UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

CENTRO DE BIOTECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA CELULAR E MOLECULAR

Análises transcritômicas em *Cryptococcus gattii* e seu impacto na definição de genes e outros elementos genômicos

Patrícia Aline Gröhs Ferrareze

Tese submetida ao Programa de Pós-Graduação em Biologia Celular e Molecular do Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul como requisito parcial para obtenção do título de Doutor em Ciências.

Orientador: Prof. Dr. Charley Christian Staats

Porto Alegre, setembro de 2019.

Este trabalho foi desenvolvido no Laboratório de Microbiologia Molecular e Celular, situado no Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul; e na *Unité Biologie des ARN des Pathogènes Fongiques*, situada no *Département Mycologie* do *Institut Pasteur*. Para seu desenvolvimento, este trabalho contou com fomento do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

AGRADECIMENTOS

- Ao professor Dr. Charley C. Staats, meu orientador, pela oportunidade, paciência, confiança e por todos os ensinamentos;
- À professora Dra. Marilene H. Vainstein, por ter aberto as portas do PPGBCM
- Ao professor Dr. Henrique Ferreira, membro da comissão de acompanhamento;
- Ao Dr. Guilhem Janbon, meu supervisor no Institut Pasteur;
- A Frédérique Moyrand e Corinne Maufrais, colegas do Institut Pasteur;
- Aos colegas do Laboratório 219, UBTEC, e do grupo de *Cryptococcus;* em especial, ao Rodrigo Streit, pela amizade e todas as ricas discussões científicas;

- Aos meus pais, que tem sido a base de toda minha jornada;
- A Tailine e ao Felipe, pelo apoio incondicional;

- A todos que contribuíram direta ou indiretamente na construção deste trabalho.

SUMÁRIO

LISTA DE ABREVIATURAS 6

LISTA DE FIGURAS 9

RESUMO 10

ABSTRACT 11

1. INTRODUÇÃO GERAL 12

1.1 Impacto do sequenciamento de genomas no entendimento da fisiologia de fungos 12

1.1.1 Estudos de transcritomas fúngicos 17

1.1.2 Impacto de transcritomas na anotação gênica 21

1.2 Elementos transponíveis 25

1.3 RNA de Interferência 31

1.4 Espécies patogênicas do gênero *Cryptococcus*: aspectos funcionais, estudos genômicos e transcritômicos 38

2. OBJETIVOS 47

2.1 Objetivos Específicos 47

3. RESULTADOS 48

3.1 Capítulo I: sRNAs as possible regulators of retrotransoposon activity in *Cryptococcus gattii* VGII 49

3.2 Capítulo II: Transcriptional analysis allows genome reannotation and reveals that *Cryptococcus gattii* VGII undergoes nutrient restriction during Infection 64

3.3 Capítulo III: Development of an automated and optimized pipeline for gene prediction in *Cryptococcus gattii* R265 80

4 DISCUSSÃO GERAL 122

5 CONCLUSÃO 131

6 REFERÊNCIAS BIBLIOGRÁFICAS 132

7 CURRICULUM VITAE RESUMIDO..... 140

LISTA DE ABREVIATURAS

AFLP: Polimorfismo de comprimento de fragmento amplificado, do inglês, *Amplified Fragment Length Polymorphism*

AGO: Argonauta

cDNA: DNA complementar, do inglês, *complementary DNA*

disiRNA: pequeno RNA interferente independente de Dicer, do inglês, *dicer-independent small interfering RNA*

dsRNA: RNA dupla fita, do inglês, *double strand RNA*

EST: marcador de sequência expressa, do inglês, *Expressed Sequence Tag*

ex-siRNA: pequeno RNA interferente exônico, do inglês, *exonic small interfering RNA*

GAG: antígeno grupo-específico, do inglês, *Groupspecific AntiGen*

Gb: Gigabases

kb: Kilobases

LINE: elementos nucleares longos intercalados, do inglês, *Long Interspersed Nuclear Elements*

lncRNA: RNA longo não-codificante, do inglês, *long non-coding RNA*

LTR: repetições longas terminais, do inglês, *Long Terminal Repeats*

Mb: Megabases

milRNA: micro RNA-tipo, do inglês, *miRNA-like*

miRNA: micro RNA

MLMT: tipagem de microsatélite multilocus , do inglês, *Multilocus microsatellite typing*

MLST: tipagem de sequência multilocus, do inglês, *Multilocus sequence typing*

mRNA: RNA mensageiro, do inglês, *messenger RNA*

MSUD: silenciamento meiotico de DNA não-pareado, do inglês, *Meiotic Silencing of Umpaired DNA*

ncRNA: RNA não-codificante, do inglês, *non-coding RNA*

nt: nucleotídeo

ORF: fase aberta de leitura, do inglês, *Open Reading Frame*

PCR: reação em cadeia da polimerase, do inglês, *Polimerase Chain Reaction*

piRNA: Piwi RNA

rasiRNAs: pequeno RNA interferente associado a repetição, do inglês, *repeat associated small interfering RNA*

RdRP: RNA polimerase dependente de RNA, do inglês, *RNA-dependent RNA polimerase*

RFLP: Polimorfismo de comprimento de fragmento de restrição, do inglês, *Restriction Fragment Length Polimorfism*

RIP: mutação pontual induzida por repetição, do inglês, *Repeat Induced Point mutation*

RISC: complexo de silenciamento induzido por RNA, do inglês, *RNA Induced Silencing Complex*

RITS: silenciamento transcricional induzido por RNA, do inglês, *RNA Induced Transcriptional Silencing*

RNAi: RNA de interferência, do inglês, *interference RNA*

RNA-Seq: sequenciamento de RNA, do inglês, *RNA sequencing*

rRNA: RNA ribossomal, do inglês, *ribosomal RNA*

SCAMP: proteína carreadora secretora de membrana, do inglês, *Secretory Carrier Membrane Protein*

SNP: polimorfismo de nucleotídeo único, do inglês, *Single Nucleotide Polymorphism*

SINE: elementos nucleares curtos intercalados, do inglês, *Short Interspersed Nuclear Elements*

siRNA: pequeno RNA interferente, do inglês, *small interfering RNA*

SIS: Silenciamento induzido pelo sexo, do inglês, *Sex-Induced Silencing*

snoRNA: pequeno RNA nucleolar, do inglês, *small nucleolar RNA*

snRNA: pequeno RNA nuclear, do inglês, *small nuclear RNA*

sRNA: pequeno RNA, do inglês, *small RNA*

SSP: pequena proteína secretada, do inglês, *Small Secreted Protein*

TE: element transponível, do inglês, *Transposable Element*

TIR: repetição invertida terminal, do inglês, *Terminal Inverted Repeat*

tRNA: RNA transportador, do inglês, *transporter RNA*

TSD: duplicação do sítio alvo, do inglês, *Target Site Duplication*

UTR: região não traduzida, do inglês, *Untranslated Region*

VG: variedade *gattii*, do inglês, *variety gattii*

VLPs: partículas tipo vírus, do inglês, *Virus-Like Particles*

VN: variedade *neoformans*, do inglês, *variety neoformans*

WGST: tipagem de sequência de genoma inteiro, do inglês, *Whole-Genome Sequence Typing*

YR: tirosina recombinase, do inglês, *Tyrosine Recombinase*

LISTA DE FIGURAS

Figura 1. Mecanismos de transposição de elementos móveis no genoma 26

Figura 2. Filogenia e quantificação de elementos transponíveis de dezoito espécies fúngicas 30

Figura 3. Presença dos componentes da via de RNAi em diferentes espécies fúngicas 37

Figura 4. Diversidade no complexo de espécies *C. gattii/C. neoformans* inferido de um conjunto concatenado de 11 loci 41

RESUMO

Cryptococcus gattii é um dos agentes etiológicos da criptococose, uma doença com cerca de 1 milhão de novos casos anuais e aproximadamente 625 mil óbitos. Embora a espécie irmã *C. neoformans* cause meningite criptocócica em indivíduos imunodeprimidos e seja a maior responsável pelo grande número de casos diagnosticados; *C. gattii* infecta indivíduos imunocompetentes causando infecção pulmonar. Neste contexto, *C. gattii* tem recebido crescente atenção desde o surto do Noroeste do Pacífico, tendo-se reunido esforços para identificar as variantes genéticas das linhagens que provocaram tal evento. Devido as diferenças entre genótipos, hipotetiza-se que a ausência da maquinaria do RNA de interferência seja um dos mecanismos causadores de mutação e consequente hipervirulência, já que as linhagens do genótipo VGII responsável pelo surto não possuem RNAi funcional; e mutantes de componentes do RNAi em *C. neoformans* podem apresentar tal fenótipo. Assim, teve este estudo o objetivo de (i) identificar a presença de retrotransposons – principais alvos da via do RNAi para a defesa da integridade genômica – em linhagens de *C. gattii*; (ii) avaliar a expressão de retrotransposons e a presença de sRNAs associados na linhagem R265 – representante hipervirulenta do subgrupo VGII; (iii) analisar o perfil metabólico da linhagem R265 durante infecção pulmonar em modelo murino de criptococose; e (iv) desenvolver uma pipeline automática de predição gênica otimizada e validada com *C. neoformans*, para aplicação em *C. gattii* R265. Como resultados, observou-se que linhagens do genótipo VGII apresentavam um menor número de sequências preditas de retrotransposons em seus genomas, quando comparadas a linhagens do VGI, a qual possui RNAi funcional. Além disso, pode-se inferir a presença de sRNAs associados a estas sequências, e a possível regulação das mesmas pela distribuição de sRNAs nas regiões LTR e regiões entre domínios. O perfil metabólico de *C. gattii* R265 durante a infecção murina, por sua vez, demonstrou a alta expressão de genes associados a vias de metabolismo de aminoácidos. Tais proteínas atuam na utilização de carbono e nitrogênio como fonte para geração de intermediários do metabolismo de carboidratos em um cenário de restrição de glicose imposto como sistema de defesa do hospedeiro. Por fim, o desenvolvimento de uma *pipeline* de predição gênica automatizada e otimizada resultou em um sistema com cerca de 70% dos genes codificantes de *C. gattii* R265 corretamente anotados.

ABSTRACT

Cryptococcus gattii is one of the etiologic agents of cryptococcosis, a disease with about 1 million new annual cases and approximately 625,000 deaths. Although sister species *C. neoformans* causes cryptococcal meningitis in immunocompromised individuals and is the major responsible for the large number of diagnosed cases; *C. gattii* infects immunocompetent individuals causing lung infection. In this context, *C. gattii* has received increasing attention since the Pacific Northwest outbreak, and efforts have been made to identify the genetic variants of *C. gattii* strains that caused such an event. Due to the differences between genotypes, it is hypothesized that the absence of interference RNA machinery is one of the mechanisms that cause mutation and consequent hypervirulence, since the genotypes of the VGII genotype responsible for the outbreak do not have a functional RNAi; and mutants of RNAi components in *C. neoformans* may exhibit such a phenotype. Thus, this study aimed to (i) identify the presence of retrotransposons - main targets of the RNAi pathway for the defense of genomic integrity - in *C. gattii* strains; (ii) evaluate the expression of retrotransposons and the presence of associated sRNAs in R265 strain - hypervirulent representative of subgroup VGII; (iii) analyze the metabolic profile of the R265 strain during lung infection in a murine model of cryptococcosis; and (iv) develop an automatic gene prediction pipeline, optimized and validated with *C. neoformans* for application on *C. gattii* R265. As a result, it was observed that VGII genotype strains presented fewer predicted retrotransposon sequences in their genomes when compared to VGI strains, which have functional RNAi. Furthermore, was possible to infer the presence of sRNAs associated with these sequences and their possible regulation by the distribution of sRNAs in LTR regions and regions between domains. The metabolic profile of *C. gattii* R265 during murine infection, in turn, demonstrated the high expression of genes associated with amino acid metabolism pathways. These proteins act on the use of carbon and nitrogen as a source for the generation of carbohydrate metabolism intermediates in a glucose restriction scenario imposed as a host defense system. Finally, the development of an automated and optimized gene prediction pipeline resulted in a system with about 70% of the correctly annotated *C. gattii* R265 coding genes.

1 INTRODUÇÃO

1.1 Impacto do sequenciamento de genomas no entendimento da fisiologia fúngica

Em 1975, Sanger e colaboradores realizaram o primeiro sequenciamento de um genoma, com a determinação da sequência de DNA de 5,4 kb do bacteriófago ϕX174, empregando nucleotídeos radio-marcados. Mais tarde, em 1986, Thomas H. Roderick cunhou o termo “genômica”, como o campo da genética e da biologia molecular para o estudo e aplicação de técnicas de sequenciamento de DNA, bem como o uso de ferramentas computacionais para a montagem, anotação e análise de genomas (HEATHER; CHAIN, 2016; KUO; BUSHNELL; GRIGORIEV, 2014).

A genômica de fungos, especificamente, estabeleceu-se a partir de 1992, com o sequenciamento do cromossomo III da levedura *Saccharomyces cerevisiae*, o qual foi, por sua vez, o primeiro genoma fúngico a ter a sua sequência determinada, em 1996. Posteriormente, a realização do sequenciamento dos genomas de *Schizosaccharomyces pombe*, *Neurospora crassa* e *Phanerochaete chrysosporium*; bem como projetos mais amplos para determinação da sequência de genomas de fungos, como do *Fungal Genome Initiative* (Broad Institute), em 2000, *Génolevures*, em 2004, e *1000 Fungal Genomes Project*, tem contribuído para um maior conhecimento e entendimento acerca de genomas fúngicos (KUO; BUSHNELL; GRIGORIEV, 2014).

Com um número estimado entre 1.5 e 5 milhões de espécies, o grupo monofilético dos Eumycota (“fungos verdadeiros”) tem uma história evolutiva estimada em mais de 1 bilhão de anos, abrangendo organismos heterotróficos com uma variada gama de interações mutualistas, comensais ou parasitas (MOHANTA; BAE, 2015). Estas, como fator e produto da seleção purificadora, positiva e neutra sobre a evolução da estrutura e organização gênica, levaram ao desenvolvimento de ciclos de vida, metabolismos, morfogêneses e ecologias inteiramente distintos (CHOI; KIM, 2017; STAJICH, 2017).

Neste contexto, diferentes interesses e necessidades são relacionados à crescente demanda de sequenciamento de genomas fúngicos, os quais, muitas vezes, funcionam como modelos mais simples para estudos de sistemas eucarióticos. O baixo custo e maior rapidez das tecnologias de sequenciamento de segunda geração (*second-generation* ou *next-generation sequencing* – NGS) como SOLiD, Illumina/Solexa e Ion Torrent em comparação com os métodos Sanger tradicionais; bem como a maior disponibilidade de ferramentas de bioinformática, tem incentivado o desenvolvimento de inúmeras pesquisas com os patógenos humanos *Candida* (JONES et al., 2004), *Coccidioides* (SHARPTON et al., 2009), *Histoplasma* (SEPÚLVEDA et al., 2017; SHARPTON et al., 2009) e *Aspergillus* (HAGIWARA et al., 2014; MACHIDA et al., 2005); dentre outros (SHARMA, 2015).

Recentemente, o desenvolvimento de tecnologias de sequenciamento de terceira geração com os *reads* longos e contíguos do Pacific Biosciences (PacBio) e Oxford Nanopore Technologies tem contribuído com a montagem e anotação de genomas eucarióticos maiores, mais complexos e com alta qualidade (COOK et al.,

2019). Outra vantagem refere-se ao sequenciamento de genomas de fungos dicarióticos: devido à alta heterozigosidade, as montagens genômicas a partir de *reads* curtos tornam-se altamente fragmentadas e sequências específicas de cada haplótipo são perdidas pelo colapso de *reads* e formação de sequências consenso. Para tanto, sequenciamentos de *reads* longos constituem uma alternativa já amplamente utilizada em diversos trabalhos, como na investigação de polimorfismos associados as diferenças entre linhagens clínicas e ambientais de *Penicillium capsulatum* (YANG et al., 2016) e na identificação de mutações que conferem resistência a fungicidas em *Botrytis cinerea* (KAN et al., 2017).

Sequências genômicas, desta forma, possibilitam análises e predições *in silico,* identificando e caracterizando sequências codificantes, transposons, regiões regulatórias, polimorfismos e outros elementos que afetam a plasticidade genômica, a qual pode conferir diferentes fenótipos frente as pressões seletivas e adaptativas do meio externo*.* O sequenciamento de genomas, assim sendo, possibilitou a identificação da base genética associada a diversos fenótipos, como, por exemplo, a expansão da família de receptores acoplados à proteína G no fitopatógeno *Magnaporthe grisea* (DEAN et al., 2005); a ausência de famílias gênicas com domínios de ligação à celulose em ascomicetos Onygenales (SHARPTON et al., 2009); a ausência do cluster de biossíntese da fumitremorgina no genoma de *Aspergillus novofumigatus* (KJÆRBØLLING et al., 2018); e a expansão da família gênica contendo domínio para subtilisina N em *Coccidioides* (SHARPTON et al., 2009).

Cerca de 1000 espécies fúngicas tem seus genomas sequenciados e publicamente disponíveis. Entretanto, apesar de mais de 90% das espécies fúngicas

conhecidas pertencerem aos filos Ascomycota e Basidiomycota, apenas 85 famílias em Ascomycota e 66 em Basidiomycota possuem mais de um genoma representativo, sendo a proporção de espécies sequenciadas correspondente ao número de espécies conhecidas; e, grande parte destas, espécies patogênicas (AYLWARD et al., 2017).

Embora *S. cerevisiae* tenha sido o primeiro genoma fúngico sequenciado, montado e anotado, devido à importância biotecnológica da espécie, sua estrutura genômica é relativamente simples, não sendo representativa para a maioria dos genomas fúngicos, os quais, muitas vezes, são compostos por grandes e numerosas regiões repetitivas, genes longos e ricos em íntrons, duplicações gênicas, polimorfismos e eventos de *splicing* alternativo (KUO; BUSHNELL; GRIGORIEV, 2014).

Assim, o sequenciamento de diferentes genomas, abrangendo microsporídeos (tamanho do genoma de cerca de 2 Mb) a Pucciniales (Basidiomycota) (tamanho do genoma de cerca de 900 Mb), possibilita a investigação dos fatores relacionados a tais expansões e reduções genômicas, quando a maior parte dos genomas publicados apresentam tamanhos entre 30 e 40 Mb. Portanto, expansão e redução genômica tem sido, diversas vezes, diretamente associados ao fenótipo. Eventos de redução são comumente relacionados a genomas de parasitas obrigatórios, como resultado da pressão seletiva da interação patógeno-hospedeiro, e evolução da especificidade. Como resultado, estes eventos podem levar ao aumento de cópias de genes, bem como a contração de famílias gênicas, como observado na alteração entre os perfis generalista e especialista, em espécies de *Metarhizium*. Os eventos independentes de redução genômica em fungos leveduriformes como *Schizosaccharomyces* e

Cryptococcus, todavia, são implicados como resultado da transição para a forma de vida unicelular (STAJICH, 2017).

A expansão genômica, diferentemente, pode ser consequência do relaxamento da seleção purificadora em decorrência do tamanho populacional efetivo da espécie, de modo que adições deletérias ao genoma podem acumular em populações menores (AYLWARD et al., 2017; MOHANTA; BAE, 2015; STAJICH, 2017). Assim como expansões genômicas em espécies patogênicas de plantas podem ser decorrentes da propagação de retrotransposons e outros elementos móveis, com a manutenção de semelhante número de genes codificantes a espécies de genomas menores (STAJICH, 2017).

Outro importante aspecto relacionado aos sequenciamentos de genomas, diz respeito aos estudos evolutivos produzidos a partir de tais dados. Estudos evolutivos são normalmente gerados através de análises filogenéticas de um restrito número de genes ortólogos altamente conservados. No entanto, este sistema pode não representar a verdadeira relação evolutiva entre organismos, já que não refletem todo conjunto gênico de uma espécie. A análise em nível genômico, desta forma, fornece diversas informações valiosas para o entendimento da fisiologia, bem como das relações evolutivas, refletidas em eventos de macro e microescala. Estudos filogenéticos com o pool completo de sequências de genes codificantes de diversos genomas de espécies de *Microsporidia* desafiam as relações de parentesco com outros grupos fúngicos; apontando uma possível origem de um ancestral protista (CHOI; KIM, 2017). Ademais, informações precisas de estruturas gênicas podem facilitar o delineamento da base molecular das variações fenotípicas ao longo da

evolução, pela identificação de eventos que alterem as taxas evolutivas, bem como levem a perda ou aquisição de novos genes ou vias metabólicas; e, por consequência, a divergência e adaptação de organismos geneticamente próximos a ciclos de vida e processos biológicos distintos; bem como a convergência de diferentes organismos a modos de vida semelhantes (STAJICH, 2017).

1.1.1 Estudos de transcritomas fúngicos

A transcritômica é o estudo de todo o conjunto de moléculas de RNAs sintetizados por uma célula ou um conjunto de células. Proposta em 1996, por Charles Auffray, a transcritômica constitui parte fundamental da genômica para a interpretação dos elementos funcionais do genoma (BHADAURIA et al., 2007; SRIVASTAVA; GEORGE; KARUTURI, 2018; WANG; GERSTEIN; SNYDER, 2009). Assim, a transcritômica tem por objetivos mais comuns: (i) identificar e caracterizar moléculas de RNA, como mRNAs, RNAs não codificantes e pequenos RNAs; (ii) reconhecer e validar a correta estrutura gênica (extremidades 5’ e 3’, regiões não traduzidas - UTRs, sítios de início da transcrição e sequências codificantes, íntrons, éxons, sítios de *splicing* canônicos *e splicing* alternativos*,* dentre outros); e (iii) detectar transcritos gênicos, e sua relativa abundância, durante diferentes fases do ciclo celular, bem como diferentes condições biológicas (WANG; GERSTEIN; SNYDER, 2009).

Desde o surgimento dos *microarrays*, na década de 1990, tem sido possível analisar os padrões de expressão de genes conhecidos (codificantes ou não-codificantes), em diferentes populações de células e condições. Fundamentados pela

técnica de hibridização dupla fita, empregando oligonucleotídeos ou sondas marcadas de cDNA ancoradas a superfície para detecção de sinal, os *microarrays* conseguem detectar a expressão de transcritos hibridizados (WANG; GERSTEIN; SNYDER, 2009). Contudo, foi apenas com o desenvolvimento de tecnologias como o sequenciamento de RNA (RNA-Seq) que a obtenção de dados mais precisos de quantificação da expressão gênica (sem o viés de genes muito ou pouco expressos), bem como a identificação de novos genes e isoformas, com a detecção de padrões de *splicing* alternativo ao longo do desenvolvimento celular, foi possibilitada. (SRIVASTAVA; GEORGE; KARUTURI, 2018).

Neste contexto, a análise de dados de RNA-Seq de diferentes espécies fúngicas possibilitou a confirmação da presença de isoformas gênicas para todas as espécies analisadas, com taxas entre 0.82% (*Candida parapsilosis*) a 38.82% (*Histoplasma capsulatum*) dos genes. A avaliação dos diferentes padrões de *splicing* alternativo, a partir das informações obtidas pelo mapeamento dos *reads* contra os genomas constatou, ainda, a prevalência do mecanismo de retenção de íntrons em fungos; bem como a maior frequência destes eventos em condições de estresse, afetando, assim, genes de diversas funções regulatórias (SIEBER et al., 2018).

Eventos essencialmente procarióticos, como a transcrição de mRNAs policistrônicos, têm sido reconhecidos em espécies de fungos pelo sequenciamento de *reads* longos. Em algumas espécies de basidiomicetos, cujos genomas são ricos em íntrons, até 8% dos genes transcritos são policistrônicos. Em *N. crassa*, no entanto, unidades de transcrição policistrônicas não foram encontradas, e especula-se a associação com o mecanismo de silenciamento induzido por mutação (RIP), devido

ao enriquecimento de duplicações gênicas *in tandem* nessas unidades (GORDON et al., 2015).

A descoberta de novos transcritos de RNA na década de 1980, após o estabelecimento do dogma central da biologia molecular (RNA como intermediário entre DNA e proteína), possibilitou a caracterização de novas funções associadas a estas moléculas, como a atividade catalítica sob forma de ribozimas e regulação da expressão gênica por micro RNAs (miRNAs) e RNAs longos não codificantes (lncRNAs) (SRIVASTAVA; GEORGE; KARUTURI, 2018).

Assim, o conjunto inicial de mRNAs, tRNAs e rRNAs deu lugar a um grupo maior e mais diverso de transcritos, o qual inclui os pequenos RNAs nucleares (snRNA), pequenos RNAs nucleolares (snoRNA), pequenos RNAs de interferência (siRNAs), miRNAs, lncRNAs e pseudogenes. Enquanto ncRNAs como tRNAs, rRNAs, snRNAs e snoRNAs podem ser constitutivamente expressos e associados a viabilidade celular, ncRNAs regulatórios são comumente expresso em estágios específicos do desenvolvimento, ou em resposta a condições ambientais (DONALDSON; SAVILLE, 2012). Em espécies como *Ustilago maydis*, a avaliação de ncRNAs tem identificado a presença de transcritos antisenso (NATs), os quais apresentam complementariedade de sequência com mRNAs; e cuja deleção de promotores associados é relacionada a redução da virulência (DONALDSON; SAVILLE, 2013). Por outro lado, em *S. cerevisiae*, NATs são produzidos em resposta à disponibilidade de nitrogênio, fontes alternativas de carbono e retrotransposons Ty1 (DONALDSON; SAVILLE, 2012).

RNAs atuam em uma variada gama de atividades celulares, com padrões de expressão/síntese específicos e fornecem uma grande quantidade de informações

acerca da fisiologia; de forma que variações na composição do transcritoma auxiliam no entendimento de alterações e mecanismos celulares de resposta em processos biológicos (SRIVASTAVA; GEORGE; KARUTURI, 2018). A transição do enriquecimento funcional de processos relacionados à tradução, divisão celular, metabolismo primário e respiração, nos primeiros dias de infecção; para indução de catabolismo proteico e autofagia, em fitopatógenos como *U. maydis*, representa bem a relação entre o transcriptoma e o ciclo de vida (LANVER et al., 2018). Da mesma forma, a transição de levedura para fungo filamentoso, em espécies patogênicas da *Candida* (MUÑOZ et al., 2019; WU et al., 2016) mostra que a alteração da expressão de genes associados a filamentação e remodelamento da parede celular levam a diminuição do estado pró inflamatório das células hospedeiras (MUÑOZ et al., 2019).

Desta forma, tem a transcritômica o intuito de entender como diferentes organismos respondem às mudanças ambientais. E neste contexto, é importante ressaltar que tanto a qualidade do genoma sequenciado, quanto a acurácia da anotação desempenham papel determinante em diversas análises; especialmente aquelas que dependem de métodos sensíveis a erros de sequência, como o estudo da evolução de genes sujeitos a seleção positiva associados a virulência e adaptação (MUÑOZ et al., 2014).

1.1.2 Impacto de transcritomas na anotação gênica

Genomas fúngicos diferem de genomas de eucariotos superiores, especialmente no que diz respeito à densidade gênica e tamanho de regiões

intrônicas. Nestes aspectos, genes fúngicos apresentam íntrons menores, e menos eventos de *splicing* alternativo; embora a retenção de íntrons ainda seja o evento de *splicing* alternativo predominante neste grupo. Quanto à densidade gênica, observa-se frequentemente em genomas de fungos, a sobreposição de UTRs de genes adjacentes no genoma; especialmente sobreposições senso-antisenso (S-AS). Nestes casos, protocolos de sequenciamento que permitam a identificação da fita de origem do transcrito (RNA-Seq *stranded*) auxiliam na detecção da correta estrutura gênica (TESTA et al., 2015).

