APPLICATION OF AN OPTIMIZED ANNOTATION PIPELINE TO THE *CRYPTOCOCCUS DEUTEROGATTII* GENOME REVEALS DYNAMIC PRIMARY METABOLIC GENE CLUSTERS AND GENOMIC IMPACT OF RNAI LOSS

Patrícia Aline GRÖHS FERRAREZE[*,‡,] Corinne MAUFRAIS[*,†], Rodrigo SILVA ARAUJO STREIT[‡], Shelby J. PRIEST[§], Christina CUOMO[**], Joseph HEITMAN[§], Charley Christian STAATS[‡,2], Guilhem JANBON[*,1,2]

* Institut Pasteur, Unité Biologie des ARN des Pathogènes Fongiques, Département de Mycologie, F-75015, Paris, France

† Institut Pasteur, HUB Bioinformatique et Biostatistique, C3BI, USR 3756 IP CNRS, F-75015, Paris, France

‡ Programa de Pós-Graduação em Biologia Celular e Molecular, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil.

§ Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, NC, 27710, USA

** Broad Institute of MIT and Harvard, Cambridge, MA 02142

1 Corresponding author

2. Both authors should be considered as senior authors

Running Tittle: *C. deuterogattii* genome annotation

Key words: *Cryptococcus deuterogattii*; Genome annotation pipeline; RNAi; Metabolic gene cluster

Address of the corresponding author:
Guilhem Janbon
Institut Pasteur, Unité Biologie des ARN des Pathogènes Fongiques, Département de Mycologie, 25 rue du Dr Roux,75015, Paris, France

Abstract

Evaluating the quality of a *de novo* annotation of a complex fungal genome based on RNA-seq data remains a challenge. In this study, we sequentially optimized a Cufflinks-CodingQuary based bioinformatics pipeline fed with RNA-seq data using the manually annotated model pathogenic yeasts *Cryptococcus neoformans* and *Cryptococcus deneoformans* as test cases. Our results demonstrate that the quality of the annotation is sensitive to the quantity of RNA-seq data used and that the best quality is obtained with 5 to 10 million reads per RNA-seq replicate. We also demonstrated that the number of introns predicted is an excellent *a priori* indicator of the quality of the final *de novo* annotation. We then used this pipeline to annotate the genome of the RNAi-deficient species *Cryptococcus deuterogattii* strain R265 using RNA-seq data. Dynamic transcriptome analysis revealed that intron retention is more prominent in *C. deuterogattii* than in the other RNAi-proficient species *C. neoformans* and *C. deneoformans*. In contrast, we observed that antisense transcription was not higher in *C. deuterogattii* than in the two other *Cryptococcus* species. Comparative gene content analysis identified 21 clusters enriched in transcription factors and transporters that have been lost. Interestingly, analysis of the subtelomeric regions in these three annotated species identified a similar gene enrichment, reminiscent of the structure of primary metabolic clusters. Our data suggest that there is active exchange between subtelomeric regions, and that other chromosomal regions might participate in adaptive diversification of *Cryptococcus* metabolite assimilation potential.

Introduction

In recent years, we have seen an astonishing multiplication of fungal genome sequences (JAMES *et al.* 2020). Long-read sequencing and adapted bioinformatics tools are quickly improving as well. It is expected that telomere-to-telomere whole-genome sequencing will soon become standard for reference genomes of diverse organisms (GIORDANO *et al.* 2017; DAL MOLIN *et al.* 2018; YADAV *et al.* 2018). Yet, fungal genomes remain difficult to annotate. Historically, most annotation tools have relied upon comparative genomics, but other pipelines utilize RNA-seq data or a combination of both approaches to propose gene annotation models (CANTAREL *et al.* 2008; HAAS *et al.* 2011; MIN *et al.* 2017; HARIDAS *et al.* 2018). These pipelines are very efficient in intron-poor species, at least for predicting coding regions. For instance, a recent MAKER-based optimized pipeline tested on 39 budding yeast genomes missed only 3.9% of genes and 4.8% of exons, on average (SHEN *et al.* 2018). However, the results were poorer in intron-rich species, for which gene annotation is challenging. Even when RNA-seq data are available, it is still very difficult to correctly predict the exon-intron structure primarily because fungal exons can be extremely short (JANBON *et al.* 2014), but also because these genomes are compact. Thus, when tested on fungal data sets, *de novo* transcriptome assemblers like Trinity (GRABHERR *et al.* 2011) or Cufflinks (TRAPNELL *et al.* 2010) tend to predict very large transcripts with no biological relevance. Nevertheless, several pipelines have been published and sequencing centers like the Joint Genome Institute (JGI) and the Broad Institute have developed specialized pipelines to produce annotation drafts, which are very useful in large-scale comparison analyses (HAAS *et al.* 2011; HARIDAS *et al.* 2018).

Some methods, like the construction of large deletion collections, or precise analysis of gene content needs more precise annotation, and the annotation strategy applied will depend on the goal of the research (MUDGE AND HARROW 2016). Manual curation of a pre-annotated genome will likely result in the highest-quality gene prediction. Some tools, like Artemis (CARVER *et al.* 2012) and Apollo (DUNN *et al.* 2019), have been used to manually curate annotation, but they are time

consuming even when several annotators are implemented. Without manual curation, it is impossible to anticipate the results from an annotation bioinformatics pipeline fed with RNA-seq data. Typically, the quality of the prediction will depend on the diversity, quantity, and quality of the data, but no *a priori* indicator exists to determine if the *de novo* gene prediction is accurate.

Pathogenic *Cryptococcus* species are basidiomycete yeasts, which cause nearly 200,000 deaths annually around the world (KWON-CHUNG *et al.* 2014). There are currently eight recognized pathogenic species of *Cryptococcus* (HAGEN *et al.* 2015; Farrer et al., 2019). Manual annotation of the *Cryptococcus neoformans* and *Cryptococcus deneoformans* reference genomes revealed complex and dynamic transcriptomes (JANBON *et al.* 2014; WALLACE *et al.* 2020). These annotations were recently completed through precise identification of the transcript leader (TL) and 3'UTR sequences through TSS-seq and 3UTR-seq analyses; these annotations are likely the most complete and detailed annotations in intron-rich fungi (WALLACE *et al.* 2020). With 99.5% of 6,795 annotated coding genes containing introns, five to six introns per coding gene, and 37,832 introns in total, an automatic annotation of these genomes would be considered highly challenging even with the large sets of RNA-seq data that have been produced (WILM *et al.* 2007; JANBON 2018; WALLACE *et al.* 2020).

In this study, we compared the performances of three annotation pipelines fed with RNA-seq data. We gradually optimized the quality of the *de novo* annotation using the well-annotated genomes of *C. neoformans* and *C. deneoformans* as ground-truth inputs. We found that the quantity of data used should not be too large and that the number of introns predicted had a positive, linear relationship with the quality of the *de novo* annotation. We used this pipeline to re-annotate the reference genome of the RNAi-deficient *Cryptococcus deuterogattii* strain R265 using RNA-seq data. Analysis of the transcriptome dynamics of these three *Cryptococcus* species revealed that although the sense/antisense transcript ratio is similar across all three species,

intron retention is higher in *C. deuterogattii*. Comparative gene content analysis identified a list of genes that are absent or largely truncated in R265, many of which have been implicated in RNAi-mediated silencing in *Cryptococcus* species. Finally, we also identified several primary metabolic gene clusters (MGCs) that are absent in R265 and associated this loss with the subtelomeric gene content. Our data suggest an active exchange of MGCs between subtelomeric regions and more central regions of the genome. This exchange might contribute to the adaptive diversification of metabolite assimilation potential in *Cryptococcus*.

MATERIALS AND METHODS

RNA-Seq sample and data production

RNA-seq libraries from four growth conditions (exponential phase at 30°C, + exponential phase at 37°C, stationary phase at 30°C, and stationary phase at 37°), conducted in triplicate, of *C. neoformans* H99 and *C. deneoformans* JEC21 used in this study have been previously described (WALLACE et al. 2020). The *C. deuterogattii* R265 strain was grown in YPD at 30°C and 37°C under agitation to exponential or early stationary phase as previously described (WALLACE et al. 2020). Briefly, early stationary phase was obtained after 18 h of growth (final OD 600 = 15) starting from at OD600 = 0.5. Each *Cryptococcus* cell preparation was spiked in with one tenth (OD/OD) of *S. cerevisiae* strain FY834 cells grown in YPD at 30°C in stationary phase. Cells were washed, snap frozen and used to prepare RNA and total DNA samples. Biological triplicates were prepared in each condition. For RNA-seq, strand-specific, paired-end cDNA libraries were prepared from 10 µg of total RNA following poly-A purification using the TruSeq Stranded mRNA kit (Illumina) according to manufacturer's instructions. cDNA fragments of ~400 bp were purified from each library and confirmed for quality by Bioanalyzer (Agilent). DNA-Seq libraries were prepared using the kit TruSeq

DNA PCR-Free (Illumina). Then, 100 bases were sequenced from both ends using an Illumina HiSeq2500 instrument according to the manufacturer's instructions (Illumina). For the mating condition, total RNA was isolated (in biological triplicates) from a *C. neoformans* cross between the congenic mating partners H99 (*MAT*α) and YL99 (*MAT***a**) (SEMIGHINI et al., 2011) or a *C. deuterogattii* cross between the congenic mating partners R265 (*MAT*α) and AIR265 (*MAT***a**) (ZHU et al. 2013). Briefly, overnight cultures were grown under standard laboratory conditions in YPD at 30°C. Overnight cultures were diluted to an OD600 = 1.0, and cells from both strains were mixed, spotted onto V8 (pH = 5) mating medium, and incubated in the dark at room temperature for 48 h. Cells were scraped from mating plates, snap frozen, and RNA was isolated using Trizol following the manufacturer's protocol. RNA quality was confirmed by Bioanalyzer (Agilent) and RNA samples were depleted of ribosomal RNA with the Ribo-Zero Gold rRNA Removal Kit for Yeast (Illumina). Strand-specific, paired-end cDNA libraries were prepared using the TruSeq Stranded mRNA kit (Illumina), and 150 bases were sequenced from both ends using an Illumina HiSeq4000 instrument according to the manufacturer's instructions (Illumina).

RNA-Seq library trimming and rRNA cleaning

The paired reads from the RNA-seq libraries were trimmed for low quality reads and Illumina TruSeq adapters were removed with Cutadapt v1.9.1 (MARTIN 2011) with the following parameters: --trim-qualities 30 –e (maximum error rate) 0.1 --times 3 --overlap 6 --minimum-length 30. The cleaning of rRNA sequences was performed with Bowtie2 v2.3.3 (LANGMEAD AND SALZBERG 2012) with default parameters; unmapped paired reads were reported using option --un-conc to identify reads that did not align with rRNA sequences.

RNA-Seq library mapping

The cleaned reads from RNA-seq paired-end libraries from *C. neoformans* H99, *C. deneoformans* JEC21, and *C. deuterogattii* R265 were mapped against their reference genomes (NCBI Genome

158 Assemblies GCA_000149245.3, GCA_000091045.1 and GCA_002954075.1) with Tophat2 v2.0.14
159 (KIM *et al.* 2013) and the following parameters: minimum intron length 30; minimum intron
160 coverage 30; minimum intron segment 30; maximum intron length 4000; maximum multihits 1;
161 microexon search; and library-type fr-firststrand or fr-secondstrand (according to the RNA-seq
162 library).

163 **Pipeline selection**

164 The RNA-seq mapped reads from *C. neoformans* H99 and *C. deneoformans* JEC21 from the EXPO30
165 condition (exponential growth at 30 C) were tested in the three pipelines for gene prediction.
166 BRAKER1 (HOFF *et al.* 2016) was performed with the default parameters plus the exclusion of
167 alternative transcripts (--alternatives-from-evidence=false) using the three replicates (A, B, and C)
168 as RNA-seq source. Cuff-CQ (Cufflinks v2.1.1 ((TRAPNELL *et al.* 2010)) /Coding Quarry v2.0 (TESTA *et al.*
169 2015)) and C3Q (Cufflinks v2.1.1/Cuffmerge/Coding Quarry v2.0) were tested with the basic
170 parameters: minimum intron length (30); maximum intron length (4000); minimum isoform fraction
171 (0.9); and overlap radius (10). The merged BAM file generated by the three replicates (A, B, and C)
172 and used in the Cuff-CQ pipeline was obtained with Samtools *merge.* C3Q was performed separately
173 for the three BAM files; the GTF files generated by the three predictions (for replicates A, B, and C)
174 were then combined by Cuffmerge and the resulting transcripts were processed by CodingQuarry.
175 The evaluation of the pipeline sensitivity and precision for gene prediction was performed by
176 comparing the predicted annotations against the H99 and JEC21 reference annotations (WALLACE
177 et al., 2020) with the GFFCompare program (PERTEA AND PERTEA 2020).

178 *For a better understanding, the C3Q pipeline with the basic Cufflinks parameters is named as C3Q1*
179 *protocol in the results section.*

180 **Cufflinks parameters selection**

The selection of the best Cufflinks parameter combination was also performed with EXPO30 RNA-seq libraries from *C. neoformans* H99 and *C. deneoformans* JEC21 according to the C3Q pipeline. For this, the Cufflinks transcript assembly generated for each EXPO30 replicate (A, B, and C) was tested with fixed and variable parameter combinations (Table 1). Subsequently, as established for the C3Q pipeline, the predicted GTFs were merged and processed by CodingQuarry. All combinations include minimum intron length 30; maximum intron length 4000; and minimum isoform fraction 0.9; since we want to remove all isoforms. The variable parameters include: pre-mRNA fraction 0.15 to 1.0; small anchor fraction 0.0; minimum fragments per transfag 1; overlap radius 1, 10 and 100; 3' trimming (--trim-3-avgcov-thresh and –trim-3-dropoff-frac) 0. The evaluation of the Cufflinks parameters for sensitivity and specificity for gene prediction was performed by comparison of the predicted annotations against the H99 and JEC21 reference annotations with the GFFCompare program.

For a better understanding, the C3Q pipeline with the "Q" Cufflinks parameters (selected combination) is named as C3Q2 protocol in the results section.

