFs from *C. albicans*), with 2,174 genes being differentially expressed in mouse kidneys at 48 h p.i. ($\log_2$ fold change ≥ 2; $P \leq 0.05$). The higher resolution achieved here has the major advantage of allowing the study of genes that might have a relevant role in infection, despite their low expression levels. Accordingly, we identified several genes with unknown functions that are differentially expressed *in vivo*. Thus, these data may represent a milestone in the understanding of *C. albicans* biology.

Time-dependent gene expression. With the validated enrichment procedure, we investigated the *C. albicans* transcriptome in two different hosts and at two different time points. Time points were chosen to differentiate between early and late responses of *C. albicans* in the two hosts. The majority of genes were similarly regulated at the two time points. We noticed that iron homeostasis genes increased their expression over time in both hosts, highlighting the relevance of iron acquisition in the pathogenesis of *C. albicans* in both hosts. *SEF1* is a transcriptional activator that positively regulates several genes mediating iron uptake under iron-depleted conditions (56). Our data reveal that *SEF1* is upregulated under all conditions tested here compared to *in vitro* growth (7- to 9-fold in *G. mellonella*; 18- to 20-fold in mice) (see File S1, "FC_mouse_16h", "FC_mouse_48h", "FC_Galleria_2h", and "FC_Galleria_24h", in the supplemental material), consistent with the expression profiles of *SEF1* target genes discussed here, including *CFL5*, *RBT5*, and *FTR1*. *HAP43* and *SFU1* are both transcriptional regulators and part of the *SEF1* regulatory circuit. *HAP43* represses iron utilization and *SFU1* represses iron uptake systems, and these genes were up- and downregulated, respectively, under all conditions tested here (see File S1, "FC_mouse_16h", "FC_mouse_48h", "FC_Galleria_2h", and

TABLE 6 List of *C. albicans* genes regulated "*in vivo*" and categorized according to enriched GO terms[a]

Group and GO term (biological process)	Enrichment fraction[c]	Log odds ratio	Adjusted *P* value	Gene list
Positive *z* scores				
Regulation of response to stimulus (GO:0048583)[b]	47/252	0.88	0.03	*HGC1*, *UME6*, *PTP3*, *CEK1*, orf19.6705, *HGT1*, *SLN1*, *FAV1*, *CST5*,*BEM2*, *PRA1*, *KEX2*, *RPN4*, *ALS1*, orf19.2565, *CLA4*, orf19.3501, orf19.4792, *GCN2*, *HSP70*, *RIM101*, *RGS2*, *MSB2*, *IQG1*, *RAX1*, *MKK2*, *CCN1*, *SAP4*, *BUD2*, *LTE1*, *CAG1*, *GPA2*, *CZF1*, *ZCF30*, *ZCF2*, *GEA2*, *SKN7*, *TEA1*, *AHR1*, *CPP1*, *TEC1*, *HRR25*, *SET1*, *CRZ1*, *RAS1*, *PPR1*, orf19.321
Iron ion homeostasis (GO:0055072)	16/41	1.95	0.00	*HMX1*, *HAP43*, *FTR1*, *IRO1*, *FRP1*, *RBT5*, *CSR1*, *CFL2*, *CSA2*,*CCC2*, *FET34*, *CSA1*, *ISU1*, *SEF1*, *ALS3*
Regulation of filamentous growth (GO:1900443)	18/60	1.57	0.03	*HGC1*, *UME6*, *CEK1*, *SLN1*, *CLA4*, *CCN1*, *GPA2*, *CZF1*, *ZCF30*,*ZCF2*, *SKN7*, *TEA1*, *AHR1*, *CPP1*, *TEC1*, *CRZ1*, *RAS1*, *PPR1*
Biofilm formation (GO:0042710)	30/137	1.11	0.04	*HWP1*, *RBT5*, *ECE1*, *TRY5*, *HYR1*, *HGC1*, *PHR1*, *ALS3*, *PGA7*, *CSA1*, *CST5*, *CSR1*, *ALS1*, *TRY4*, *BCR1*, *ARG81*, *CSH1*, *SUN41*, *CSA2*, *ZNC1*, *TRY6*,*PBR1*, *BRG1*, *VPS1*, *CZF1*, *AHR1*, *TEC1*, *IPT1*, *RAS1*, *GUP1*
Pathogenesis (GO:0009405)	41/225	0.85	0.09	*HWP1*, *PHR1*, *UME6*, *SOD5*, *FTR1*, *ALS3*, *FET34*, *CEK1*, *SAP5*,*ICL1*, *RBT4*, *HAP43*, *KEX2*, *IFF4*, *ALS1*, *IRE1*, *CLA4*, *HEX1*, *CSH1*, *SFL2*, *HGT4*, *RIM101*, *IRO1*, *SUN41*, *CHS3*, *BRG1*, *SAP4*, *MTLA1*, *CTF1*, *BUD2*, *MIT1*, *IRS4*, *SLK19*, *GPA2*, *AHR1*, *CPP1*, *CKA2*, *TEC1*, *DUR1*,*2*, *SET1*, *RAS1*
Biological adhesion (GO:0022610)	21/86	1.27	0.08	*HWP1*, *RBT5*, *HYR1*, *PHR1*, *ALS3*, *SAP5*, *DFI1*, *PRA1*, *IFF4*,*ALS1*, *BCR1*, *CSH1*, *SUN41*, *MSB1*, *PBR1*, *SAP4*, *IRS4*, *AHR1*, *TEC1*, *SAP10*, *RAS1*
Negative *z* scores				
Cellular amino acid biosynthesis (GO:0008652)	44/132	2.02	0.00	*ILV3*, *HOM2*, *ARO3*, *ILV6*, *HIS1*, *GLO2*, *HIS5*, *LEU4*, *PDB1*, *TRP2*, *ANB1*, *THR4*, orf19.1626, *HOM6*, orf19.2306, *PRO1*, *CYS4*, *ARO2*, *HOM3*, *HIS4*, orf19.2286, *BAT21*, orf19.3810, *ASN1*, *MET16*, *ILV1*, *SER1*, *ARO4*, *MET2*, *TRP4*, *GLN1*, *MMD1*, *THR1*, *ARO1*, *SER2*, *GLT1*, orf19.7152, *ILV5*, *SAM2*, *SER33*, *MIS11*, *GSH2*, *CBF1*, *IDP1*
Glycolysis (GO:0006096)	11/16	3.07	0.00	*CDC19*, *FBA1*, *GPM1*, *TPI1*, *ENO1*, *TDH3*, *PGK1*, *PDA1*, *ADH1*,*PFK1*, *PGI1*
Induction of host defense response (GO:0044416)	9/21	2.39	0.00	*CDC19*, *FBA1*, *TPI1*, *ENO1*, *TDH3*, *PGK1*, *SSB1*, *ADH1*, *IMH3*
Mitochondrial electron transport (GO:0006122)	6/10	2.87	0.01	*RIP1*, *CYT1*, *QCR2*, *QCR8*, orf19.913.2, *CYC1*
Translation (GO:0006412)	95/404	1.52	0.00	*RPP1A*, *RPS3*, orf19.3649, orf19.5812, *RPS24*, *RPP2A*, *RPS17B*, *RPL42*, *RPL2*, *RPL37B*, *RPL39*, *DED81*, *RPP2B*, *ANB1*, orf19.5943.1, *SUI3*, *RPL20B*, *RPS19A*, *RPL28*, *RPL30*, *RPS21*, *ARC1*, *RPL16A*, *RPS23A*, *RPL18*, orf19.6220.4, *RPL15A*, *RPL5*, *RPS1*, *SSZ1*, *TYS1*, *RPS4A*, *RPL14*, *TMA19*, *RPL27A*, *RPL43A*, *RPS18*, *RPL10*, orf19.5684, *SSB1*, orf19.4149.1, *TEF2*, *RPS20*, *DRG1*, *EIF4E*, *CAM1-1*, *RPL23A*, *RPL17B*, *RPL3*, orf19.3341, *RPL19A*, *RPL13*, *RPL21A*, *RPS9B*, *RPS28B*, *GRS1*, *SIS1*, *GCD11*, *RPS21B*, orf19.6882.1, *RPS15*, *RPS7A*, *MMD1*, *HCR1*, *RPS8A*, *TIF*, *YST1*, *SUP35*, *RPS12*, *GUS1*, *RPL6*, *RPL4B*, *RPS6A*, *FRS1*, *RPS5*, *KRS1*, *RPS16A*, *RPL25*, *RPL9B*, *RPL40B*, *RPS22A*, *CDC60*, *UBI3*, *RPL10A*, orf19.3572.3, *TEF1*, *RPL38*, *EFB1*, *THS1*, *RPP0*, *TIF34*, *RPP1B*, *GCD6*, orf19.2478.1, *RPL11*