Anotações de genomas fúngicos têm sido comumente derivadas de predições *ab initio*, ou seja, a predição das estruturas gênicas baseia-se em sequências do genoma; ou pela homologia de sequência proteica com espécies relacionadas. Para predições *ab initio*, modelos estatísticos são treinados para encontrar assinaturas de éxons, íntrons, sítios de *splicing*, códons de ínicio e terminação e de DNA não codificante de regiões intergênicas. O treinamento destes modelos, por sua vez, é geralmente baseado em estruturas gênicas conhecidas de espécies de referência, ou de espécies relacionadas que possuam genomas anotados (HAAS et al., 2011). Em genomas de leveduras, por exemplo, a inferência de estruturas é relativamente simples, dada a pequena quantidade de genes com introns (~4%). Contudo, para a anotação completa destes organismos, o uso de outros modelos gênicos de fungos leva à predição de numerosos íntrons não existentes (PROUX-WÉRA et al., 2012).

Embora métodos de predição baseados em homologia constituam uma forte evidência para a localização de proteínas ortólogas em genomas de espécies relacionadas; para muitos fungos, espécies próximas não foram analisadas de forma

acurada, de modo que o conjunto de dados passível de ser utilizado é pequeno ou não confiável. Ademais, a ausência de homologia, ou a baixa similaridade de sequências pode levar a predições parciais de estruturas gênicas; bem como a não predição de sequências específicas de um genoma. Ferramentas como o BLAST, embora rápidas, não são altamente específicas para a detecção dos limites de éxons, pois não preconizam a identificação de *Open Reading Frames* (ORFs) nas sequências avaliadas. Assim, programas, como Exonerate e GeneWise, são frequentemente usados para o alinhamento de sequências de proteínas de referência a outros genomas, a fim de obter informações mais precisas a respeito de sítios de *splicing* e estrutura gênica (YANDELL; ENCE, 2012).

A necessidade de sistemas de predição acurados, automatizados e que possam se adaptar às características específicas de cada genoma, por consequência, é recorrente em um cenário em que a anotação manual ainda é considerada o método mais confiável para a produção de uma anotação de alta qualidade. Assim, a possibilidade de incorporação de dados de RNA-Seq, que reflitam não apenas o estado fisiológico do organismo, mas também as características reais de composição e funcionalidade de diferentes genes e genomas, torna-se um desafio na otimização destas *pipelines* (PROUX-WÉRA et al., 2012; TESTA et al., 2015).

Diversos programas utilizam montagens de transcritos a partir de *reads* de RNA-Seq para a correção e validação das coordenadas de íntrons e éxons em estruturas gênicas preditas. Igualmente, tais dados podem ser usados para o treinamento destes programas em predições *ab initio*. A utilização de bibliotecas de RNA-Seq para a predição gênica geralmente é precedida pela montagem dos

transcritos. Este processo, por sua vez, pode ser realizado de forma dependente ou independente do genoma. A forma independente, através de uma montagem *de novo* dos *reads* sequenciados, com ferramentas como ABySS, SOAPdenovo e Trinity; pede o posterior alinhamento destes transcritos ao genoma, como ESTs. A montagem dependente, entretanto, realiza o mapeamento dos *reads* sequenciados diretamente contra o genoma, com programas alinhadores como TopHat, GMAP e STAR; seguido pela montagem dos alinhamentos em transcritos, com programas como o Cufflinks (YANDELL; ENCE, 2012).

Preditores híbridos, como o programa CodingQuarry, atuam em duas etapas. A primeira etapa consiste da incorporação de predições diretas das sequências de transcritos (geradas pelo programa Cufflinks) de regiões suportadas por dados de RNA-Seq. A segunda etapa, utiliza tais dados no treinamento do algoritmo de identificação (cuja estatística é baseada no modelo oculto de Markov) para predições com a sequência genômica (*ab initio*), a partir da fixação de valores específicos para o reconhecimento dos códons de início e parada, sequência codificante e sítios de *splicing* (TESTA et al., 2015).

O programa AUGUSTUS, por sua vez, permite ao usuário incorporar dados de RNA-Seq, os quais são utilizados para melhorar a acurácia da predição, através de treinamento do algoritmo de identificação. A pipeline automática BRAKER1, de forma similar, realiza o treinamento iterativo e predição das estruturas gênicas iniciais com o programa GeneMark-ET; e, posteriormente, utiliza os genes preditos para o treinamento da predição *ab initio*, com a integração de informações do RNA-Seq à predição final (HOFF et al., 2016).

Nos últimos anos, diversos estudos têm usado dados de RNA-Seq para a correção de estruturas gênicas anotadas *ab initio*, identificação de regiões UTR, isoformas, sítios de *splicing* alternativo, polimorfismos e novos genes; além de validar estudos filogenéticos que relacionam alterações fisiológicas e especiação a eventos de retração e expansão de famílias gênicas (MUSZEWSKA et al., 2017; WU et al., 2016). A anotação do genoma de *Histoplasma capsulatum* G186A com o software PASA, por exemplo, levou à predição de 9359 genes, dos quais 77% apresentam íntrons. Neste mesmo estudo, a comparação de estruturas de genes discordantes preditas *ab initio* e estruturas baseadas em RNA-Seq demonstrou que esta última melhora significativamente a acurácia da predição da estrutura gênica (EDWARDS et al., 2013).

No mesmo contexto, a montagem do genoma e anotação de três linhagens do fitopatógeno *Leptosphaeria maculans* levou à predição de mais de 12000 genes codificantes de proteínas; dentre os quais um expressivo número de genes efetores codificantes de pequenas proteínas secretadas (SSPs). Geralmente presentes em fases específicas do parasitismo, as SSPs não são comumente identificadas durante a predição gênica, de forma que o número aumentado de sequências reflete a maior acurácia da anotação baseada em RNA-Seq (DUTREUX et al., 2018).

1.2 Elementos transponíveis

Uma das grandes possibilidades que as tecnologias de sequenciamento e montagem de genomas proporcionaram foi a análise da composição e efeito de

sequências repetitivas, especialmente elementos transponíveis, sobre a estabilidade genômica e expressão gênica. Elementos transponíveis (TE) são sequências repetitivas de DNA que se propagam no genoma através de mecanismos “copia e cola” (TE de classe I) ou “corta e cola” (TE de classe II) (Figura 1).

Elementos transponíveis de Classe II, classicamente, codificam transposases de domínio DDE para os processos de transposição de sequências de DNA. No entanto, podem apresentar diferentes mecanismos em algumas famílias de elementos. A presença de motivos conservados no domínio nuclease, bem como as características das repetições invertidas terminais (TIRs) e duplicações curtas (4–6 pb) geradas em sítios alvo (TSDs) determinam as famílias dos elementos, as quais são agrupadas conforme o mecanismo de transposição. Transposons de DNA das superfamílias Tc1/Mariner, hAT, MULE, PiggyBac, PIF-Harbinger e Merlin, mais comumente encontrados em fungos, utilizam uma nuclease DDE ou transposase para a clivagem da sequência dupla fita de DNA do genoma e posterior inserção no sítio alvo. Elementos do tipo Maverick/Polintons, por sua vez, apresentam replicação do tipo círculo rolante; ao passo que Helitrons não tem seu mecanismo de transposição completamente elucidado (MUSZEWSKA et al., 2017).

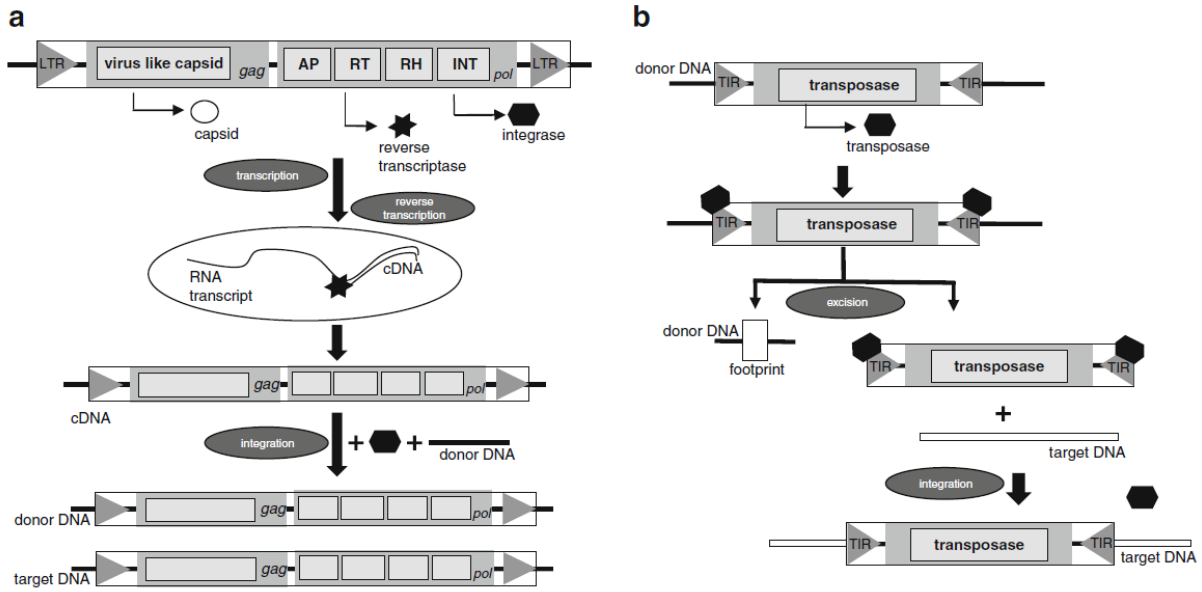

Figura 1. Mecanismos de transposição de elementos móveis no genoma. A Replicação clássica de retrotransposons LTR (Classe I). **B** Replicação clássica de transposons de DNA (Classe II). Abreviações LTR (repetições terminais longas), AP (aspartil protease), RT (transcriptase reversa), RH (RNase H), INT (integrasse), TIR (repetições terminais invertidas). Obtido de (PAUN; KEMPKEN, 2015).

Elementos transponíveis de classe I, ou retrotransposons, são replicados através de um intermediário de RNA (Figura 1), sendo, geralmente, sequências autônomas que codificam sua própria transcriptase reversa e tem, tipicamente entre 5 a 10 kb de tamanho. Estas sequências podem ser divididas em cinco principais classes: (i) retrotransposons LTR (*Long Terminal Repeats*); (ii) retrotransposons não-LTR ou LINE (*Long Interspersed Nuclear Elements*); (iii) elementos não-autônomos SINE (*Short Interspersed Nuclear Elements*); (iv) retotransposons PLE (*Penelope-like elements*); e (iv) retrotransposons YR (“tirosina recombinase”). Dentre as cinco classes, os retrotransposons LTR constituem o grupo predominante de elementos móveis nos genomas fúngicos. No entanto, isso se deve, possivelmente, a algumas

características intrínsecas a sequência e composição das famílias de elementos (CHÉNAIS et al., 2012; GOODWIN; POULTER, 2001).

Retrotransposons LINE são compostos por duas ORFs, a primeira codifica uma proteína de ligação ao RNA, a qual se associa com o transcrito intermediário. A segunda, codifica a nuclease e transcriptase reversa, podendo também codificar a RNAse H. Embora sejam os elementos mais semelhantes a organização dos retrotransposons LTR, não possuem repetições terminais e tem sido mais associado a genomas de eucariotos superiores, especialmente animais. Nestes organismos, apresentam taxas evolutivas menores, se comparados a retrotransposons de genomas fúngicos. Além disso, não compõem mais de 0.5% de alguns genomas representativos, os quais possuem, geralmente, entre 10 e 15% de DNA repetitivo. *S. cerevisiae*, bem como *S. pombe*, não apresentam retrotransposons não-LTR preditos (NOVIKOVA; FET; BLINOV, 2009). Elementos SINE, de forma contrária, são pequenas sequências não autônomas, de aproximadamente 300 pb, os quais dependem de proteínas codificadas por outros elementos para a sua própria transposição. Possuem características do promotor interno da RNA polimerase III.

Retrotransposons PLE, por sua vez, apresentam domínio de endonuclease GIY-YIG, habilidade de reter introns e uma transcriptase reversa pertencente a um clado irmão das telomerases. Prevalente em animais, sua presença no genoma de fungos foi identificada apenas em *Coprinus cinereus* e *P. chrysosporium* (ARKHIPOVA, 2006). Por fim, retrotransposons YR, possuem três genes: *gag*, *pol* (composto por uma transcriptase reversa e uma RNAse H) e tirosina recombinase,

flanqueados com repetições terminais (MUSZEWSKA; STECZKIEWICZ; GINALSKI, 2013). São encontrados em algumas poucas espécies.

Retrotransposons LTR possuem, de forma geral, dois genes: *gag* (*Groupspecific AntiGen*) e *pol*, normalmente localizados em duas diferentes ORFs; no entanto, em alguns retrotransposons LTR, uma única ORF com *gag* e *pol* é separada por uma mudança da fase de leitura ou um códon de parada (HAVECKER; GAO; VOYTAS, 2004). Genes *gag* codificam as proteínas que formam as VLPs (*Virus-Like Particles*). Genes *pol*, por sua vez codificam o polipeptídeo composto por (i) integrase, (ii) ribonuclease H, (iii) transcriptase reversa, e (iv) aspartil protease (ALZOHAIRY et al., 2013). Em cada extremidade dos retrotransposons LTR, existem longas repetições (geralmente entre 100 e 500 nt) necessárias ao ciclo de replicação, as quais contém o promotor e regiões regulatórias da transcrição (DONNART et al., 2017).

Conforme a similaridade das sequências da transcriptase reversa e organização das subunidades do gene *pol*, os retrotransposons LTR podem ser agrupados em quatro grandes famílias: Pseudoviridae, do qual fazem parte os retrotransposons tipo Copia (Ty1, Ty2, Ty4 e Ty5); Metaviridae, composta pelos retrotransposons tipo Gypsy (Ty3); Belpaoviridae, composta pelos retrotransposons tipo Bel/Pao; e Retroviridae, do qual fazem parte os retrovírus endógenos (DODONOVA et al., 2019). Dentre estas famílias de retrotransposons LTR, Gypsy /Ty3 e Copia /Ty1 são as mais prevalentes em genomas fúngicos, podendo apresentar inserção sítio-alvo específica, de forma que retrotransposons Ty1 (Copia) são inseridos preferencialmente em regiões à montante de genes transcritos pela RNA polimerase III; Ty3 (Gypsy) estão próximos a sítios de início de transcrição de tRNAs;

e Ty5 (Copia) tendem a regiões de heterocromatina, como telômeros e loci de *mating* (MUSZEWSKA; HOFFMAN-SOMMER; GRYNBERG, 2011; NEUVÉGLISE et al., 2002). Retrotransposons LTR Tf1 (Ty3) de *S. pombe*, entretanto, inserem-se preferencialmente em regiões próximas a promotores de genes associados ao estresse, aumentando sua expressão (LEEM et al., 2008); enquanto elementos móveis de *Coccidioides immitis* e *Coccidioides posadasii* (majoritariamente Gypsy/Ty3) estão próximos a genes de fosforilação (KIRKLAND; MUSZEWSKA; STAJICH, 2018).

Dentre todas as classes de elementos transponíveis, os retrotransposons (Classe I) são encontrados com maior abundância em espécies fúngicas (Figura 2), particularmente LTR retrotransposons Ty3/Gypsy (DESJARDINS et al., 2011; KIRKLAND; MUSZEWSKA; STAJICH, 2018; MUSZEWSKA et al., 2017; MUSZEWSKA; HOFFMAN-SOMMER; GRYNBERG, 2011), dos quais, grande parte é classificada no subclado *Chromovirus* devido a presença de um cromodomínio de remodelamento da cromatina e regulação da expressão gênica, próximo a região C-terminal da integrase (MUSZEWSKA; HOFFMAN-SOMMER; GRYNBERG, 2011). Transposons da superfamília Tc1/Mariner, por sua vez, são os elementos de Classe II mais comumente encontrados no genoma de fungos, sendo correlacionados com a expansão do tamanho genômico. Sob condições normais, elementos transponíveis permanecem silenciados. Entretanto, condições estressantes podem desencadear a ativação destes elementos, criando, muitas vezes, uma correlação entre a abundância de elementos transponíveis no genoma e o modo de vida do organismo (MUSZEWSKA et al., 2017). Alguns estudos têm elucidado os efeitos significativos de elementos transponíveis como uma força evolutiva, levando à perda de sintenia,

inativação ou remoção de genes avirulentos e expansão genômica (CHÉNAIS et al., 2012; NUNES et al., 2011).

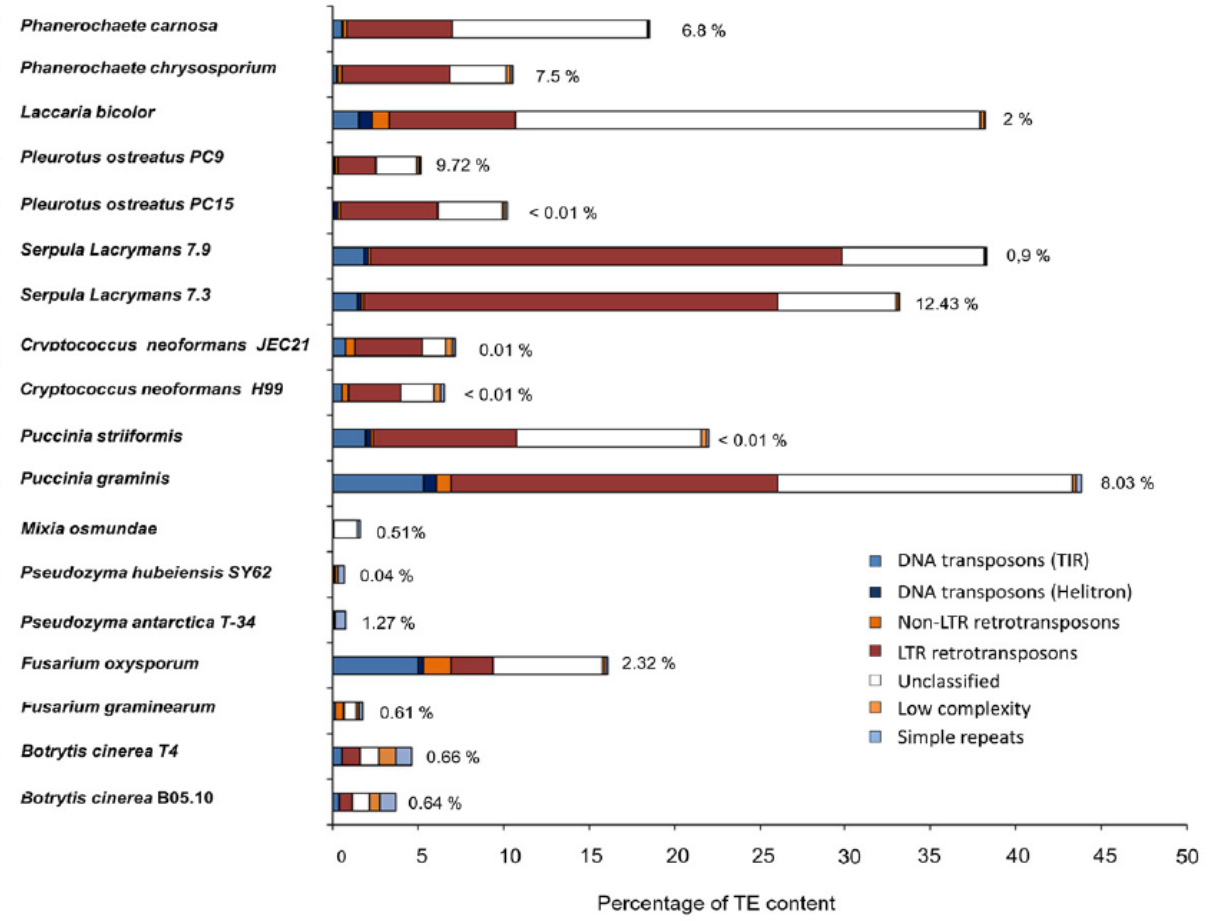

Figura 2. Filogenia e quantificação de elementos transponíveis de dezoito espécies fúngicas. Árvore construída com base em 551 genes; e inferida pelo programa RAxML com método de máxima verossimilhança e 100 *bootstraps*. Percentagens dos gaps de montagem são apresentados ao lado das barras. Linhas tracejadas são usadas para alinhar cada ramificação à ponta. Adaptado de (CASTANERA et al., 2016).

A inserção de elementos transponíveis próximo a genes pode levar a mudanças nos padrões de expressão gênica, enquanto inserções dentro dos genes, modificam a

estrutura gênica (KIRKLAND; MUSZEWSKA; STAJICH, 2018). Em algumas espécies, a metilação de sequência de elementos transponíveis leva à repressão de genes adjacentes, flanqueados por essas sequências em uma janela de 1 kb (CASTANERA et al., 2016; KIRKLAND; MUSZEWSKA; STAJICH, 2018). Além disso, elementos transponíveis também podem atuar na regulação (ou serem regulados) de clusters de genes próximos, como observado para os genes do cluster da penicilina em *Aspergillus nidulans* (SHAABAN et al., 2010); e de metabólitos secundários do fungo mutualista *Epichloë festucae* (CHUJO; SCOTT, 2014), e do fitopatógeno *Colletotrichum truncatum* (RAO et al., 2018).

1.3 RNA de interferência

A abundância de elementos transponíveis em um genoma depende tanto da capacidade de propagação destes elementos, quanto da capacidade do organismo de inativar este processo. Devido aos efeitos potencialmente deletérios para o *fitness*, organismos hospedeiros frequentemente desenvolvem mecanismos de defesa que limitam a atividade de tais elementos, como a metilação do DNA para formação de heterocromatina, silenciamento pela via do RNA de interferência, repressão por pequenos RNAs derivados de tRNA; e mutações pontuais induzidas por repetição (GOODWIN; POULTER, 2001; MUSZEWSKA et al., 2019).

A função de pequenos RNAs (sRNAs) como um mecanismo de defesa contra elementos móveis é amplamente conservada na via mediada pelo RNA de interferência. Em organismos eucarióticos, sRNAs com função regulatória podem ser

genericamente classificados em três classes: (i) siRNAs; (ii) miRNAS; e (iii) piRNAS. Diferentemente das duas primeiras classes, piRNAs constituem uma classe de sRNAs de silenciamento específica de animais, a qual se liga a proteínas Argonauta do clado PIWI, e não AGO (OZATA et al., 2019).

Pequenos RNAs de interferência (siRNAs) podem ser produzidos a partir de transcritos exógenos (dsRNA viral); transcritos endógenos de sequências repetitivas (como elementos transponíveis); ou transcritos que formam longos grampos; podendo ser dependentes da RNA polimerase dependente de RNA (RdRP) para a produção dos transcritos dupla fita. Além disso, apresentam total complementariedade de bases com os mRNAs alvo, podendo levar a clivagem destes, ou o silenciamento transcricional (DANG et al., 2011).

A análise de sRNAs em fungos tem possibilitado a identificação de sRNAs contendo o 5'-CAP e poliadenilados, bem como sRNAs associados a genes codificantes, tRNAs, rRNAs, snRNAs, elementos repetitivos e regiões intergênicas (Tabela 1); de forma que sRNAs podem estar associados ao silenciamento de retrotransposons, bem como a regulação de diversos processos biológicos (CARRERAS-VILLASEÑOR et al., 2013; NUNES et al., 2011). Curiosamente, espécies de leveduras que não apresentam conservados todos os componentes canônicos da via de RNAi podem apresentar sRNAs, sendo, grande parte, sRNAs associados a retrotransposons (DRINNENBERG et al., 2009).

Pequenos RNAs de interferência independentes de Dicer (disiRNAs) se originam de regiões genômicas não-repetivas (regiões gênicas ou intergênicas), pareiam com as fitas senso e antisenso do DNA em *N. crassa* e tem sua biogênese

também independente de RdRP e Argonauta. Pequenos RNAs exônicos (ex-siRNAs) de *Mucor circinelloides* e *Rhizophagus irregularis* (SILVESTRI et al., 2019) apresentam complementariedade de sequência com éxons e são dependentes de RdRP e Dicer (DANG et al., 2011).

Tabela 1 Classes de sRNAs dependentes de RNAi em fungos. Adaptado de (NICOLÁS; RUIZ-VÁZQUEZ, 2013)

Função	Nome	Acrônimo	Indutor	Primeiramente descrito
Defesa	Pequenos RNAs interferentes	SiRNAs	Transgenes integrativos	*Neurospora crassa*
			Transgenes não integrativos	*Mucor circinelloides*
			Transposons	*Neurospora crassa*
			Vírus	*Cryphonectria parasítica*
	Pequenos RNAs Interferentes associados ao MSUD	masiRNAs	DNA não pareado	*Neurospora crassa*
	Pequenos RNAs Interferentes do SIS	SIS siRNAs	Transgenes repetitivos	*Cryptococcus neoformans*
Regulação gênica endógena	Pequenos RNAs Interferentes Exônicos	ex-siRNAs	Transcrição regular	*Mucor circinelloides*
	Micro RNAs *like*	milRNAs	Transcrição regular	*Neurospora crassa*
	Pequenos RNAs Interferentes	siRNAs	Transcrição em heterocromatina	*Schizosaccharomyces pombe*
	Pequenos RNAs interatuantes a QDE-2	qiRNAs	Dano no DNA	*Neurospora crassa*

Os miRNAs, por sua vez, são gerados a partir de genes de miRNAs, os quais geram transcritos de fita simples que formam estruturas em grampo; podem não apresentar total complementariedade de bases com o alvo e levam à degradação do

mRNA ou repressão da tradução (DANG et al., 2011). Embora não presentes em fungos (miRNAs canônicos), a identificação de miRNAs-like (milRNAs) em *N. crassa*, permitiu o reconhecimento de moléculas produzidas por um mecanismo similar. Dentre as classes de milRNAs, quatro são descritas de acordo com a importância das proteínas Dicer e Argonauta para a produção de pré-milRNAs e milRNAs maduros. Deste modo, milR-2 são independentes de Dicer e dependentes de Argonauta; milR-3 e milR-4 são completa ou parcialmente dependentes de Dicer e independentes de Argonauta para a produção do milRNA maduro; e milR-1, é completamente dependente de Dicer e Argonauta (LEE et al., 2010). Similarmente, milRNAs tem sido descritos em outras espécies como *Penicillium marneffei* (LAU et al., 2013), *Fusarium oxysporum* (CHEN et al., 2014a), *Fusarium graminearum* (CHEN et al., 2015), *Aspergillus fumigatus* (ÖZKAN et al., 2017) e *Trichophyton rubrum* (WANG et al., 2018).

A expressão diferencial de milRNAs e siRNAs em diferentes fases de desenvolvimento ou condições ambientais, indicam que estas moléculas podem estar envolvidas na regulação de diversos processos celulares, como a manutenção da integridade genômica contra elementos móveis (DRINNENBERG et al., 2009; NGUYEN et al., 2018; NICOLAS et al., 2010; NOLAN et al., 2005; NUNES et al., 2011), resposta ao estresse (MENG et al., 2017; TRIEU et al., 2015; YANG, 2015) e regulação do processo morfológico (LAU et al., 2013; NICOLÁS et al., 2015; RAMAN et al., 2017; WANG et al., 2018). Em epimutantes de *M. circinelloides,* sRNAs antisenso específicos ao gene *fkbA* conduzem à degradação do mRNA alvo e conferem resistência aos antifúngicos FK506 e rapamicina (CHANG et al., 2019). *Sclerotinia sclerotiorum*

(DERBYSHIRE et al., 2018) *e B. cinerea* (WEIBERG et al., 2013) silenciam componentes do sistema imune em diversos hospedeiros através da produção de sRNAs associados a retrotransposons.