196 **Table 1.**

	Cufflinks parameter combinations
A	--max-intron-length 4000 --min-intron-length 30 --min-isoform-fraction 0.9
B	--max-intron-length 4000 --min-intron-length 30 --min-isoform-fraction 0.9 --min-frags-per-transfag 1
C	--max-intron-length 4000 --min-intron-length 30 --min-isoform-fraction 0.9 --pre-mrna-fraction 0.25
D	--max-intron-length 4000 --min-intron-length 30 --min-isoform-fraction 0.9 --overlap-radius 10
E	--max-intron-length 4000 --min-intron-length 30 --min-isoform-fraction 0.9 --overlap-radius 100
F	--max-intron-length 4000 --min-intron-length 30 --min-isoform-fraction 0.9 --trim-3-avgcov-thres 0 --trim-3-dropoff-frac 0.0
G	--max-intron-length 4000 --min-intron-length 30 --min-isoform-fraction 0.9 --pre-mrna-fraction 0.25 --overlap-radius 10 --trim-3-avgcov-thresh 0 --trim-3-dropoff-frac 0.0
H	--max-intron-length 4000 --min-intron-length 30 --min-isoform-fraction 0.9 --pre-mrna-fraction 0.25 --overlap-radius 10 --trim-3-avgcov-thresh 0 --trim-3-dropoff-frac 0.0 --min-frags-per-transfag 1
I	--max-intron-length 4000 --min-intron-length 30 --min-isoform-fraction 0.9 --pre-mrna-fraction 0.25 --overlap-radius 1 --trim-3-avgcov-thresh 0 --trim-3-dropoff-frac 0.0
J	--max-intron-length 4000 --min-intron-length 30 --min-isoform-fraction 0.9 --pre-mrna-fraction 0.25 --overlap-radius 10 --trim-3-avgcov-thresh 0 --trim-3-dropoff-frac 0.0 --min-frags-per-transfag 1
K	--max-intron-length 4000 --min-intron-length 30 --min-isoform-fraction 0.9 --pre-mrna-fraction 0.25 --overlap-radius 1 --trim-3-avgcov-thresh 0 --trim-3-dropoff-frac 0.0 --small-anchor-fraction 0.0
L	--max-intron-length 4000 --min-intron-length 30 --min-isoform-fraction 0.9 --pre-mrna-fraction 0.50 --overlap-radius 1 --trim-3-avgcov-thresh 0 --trim-3-dropoff-frac 0.0 --small-anchor-fraction 0.0
M	--max-intron-length 4000 --min-intron-length 30 --min-isoform-fraction 0.9 --pre-mrna-fraction 0.75 --overlap-radius 1 --trim-3-avgcov-thresh 0 --trim-3-dropoff-frac 0.0 --small-anchor-fraction 0.0
N	--max-intron-length 4000 --min-intron-length 30 --min-isoform-fraction 0.9 --pre-mrna-fraction 0.80 --overlap-radius 1 --trim-3-avgcov-thresh 0 --trim-3-dropoff-frac 0.0 --small-anchor-fraction 0.0
O	--max-intron-length 4000 --min-intron-length 30 --min-isoform-fraction 0.9 --pre-mrna-fraction 0.90 --overlap-radius 1 --trim-3-avgcov-thresh 0 --trim-3-dropoff-frac 0.0 --small-anchor-fraction 0.0

P	--max-intron-length 4000 --min-intron-length 30 --min-isoform-fraction 0.9 --pre-mrna-fraction 1.0 --overlap-radius 1 --trim-3-avgcov-thresh 0 --trim-3-dropoff-frac 0.0 --small-anchor-fraction 0.0
Q	--max-intron-length 4000 --min-intron-length 30 --min-isoform-fraction 0.9 --pre-mrna-fraction 0.85 --overlap-radius 1 --trim-3-avgcov-thresh 0 --trim-3-dropoff-frac 0.0 --small-anchor-fraction 0.0

Gene predictions with H99 and JEC21 RNA-seq libraries

The validation of this gene prediction system was evaluated by applying the C3Q pipeline with the best selected Cufflinks parameters ("Q" combination) to all RNA-seq libraries from *C. neoformans* H99 and *C. deneoformans* JEC21. For H99, the fifteen libraries obtained from the five growth conditions were used (Exponential phase at 30°C, Exponential phase at 37°C; Stationary phase at 30°C, Stationary phase at 37°C and Mating). For JEC21, we tested twelve libraries obtained from four growth conditions (Exponential phase at 30°C, Exponential phase at 37°C; Stationary phase at 30°C and Stationary phase at 37°C).

The evaluation of the sensitivity and specificity for gene prediction was performed by comparison of the predicted annotations against H99 and JEC21 reference annotations with the GFFCompare program.

For a better understanding, the C3Q pipeline with the "Q" Cufflinks parameters and the RNA-seq libraries for all the sequenced conditions ("ES3037M" for C. neoformans H99 and "ES3037" for C. deneoformans JEC21) is named as C3Q3 protocol in the results section.

*ES3037M: Exponential phase at 30°C (EXPO30) + Exponential phase at 37°C (EXPO30) + Stationary phase at 30°C (STAT30) + Stationary phase at 37°C (STAT37) + Mating

*ES3037: Exponential phase at 30°C (EXPO30) + Exponential phase at 37°C (EXPO30) + Stationary phase at 30°C (STAT30) + Stationary phase at 37°C (STAT37)

Effect of different conditions on predictions

To evaluate the effect of the growth conditions on the predicted annotation, we used combinations of RNA-seq libraries derived from two, three, and four of the growth conditions for *C. neoformans* H99 and of two and three of the growth conditions for *C. deneoformans* JEC21. The predictions for each combination were performed according to the C3Q pipeline and "Q" Cufflinks parameters. The evaluation of the sensitivity and specificity for gene prediction was performed by comparison of the predicted annotations against H99 and JEC21 reference annotations with the GFFCompare program.

Evaluation of the effect of the sequencing depth on gene prediction quality

The evaluation of the effect of the sequencing depth on gene prediction was performed by down sampling the three RNA-seq libraries from the EXPO30 condition (replicates A, B, and C) with the tool *PositionBasedDownsampleSam* from Picard package (https://broadinstitute.github.io/picard/). In this analysis, *C. neoformans* H99 and *C. deneoformans* JEC21 were used. According to a random algorithm that downsamples BAM files, we used defined fractions of 1, 5, 7.5, 10, 15, 20, 30 and 40 million reads for each replicate. Subsequently, the predictions were performed according to the C3Q pipeline with the Cufflinks "Q" parameter combination using the downsampled files. Evaluation of the sensitivity and specificity of gene prediction was performed by comparison of the predicted annotations against H99 and JEC21 reference annotations with the GFFCompare program.

Gene predictions with downsampled H99 and JEC21 RNA-seq libraries

Gene prediction using the downsampled BAM files from the RNA-seq conditions was performed according to the C3Q pipeline with "Q" Cufflinks parameters and the downsampled RNA-Seq alignment files for all *C. neoformans* H99 (Exponential phase at 30°C, Exponential phase at 37°C,

Stationary phase at 30°C, Stationary phase at 37°C and Mating) and *C. deneoformans* JEC21 (Exponential phase at 30°C, Exponential phase at 37°C; Stationary phase at 30°C and Stationary phase at 37°C) growth conditions. The downsampling of each replicate to 7.5 million reads was performed with the Picard package, as previously described. Evaluation of the sensitivity and specificity of gene prediction was performed by comparison of the predicted annotations against H99 and JEC21 reference annotations with the GFFCompare program.

For a better understanding, the C3Q pipeline with the "Q" Cufflinks parameters and the subsampled BAM files from RNA-seq libraries for all the growth conditions ("ES3037M" for C. neoformans H99 and "ES3037" for C. deneoformans JEC21) is named as C3Q4 protocol in the results section.

Characterization of novel and missed loci

The identification of novel and missed loci was performed with the GFFCompare program using the reference annotations from *C. neoformans* H99 and *C. deneoformans* JEC21 and the predicted C3Q gene annotations. Evaluation of the functional annotation (function, presence of domain signatures) of these sequences was performed by Blastp and Interproscan search from Blast2GO (CONESA *et al.* 2005). The expression quantification of these sequences was performed with HTSeq-count (ANDERS *et al.* 2014) with the following parameters *--stranded yes -f bam -r pos -t CDS*.

Deletion of dubious novel loci from predictions

Deletion of dubious novel sequences was tested with predicted transcripts of 100 nt, 150 nt, 200 nt and 300 nt, as well as intronless sequences of 300 nt from *C. neoformans* H99 and *C. deneoformans* JEC21 C3Q predictions. The sequence deletion and evaluation of the results was performed with an in-house AWK script and the GFFCompare program. Deletion of genome-predicted sequences without supporting reads and those with low FPKM values were performed and evaluated with an in-house AWK script combined with the HTSeq-count and the GFFCompare program. Deletion of alternative transcripts from multi-transcript loci was also performed with an in-house AWK script

and GFFCompare. In this process, we selected for transcripts predicted by Cufflinks with supporting RNA-seq evidence. Of these selected transcripts, the longest transcript was chosen. For the other genes predicted only from genome sequencing (without RNA-seq evidence), the longest transcript was selected.

We assessed the sensitivity and specificity of the C3Q predictions for *C. neoformans* H99 and *C. deneoformans* JEC21 against their reference annotations to analyze the effect of dubious sequence deletion. Filter combinations with the low numbers of remnant novel transcripts and smaller reduction in the prediction quality parameters were favored.

For a better understanding, the C3Q pipeline with the "Q" Cufflinks parameters, the subsampled BAM from RNA-seq libraries for all the sequenced conditions ("ES3037M" for C. neoformans H99 and "ES3037" for C. deneoformans JEC21), and the sequence filtering (sequences up to 150 nt, intronless sequences up to 300 nt, genome-predicted sequences without reads and alternative transcripts) is named as C3Q5 protocol in the results section.

Retrieval of deleted and non-predicted loci

The mapping of *C. neoformans* H99 protein sequences in the *C. deneoformans* JEC21 genome and JEC21 protein sequences in the H99 genome by Exonerate v2.2.0 program (https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate) with the following parameters (*protein2genome --percent 30 -- bestn 1 --minintron 30 – maxintron 4000 --showalignment false --showvulgar false --showtargetgff true --refine region --subopt false)* was performed to recover sequences deleted in the previous filtering step with conserved orthology in *Cryptococcus*. For this purpose, the mapped gene coordinates matching previously predicted sequences (GFFCompare program) were used to add these deleted genes to the annotation with an in-house AWK script. The addition of non-predicted genes was performed by comparing the mapped protein sequence coordinates and the genomic regions without predicted genes.

For a better understanding, the C3Q pipeline with the "Q" Cufflinks parameters, the subsampled BAM from RNA-Seq libraries for all the sequenced conditions ("ES3037M" for C. neoformans H99 and "ES3037" for C. deneoformans JEC21), the sequence filtering (sequences up to 150 nt, intronless sequences up to 300 nt, genome-predicted sequences without reads and alternative transcripts), and the Exonerate-based retrieval of deleted and non-predicted genes is named as C3Q6 protocol in the results section.

Automatization of the C3Q pipeline

The C3Q pipeline, an automatic gene predictor, was built with Python3 code *(C3Q_gene-predictor.py)* and is available in Github (https://github.com/UBTEC/C3Q)

The C3Q pipeline includes all established parameters for *Cryptococcus* genome annotation (C3Q6 protocol):

- The Cufflinks assembly of transcripts for each RNA-seq library.

- The merging of the generated GTF files by Cuffmerge.

- The GTF conversion to GFF format (needed for CodingQuarry).

- The training and genome prediction by CodingQuarry, using the merged GFF file and the reference genome.

- The sequence filtering: deletion of small transcripts up to 150 nt and intronless transcripts up to 300 nt; deletion of genome-predicted sequences without supporting reads and deletion of alternative transcripts from multi-transcript loci.

- The retrieval of deleted and non-predicted orthologous/paralogous sequences by Exonerate (modified version with GFF3 support from https://github.com/hotdogee/exonerate-gff3).

Gene prediction in *C. deuterogattii* R265

Gene prediction in *C. deuterogattii* R265 was performed with the C3Q pipeline (C3Q6 protocol) using the *C3Q_gene-predictor.py* script. For this, the five RNA-seq triplicate libraries from *C. deuterogattii* R265 (Exponential phase at 30°C, Exponential phase at 37°C, Stationary phase at 30°C, Stationary phase at 37°C and Mating) were subsampled to 7.5 million reads each, and input into the script in addition to the *C. neoformans* H99 and *C. deneoformans* JEC21 protein sequences for the Exonerate step.

Concomitantly, manual correction of genes from chromosomes 9 and 14 was performed with the software Artemis (CARVER et al., 2012), the R265 genome (NCBI assembly GCA_002954075.1), and the stranded paired-end RNA-seq data from *C. deuterogattii* R265 in the five growth conditions.

The predicted annotation was evaluated by comparing it to the manually corrected genes from chromosomes 9 and 14, as well as the *C. deuterogattii* R265 annotations from Broad (NCBI assembly GCA_000149475.3) and Ferrareze et al., 2017.

CDS gene coordinates from old annotations were also identified in the new sequenced genome with Exonerate aligner (*coding2genome*). The predicted novel genes were named with CNBG ID numbers above 10000. The statistics of the gene annotations ware generated by AGAT script *agat_sp_statistics.pl* (https://github.com/NBISweden/AGAT). The final annotation is available in file S1.

Comparison of orthologue groups across *Cryptococcus* species

Ortho-groups and genes unique to *C. neoformans* H99, *C. deneoformans* JEC21 and *C. deuterogattii* R265 were evaluated with Orthofinder v2.3.3 configured to use the Blast aligner. Gene size comparisons were performed with orthologues and paralogues (if the true orthologue was not known) obtained from the OrthoFinder analysis (EMMS AND KELLY 2019), as well as gene sizes. For the ratio calculation, the size (nt) of the R265 gene was divided by the size (nt) of the H99 and JEC21 orthologous genes. The analysis of conserved domains in unique sequences and the functional

annotation of *C. deuterogattii* R265 were performed with Blast2GO (Blastp, Interproscan and GO mapping).