[a] Log odds ratios and adjusted *P* values were obtained by performing GO term enrichment analysis with GOEAST (89). Only selected GO terms are listed.
[b] Corresponding GO term numbers are given in parentheses.
[c] Enrichment fraction was obtained by the ratio between the gene lists and the total number of genes present in a given GO term.

"FC_Galleria_24h"), which is in agreement with other mouse studies (57, 58) but extends this important fungal pathogenesis mechanism to the insect mini-host model.

Even if the global trend of regulation is similar between time points and the two hosts, some genes were inversely regulated between early and late stages of infection. One interesting gene among them is *WOR1*, a regulator of the white-opaque switch in *C. albicans*, which is upregulated at early time points but downregulated at late time points. The white-opaque switch, besides its role in facilitating mating between opposite mating types, is coupled with the expression of several genes involved in adhesion, drug resistance, and metabolism (59). *WOR1* expression is critical for this switch in homozygous **a** or α cells; however, ectopic expression of *WOR1* in **a**/α cells can also induce switching, though with limited capacity. In our experiments, we observed that *WOR1* is 6-fold upregulated in mice at the early time point compared to *in vitro* conditions but then 28-fold downregulated at the late time point. It was recently reported that *WOR1* can be expressed in **a**/α cells upon passage in the gastrointestinal tract (60); however, this has not been reported yet for kidneys. *WOR1* upregulation triggers the expression of several genes important for *C. albicans* gut colonization (60) and is also involved in biofilm formation (61). The *WOR1* inverse regulation was surprising, since most of the opaque genes are upregulated in mice at both time points. This suggests that other morphogenesis regulators may compensate for sustained expression of the opaque genes.

LAP3 (encoding a putative aminopeptidase) was among other genes inversely regulated between time points in *G. mellonella* (Fig. 3a). Interestingly, *LAP3* is among the genes positively regulated by *SEF1* (56). Considering the permanent upregulation of *SEF1*, this suggests that *LAP3* is under the control of another, yet-unknown regulator(s). *POX1-3* (encoding acyl coenzyme A

oxidase enriched in stationary phase) and *HGT3* (encoding a high-affinity glucose transporter) are inversely regulated between time points in mice (Fig. 3b) and thus reflect metabolic adaptation upon the course of infection in this model.

Host infection transcriptomic signature. The choice of hosts (murine and insect models) was motivated by the idea that minihost models such as *G. mellonella* could replace mammalian models in some experimental infections with *C. albicans* due to cost and ethical reasons. Several laboratories have investigated fungal virulence and antifungal drug activity with *G. mellonella* (62–67). The results obtained from this insect model were consistent with those obtained from the mouse systemic model of infection (62, 64; reviewed in reference 40) and with the pathogenicity of *C. albicans* strains in human patients (68). The immune system of *G. mellonella* can be compared with the innate immunity of mammals, and the larval immune response to microorganisms can be assessed based on antimicrobial peptide (AMP) production or hemocyte counts (reviewed in reference 40). The infection models used here are evolutionarily highly divergent, but *C. albicans* transcriptomic responses in these hosts showed significant correlations. A data mining approach combining time points and hosts allowed us to delimit on one hand an *in vivo*-specific gene expression signature (1,169 genes), independent of the infected host, compared to *in vitro*-grown cells, but on the other hand highlighted some highly significant host-specifically regulated genes.

Considering the 1,169 *in vivo* regulated genes, we performed two types of data mining analysis, GO term analysis and GSEA. Interestingly, GO term analysis highlighted genes in the category "induction of host defense response," including *CDC19*, *FBA1*, *TPI1*, *ENO1*, *TDH3*, *PGK1*, *SSB1*, *ADH1*, and *IMH3*. The products of these genes are antigenic in animal models (69–71), and therefore, their downregulation *in vivo* can be taken as a fungal strategy to diminish adaptive host defenses. GSEA showed that the set of 1,169 *in vivo* specific genes extensively overlapped previously published genome-wide transcriptional analysis of *C. albicans* exposed to *in vitro* conditions mimicking the host environment. Moreover, the set of *in vivo*-regulated genes also overlapped data sets originating from transcriptional analysis from host samples, including zebrafish, mouse, and human samples. The conserved expression profile of *C. albicans* found here for mice and *G. mellonella* suggests that the pathogenesis mechanisms of *C. albicans* may be similar in both hosts. On the other hand, since *C. albicans* is, as far as the literature reports, not a natural pathogen of insects, the mammal-specific host response of *C. albicans* may be fortuitously sufficient for pathogenesis in *G. mellonella*. In any case, absence of important virulence factors in *C. albicans* may result in similar disease outcomes in both hosts. As mentioned earlier, several studies have concluded that the response of mice to specific *C. albicans* mutants can be phenocopied in *G. mellonella* infections. Our transcriptional data showed for the first time at the transcriptional level this convergence between the two animal models.