RNAs de interferência desempenham papel crítico na regulação gênica, estrutura cromossômica e estabilidade genômica, de forma que a função conservada da via de RNAi é proteger o genoma da invasão de elementos móveis (WANG et al., 2010). A via do RNAi medeia a degradação de mRNA dependente de homologia com pequenas moléculas de RNA e é um mecanismo regulatório chave que controla a transcrição e tradução em organismos eucarióticos. As enzimas necessárias ao silenciamento por RNA são numerosas e podem variar de acordo com a espécie, mas, universalmente, incluem as proteínas Argonauta, com domínios PAZ e PIWI, de ligação a pequenos RNAs e clivagem de mRNAs; a Ribonuclease Dicer, a qual produz os pequenos RNAs de interferência a partir de precursores dupla fita (dsRNAs); e a RdRP, a qual produz dsRNAs (MOAZED, 2009). Neste processo, um RNA fita dupla, produzido por uma RdRP é inicialmente processado em siRNAs de 21-28 nucleotídeos pela enzima Dicer e entra na via canônica. O siRNA é então incorporado ao Complexo da Silenciamento Induzido por RNA (RISC), o qual contém uma proteína Argonauta com atividade de degradação de RNA. Após entrar no complexo RISC, os siRNAs promovem a clivagem de mRNAs homólogos (ARMAS-TIZAPANTZI; MONTIEL-GONZÁLEZ, 2016).

Componentes canônicos da via do RNAi estão ausentes nas leveduras *S. cerevisiae, Candida glabrata, Kluyveromyces lactis, Ashbya gossypii* e no fungo

filamentoso *U. maydis*, dentre outros (Figura 3). Espécies, como *Candida albicans, Candida tropicalis*, *Saccharomyces castellii e Kluyveromyces polysporus*, apresentam os componentes de forma incompleta (DRINNENBERG et al., 2009). Enquanto leveduras de brotamento parecem ter perdido, universalmente, a RdRP, e mantém uma via do RNA de interferência não canônica (ou então uma via não funcional); fungos filamentosos apresentam grande conservação e diversidade de funções (DANG et al., 2011).

Genomas fúngicos, ascomicetos e basidiomicetos, possuem, tipicamente, três genes de proteínas Argonauta e três RdRPs. *S. pombe* possui um único gene para AGO, Dicer e RdRP, as quais estão envolvidas no silenciamento pós-transcricional e heterocromatinização (HALIC; MOAZED, 2010). *A. nidulans* realiza o silenciamento sem a presença da RdRP (HAMMOND; KELLER, 2005). A presença de uma via de RNAi funcional desempenha um papel chave na defesa de genomas eucarióticos. Em *Cryptococcus neoformans*, mutações nulas em genes da via de RNAi resultam no aumento da expressão e mobilização de retrotransposons (JANBON et al., 2010). A deleção dos genes MoDCL-2 (Dicer), MoRdRP (RdRP) e MoAgo3 (Argonauta) em *Magnaporthe oryzae* leva à diminuição de sRNAs associados a repetições, especialmente para retrotransposons LTR e Gypsy (RAMAN et al., 2017).

Entretanto, uma correlação positiva entre a presença de mecanismos de defesa e a abundância de elementos transponíveis pode ser encontrada, de forma que fungos nos quais RNAi e RIP estão ausentes, um número menor de elementos transponíveis é observado; bem como fungos que possuem tais mecanismos de defesa apresentam

menos elementos de classe II com sequências conservadas de transposases (MUSZEWSKA et al., 2017).

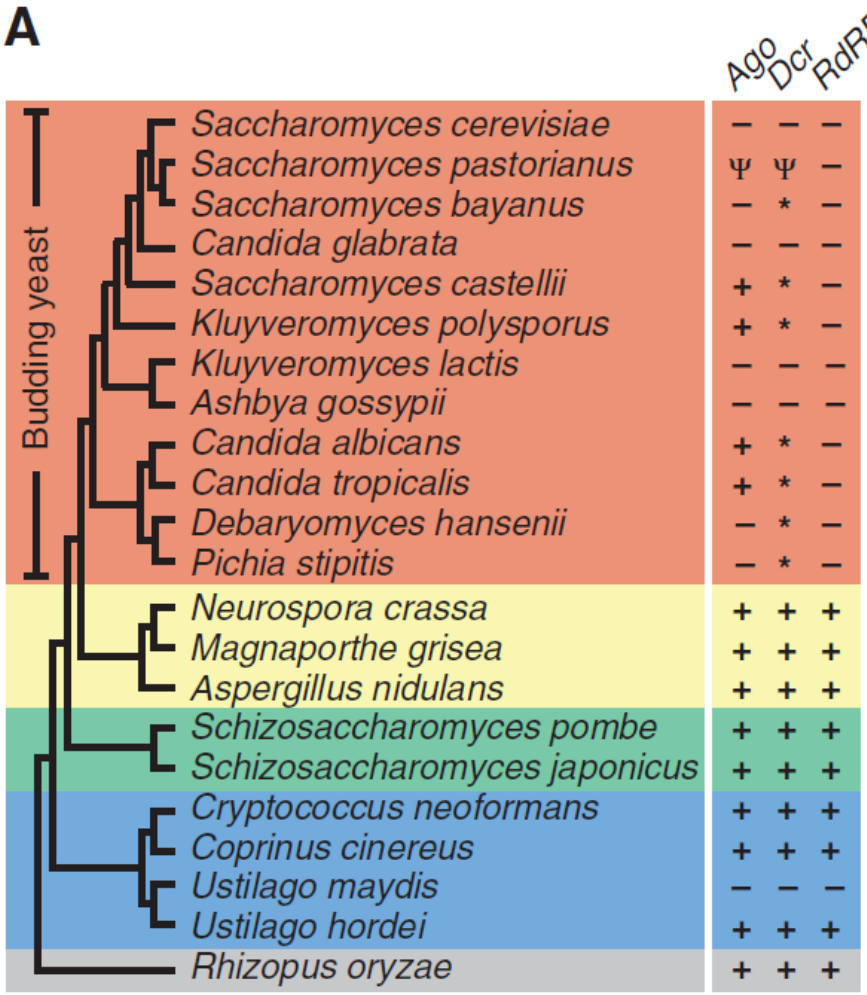

Figura 3. Presença dos componentes da via de RNAi em diferentes espécies fúngicas. Cladograma mostrando Basidiomycota (azul), Zygomycota (cinza) e Ascomycota. Subdividida em Saccharomycotina (leveduras de brotamento, laranja), Pezizomycotina (amarelo) e Taphrinomycotina (verde). A presença de genes canônicos do RNAi é indicado (+). Todos os genomas têm um ortólogos RNT1, e muitos outros tem um segundo gene contendo domínio RNaseIII (*), o qual tem atividade Dicer em *S. castellii*. Pseudogenes estão indicados (Ψ). Proteínas universais da via canônica de RNAi: Argonauta (Ago), Dicer (Dcr) e RNA polimerase RNA-dependente (RdRP). Adaptado de (DRINNENBERG et al., 2009).

1.4 Espécies patogênicas do gênero *Cryptococcus*: aspectos funcionais, estudos genômicos e transcritômicos

Cryptococcus neoformans e *Cryptococcus gattii* constituem os dois agentes etiológicos da criptococose, doença transmitida pela inalação de esporos ou partículas infecciosas presentes em variados nichos ecológicos, como fezes de aves (especialmente pombos) e matéria em decomposição de diversas espécies de árvores. *Cryptococcus neoformans* é o mais importante causador de meningite fúngica no mundo, com aproximadamente cinco mil casos anuais só na América Latina, e uma taxa de mortalidade de aproximadamente 50%. No Brasil, a meningite criptocócica acomete entre 1000 e 2500 indivíduos anualmente, dos quais, 47% são levados a óbito (décima terceira causa de morte mais frequente). Mesmo com poucos dados regionais, a criptococose mantem-se como uma das patologias fúngicas mais incidentes. Em nível mundial, 1 milhão de novos casos emergem anualmente, com um número estimado de 625000 mortes; todavia, a indistinção clínica entre *C. neoformans* e *C. gattii* durante o diagnóstico da criptococcose pode subestimar o real número de casos associados a este último (BIELSKA; MAY, 2015; FIRACATIVE et al., 2018). Estima-se que 20% dos casos de criptococose sejam causados por linhagens de *C. gattii* (MEYER et al., 2011); contudo, estudos em modelos murinos de criptococose têm mostrado que não há diferença significativa entre as taxas de mortalidade de camundongo infectados com *C. neoformans* H99 e com a linhagem hipervirulenta R265 de *C. gattii* (OKUBO et al., 2013).

C. neoformans apresenta distribuição global e causa meningoencefalite em indivíduos imunocomprometidos; *C. gattii,* contudo, mostrava-se restrito a regiões tropicais e subtropicais, até o surto da Ilha de Vancouver (Columbia Britânica) e do Noroeste do Pacífico, em 1999, causando infecção pulmonar. Desde então, têm sido promovidos numerosos esforços para o entendimento dos determinantes genéticos que ocasionaram o surgimento e propagação das linhagens hipervirulentas que levaram ao evento (MEYER, 2015). Entre 1999 e 2011, 240 casos humanos de criptococose (principalmente em indivíduos imunocompetentes) e centenas de casos em animais foram reportados na região, criando uma média anual de incidência de 5.8 casos por milhão de pessoas (D'SOUZA et al., 2011).

Enquanto linhagens de *C. neoformans* podem ser classificados em duas variedades, *C. neoformans* variedade *grubii* (sorotipo A ou genótipos VNI e VNII), representado pela linhagem H99; e *C. neoformans* variedade *neoformans* (sorotipo D ou genótipo VNIV), representado pela linhagem de referência JEC21; análises genéticas de RFLP, PCR *fingerprinting*, AFLP, MLMT, MLST e WGST reconheceram 4 tipos moleculares de *C. gattii* ("variedade *gattii*" VGI, VGII, VGIII e VGIV); os quais diferem genética, epidemiológica e geograficamente (MEYER et al., 2011). Linhagens associadas aos genótipos VGI e VGII são as maiores causadoras de criptococose em indivíduos imunocompetentes, enquanto àquelas pertencentes aos genótipos VGIII e VGIV raramente causam infecções e afetam majoritariamente indivíduos imunodeprimidos (FARRER et al., 2015). Neste contexto, 97% dos casos documentados de criptococose do surto do Noroeste do Pacífico foi causada por três subgrupos quase clonais do tipo molecular VGII: VGIIa (oriundo das Américas do Sul

e Norte e responsável por mais de 80% dos casos), VGIIb (oriundo das Américas do Sul e Norte, Austália, Ásia e Europa) e VGIIc (oriundo do Noroeste do Pacífico) (BIELSKA; MAY, 2015; MEYER, 2015).

Cryptococcus neoformans e *C. gattii* compartilham todos os maiores fatores de virulência, como a produção da cápsula polissacarídica, presença de melanina na parede celular, urease, fosfolipase B e crescimento à 37 ºC. Além disso, clinicamente, *C. gattii* leva a maior incidência de granulomas no cérebro e pulmão; os quais demandam um tratamento antifúngico mais prolongado (D'SOUZA et al., 2011).

O desenvolvimento do processo infeccioso é similar entre espécies, sendo o mesmo iniciado pela inalação de propágulos infecciosos, os quais alcançam o parênquima alveolar e são tipicamente fagocitados por macrófagos (e outras células de defesa presentes nos estágios iniciais da infecção pulmonar, como neutrófilos e células dendríticas), dentro dos quais as células criptococócicas conseguem sobreviver e se replicar (BYRNES et al., 2010; LEOPOLD WAGER et al., 2016). Estas células do sistema imune inato possuem atividade apresentadora de antígenos e induzem uma resposta adaptativa no hospedeiro (LEOPOLD WAGER et al., 2016; ZHANG; SUN; SHI, 2015). A capacidade de *Cryptococcus* de evitar a fagocitose (e escapar da mesma por um mecanismo não lítico), bem como sobreviver e se propagar dentro dos fagócitos contribui para a disseminação das células fúngicas do pulmão pela corrente sanguínea, podendo atravessar a barreira hematoencefálica e alcançar o sistema nervoso central (DELEON-RODRIGUEZ; CASADEVALL, 2016; JOHNSTON; MAY, 2013). Interessantemente, linhagens de VGII diferem de VGI em

sua forma clínica, apresentando maior taxa de sintomas associados a quadros de infecção pulmonar que sintomas do sistema nervoso central (FARRER et al., 2015).

Recentemente, a variação genética entre os genótipos de *C. gattii* (bem como de *C. neoformans*) levou à proposição de reclassificação dos tipos moleculares VGI, VGII, VGIII e VGIV como espécies, e a consequente nomenclatura dos mesmos como *C. gattii, C. deuterogattii, C. bacillisporus* e *C. tetragattii*, respectivamente (Figura 4).

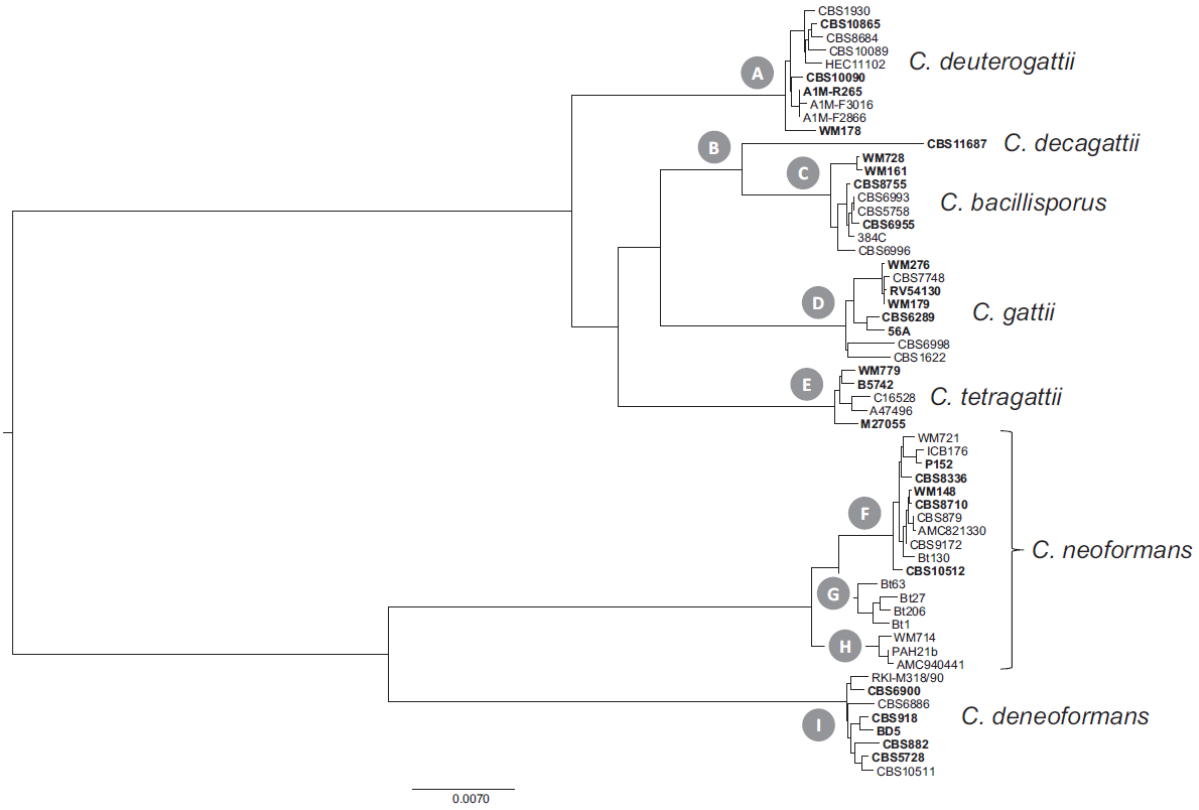

Figura 4. Diversidade no complexo de espécies *C. gattii/C. neoformans* inferido de um conjunto concatenado de 11 loci. *C. gattii* (D), *C. deuterogatii* (A), *C. bacillisporus* (C), *C. tetragattii* (E) e *C. decagattii* (B) correspondem aos genótipos VGI, VGII, VGIII, VGIV e VGIV/VGIIIc, respectivamente. *C neoformans* F, G e H correspondem aos genótipos VNI, VNB/VNII e VNII (sorotipo A). *C. deneoformans* (I) corresponde ao genótipo VNIV (sorotipo D). Obtido de (HAGEN et al., 2015).

Aliado a isso, estão os dados que demonstram a pouca troca genética entre as linhagens dos quatro genótipos, sugerindo que as mesmas estão relativamente isoladas desde a divergência, apesar da sobreposição de nichos e distribuição geográfica (D'SOUZA et al., 2011; FARRER et al., 2015; HAGEN et al., 2015). Assim, uma considerável variação genética é encontrada entre genótipos de *C. gattii* (93% de identidade e 52 SNPs por kb), ao passo que a comparação intralinhagem apresenta 97% de identidade em média e menos de 6 SNPs por kb. Embora a estrutura cromossomal das linhagens seja altamente conservada, 4 grandes rearranjos cromossomais foram suportados por múltiplas linhagens, incluindo uma fusão cromossomal única ao VGII, duas translocações de 140 e 700 kb únicas ao VGIII, e uma translocação de 450 kb única para o VGIV; as quais podem impactar a capacidade de troca genética entre linhagens (FARRER et al., 2015).

Apesar de todos os tipos moleculares causem a criptococose, diferenças entre grupos e subgrupos em relação às capacidades de virulência tem sido identificadas, como a aumentada taxa de proliferação intracelular de linhagens VGIIc e a maior capacidade de linhagens VGIIa para sobreviver e se proliferar dentro de células fagocíticas do hospedeiro (FARRER et al., 2016). Ademais, as linhagens de VGII possuem um maior repertório de proteínas de membrana secretoras (*SCAMPs*), de domínios Fop (*Friend of Prmt1*) de associação à cromatina e de domínio DUF1932 de função desconhecida; alcançando um número de 119 genes linhagem-específicos com enriquecimento de domínios para as mais diversas funções. A perda de famílias de domínios de transportadores de cobre (Ctr) e enolases nos genótipos VGIII e VGIV (os quais apresentam 79 e 170 genes linhagem-específicos, respectivamente),

contudo, podem estar associados às baixas taxas de infecção em indivíduos imunocompetentes; ao passo que o baixo número de genes perdidos em linhagens VGI pode explicar a sobrevivência desta linhagem em diversos nichos (FARRER et al., 2015).

A análise dos genes perdidos de linhagens VGII apresenta, por sua vez, enriquecimento de domínios associados aos componentes da via do RNA de interferência (PAZ, Piwi e DUF1785), devido à ausência das proteínas Argonauta (AGO1 e AGO2), Dicer1 e RPdR. Além disso, outras proteínas envolvidas na resposta de defesa de *Cryptococcus* ao estresse antioxidante, como peroxidases e a subunidade VIb da citocromo oxidase c (COX6B) também estão ausentes (FARRER et al., 2015).

Componentes da via do RNAi em *Cryptococcus*, curiosamente, parecem ter sido perdidos diversas vezes durante a evolução. Enquanto o sorotipo D de *C. neoformans* contém dois genes parálogos para Argonauta (*AGO1* e *AGO2*), Dicer (*DCR1* e *DCR2*) e um gene para a *RdRP* (RNA Polimerase RNA-dependente); o sorotipo A representado por H99 apresenta um único gene Argonauta (*AGO1*), uma *RdRP* e dois genes parálogos para a Dicer. *C. gattii*, no entanto, possui a via funcional para os genótipos VGI, VGIII e VGIV; e não funcional para VGII, devido a ausência da Argonauta e presença de DCR1 e RdRP como pseudogenes, restando apenas uma DCR2 canônica (FERETZAKI et al., 2016).

O novo sequenciamento do genoma de *C. gattii* linhagem R265, representativa do genótipo VGII, por Yadav e colaboradores (2018) com o uso das tecnologias PacBio e Nanopore, gerou uma nova montagem genômica completa de ≈17 Mb, com 14

cromossomos, em substituição à montagem prévia de 27 *scaffolds*. De forma similar, a comparação do novo genoma de *C. gattii* com *C. neoformans* revelou rearranjos genômicos entre as espécies. Adicionalmente, a avaliação dos centrômeros permitiu a identificação de regiões de baixa transcrição compostas por retrotransposons Tcn1 (Ty3/gypsy) –Tcn6 (Ty1-copia) altamente conservados (70% de identidade) nas espécies patogênicas de *Cryptococcus* (*C. neoformans* H99, JEC21 e *C. gattii* R265) (YADAV et al., 2018).

Dados semelhantes foram encontrados com a identificação de 13 eventos de rearranjo cromossomal desencadeados pela atividade de retrotransposons Tcn nos genomas de linhagens de diferentes genótipos de *C. gattii* (FARRER et al., 2015); e entre *C. gattii* e *C. neoformans*, na associação dos retrotransposons Tcn1 e Tcn6 a rearranjos nas junções (região centromérica) dos cromossomos 4 e 10 (D'SOUZA et al., 2011).

Como observado em *C. neoformans*, embora a perda de funcionalidade dos componentes canônicos da via de RNAi não altere a viabilidade celular em diferentes temperaturas, condições nutricionais ou estresse, a mutação do gene *rdrp* leva, por sua vez, ao aumento de abundância de vários retrotransposons endógenos, resultando em um aumento significativo dos eventos de transposição (JANBON et al., 2010). Linhagens mutantes da via de RNAi também apresentaram elevada indução da expressão de retrotransposons LTR de regiões centroméricas quando em condições de *mating* ou de limitação de nutrientes. Semelhante ao *quelling* de *N. crassa*, *C. neoformans* possui o SIS (*Sex-Induced Silencing*), um mecanismo de silenciamento

mediado por RNAi que ocorre durante o ciclo sexual e protege o genoma da propagação de elementos transponíveis durante o *mating* (WANG et al., 2010).

A identificação de sRNAs homólogos a transgenes repetitivos e elementos transponíveis, tanto durante o ciclo sexual quanto na fase vegetativa, evidencia a regulação destas sequências pela via do RNAi (WANG et al., 2012). Todavia, o silenciamento gênico dependente de RNAi ocorre em uma frequência cerca de 250 vezes maior durante o *mating* (reprodução unissexual α-α ou de *mating* oposto a-α), elucidando a importância do SIS na defesa da integridade genômica (WANG et al., 2010, 2012). Curiosamente, o *mating type* α é relacionado à virulência de *C. neoformans*, e as taxas de transposição e de expressão de retrotransposons não só são maiores em mutantes rdp1 como também, são ainda maiores durante a reprodução unissexual α-α, de forma que a potencial carga mutacional derivada da hiperatividade de retrotransposons necessita de um mecanismo robusto de silenciamento. Neste contexto, a perda de uma via de RNAi funcional em *C. gattii* pode ter sido benéfica para a evolução da espécie, de modo que as maiores taxas de recombinação de linhagens do genótipo VGII podem estar associadas com a geração de um genótipo hipervirulento (WANG; DARWICHE; HEITMAN, 2013).

Entretanto, qualquer abordagem mais detalhada do impacto da perda da maquinaria de RNAi para a evolução do genoma de *C. gattii* R265, bem como a análise de mecanismos alternativos de silenciamento transcricional para a regulação de elementos transponíveis e vias metabólicas só pode ser inteiramente avaliado com o auxílio de uma anotação gênica acurada. Publicado em 2011 (D'SOUZA et al., 2011), o primeiro genoma sequenciado de *C. gattii* R265 não apresentava montagem

completa, e possuía anotação gênica baseada primariamente na homologia com a anotação de *C. gattii* WM276 (D'SOUZA et al., 2011). *C. gattii* WM276, por sua vez, foi anotado segundo homologia com modelos gênicos e ESTs de *C. neoformans* JEC21 (LOFTUS et al., 2005), seguido de análise manual.

Ainda, é importante notar que apenas em 2014 (JANBON et al., 2014) foi publicada a versão corrente da anotação de *C. neoformans* H99 (a qual é mantida como anotação de referência para a espécie), com base em dados de sequenciamento genômico e RNA-Seq de diferentes condições; ao passo que a atualização da anotação de JEC21 ainda não está disponível.

2 OBJETIVOS

O presente trabalho teve por objetivo realizar estudos transcritômicos com *Cryptococcus gattii* R265, a fim de avaliar processos fisiológicos e de manutenção da estabilidade genômica, tendo como base o desenvolvimento de uma *pipeline* de predição gênica automática e otimizada para *Cryptococcus*.

2.1 Objetivos Específicos

 Identificar sequências de retrotransposons em genomas pertencentes aos quatro genótipos de *C. gattii*;

 Avaliar a expressão de retrotransposons em *C. gattii* R265 e correlacionar com a presença de sRNAs associados;

 Predizer o perfil metabólico de *C. gattii* R265 durante infecção em modelo murino utilizando dados de transcritômica desta levedura coletada do sítio de infecção;

 Desenvolver uma *pipeline* automática de predição gênica automática, otimizada e validada com *C. neoformans*;

 Aplicar a *pipeline* de predição para a anotação de sequências codificantes do novo genoma sequenciado de *C. gattii* R265

3 RESULTADOS

Os resultados da presente tese estão distribuídos em três capítulos.

O primeiro capítulo (item 3.1) trata da identificação de retrotransposons em 17 linhagens de *C. gattii*; bem como a avaliação da expressão de retrotransposons e sRNAs associados em *C. gattii* R265. Os resultados apresentados neste capítulo foram publicados na forma de um artigo científico no periódico *BMC Genomics* (Ferrareze et al., 2017).

O segundo capítulo (item 3.2) trata da publicação da anotação manual de *C. gattii* R265 e análise do perfil metabólico de *C. gattii* R265 durante infecção em modelo murinho. Os resultados apresentados neste segundo capítulo foram publicados na forma de um artigo científico no periódico *Microorganisms* (Ferrareze et al., 2017).

O terceiro capítulo (item 3.3) refere-se ao desenvolvimento de uma pipeline automática de anotação, otimizada e validada com *C. neoformans* para aplicação em *C. gattii* R265. Os resultados apresentados neste terceiro capítulo estão sob forma de um manuscrito a ser submetido à publicação em breve.

3.1 Capítulo I: sRNAs as possible regulators of retrotransposon activity in *Cryptococcus gattii* VGII.

Este capítulo apresenta o artigo publicado na revista *BMC Genomics* em 12 de abril de 2017 (Ferrareze et al., 2017).

Ferrareze et al. BMC Genomics (2017) 18:294
DOI 10.1186/s12864-017-3688-4

BMC Genomics

RESEARCH ARTICLE Open Access

sRNAs as possible regulators of retrotransposon activity in *Cryptococcus gattii* VGII

Patrícia Aline Gröhs Ferrareze[1], Rodrigo Silva Araujo Streit[2], Francine Melise dos Santos[1], Augusto Schrank[1,2], Livia Kmetzsch[1,2], Marilene Henning Vainstein[1,2] and Charley Christian Staats[1,2*]

Abstract

Background: The absence of Argonaute genes in the fungal pathogen *Cryptococcus gattii* R265 and other VGII strains indicates that yeasts of this genotype cannot have a functional RNAi pathway, an evolutionarily conserved gene silencing mechanism performed by small RNAs. The success of the R265 strain as a pathogen that caused the Pacific Northwest and Vancouver Island outbreaks may imply that RNAi machinery loss could be beneficial under certain circumstances during evolution. As a result, a hypermutant phenotype would be created with high rates of genome retrotransposition, for instance. This study therefore aimed to evaluate *in silicio* the effect of retrotransposons and their control mechanisms by small RNAs on genomic stability and synteny loss of *C. gattii* R265 through retrotransposons sequence comparison and orthology analysis with other 16 *C. gattii* genomic sequences available.

Results: Retrotransposon mining identified a higher sequence count to VGI genotype compared to VGII, VGIII, and VGIV. However, despite the lower retrotransposon number, VGII exhibited increased synteny loss and genome rearrangement events. RNA-Seq analysis indicated highly expressed retrotransposons as well as sRNA production.