Gene orientation analysis

To determine the frequency of tandem genes with the same orientation, we searched for groups of two, three, four, or five genes assigned to the same DNA strand in the GFF file from *C. deuterogatti* R265 (with our new annotation), *C. neoformans* H99 (Genome assembly reference GCF_000149245.1), and *C. deneoformans* JEC21 (Genome assembly reference GCF_000091045.1) annotations. This was performed by analyzing the orientation of each gene pair in the GFF file from R265, H99 and JEC21, recording the frequency of genes converging (tail-to-tail), diverging (head-to-head), and in the same orientation (head-to-tail) in the whole genome and for each chromosome. This was executed using an in-house Python script (script *gene_organization_analysis.py*).

Antisense transcription analysis

To evaluate the antisense transcription in the genomes analyzed, we first generated a reversed annotation, which consisted of a GFF file with genes assigned to the opposite strand of their actual annotation. With the annotation and the reverse annotation, we analyzed the percentage of antisense transcription for each protein coding gene using the software HTseq-count using the following attributes *(-f bam -r pos - s yes -t CDS -i ID -m intersection-nonempty --nonunique none (.bam) (.gff))* and the distinct RNA-seq libraries. Sense/antisense counts ratios for each gene for each condition were plotted. The script used for generation of a reverse GFF is available (*reverse_gff.py*).

Intron retention evaluation

For a given intron, an IR index was calculated by determining the ratio of spliced:non-spliced reads at the upstream and downstream exon-intron junctions and choosing the lowest of these two numbers. These IR indices were calculated using RNA-seq obtained from cells growing in each of the

four growth conditions. An intron was considered to be regulated by intron retention when the IR index was at least 0.01. We restricted our analysis to introns with more than 10 spliced reads.

Statistical analysis

The proportion of genes with intron retention regulation was compared amongst the distinct conditions using one way ANOVA followed by multi-comparison analysis corrected by FDR. The X-squared analysis was conducted using R (version 4.0.2) and plots were prepared using the corrplot package (version 0.84).

Availability and accession number

Raw and summarized sequencing data are available at SRA with the accession number: PRJNA660459. The C3Q pipeline is available in Github (https://github.com/UBTEC/C3Q); Supplemental files available at FigShare. The final annotation of *C. deuterogattii* genome was submitted to NCBI and is available on accessions CP025759.1 to CP025772.1.

RESULTS

Pipeline selection

We first compared the performance of two previously published annotation pipelines used in coding gene *de novo* annotation in intron-rich fungal genomes using RNA-seq data. The BRAKER1 pipeline, which combines GeneMark-ET (LOMSADZE *et al.* 2014) and Augustus (STANKE *et al.* 2008) and is already optimized with the best prediction parameters (HOFF *et al.* 2016), was compared with a genome annotation pipeline composed of Cufflinks v2.1.1 (TRAPNELL *et al.* 2010) and CodingQuarry v2.0 (TESTA *et al.* 2015). We used the *C. neoformans* H99 and *C. deneoformans* JEC21 genomes as controls to assess of the performance of both pipelines (GONZALEZ-HILARION *et al.* 2016; WALLACE *et al.* 2020).

For this analysis, we used only RNA-seq data obtained in biological triplicate from cells grown to exponential growth phase at 30°C in complete medium (YPD) (EXPO30) (WALLACE *et al.* 2020). Previously described BAM files obtained after alignment of trimmed reads to the *C. neoformans* H99 genome were input into the BRAKER1 and Cufflink-CodingQuarry pipelines (WALLACE *et al.* 2020). For the Cufflink-CodingQuarry-based analyses, we used two alternative protocols. In the first case, we first merged the BAM files from each of the three replicates (CUFF-CQ protocol). In the second case, each replicate BAM file was used to generate a unique GTF prediction file, these files were then merged using Cuffmerge and used by CodingQuarry as a single transcript source (C3Q1 protocol) (Figure 1).

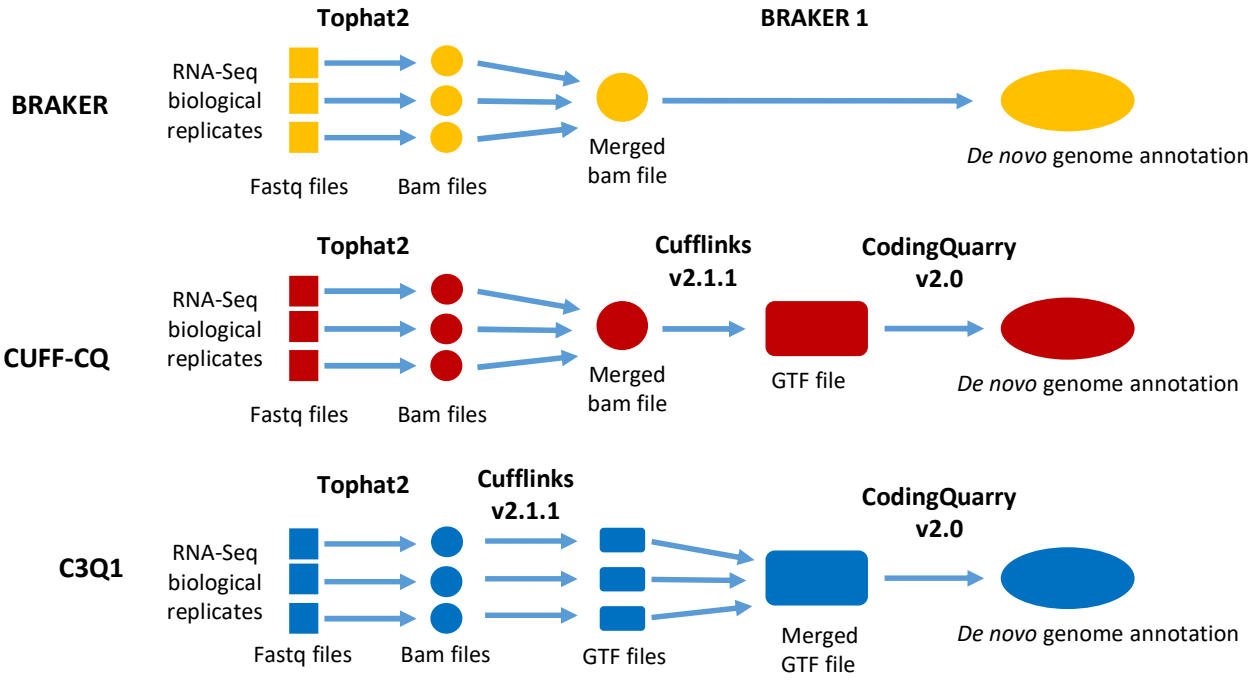


Figure 1. Schematic of the different pipelines tested in this study.

To compare the quality of these pipelines for identification of coding genes, we calculated their sensitivity (percentage of coding genes present in the reference annotation overlapping with one coding gene in the *de novo* annotation) and their specificity (percentage of predicted coding

genes overlapping with one coding gene in the reference annotation). These comparisons revealed that BRAKER was much more sensitive than either Cufflinks-CodingQuarry protocol, missing only 91 coding genes in the *C. neoformans* genome (Figure 2A, Supplementary Table S1). However, the BRAKER pipeline was less specific (91.4%), predicting 622 coding genes absent in the reference annotation (Figure 2B). In contrast, both Cufflink-CodingQuarry protocols missed more coding genes (743 and 447 genes for CUFF-CQ and C3Q protocols, respectively), but had a higher (95%) specificity (Figures 2A,2B). We observed a similar pattern when we looked at CDS introns and CDS exons within the identified references genes. Again, the BRAKER pipeline was very sensitive, with only 0.4% missed introns (n=157) and 0.4% missed exons (n=164) in the prediction but had poor specificity, with 4471 novel introns and 3065 exons predicted but not present in the reference annotation (Figures 2A,B; Table S1). On the other hand, both Cufflink-CodingQuary-based protocols missed more introns (n=4944 and n=3238 for CUFF-CQ and C3Q1 protocols, respectively) and more exons (n=4281 and n=2777, respectively) but both were more specific, predicting less than 200 introns or exons not present in the reference annotation. Overall, both Cufflink-CodingQuary-based protocols returned more conservative results; they were more specific in the predicted gene structures and identified a smaller number of new insertions (novel exons/introns) and new genes. These more conservative predictions came at the cost of missing a larger number of features than the BRAKER protocol.

To assess all of these performance parameters and select the highest-performing protocol for further optimization, we considered the sensitivity and specificity of accurately predicting gene structure (perfect exon/intron organization) for each of the three pipelines. The C3Q1 protocol was the most sensitive, perfectly predicting the exon-intron layout of 66.5% (n=4516) of *C. neoformans* H99 genes, compared to 65.2% and 51.9% perfect predictions from the BRAKER and CUFF-CQ protocols, respectively (Figure 2C). This was also the most specific protocol with 63.9% of the predicted genes perfectly fitting the reference gene structures, compared to 61.3% and 52.6%

of the predictions made by the BRAKER and CUFF-CQ protocols, respectively (Figure 2C). To better compare the quality of these pipelines, we considered a quality index that multiplied the sensitivity by the specificity of predicted gene structure predictions (Figure 2D). Our results demonstrated that the C3Q1 pipeline was the best, with a quality index of 0.42. We performed the same analysis with the *C. deneoformans* JEC21 genome annotation data and obtained similar results, confirming the C3Q1 protocol was the best protocol for further optimization (Figure S1).

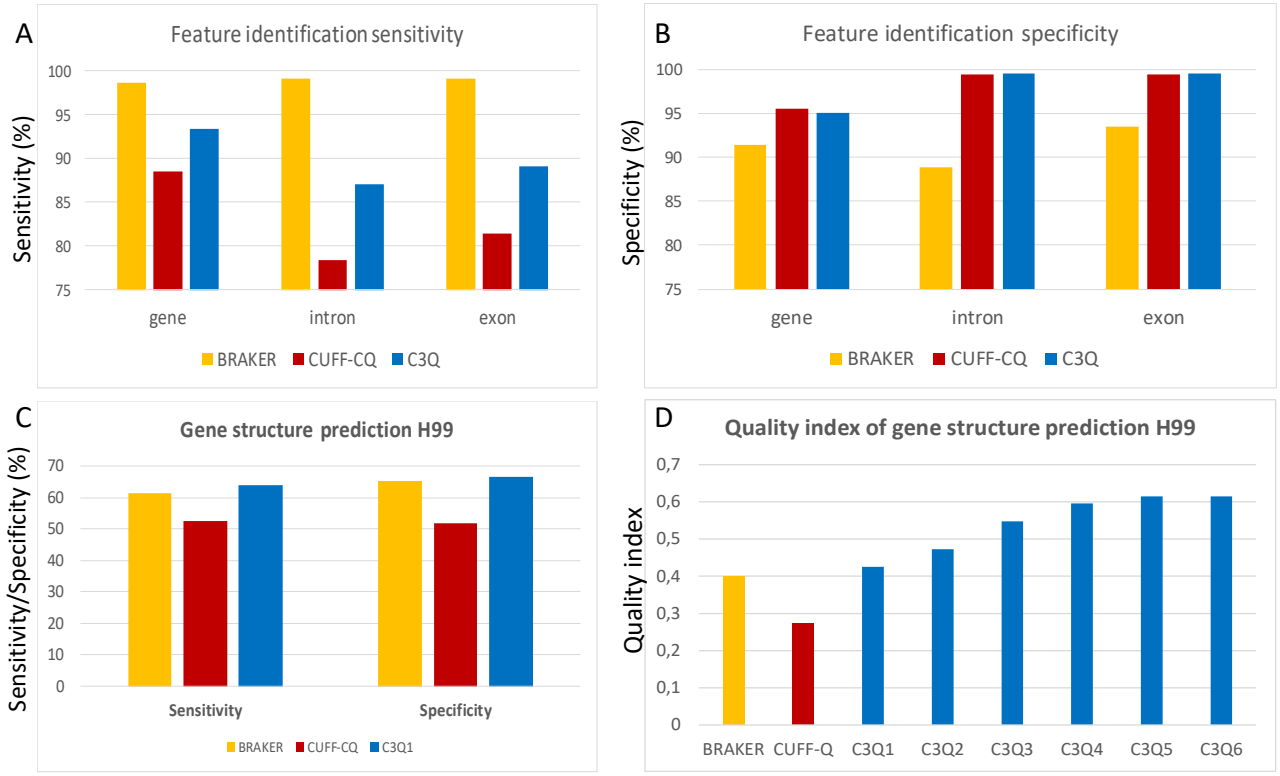


Figure 2: Sensitivity (A) and specificity (B) of the different tested pipelines for *C. neoformans* H99 genomic feature identification. For introns and exons, calculations were done using only genes that were both identified by the pipelines and present in the reference annotation. (C) Sensitivity and specificity of gene structure predictions using the three annotation pipelines. (D) Optimization of the C3Q pipeline. C3Q1 is the pipeline using default settings. C3Q2 through C3Q6 refer to the results obtained after each step of the pipeline optimization.

435 **Optimization of the C3Q1 pipeline**

436 *Effect of Cufflinks settings*

437 To improve both the number of perfectly predicted gene structures and the percentage of
438 predicted loci in perfect agreement with the reference coding gene structures, we considered 17
439 combinations of Cufflinks settings. We varied parameters including 1) the minimum distance
440 between transfags allowed to merge, 2) trimming of 3' ends of reads, 3) filtering of alignments
441 that lie within intronic intervals, 4) filtering of suspicious spliced reads, 5) minimum RNA-seq
442 fragments allowed to assemble transfags, and 6) filtering of alignments that lie within intronic
443 intervals in the same set of RNA-seq data. These Cufflinks parameter modifications reduced the
444 number of missed genes and increased the number of reference genes identified from 6348 to
445 6462 (Table S1). Using the final settings, the pipeline C3Q2 quality index reached a score of 0.473,
446 with 70.8% of reference gene intron-exon structures perfectly predicted and 66.8% of the
447 predicted genes perfectly matching the reference exon-intron gene structures (Figure 2D).