However, our data also highlight a few host-specific characteristics of the *C. albicans* transcriptional response. Among mouse-specific and upregulated genes are *MNN4-4* (encoding a mannosyltransferase) and *DAG7* (encoding a secreted protein; both are downregulated in mice but upregulated in *G. mellonella*). While the putative function of *DAG7* is unknown, it contains a barwin-like endoglucanase domain (IPR014733). *MNN4-4* and *DAG7* may therefore modify cell wall composition in *C. albicans* according to the specific host environment. In addition, *TLO1* stands out as a gene with an inverse regulation in *G. mellonella*. *TLO1* is known as a member of the telomere-proximal genes with homology to the Med2 subunit of Mediator (72). *TLO1* from a closely related species (*C. dubliniensis*) was shown to be critical for expression of genes relevant to virulence (adhesins, iron acquisition genes, amino acid catabolic genes, etc.), thus suggesting that *TLO1* in *C. albicans* may have similar functions (52). The difference of expression of *TLO1* between the two hosts investigated may suggest that this gene has a niche-specific role. Accordingly, the list of *Galleria*-specific genes contains several *TLO* genes (*CTA24/TLO12*, *CTA2/TLO3*, and *TLO1*). The downregulation of these other *TLO* genes in *G. mellonella* may reflect a specialized role for TLO family members. From the list of *Galleria*-specific genes, we noticed that several cell wall-associated genes (*PGA57*, *PGA50*, and *PGA1*), genes encoding amino acid and amine transporters (*GAP2* and *TPO5*), and genes encoding 3 different cyclins (*PCL1*, *PCL2*, and *CLN3*) were upregulated. While differential regulation of cell wall genes may be associated with cell wall modifications, the specific regulation of transporters may be understood as alterations in nutrient acquisition. The regulation of both *PCL1* (encoding a cyclin that was shown to be critical for hypha formation at high temperature or HSP90 inhibition) and *CLN3* (encoding a cyclin required for yeast and hyphal growth) suggests a more dynamic morphogenesis in *G. mellonella* than in mice at the time points investigated (73, 74). Among downregulated *Galleria*-specific genes, several genes involved in the control of gene expression are present, including *TYE7* (a bHLH [basic Helix-Loop-Helix] transcription factor controlling glycolysis) (75). Consistent with this observation, several glycolytic genes, such as *PDX1* (encoding pyruvate dehydrogenase complex protein), *PYC2* (encoding pyruvate carboxylase), *SHA3* (encoding a Ser/Thr kinase involved in glucose transport), *OSM1* (encoding a flavoprotein subunit of fumarate reductase) and *HXK2* (encoding hexokinase II), are also *Galleria*-specific downregulated genes.

Finally, we compared our results with a recent study by Xu et al. (24) in which the expression of a small set of genes (248) was measured from kidney lysates taken at early (12 h) and late (48 h) infection time points. We could match a total of 85 genes between the two studies, thus including 49% of the regulated genes identified in the work of Xu et al. (24). As shown in Fig. 6, early gene expression patterns were highly concordant between both studies ($r = 0.89$). Given that time points designated "early" differ between the two studies (12 h versus 16 h), these correlation coefficients highlight a good agreement. This observation further strengthens the validation of the method chosen here for RNA enrichment.

Dual RNA-seq of host and pathogen. Dual RNA-seq of host and pathogen during infection would be optimal for maximizing the utility of transcriptomics studies. However, as mentioned earlier, this is often limited by the low proportions of pathogen in the host tissues. Two recent studies have attempted dual RNA-seq with *C. albicans*-infected host samples. For example, Liu et al. (38) sequenced the transcriptomes of infected mouse kidneys but could draw conclusions only on the host side. Bruno et al. (39) studied a murine vulvovaginal candidiasis (VVC) model, but the reads mapped to *C. albicans* constituted less than 0.1% of the samples. As a result, the authors could not conduct a genome-wide analysis of *C. albicans* but could analyze only 52 genes. On the other hand, both of these studies were highly successful in

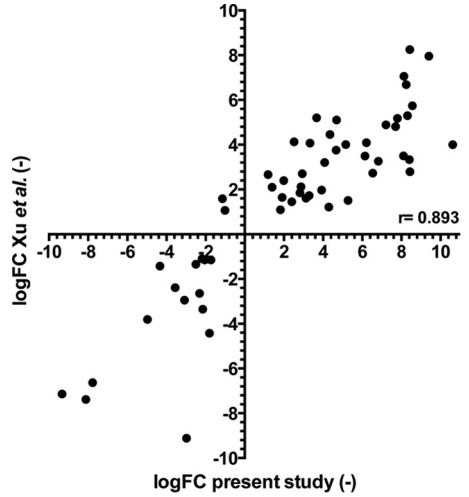

FIG 6 Correlations between $\log_2$ fold change (logFC) data generated in this study and from reference 24, taking early gene expression patterns from both studies (Xu et al. [24], mouse kidneys 12 h p.i. versus *in vitro* stationary-phase culture; Amorim-Vaz et al. [62], mouse kidneys 16 h p.i. versus *in vitro* exponential-phase culture). *r*, Pearson correlation coefficient, calculated with Prism 6.0.

analyzing the host response to the pathogen. In the present work, we were able to comprehensively examine the transcriptional profile of the pathogen after applying the enrichment procedure. Enriched samples still contained 10 to 50% host sequences. However, we observed a nonnegligible bias for the mouse transcriptome when comparing enriched and nonenriched samples (data not shown). We concluded that enriched RNA-seq libraries can be used only to analyze the pathogen response but that the same RNA extracts can be used to prepare nonenriched libraries and therefore to study both pathogen and host in any context of infection. This will allow the dissection of the host-pathogen cross talk at a transcriptional level in more detail.

Conclusions. The RNA enrichment technology used here was first designed to target the human exome and is now widely used for this purpose. The enrichment technology used here has, to our knowledge, not been used until now to enrich for microbial RNA from host tissues. In principle, the method could be implemented in other types of microbial systems in which the microbial RNA is found in small amounts in host samples. With regard to *C. albicans* biology, this method can be used to analyze its kinetic of infection at a transcriptional level in other organs besides kidneys. On the other hand, the enrichment procedure may help the enrichment of RNA from mutants with virulence and/or colonization defects in specific hosts. In the future, we aim to enrich for RNA of *C. albicans* mutants with decreased virulence, and even if fungal RNA is even further diluted by host RNA, the analysis of mutant transcriptomes may be still be possible. These experiments are currently undertaken in our laboratory. Moreover, fungal RNA enrichments can be envisaged on human biopsy samples. This would indeed contribute to the analysis of fungal transcriptional response directly in the context of human infection, since our understanding is now limited to experimental systems that only partially reflect human disease.

MATERIALS AND METHODS

Mouse infections and organ collection. All animal experiments were performed at the University of Lausanne and at the University Hospital Centre under the surveillance and with the approval of the institutional Animal Use Committee, Affaires Vétérinaires du Canton de Vaud, Switzerland, authorization no. 1734.3, according to decree 18 of the federal law on animal protection. Infection of 6-week-old BALB/c mice (*Mus musculus*) was performed as described previously (62). Part of the *in vitro* cell culture used for infection was saved for RNA extraction, constituting the *in vitro* samples used for the RNA-seq analysis. Six to eight mice were injected through the tail vein with 250 μl of cell suspension, and two mice were injected with phosphate-buffered saline (PBS; 137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4) (mock infection) in each experiment. At 16 h and 48 h postinfection (p.i.), 3 or 4 infected mice and one mock-infected mouse were randomly chosen (no blinding) and sacrificed (see Fig. S1 in the supplemental material). Kidneys were collected, immediately halved, and placed in vials containing 1 ml of RNAlater solution (Life Technologies). This reagent immediately stabilizes RNA in tissue samples to preserve the gene expression profile. Samples were kept on ice and then at −80°C until the time of RNA extraction. This experiment was performed twice. The number of animals used was chosen so that each sample represented an average of 3 animals, to reduce interindividual noise.