Conclusions: Genome rearrangement and synteny loss may suggest a greater retrotransposon mobilization caused by RNAi pathway absence, but the effective presence of sRNAs that matches retrotransposon sequences means that an alternative retrotransposon silencing mechanism could be active in genomic integrity maintenance of *C. gattii* VGII strains.

Keywords: *Cryptococcus gattii*, RNAi, Retrotransposons, Synteny

Background

Transposable elements (TE) are repeated DNA sequences that move and propagate in genome by "copy-and-paste" (TE class I) and "cut-and-paste" (TE class II) mechanisms [1]. TE class I, or retrotransposons, replicate through an RNA intermediate. Autonomous sequences encode their own reverse transcriptase and are typically 5–10 kilobases in length. There are two classes of retrotransposons, based on the presence of LTR (long terminal repeats) or its absence (non-LTR retrotransposons – LINE/SINE elements) [2]. At each end of LTR retrotransposons, there are repeats necessary for their replication cycle which contain sequences that regulate transcription. All LTR retrotransposons can have *gag* and *pol* genes; *gag* encodes proteins (GAGs; Group-specific AntiGen) that form VLPs (Virus-Like Particles), and *pol* encodes a polypeptide composed of (i) integrase, (ii) ribonuclease H, (iii) reverse transcriptase, and (iv) aspartyl protease enzymes [3]. *Gag* and *pol* genes are usually located in two open reading frames (ORFs), but in some LTR retrotransposons, a single *gag*/*pol* ORF is separated by a frameshift or a stop codon [4].

Transposable elements most commonly found in fungal species are retrotransposons. Some studies have elucidated significant effects of transposable elements as an evolutionary force, leading to synteny loss, avirulent gene deletion, and genome expansion [1, 5]. Insertion of transposable elements near genes can lead to

* Correspondence: staats@cbiot.ufrgs.br
[1]Programa de Pós-Graduação em Biologia Celular e Molecular, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul (UFRGS), 91501-970 Porto Alegre, RS, Brazil
[2]Departamento de Biologia Molecular e Biotecnologia, Instituto de Biociências, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil

BioMed Central

© The Author(s). 2017 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

changes in gene expression patterns, while insertions within them can modify gene structure. Recombination between LTR elements in different sites leads to chromosomal rearrangements on a large scale. Because of their potentially deleterious effects, host organisms frequently evolve mechanisms to limit the activity of such elements [2].

The function of small RNAs as a defense mechanism against mobile elements is widely conserved among double-stranded RNA-mediated interference (RNAi) routes [6]. In recent years, several studies have examined the role of miRNAs in various cellular processes such as genomic integrity maintenance, DNA damage responses to biotic stress, and morphological process regulation. Interfering RNAs play critical roles in gene regulation, chromosomal structure, and genomic stability, such that the RNAi conserved function is to protect the genome from mobile element invasion [6]. The RNAi pathway mediates homology-dependent degradation of an mRNA with small RNA molecules and is a key regulatory mechanism that controls transcription and translation in eukaryotic organisms. The enzymes required for RNA silencing are numerous and may vary among species, but universally include Argonaute proteins that bind small RNAs, Dicer ribonuclease that produces small interfering RNAs (siRNA) from double stranded RNA precursors (dsRNA), and RNA-dependent RNA polymerase that produces dsRNAs [7]. In this process, a double stranded RNA produced by an RNA-dependent RNA polymerase is first processed into small interfering RNAs of 21–28 nucleotides by Dicer RNase III and enters the canonical pathway. This siRNA is incorporated into a nuclease complex, the RNA-induced silencing complex (RISC), containing an Argonaut protein with RNA-degrading activity. After the RISC complex, siRNAs promote the cleavage of homologous mRNAs [8]. These factors play a key role in eukaryote genome defense: null mutations in their corresponding genes result in increased transposons expression and mobilization, decrease and loss of endogenous retrotransposon-derived siRNAs, and drug resistance mutations induced by retrotransposons [9–11].

Two species from the genus *Cryptococcus* are the etiologic agent of cryptococcosis. *Cryptococcus neoformans* has worldwide distribution and affects immunocompromised individuals, mainly causing cryptococcal meningoencephalitis. *Cryptococcus gattii*, prior to the Pacific Northwest outbreak, was known to be endemic in tropical and subtropical areas, typically infecting healthy hosts with pulmonary cryptococcosis [12, 13]. Estimated numbers calculate that *C. neoformans* cryptococcosis kills 650,000 immunocompromised HIV/AIDS patients every year worldwide [13]. However, the common clinical indistinction between *C. neoformans* and *C. gattii* during cryptococcosis diagnosis can underestimate the real number of cases associated with *C. gattii* cryptococcosis. Studies in a murine model of pulmonary cryptococcosis have demonstrated that there is no significant difference in mortality rates of mice infected with *C. neoformans* H99 and the hypervirulent strain *C. gattii* R265 [14].

The *C. neoformans* serotype D contains two paralogous genes for Argonaute (*AGO1* and *AGO2*) and Dicer (*DCR1* and *DCR2*), and one gene for RdRP (*RDP1*). In contrast, serotype A has a single Argonaut gene (*AGO1*) and RdRP (*RDP1*) and two paralogous genes for Dicer (*DCR1* and *DCR2*). *C. gattii* has four molecular types (VG "variety gattii"): VGI and VGII are the main pathogens of cryptococcosis in immunocompetent hosts, while VGIII and VGIV rarely cause infections and affect immunocompromised hosts [15]. VGI, VGIII, and VGIV genotypes have all the components of an RNAi system, while VGII has a canonical *DCR2* and *DCR1/RDP1* pseudogenes [16]. Bielska and May [13] have shown that 97% of documented cases of cryptococcosis in the Vancouver Island outbreak were caused by VGII, with VGIIa prevalent in more than 80% of the cases.

The absence of Argonaute genes in the primary pathogen *C. gattii* R265, and other VGII strains, indicates that these genotypes do not have a functional RNAi pathway [16] . The success of the R265 strain as a pathogen, in causing the Vancouver Island and Pacific Northwest outbreaks, may imply that RNAi machinery loss can be beneficial under certain circumstances during evolution [6]. This can be linked to the higher recombination rates of VGII strains, which are associated with the generation of novel and possibly more virulent genotypes [17]. The loss of RNAi could thus allow greater mobility of transposable elements, as has been reported in *C. neoformans* [18]. However, in the RNAi-lacking budding yeast *Saccharomyces cerevisiae*, a predominant proportion of sequenced small RNAs mapped to transposable elements [5]. Recent analysis showed that some yeasts, which have presumably lost the RNAi route, actually have siRNAs and use non-canonical Dicer proteins to control Argonaute-like proteins without RdRP [8]. Therefore, an alternative retrotransposon silencing mechanism by sRNAs may act through a non-canonical RNAi pathway or through independent control systems.

Thus, the study of retrotransposon effects in *C. gattii* genomes, as a hypervirulent strain, with potentially high activity of transposable elements unrelated to any silencing mechanism, may contribute to increased understanding of the role of retrotransposons and sRNAs in genome evolution.

Results

Retrotransposon mining in *C. gattii* genome sequences

Retrotransposon structure usually consists of two almost identical LTRs flanking two ORFs to the *gag* and *pol*

genes [19]. Retrotransposon mining programs rely on the search for several structural features of LTR retrotransposons, such as the size range of the LTR sequences, the distance between the two LTRs of an element, among others [20]. The number of LTR-retrotransposons identified in 17 *C. gattii* genomes ranged from 21 to 78 sequences (Fig. 1). Their structural composition was confirmed using NCBI Blast (Conserved Domains) for the presence of *gag* and *pol* domains (Additional file 1). In this search, non-autonomous LTR sequences (which do not encode necessary proteins for transposition) were excluded. Reciprocal BLASTn (data not shown) and domain arquitecture present in the predicted proteins (Additional file 1) revealed that such sequences groups are unique. As shown in Fig. 1, VGI strains (WM276, NT10, E566, EJB2, and RU294), which putatively contain a complete RNAi pathway, present a higher retrotransposon count (especially WM276 and NT10) when compared to VGII (R265, CA1014, CBS7750, CBS10090, RAM5, LA55, MMRL2647, 99473, and 2001935), VGIII (CA1280 and CA1873), and VGIV (IND107) strains.

Retrotransposon orthology and synteny evaluation

Retrotransposon integration is a non-random process in the genome, as many transposable elements integrate into specific sequences as intergenic regions or upstream of RNA polymerase III-transcribed genes [21]. *Copia* elements (Ty1 family: Pseudoviridae) present specificity for pre-existing LTR transposons (resulting in a TE cluster formation) and transcriptionally active regions, while *gypsy* elements (Ty3 family: Metaviridae) are reported in heterochromatin [22]. Preferential insertion of TE clusters in regions with low recombination rates can be explained by weak selection against these sites [21]. For these reasons, most filamentous fungi carry more Ty3/*gypsy* than Ty1/*copia* elements [22]. The presence of retrotransposon-enriched regions (TE clusters) can be observed by sequence distribution along the 14 chromosomes of WM276 (Additional file 1). However, the 73 retrotransposon elements detected in the NT10 genome are generally incomplete (absence of conserved domain) and dispersed along the genome. A slight prevalence of Ty3/*gypsy* was found in the element composition of *C. gattii* strains.

The WM276 strain has 78 retrotransposon sequences, of which 24 are unique and 29 shared only with VGI strains. The R265 strain has 34 sequences, five of which are unique with 14 exclusive to VGII strains. Unique sequences indicate recent transposition events, since they do not have identifiable orthologues. Figure 2 shows the relationship between sequence amount and orthologue number. VGII strains (mainly CA1014) share more sequences with R265 than other genotypes. Similarly, WM276 shares 49 sequences with VGI strains. These results indicate the existence of transposition events after genotype diversification. Data analysis has allowed the observation of the prevalence (almost exclusivity) of *gag* domains in Ty1 family sequences in the absence of aspartic protease. However, Ty3 has aspartic protease domains and is lacking *gag*. The presence of RVT_1 (Pfam: PF00078) and RVT_2 (Pfam: PF07727) reverse transcriptase families have been observed in *gypsy* and *copia*, respectively.

The chromodomains identified next to retrotransposons WM276.1, 53, 72, and 78 can be related to the

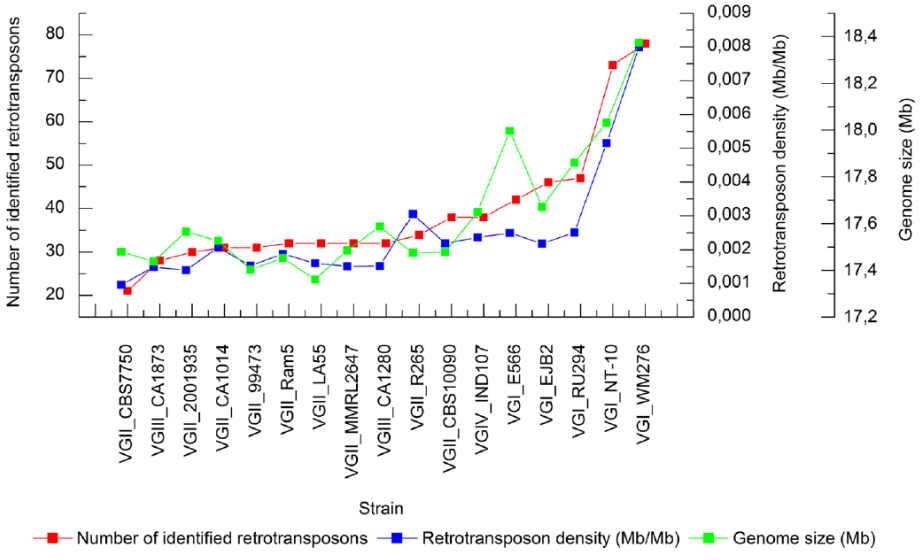

Fig. 1 Retrotransposon number, density and genome size comparison. The *red line* indicates the number of identified retrotransposons in each strain; *blue line* represents retrotransposon density in *C. gattii* genomes. The *green line* shows the size of *C. gattii* assembled genomes

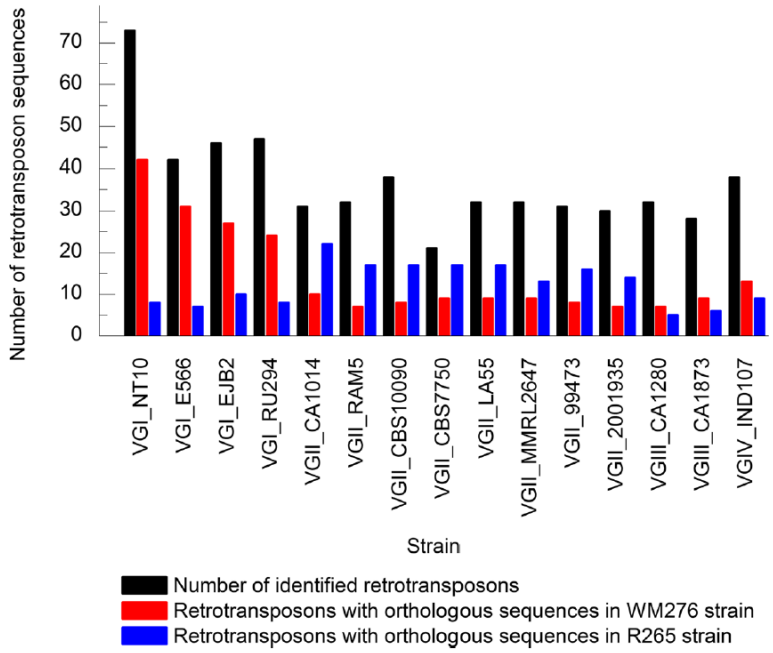

Fig. 2 Orthologous sequences count. Orthologous sequences from *C. gattii* WM276 and R265 shared with VGI, VGII, VGIII, and VGIV strains

Chromovirus genus of the Metaviridae (Ty3) family. These chromodomains, absent in R265 strain, are found in functional proteins as well as *gypsy* LTR retrotransposons. The canonical function includes chromatin remodeling and gene expression regulation. However, chromodomains in retrotransposons are poorly understood [23].

Genome rearrangements naturally accumulate in different strains. Successive rearrangements in *C. gattii* and *C. neoformans* genomes may have contributed to sexual isolation and speciation [24]. To determine the effect of retrotransposon mobilization in VGI and VGII strains, their synteny conservation was evaluated (Additional file 1). Orthologous elements shared by VGI and VGII genotypes were analyzed for synteny on a multiple genome alignment. Multiple alignments were performed using the Mauve program with WM276 as a reference for supercontig reordering. VGI strain alignment identified 29 synteny blocks in WM276 (Additional file 2) while VGII alignment found 54 synteny blocks in R265 (Additional file 3), indicating greater genome fragmentation or a better VGI genome sequence assembly. Both alignments had few rearrangements, with a predominant collinear genome. Coordinates of orthologous retrotransposons were used to classify syntenic and non-syntenic sequences.

Long terminal repeats from LTR retrotransposons are known to promote crossing over at non-homologous sites, leading to chromosomal rearrangement. Likewise, mutations arising from mobile element insertions in conserved blocks can cause synteny loss [21]. TE clusters from WM276 were shown to be responsible for synteny breaks at 5 genome points (Additional files 4, 5, 6, 7 and 8). The same number was found in R265 (Additional files 9, 10, 11, 12 and 13), suggesting that R265 active retrotransposons, even if in a small identifiable number, are more influential on genome integrity than WM276 sequences.

C. gattii R265 strain retrotransposons have 133 orthologous sequences in 8 VGII genomes. The majority of these (65.41%) preserve synteny with the R265 genome (Fig. 3a), particularly the CBS7750 strain, which has 17 orthologous sequences in syntenic blocks with R265. In comparison, VGI, despite its higher elements count, has almost 80% of orthologous sequences in synteny with the WM276 genome (Fig. 3b).

Among the widely distributed orthologous sequences of R265.3, R265.19, and R265.22 retrotransposons, the multiple genome alignment of 17 *C. gattii* strains shows synteny conservation for most of the *C. gattii* strains (Additional file 14). However, some variations suggest genome rearrangement events (Additional file 15). Orthologs of retrotransposon R265.3 identified in RU294 (VGI), IND107 (VGIV), and CA1280 (VGIII) are localized in the same block and generate their own syntenic group, while orthologs of retrotransposon R265.19 are kept syntenic in all genomes of these three genotypes. Among VGII, orthologs of retrotransposon R265.3 and orthologs of retrotransposon R265.19 sequences were located in opposite positions, such as an ectopic recombination result. This sequence inversion was found in CA1014, CBS10090, MMRL2647, and 2001935. Moreover, CA1014 and CBS10090 each present an exclusive

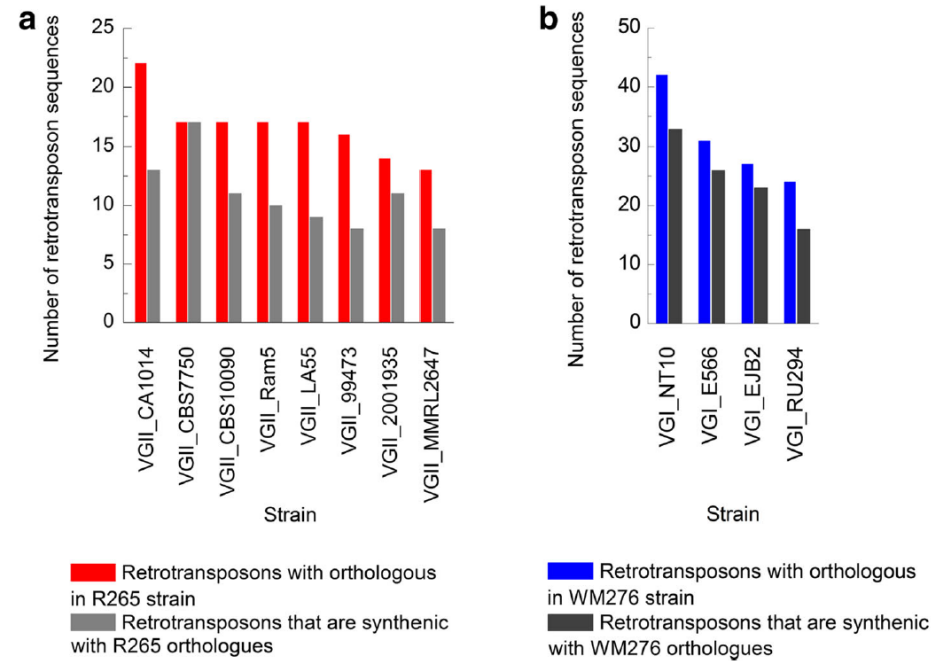

Fig. 3 Sequence count of orthologous and syntenic retrotransposons. *Bars* indicate the total amount of orthologous and syntenic sequences shared by VGII (**a**) and VGI (**b**) strains with R265 and WM276 genomes, respectively

syntenic location to sequence orthologs of retrotransposon R265.22.

Synteny conservation at retrotransposon orthologous sites indicates low mobilization rates and reveals that an RNAi pathway effectively acts in WM276 to protect the genome against retrotransposon propagation. In this context, it is feasible to assume that there are additional mechanisms protecting the VGII strain genomes from mobile elements.

Phylogenetic inference

Among the 34 retrotransposons belonging to the R265 genome, three RT sequences (named R265.3, R265.19, and R265.22) have orthologues present in all 17 *C. gattii* strains. Of these, only one (R265.19) was retained in *C. neoformans* H99. The RT R265.19 has two conserved domains (reverse transcriptase and RNase H) in all orthologous sequences (except in H99, where the sequence is inserted in a TE cluster) and was used as a model for the phylogenetic evaluation of *C. gattii* retrotransposon propagation. The evolutionary model selected for nucleotide substitution was Kimura 2-parameters [25]. The phylogram (Fig. 4) shows *C. gattii* diversification through the evolutionary path taken by reverse transcriptase (Fig. 4a) and RNase H (Fig. 4b) sequences. In both cases, monophyletic clades for 4 *C. gattii* genotypes are identifiable. Although the reverse transcriptase domain seems to have undergone more modifications, both trees indicate the relationship between VGI, VGIII, and VGIV, as well as the basal separation of the VGII clade. Early divergence and isolation of VGII molecular types as well as the late separation of VGI, VGIII, and VGIV from a common ancestor has been proposed by Farrer and colleagues [15] in an evolutionary relationship reconstruction for *C. gattii* genotypes.

Expression evaluation

The activation of transposable elements in response to stress conditions usually involves regulatory sequences at promoter regions, which are similar to motifs required for the induction of stress-responsive genes [26] as well as the preferential integration of some type of retrotransposon sequence into the promoters of stress response genes [27]. Different stress conditions, such as ionizing radiation, reactive oxygen species (ROS), DNA damage, and nutrient starvation can activate transcription and retrotransposition [26, 28]. Hence, element mobilization provide a means for the rapid evolution of gene regulatory networks, mediating the adaptive response to stress conditions [27]. Therefore, we speculated that stress conditions caused by zinc starvation in wild-type (WT) cells and cells lacking the *ZAP1* gene mutant would alter retrotransposon expression. Zinc restriction was show to induces oxidative stress in yeast cells [29–31]. *ZAP1* codes for a master zinc regulator in *C. gattii* and its is characterized in some fungal species [32]. Its absence was show to lead to reduction of anti-

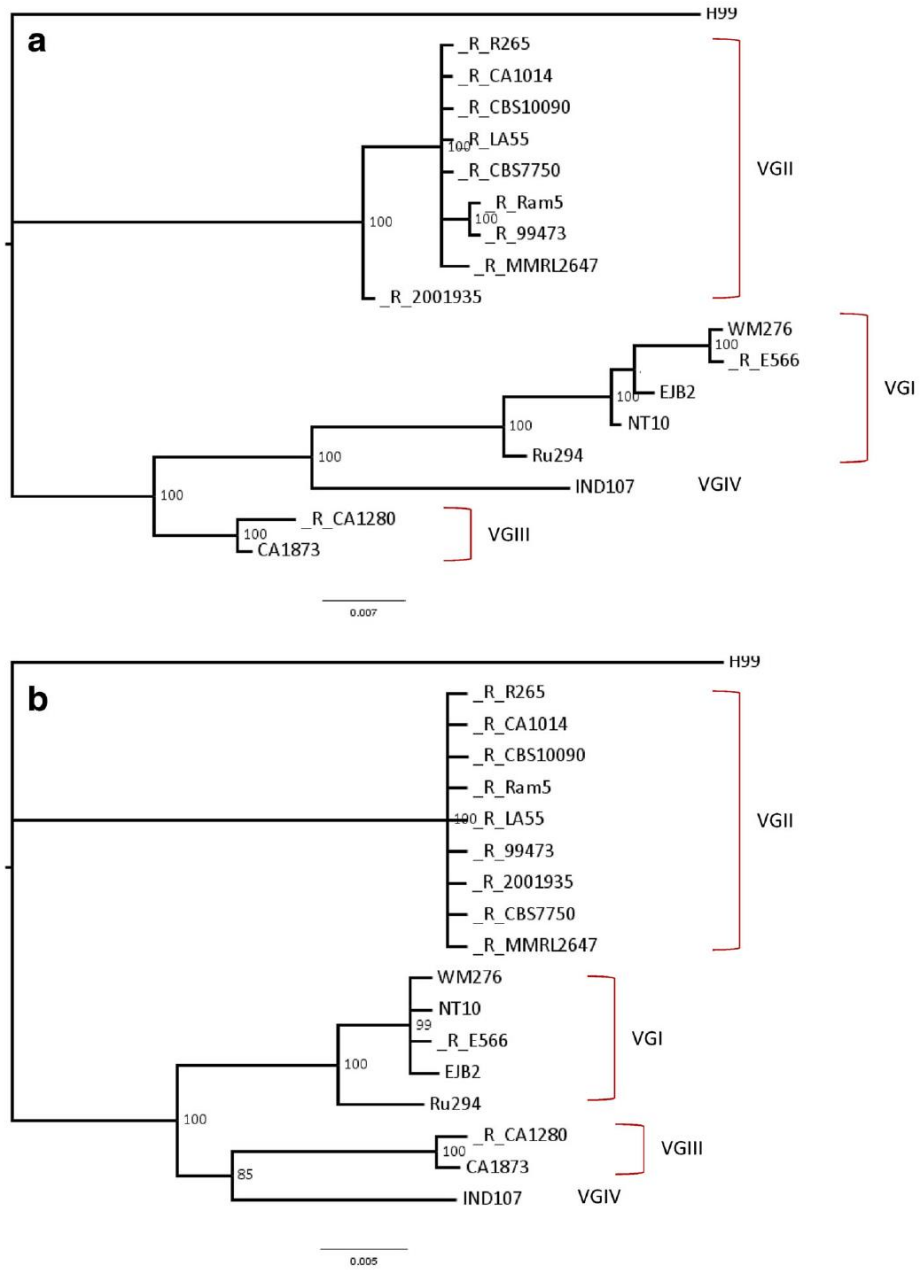

Fig. 4 Phylogenetic tree for reverse transcriptase (RT) and RNase H domains. Domains of reverse transcriptase (**a**) and ribonuclease H (**b**) from *C. gattii* retrotransposons to orthologous R265.19 sequence. Sequences that were direction-adjusted by MAFFT aligner are indicated by the "R" before strain name. Model inference was performed by jModelTest and phylogeny reconstruction by Mr. Bayes using the Kimura-2-parameters substitution model

oxidative responses in *C. gattii* cells [33] as well in *S. cerevisiae* [29, 34].

RNA-Seq data from *C. gattii* R265 WT and a *zap1* mutant grown in YNB media with zinc deprivation were used to evaluate retrotransposon expression and silencing in stress conditions. The read counts of aligned libraries were quantified and the log2 fold change of differential expression analysis was quantified and normalized to Reads Per Kilobase of tanscript per Million mapped reads (RPKM) values for transcription measurement (Table 1, Additional file 16). No drastic changes in the RPKM values or the number of mapped reads distribution could be detected for both conditions (Additional files 17 and 18). Despite the variability of expression among different retrotransposons, no differential expression of such elements between WT and zap1 strains could be detected. The only two exceptions were RTs R265.26 and R265.38, which are only detected in WT or

Table 1 RPKM and log2 fold change values in retrotransposon sequences

Retro transposon	RPKM sRNA_WT	RPKM mRNA_WT	RPKM mRNA_ZAP1	Log2 fold change (mRNA) ZAP1xWT
R265.1	2212.17	165.12	48.09	−1.78
R265.10	2728.76	182.41	318.81	0.80
R265.11	8396.39	14638.41	15671.02	0.10
R265.12	13322.54	133.26	16.63	−3.00
R265.13	3991.29	2050.13	677.89	−1.60
R265.14	8884.66	80.78	70.59	−0.19
R265.15	1676.88	0	0	N/A
R265.16	19299.97	6745.21	11206.42	0.73
R265.17	808.77	0	27	N/A
R265.18	5015.68	107.47	125.22	0.22
R265.19	48521.82	29562.31	36487.05	0.30
R265.2	2923.34	360.06	125.86	−1.52
R265.20	2567.82	168.44	183.99	0.13
R265.21	4481.6	23880.77	25869.49	0.11
R265.22	45424.93	259.33	226.62	−0.19
R265.23	3496.62	28.71	25.09	−0.19
R265.24	1074.35	145.76	42.46	−1.78
R265.25	18917.41	9940.13	12131.12	0.29
R265.26	2075.06	511.17	0	N/A
R265.27	4526.67	103.71	30.21	−1.78
R265.28	1553.01	40.86	35.71	−0.19
R265.29	5431.2	496.81	186.06	−1.42
R265.3	19795.43	7875.53	9215.24	0.23
R265.30	9472.7	2621.35	52.66	−5.64
R265.31	5159.63	129824.4	90122.72	−0.53
R265.34	4506.72	0	0	N/A
R265.36	103818.7	10138.81	20509.48	1.02
R265.37	39301.07	5627.55	22949.74	2.03
R265.38	3792.86	0	35.2714	N/A
R265.4	4525.26	75.71	132.32	0.80
R265.5	24813.18	22065.62	31063.56	0.49
R265.6	37947.45	514.37	345.77	−0.57
R265.7	22013.22	17762.5	16862.24	−0.07
R265.8	3893.5	241.84	1704.8	2.82

zap1 conditions, respectively. This suggests that the absence of Zap1 transcription factor in a zinc-limiting environment alters the activity of such retrotransposons. Thus, it is possible that motif recognition on the retrotransposon promoter or a genomic localization near to Zap-1 target genes may be the cause of expression decline.