448

449 *Effect of the RNA-seq data set*

450 We tested the C3Q2 optimized pipeline using additional RNA-seq data obtained under five
451 different conditions in triplicate: stationary growth at 30°C (STAT30) and 37°C (STAT37),
452 exponential growth at 30°C (EXPO30) and 37°C (EXPO37), and growth under mating conditions
453 (Mating). Each RNA-seq data set generated a similar number of predicted transcripts, ranging
454 between 7049 genes using the STAT37 set up to 7199 loci using the EXPO30 data set (Table S1).
455 When compared to the reference set of genes, the number of predicted annotations were also
456 very similar (Table S1). As expected, including more samples improved the annotation quality. The
457 usage of the five conditions improved the C3Q3 annotation quality index to 0.547 despite the fact
458 that more predicted genes not present in the reference genome were identified using this pipeline

(n = 510) (Figures 2D, S2, Table S1). Similar results were obtained using the *C. deneoformans* JEC21 annotation and RNA-seq data (Figure S1, Table S1).

Evaluation of RNA-Seq data set size in gene prediction quality

Analysis of the results obtained using the C3Q2 pipeline fed with individual replicates of the EXPO30 RNA-seq data counterintuitively suggested the size of the initial BAM file might be negatively correlated with the quality of the final prediction (Tables S1). Identical analysis performed with *C. deneoformans* RNA-seq gave a similar result, suggesting the sequencing depth may substantially affect the quality of predictions and should be considered as a possible parameter of optimization for gene prediction pipelines. To improve the analysis of the effect of the size of the data set on the quality of gene prediction, the C3Q2 pipeline was tested with different representative fractions of reads from a single EXPO30 replicate. Thus, replicate samples with 1, 5, 7.5, 10, 15, 20, 30, and 40 million reads were used for *de novo* annotation of the *C. neoformans* H99 genome, and the quality of the gene predictions were compared. We performed this analysis using the same strategy for *C. deneoformans.* As shown in Figure 3, the quality of the gene structure prediction was highly dependent on the size of the RNA-seq initial data set in both species and strongly anti-correlated with the number of Cufflinks-assembled transcripts (Table S1). The highest-quality predictions were obtained with replicate samples with only 5-10 million reads. Using this adjusted read depth, the prediction showed improvement in nearly all parameters, including for missed genes, missed exons, and missed introns (Table S1).

We adjusted the number of reads to 7.5 million for each replicate in each condition and used these adjusted RNA-seq data sets for *de novo* annotation of the *C. neoformans* and *C. deneoformans* reference genomes. As expected, the gene predictions obtained with the C3Q4 pipeline were further improved with a quality index of 0.593 and 0.596, for *C. neoformans* and *C.*

deneoformans annotations, respectively (Figure 2D, Figure S1, Table S1). In *C. neoformans*, 81.9% of the reference gene structure was perfectly predicted and only 1.9% (n=129) of genes were missed.

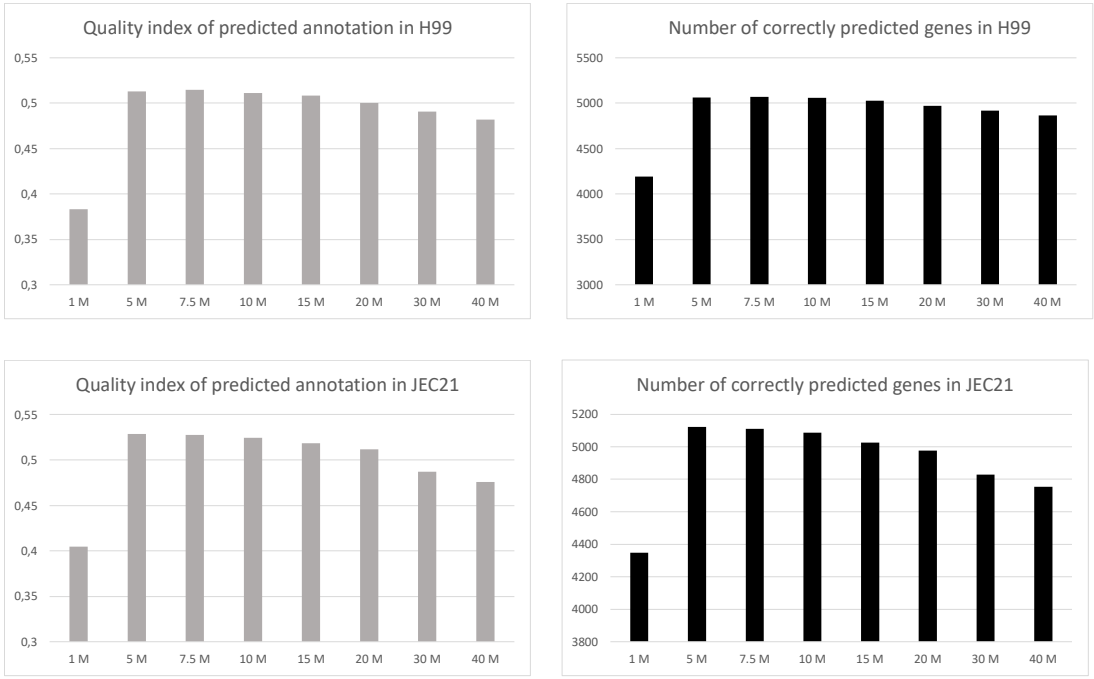


Figure 3. Effect of the size of the BAM file on the quality of the predicted annotation and on the number of correctly predicted genes in *C. neoformans* H99 and *C. deneoformans* JEC21.

Gene filtering

Each optimization step improving the quality of the gene prediction was also associated with an increase of the number of predicted genes not present in the previously annotated reference genome (Table S1). Using the C3Q5 protocol, 717 (703 loci) and 774 (762 loci) additional genes were predicted in *C. neoformans* and *C. deneoformans*, respectively, compared to the reference annotation. The majority of these genes are likely to be misannotations. One hundred and six of the sequences had domain signatures of transposable elements, suggesting they correspond to fragments of transposons or retrotransposons unannotated in the reference H99 genome. To filter out some of the novel predicted genes, we looked at their structure and

coverage. We compared the characteristics of these false-positive genes to the reference genes and found that most novel predicted genes were short (219 nt mean length, 112 nt median length), poorly expressed, and intronless. We tested different filters alone and in combination to eliminate as many false positive genes as possible without affecting the number of correctly predicted ones; the results are presented in Table S1. In both species, the best combination of filters eliminated all spliced coding regions smaller than 150 nt, all intronless genes smaller than 300 nt, and all genome-predicted genes not supported by any RNA-seq reads. Due to the presence of secondary transcripts at some loci, many of which were generated due to differences in the RNA-seq-predicted and genome-predicted transcripts for the same gene, a fourth filtering step was performed. In this step, to ensure that there was only one transcript per loci, the longest RNA-seq-predicted transcript or the longest genome-predicted transcript (for loci without RNA-seq evidence) was selected as a representative for the gene CDS coordinates. After this fourth filtering step, the number of predicted genes not present in the reference annotation was down to 409 and 427 genes in H99 and JEC21, respectively, and the quality index of the annotation increased to 0.614 and 0.608, respectively (Figures 2D, S1; Table S1).

Exonerate-based recovery of missed genes

Improvement of the pipelines was associated with an increase in the sensitivity of gene identification. In the initial C3Q1 protocol, 447 reference genes were missed, whereas only 162 H99 genes and 132 JEC21 genes were missed with the C3Q5 pipeline. Blast2GO analysis of the protein sequences encoded by the missed genes identified 16 proteins with conserved domains suggesting that it might be possible to identify some of them through comparative sequence analysis. Another 111 sequences were defined as hypothetical proteins. We first used the sequence alignment program Exonerate (SLATER AND BIRNEY 2005) and the JEC21 reference proteome as a reference to try to recover missed coding genes in H99. As expected, this analysis

identified a number of missed loci, but also added a number of unpredicted loci thus reducing the quality of the annotation. In the final C3Q6 pipeline, we chose to restrict this Exonerate analysis to genes that had been filtered out in the last step of the C3Q5 pipeline. We ultimately identified 14 and 9 novel genes in H99 and JEC21, respectively. Overall, the C3Q6 optimized pipeline was able to identify nearly 98% of genes in H99, contributing only 410 (~6%) novel genes. Importantly, the exon-intron structure of the predicted genes was predicted perfectly for >81% of the reference genes in both species.

Intron number is predictive of the quality of the C3Q6 predicted annotation

During the course of the C3Q pipeline optimization, we obtained 88 versions of the H99 genome annotation. We carefully examined the different characteristics of these annotations, looking for a parameter predictive of their quality. First, we plotted the number of predicted loci against the quality of the annotation, but we did not observe any correlation. Similar results were obtained when we looked at missed or novel loci, suggesting that these parameters were also not indicative of the annotation quality. However, when the numbers of introns predicted were plotted against the quality of the annotation, we obtained a linear correlation (Figure 4). This correlation was lost during the filtration steps (red dots), which tend to reduce the number of introns. Similar results were obtained for JEC21, suggesting that the number of introns is a good parameter to consider when evaluating the quality of the annotation using the C3Q pipeline.

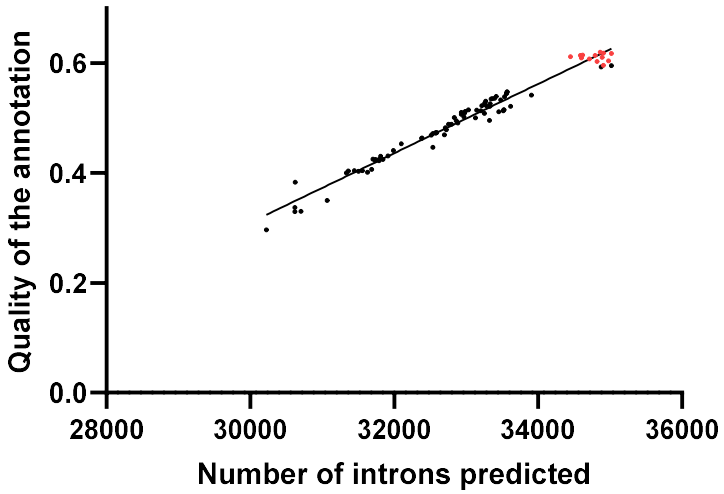


Figure 4. Relationship between the quality index of the H99 annotation and the number of introns and transcripts predicted. The red points correspond to the filtering steps of the optimization pipeline.

Genome annotation of the *Cryptococcus deuterogattii genome*

We used the C3Q6 optimized pipeline to generate a new genome annotation for the *C. deuterogattii* reference strain R265. This reference strain was isolated in 2001 from the bronchoalveolar lavage fluid of an infected patient from the Vancouver Island outbreak (KIDD *et al.* 2004). Because of its outbreak origin and the loss of a functional RNAi pathway (D'SOUZA *et al.* 2011), *C. deuterogattii* has been the focus of a number of studies in recent years (CHENG *et al.* 2009; MA *et al.* 2009; NGAMSKULRUNGROJ *et al.* 2012; HUSTON *et al.* 2013; LAM *et al.* 2019). The R265 genome has been previously annotated three times (D'SOUZA *et al.* 2011; FARRER *et al.* 2015; FERRAREZE *et al.* 2017), but a recent release of telomere-to-telomere genome sequence data (YADAV *et al.* 2018) motivated us to generate an updated annotation. We generated RNA-seq data in biological triplicate from cells grown under five conditions (exponential growth phase at 30°C and

37°C, stationary growth phase at 30°C and 37°C, and under mating conditions) as previously described for *C. neoformans* H99 and JEC21 (WALLACE *et al.* 2020). Reads were trimmed, aligned to the reference genomes (Table S2), and input into the optimized C3Q6 genome annotation pipeline.

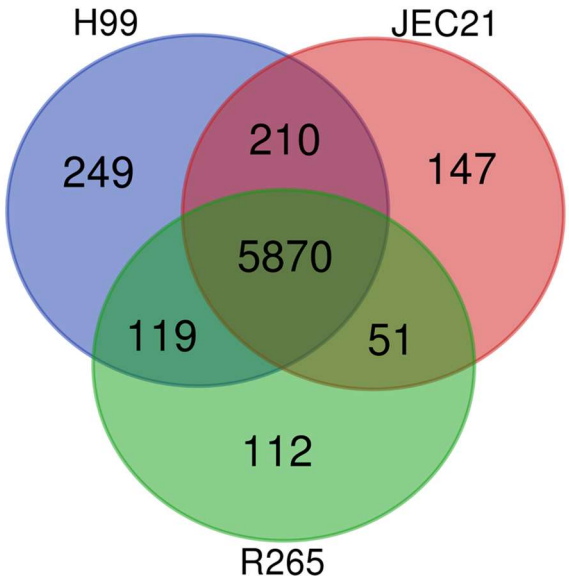
To gain further insight into the quality of our updated R265 annotation, the structure predictions of genes for *C. deuterogattii* R265 chromosomes 9 and 14 were manually curated through visual examination of read alignments using Artemis (CARVER *et al.* 2012) as previously described (JANBON *et al.* 2014; GONZALEZ-HILARION *et al.* 2016). We compared this manually curated annotation of chromosomes 9 and 14 with the prediction obtained from the C3Q6 genome annotation pipeline of these two chromosomes. This analysis revealed a quality index of this annotation of 0.51, with 68% of all predicted loci correctly predicted and 75% of the manually curated genes on these two chromosomes correctly predicted. In *C. neoformans* and *C. deneoformans*, the C3Q6 genome annotation pipeline missed very few genes (1.4% missed) and predicted a small number of false-positive genes (6.3%) (Table S3). As expected, the quality of the C3Q6 annotation was much better than previously published annotations (FARRER *et al.* 2015; FERRAREZE *et al.* 2017) (Table S3).