***G. mellonella* infections.** As previously described (62), *G. mellonella* larvae were purchased from Bait Express GmbH (Basel, Switzerland). Upon arrival, the larvae were stored at 12°C in the dark with wood shavings, and larvae were used within a maximum of 2 weeks. Larvae weighing 300 to 400 mg were used for the experiments. *C. albicans* SC5314 strain was grown overnight under agitation at 30°C in yeast extract-peptone-dextrose (YEPD), then diluted 100-fold in YEPD, and grown to an approximate density of 1.5×10^7 cells/ml (measured by optical density). Cultures were then washed twice in PBS and resuspended in 5 ml PBS. The concentration of the culture was measured by optical density, and the culture diluted in PBS to 1.25×10^7 cells/ml. Part of the *in vitro* cell culture was saved for RNA extraction, constituting the *in vitro* samples used for the RNA-seq analysis. Six to eight larvae were injected through the last left proleg, using a Myjector U-100 insulin syringe (Terumo, Europe), with 40 μl of cell suspension, and two larvae were injected with PBS (mock infection). Larvae were kept at 30°C in the dark. At 2 h and 24 h p.i., 3 or 4 infected larvae and one mock-infected larva were randomly chosen (no blinding), sacrificed (see Fig. S1 in the supplemental material), and directly used for RNA extraction. This experiment was performed twice. The number of animals used was chosen so that each sample represented an average of 3 animals, to reduce interindividual noise.

RNA extraction. When cell suspensions were prepared for infection of mice or larvae, 50 ml of the 1.5×10^7cell/ml suspensions was kept for direct RNA extraction of the *in vitro* culture. RNA was extracted from *in vitro* cultures, mouse kidneys, and *G. mellonella* larvae as previously described (62). The list of RNA extracts and the corresponding conditions are listed in Table S2 in the supplemental material. After analysis of RNA quality (see below), the 3 or 4 RNA samples from the same animal species and same time point were combined and further analyzed as a single sample (see Table S2 in the supplemental material). Therefore, each final RNA sample constituted an average of 3 or 4 biological replicates (see Fig. S1 in the supplemental material). This was done in duplicate.

Analysis of RNA integrity. RNA quality was analyzed in a 2100 Bioanalyzer system (Agilent Technologies) according to manufacturer's instructions. RNA was denatured at 70°C for 2 min prior to analysis. Since it was not always possible to calculate an RNA integrity value (RIN), samples were included in the study according to visual examination of the Bioanalyzer profiles (two clear peaks for 18S and 25/28 S for *C. albicans* and mouse samples and one clear peak at 18S for *G. mellonella* samples, with no obvious degradation).

qPCR quantification of *C. albicans* transcripts. (i) Estimation of proportion of fungal RNA in the total *in vivo*-extracted RNA samples infected with different inocula and at different time points. Reverse transcriptase reactions were carried out as previously described (62). This

cDNA was used to determine the percentage of *C. albicans* transcripts in each sample by real-time quantitative PCR (qPCR) targeting *ACT1* of *C. albicans* using primers ACT1-RT-F and ACT1-RT-R and a TaqMan probe ACT1-RT-P (see Table S3 in the supplemental material for details on the primers and probe used). Quantitative PCR was performed in a StepOnePlus real-time PCR system (Applied Biosystems) instrument. Different concentrations of *in vitro*-extracted *C. albicans* cDNA were used in each qPCR to establish a calibration curve, which was then used to calculate the percentage of fungal RNA in the *in vivo*-extracted RNA samples.

(ii) Validation of RNA sequencing data. To validate RNA-seq data, expression levels of 8 genes (*DAG7*, *HWP1*, *TLO1*, *TRY5*, *UPC2*, *YWP1*, *ZRT1*, and orf19.7455) were determined by qPCR. One microgram (determined with a NanoDrop 1000 spectrophotometer; Thermo, Fisher Scientific) of the RNA used for RNA-seq was reverse transcribed using random hexamers as a priming method (Transcriptor high-fidelity cDNA synthesis kit; Roche). Subsequent qPCRs were performed with a 0.2 μM concentration of each primer and a 0.2 μM concentration of probe for *ACT1*, *DAG7*, *HWP1*, *TLO1*, *UPC2*, and *ZRT1* or a 0.4 μM concentration of each primer and a 0.1 μM concentration of probe for *TRY5* and *YWP1* and iTAQ Supermix with ROX (Amine-reactive carboxy-x-rhodamine) (BioRad, Reinach, Switzerland) according to the manufacturer's instructions. Sequences of primers and probes are shown in Table S3 in the supplemental material. Three standard curves were calculated for each gene: one using cDNA from *in vitro*-grown *C. albicans* (used to calculate gene expression levels of *in vitro C. albicans* samples), one using cDNA from noninfected mice spiked with 1% of cDNA from *in vitro*-grown *C. albicans* (used to calculate gene expression levels in samples from infected mice), and one using cDNA from noninfected *G. mellonella* spiked with 1% of cDNA from *in vitro*-grown *C. albicans* (used to calculate gene expression levels in samples from infected larvae). The expression level of *ACT1* was used for normalization. All reactions were repeated twice.

Initial probe design. Nonoverlapping head-to-tail 120-nucleotide probes were designed using the eArray software (Agilent Technologies, Santa Clara, CA). A total of 55,342 bait probes were designed to cover 6,094 *C. albicans* ORFs (assembly 21 SC5314; 16-Mb haploid genome; 6,218 ORFs). The first 250 nucleotides of each gene were not covered in the bait design, resulting in an average of 9 probes for each ORF. Using Megablast (v2.2.26) (76), it was verified that all genes of *C. albicans* were matched by at least one probe and that only a negligible fraction of the probes could be mapped on the mouse and human cDNA sequences from Ensembl and on the available *G. mellonella* expressed sequence tags (ESTs) and cDNA sequences from NCBI.

Feature selection and gene classification. We used Perl and R to compute the numeric gene features and carry out the analysis. The Dustmasker module from NCBI-BLAST (v2.2.29) (77) was used to acquire the low-complexity regions. RNALfold (v2.1.8) (78) was applied to calculate the folding free energy. The Megablast module from BLAST (v2.2.26) (76) was used to align the probes against themselves and to form clusters of similar probes, and the number of sequences clustered together was taken as a measure of redundancy.

For the feature selection process, we investigated five different search methods, with the correlation-based feature subset evaluator, proposed by the WEKA workbench (v3.6.11) (79): best first, greedy stepwise, linear forward selection, scatter search, and subset size forward selection. We then built a linear weighted support vector machine model (43) with the Kernlab R package (v0.9.19) (80) to assess the importance of the selected features and to classify the genes accordingly.

Preparation of RNA-seq libraries. RNA libraries for RNA-seq were prepared using the SureSelect^{XT} multiplexed sequencing kit with RNA target enrichment for Illumina or the SureSelect multiplexed sequencing kit with strand-specific RNA library preparation for Illumina (Agilent Technologies), for enriched and nonenriched samples, respectively, according to the manufacturer's instructions. Briefly, mRNA was purified by poly(A) capture and enzymatically fragmented. Next, double-stranded cDNA was produced with adapters ligated to both ends. The library was then amplified using provided primers which hybridize to the previously inserted adapters, therefore allowing a linear amplification of all transcripts present in the sample. In the case of nonenriched libraries, RNA-seq indexes were also inserted during this PCR. Each library received a different index (see Table S2 in the supplemental material). This index allows several libraries to be sequenced together (multiplexing), and the index sequence was used to distinguish between samples. For enriched libraries, double-stranded cDNA ligated to adapters was also amplified by PCR according to the manufacturer's instructions and was then incubated at 65°C for 24 h with a set of biotinylated oligonucleotides specifically designed to capture *C. albicans* transcripts (baits), as described above. The hybridized sequences were captured with magnetic streptavidin beads. They were next linearly amplified using provided primers and indexed in a new PCR.

For analysis of *in vitro* samples, RNA from noninfected mouse or *G. mellonella* was spiked with 1% of *in vitro C. albicans* RNA samples. Then, these spiked *in vitro* samples were subjected to the same enrichment procedure as *in vivo* samples.