Retrotransposon propagation leads to genomic instability, and as a consequence, genetic diversity and phenotypic variation [35]. However, to avoid deleterious effects, there are several multiple transcriptional and posttranscriptional repressing mechanisms [36]. Among these mechanisms are those mediated by small RNAs, as previously described for the RNAi proficient *C. neoformans* H99 strain [9]. We therefore speculated that sRNAs could be controlling the expression of retrotransposons in *C. gattii* R265. We determined the sequence of small RNAs recovered form *C. gattii* R265 grow in zinc limitation conditions. Read length analysis revealed an abundance pattern somewhat similar to those observed in RNAi-proficient *C. neoformans* H99. While *C. gattii* most abundant reads ranges from 14 to 26 nt (data

not shown), the majority of reads form *C. neoformans* library ranges from 21 to 23 nt [9]. The presence of high mRNA expression with low sRNA associated levels (R265.31, R265.21) and high sRNAs levels with low expressions of mRNAs (R265.36, R265.22, R265.37, R265.6) may indicate an sRNA-regulated silencing mechanism for retrotransposon propagation. The correlation of mRNA x sRNA RPKM values from WT libraries is exhibited in Fig. 5.

The mapping profile of sRNAs and mRNAs from the WT condition in the two most highly expressed full-length retrotransposons (R265.5 and R265.7) can be observed in Fig. 6. Domain loss in many sequences, as well as the mutations and frameshifts in coding regions that render several domains non-functional, hampers the correct evaluation of domain expression patterns for all retrotransposons. Visually, sRNAs seem to be widely distributed and mRNAs are more abundant in LTR-5′ (Additional files 19, 20 and 21). Statistically, the comparison between read counts of mRNAs distributed in LTRs, protein domains, and spacer regions (SR) exhibits significant a difference ($p < 0{,}0001$ to $p < 0{,}05$) due to the increased values of LTR-5′ read counts compared to other regions. As previously noted, long terminal repeats (LTRs) are important for transcription regulation (LTR-5′, specifically, contains the promoter sequence) and usually contain high amounts of aligned reads. The absence of statistical variation inside the internal region of retrotransposons (consistent with the entire retrotransposon transcription) and the higher density of RNA-Seq reads mapping to LTR-5′ was previously observed [37] in a large-scale LTR retrotransposon analysis of transcriptome data.

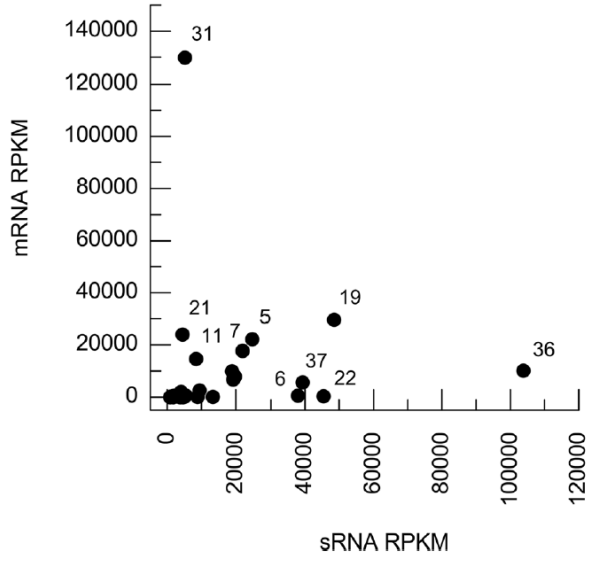

Fig. 5 mRNA and sRNA retrotransposon correlation. RPKM values of *C. gattii* R265 retrotransposons for mRNA and sRNA expression of wild-type (WT) *C. gattii* R265 in a zinc deprivation culture

The sRNA profile analysis of raw read counts with R265.5 and R265.7 sequences reveals a decrease in aligned reads to protein domain sequences, especially reverse transcriptase (RT) and protease (PROT). As reported by Domingues and colleagues [38], most LTR retrotransposon families do not have high sRNA counts within the coding domains. The increased sRNAs read count in both LTRs (LTR-5′ with statistical significance) and spacer regions (SR1, SR2, and SR3) shows the possible target regions for sRNA activity in a silencing mechanism.

Discussion

Retrotransposons are ubiquitously distributed in fungal species, albeit their structure is not conserved. The *pol* gene encodes a polyprotein composed of an aspartic protease, a reverse transcriptase, a ribonuclease H, and an integrase; the order and number of domains may differ between retrotransposon families, and most of the elements have incomplete *pol* genes. In addition, retrotransposons are generally characterized by the presence of LTR [22]. As previously described [2], *C. gattii* retrotransposon sequences can be grouped into (i) LTRs with associated, but incomplete, internal regions; and (ii) partial internal regions with no identified LTRs.

The comparative analysis of *C. gattii* retrotransposons showed here revealed a large diversity of retrotransposons among diverse genotypes. It is noteworthy that the number of retrotransposons families number changes in a genotype-associated fashion. The VGI genotype strains possess the higher detectable retrotransposons and consequently the higher density compared to strains from VGII, VGIII and VGIV genotypes. This diversification of the number of retrotransposon families among different fungal strains is in agreement with previous reports comparing the prevalence of such mobile elements among closely related species and strains of *Coccidioides immitis* and *Coccidioides posadasii* [39], as well for *Paracoccidioides* spp. [40].

The genomic sequence of 16 *C. gattii* strains are available and their analysis described. Comparative genomic analysis revealed that chromosomal structure was very conserved among the four varieties (VGI, VGII, VGIII, and VGIV). Furthermore, this conservation is even more pronounced among VGII strains [15]. As expected, our retrotransposon conservation analysis revealed the majority of these mobile elements present in R265 strain have orthologs in other *C. gattii* strains, most of them displaying also synteny.

Strains of the VGII genotype do not have a functional RNAi pathway, a highly conserved mechanism to control and inhibit retrotransposon expression through a

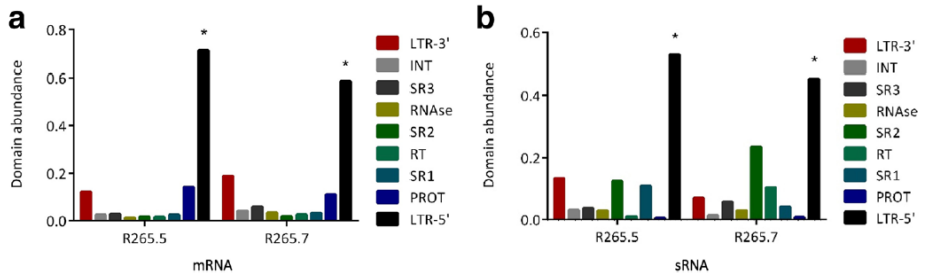

Fig. 6 mRNA and sRNA reads profile mapping. Distribution of aligned reads from **a** mRNA and **b** sRNA libraries in terminal repeats, protein domains, and spacer regions inside retrotransposon sequences. Abbreviations: Long Terminal Repeats (LTR), Integrase (INT), Ribonuclease H (RNAse), Reverse Transcriptase (RT), Aspartyl Protease (PROT)

homology-dependent gene-silencing complex [16]. It is therefore feasible to assume that VGII strains do not have an active RNAi pathway and that increased retrotransposons numbers are a consequence of increased retrotransposon expression and mobility, as seen in RNAi-deficient mutant of *C. neoformans* [8]. Unexpectedly, VGII strains displayed a smaller number of retrotransposon sequences compared to RNAi-proficient cryptococcal genotypes. Although R265 hypervirulence seems related to the absence of this regulatory mechanism, this observation suggests the existence of alternative control pathways or non-canonical Argonaute proteins, or the possibility that *C. gattii* (and not only VGII) has a more efficient (RNAi independent) system to control mobile element propagation. A predominance of LTR and DNA elements, LINEs, and low-complexity repeats in certain VGI strains was observed in *C. gattii* genotype analysis [15]. Furthermore, *AGO* gene absence in R265 and other VGII strains was not associated with gaps in the R265 genome [24]. However, it is unclear how far the incomplete assembly of genomes may alter the retrotransposon count. WM276 and NT10 have more retrotransposons and a well-assembled genome (WM276). Nevertheless, EJB2, E566, RU294, as both VGIII and VGIV strains, have RNAi pathways and exhibit similar sequence counts. The same explanation was suggested by Muszewska and colleagues [22] when analyzing differences in retrotransposon amounts between distinct strains of two *Coccidioides* species. Their findings further support that identified element amounts correlate with genomic sequences assembly lengths.

The retrotransposon expression analysis showed here revealed that all retrotransposons have detectable transcripts in distinct stress conditions. An uneven distribution of reads was detected among the *C. gattii* R265 transposons, with the large majority of reads matching one of the LTRs. This is in accordance to RNA-Seq assays using mRNA isolated from fission yeast [37].

To further evaluate retrotransposons expression, we performed a comparative transcriptome analysis using RNA-Seq data from *C. gattii* WT and *zap1* mutant strains cultured under zinc restriction [33]. Iron, copper, and zinc are essential metals for many processes in fungal pathogens. Zinc, specifically, is involved in transcriptional control, ROS detoxification, carbohydrate oxidation, and alcoholic fermentation [41, 42]. Thereby, in *Cryptococcus*, the ZIP family of plasma membrane transporters mediates zinc acquisition, especially in zinc-limiting conditions. The expression regulation of these transporter-encoding genes is performed by the transcriptional factor *ZAP1*. Our group previously demonstrated a severe impairment in ROS detoxification in cryptococcal cells lacking *ZAP1* gene [33]. This can be explained by the absence of indirect activation of many biological processes affected by zinc depletion in the *zap1* mutant, since the complete removal of this metal harms all cellular functions of zinc-dependent metalloproteins. In this way, the comparison of WT and *zap1* mutant cells allowed us to explore two distinct stress related conditions. However, our comparative analysis allowed the identification of only 2 stress-regulated retrotransposons.

Control of retrotransposon expression in fungal species is poorly characterized. Heterochromatin formation by H3mK9 histone methyltransferase recruitment in *Schizosaccharomyces pombe* is mediated by specific siRNAs acting in an Argonaute complex [43, 44]. *Saccharomyces cerevisiae* (which lacks the RNAi machinery for silencing mobile elements) produces antisense non-coding RNAs to regulate gene expression at the transcriptional level [26] and post-transcriptionally, as an auxiliary in copy number control mechanism. This maintains lower levels of mature integrase and reverse transcriptase, coupled with a truncated gag [45]. At the transcriptional level, Ty1 antisense RNAs regulate Ty1 retrotransposon expression by alterations in chromatin function [45].

In *C. neoformans*, retrotransposon activity is controlled by the RNAi pathway [8]. Stalled spliceosomes associate with a complex named SCANR, leading to the formation of siRNAs responsible for retrotransposon silencing via AGO dependent RNA degradation [9]. As *C. gattii* VGII strains lack components of RNAi pathway, it is reasonable to assume that this strategy is not used by VGII strains to control retrotransposons activity. However, the detection of sRNAs that map to retrotransposon sequences with low expression claims that additional mechanisms for retrotransposon activity control may exist in *C. gattii* VGII strains. It is noteworthy that *C. gattii* R265 possess a functional *DCR2* gene, which codes the enzyme responsible for the processing of siRNA precursors. In *C. neoformans,* Dcr2 was shown to be important for RNAi-mediated gene silencing [8]. In addition, the *C. gattii* VGII genomes accumulate a large amount of lost genes [16] and a hypervirulent phenotype. The absence of a canonical RNAi pathway can be a major influence on this profile. Nevertheless, the production of sRNAs means that *C. gattii* R265 has a secondary, alternative retrotransposon silencing mechanism, which guarantees genetic variation without the deleterious invasion of parasitic elements.

Conclusions

WM276 (VGI) is the only strain with a well-assembled genome and is therefore used as a parameter for predicting results in *C. gattii* analyses. Differences in retrotransposon count can be explained by incomplete genome assemblies. This is mainly because VGIII and VGIV strains, which are genetically more similar to VGI, have a functional RNAi pathway, but also have low numbers of retrotransposons (as VGII) when compared to WM276 and NT10. Sequences grouped in TE clusters are a strong feature of WM276 retrotransposons. Retrotransposon orthology showed the prevalence of shared sequences with related strains, a result of "ancient" retrotransposons from before *C. gattii* diversification. The unique sequences of WM276, if not a genomic assembly consequence, may indicate recent transposition events. Synteny conservation reflects RNAi importance in the control of mobile element propagation. VGI molecular type has superior rates of synteny maintenance within genotypes than VGII, and the same number of retrotransposon-associated synteny block breakpoints. This confirms that R265 retrotransposons exert an effect on genome stability. Besides differences in TE content, the phylogenetic inference of the R265.19 element demonstrated that retrotransposon propagation by nucleotide substitution analysis resembles *C. gattii* strain expansion, also showing a greater conservation of RNase and reverse transcriptase in VGII. In addition, expression evaluation confirms the existence of retrotransposons expression regulation.

Further studies are necessary to fully identify the mobile elements in *C. gattii* and, especially, to completely assemble its genome(s). However, the presence of sRNAs as possible agents to control sequence mobilization creates a new basis for research on alternative pathways, *AGO-like* domains, and other mechanisms that promote non-canonical retrotransposon silencing in this medically relevant fungus.

Methods

Retrotransposon mining

Retrotransposon mining with LTR_Finder v1.0.6 [46], LTR_Harvest v1.5.7 [47], LTR_STRUC v1.1 [48], LTR_Seq [49], MGEScan v2 [50], and Transposon_PSI (http://transposonpsi.sourceforge.net) programs were used in *C. gattii* genomes with default parameters. The 17 *C. gattii* genomes (Table 2) from VGI, VGII, VGIII, and VGIV molecular types used in this work are available in the NCBI database .

New sequences were discovered by searching with the Blastn algorithm v2.2.31 (megablast) [51, 52] of known retrotransposons against genomes. Sequence confirmation by retrotransposon domain presence (gag, integrase, reverse transcriptase, RNase H, and aspartic protease) was evaluated using Blast Conserved Domains [53].

Retrotransposon orthology between genomes

Sequences that are putatively orthologous to R265 and WM276 retrotransposons were inferred by *best bi-*

Table 2 *C. gattii* strains used in retrotransposon mining

Molecular type	Strain	NCBI Reference number
VGI	WM276	GCA_000185945.1
VGI	NT10	GCA_000935105.1
VGI	E566	GCA_000875815.1
VGI	EJB2	GCA_000835745.1
VGI	RU294	GCA_000836355.1
VGII	99473	GCA_000836455.1
VGII	2001935	GCA_000835815.1
VGII	CA1014	GCA_000875795.1
VGII	CBS7750	GCA_000499585.1
VGII	CBS10090	GCA_000835765.1
VGII	LA55	GCA_000836315.1
VGII	MMRL2647	GCA_000875855.1
VGII	R265	GCA_000149475.3
VGII	RAM5	GCA_000836375.1
VGIII	CA1280	GCA_000836335.1
VGIII	CA1873	GCA_000855695.1
VGIV	IND107	GCA_000835755.1

directional hits (BBH - with minimum 90% query cover, 90% identity) in a Blastn search (megablast) with retrotransposon sequences against genomes. R265 retrotransposons distributed in all *C. gattii* strains (R265.3, R265.19, and R265.22) were tested for orthologous sequences in *C. neoformans* H99 (GCA_000149245.3) by the same methodology.

Synteny evaluation

Multiple genome alignment for synteny analysis of widely distributed R265.3, R265.19, and R265.22 orthologous sequences was performed in Mauve aligner v20150226 [54] with default parameters for all 17 *C. gattii* strains. Intrinsic synteny by orthologous retrotransposon localization inside VGI and VGII genotypes was analyzed on an independent multiple genome alignment of VGI and VGII strains. Syntenic blocks between genomes were represented by the same color.

Phylogenetic inference

Reverse trancriptase and RNase H domain sequences (nucleotide) from R265.19 and orthologous retrotransposons (17 *C. gattii* sequences and one *C. neoformans* H99 orthologue, as an outgroup) were separately aligned using MAFFT v7 [55] with default parameters for alignment strategy and adjusted sequence direction option. jModelTest v2.1.10 [56] was utilized to determine the best evolutionary model describing the data and Mr. Bayes v3.2.6 [57] was used for Bayesian analysis of phylogenetic inference. Computed trees were visualized with FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/) .

C. gattii R265 RNA-Seq of small RNAs

The *C. gattii* strain R265 was cultured in YPD media (2% glucose; 2% peptone; 1% yeast extract) for 18 h at 200 rpm and 30 °C. Next, the cells were centrifuged (5000 × *g* – 5 min) and washed with PBS (NaCl 137 mM; KCl 2,7 mM; Na_2HPO_4 10 mM; KH_2PO_4 1,8 mM; pH 7,4). The cell pellet was resuspended in 20 mL of 1x YNB (*Yeast Nitrogen Base*) and diluted to 10^7 cells ($OD_{600} = 1$). For the zinc deprivation assay, the quantified cells were inoculated in 100 mL of YNB 1x + 10 μM TPEN and kept for 18 h at 200 rpm. Afterwards, the cells were centrifuged (5000 g–5 min), frozen in liquid nitrogen, and placed in an ultrafreezer (–80 ° C) for lyophilization.

RNA extraction was performed with Trizol® (Invitrogen) reagent according to the manufacturer's protocol. Sample quantification was evaluated by fluorometric analysis with Qubit and Quant-It (Invitrogen), and samples were stored in RNA stable reagent (Biometrica). sRNAs were purified from total RNA samples, processed, and sequenced by Fasteris Life Sciences SA, Plan-les-Ouates, Switzerland using Solexa technology on an Illumina Hi Seq 2000 platform.

sRNAs prediction

Reads from the zinc deprivation were filtered. Sequences from ribosomal RNAs (rRNAs), transporter RNAs (tRNAs), small nuclear RNAs (snRNAs), and small nucleolar RNAs (snoRNAs) present in the Rfam data bank [58] were excluded. Only sequences between 12 and 28 nucleotides were retained in the library. These sequences were aligned to Retrotransposon sequences using Bowtie2 [59]. Domains containing values of RPKM equal or larger than 500 RPKM were considered as possible sRNA targets.

Expression evaluation

The *C. gattii* R265 mRNA libraries of WT and *ZAP1* depleted mutant (both grown in zinc deprivation conditions) used in this study were obtained from Schneider and colleagues [33].

RNA-Seq libraries (mRNAS and sRNAs) were aligned against the *C. gattii* R265 genome with the Tophat v2.1.1 program [60] unique alignment option (*max-multihits = 1*). Mapped reads were visualized with IGV v2.3.78 [61] and quantified with Samtools v0.1.19 [62]. RPKM (*reads Per kilobase per million*) values were manually calculated and Log2 fold change was determined using the DESeq2 package of R [63]. Domain abundance inside retrotransposons was quantified by the division of raw read counts from each region by the sequence total read count. The statistical analysis of results was performed by ANOVA.

Additional files

Additional file 1: Retrotransposon sequence list from all 17 *C. gattii* strains. The retrotransposon list contains detailed descriptions of molecular type, *C. gattii* strain, NCBI reference for supercontig, sequence coordinates, LTR retrotransposon family and domains; as well as, the annotation of orthologous sequences in WM276 and R265 strains. The synteny analysis columns indicate presence ("YES") and absence ("NO") of synteny; and "NA" to not analyzed data. (XLS 118 kb)

Additional file 2: C. gattii multiple genome alignment. Multiple genome alignment of VGI, VGII, VGIII and VGIV molecular types with Mauve aligner. Syntenic blocks shared by genomes are represented with the same color and are connected by lines; red lines indicate chromosome or supercontig boundaries. (JPG 592 kb)

Additional file 3: VGI multiple genome alignment. Multiple genome alignment of strains from VGI molecular type with Mauve aligner. Syntenic blocks shared between genomes are represented with the same color and are connected by lines; red lines indicate chromosome or supercontig boundaries. (JPG 1033 kb)

Additional file 4: Retrotransposon-associated synteny breakpoints 1 in WM276. Figure shows a multiple genome alignment of VGI strains. Syntenic blocks shared by genomes are represented with the same color and are connected by lines; red lines indicate chromosome or supercontig boundaries. The location of associated retrotransposons WM276.1, WM276.2, WM276.3 and WM276.4 in chromosome A of WM276 strain is represented

by the black rectangle. Microsynteny disruption by retrotransposon insertion can be observed as white spaces inside blocks. (JPG 308 kb)

Additional file 5: Retrotransposon-associated synteny breakpoints 2 in WM276. Figure shows a multiple genome alignment of VGI strains. Syntenic blocks shared by genomes are represented with the same color and are connected by lines; red lines indicate chromosome or supercontig boundaries. The location of associated retrotransposons WM276.26, WM276.75, WM276.27 and WM276.28 in chromosome E of WM276 strain is represented by the black rectangle. Microsynteny disruption by retrotransposon insertion can be observed as white spaces inside blocks. (JPG 304 kb)

Additional file 6: Retrotransposon-associated synteny breakpoints 3 in WM276. Figure shows a multiple genome alignment of VGI strains. Syntenic blocks shared by genomes are represented with the same color and are connected by lines; red lines indicate chromosome or supercontig boundaries. The location of associated retrotransposons WM276.44, WM276.46 and WM276.47 in chromosome H of WM276 strain is represented by the black rectangle. Microsynteny disruption by retrotransposon insertion can be observed as white spaces inside blocks. (JPG 328 kb)

Additional file 7: Retrotransposon-associated synteny breakpoints 4 in WM276. Figure shows a multiple genome alignment of VGI strains. Syntenic blocks shared by genomes are represented with the same color and are connected by lines; red lines indicate chromosome or supercontig boundaries. The location of associated retrotransposons WM276.52, WM276.53, WM276.54 and WM276.55 in chromosome J of WM276 strain is represented by the black rectangle. Microsynteny disruption by retrotransposon insertion can be observed as white spaces inside blocks. (JPG 356 kb)

Additional file 8: Retrotransposon-associated synteny breakpoints 5 in WM276. Figure shows a multiple genome alignment of VGI strains. Syntenic blocks shared by genomes are represented with the same color and are connected by lines; red lines indicate chromosome or supercontig boundaries. The location of associated retrotransposons WM276.72 and WM276.73 in chromosome N of WM276 strain is represented by the black rectangle. Microsynteny disruption by retrotransposon insertion can be observed as white spaces inside blocks. (JPG 329 kb)

Additional file 9: Retrotransposon-associated synteny breakpoints 1 in R265. Figure shows a multiple genome alignment of VGII strains. Syntenic blocks shared by genomes are represented with the same color and are connected by lines; red lines indicate chromosome or supercontig boundaries. The location of associated retrotransposons R265.12 and R265.13 in supercontig 7 of R265 strain is represented by the black rectangle. Microsynteny disruption by retrotransposon insertion can be observed as white spaces inside blocks. (JPG 483 kb)

Additional file 10: Retrotransposon-associated synteny breakpoints 2 in R265. Figure shows a multiple genome alignment of VGII strains. Syntenic blocks shared by genomes are represented with the same color and are connected by lines; red lines indicate chromosome or supercontig boundaries. The location of associated retrotransposons R265.29, R265.30 and R265.31 in supercontig 25 of R265 strain is represented by the black rectangle. Microsynteny disruption by retrotransposon insertion can be observed as white spaces inside blocks. (JPG 436 kb)

Additional file 11: Retrotransposon-associated synteny breakpoints 3 in R265. Figure shows a multiple genome alignment of VGII strains. Syntenic blocks shared by genomes are represented with the same color and are connected by lines; red lines indicate chromosome or supercontig boundaries. The location of associated retrotransposons R265.23 and R265.24 in supercontig 17 of R265 strain is represented by the black rectangle. Microsynteny disruption by retrotransposon insertion can be observed as white spaces inside blocks. (JPG 493 kb)

Additional file 12: Retrotransposon-associated synteny breakpoints 4 in R265. Figure shows a multiple genome alignment of VGII strains. Syntenic blocks shared by genomes are represented with the same color and are connected by lines; red lines indicate chromosome or supercontig boundaries. The location of associated retrotransposons R265.26, R265.27, R265.34 and R265.28 in supercontig 21 of R265 strain is represented by the black rectangle. Microsynteny disruption by retrotransposon insertion can be observed as white spaces inside blocks. (JPG 508 kb)

Additional file 13: Retrotransposon-associated synteny breakpoints 5 in R265. Figure shows a multiple genome alignment of VGII strains. Syntenic blocks shared by genomes are represented with the same color and are connected by lines; red lines indicate chromosome or supercontig boundaries. The location of associated retrotransposons R265.38 and R265.20 in supercontig 13 of R265 strain is represented by the black rectangle. Microsynteny disruption by retrotransposon insertion can be observed as white spaces inside blocks. (JPG 487 kb)

Additional file 14: VGII multiple genome alignment. Multiple genome alignment of strains from VGII molecular type with Mauve aligner. Syntenic blocks shared by genomes are represented with the same color and are connected by lines; red lines indicate chromosome or supercontig boundaries. (JPG 2507 kb)

Additional file 15: Synteny analysis of R265.3, R265.19 and R265.22 orthologous sequences. Localization of orthologous retrotransposons of R265.3, R265.19, and R265.22 sequences are indicated by black lines in a multiple genome alignment with strains of VGI, VGII, VGIII, and VGIV molecular types. Orthologous sequences in each strain are named by the retrotransposon number. Syntenic blocks shared by genomes are represented with the same color and are connected by lines; red lines indicate chromosome or supercontig boundaries. (JPG 1397 kb)

Additional file 16: Raw read counts of mRNAs and sRNAs. Raw counts of mapped reads from mRNA and sRNA libraries to *C. gattii* R265 retrotransposons by the TopHat aligner. (XLSX 10 kb)

Additional file 17: Read counts values of retrotransposon sequences. Raw read counts distribution of retrotransposon sequences in WT and ZAP1 conditions. Distribution values are shown in a log scale. (JPG 27 kb)

Additional file 18: RPKM values of retrotransposon sequences. RPKM distribution of retrotransposon sequences in WT and ZAP1 conditions. Distribution values are shown in a log scale. (JPG 31 kb)

Additional file 19: Reads mapping profile of in full-length retrotransposons. Visualization of mapped reads profile from mRNA and sRNA libraries in full-length retrotransposons R265.5 and R265.7. Abbreviations: Long Terminal Repeats (LTR), Integrase (INT), Ribonuclease H (RNAse), Reverse Transcriptase (RT), Aspartyl Protease (PROT). (JPG 289 kb)

Additional file 20: Reads mapping profile of highly expressed retrotransposons. Visualization of mapped reads profile from mRNA and sRNA libraries in highly expressed sequences R265.19 and R265.31. Abbreviations: Long Terminal Repeats (LTR), Integrase (INT), Ribonuclease H (RNAse), Reverse Transcriptase (RT), Aspartyl Protease (PROT). (JPG 285 kb)

Additional file 21: Reads mapping profile of highly repressed retrotransposons. Visualization of mapped reads profile from mRNA and sRNA libraries in highly repressed sequences R265.36 and R265.37. Abbreviations: Long Terminal Repeats (LTR), Integrase (INT), Ribonuclease H (RNAse), Reverse Transcriptase (RT), Aspartyl Protease (PROT). (JPG 287 kb)

Abbreviations

dsRNA: Double stranded RNA; GAGs: Group-specific AntiGen; LTR: Long terminal repeats; OD_{600}: Optical density at 600 nm; PBS: Phosphate buffered saline; RdRP: RNA dependent RNA polymerase; RISC: RNA-induced silencing complex; RNAi: Interfering RNA; ROS: Reactive oxygen species; RPKM: Reads Per Kilobase of tanscript per Million mapped reads; siRNA: Small interfering RNA; sRNA: Small RNA; TE: Transposable elements; TPEN: N,N,N′,N′-tetrakis(2-pyridinylmethyl)-1,2-ethanediamine; VG: Variety *gattii*; VLPs: Virus-Like Particles; WT: Wild type; YNB: Yeast nitrogen base; YPD: Yeast peptone dextrose

Acknowledgment

The authors would to thank Dr. Rodrigo Ligabue Braun for critical reading of the manuscript.