Manual curation of R265 annotation

To systematically analyze critical points of the automated *C. deuterogattii* R265 gene prediction, four sets of data were evaluated and selected for manual correction: 1) Exonerate-retrieved sequences (deleted and non-predicted), 2) predicted novel loci, 3) genes predicted in merged/split loci, and 4) small and potential pseudogenes. During the course of this manual curation of chromosomes 9 and 14, visual examination of the aligned reads revealed a number of loci at which the genome sequence did not entirely align with the RNA-seq reads, suggesting there were errors in the reference genome assembly. These errors were responsible for gene shortening

584 or splitting and might partially explain the lower quality index score calculated for the R265
585 predicted annotation of chromosomes 14 and 9 compared to the quality scores obtained using
586 similar data from JEC21 and H99. To systematically identify these types of annotation mistakes, we
587 compared the size of the *C. deuterogattii* R265 predicted genes with their *C. neoformans* H99/*C.*
588 *deneoformans* JEC21 orthologous counterparts. We identified 729 genes in R265 that were
589 significantly smaller than their *C. neoformans* and *C. deneoformans* orthologues (size ratio < 0.8).
590 Visual examination of these loci revealed that most of them were incorrectly predicted and
591 needed manual curation. Manual curation was also performed for 67 genes that were significantly
592 smaller than only one of their orthologues (*C. neoformans* or *C. deneoformans*). This manual
593 curation also identified 125 genes which would have otherwise been challenging to predict due to
594 genome sequence errors mistakenly affecting orthologue size ratios (Table S4).
595 Overall, our new version of R265 genome annotation contains 6,405 coding genes with
596 33,619 introns in CDS regions (34,512 introns including the UTRs). The manually corrected genes
597 from chromosomes 9 and 14 were added and replaced the predicted genes from these regions,
598 improving the quality of the final annotation. Of the 6,405 genes predicted with the C3Q6 pipeline,
599 the CDS structure was modified for 873 coding through manual curation. Annotation of 3'UTR
600 and/or the transcript leader sequence was performed for 1210 genes from the manually curated
601 chromosomes (9 and 14) and the 873 manually curated genes with modified CDS structures.
602 Furthermore, we annotated 55 lncRNAs and used tRNAscan (LOWE AND CHAN 2016) to annotate 161
603 tRNAs (Table S3). We also removed all genes predicted to reside within centromeric regions, and
604 used previously published, manually curated annotations for these regions (YADAV *et al.* 2018).
605
606 **Putative pseudogenes and missing genes**
607 We compared the gene content across the three annotated *Cryptococcus* genomes. We
608 identified 5870 ortho-groups common to all three species (Figure 5). We found a similar number

609 of R265-specific genes to the number of specific genes identified in H99 and JEC21, which is likely
610 an indicator of the high quality of this annotation. Of interest, this analysis revealed 210 ortho-
611 groups missing in *C. deuterogattii* R265, but present in both the *C. neoformans* H99 and the *C.*
612 *deneoformans* JEC21 genomes (Table S5). This list of genes was curated first through Exonerate-
613 based analysis and then through manual examination of the syntenic loci.
614



615

616 Figure 5. Comparative gene content of the annotated *C. neoformans* H99, *C. deneoformans* JEC21,
617 and *C. deuterogattii* R265 genomes. Ortho-groups specific or common to the different species
618 were identified and numbered.
619
620

621 *C. deuterogattii* R265 has previously been shown to lack a functional RNA interference
622 pathway (D'SOUZA *et al.* 2011; BILLMYRE *et al.* 2013). Accordingly, the genes encoding one Dicer
623 (*DCR1*) and an Argonaute protein (*AGO1*) have been lost, and the genes encoding an RNA
624 dependent RNA polymerase gene (*RPD1*) and an RNAi essential zinc finger protein (*ZNF3*) are
625 truncated and probably not functional in this strain (D'SOUZA *et al.* 2011; FERETZAKI *et al.* 2016). The

626 identification of truncated or absent genes in the R265 genome has been as a strategy to identify
627 additional, novel components of the RNAi pathway in *C. neoformans* (FERETZAKI *et al.* 2016). To
628 identify genes specifically lost in *C. deuterogattii*, we considered the genes not predicted by our
629 pipeline but present in the other *Cryptococcus* species annotations available in FungiDB (BASENKO
630 *et al.* 2018; FARRER ET AL. 2015; D'SOUZA ET AL. 2011). We considered here *C. tetragattii* strain
631 IND107, *C. bacillisporus* strain CA1873, and *C. gattii* strains WM276, NT-10, and EJB2; no *C.*
632 *decagatiii* annotation was available at the time of this study. We identified 17 ortho-groups that
633 were absent in the R265 genome but present in all other species (Table 1). As expected, one
634 ortho-group corresponds to an Argonaute protein (ortho-group OG0000415). We also confirmed
635 that the gene *FCZ28,* which encoded a transcription factor essential for the sex-induced-silencing
636 RNAi pathway in *C. neoformans*, was specifically absent in the R265 genome (FERETZAKI *et al.* 2016).
637 In contrast, the gene *GWO1*, previously identified as specifically lost in R265 and coding for an
638 Ago1-interacting protein (although deletion mutants have normal siRNA profiles) was not present
639 in this new list due to its absence in the IND107 genome (DUMESIC *et al.* 2013). The orthologue size
640 ratio analysis performed to pinpoint genome sequence mistakes eventually identified 119 R265
641 genes with a size ratio lower than 0.8 compared to both *C. neoformans* and *C. deneoformans*
642 orthologs or shortened in one species and absent in the other (Tables 2; S7). Although some loci
643 are likely pseudogenes, we decided not to annotate them as such because there is no strict
644 structural definition of what constitutes a pseudogene (TUTAR 2012) and we cannot evaluate the
645 functionality of a gene based on its structure alone.
646

H99 gene ID	R265 gene ID	Size ratio	Putative function	Role in RNAi	ref
CNAG_00505	absent		transcription factor (*FZC28*)	yes	(1)
CNAG_01061	absent		serine/threonine protein kinase (*FRK102*)	?	(2)
CNAG_02207	absent		glycosyl hydrolase	?	
CNAG_03734	absent		chromodomain-containing protein (*CDP1*)	?	(1)
CNAG_04016	absent		Identified spore protein 5 (*ISP5*)	?	(3)
CNAG_04596 , CNAG_04619	absent		prolyl endopeptidase	?	
CNAG_04609	absent		Argonaute protein (*AGO1*)	yes	(4)
CNAG_05158	absent		hypothetical protein	?	-
CNAG_05265	absent		hypothetical protein	?	
CNAG_05449	absent		copper metallothionein 1 (*MTN1*)	?	(5)
CNAG_05657	absent		2,4-dienoyl-CoA reductase	?	
CNAG_06233	absent		hypothetical protein	?	
CNAG_06395	absent		hypothetical protein	?	
CNAG_06609	absent		2-polyprenyl-6-methoxyphenol hydroxylase (*ORX1*)	?	(1)
CNAG_07556	absent		hypothetical protein	?	
CNAG_07702	absent		F-box containing protein	?	
CNAG_07959	absent		GTPase-activator protein (GAP)	?	
CNAG_03466	CNBG_2143	0.083	RNA-dependent RNA polymerase 1 (*RDP1*)	yes	(4)
CNAG_02700	CNBG_9326	0.092	C2H2-type zinc finger transcription factor (*ZNF3*)	yes	(1)
CNAG_01423	CNBG_5946	0.102	*QIP1*	yes	(6)
CNAG_04146	CNBG_2894	0.121	SET domain-containing protein	?	
CNAG_06486	CNBG_4982	0.137	*GWC1*	yes	(6)
CNAG_03911	CNBG_9603	0.139	carboxylesterase domain-containing protein	?	
CNAG_06497	CNBG_4974	0.162	microsomal epoxide hydrolase (*MEH1*)	?	(1)
CNAG_01992	CNBG_2960	0.190	SET domain-containing protein	?	
CNAG_03117	CNBG_2464	0.191	hypothetical protein	?	
CNAG_07344	CNBG_9031	0.197	Ras guanyl-nucleotide exchange factor	?	
CNAG_01406	CNBG_5961	0.198	hypothetical protein	?	
CNAG_03414	CNBG_10064	0.201	REX4-like exonuclease domain containing protein	?	
CNAG_04184	CNBG_2860	0.224	transcription factor (*FZC47*)	no	(1)
CNAG_03193	CNBG_9232	0.233	hypothetical protein	?	
CNAG_04400	CNBG_9268	0.234	Ribosomal protein S10p/S20e	?	
CNAG_03938	CNBG_5530	0.243	*Cryptococcus* pheromone receptor 2 (*CPR2*)	yes	(1)
CNAG_00123	CNBG_9042	0.286	hypothetical protein	?	
CNAG_06159	CNBG_4866	0.315	hypothetical protein	?	
CNAG_01004	CNBG_0584	0.316	Rho/Rac/Cdc42-like GTPases	?	
CNAG_06509	CNBG_4963	0.318	hypothetical protein	?	

Table 2: Genes with putative or known roles in RNAi identified as genes of H99 with orthologues in all *Cryptococcus* species but absent or severely truncated (and thus putative pseudogenes) in *C. deuterogattii* R265 as compared to JEC21 and H99 (proteins with a ratio <0.33 are presented. The full table of shortened genes is presented as Table S6). (1) FERETZAKI et al 2016; (2) Lee et al. 2016; (3) HUANG et al. 2015; (4) WANG et al. 2010 ; (5) DING et al. 2011 ; (6) DUMESIC et al. 2013.

As expected, the RNAi genes *RPD1 (WANG et al. 2010)*, *ZNF3 (FERETZAKI et al. 2016)*, *CPR2 (FERETZAKI et al. 2016)*, *QIP1 (DUMESIC et al. 2013)*, *GWC1 (DUMESIC et al. 2013)*, *RDE4,* and *RDE5* (BURKE *et al.* 2019) were present in this list, confirming that a large number RNAi genes are not functional or are absent in R265. Conversely, *RDE1*, *RDE2,* and *RDE3* (BURKE *et al.* 2019), which were recently implicated in RNAi in *C. neoformans*, all possess an orthologue of similar size in R265 (CNBG_3369, CNBG_4718, and CNBG_1922, respectively). Of note, in this version of the R265 annotation, the *DMT5* (CNBG_3156) gene encoding a putative DNA methyltransferase is not truncated as previously published (YADAV *et al.* 2018; CATANIA *et al.* 2020) and appears to be expressed and functional.

Gene organization, antisense transcription, and alternative splicing in R265

The absence of RNAi in R265 was recently shown to be associated with a modification of the chromosome structure: shorter centromeres and the loss of any full-length transposable elements (YADAV *et al.* 2018). Here, we examined the possible consequences of RNAi loss on the transcriptome aside from the expected absence of siRNA. We first hypothesized that the absence of RNAi in *C. deuterogattii* could result in increased antisense transcription, as it might be the source of double-stranded RNA; increased antisense transcription in RNAi-deficient species has also been observed in *Saccharomyces* species (ALCID and TSUKIYAMA 2016). We thus evaluated the sense/antisense transcript ratio at coding gene loci. Because the 3'UTR and TL sequences were only partially annotated in the R265 genome, we restricted our analysis to the CDS regions. We compared the ratio of read numbers of sense and antisense transcripts corresponding to all coding regions in *C. neoformans*, *C. deneoformans,* and *C. deuterogattii* under four growth conditions. When cells were grown to exponential growth phase at 30°C (E30), most of the expressed genes

(92.2 %) had antisense transcription in *C. neoformans*, but antisense transcripts were expressed at a very low level (1.2% of transcription antisense *vs* sense). Antisense transcription prevalence and expression levels were similar in the two other species (92.6% and 95.2% of genes with antisense transcription, 3.2% and 2.5% of antisense *vs* sense transcription in *C. deneoformans* and *C. deuterogattii,* respectively). These ratios changed in different growth conditions, particularly increased temperature at both exponential and stationary growth phase. However, this analysis did not provide evidence of a link between the level of antisense transcription and the absence of RNAi in R265 because the RNAi-proficient JEC21 strain had the highest antisense/sense transcription ratio across all conditions tested (Figure 6).

We then analyzed gene orientation in the three species, evaluating the number of genes coupled in a tail-to-tail orientation as this orientation should favor antisense transcription over head-to-tail or head-to-head orientations. Indeed, as shown in Figure 6B, there was a clear selection against tail-to-tail gene orientation in *C. deuterogattii*, thus limiting antisense transcription (c-squared = 103.79, df = 4, p-value < 2.2e-16). In contrast, this orientation is favored in JEC21, which might explain the higher level of antisense transcription.

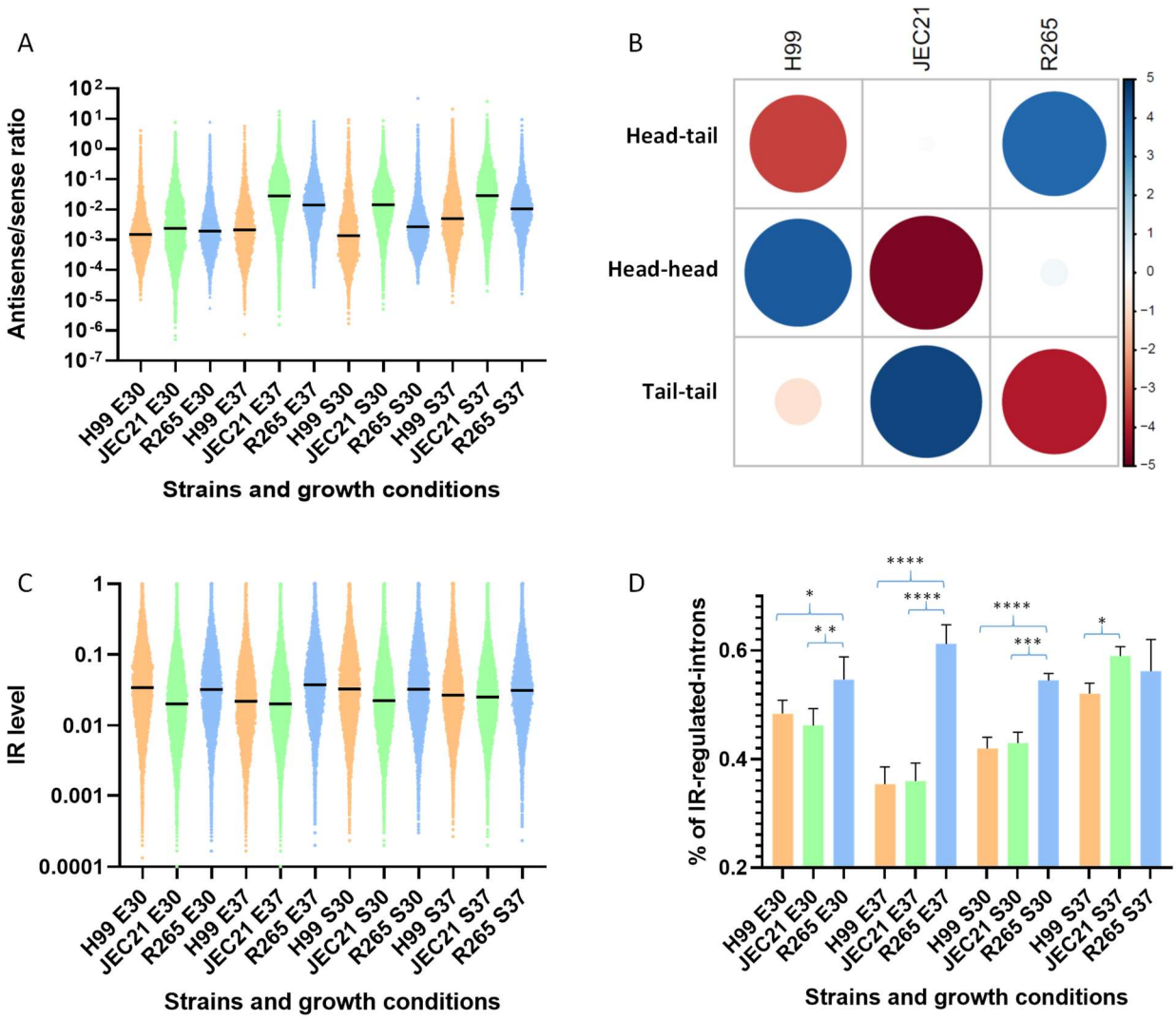
The SCANR model predicts that siRNAs are produced in response to poorly spliced introns that stall the spliceosome complex, which should result in lower levels of expression for the corresponding gene (DUMESIC *et al.* 2013). To explore whether loss of RNAi could have affected intron retention (IR) in *C. deuterogatiii*, we compared the number of CDS introns regulated by intron retention in this species and two RNAi-proficient ones. Interestingly, in three conditions the percentage of introns regulated by IR was higher in R265 than in JEC21 or H99 (Figure 6C). For instance, when cells were grown to exponential phase at 37°C, 44.5% of R265 introns are regulated by IR as compared to 21.7% and 20.8% in *C. neoformans* and *C. deneoformans*, respectively. In contrast, the IR indices were similar across the three species when cells were grown at 30°C. However, at 37°C in either exponential and stationary growth phase, the median