No batch effect was observed between libraries prepared in different days (see Fig. S4 in the supplemental material).

Before sequencing, libraries were analyzed with a fragment analyzer automated CE system (Advanced Analytical) to assess quality and fragment size and with a Qubit fluorometer (Invitrogen) to determine cDNA concentration. Libraries were kept at −20°C until they were sequenced.

RNA sequencing. Cluster generation was performed with the resulting libraries using the Illumina TruSeq PE cluster kit v3 reagents and sequenced on the Illumina HiSeq 2500 system using TruSeq SBS kit v3 reagents. Sequencing data were processed using Illumina Pipeline software version 1.82. Purity-filtered reads were adapters- and quality-trimmed with Cutadapt (v1.2.1) (81) and filtered for low complexity with Prinseq (v0.20.3) (82). Reads were aligned against *Candida albicans* genome SC5314 version A21-s02-m09-r07 using TopHat2 (v2.0.9) (83). The number of read counts per gene locus was summarized with htseq-count (v0.5.4p3) (84).

Data normalization and differential expression analysis were performed in R (v3.1.1), using Bioconductor packages. The read count data were normalized with the TMM (trimmed mean of M-values) method available in the R Bioconductor package edgeR (85) and subsequently transformed to $\log_2$ counts per million by Voom, a method implemented in the R bioconductor package Limma (86). A linear model with one factor per condition was applied to the transformed data using Limma (87) (see File S1, "normalized_gene_expression" in the supplemental material). The conditions were the following: *in vitro*, 2 h p.i. in *G. mellonella*, 24 h p.i. in *G. mellonella*, 16 h p.i. in *M. musculus*, 48 h p.i. in *M. musculus* (all in duplicate). Four contrasts, representing the difference between *in vivo* conditions and *in vitro* conditions, were extracted from the linear model, resulting in a moderated *t* statistic for every gene in every *in vivo* condition.

A two-step meta-analytical approach was used to identify genes affected *in vivo* regardless of the host model and the time point (see the R scripts in File S3 in the supplemental material). First, as described by Wirapati et al. (88), the *P* values for the two time points of a host were converted to *z* scores taking into account the sign of the *t* statistic and combined meta-analytically into one *z* score per host. Then, the resulting *G. mellonella* and *M. musculus* z scores were again combined meta-analytically into one global *z* score reflecting the chance that a particular gene is affected *in vivo*. *P* values calculated from the meta-analysis were adjusted using Bonferroni corrections, and adjusted *P* values of ≤0.05 were considered significant.

Sequence data accession number. RNA-seq raw data reported here are accessible under the BioProject accession number SRP058281.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.00942-15/-/DCSupplemental.

File S1, XLSX file, 2.7 MB.
File S2, XLSX file, 2.4 MB.
File S3, DOCX file, 0.03 MB.
Figure S1, TIF file, 2.1 MB.
Figure S2, TIF file, 1.5 MB.
Figure S3, TIF file, 1.9 MB.
Figure S4, TIF file, 1.2 MB.
Table S1, DOCX file, 0.01 MB.
Table S2, DOCX file, 0.02 MB.
Table S3, DOCX file, 0.01 MB.

ACKNOWLEDGMENTS

This work was supported by Swiss National Science Foundation grant CRSII3_141848 (Sinergia) to Dominique Sanglard. Sequencing was performed at the Lausanne Genomic Technologies Facility. The computations were performed at the Vital-IT Center for high-performance computing (http://www.vital-it.ch) of the SIB Swiss Institute of Bioinformatics. SIB receives financial supports from the Swiss Federal Government through the State Secretariat for Education and Research (SER).

We thank Floriane Consales for helping with SureSelect library preparation and Sandra Calderon for preparing *C. albicans* genome annotations.