Funding

Financial support for this study was provided by Coordenação de Apoio de Pessoal de Nível Superior – CAPES (Process no. 23038.010041/2013-13) and Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPQ (Process no. 312408/2014-5). The funding agencies had no role in the design of the study and collection, analysis, and interpretation of data, as well as in writing the manuscript.

Availability of data and materials
The data supporting the conclusions of this article will be available at the Sequence Read Archive (SRA) under codes SRR5209627, SRR5209175, and SRR5209145 in the NCBI GenBank repository as BioProject PRJNA368962.

Authors' contributions
Conceived and designed the experiments: PAGF RSAS FMS LK AS MHV CCS. Performed the experiments: PAGF RSAS FMS. Analyzed the data: PAGF CCS. Contributed reagents/materials/analysis tools: LK AS MHV CCS. Wrote the paper: PAGF CCS. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Nothing to declare.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 31 January 2017 Accepted: 6 April 2017
Published online: 12 April 2017

References
1. Chénais B, Caruso A, Hiard S, Casse N. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. Gene. 2012;509:7–15.
2. Goodwin TJ, Poulter RT. The diversity of retrotransposons in the yeast *Cryptococcus neoformans*. Yeast. 2001;18:865–80.
3. Alzohairy AM, Gyulai G, Jansen RK, Bahieldin A. Transposable elements domesticated and neofunctionalized by eukaryotic genomes. Plasmid. 2013; 69:1–15.
4. Havecker ER, Gao X, Voytas DF. The diversity of LTR retrotransposons. Genome Biol. 2004;5:225.
5. Nunes CC, Gowda M, Sailsbery J, Xue M, Chen F, Brown DE, Oh Y, Mitchell TK, Dean RA. Diverse and tissue-enriched small RNAs in the plant pathogenic fungus, *Magnaporthe oryzae*. BMC Genomics. 2011;12:288.
6. Wang X, Hsueh YP, Li W, Floyd A, Skalsky R, Heitman J. Sex-induced silencing defends the genome of *Cryptococcus neoformans* via RNAi. Genes Dev. 2010;24:2566–82.
7. Moazed D. Small RNAs in transcriptional gene silencing and genome defence. Nature. 2009;457:413–20.
8. Janbon G, Maeng S, Yang DH, Ko YJ, Jung KW, Moyrand F, Floyd A, Heitman J, Bahn YS. Characterizing the role of RNA silencing components in *Cryptococcus neoformans*. Fungal Genet Biol. 2010;47:1070–80.
9. Dumesic PA, Natarajan P, Chen C, Drinnenberg IA, Schiller BJ, Thompson J, Moresco JJ, Yates JR, Bartel DP, Madhani HD. Stalled spliceosomes are a signal for RNAi-mediated genome defense. Cell. 2013;152:957–68.
10. Shi H, Djikeng A, Tschudi C, Ullu E. Argonaute protein in the early divergent eukaryote *Trypanosoma brucei*: control of small interfering RNA accumulation and retroposon transcript abundance. Mol Cell Biol. 2004; 24:420–7.
11. Boesler B, Meier D, Förstner KU, Friedrich M, Hammann C, Sharma CM, Nellen W. Argonaute proteins affect siRNA levels and accumulation of a novel extrachromosomal DNA from the *Dictyostelium retrotransposon* DIRS-1. J Biol Chem. 2014;289:35124–38.
12. Chaturvedi V, Chaturvedi S. *Cryptococcus gattii:* a resurgent fungal pathogen. Trends Microbiol. 2011;19:564–71.
13. Bielska E, May RC. What makes *Cryptococcus gattii* a pathogen? FEMS Yeast Res. 2016;16:fov106.
14. Okubo Y, Wakayama M, Ohno H, Yamamoto S, Tochigi N, Tanabe K, Kaneko Y, Yamagoe S, Umeyama T, Shinozaki M, et al. Histopathological study of murine pulmonary cryptococcosis induced by *Cryptococcus gattii* and *Cryptococcus neoformans*. Jpn J Infect Dis. 2013;66:216–21.
15. Farrer RA, Desjardins CA, Sakthikumar S, Gujja S, Saif S, Zeng Q, Chen Y, Voelz K, Heitman J, May RC, et al. Genome evolution and innovation across the four major lineages of *Cryptococcus gattii*. MBio. 2015;6:e00868–00815.
16. Feretzaki M, Billmyre RB, Clancey SA, Wang X, Heitman J. Gene network polymorphism illuminates loss and retention of novel RNAi silencing components in the *Cryptococcus* pathogenic species complex. PLoS Genet. 2016;12:e1005868.
17. Byrnes 3rd EJ, Li W, Lewit Y, Ma H, Voelz K, Ren P, Carter DA, Chaturvedi V, Bildfell RJ, May RC, et al. Emergence and pathogenicity of highly virulent *Cryptococcus gattii* genotypes in the northwest United States. PLoS Pathog. 2010;6:e1000850.
18. Billmyre RB, Calo S, Feretzaki M, Wang X, Heitman J. RNAi function, diversity, and loss in the fungal kingdom. Chromosome Res. 2013;21:561–72.
19. Servant G, Pennetier C, Lesage P. Remodeling yeast gene transcription by activating the Ty1 long terminal repeat retrotransposon under severe adenine deficiency. Mol Cell Biol. 2008;28:5543–54.
20. Lerat E. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity. 2010;104:520–33.
21. Thon MR, Pan H, Diener S, Papalas J, Taro A, Mitchell TK, Dean RA. The role of transposable element clusters in genome evolution and loss of synteny in the rice blast fungus Magnaporthe oryzae. Genome Biol. 2006;7:R16.
22. Muszewska A, Hoffman-Sommer M, Grynberg M. LTR retrotransposons in fungi. PLoS One. 2011;6:e29425.
23. Novikov A, Smyshlyaev G, Novikova O. Evolutionary history of LTR retrotransposon chromodomains in plants. Int J Plant Genomics. 2012;2012: 874743.
24. D'Souza CA, Kronstad JW, Taylor G, Warren R, Yuen M, Hu G, Jung WH, Sham A, Kidd SE, Tangen K, et al. Genome variation in *Cryptococcus gattii*, an emerging pathogen of immunocompetent hosts. MBio. 2011;2:e00342–00310.
25. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980;16:111–20.
26. Servant G, Pinson B, Tchalikian-Cosson A, Coulpier F, Lemoine S, Pennetier C, Bridier-Nahmias A, Todeschini AL, Fayol H, Daignan-Fornier B, et al. Tye7 regulates yeast Ty1 retrotransposon sense and antisense transcription in response to adenylic nucleotides stress. Nucleic Acids Res. 2012;40:5271–82.
27. Zaratiegui M. Influence of long terminal repeat retrotransposons in the genomes of fission yeasts. Biochem Soc Trans. 2013;41:1629–33.
28. VanHoute D, Maxwell PH. Extension of *Saccharomyces paradoxus* chronological lifespan by retrotransposons in certain media conditions is associated with changes in reactive oxygen species. Genetics. 2014;198:531–45.
29. Eide DJ. Homeostatic and adaptive responses to zinc deficiency in Saccharomyces cerevisiae. J Biol Chem. 2009;284:18565–9.
30. Wu CY, Roje S, Sandoval FJ, Bird AJ, Winge DR, Eide DJ. Repression of sulfate assimilation is an adaptive response of yeast to the oxidative stress of zinc deficiency. J Biol Chem. 2009;284:27544–56.
31. Wu CY, Bird AJ, Winge DR, Eide DJ. Regulation of the yeast TSA1 peroxiredoxin by ZAP1 is an adaptive response to the oxidative stress of zinc deficiency. J Biol Chem. 2007;282:2184–95.
32. Ballou ER, Wilson D. The roles of zinc and copper sensing in fungal pathogenesis. Curr Opin Microbiol. 2016;32:128–34.
33. Schneider RO, Fogaça NS, Kmetzsch L, Schrank A, Vainstein MH, Staats CC. Zap1 regulates zinc homeostasis and modulates virulence in *Cryptococcus gattii*. PLoS One. 2012;7:e43773.
34. Zhao H, Eide DJ. Zap1p, a metalloregulatory protein involved in zinc-responsive transcriptional regulation in *Saccharomyces cerevisiae*. Mol Cell Biol. 1997;17:5044–52.
35. Cavrak W, Lettner N, Jamge S, Kosarewicz A, Bayer LM, Mittelsten Scheid O. How a retrotransposon exploits the plant's heat stress response for its activation. PLoS Genet. 2014;10:e1004115.
36. Nakayashiki H. The Trickster in the genome: contribution and control of transposable elements. Genes Cells. 2011;16:827–41.
37. Mourier T, Willerslev E. Large-scale transcriptome data reveals transcriptional activity of fission yeast LTR retrotransposons. BMC Genomics. 2010;11:167.
38. Domingues DS, Cruz GM, Metcalfe CJ, Nogueira FT, Vicentini R, Alves CS, Van Sluys MA. Analysis of plant LTR-retrotransposons at the fine-scale family level reveals individual molecular patterns. BMC Genomics. 2012;13:137.
39. Desjardins CA, Champion MD, Holder JW, Muszewska A, Goldberg J, Bailao AM, Brigido MM, Ferreira ME, Garcia AM, Grynberg M, et al. Comparative genomic analysis of human fungal pathogens causing paracoccidioidomycosis. PLoS Genet. 2011;7:e1002345.
40. Soares MA, de Carvalho Araujo RA, Marini MM, de Oliveira LM, de Lima LG, De Souza Alves V, Felipe MS, Brigido MM, de Almeida Soares CM, da Silveira JF, et

al. Identification and characterization of expressed retrotransposons in the genome of the Paracoccidioides species complex. BMC Genomics. 2015;16:376.
41. Caza M, Kronstad JW. Shared and distinct mechanisms of iron acquisition by bacterial and fungal pathogens of humans. Front Cell Infect Microbiol. 2013;3:80.
42. Tristão GB, Assunção LP, Dos Santos LP, Borges CL, Silva-Bailão MG, Soares CM, Cavallaro G, Bailão AM. Predicting copper-, iron-, and zinc-binding proteins in pathogenic species of the *Paracoccidioides* genus. Front Microbiol. 2014;5:761.
43. Volpe TA, Kidner C, Hall IM, Teng G, Grewal SI, Martienssen RA. Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. Science. 2002;297:1833–7.
44. Lippman Z, May B, Yordan C, Singer T, Martienssen R. Distinct mechanisms determine transposon inheritance and methylation via small interfering RNA and histone modification. PLoS Biol. 2003;1:E67.
45. Saha A, Mitchell JA, Nishida Y, Hildreth JE, Ariberre JA, Gilbert WV, Garfinkel DJ. A trans-dominant form of Gag restricts Ty1 retrotransposition and mediates copy number control. J Virol. 2015;89:3922–38.
46. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007;35:W265–8.
47. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics. 2008;9:18.
48. McCarthy EM, McDonald JF. LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics. 2003;19:362–7.
49. Kalyanaraman A, Aluru S. Efficient algorithms and software for detection of full-length LTR retrotransposons. J Bioinform Comput Biol. 2006;4:197–216.
50. Lee H, Lee M, Mohammed Ismail W, Rho M, Fox GC, Oh S, Tang H. MGEScan: a Galaxy-based system for identifying retrotransposons in genomes. Bioinformatics. 2016;32:2502–4.
51. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
52. Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA. Database indexing for production MegaBLAST searches. Bioinformatics. 2008;24:1757–64.
53. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, et al. CDD: NCBI's conserved domain database. Nucleic Acids Res. 2015;43:D222–6.
54. Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004;14:1394–403.
55. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
56. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 2012;9:772.
57. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001;17:754–5.
58. Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, Eddy SR, Gardner PP, Bateman A. Rfam 11.0: 10 years of RNA families. Nucleic Acids Res. 2013;41:D226–32.
59. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 2012;9:357–59.
60. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36.
61. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14:178–92.
62. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
63. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit


BioMed Central

3.2 Capítulo II: Transcriptional analysis allows genome reannotation and reveals that *Cryptococcus gattii* VGII undergoes nutrient restriction during infection.

Este capítulo apresenta o artigo publicado na revista *Microorganisms* em 23 de agosto de 2017 (Ferrareze et al., 2017).

microorganisms

MDPI

Article

Transcriptional Analysis Allows Genome Reannotation and Reveals that *Cryptococcus gattii* VGII Undergoes Nutrient Restriction during Infection

Patrícia Aline Gröhs Ferrareze [1,†], Rodrigo Silva Araujo Streit [2,†], Patricia Ribeiro dos Santos [1,†], Francine Melise dos Santos [1], Rita Maria Cunha de Almeida [3], Augusto Schrank [1], Livia Kmetzsch [1], Marilene Henning Vainstein [1,*] and Charley Christian Staats [1,*]

1 Programa de Pós-Graduação em Biologia Celular e Molecular, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre 91501970, Brazil; p.ferrareze@outlook.com (P.A.G.F.); patriciaribsant@yahoo.com.br (P.R.d.S.); francinemelise@hotmail.com (F.M.d.S.); aschrank@cbiot.ufrgs.br (A.S.); liviak@cbiot.ufrgs.br (L.K.)

2 Departamento de Biologia Molecular e Biotecnologia, Instituto de Biociências, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre 91501970, Brazil; rodrigo.streit@hotmail.com

3 Instituto de Física, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre 91501970, Brazil; rita@if.ufrgs.br

* Correspondence: mhv@cbiot.ufrgs.br (M.H.V.); staats@cbiot.ufrgs.br (C.C.S.); Tel.: +55-51-3308-6060 (M.H.V.); +55-51-3308-6080 (C.C.S.)

† These authors contributed equally to this work.

Received: 24 July 2017; Accepted: 18 August 2017; Published: 23 August 2017

Abstract: *Cryptococcus gattii* is a human and animal pathogen that infects healthy hosts and caused the Pacific Northwest outbreak of cryptococcosis. The inhalation of infectious propagules can lead to internalization of cryptococcal cells by alveolar macrophages, a niche in which *C. gattii* cells can survive and proliferate. Although the nutrient composition of macrophages is relatively unknown, the high induction of amino acid transporter genes inside the phagosome indicates a preference for amino acid uptake instead of synthesis. However, the presence of countable errors in the R265 genome annotation indicates significant inhibition of transcriptomic analysis in this hypervirulent strain. Thus, we analyzed RNA-Seq data from in vivo and in vitro cultures of *C. gattii* R265 to perform the reannotation of the genome. In addition, based on in vivo transcriptomic data, we identified highly expressed genes and pathways of amino acid metabolism that would enable *C. gattii* to survive and proliferate in vivo. Importantly, we identified high expression in three APC amino acid transporters as well as the GABA permease. The use of amino acids as carbon and nitrogen sources, releasing ammonium and generating carbohydrate metabolism intermediaries, also explains the high expression of components of several degradative pathways, since glucose starvation is an important host defense mechanism.

Keywords: *Cryptococcus gattii*; R265; transcriptome; genome; annotation; amino acid; bronchoalveolar lavage; cryptococcosis

1. Introduction

Cryptococcus gattii, along with its sibling species *Cryptococcus neoformans*, are the etiological agents of cryptococcosis, a life-threatening disease that particularly affects the lungs and central nervous system. Cryptococcosis is a major health problem on the African continent and other tropical regions, with an extremely high fatality ratio among immunocompromised patients [1] and is, therefore, considered one of the most prevalent fatal fungal diseases worldwide [2]. Although accountable for

20% of cryptococcosis cases [3], *C. gattii* is also capable of infecting immunocompetent patients [4]. In addition, the identification of hypervirulent strains in the outbreaks on Vancouver Island [5] and in the United States [6] has driven efforts to identify the genetic and molecular differences responsible for the increased virulence of *C. gattii* [4].

Cryptococcosis is initiated by the inhalation of infectious propagules (as spores and dried yeast cells). After reaching the alveolar parenchyma, yeast cells are typically phagocytosed by macrophages, within which *C. gattii* can survive and replicate [7]. In addition, cryptococcal cells are considered a facultative intracellular pathogen of neutrophils and dendritic cells, defense cells which act during the early stages of murine pulmonary infection [8]. These innate immune cells have antigen-presenting activity and produce proinflammatory cytokines, inducing an adaptive immune response in the host [8,9]. The capability of *Cryptococcus* cells to avoid phagocytosis, as well as to survive and replicate at a high rate inside phagocytes, contributes to the dissemination of fungal cells from the lung via the bloodstream [10,11]. Among all the virulence factors of *Cryptococcus*, the most important element for preventing phagocytosis is the capsule, as its immunosuppressive functions are directly involved in cryptococcal survival within phagocytic cells. Internalized cryptococcal cells can inhibit the acidification of phagosomes in macrophages, leading to phagosome-lysosome fusion membrane disruption [10,12]. Phagolysosomal membrane damage, in turn, enables *Cryptococcus* to escape the phagolysosome, thus accessing the nutrients in the cytoplasm [11].

The nutritional immunity concept refers to an innate immune activity characterized by limitation by the host of essential nutrients in order to inhibit pathogen survival and proliferation. In this context, metals such as iron, copper, and zinc, which are necessary cofactors for at least one-third of proteins, constitute an important target of this mechanism [13,14]. To adapt to the stringent conditions imposed by the host, *C. neoformans* triggers the expression of membrane transporters for hexoses, amino acids, fatty acids, metals, ammonium, and phosphate, as well as stress response genes [15,16]. Additionally, alternative carbon sources such as acetate, lactate, fatty acids, or amino acids are used as primary carbon sources by cryptococcal cells during interaction with host cells [15].

Thus, considering the nutrient limitation imposed on *C. neoformans* cells in the lungs and that the *C. gattii* VGII representative strain primarily causes pulmonary disease [17], we performed transcriptional analysis of *C. gattii* R265 strain, recovered from bronchoalveolar lavage of infected mice. Functional mapping of expressed genes revealed that several genes involved in the uptake and metabolism of nutrients displayed high expression, suggestive of adaptation to a limited availability of carbon and nitrogen sources.

2. Materials and Methods

2.1. Ethics Statement

The use of animals in this work was approved by the Universidade Federal do Rio Grande do Sul Ethics Committee for Use of Animals (CEUA—protocol number 18807, approved on 10 May 2011). Mice were housed in groups of four in filtered top-ventilated cages, with a 12 h dark/light cycle and food and water ad libitum. The animals were cared for according to the guidelines of the Brazilian National Council for Animal Experimentation Control (CONSEA) and the Brazilian College of Animal Experimentation (COBEA). All efforts were made to minimize animal suffering.

2.2. RNA-seq Data

For bronchoalveolar lavage (BAL), 30 BALB/c mice were anesthetized with 100 mg/kg ketamine and 16 mg/kg xylazine in phosphate-buffered saline (PBS; NaCl 137 mM; KCl 2.7 mM; Na_2HPO_4 10 mM; KH_2PO_4 8 mM; pH 7.4) and nasally infected with 1×10^6 *C. gattii* R265 cells in PBS. After 24 h, bronchoalveolar lavage was performed via tracheal cannula using three consecutive lavages of 1 mL of PBS. Following centrifugation, the recovered cells were merged in a single biological sample and lysed by washing with cold ultrapure H_2O. RNA extraction from recovered yeast cells was performed

using glass beads and an RNeasy Mini kit according to the manufacturer instructions (Qiagen, Hilden, Germany). RNA integrity and concentration were assessed by electrophoresis on a 1% agarose gel and by fluorometric analysis using a Qubit fluorometer and a Quant-iT RNA assay kit according to the manufacturer's instructions (Invitrogen, San Diego, CA, USA). mRNA was purified from total RNA, processed, and a single-end sequencing was performed using Solexa technology on an Illumina Genome Analyzer GAII (Fasteris Life Sciences SA, Plan-les-Ouates, Switzerland). After quality analysis employing FastQC [18], low quality bases were filtered using FastX-Toolkit [19].

The *C. gattii* R265 libraries of WT and *ZAP1* null mutant (both grown in zinc deprivation conditions) used in this study were obtained from Schneider and colleagues [20], available at the NCBI SRA databases under accession codes SRX2523180 and SRX2522699, respectively.

2.3. Reads Alignment and Gene Prediction Refinement

The genome and transcript annotation of the R265 strain was downloaded from the Broad Institute archive [21] in August 2015. The reads were aligned against the genomic sequence using Tophat v2.0.13 software [22], with the following settings: minimum intron length, 10; maximum intron length, 5000; minimum segment intron, 10; and maximum segment intron, 5000. The alignment file was uploaded to Cufflinks v2.2.1 software [23] to generate the initial gene predictions, with the following settings: minimum intron length, 10; maximum intron length, 500; overlap radius, 10; and minimum isoform fraction, 0.4. Next, we performed CodingQuarry [24] predictions using the alignment file from Tophat and the gene predictions of Cufflinks to produce the preliminary transcriptome annotation. The alignment file, the new transcriptome annotation, and the transcriptome annotation from the Broad Institute were then uploaded to the Integrative Genomic Viewer software [25] for manual revision of both annotations. Exon/intron boundary predictions with less than 30 reads aligned, but those predicted in the previous annotation retained their structure as in the previous annotation. Exon/intron boundaries with no read support and no previous annotation were excluded. In addition, the untranslated regions (UTRs) were manually added for all predicted gene models based on the alignment. A fluxogram describing the gene prediction strategy is presented in Figure 1.

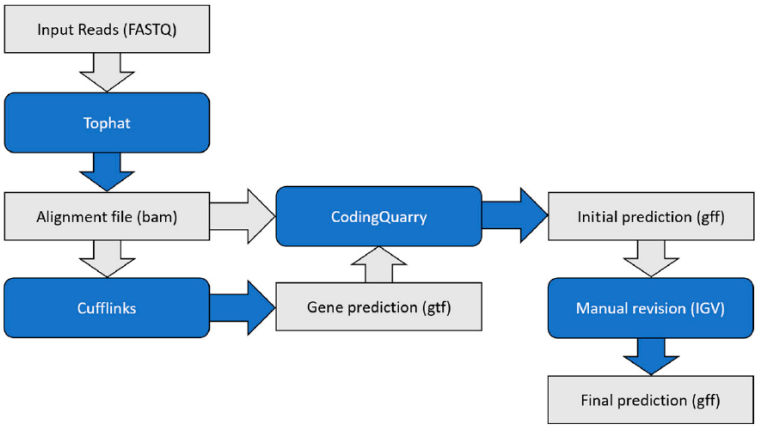

Figure 1. Gene prediction workflow. Reads were first aligned against the genome sequence using Tophat software. The alignment file (BAM) generated was then used to predict gene models (gtf) using Cufflinks software. Finally, both the alignment file and the gene models were loaded to CodingQuarry to generate the initial gene models, which are manually revised to generate a final prediction (gff).

2.4. RNA Isolation and RT-PCR

C. gattii strain R265 cells were incubated in yeast peptone dextrose (YPD) media overnight at 200 rpm and 30 °C. Cells were then centrifuged (5000× *g* for 5 min) and washed in PBS. The cell pellet was suspended in 20 mL of yeast nitrogen base (YNB) and diluted to

1×10^6 cells/mL. The cells were inoculated in 100 mL of YNB plus 10 μM of zinc chelator *N*,*N*,*N*′,*N*′-tetrakis(2-pyridylmethyl)ethane-1,2-diamine (TPEN) and incubated for 4 h at 200 rpm and 30 °C. The cells were collected by centrifugation (5000× *g* for 5 min), frozen in liquid nitrogen, and placed in an ultrafreezer (−80 °C) for lyophilization.

RNA isolation was performed using Trizol (Invitrogen–Life Technologies, Carlsbad, CA, USA) after cellular lysis via mortar and pestle. RNA integrity and quantification were assessed by electrophoresis on a 0.8% agarose gel and by spectrophotometry on a NanoDrop 2000 (Thermo Fisher Scientific, Wilmington, DE, USA). cDNA was prepared from DNAse-treated total RNA samples (1 μg) using Improm II Reverse Transcriptase (Promega, Madison, WI, USA) and oligo-dT. PCR was performed on a ProFLex PCR system (Applied Biosystems–Life Technologies, Carlsbad, CA, USA) with the following thermal cycling conditions: an initial step of 94 °C for 5 min followed by 30 cycles of 94 °C for 15 s, 55 °C for 15 s, and 72 °C for 60 s. All PCR products were subjected to agarose gel electrophoresis and visualized by staining with ethidium bromide.

2.5. BAL Expression Analysis

For expression analysis, the RNA-Seq library of *C. gattii* R265 from murine bronchoalveolar lavage was aligned against the R265 genome (NCBI accession code GCA_000149475.3) using Tophat [22] and the unique alignment option (max-multihits = 1). The expression values were measured by the fragments per kilobase per million (FPKM) using Cufflinks [23] and the proposed *C. gattii* R265 genome annotation.

2.6. Functinal Enrichment Analysis

For the detection of enriched functional KEGG and Gene Ontology terms, we selected genes classified as the most abundant according to the FPKM value distribution. Genes were collected in the FungiDB server [26], their orthologs from *C. gattii* WM276 identified in the same platform, and gene set enrichment analysis was conducted for the Biological Process and KEGG pathways. Only terms containing Benjamini-corrected *p*-values ≤ 0.05 were considered to be statistically enriched.

2.7. Transcriptogram

The transcriptogram was generated using the Transcriptogramer [27] program. For gene ordination, we used the STRING protein interaction data from *C. gattii* WM276 [28] with score ≥ 0.800. Expression analysis was performed using BAL FPKM expression data from Cufflinks. The enrichment analysis was evaluated using KEGG data for pathways and genes of *C. gattii* WM276 [29]. The conversion of WM276 genes to R265 was performed by a Blastp [30] orthologous search, with the best bidirectional hit model.

2.8. KEGG Pathway Mapping

For gene expression visualization within the phenylalanine, tyrosine, and tryptophan biosynthetic pathway, the Pathview server [31] was used. For KEGG pathway mapping, we used the WM276 gene names to R265 orthologs as well as the BAL expression values. The Pathview parameters were adjusted to a FPKM limit of 1000.

3. Results

3.1. Refinement of C. gattii R265 Genome Annotation

In order to understand the transcriptional profiling of *C. gattii* VGII on the site of infection, we performed RNA-Seq experiments to identify the most abundant transcripts. Purified mRNA samples from cryptococcal cells recovered from BAL fluid of BALB/c mice 24 h post infection were pooled and sequenced using the Solexa technology. After filtering low quality reads and/or bases, we obtained a total of 43,094,699 reads with a size of 100 nt. Our first attempt to align reads originating

from the BAL library revealed that a consistent proportion of reads (81.41%) could not be aligned. In addition, we noted that only 65.04% of the reads aligned to genes, and 18.47% of the reads aligned to introns and intergenic regions. This led us to speculate that the current genomic annotation of *C. gattii* might contains errors, confirming our previous findings for the ZAP1 gene [20]. Therefore, we generated a new annotation based on transcriptomic evidence. Using our RNA-Seq dataset from three different growth conditions (*C. gattii* WT and ZIP1 null mutant exposed to a low zinc environment for 2 h, as well as cryptococcal cells recovered from BAL), we executed the CodingQuarry pipeline to generate new gene models for the R265 strain of *C. gattii*, which consisted of three steps (Figure 1). In the first step, all reads were aligned together against the genome of the R265 strain using Tophat software. In the second step, the alignment was loaded to Cufflinks software for an initial prediction. In the last step, both the predicted gene models of Cufflinks and the alignment from Tophat were loaded to the CodingQuarry software to obtain the refined gene models. Using this approach, we obtained 6956 protein coding gene models, which were then manually revised using IGV software.