706 value of the IR index in R265 was higher than those in both *C. neoformans* and *C. deneoformans*.

707 Overall, these data suggest that IR is better tolerated in R265 than in H99 or JEC21; this result

708 aligns with the SCANR model of siRNA production and gene regulation in *Cryptococcus*.

709



710

711 Figure 6. (A) Antisense/sense transcription ratios in *C. neoformans* (H99), *C. deneoformans* (JEC21)

712 and *C. deuterogattii* (R265). RNA-seq data obtained from cells grown to exponential phase at 30°C

713 (E30) and 37°C (E37) or stationary phase at 30°C (S30) and 37°C (S37) were used. (B) Statistical

714 analysis (Pearson's Chi-squared test) revealed a species-specific bias in gene orientation. Circle size

715 is proportional to the standardized residuals, with absolute values higher than 2 representing

716 statistical significance (Sharpe 2015). Positive values (blue circles) in cells specify a positive

717 association between the corresponding row and column variables. Negative residuals are
718 represented by red circles. This implies a negative association between the corresponding row
719 and column variables. (C) Intron retention level in each species according to growth condition.
720 Black bars represent median values. (D) Percentage of CDS introns regulated by IR in each species
721 according to growth condition. The results of the statistical analysis (ANOVA one-way multi-
722 comparison analysis corrected by FDR). * (q value < 0.05), ** (q value < 0.01), *** q value <0.001),
723 **** (q value < 0.0001).
724

725 **Subtelomeric gene organization and cluster exchange in *Cryptococcus***

726 Our analysis identified 210 orthologue groups present in both *C. neoformans* and *C.*
727 *deneoformans* but absent in *C. deuterogattii*. Interestingly some of these lost genes are clustered
728 in these genomes. We identified 21 clusters of lost genes with consecutive elements in both *C.*
729 *neoformans* and *C. deneoformans* reference genomes. One of these lost clusters has been
730 previously described and has been reported to contain homologues of several *GAL* genes (*GAL1*,
731 *UGE2*, and *GAL7*) and a gene encoding a sugar transporter of the major facilitator superfamily
732 (MFS) (CNAG_07897) (SLOT AND ROKAS 2010). We also identified a fifth gene in this cluster
733 (CNAG_06055) which encodes a putative alpha-1,4-galactosidase (Figure S2A). *C. neoformans* and
734 *C. deneoformans* also possess unclusterered paralogues of the genes *UGE2* (*UGE1*, CNAG00697),
735 *GAL1* (*GAL101*, CNAG_03946), and *GAL7* (*GAL701*, CNAG_03875). Previous studies have shown
736 *UGE2* is required for growth on galactose, whereas it paralogous gene *UGE1* is necessary for
737 growth at 37°C and glucuronoxylomannogalactan (GXMGal) biosynthesis, which makes up an
738 important fraction of the *Cryptococcus* polysaccharide capsule (MOYRAND *et al.* 2008). Interestingly,
739 we previously reported that a *uge2Δ* mutant strain was able to grow on galactose at 37°C,
740 suggesting that *UGE1* is able to compensate in the absence of *UGE2* at 37°C. The GAL cluster with

five genes has also been lost in all other *Cryptococcus* species that were assessed. Thus, the *C. gattii* clade species possess the only non-clustered paralogues of the GAL pathway; yet, they are all able to grow on galactose as a sole carbon source, suggesting these genes are involved in both GXMGal synthesis and galactose assimilation in this species (Figure S2B).

Gene ontology (GO) term enrichment analysis (PRIEBE *et al.* 2014) of 52 genes within 18 non-subtelomeric clusters that were absent in R265 revealed a statistically significant enrichment of genes coding for proteins implicated in transport and transcription regulation (Figure 7A). Functional annotation of these genes confirmed this result (Table S7). We identified 13 clusters containing at least one gene coding for a putative transporter, including six MFS-type transporters, and eight clusters containing at least one gene coding for an annotated or putative transcription factor (TF), including six fungal Zn(2)-Cys(6) binuclear cluster domain-containing TFs. Overall, seven clusters contain both a transporter and a TF (Figure 7B, Figure S3). Strikingly, this association between transporters and TFs resembles the organization of primary metabolic gene clusters (MGCs) (ROKAS *et al.* 2018). Because three MGCs were located within subtelomeric loci, we compared the gene content within subtelomeric regions to the gene content of the lost clusters. We considered the 20 most distal genes of each chromosome arm in H99. GO term enrichment analysis of these 560 H99 subtelomeric genes revealed very similar profiles to the profiles obtained for the cluster genes. Again, genes coding for proteins implicated in transport in subtelomeric regions were significantly enriched (Figure 7C).

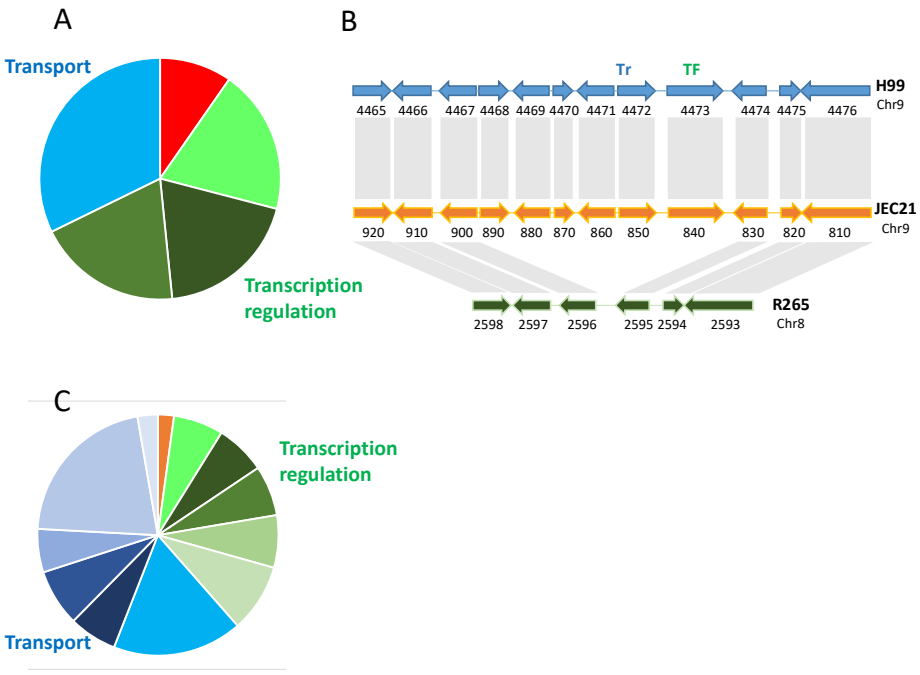


Figure 7. (A) GO term enrichment analysis of genes in clusters absent in R265. Green colors indicate GO terms associated with transcription regulation (GO:0006012, GO:0000981, GO:0006366, GO:0006357). Blue colors indicate GO terms associated with transport (GO:0055085). Orange colors indicate GO terms associated with galactose metabolism (GO:0006012). (B) Example of the organization of an MGC-like cluster absent in R265. CNAG_04468 (CNI00890) encodes a putative tartrate dehydrogenase, CNAG_04469 (CNI99880) encodes a putative 4-aminobutyrate transaminase, CNAG_04470 (CNI00870) encodes a putative halo-acid dehalogenase, CNAG_04471 (CNI00860) encodes an FAD-dependent oxidoreductase superfamily protein, CNAG_04472 (CNI00850) encodes an MFS protein, and CNAG_04473 (CNI00840) encodes a TF with a fungal Zn(2)-Cys(6) binuclear cluster domain. (C) GO term enrichment analysis of subtelomeric genes in H99. Green colors indicate GO terms associated with transcription regulation (GO:0051213, GO:0000981, GO:0006366, GO:0006357, GO:0006351, GO:0006355). Blue colors indicate GO terms associated with transport (GO:0055085, GO:0022891, GO:0022857, GO:0005215, GO:0016021, GO:0008643, GO:0006810). The orange color indicates a GO term associated with dioxygenase activity (GO:0051213).

Functional annotation of these subtelomeric genes confirmed this enrichment of transporters and TFs (Table S8). We found an unexpected number of genes encoding annotated or putative TFs (n = 33) and transporters (n = 68) within these regions of the H99 genome. Most of these TFs and transporters belong to the fungal Zn(2)-Cys(6) binuclear cluster domain-type (n= 24) and MFS-type (n= 49) families, respectively. Comparison of the organization of *C. neoformans* H99 subtelomeric loci to those in *C. deneoformans* JEC21 revealed a very similar organization, and only four mosaic subtelomeric regions were identified with genes from at least two different regions in H99; few genes were present in H99 but absent in JEC21 (Figure 8). However, we did identify two duplicated regions in the JEC21 subtelomeric regions. The first duplicated locus consists of six genes with orthologues in subtelomeric region of the left arm of Chr 5 in H99. The second duplicated region has been previously described (FRASER *et al.* 2005). It is located in the left arms of Chrs 8 and 12 and resulted from a telomere-telomere fusion that occurred during the construction of the JEC20/JEC21 congenic mating pair. Interestingly, a TF with a fungal Zn(2)-Cys(6) binuclear cluster domain (*FZC2*, CNAG_05255) and a putative amino acid transporter (CNAG_05254) are present within this repeated region. Conversely, genes in the subtelomeric regions of the right arms of H99 Chrs 4 and 14 are orthologues of genes located within a central part of JEC21 Chr 8 (Figure S3), suggesting an additional telomere-telomere fusion event. In contrast, the subtelomeric regions in R265 have undergone more rearrangements compared to JEC21 – in R265 there are fifteen mosaic subtelomeric regions that contain genes from at least two different regions in H99. We also identified nine genes within six R265 subtelomeric regions whose orthologues are located far from the telomeres in H99. Interestingly, functional annotation of the R265-specific subtelomeric gene clusters (n = 12) (Figure 8; Table S9), revealed an enrichment of genes encoding TFs (n = 2) and transporters (n = 6).

801

802

803

804

805

806

807

808

***C. neoformans* H99**

Left arm Right arm

1
2
3
4
5
6
7
8
9
10
11
12
13
14

***C. deneoformans* JEC21**

Left arm Right arm

1
2
3
4
5
6
7
8
9
10
11
12
13
14

duplication

***C. deuterogattii* R265**

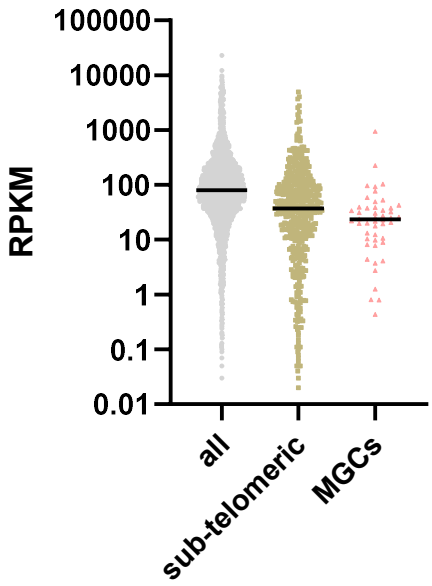
Left arm Right arm

1
2
3
4
5
6
7
8
9
10
11
12
13
14

809

Figure 8. Subtelomeric gene organization in *Cryptococcus*. The 20 most distal genes at each subtelomeric locus were considered. The color code identifies each subtelomeric regions in H99 and orthologous genes in the other species. The positions of these orthologues in the H99 subtelomeric regions are given (TEL-RX or TEL-LX correspond to genes positioned within the right or left arm of chromosome X). When the orthologous gene is not located within a subtelomeric region, its locus named is given. Black boxes correspond to genes present in *C. deneoformans* or *C. deuterogattii* but absent in *C. neoformans*. Red and green boxes indicate duplicated sets of genes. Blue dots indicate transporters. Green dots indicate transcription factors.