REFERENCES

1. **Odds FC**. 1988. The ecology of Candida and epidemiology of candidosis. Candida and candidosis: a review and bibliography. Balliere Tindall, London, United Kingdom.
2. **Pfaller MA, Diekema DJ**. 2010. Epidemiology of invasive mycoses in North America. Crit Rev Microbiol **36:**1–53. http://dx.doi.org/10.3109/10408410903241444.
3. **Lepak A, Nett J, Lincoln L, Marchillo K, Andes D**. 2006. Time course of microbiologic outcome and gene expression in Candida albicans during and following in vitro and in vivo exposure to fluconazole. Antimicrob Agents Chemother **50:**1311–1319. http://dx.doi.org/10.1128/AAC.50.4.1311-1319.2006.
4. **Barker KS, Crisp S, Wiederhold N, Lewis RE, Bareither B, Eckstein J, Barbuch R, Bard M, Rogers PD**. 2004. Genome-wide expression profiling reveals genes associated with amphotericin B and fluconazole resistance in experimentally induced antifungal resistant isolates of Candida albicans. J Antimicrob Chemother **54:**376–385. http://dx.doi.org/10.1093/jac/dkh336.
5. **Rogers PD, Barker KS**. 2002. Evaluation of differential gene expression in fluconazole-susceptible and -resistant isolates of Candida albicans by cDNA microarray analysis. Antimicrob Agents Chemother **46:**3412–3417. http://dx.doi.org/10.1128/AAC.46.11.3412-3417.2002.
6. **Copping VM, Barelle CJ, Hube B, Gow NA, Brown AJ, Odds FC**. 2005. Exposure of Candida albicans to antifungal agents affects expression of SAP2 and SAP9 secreted proteinase genes. J Antimicrob Chemother **55:**645–654. http://dx.doi.org/10.1093/jac/dki088.
7. **De Backer MD, Ilyina T, Ma XJ, Vandoninck S, Luyten WH, Vanden Bossche H**. 2001. Genomic profiling of the response of Candida albicans to itraconazole treatment using a DNA microarray. Antimicrob Agents Chemother **45:**1660–1670. http://dx.doi.org/10.1128/AAC.45.6.1660-1670.2001.
8. **García-Sánchez S, Aubert S, Iraqui I, Janbon G, Ghigo JM, d'Enfert C**. 2004. Candida albicans biofilms: a developmental state associated with specific and stable gene expression patterns. Eukaryot Cell **3:**536–545. http://dx.doi.org/10.1128/EC.3.2.536-545.2004.
9. **Nobile CJ, Mitchell AP**. 2006. Genetics and genomics of Candida albicans biofilm formation. Cell Microbiol **8:**1382–1391. http://dx.doi.org/10.1111/j.1462-5822.2006.00761.x.
10. **Nantel A, Dignard D, Bachewich C, Harcus D, Marcil A, Bouin A-P, Sensen CW, Hogues H, van het Hoog M, Gordon P, Rigby T, Benoit F, Tessier DC, Thomas DY, Whiteway M**. 2002. Transcription profiling of Candida albicans cells undergoing the yeast-to-hyphal transition. Mol Biol Cell **13:**3452–3465. http://dx.doi.org/10.1091/mbc.E02-05-0272.
11. **Carlisle PL, Kadosh D**. 2013. A genome-wide transcriptional analysis of morphology determination in Candida albicans. Mol Biol Cell **24:**246–260. http://dx.doi.org/10.1091/mbc.E12-01-0065.
12. **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. http://dx.doi.org/10.1091/mbc.E05-05-0435.
13. **Enjalbert B, Nantel A, Whiteway M**. 2003. Stress-induced gene expression in Candida albicans: absence of a general stress response. Mol Biol Cell **14:**1460–1467. http://dx.doi.org/10.1091/mbc.E02-08-0546.
14. **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. http://dx.doi.org/10.1111/j.1365-2958.2004.04350.x.
15. **Fradin C, Kretschmar M, Nichterlein T, Gaillardin C, D'enfert C, Hube B**. 2003. Stage-specific gene expression of Candida albicans in human blood. Mol Microbiol **47:**1523–1543. http://dx.doi.org/10.1046/j.1365-2958.2003.03396.x.
16. **Lorenz MC, Bender JA, Fink GR**. 2004. Transcriptional response of Candida albicans upon internalization by macrophages. Eukaryot Cell **3:**1076–1087. http://dx.doi.org/10.1128/EC.3.5.1076-1087.2004.
17. **Fradin C, De Groot P, Maccallum D, Schaller M, Klis F, Odds FC, Hube B**. 2005. Granulocytes govern the transcriptional response, morphology and proliferation of Candida albicans in human blood. Mol Microbiol **56:**397–415. http://dx.doi.org/10.1111/j.1365-2958.2005.04557.x.
18. **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 U S A **100:**11007–11012. http://dx.doi.org/10.1073/pnas.1834481100.
19. **Zakikhany K, Naglik JR, Schmidt-Westhausen A, Holland G, Schaller M, Hube B**. 2007. In vivo transcript profiling of Candida albicans identifies a gene essential for interepithelial dissemination. Cell Microbiol **9:**2938–2954. http://dx.doi.org/10.1111/j.1462-5822.2007.01009.x.
20. **Martin R, Wächtler B, Schaller M, Wilson D, Hube B**. 2011. Host-pathogen interactions and virulence-associated genes during Candida albicans oral infections. Int J Med Microbiol **301:**417–422. http://dx.doi.org/10.1016/j.ijmm.2011.04.009.
21. **Spiering MJ, Moran GP, Chauvel M, Maccallum DM, Higgins J, Hokamp K, Yeomans T, d'Enfert C, Coleman DC, Sullivan DJ**. 2010. Comparative transcript profiling of Candida albicans and Candida dubliniensis identifies SFL2, a C. albicans gene required for virulence in a reconstituted epithelial infection model. Eukaryot Cell **9:**251–265. http://dx.doi.org/10.1128/EC.00291-09.
22. **Fanning S, Xu W, Solis N, Woolford CA, Filler SG, Mitchell AP**. 2012. Divergent targets of *Candida albicans* biofilm regulator Bcr1 *in vitro* and *in vivo*. Eukaryot Cell **11:**896–904. http://dx.doi.org/10.1128/EC.00103-12.
23. **Cheng S, Clancy CJ, Xu W, Schneider F, Hao B, Mitchell AP, Nguyen MH**. 2013. Profiling of Candida albicans gene expression during intra-abdominal candidiasis identifies biologic processes involved in pathogenesis. J Infect Dis **208:**1529–1537. http://dx.doi.org/10.1093/infdis/jit335.
24. **Xu W, Solis NV, Ehrlich RL, Woolford CA, Filler SG, Mitchell AP**. 2015. Activation and Alliance of Regulatory Pathways in C. albicans during mammalian infection. PLoS Biol **13:**e1002076. http://dx.doi.org/10.1371/journal.pbio.1002076.
25. **Andes D, Lepak A, Pitula A, Marchillo K, Clark J**. 2005. A simple approach for estimating gene expression in Candida albicans directly from a systemic infection site. J Infect Dis **192:**893–900. http://dx.doi.org/10.1086/432104.
26. **Thewes S, Kretschmar M, Park H, Schaller M, Filler SG, Hube B**. 2007. In vivo and ex vivo comparative transcriptional profiling of invasive and non-invasive Candida albicans isolates identifies genes associated with tissue invasion. Mol Microbiol **63:**1606–1628. http://dx.doi.org/10.1111/j.1365-2958.2007.05614.x.
27. **Nett JE, Lepak AJ, Marchillo K, Andes DR**. 2009. Time course global gene expression analysis of an in vivo Candida biofilm. J Infect Dis **200:**307–313. http://dx.doi.org/10.1086/599838.
28. **Rosenbach A, Dignard D, Pierce JV, Whiteway M, Kumamoto CA**. 2010. Adaptations of Candida albicans for growth in the mammalian intestinal tract. Eukaryot Cell **9:**1075–1086. http://dx.doi.org/10.1128/EC.00034-10.
29. **Walker LA, MacCallum DM, Bertram G, Gow NA, Odds FC, Brown AJ**. 2009. Genome-wide analysis of Candida albicans gene expression patterns during infection of the mammalian kidney. Fungal Genet Biol **46:**210–219. http://dx.doi.org/10.1016/j.fgb.2008.10.012.
30. **Chen YY, Chao CC, Liu FC, Hsu PC, Chen HF, Peng SC, Chuang YJ, Lan CY, Hsieh WP, Wong DS**. 2013. Dynamic transcript profiling of

Candida albicans infection in zebrafish: a pathogen-host interaction study. PLoS One **8:**e72483. http://dx.doi.org/10.1371/journal.pone.0072483.