During the manual revision, 509 gene models were excluded from the annotation due to a lack of alignment support or because their *open reading frames* did not show homology to any other fungal genes. Furthermore, 569 gene models predicted by the previous annotation but with a low count of reads spanning the intron-exon boundaries (less than 30 reads) had their previous annotation's structure retained, as it was not possible to assure the accuracy of both prediction and revision. Finally, information on the 5′-UTR and 3′-UTR of the revised genes was manually added, as the prediction tools could not perform this step. After the revision, a total of 6411 gene models were proposed as the new transcriptome annotation (GFF File S1).

From the 6428 gene models comprising the annotation provided by the BROAD institute, we verified that only 3187 had their structure supported by the alignment data, meaning that more than a third of the gene models presented some type of error. Of these, 2563 genes contained misplaced intron-exon boundaries, sometimes leading the automated prediction program to exclude large extensions of the transcribed region in order to maintain the ORF that was initially predicted (Figure S1). In addition to the structural errors, 55 genes were excluded from the previous annotation. Two of these were excluded due to the identification of pseudogenes. The remaining 53 were excluded as a result of gene model fusion that incorrectly predicted separation in the last annotation. Thirty-eight potential new genes were also added, three of which were derived from the split of genes in the previous annotation that showed homology with genes from other strains of *C. gattii* and *C. neoformans*. In addition, we identified 257 punctual errors on the genome annotation of the strain R265 within the transcribed regions, which were associated with missing nucleotides, misplaced nucleotides, and wrong nucleotides (FASTA File S1).

3.2. Experimental Validation of the Gene Models

In order to confirm the new annotation, we selected examples from the three major types of changes in the annotation: (i) genes that presented incorrectly identified intron-exon boundaries; (ii) genes that encompassed two split genes; and (iii) genes that encompassed one fused gene. Using the RNA-Seq data from the zinc limitation condition, we selected the most abundant gene in each group for confirmation, with the exception of group (iii), in which both genes comprising the gene with the highest expression in the previous annotation were selected. The genes selected for experimental validation were (i) CNBG_3432, (ii) CNBG_5018, and (iii) CNBG_0818 and CNBG_9683. RNA was isolated from cells cultured under the zinc limitation condition (YNB + 10 μM TPEN) for 2 h, and RT-PCR was used to confirm the new gene models (Figure 2).

3.3. Transcriptional Profiling of C. gattii Recovered from Murine BAL

In order to determine RNA abundance in *C. gattii* cells at the site of infection, we recovered cryptococcal cells from mouse lung 24 h after infection. RNA was isolated from fungal cells and the sequences determined using Solexa technology. Measurements of transcript abundance using

FPKM revealed that almost all genes could be detected. From the 6411 genes predicted in the *C. gattii* R265 genome, a total of 6257 presented FPKM values ≥ 1.0. However, the transcription levels of all genes ranged from 0 to 89,303.4. In addition, FPKM value distribution in quartiles revealed that 50% of transcripts had expression between 10 and 100 (quartiles 2 and 3), while the expression of the remaining genes (quartiles 1 and 4) ranged to two or three orders of magnitude, respectively (Figure 3 and Table S1).

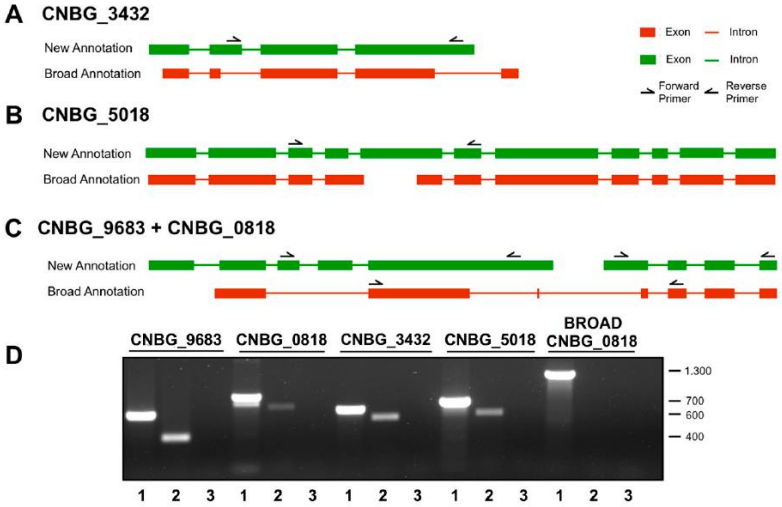

Figure 2. Experimental confirmation of the gene models. (**A**–**C**) Four genes were selected to confirm the changes in annotation provided by the RNA-Seq validation. Additionally, a primer pair designed to specifically amplify the predicted Broad annotation of the gene CNBG_0818 was design to confirm that it is in fact two genes. (**D**) The selected genes were confirmed by RT-PCR. For each gene, one genomic DNA (1) template, one cDNA template (2), and a negative control (3) were evaluated. All but Broad CNBG_0818 were amplified, indicating that the new annotation was correct. Values on the left refer to the band sizes (bp) of the molecular size marker. The green/red rectangles represent the exons, and the green/red lines between them are the introns; according the new and old (Broad) annotations, respectively.

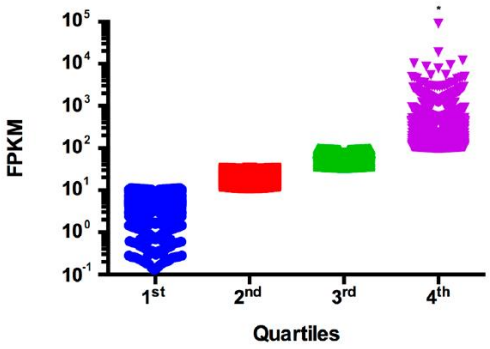

Figure 3. Transcriptional landscape of *C. gattii* R265 genes during pulmonary infection. FPKM values were determined from reads aligned to the *C. gattii* R265 genome using Cufflinks. FPKM values were distributed into quartiles and statistically analyzed using ANOVA followed by Tukey's multiple comparisons test. The quartiles 1 (blue), 2 (red) and 3 (green) do not present statistical difference. The fourth quartile (purple) concentrates the abundant transcripts and has significant statistical difference (*) compared with others.

3.4. Functional Profiling of C. gattii Transcriptome Recovered from Murine BAL

In order to functionally profile the most abundant genes of *C. gattii* recovered from BAL, we performed Gene Ontology enrichment using FungiDB [26] version 32. As the Gene Ontology annotation for *C. gattii* R265 is incomplete, we collected the IDs of *C. gattii* WM276 orthologs to most prevalent transcripts in the *C. gattii* R265 BAL library. The enriched terms refer mainly to processes involved in primary metabolism, such as those associated with synthesis and uptake of fundamental metabolites for cell survival (Tables 1 and S2).

Table 1. Gene ontology enrichment analysis of the most abundantly expressed genes in *C. gattii* recovered from BAL. The top ten enriched Biological Process terms are shown. Only Benjamini-corrected *p*-values < 0.5 were considered as significantly enriched.

ID	Name	Benjamini Corrected *p*-Value
GO:1901566	Organonitrogen compound biosynthetic process	8.59×10^{-11}
GO:1901564	Organonitrogen compound metabolic process	3.73×10^{-10}
GO:0006412	Translation	9.05×10^{-10}
GO:0043043	Peptide biosynthetic process	1.53×10^{-9}
GO:0043604	Amide biosynthetic process	2.40×10^{-9}
GO:0006518	Peptide metabolic process	2.65×10^{-9}
GO:0043603	Cellular amide metabolic process	6.16×10^{-9}
GO:0044271	Cellular nitrogen compound biosynthetic process	4.24×10^{-6}
GO:0055114	Oxidation-reduction process	5.89×10^{-6}
GO:0008152	Metabolic process	6.56×10^{-6}

We expanded our functional profiling analysis using a transcriptogram [27], a tool that can measure the relative abundance of a given process for which a set of transcripts is assigned. The Transcriptogramer was used to take a snapshot of the metabolic pathways expressed during *C. gattii* infection. As the String and KEGG data were from *C. gattii* WM276, we collected information on orthologs from BLASTp analysis and generated an annotation file for *C. gattii* gene products based on their orthology with *C. gattii* WM276 gene products. The String networks provided information on protein interaction for 3076 genes (interaction score $\geq$ 0.800). We then obtained KEGG information for 2813 out of these 3076 genes for enrichment analysis (Figure 4A). As processes related to information processing (translation and ribosome biogenesis) displayed high expression values in the transcriptogram and this hinders the observation of enriched pathways, we determined the median of expression values from the genes associated to the remaining pathways. We then focused on the pathways whose associated gene expression was higher than the median (Figure 4B).

3.5. Virulence Gene Expression in C. gattii Recovered from Murine BAL

In order to evaluate the expression of virulence genes, we queried the pathogen–host interactions database (PHI-base) [32] using proteins coded by genes whose transcripts were defined as the most abundant in *C. gattii* recovered from BAL (FPKM $\geq$ 98.96). A total of 85 genes whose expression was classified as highly abundant were assigned as virulence factors based on this analysis (Table S3). The most abundant virulence factors were related to an extracellular elastinolytic metalloproteinase (CNBG_6001; FPKM = 4679.19) and a thiol-specific antioxidant protein 1 (CNBG_2132; FPKM 2709.54).

3.6. Nutrient Uptake Gene Expression in C. gattii Recovered from Murine BAL

A growing body of evidence suggests that cryptococcal cells undergo nutritional limitations in infection conditions and must, therefore, express nutrient uptake transporters on the cell surface in order to survive the harsh environment of the infection milieu [16,33–35]. Therefore, we determined the presence of transporter-coding transcripts among the most abundant genes. According to *C. gattii* R265 genome annotation, a total of 336 genes have the term "transporter" associated with at least one

field in their annotation. Of these, 88 genes displayed FPKM values above 98.96, placing them as highly expressed genes (Table S4). The two most abundant transporters refer to a ribonucleotide transporter (ADP, ATP carrier protein; CNBG_4811; FPKM = 9360.59), and a glucose transporter (CNBG_0170; FPKM = 4905.2). However, transporters for other sugars, metals, and other compounds were found, suggesting that *C. gattii* responds to nutrient deprivation in infection conditions.

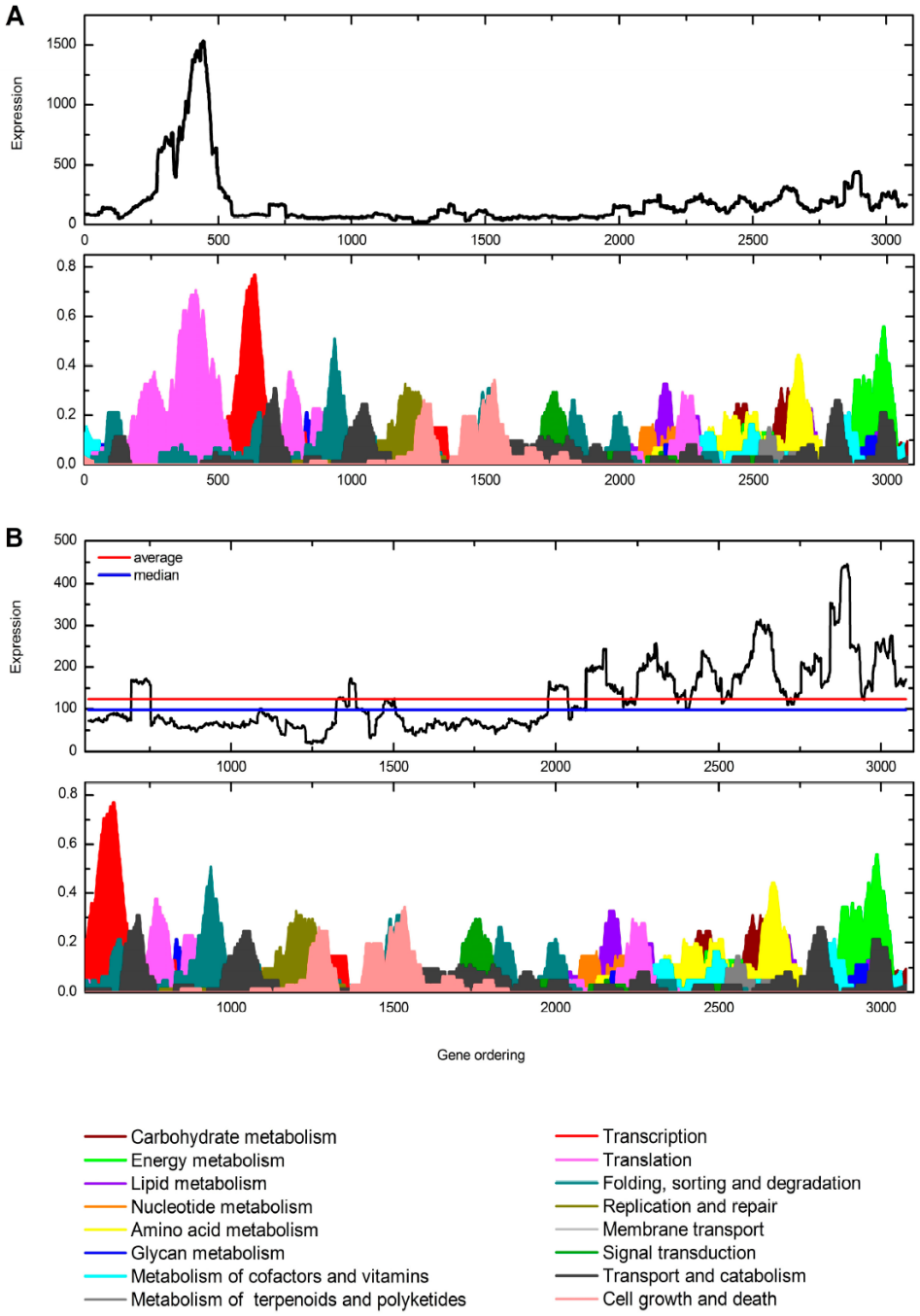

Figure 4. Transcriptogram of BAL RNA-Seq data of *C. gattii* R265. The upper graph shows the expression values (y) of ordered genes (x). The bottom graph indicates the KEGG pathways related to the ordered genes. (**A**) Complete transcriptogram; and (**B**) transcriptogram with the median and average expression values.

3.7. Nitrogen and Amino Acid Metabolism

Based on gene ontology enrichment analysis and transcriptogram profiling, we noted that amino acid metabolic pathways were enriched in cryptococcal cells recovered from BAL. This was confirmed by KEGG pathway enrichment analysis (Table S5), in which the following amino acid metabolic pathways appeared as enriched: cysteine and methionine metabolism (ec00270); glycine, serine, and threonine metabolism (ec00260); phenylalanine, tyrosine, and tryptophan biosynthesis (ec00400); alanine, aspartate, and glutamate metabolism (ec00250); valine, leucine, and isoleucine biosynthesis (ec00290); and valine, leucine, and isoleucine degradation (ec00280). This suggests that for some amino acids, the host imposes a nutrient deficient condition. Thus, we evaluated the expression of each gene associated with amino acid metabolic pathways to infer the transport and metabolism of nitrogen and amino acids. Several amino acids and organic and inorganic nitrogen uptake transporters were detected with high expression values (FPKM $\geq$ 98.96; Table 2).

Table 2. FPKM values of ammonium permeases, amino acid transporters, and GATA transcription factors.

GENE	DESCRIPTION	FPKM BAL
CNBG_0332	Ammonium permease 1 (AMT1)	415.458
CNBG_6023	Ammonium permease 2 (AMT2)	40.3613
CNBG_1602	Gamma-aminobutyric acid transporter	677.554
CNBG_3901	Gamma-aminobutyric acid transporter	268.427
CNBG_4571	Gamma-aminobutyric acid transporter	61.0116
CNBG_4665	Gamma-aminobutyric acid transporter	4.20125
CNBG_4156	Choline transporter	74.1659
CNBG_5513	L-methione transporter	11.9407
CNBG_4785	General amino acid transporter (AAP2)	441.806
CNBG_1371	General amino acid transporter (AAP4)	416.72
CNBG_9416	General amino acid transporter (AAP1)	363.575
CNBG_6051	General amino acid transporter (AAP6)	32.0951
CNBG_1350	Gamma-aminobutyric acid transporter (AAP8)	18.3341
CNBG_2012	Neutral amino acid permease	258.968
CNBG_1852	Neutral amino acid permease	194.189
CNBG_2927	Ure2p	21.9955
CNBG_4137	Bwc2	148.2437
CNBG_9614	Cir1	6.505
CNBG_0368	Gat1	29.7171
CNBG_3885	Gat201	340.642

In addition, we found that several genes related to amino acid metabolism were present in the most abundant transcripts in *C. gattii* recovered from BAL (Table S5). We focused on phenylalanine, tyrosine, and tryptophan biosynthesis genes, as these appear enriched in KEGG enrichment analysis. According to the KEGG database, 60 genes present in *C. gattii* WM276 could be assigned to this pathway. From these, a total of 56 orthologs and two lineage-specific genes were found in *C. gattii* R265 according to the FungiDB. The expression of these genes was detected in BAL (Figure S2). Despite some genes presenting low FPKM values, the majority (62%) had expression values above the median of all transcript values (FPKM = 34.22), suggesting that aromatic amino acids are likely synthesized during cryptococcal infection of lungs.

We also investigated valine, leucine, and isoleucine metabolism, as both the biosynthesis and degradation of these amino acids appears as enriched among the most abundant transcripts in *C. gattii* cells recovered from BAL. For the biosynthesis pathway, 15 genes from *C. gattii* WM276 and *C. gattii* R265 were assigned based on the EC codes. For the degradation pathway, 44 genes present in *C. gattii* WM276 were assigned to this pathway. All the genes related to these pathways had their transcripts detected, suggesting that the corresponding amino acids are low in concentration

during cryptococcal infection. However, as the mean FPKM values for valine, leucine, and isoleucine biosynthesis-associated genes were slight lower than those observed for their degradation (Figure 5), it is possible that a continuous process of synthesis and degradation takes place to generate acetyl-CoA during host adaptation as all of these pathways result in acetyl-CoA production.

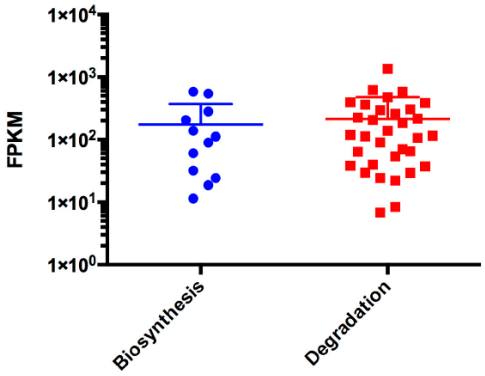

Figure 5. Transcriptional landscape of *C. gattii* R265 genes associated to valine, leucine, and isoleucine metabolism during pulmonary infection. FPKM values were determined from reads aligned to *C. gattii* R265 genome using Cufflinks software. Genes associated with biosynthesis (blues dots) and degradation (red squares) processes were recovered from The FungiDB.

The Gat1 transcription factor (CNBG_0368) is a specific nitrogen starvation regulator, and is expressed in BAL, despite in low levels (FPKM = 29.71). For means of comparisons, we determined the expression values of the *C. gattii* VGII orthologs (Table S6) to the *C. neoformans* transcription factors [36]. The median value of expression of *C. gattii* transcription factors is 20.25, which FPKM values ranging from 0 to 552.08 (Table S6). Along with the ammonium permease expression values, this may indicate that this condition is not subject to nitrogen starvation. Therefore, the high expression of GABA permease (CNBG_1602) indicates ammonium starvation and the utilization of secondary nitrogen sources, according to the expression of Ure2p. Moreover, *C. gattii* R265 showed high expression of some amino acid transporters (Table 2), suggesting amino acid uptake during host-pathogen interaction.

Taken together, the results suggest intense modulation of amino acid metabolism during the *C. gattii* R265 infection process.

4. Discussion

The task of describing the complete set of transcripts that an organism expresses can be extremely challenging and laborious. In some studies in which RNA-Seq is used to improve the transcript annotation, the corrections are performed manually, using the alignment as a guide [37]. Although this approach can improve the gene models and identify new genes, it demands a significant effort, especially where there are thousands of genes with incorrect structures. Therefore, we generated an entirely new set of transcripts using an alignment-based prediction tool, CodingQuarry. This program was designed to deal with the singularities of the fungal genome, such as the reduced intron size and the proximity of the genes, where other tools, such as Cufflinks, fail [24,37,38]. Although there was still a need for manual revision as not all the predicted models were correct, it was clear that a new prediction was more time-efficient than manually correcting thousands of genes.

With the final RNA-Seq based prediction, we corrected more than half of the transcripts predicted by the previous annotation and also identified 35 new genes. This result is similar to that obtained with transcriptome reannotation of the strain H99 of *C. neoformans* using RNA-Seq, in which almost 40% of the transcripts were changed and 55 new genes were found [39]. In contrast, other fungal species as *Candida albicans* and *Fusarium graminearum* showed less improvement with RNA-Seq-based correction [37,40]. This difference could be explained by the homology basis of automated annotation

as, being basidiomycetes, *C. gattii* and *C. neoformans* are evolutionally more distant from the other well-described fungi. Therefore, we expect that the new transcript annotation of the *C. gattii* VGII R265 strain may allow the improvement of the annotation of other *C. gattii* strains. In addition, such gene models can be used by the cryptococcal research community. Public available platform, as the FungiDB, can be fed with the annotation of gene models and with transcriptome data to provide researchers fundamental information about expression of cryptococcal genes during the infection process. However, we are aware that we cannot refine the whole transcriptome. In this way, further experiments to profile the transcriptome would strengthen the reannotation and provide additional data for the correction of misannotated genes.

The new annotation, therefore, enabled new transcriptome analysis of *C. gattii* R265. We therefore evaluated some enriched pathways in these fungal cells obtained by bronchoalveolar lavage in a murine model of cryptococcosis. Although the nutrient composition of the phagosome is poorly defined, it is well accepted that it is a nutrient-poor environment [41]. Thus, several studies in *C. neoformans* have shown that such cells develop a strategy to cope with the scarcity of nutrients, since *S. cerevisiae* and *C. albicans* show that few, or no, amino acid biosynthesis genes are upregulated [15,42,43].

Nitrogen catabolite repression (NCR) is a mechanism that controls the utilization of optimal nitrogen sources in *Cryptococcus* species. In this context, NCR-regulated genes are repressed to ensure that secondary nitrogen source degrading pathways are not expressed when preferred nitrogen sources are available [44]. Thus, during nitrogen starvation, the expression of permeases and catabolic enzymes is activated by a specific GATA-factor family of transcription factors [45]. *S. cerevisiae* has two GATA factors involved in nitrogen metabolism, Gat1 and Gln3 [45]; *C. gattii* R265 has a Gat1 ortholog, the gene CNBG_0368 [46]. The utilization of amino acids and other nitrogenous sources requires their internalization by membrane permeases.

Ammonium uptake in *C. gattii* R265 is mediated by the low- and high-affinity ammonium permeases Amt1 and Amt2. Interestingly, the FPKM values of Amt2, which is transcriptionally induced in response to ammonium-limiting levels [47], does not indicate nitrogen deprivation under BAL conditions. However, Amt2 is also induced by low levels of alternative nitrogen sources [47], which might indicate that non-preferred nitrogen compounds, likely acquired by amino acid uptake and degradation, are supporting the minimum nitrogen requirements of the cell and avoiding Amt2 overexpression. Similarly, Ure2p (CNBG_2927) levels are low. In *S. cerevisiae*, the high intracellular concentration of Ure2p indicates nitrogen excess, and its inactivation by nitrogen limitation leads to NCR de-repression and GATA factor activation [48].

Another gene, the gamma-aminobutyric acid transporter, showed high expression in BAL. As observed by Luzzani and colleagues [49], GABA can induce UGA4 expression when cells are grown in nitrogen-poor conditions, but not when they are grown with ammonium, since GABA is used as a poor nitrogen source by *S. cerevisiae* [50] in an NCR-dependent mechanism.

C. neoformans expresses 10 genes encoding cytoplasm amino acid permease genes of the APC (amino acid–polyamine-choline transporter) superfamily: eight encode global permeases (AAP1 to AAP8) and two sulfur amino acid permeases (Mup1 and Mup3) with a high and low affinity for methionine and cysteine. The small number of amino acid permeases encoded by *Cryptococcus* genomes can be related to the low enzyme-substrate affinity [51]. Among the permeases, AAP3, AAP5, and AAP7 do not have true orthologues in R265. AAP6 (CNBG_6051), which displayed no transcriptional change according to nitrogen source (ammonium sulfate or amino acids) in *C. neoformans* studies [51], presented low levels. In *C. neoformans*, the deletion of AAP4 and AAP5 genes result in the highest impact on growth, indicating that AAP4 and AAP5 are highly redundant and essential for amino acid uptake, especially at 37 °C. Moreover, it was demonstrated that AAP4/AAP5 amino acid permeases are required as a virulence factor, since they participate in capsule production and stress resistance, and the aap4Δ/aap5Δ double mutant is avirulent in mouse and *Galleria mellonella* models. Despite the redundancy of AAP4/AAP5, *C. gattii* R265 appears to have just one ortholog to

AA4 permease, CNBG_1371; this is consistent with the observation that just one is necessary for the thermal and oxidative stress response [51].

According to our analysis, some pathways related to amino acid biosynthesis were predicted to be enriched based on the FPKM of related genes from cryptococcal cells recovered from BAL. However, according to KEGG mapping, the enzymes involved in ammonium release by glutamate degradation are more abundant, suggesting that the in vivo condition prioritizes amino acid acquisition for degradation instead of synthesis. In addition, the production of acetyl-CoA from valine, leucine, isoleucine, and fatty acid degradation pathways, as well as from pyruvate and acetate by *Cryptococcus* during infection, is essential for the synthesis of chitin in the cell wall and O-acetylation of the capsule [52]. Similarly, all enriched degradative pathways that result in oxaloacetate formation contribute to gluconeogenesis [53].

The analysis of metabolites following co-incubation of *C. neoformans* with the lung epithelial cells of Liew and colleagues [52] showed that some compounds such as L-cysteine, lactic acid, pantothenic acid, fumaric acid, L-tyrosine, D-fructose, DL-3-phenyllactic acid, and 3-hydroxyisovaleric acid were potentially secreted by *C. neoformans* in supernatant culture media. In our study, we observed the enrichment of cysteine metabolism pathway in BAL, which could potentially explain the metabolic secretion during co-incubation.

In summary, we have generated a new annotation for the R265 strain of *C. gattii*, significantly improving the previous automated annotation. As this is the first RNA-Seq-based transcriptome annotation of a *C. gattii* strain, we believe that this annotation represents a valuable resource for the research community and will help to improve the annotation of other strains of *C. gattii*. Furthermore, all the results obtained in this work appear to correlate with amino acid uptake promoted by *Cryptococcus* cells in vivo. Furthermore, the accented degradation profiles confirm acquisition and degradation instead of the biosynthesis of nitrogen sources. The use of amino acids as carbon sources, generating carbohydrate metabolism intermediaries, also explains the high expression of many degradative pathways, since glucose starvation is an important host defense mechanism.