Subtelomeric regions have been shown to be silenced by H3K27me3 histone modifications in *C. neoformans*, and a large number of genes that are upregulated upon deletion of the H3K27 methyltransferase (encoded by *EZH2*) are located within subtelomeric regions (DUMESIC *et al.* 2015). Accordingly, we observed that expression of the 580 most proximal genes was generally lower than the expression of the most telomere-distal genes (Figure 9). Interestingly, we found that the H99 genes present within the MGCs that were lost in R265 were also poorly expressed. However, none of these genes were upregulated upon *EZH2* deletion (DUMESIC *et al.* 2015), suggesting that they are not directly regulated by H3K7me3. In summary, these data suggest that dynamic exchanges of MGCs between subtelomeric regions occurred during *Cryptococcus* speciation. These results also suggest that MGC exchanges between subtelomeric loci and more central parts of chromosomes might be associated with new assimilation capacities.



831

832 Figure 9. Expression of genes according to their position on the chromosome. Subtelomeric genes

833 are the 20 most distal genes of each chromosome arm. The H99 genes present within the non-

834 subtelomeric cluster of genes lost in R265 are indicated as MGCs.

835

836

Discussion

Although a number of bioinformatic pipelines have been published in recent years, accurate annotation of fungal genomes is still difficult due to their complexity and compactness (HAAS *et al.* 2011). In this study, we have carefully optimized a previously published Cufflinks-CodingQuarry-based annotation pipeline and tested it on two complex fungal genomes. This pipeline largely outcompeted the BRAKER1 pipeline when applied to two *Cryptococcus* reference genomes (*C. neoformans* H99 and *C. deneoformans* JEC21) and would likely outcompete many other pipelines use to annotate fungal genomes *de novo* (MIN *et al.* 2017).

Our optimization process revealed three notable points. First and counterintuitively, increasing the quantity of data did not always result in a better annotation. This is likely because Cufflinks tends to make huge clusters when large data sets are input; these clusters might be eliminated during the transcript identification step. Accordingly, we found that the number of predicted transcripts decreased when too much data was used. Second, we found a nearly linear relationship between the number of introns predicted and the quality of the annotation. However, this correlation did not hold when two of the pipelines were compared; the BRAKER pipeline predicted more introns than the C3Q pipeline, along with predicting many more genes. Nevertheless, the correlation between intron number and annotation quality provided a facile way to evaluate the reliability of a *de novo* annotation, which might be affected, for instance, by the quality of the RNA-seq data. Third, we found the final step of comparative genomics did not always improve the quality of the annotation. In our assay, the Exonerate-based analysis step using the whole proteome of a reference species primarily introduced errors into the annotation. This was probably due to the fact that even when manually curating genome annotations, a number of dubious genes remain, which are then transferred to the new genome annotation. In fact, a systematic usage of a comparative annotation step following a *de novo* RNA-seq annotation

would likely result in a dramatic expansion of dubiously annotated genes in fungal genomes. Accordingly, it is noticeable that the number of predicted coding genes in R265 (n=6405) is lower than the ones predicted in H99 (n=6795) and JEC21 (n=6639) although we ignore whether these differences have some biological relevance or if they are due to the different strategies used to annotate these genomes.

During the annotation of the R265 genome, we manually curated a subset of genes that were lost in R265 compared to all of the other *Cryptococcus* species as well as a set of putative pseudogenes. The identification of genes specifically lost or pseudogenized in R265 has previously been used as a strategy to identify novel RNAi components in *C. neoformans* (FERETZAKI *et al.* 2016). Accordingly, most known RNAi genes are present in these sets of lost and pseudogenized genes (BILLMYRE *et al.* 2013). However, some genes, like *RDE1* (BURKE *et al.* 2019), which is necessary for siRNA production, are present and functional in R265, suggesting that it may have roles independent from RNAi silencing. On the other hand, *GWO1*, which is also considered to be an RNAi pathway component, is also absent in the *C. tetragattii* strain IND107 and is therefore absent in our list as well. One possible explanation is that Gwo1 alone or in complex with Ago1 could play another role independent of RNAi. Another possibility is that a Gwo1-dependent RNAi pathway has also been lost in *C. tetragattii*. Nevertheless, this analysis confirms that looking for specific gene loss in a fungal species deficient for a certain pathway remains a promising strategy for the identification of genes implicated in this pathway in other proficient species. In the present case, it would be interesting to see how many of the R265 truncated genes are functional in other *C. gattii* species, although it would demand a complex manual curation, which is beyond the scope of this paper.

Our study revealed that loss of RNAi in R265 is associated with few general transcriptome modifications compared to the transcriptomes of JEC21 and H99, aside from the predictable

absence of siRNA. This might be because we did not annotate most non-coding features like lncRNAs, transcript leaders, and 3'UTRs. Yet, quantification of the sense/antisense transcription ratio at CDS did not reveal any differences between R265 and the other *Cryptococcus* species analyzed, suggesting that this ratio does not depend on the RNAi status of the species in this genus. This is in agreement with the fact that siRNAs in *C. neoformans* primarily target transposons and retrotransposons (JANBON *et al.* 2010; WANG *et al.* 2010; DUMESIC *et al.* 2013), whereas antisense transcription is associated with nearly all of the genes as we have shown. This result also suggests antisense transcription in *Cryptococcus* only rarely results in the production of double-stranded RNA. Dumesic and colleagues showed that delayed splicing is a source of siRNA production in *C. neoformans* (DUMESIC *et al.* 2013). We thus anticipated that the absence of RNAi would increase the level of intron retention. In agreement with previous reports in *C. deneoformans*, we found that IR level was regulated by growth conditions in both *C. neoformans* and *C. deuterogattii* (GONZALEZ-HILARION *et al.* 2016). However, the number of introns regulated by IR was markedly larger in R265 suggesting that IR is better tolerated in this RNAi-deficient species. We also expected that some compensatory mechanisms might be acting to control the level of IR because IR rates were largely similar across the three species analyzed even though it was higher in R265, at least at 37°C. It is important to note the remarkable effect of temperature on both IR and antisense transcription, which might be related to a recent report that transposons are specifically mobilized at this temperature in *C. deneoformans* (GUSA *et al.* 2020).

While most loci that have been lost in R265 compared to other *Cryptococcus* species contain only a single gene, we also identified gene clusters that were missing in R265. Analysis of the gene content within these clusters revealed a strong enrichment of genes coding for proteins implicated in transport and transcriptional regulation. This finding was reminiscent of patterns identified in metabolic gene clusters (MGCs) involved in primary metabolism, which typically contain transcription factors and transporters (ROKAS *et al.* 2018). MGCs can be lost or gained in

910 fungi and several examples of instances of horizontal transfer of whole MGCs from one species to
911 another have been published (SLOT AND ROKAS 2010; ROKAS *et al.* 2018; WANG *et al.* 2019). In
912 filamentous fungi, the majority of MGCs are located within subtelomeric regions, which are largely
913 subjected to inter-chromosomal reshuffling (GLADIEUX *et al.* 2014). Two examples of lineage-
914 specific gene clusters harboring both transcription factors and transporters have been previously
915 reported in *C. neoformans* (RHODES *et al.* 2017), suggesting dynamic gene cluster gain and loss
916 events even with a single species in *Cryptococcus*. Interestingly, these *C. neoformans* lineage-
917 specific clusters also contain transcription factors and transporters (RHODES *et al.* 2017). A more
918 recent report suggests that genes within one of *these C. neoformans* clusters are co-regulated, as
919 is expected from a typical MGC (YU *et al.* 2020). In *Cryptococcus*, we found that the subtelomeric
920 regions were also enriched for characteristic MGC genes as well, and comparisons of subtelomeric
921 gene organization across the three *Cryptococcus* species suggested active reshuffling. This was in
922 agreement with previous data showing that subtelomeric genes are under strong evolutionary
923 pressure in *Cryptococcus* (DESJARDINS *et al.* 2017). We found a large number of genes encoding
924 transporters and TFs of unknown function in *Cryptococcus* subtelomeric regions. Surprisingly, most
925 of the TF genes identified in these MGCs within subtelomeric regions as well as in MGCs far from
926 telomeres were not annotated as TFs and were not included when a systematic TF deletion
927 collection was constructed and studied (JUNG *et al.* 2015). It therefore seems that the TF repertoire
928 in *Cryptococcus* may be larger than currently appreciated. Similarly, besides myo-inositol
929 transporters, which have been previously reported to be localized within subtelomeric regions
930 (XUE *et al.* 2010), the substrates of most transporters located in these regions remain unknown.
931 Genes within subtelomeric regions are silenced by H3K27me3 epigenetic marks and,
932 accordingly, are expressed at lower levels than genes located in more central regions of the
933 chromosomes. Similarly, genes within the subtelomeric clusters lost in R265 were poorly
934 expressed. Yet, their expression levels did not significantly change following deletion of the gene

935 encoding the H3K27me3 methyltransferase *EZH2,* suggesting they are either not regulated by
936 H3K27me3 or that additional changes are needed to activate expression of these genes like those
937 previously described in *Fusarium graminearum* (CONNOLLY *et al.* 2013). If this is the case, the
938 regulation of *GAL* genes by galactose might represent a good example of how genes within the
939 MGCs could be regulated in *Cryptococcus* (WICKES AND EDMAN 1995; MOYRAND *et al.* 2008; RUFF *et al.*
940 2009). Besides the *GAL* cluster, the function and regulation of most of MGC genes in *Cryptococcus*
941 are unknown. Nevertheless, our results suggest active exchange between subtelomeric regions
942 and more central parts of chromosomes in *Cryptococcus*, potentially reshaping primary
943 metabolism for adaptation to different environmental niches. They also emphasize how both
944 complete genome and precise annotations are needed to study these dynamics in fungi.
945

Acknowledgements

PAGF was supported a CAPES exchange grant (Advanced Network of Computational Biology - RABICÓ - Grant no. 23038.010041/2013-13). This work was supported by a CAPES COFECUB grant n°39712ZK to GJ, by a CNPq grant (309897/2017-3) to CCS, and by a NIH/NIAID R37 MERIT Award AI39115-23 and a NIH/NIAID R01 Award AI50113-16 to J.H. J.H. is co-director and fellow of CIFAR program Fungal Kingdom: Threats & Opportunities. S.J.P was supported by the NIH/NIAID F31 Fellowship 1F31AI143136-02. Members of the Heitman laboratory are acknowledged for valuable discussion.

957

958

959

960 **References**

961

962 Alcid, E. A. and T. Tsukiyama, 2016 Expansion of antisense lncRNA transcriptomes in budding yeast
963 species since the loss of RNAi. *Nature Structural and Molecular Biology* 23**:** 450-455.

964 Anders, S., P. T. Pyl and W. Huber, 2014 HTSeq—a Python framework to work with high-throughput
965 sequencing data. *Bioinformatics* 31**:** 166-169.

966 Basenko, E. Y., J. A. Pulman, A. Shanmugasundram, O. S. Harb, K. Crouch *et al.*, 2018 FungiDB: an
967 integrated bioinformatic resource for fungi and oomycetes. *Journal of Fungi* (Basel,
968 Switzerland) 4**:** 39.

969 Billmyre, R. B., S. Calo, M. Feretzaki, X. Wang and J. Heitman, 2013 RNAi function, diversity, and loss
970 in the fungal kingdom. *Chromosome Research* 21**:** 561-572.

971 Burke, J. E., A. D. Longhurst, P. Natarajan, B. Rao, J. Liu *et al.*, 2019 A non-Dicer RNase III and four
972 other novel factors required for RNAi-mediated transposon suppression in the human
973 pathogenic yeast *Cryptococcus neof*ormans. *G3* (Bethesda, Md.) 9**:** 2235-2244.

974 Cantarel, B. L., I. Korf, S. M. C. Robb, G. Parra, E. Ross *et al.*, 2008 MAKER: an easy-to-use annotation
975 pipeline designed for emerging model organism genomes. *Genome Research* 18**:** 188-196.

976 Carver, T., S. R. Harris, M. Berriman, J. Parkhill and J. A. McQuillan, 2012 Artemis: an integrated
977 platform for visualization and analysis of high-throughput sequence-based experimental
978 data. *Bioinformatics* 28.

979 Catania, S., P. A. Dumesic, H. Pimentel, A. Nasif, C. I. Stoddard *et al.*, 2020 Evolutionary persistence
980 of DNA methylation for millions of years after ancient loss of a *de novo* methyltransferase.
981 *Cell* 180**:** 263-277.e220.

Cheng, P.-Y., A. Sham and J. W. Kronstad, 2009 *Cryptococcus gattii* isolates from the British Columbia cryptococcosis outbreak induce less protective inflammation in a murine model of infection than *Cryptococcus neoformans*. *Infection and Immunity* 77**:** 4284.

Conesa, A., S. Götz, J. M. García-Gómez, J. Terol, M. Talón *et al.*, 2005 Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21**:** 3674-3676.

Connolly, L. R., K. M. Smith and M. Freitag, 2013 The *Fusarium graminearum* histone H3 K27 methyltransferase KMT6 regulates development and expression of secondary metabolite gene clusters. *PLOS Genetics* 9**:** e1003916.

D'Souza, C. A., J. W. Kronstad, G. Taylor, R. Warren, M. Yuen *et al.*, 2011 Genome variation in *Cryptococcus gattii*, an emerging pathogen of immunocompetent hosts. *mBio* 2**:** e00342-00310.

Dal Molin, A., A. Minio, F. Griggio, M. Delledonne, A. Infantino *et al.*, 2018 The genome assembly of the fungal pathogen *Pyrenochaeta lycopersici* from Single-Molecule Real-Time sequencing sheds new light on its biological complexity. *PLOS One* 13**:** e0200217.

Desjardins, C. A., C. Giamberardino, S. M. Sykes, C.-H. Yu, J. L. Tenor *et al.*, 2017 Population genomics and the evolution of virulence in the fungal pathogen Cryptococcus neoformans. *Genome research* 27**:** 1207-1219.

Ding, C., Yin, J., Tovar, E.M.M., Fitzpatrick, D.A., Higgins, D.G. and Thiele, D.J. (2011) The copper regulon of the human fungal pathogen *Cryptococcus neoformans* H99. *Mol Microbiol*, **81**, 1560-1576.

Dumesic, Phillip A., Christina M. Homer, James J. Moresco, Lindsey R. Pack, Erin K. Shanle *et al.*, 2015 Product binding enforces the genomic specificity of a yeast polycomb repressive complex. *Cell* 160**:** 204-218.