31. **Draghici S, Khatri P, Eklund AC, Szallasi Z.** 2006. Reliability and reproducibility issues in DNA microarray measurements. Trends Genet **22:** 101–109. http://dx.doi.org/10.1016/j.tig.2005.12.005.
32. **Geiss GK, Bumgarner RE, Birditt B, Dahl T, Dowidar N, Dunaway DL, Fell HP, Ferree S, George RD, Grogan T, James JJ, Maysuria M, Mitton JD, Oliveri P, Osborn JL, Peng T, Ratcliffe AL, Webster PJ, Davidson EH, Hood L, Dimitrov K.** 2008. Direct multiplexed measurement of gene expression with color-coded probe pairs. Nat Biotechnol **26:**317–325. http://dx.doi.org/10.1038/nbt1385.
33. **SEQC/MAQC-III Consortium.** 2014. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotech **32:**903–914.
34. **Bruno VM, Wang Z, Marjani SL, Euskirchen GM, Martin J, Sherlock G, Snyder M.** 2010. Comprehensive annotation of the transcriptome of the human fungal pathogen Candida albicans using RNA-seq. Genome Res **20:**1451–1458. http://dx.doi.org/10.1101/gr.109553.110.
35. **Dhamgaye S, Devaux F, Manoharlal R, Vandeputte P, Shah AH, Singh A, Blugeon C, Sanglard D, Prasad R.** 2012. In vitro effect of malachite green on Candida albicans involves multiple pathways and transcriptional regulators UPC2 and STP2. Antimicrob Agents Chemother **56:**495–506. http://dx.doi.org/10.1128/AAC.00574-11.
36. **Hnisz D, Bardet AF, Nobile CJ, Petryshyn A, Glaser W, Schöck U, Stark A, Kuchler K.** 2012. A histone deacetylase adjusts transcription kinetics at coding sequences during *Candida albicans* morphogenesis. PLoS Genet **8:**e1003118. http://dx.doi.org/10.1371/journal.pgen.1003118.
37. **Tierney L, Linde J, Müller S, Brunke S, Molina JC, Hube B, Schöck U, Guthke R, Kuchler K.** 2012. An interspecies regulatory network inferred from simultaneous RNA-seq of Candida albicans invading innate immune cells. Front Microbiol **3:**85. http://dx.doi.org/10.3389/fmicb.2012.00085.
38. **Liu Y, Shetty AC, Schwartz JA, Bradford LL, Xu W, Phan QT, Kumari P, Mahurkar A, Mitchell AP, Ravel J, Fraser CM, Filler SG, Bruno VM.** 2015. New signaling pathways govern the host response to C. albicans infection in various niches. Genome Res **25:**679–689. http://dx.doi.org/10.1101/gr.187427.114.
39. **Bruno VM, Shetty AC, Yano J, Fidel PL, Noverr MC, Peters BM.** 2015. Transcriptomic analysis of vulvovaginal candidiasis identifies a role for the NLRP3 Inflammasome. mBio **6:**e00182-15. http://dx.doi.org/10.1128/mBio.00182-15.
40. **Coste AT, Amorim-Vaz S.** 2015. Animal models to study fungal virulence and antifungal drugs, p 289–315. *In* Coste AT, Vandeputte P (ed), Antifungals: from genomics to resistance and the development of novel agents. Caister Academic Press, Norfolk, United Kingdom.
41. **Saeys Y, Inza I, Larrañaga P.** 2007. A review of feature selection techniques in bioinformatics. BioInformatics **23:**2507–2517. http://dx.doi.org/10.1093/bioinformatics/btm344.
42. **Butler G, Rasmussen MD, Lin MF, Santos MA, Sakthikumar S, Munro CA, Rheinbay E, Grabherr M, Forche A, Reedy JL, Agrafioti I, Arnaud MB, Bates S, Brown AJ, Brunke S, Costanzo MC, Fitzpatrick DA, de Groot PW, Harris D, Hoyer LL, Hube B, Klis FM, Kodira C, Lennard N, Logue ME, Martin R, Neiman AM, Nikolaou E, Quail MA, Quinn J, Santos MC, Schmitzberger FF, Sherlock G, Shah P, Silverstein KA, Skrzypek MS, Soll D, Staggs R, Stansfield I, Stumpf MP, Sudbery PE, Srikantha T, Zeng Q, Berman J, Berriman M, Heitman J, Gow NA, Lorenz MC, Birren BW, Kellis M.** 2009. Evolution of pathogenicity and sexual reproduction in eight Candida genomes. Nature **459:**657–662. http://dx.doi.org/10.1038/nature08064.
43. **Vapnik V.** 1998. Statistical learning theory. Wiley Interscience, New York, NY.
44. **Mayer FL, Wilson D, Hube B.** 2013. Candida albicans pathogenicity mechanisms. Virulence **4:**119–128. http://dx.doi.org/10.4161/viru.22913.
45. **Zheng X, Wang Y, Wang Y.** 2004. Hgc1, a novel hypha-specific G1 cyclin-related protein regulates Candida albicans hyphal morphogenesis. EMBO J **23:**1845–1856. http://dx.doi.org/10.1038/sj.emboj.7600195.
46. **Banerjee M, Thompson DS, Lazzell A, Carlisle PL, Pierce C, Monteagudo C, López-Ribot JL, Kadosh D.** 2008. UME6, a novel filament-specific regulator of Candida albicans hyphal extension and virulence. Mol Biol Cell **19:**1354–1365. http://dx.doi.org/10.1091/mbc.E07-11-1110.
47. **Lan CY, Rodarte G, Murillo LA, Jones T, Davis RW, Dungan J, Newport G, Agabian N.** 2004. Regulatory networks affected by iron availability in Candida albicans. Mol Microbiol **53:**1451–1469. http://dx.doi.org/10.1111/j.1365-2958.2004.04214.x.
48. **Almeida RS, Wilson D, Hube B.** 2009. Candida albicans iron acquisition within the host. FEMS Yeast Res **9:**1000–1012. http://dx.doi.org/10.1111/j.1567-1364.2009.00570.x.
49. **Wächtler B, Wilson D, Haedicke K, Dalle F, Hube B.** 2011. From attachment to damage: defined genes of Candida albicans mediate adhesion, invasion and damage during interaction with oral epithelial cells. PLoS One **6:**e17046. http://dx.doi.org/10.1371/journal.pone.0017046.
50. **Phan QT, Myers CL, Fu Y, Sheppard DC, Yeaman MR, Welch WH, Ibrahim AS, Edwards JE, Filler SG.** 2007. Als3 is a Candida albicans invasin that binds to cadherins and induces endocytosis by host cells. PLoS Biol **5:**e64. http://dx.doi.org/10.1371/journal.pbio.0050064.
51. **Leidich SD, Ibrahim AS, Fu Y, Koul A, Jessup C, Vitullo J, Fonzi W, Mirbod F, Nakashima S, Nozawa Y, Ghannoum MA.** 1998. Cloning and disruption of caPLB1, a phospholipase B gene involved in the pathogenicity of Candida albicans. J Biol Chem **273:**26078–26086. http://dx.doi.org/10.1074/jbc.273.40.26078.
52. **Haran J, Boyle H, Hokamp K, Yeomans T, Liu Z, Church M, Fleming AB, Anderson MZ, Berman J, Myers LC, Sullivan DJ, Moran GP.** 2014. Telomeric ORFs (TLOs) in Candida spp. encode mediator subunits that regulate distinct virulence traits. PLoS Genet **10:**e1004658. http://dx.doi.org/10.1371/journal.pgen.1004658.
53. **Nobile CJ, Fox EP, Nett JE, Sorrells TR, Mitrovich QM, Hernday AD, Tuch BB, Andes DR, Johnson AD.** 2012. A recently evolved transcriptional network controls biofilm development in Candida albicans. Cell **148:**126–138. http://dx.doi.org/10.1016/j.cell.2011.10.048.
54. **Finkel JS, Xu W, Huang D, Hill EM, Desai JV, Woolford CA, Nett JE, Taff H, Norice CT, Andes DR, Lanni F, Mitchell AP.** 2012. Portrait of Candida albicans adherence regulators. PLoS Pathog **8:**e1002525. http://dx.doi.org/10.1371/journal.ppat.1002525.
55. **Sudbery PE.** 2011. Growth of Candida albicans hyphae. Nat Rev Microbiol **9:**737–748. http://dx.doi.org/10.1038/nrmicro2636.
56. **Chen C, Pande K, French S, Tuch B, Noble S.** 2011. An iron homeostasis regulatory circuit with reciprocal roles in Candida albicans commensalism and pathogenesis. Cell Host Microbes **10:**118–135. http://dx.doi.org/10.1016/j.chom.2011.07.005.
57. **Chen C, Noble SM.** 2012. Post-transcriptional regulation of the Sef1 transcription factor controls the virulence of Candida albicans in its mammalian host. PLoS Pathog **8:**e1002956. http://dx.doi.org/10.1371/journal.ppat.1002956.
58. **Noble SM.** 2013. Candida albicans specializations for iron homeostasis: from commensalism to virulence. Curr Opin Microbiol **16:**708–715. http://dx.doi.org/10.1016/j.mib.2013.09.006.
59. **Zordan RE, Galgoczy DJ, Johnson AD.** 2006. Epigenetic properties of white-opaque switching in Candida albicans are based on a self-sustaining transcriptional feedback loop. Proc Natl Acad Sci U S A **103:**12807–12812. http://dx.doi.org/10.1073/pnas.0605138103.
60. **Pande K, Chen C, Noble SM.** 2013. Passage through the mammalian gut triggers a phenotypic switch that promotes *Candida albicans* commensalism. Nat Genet **45:**1088–1091. http://dx.doi.org/10.1038/ng.2710.
61. **Fox EP, Cowley ES, Nobile CJ, Hartooni N, Newman DK, Johnson AD.** 2014. Anaerobic bacteria grow within *Candida albicans* biofilms and induce biofilm formation in suspension cultures. Curr Biol **24:**2411–2416. http://dx.doi.org/10.1016/j.cub.2014.08.057.
62. **Amorim-Vaz S, Delarze E, Ischer F, Sanglard D, Coste AT.** 2015. Examining the virulence of *Candida albicans* transcription factor mutants using Galleria mellonella and mouse infection models. Front Microbiol **6:**367. http://dx.doi.org/10.3389/fmicb.2015.00367.
63. **Favre-Godal Q, Dorsaz S, Queiroz EF, Conan C, Marcourt L, Wardojo BP, Voinesco F, Buchwalder A, Gindro K, Sanglard D, Wolfender JL.** 2014. Comprehensive approach for the detection of antifungal compounds using a susceptible strain of Candida albicans and confirmation of in vivo activity with the Galleria mellonella model. Phytochemistry **105:** 68–78. http://dx.doi.org/10.1016/j.phytochem.2014.06.004.
64. **Brennan M, Thomas DY, Whiteway M, Kavanagh K.** 2002. Correlation between virulence of Candida albicans mutants in mice and Galleria mellonella larvae. FEMS Immunol Med Microbiol **34:**153–157. http://dx.doi.org/10.1111/j.1574-695X.2002.tb00617.x.
65. **Mylonakis E, Moreno R, El Khoury JB, Idnurm A, Heitman J, Calderwood SB, Ausubel FM, Diener A.** 2005. Galleria mellonella as a model system to study Cryptococcus neoformans pathogenesis. Infect Immun **73:**3842–3850. http://dx.doi.org/10.1128/IAI.73.7.3842-3850.2005.

66. **Thomaz L, García-Rodas R, Guimarães AJ, Taborda CP, Zaragoza O, Nosanchuk JD**. 2013. Galleria mellonella as a model host to study Paracoccidioides lutzii and Histoplasma capsulatum. Virulence **4:**139–146. http://dx.doi.org/10.4161/viru.23047.
67. **Coleman JJ, Muhammed M, Kasperkovitz PV, Vyas JM, Mylonakis E**. 2011. Fusarium pathogenesis investigated using Galleria mellonella as a heterologous host. Fungal Biol **115:**1279–1289. http://dx.doi.org/10.1016/j.funbio.2011.09.005.
68. **Cotter G, Doyle S, Kavanagh K**. 2000. Development of an insect model for the in vivo pathogenicity testing of yeasts. FEMS Immunol Med Microbiol **27:**163–169. http://dx.doi.org/10.1111/j.1574-695X.2000.tb01427.x.
69. **Titz B, Thomas S, Rajagopala SV, Chiba T, Ito T, Uetz P**. 2006. Transcriptional activators in yeast. Nucleic Acids Res **34:**955–967. http://dx.doi.org/10.1093/nar/gkj493.
70. **Fernández-Arenas E, Molero G, Nombela C, Diez-Orejas R, Gil C**. 2004. Low virulent strains of Candida albicans: unravelling the antigens for a future vaccine. Proteomics **4:**3007–3020. http://dx.doi.org/10.1002/pmic.200400929.
71. **Pitarch A, Abian J, Carrascal M, Sánchez M, Nombela C, Gil C**. 2004. Proteomics-based identification of novel Candida albicans antigens for diagnosis of systemic candidiasis in patients with underlying hematological malignancies. Proteomics **4:**3084–3106. http://dx.doi.org/10.1002/pmic.200400903.
72. **Anderson MZ, Baller JA, Dulmage K, Wigen L, Berman J**. 2012. The three clades of the telomere-associated TLO gene family of Candida albicans have different splicing, localization, and expression features. Eukaryot Cell **11:**1268–1275. http://dx.doi.org/10.1128/EC.00230-12.
73. **Shapiro RS, Sellam A, Tebbji F, Whiteway M, Nantel A, Cowen LE**. 2012. Pho85, Pcl1, and Hms1 signaling governs Candida albicans morphogenesis induced by high temperature or Hsp90 compromise. Curr Biol **22:**461–470. http://dx.doi.org/10.1016/j.cub.2012.01.062.
74. **Chapa y Lazo B, Bates S, Sudbery P**. 2005. The G1 cyclin Cln3 regulates morphogenesis in Candida albicans. Eukaryot Cell **4:**90–94. http://dx.doi.org/10.1128/EC.4.1.90-94.2005.
75. **Sellam A, van het Hoog M, Tebbji F, Beaurepaire C, Whiteway M, Nantel A**. 2014. Modeling the transcriptional regulatory network that controls the early hypoxic response in Candida albicans. Eukaryot Cell **13:**675–690. http://dx.doi.org/10.1128/EC.00292-13.
76. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ**. 1990. Basic local alignment search tool. J Mol Biol **215:**403–410. http://dx.doi.org/10.1016/S0022-2836(05)80360-2.
77. **Morgulis A, Gertz EM, Schäffer AA, Agarwala R**. 2006. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J Comput Biol **13:**1028–1040. http://dx.doi.org/10.1089/cmb.2006.13.1028.
78. **Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL**. 2011. ViennaRNA package 2.0. Algorithms Mol Biol **6:**26. http://dx.doi.org/10.1186/1748-7188-6-26.
79. **Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH**. 2009. The WEKA data mining software. ACM SIGKDD Explor Newsl **11:**10. http://dx.doi.org/10.1145/1656274.1656278.
80. **Karatzoglou A, Smola A, Hornik K, Zeileis A**. 2004. Kernlab—an S4 package for kernel methods in R. J Stat Softw **11:**1–20.
81. **Martin M**. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EmBnet J **17:**10. http://dx.doi.org/10.14806/ej.17.1.200.
82. **Schmieder R, Edwards R**. 2011. Quality control and preprocessing of metagenomic datasets. BioInformatics **27:**863–864. http://dx.doi.org/10.1093/bioinformatics/btr026.
83. **Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL**. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol **14:**R36. http://dx.doi.org/10.1186/gb-2013-14-4-r36.
84. **Anders S, Pyl PT, Huber W**. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. BioInformatics **31:**166–169. http://dx.doi.org/10.1093/bioinformatics/btu638.
85. **Robinson MD, Oshlack A**. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol **11:**R25. http://dx.doi.org/10.1186/gb-2010-11-3-r25.
86. **Law CW, Chen Y, Shi W, Smyth GK**. 2014. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol **15:**R29. http://dx.doi.org/10.1186/gb-2014-15-2-r29.
87. **Smyth GK**. 2005. Limma: linear models for microarray data, p 397–420. *In* Gentleman R, Carey V, Dudoit S, Irizarry R, Huber H (ed), Bioinformatics and computational biology solutions using R and Bioconductor. Springer, New York, NY.
88. **Wirapati P, Sotiriou C, Kunkel S, Farmer P, Pradervand S, Haibe-Kains B, Desmedt C, Ignatiadis M, Sengstag T, Schütz F, Goldstein DR, Piccart M, Delorenzi M**. 2008. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast Cancer Res **10:**R65. http://dx.doi.org/10.1186/bcr2124.
89. **Zheng Q, Wang XJ**. 2008. GOEAST: a web-based software toolkit for gene ontology enrichment analysis. Nucleic Acids Res **36:**W358–W363. http://dx.doi.org/10.1093/nar/gkn276.