1006 Dumesic, P. A., P. Natarajan, C. Chen, I. A. Drinnenberg, B. J. Schiller *et al.*, 2013 Stalled spliceosomes
1007 are a signal for RNAi-mediated genome defense. *Cell* 152**:** 957-968.
1008 Dunn, N. A., D. R. Unni, C. Diesh, M. Munoz-Torres, N. L. Harris *et al.*, 2019 Apollo: Democratizing
1009 genome annotation. *PLOS Computational Biology* 15**:** e1006790.
1010 Emms, D. M., and S. Kelly, 2019 OrthoFinder: phylogenetic orthology inference for comparative
1011 genomics. *Genome Biology* 20: 238.
1012 Farrer, R. A., C. A. Desjardins, S. Sakthikumar, S. Gujja, S. Saif *et al.*, 2015 Genome evolution and
1013 innovation across the four major lineages of *Cryptococcus gattii*. *mBio* 6**:** e00868-00815.
1014 Farrer, R.A., M. Chang, M.J. Davis, L. van Dorp, D. Yang *et al*., 2019. A new lineage of *Cryptococcus*
1015 *gattii* (VGV) discovered in the Central Zambezian Miombo Woodlands. *mBio* 10:e02306-19.
1016 Feretzaki, M., R. B. Billmyre, S. A. Clancey, X. Wang and J. Heitman, 2016 Gene network
1017 polymorphism illuminates loss and retention of novel RNAi silencing components in the
1018 *Cryptococcus* pathogenic species complex. *PLOS Genetics* 12**:** e1005868-e1005868.
1019 Ferrareze, P. A. G., R. S. A. Streit, P. R. dos Santos, F. M. dos Santos, R. M. C. de Almeida *et al.*, 2017
1020 Transcriptional analysis allows genome reannotation and reveals that *Cryptococcus gattii*
1021 VGII undergoes nutrient restriction during Infection. *Microorganisms* 5**:** 49.
1022 Fraser, J. A., J. C. Huang, R. Pukkila-Worley, J. A. Alspaugh, T. G. Mitchell *et al.*, 2005 Chromosomal
1023 translocation and segmental duplication in *Cryptococcus neoformans*. *Eukaryotic Cell* 4**:** 401-
1024 406.
1025 Giordano, F., L. Aigrain, M. A. Quail, P. Coupland, J. K. Bonfield *et al.*, 2017 *De novo* yeast genome
1026 assemblies from MinION, PacBio and MiSeq platforms. *Scientific Reports* 7**:** 3935.
1027 Gladieux, P., J. Ropars, H. Badouin, A. Branca, G. Aguileta *et al.*, 2014 Fungal evolutionary genomics
1028 provides insight into the mechanisms of adaptive divergence in eukaryotes. *Molecular*
1029 *Ecology* 23**:** 753-773.

1030 Gonzalez-Hilarion, S., D. Paulet, K.-T. Lee, C.-C. Hon, P. Lechat *et al.*, 2016 Intron retention-
1031 dependent gene regulation in *Cryptococcus neoformans*. *Scientific Reports* 6**:** 32252.
1032 Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson *et al.*, 2011 Full-length
1033 transcriptome assembly from RNA-seq data without a reference genome. *Nature*
1034 *Biotechnology* 29**:** 644-652.
1035 Gusa, A., J. D. Williams, J.-E. Cho, A. F. Averette, S. Sun *et al.*, 2020 Transposon mobilization in the
1036 human fungal pathogen *Cryptococcus* is mutagenic during infection and promotes drug
1037 resistance in vitro. *Proceedings of the National Academy of Sciences of the USA* 117**:** 9973.
1038 Haas, B. J., Q. Zeng, M. D. Pearson, C. A. Cuomo and J. R. Wortman, 2011 Approaches to fungal
1039 genome annotation. *Mycology* 2**:** 118-141.
1040 Hagen, F., K. Khayhan, B. Theelen, A. Kolecka, I. Polacheck *et al.*, 2015 Recognition of seven species
1041 in the *Cryptococcus gattii*/*Cryptococcus neoformans* species complex. *Fungal Genetics and*
1042 *Biology* 78**:** 16-48.
1043 Haridas, S., A. Salamov and I. V. Grigoriev, 2018 Fungal Genome Annotation, pp. 171-184 in *Fungal*
1044 *Genomics: Methods and Protocols*, edited by R. P. de Vries, A. Tsang and I. V. Grigoriev.
1045 Springer New York, New York, NY.
1046 Hoff, K. J., S. Lange, A. Lomsadze, M. Borodovsky and M. Stanke, 2016 BRAKER1: unsupervised RNA-
1047 seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* (Oxford,
1048 England) 32**:** 767-769.
1049 Huang, M., Hebert, A.S., Coon, J.J. and Hull, C.M. (2015) Protein composition of infectious spores
1050 reveals novel sexual development and germination factors in *Cryptococcus*. *PLoS Genet*, **11**,
1051 e1005490-e1005490.
1052 Huston, S. M., S. S. Li, D. Stack, M. Timm-McCann, G. J. Jones *et al.*, 2013 *Cryptococcus gattii* is killed
1053 by dendritic cells, but evades adaptive immunity by failing to induce dendritic cell
1054 maturation. *The Journal of Immunology* 191**:** 249.

James, T. Y., J. E. Stajich, C. T. Hittinger and A. Rokas, 2020 Toward a fully resolved fungal tree of life. *Annual Review of Microbiology* 74.

Janbon, G., 2018 Introns in *Cryptococcus*. *Memorias do Instituto Oswaldo Cruz* 113**:** e170519-e170519.

Janbon, G., S. Y. Maeng, D-F., Y.-J. Ko, K.-W. Jung, F. Moyrand *et al.*, 2010 Characterizing the role of RNA silencing components in *Cryptococcus neoformans*. *Fungal Genetics and Biology* 47**:** 1070-1080.

Janbon, G., K. L. Ormerod, D. Paulet, E. J. Byrnes III, G. Chatterjee *et al.*, 2014 Analysis of the genome and transcriptome of *Cryptococcus neoformans* var. *grubii* reveals complex RNA expression and microevolution leading to virulence attenuation. *PLOS Genetics* 10**:** e1004261.

Jung, K.-W., D.-H. Yang, S. Maeng, K.-T. Lee, Y.-S. So *et al.*, 2015 Systematic functional profiling of transcription factor networks in *Cryptococcus neoformans*. *Nature Communications* 6**:** 6757.

Kidd, S. E., F. Hagen, R. L. Tscharke, M. Huynh, K. H. Bartlett *et al.*, 2004 A rare genotype of *Cryptococcus gattii* caused the cryptococcosis outbreak on Vancouver Island (British Columbia, Canada). *Proceeding of the National Academy of Sciences of the USA* 101**:** 17258-17263.

Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley *et al.*, 2013 TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* 14**:** R36-R36.

Kwon-Chung, K. J., J. A. Fraser, T. L. Doering, Z. Wang, G. Janbon *et al.*, 2014 *Cryptococcus neoformans* and *Cryptococcus gattii*, the etiologic agents of cryptococcosis. *Cold Spring Harbor Perspectives in Medicine* 4**:** a019760.

Lam, W. C., R. Upadhya, C. A. Specht, A. E. Ragsdale, C. R. Hole *et al.*, 2019 Chitosan biosynthesis and virulence in the human fungal pathogen *Cryptococcus gattii*. *mSphere* 4**:** e00644-00619.

1079 Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. *Nature Methods*
1080 9**:** 357-359.
1081 Lee, K.-T., So, Y.-S., Yang, D.-H., Jung, K.-W., Choi, J., *et al.* (2016) Systematic functional analysis of
1082 kinases in the fungal pathogen *Cryptococcus neoformans*. *Nature Communications*, **7**, 12766.
1083 Lomsadze, A., P. D. Burns and M. Borodovsky, 2014 Integration of mapped RNA-seq reads into
1084 automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research* 42**:** e119-
1085 e119.
1086 Lowe, T. M., and P. P. Chan, 2016 tRNAscan-SE On-line: integrating search and context for analysis
1087 of transfer RNA genes. *Nucleic Acids Research* 44**:** W54-W57.
1088 Ma, H., F. Hagen, D. J. Stekel, S. A. Johnston, E. Sionov *et al.*, 2009 The fatal fungal outbreak on
1089 Vancouver Island in characterized by enhanced intracellular parasitism driven by
1090 mitochondrial regulation. *Proceeding of the National Academy of Sciences of the USA* 106**:**
1091 12980-12985.
1092 Martin, M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads.
1093 EMBnet.journal; Vol 17, No 1: Next Generation Sequencing Data Analysis.
1094 Min, B., I. V. Grigoriev and I.-G. Choi, 2017 FunGAP: Fungal Genome Annotation Pipeline using
1095 evidence-based gene model evaluation. *Bioinformatics* 33**:** 2936-2937.
1096 Moyrand, F., I. Lafontaine, T. Fontaine and G. Janbon, 2008 *UGE1* and *UGE2* regulate the UDP-
1097 glucose/UDP-galactose equilibrium in *Cryptococcus neoformans*. *Eukaryotic Cell* 7**:** 2069-
1098 2077.
1099 Mudge, J. M., and J. Harrow, 2016 The state of play in higher eukaryote gene annotation. *Nature*
1100 *Reviews Genetics* 17**:** 758-772.
1101 Ngamskulrungroj, P., Y. Chang, E. Sionov and K. J. Kwon-Chung, 2012 The primary target organ of
1102 *Cryptococcus gattii* Is different from that of *Cryptococcus neoformans* in a murine model.
1103 *mBio* 3**:** e00103-00112.

Pertea, G., and M. Pertea, 2020 GFF Utilities: GffRead and GffCompare. *F1000Research* 9**:** 304.

Priebe, S., C. Kreisel, F. Horn, R. Guthke and J. Linde, 2014 FungiFun2: a comprehensive online resource for systematic analysis of gene lists from fungal species. *Bioinformatics* 31**:** 445-446.

Rhodes, J., C. A. Desjardins, S. M. Sykes, M. A. Beale, M. Vanhove *et al.*, 2017 Tracing genetic exchange and biogeography of *Cryptococcus neoformans* var. *grubii* at the global population level. *Genetics***:** 10.1534/genetics.1117.203836.

Rokas, A., J. H. Wisecaver and A. L. Lind, 2018 The birth, evolution and death of metabolic gene clusters in fungi. *Nature Reviews Microbiology* 16**:** 731-744.

Ruff, J. A., J. K. Lodge and L. G. Baker, 2009 Three galactose inducible promoters for use in *C. neoformans* var. *grubii*. *Fungal Genetics and Biology* 46**:** 9-16.

Semighini C.P., A.F. Averette, J.R. Perfect, and J. Heitman, 2011. Deletion of *Cryptococcus neoformans* AIF ortholog promotes chromosome aneuploidy and fluconazole-resistance in a metacaspase-independent manner. *PLOS Pathogens* 7: e1002364.

Sharpe, D. 2015 Chi-Square Test is Statistically Significant: Now What?, Practical Assessment. *Research, and Evaluation*: Vol. 20 , Article 8. DOI: https://doi.org/10.7275/tbfa-x148.

Shen, X.-X., D. A. Opulente, J. Kominek, X. Zhou, J. L. Steenwyk *et al.*, 2018 Tempo and mode of genome evolution in the budding yeast subphylum. *Cell* 175**:** 1533-1545.e1520.

Slater, G. S., and E. Birney, 2005 Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics***:** 31.

Slot, J. C., and A. Rokas, 2010 Multiple *GAL* pathway gene clusters evolved independently and by different mechanisms in fungi. *Proceeding of the National Academy of Sciences of the USA* 107**:** 10136-10141.

Stanke, M., M. Diekhans, R. Baertsch and D. Haussler, 2008 Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* 24**:** 637-644.

Testa, A. C., J. K. Hane, S. R. Ellwood and R. P. Oliver, 2015 CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. *BMC Genomics* 16**:** 170.

Törönen, P., A. Medlar and L. Holm, 2018 PANNZER2: a rapid functional annotation web server. *Nucleic Acids Research* 46**:** W84-W88.

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan *et al.*, 2010 Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology 28**:** 511-515.

Tutar, Y., 2012 Pseudogenes. *Comparative and Functional Genomics* 2012**:** 424526-424526.

Wallace, E. W. J., C. Maufrais, J. Sales-Lee, L. R. Tuck, L. de Oliveira *et al.*, 2020 Quantitative global studies reveal differential translational control by start codon context across the fungal kingdom. *Nucleic Acids Research* 48**:** 2312-2331.

Wang, M., H. Fu and R. Ruan, 2019 A small horizontally transferred gene cluster contributes to the sporulation of *Alternaria alternata*. *Genome Biology and Evolution* 11**:** 3436-3444.

Wang, X., Y. P. Hsueh, W. Li, A. Floyd, R. Skalsky *et al.*, 2010 Sex-induced silencing defends the genome of *Cryptococcus neoformans* via RNAi. *Genes & Development* 24**:** 2566-2582.

Wickes, B. L., and J. C. Edman, 1995 The *Cryptococcus neoformans GAL7* gene and its use as an inducible promoter. *Molecular Microbiology* 16**:** 1099-1109.

Wilm, A., D. G. Higgins, F. Valentin, G. Blackshields, H. McWilliam *et al.*, 2007 Clustal W and Clustal X version 2.0. *Bioinformatics* 23**:** 2947-2948.

Xue, C., T. Liu, L. Chen, W. Li, I. Liu *et al.*, 2010 Role of an expanded inositol transporter repertoire in *Cryptococcus neoformans* sexual reproduction and virulence. *mBio* 1**:** e00084-00010.

Yadav, V., S. Sun, R. B. Billmyre, B. C. Thimmappa, T. Shea *et al.*, 2018 RNAi is a critical determinant of centromere evolution in closely related fungi. *Proceeding of the National Academy of Sciences of the USA* 115**:** 3108.

1154 Yu, C.-H., Y. Chen, C. A. Desjardins, J. L. Tenor, D. L. Toffaletti *et al.*, 2020 Landscape of gene
1155 expression variation of natural isolates of Cryptococcus neoformans in response to
1156 biologically relevant stresses. *Microbial Genomics* 6.
1157 Zhu, P., B. Zhai, X. Lin, and A. Idnurm, 2013. Congenic strains for genetic analysis of virulence traits
1158 in *Cryptococcus gattii*. *Infection and Immunity* 81: 2616-2625.
1159
1160