Supplementary Materials: The following are available online at www.mdpi.com/2076-2607/5/3/49/s1. **GFF File S1:** *C. gattii* R265 genome annotation—GFF file, **Figure S1:** Identification of annotation errors, **FASTA File S1:** Sequence file of *C. gattii* R265 genome annotation, **Table S1:** Cufflinks calculation of FPKM values from *C. gattii* R265 genes in BAL condition, **Table S2:** Statistical analysis of GO enrichment in the most abundant transcripts, **Table S3:** FPKM values of *C. gattii* R265 genes related to virulence, stress response, capsule, as well as mannoproteins, **Table S4**: FPKM values all identified transporters coding genes in *C. gattii* R265 genome, **Table S5:** FPKM values of amino acid metabolism related genes (biosynthesis and degradation), **Figure S2:** KEGG mapping of genes related to phenylalanine, tyrosine, and tryptophan biosynthesis, **Table S6:** FPKM values of predicted transcription factors.

Acknowledgments: Financial support for this study was provided by Coordenação de Apoio de Pessoal de Nível Superior—CAPES (process no. 23038.010041/2013-13) and Conselho Nacional de Desenvolvimento Científico e Tecnológico—CNPQ (process no. 312408/2014-5).

Author Contributions: P.A.G.F., R.S.A.S., F.M.d.S., P.R.d.S., R.M.C.d.A., and C.C.S. conceived and designed the experiments; P.A.G.F., R.S.A.S., F.M.d.S., P.R.d.S., R.M.C.d.A., and C.C.S. performed the experiments; P.A.G.F., R.S.A.S., and C.C.S. analyzed the data; R.M.C.d.A., A.S., L.K., M.H.V., and C.C.S. contributed reagents/materials/analysis tools; and P.A.G.F., R.S.A.S., and C.C.S. wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data, in the writing of the manuscript, and in the decision to publish the results.

References

1. Vallabhaneni, S.; Mody, R.K.; Walker, T.; Chiller, T. The Global Burden of Fungal Diseases. *Infect. Dis. Clin. N. Am.* **2016**, *30*, 1–11. [CrossRef] [PubMed]
2. Rajasingham, R.; Smith, R.M.; Park, B.J.; Jarvis, J.N.; Govender, N.P.; Chiller, T.M.; Denning, D.W.; Loyse, A.; Boulware, D.R. Global burden of disease of HIV-associated cryptococcal meningitis: An updated analysis. *Lancet Infect. Dis.* **2017**. [CrossRef]

3. Meyer, W.; Gilgado, F.; Ngamskulrungroj, P.; Trilles, L.; Hagen, F.; Castañeda, E.; Boekhout, T. Molecular Typing of the *Cryptococcus neoformans/Cryptococcus gattii* Species Complex. In *Cryptococcus*; American Society of Microbiology: Washington, DC, USA, 2011; pp. 327–357.
4. Chaturvedi, V.; Chaturvedi, S. *Cryptococcus gattii*: A resurgent fungal pathogen. *Trends Microbiol.* **2011**, *19*, 564–570. [CrossRef] [PubMed]
5. Kidd, S.E.; Hagen, F.; Tscharke, R.L.; Huynh, M.; Bartlett, K.H.; Fyfe, M.; Macdougall, L.; Boekhout, T.; Kwon-Chung, K.J.; Meyer, W. A rare genotype of *Cryptococcus gattii* caused the cryptococcosis outbreak on Vancouver Island (British Columbia, Canada). *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 17258–17263. [CrossRef] [PubMed]
6. Byrnes, E.J.; Marr, K.A. The Outbreak of *Cryptococcus gattii* in Western North America: Epidemiology and Clinical Issues. *Curr. Infect. Dis. Rep.* **2011**, *13*, 256–261. [CrossRef] [PubMed]
7. Byrnes, E.J.; Li, W.; Lewit, Y.; Ma, H.; Voelz, K.; Ren, P.; Carter, D.A.; Chaturvedi, V.; Bildfell, R.J.; May, R.C.; et al. Emergence and pathogenicity of highly virulent *Cryptococcus gattii* genotypes in the northwest United States. *PLoS Pathog.* **2010**, *6*, e1000850. [CrossRef] [PubMed]
8. Leopold Wager, C.M.; Hole, C.R.; Wozniak, K.L.; Wormley, F.L., Jr. *Cryptococcus* and Phagocytes: Complex Interactions that Influence Disease Outcome. *Front. Microbiol.* **2016**, *7*, 105. [CrossRef] [PubMed]
9. Zhang, M.; Sun, D.; Shi, M. Dancing cheek to cheek: *Cryptococcus neoformans* and phagocytes. *Springerplus* **2015**, *4*, 410. [CrossRef] [PubMed]
10. DeLeon-Rodriguez, C.M.; Casadevall, A. *Cryptococcus neoformans*: Tripping on acid in the phagolysosome. *Front. Microbiol.* **2016**, *7*, 1–9. [CrossRef] [PubMed]
11. Johnston, S.A.; May, R.C. *Cryptococcus* interactions with macrophages: Evasion and manipulation of the phagosome by a fungal pathogen. *Cell Microbiol.* **2013**, *15*, 403–411. [CrossRef] [PubMed]
12. Tucker, S.C.; Casadevall, A. Replication of *Cryptococcus neoformans* in macrophages is accompanied by phagosomal permeabilization and accumulation of vesicles containing polysaccharide in the cytoplasm. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 3165–3170. [CrossRef] [PubMed]
13. Potrykus, J.; Ballou, E.R.; Childers, D.S.; Brown, A.J. Conflicting interests in the pathogen-host tug of war: Fungal micronutrient scavenging versus mammalian nutritional immunity. *PLoS Pathog.* **2014**, *10*, e1003910. [CrossRef] [PubMed]
14. Kehl-Fie, T.E.; Chitayat, S.; Hood, M.I.; Damo, S.; Restrepo, N.; Garcia, C.; Munro, K.A.; Chazin, W.J.; Skaar, E.P. Nutrient metal sequestration by calprotectin inhibits bacterial superoxide defense, enhancing neutrophil killing of Staphylococcus aureus. *Cell Host Microbe* **2011**, *10*, 158–164. [CrossRef] [PubMed]
15. Hu, G.; Cheng, P.Y.; Sham, A.; Perfect, J.R.; Kronstad, J.W. Metabolic adaptation in *Cryptococcus neoformans* during early murine pulmonary infection. *Mol. Microbiol.* **2008**, *69*, 1456–1475. [CrossRef] [PubMed]
16. Derengowski, L.d.S.; Paes, H.C.; Albuquerque, P.; Tavares, A.H.; Fernandes, L.; Silva-Pereira, I.; Casadevall, A. The transcriptional response of *Cryptococcus neoformans* to ingestion by *Acanthamoeba castellanii* and macrophages provides insights into the evolutionary adaptation to the mammalian host. *Eukaryot. Cell* **2013**, *12*, 761–774. [CrossRef] [PubMed]
17. Ngamskulrungroj, P.; Chang, Y.; Sionov, E.; Kwon-Chung, K.J. The primary target organ of *Cryptococcus gattii* is different from that of *Cryptococcus neoformans* in a murine model. *MBio* **2012**, *3*, e00103-12. [CrossRef] [PubMed]
18. Babraham Bioinformatics. FastQC: A Quality Control Tool for High throughput Sequence Data. Available online: https://www.bioinformatics.babraham.ac.uk/projects/fastqc (accessed on 1 July 2017).
19. Hannon Lab. FASTX-Toolkit. Available online: http://hannonlab.cshl.edu/fastx_toolkit (accessed on 1 July 2017).
20. De Oliveira Schneider, R.; de Souza Süffert Fogaça, N.; Kmetzsch, L.; Schrank, A.; Vainstein, M.H.; Staats, C.C. Zap1 regulates zinc homeostasis and modulates virulence in Cryptococcus gattii. *PLoS ONE* **2012**, *7*, e43773.
21. Broad Institute. *Cryptococcus Gattii* R265 Genome and Annotation. Available online: http://archive.broadinstitute.org/ftp/pub/annotation/fungi/cryptococcus_gattii/genomes/cryptococcus_gattii_r265 (accessed on 1 July 2015).
22. Kim, D.; Pertea, G.; Trapnell, C.; Pimentel, H.; Kelley, R.; Salzberg, S.L. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **2013**, *14*, R36. [CrossRef] [PubMed]

23. Trapnell, C.; Roberts, A.; Goff, L.; Pertea, G.; Kim, D.; Kelley, D.R.; Pimentel, H.; Salzberg, S.L.; Rinn, J.L.; Pachter, L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **2012**, *7*, 562–578. [CrossRef] [PubMed]
24. Testa, A.C.; Hane, J.K.; Ellwood, S.R.; Oliver, R.P. CodingQuarry: Highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. *BMC Genom.* **2015**, *16*, 170. [CrossRef] [PubMed]
25. Thorvaldsdóttir, H.; Robinson, J.T.; Mesirov, J.P. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* **2013**, *14*, 178–192. [CrossRef] [PubMed]
26. Stajich, J.E.; Harris, T.; Brunk, B.P.; Brestelli, J.; Fischer, S.; Harb, O.S.; Kissinger, J.C.; Li, W.; Nayak, V.; Pinney, D.F.; et al. FungiDB: An integrated functional genomics database for fungi. *Nucleic Acids Res.* **2012**, *40*, D675–D681. [CrossRef] [PubMed]
27. Rybarczyk-Filho, J.L.; Castro, M.A.A.; Dalmolin, R.J.S.; Moreira, J.C.F.; Brunnet, L.G.; De Almeida, R.M.C. Towards a genome-wide transcriptogram: The *Saccharomyces cerevisiae* case. *Nucleic Acids Res.* **2010**, *39*, 3005–3016. [CrossRef] [PubMed]
28. String Database. Protein Network Data from *Cryptococcus gattii* WM276. Available online: https://string-db.org/cgi/download.pl?UserId=6Sz4kwKfQN72&sessionId=DlfdC4uHXzBV&species_text=Cryptococcus+gattii+WM276 (accessed on 1 July 2017).
29. KEGG Pathways. *Cryptococcus gattii* WM276. Available online: http://www.kegg.jp/kegg-bin/search_pathway_text?map=cgi&keyword=&mode=1&viewImage=true (accessed on 1 July 2017).
30. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinform.* **2009**, *10*, 421. [CrossRef] [PubMed]
31. Luo, W.; Pant, G.; Bhavnasi, Y.K.; Blanchard, S.G.; Brouwer, C. Pathview Web: User friendly pathway visualization and data integration. *Nucleic Acids Res.* **2017**, *175*, 1292–1302. [CrossRef] [PubMed]
32. Urban, M.; Cuzick, A.; Rutherford, K.; Irvine, A.; Pedro, H.; Pant, R.; Sadanadan, V.; Khamari, L.; Billal, S.; Mohanty, S.; et al. PHI-base: A new interface and further additions for the multi-species pathogen–host interactions database. *Nucleic Acids Res.* **2017**, *45*, D604–D610. [CrossRef] [PubMed]
33. Dos Santos, F.M.; Piffer, A.C.; Schneider, R.D.O.; Ribeiro, N.S.; Garcia, A.W.A.; Schrank, A.; Kmetzsch, L.; Vainstein, M.H.; Staats, C.C. Alterations of zinc homeostasis in response to *Cryptococcus neoformans* in a murine macrophage cell line. *Future Microbiol.* **2017**, *12*. [CrossRef] [PubMed]
34. Jung, W.H.; Sham, A.; Lian, T.; Singh, A.; Kosman, D.J.; Kronstad, J.W. Iron source preference and regulation of iron uptake in *Cryptococcus neoformans*. *PLoS Pathog.* **2008**, *4*, e45. [CrossRef] [PubMed]
35. Caza, M.; Kronstad, J.W. Shared and distinct mechanisms of iron acquisition by bacterial and fungal pathogens of humans. *Front. Cell. Infect. Microbiol.* **2013**, *3*, 80. [CrossRef] [PubMed]
36. Jung, K.; Yang, D.; Maeng, S.; Lee, K.; So, Y.; Hong, J.; Choi, J.; Byun, H.; Kim, H.; Bang, S.; et al. Systematic functional profiling of transcription factor networks in *Cryptococcus neoformans*. *Nat. Commun.* **2015**, *6*, 1–14. [CrossRef] [PubMed]
37. Zhao, C.; Waalwijk, C.; De Wit, P.J.G.M.; Tang, D.; Van Der Lee, T. RNA-Seq analysis reveals new gene models and alternative splicing in the fungal pathogen *Fusarium graminearum*. *BMC Genom.* **2013**, *14*, 21. [CrossRef] [PubMed]
38. Schliebner, I.; Becher, R.; Hempel, M.; Deising, H.B.; Horbach, R. New gene models and alternative splicing in the maize pathogen *Colletotrichum graminicola* revealed by RNA-Seq analysis. *BMC. Genom.* **2014**, *15*, 842. [CrossRef] [PubMed]
39. Janbon, G.; Ormerod, K.L.; Paulet, D.; Byrnes, E.J.; Yadav, V.; Chatterjee, G.; Mullapudi, N.; Hon, C.C.; Billmyre, R.B.; Brunel, F.; et al. Analysis of the Genome and Transcriptome of *Cryptococcus neoformans* var. *grubii* Reveals Complex RNA Expression and Microevolution Leading to Virulence Attenuation. *PLoS Genet.* **2014**, *10*, e1004261. [CrossRef] [PubMed]
40. Bruno, V.M.; Wang, Z.; Marjani, S.L.; Euskirchen, G.M.; Martin, J.; Sherlock, G.; Snyder, M. Comprehensive annotation of the transcriptome of the human fungal pathogen *Candida albicans* using RNA-seq. *Genome Res.* **2010**, *20*, 1451–1458. [CrossRef] [PubMed]
41. Haas, A. The Phagosome: Compartment with a License to Kill. *Traffic* **2007**, *8*, 311–330. [CrossRef] [PubMed]
42. Fan, W.; Kraus, P.R.; Boily, M.J.; Heitman, J. *Cryptococcus neoformans* gene expression during murine macrophage infection. *Eukaryot. Cell* **2005**, *4*, 1420–1433. [CrossRef] [PubMed]

43. Chen, Y.; Toffaletti, D.L.; Tenor, J.L.; Litvintseva, A.P.; Fang, C.; Mitchell, T.G.; McDonald, T.R.; Nielsen, K.; Boulware, D.R.; Bicanic, T.; et al. The *Cryptococcus neoformans* transcriptome at the site of human meningitis. *MBio* **2014**, *5*, e01087-13. [CrossRef] [PubMed]
44. Russel Lee, I.; Chow, E.W.L.; Morrow, C.A.; Djordjevic, J.T.; Fraser, J.A. Nitrogen metabolite repression of metabolism and virulence in the human fungal pathogen *Cryptococcus neoformans*. *Genetics* **2011**, *188*, 309–323.
45. Kmetzsch, L.; Staats, C.C.; Simon, E.; Fonseca, F.L.; Oliveira, D.L.; Joffe, L.S.; Rodrigues, J.; Lourenco, R.F.; Gomes, S.L.; Nimrichter, L.; et al. The GATA-type transcriptional activator Gat1 regulates nitrogen uptake and metabolism in the human pathogen *Cryptococcus neoformans*. *Fungal Genet. Biol.* **2011**, *48*, 192–199. [CrossRef] [PubMed]
46. Ngamskulrungroj, P.; Chang, Y.; Roh, J.; Kwon-Chung, K.J. Differences in nitrogen metabolism between *Cryptococcus neoformans* and *C. gattii*, the two etiologic agents of cryptococcosis. *PLoS ONE* **2012**, *7*, e34258. [CrossRef] [PubMed]
47. Rutherford, J.C.; Lin, X.; Nielsen, K.; Heitman, J. Amt2 permease is required to induce ammonium-responsive invasive growth and mating in *Cryptococcus neoformans*. *Eukaryot. Cell* **2008**, *7*, 237–246. [CrossRef] [PubMed]
48. Cunningham, T.S.; Andhare, R.; Cooper, T.G. Nitrogen catabolite repression of DAL80 expression depends on the relative levels of Gat1p and Ure2p production in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **2000**, *275*, 14408–14414. [CrossRef] [PubMed]
49. Luzzani, C.; Cardillo, S.B.; Moretti, M.B.; García, S.C. New insights into the regulation of the Saccharomyces cerevisiae UGA54 gene: Two parallel pathways participate in carbon-regulated transcription. *Microbiology* **2007**, *153*, 3677–3684. [CrossRef] [PubMed]
50. Cardillo, S.B.; Moretti, M.B.; García, S.C. Uga3 and Uga35/Dal81 transcription factors regulate UGA4 transcription in response to γ-Aminobutyric acid and Leucine. *Eukaryot. Cell* **2010**, *9*, 1262–1271. [CrossRef] [PubMed]
51. Martho, K.F.C.; De Melo, A.T.; Takahashi, J.P.F.; Guerra, J.M.; Da Silva Santos, D.C.; Purisco, S.U.; Melhem, M.D.S.C.; Dos Anjos Fazioli, R.; Phanord, C.; Sartorelli, P.; et al. Amino acid permeases and virulence in *Cryptococcus neoformans*. *PLoS ONE* **2016**, *11*, e0163919. [CrossRef] [PubMed]
52. Liew, K.L.; Jee, J.M.; Yap, I.; Yong, P.V.C. In vitro analysis of metabolites secreted during infection of lung epithelial cells by *Cryptococcus neoformans*. *PLoS ONE* **2016**, *11*, e0153356. [CrossRef] [PubMed]
53. Lorenz, M.C.; Bender, J.A.; Fink, G.R. Transcriptional Response of Candida albicans upon Internalization by Macrophages Transcriptional Response of *Candida albicans* upon Internalization by Macrophages. *Eukaryot. Cell* **2004**, *3*, 1076–1087. [CrossRef] [PubMed]



© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

3.3 Capítulo III: Development of an automated and optimized pipeline for gene prediction in *Cryptococcus gattii* R265

Manuscript in preparation

Development of an automated and optimized pipeline for gene prediction in *Cryptococcus gattii* R265

Patrícia Aline Gröhs Ferrareze[1], Corinne Maufrais[2], Frederique Moyrand[2], Charley Christian Staats[1] and Guilhem Janbon[2]

[1] Laboratório de Microbiologia Molecular e Celular, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul, Porto Alegre – Brazil

[2] Unité Biologie des ARN des Pathogènes Fongiques, Département Mycologie, Institut Pasteur, Paris - France

ABSTRACT

The development of automated pipelines to gene prediction in intron-rich fungal genomes is still a challenge in this days. In this sense, RNA-Seq based pipelines may solve the hard task of intron-exon definitions as well as, gene boundaries in gene dense genomes. *Cryptococcus* species, in turn, represent good models to development and optimization of a gene prediction pipeline for complex fungal genomes. *Cryptococcus neoformans* H99 and JEC21, firstly, due of the publication of reliable gene annotations that can be used as reference; *C. gattii*, secondly, due to necessity of understand the genetic determinants of the hyper virulent strains from Pacific Northwest outbreak. Thus, this study aims to describe the development of a *Cryptococcus* optimized pipeline for gene prediction with RNA-Seq data; as well as, the successfully application of this protocol to *C. gattii* R265 with 70% of accuracy in gene structures.

INTRODUCTION

The development of the next-generation high-throughput sequencing technologies started a new era in transcriptomic studies. As gene prediction and genome annotation is a hard task in non-model organisms, it is usually necessary the

combination of multiple evidence sources. In this way, RNA sequencing (RNA-Seq) is an important tool to improve the accuracy of genome annotations by correction of *ab initio* predicted gene structures; and the detection of alternative splicing isoforms as well as of new transcripts (MCGETTIGAN, 2013).

The genome assembly and annotation of *Cryptococcus* species has been provided by the Broad Institute since 2011, serving as the basis of many comparative and functional studies, which demands high accuracy of this data. In order to generate the gene models, the Broad Institute adopted an automated annotation methodology, using homology data as parameters for their tools (D'SOUZA et al., 2011). Despite the last revision of the transcriptome annotation and the improvement of the gene models accomplished by it (FARRER et al., 2015), there are inherent errors within this methodology, like the reduced precision when predicting the structure of non-conserved genes and the difficulty on finding new genes, which puts the reliability of the annotation in check. As an example, previous work from our group led to the identification of an annotation error of the gene *ZAP1* of the R265 strain of *C. gattii*, in which the predicted structure of exons and introns did not match with the RNA sequencing (RNA-seq) data (SCHNEIDER et al., 2012).

Recently, a new *C. gattii* R265 genome assembly was released, emcompassing data from PacBio and Nanopore sequencing. In this way, the 27 supercontigs previously available are now assembled *de novo* in 14 chromosomes and fixed several errors in the genome sequence and structure (YADAV et al., 2018). In *C. neoformans*, the application of RNA-seq methodologies have greatly improved the previous automatic annotation, not only correcting the structure a considerable number of gene

models, but also by providing the information about untranslated regions (UTRs) and identifying new genes (JANBON et al., 2014). Thus, this study aimed to develop an automated pipeline to gene prediction based in RNA-Seq data evidence. In order to successfully apply with *C. gattii* R265, which have an intron rich genome, this pipeline was optimized, standardized and validated with the sibling species *C. neoformans* in which gene coordinates were manually curated, serving as a gold standard for comparative purposes.

MATERIALS AND METHODS

Culture and RNA Sequencing

C. neoformans strain H99 and JEC21 and *C. gattii* R265 and were grown in YPD at 30°C 12 or 37°C under agitation up to exponential or early stationary phase as previously described (JANBON et al., 2014). 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 (82) cells grown in YPD at 30°C in stationary phase. Cells were washed, snap frozen and used to prepare RNA and total DNA samples. Each condition was used to prepare biological triplicate samples. For RNA-Seq, strand specific, paired-end cDNA libraries were prepared from 10 µg of total RNA by poly A selection 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).

RNA-Seq libraries trimming and rRNA cleaning

The paired reads from RNA-Seq libraries were trimmed for low quality reads and Illumina TruSeq adapters removal with Cutadapt v1.9.1 (MARTIN, 2011) with the followed 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; SALZBERG, 2012) with default parameters; and report of unmapped paired reads (--*un-conc*) as output file for selection of reads that did not align with rRNA sequences.

RNA-Seq mapping

The H99 and JEC21 sequence reads from RNA-Seq paired-end libraries were mapped against the H99 and JEC21 reference genomes (NCBI Genome assemblies GCA_000149245.3 and GCA_000091045.1), respectivelly, using Tophat2 v2.0.14 (KIM et al., 2013) with the following parameters: --min-intron-length 30; --min-coverage-intron 30; --min-segment-intron 30; --max-intron-length 4000; --max-multihits 1; --microexon-search; and --library-type fr-firststrand or fr-secondstrand (according RNA-Seq library).

Pipeline selection

The RNA-Seq mapped read pairs were tested in three pipelines for gene prediction (Figure 1). BRAKER1 (HOFF et al., 2016) was performed with the default parameters plus the exclusion of alternative transcripts (--alternatives-from-evidence=false), and the three merged replicates from EXPO30 condition as RNA-Seq source. The other two pipelines, here named “Cuff-CQ” and “C3Q”, are composed by Cufflinks v2.1.1 (TRAPNELL et al., 2010) and CodingQuarry v2.0 (TESTA et al., 2015). These pipelines (“Cuff-CQ”: Cufflinks + CodingQuarry; “C3Q”: Cufflinks + Cuffmerge + CodingQuarry) were tested with the Cufflinks parameters: --min-intron-length 30; --max-intron-length 4000; --min-isoform-fraction 0.9; and –overlap-radius 10. The merged BAM file from the three replicates used in Cuff-CQ pipeline was obtained with Samtools *merge.* C3Q was performed separately with the BAM files from each replicate; the GTF files generated by the three predictions (for replicates A, B, and C) were then combined by Cuffmerge and the resulting transcripts were processed by CodingQuarry. The evaluation of the pipelines sensibility and precision to gene prediction was performed by comparison of the predicted annotations against H99 and JEC21 reference annotations (WALLACE et al., 2019) with GFFcompare program (http://ccb.jhu.edu/software/stringtie/gffcompare.shtml).

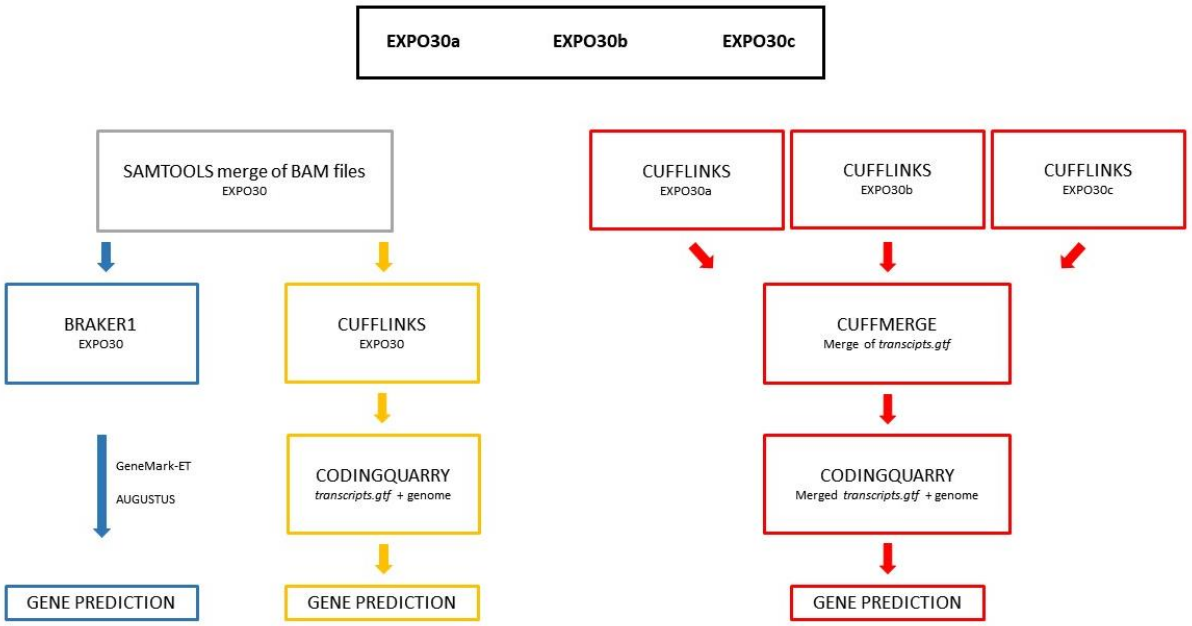

Figure 1. Gene prediction pipelines performed with *C. neoformans* RNA-Seq libraries. BRAKER1 pipeline (blue) is composed of GeneMark-ET and AUGUSTUS programs, in an unsupervised and automatic way. Cuff-CQ (orange) is composed of Cufflinks and CodingQuarry programs. C3Q (red) is composed of Cufflinks, Cuffmerge and CodingQuarry programs. EXPO30 "a", "b", and "c": RNA-Seq library replicates from condition Exponential 30 ºC; *transcripts.gtf:* Cufflinks output of assembled transcripts.

Cufflinks parameters selection

The selection of the Cufflinks best parameters combination was performed with EXPO30 RNA-Seq libraries of mapped reads and C3Q pipeline (Cufflinks + Cuffmerge + CodingQuarry). For this, the Cufflinks transcripts assembly generated for each EXPO30 replicate was tested with fixed and variable parameter combinations (Table 1). Posteriorly, as established for the C3Q pipeline, the predicted GTFs were merged and processed by Coding Quarry. 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 sensibility and precision to gene prediction was performed by comparison of the predicted annotations against H99 and JEC21 reference annotations with GFFcompare program.

Table1. Cufflinks parameters combinations tested to pipeline optimization according *Cryptococcus* transcriptome.

Code	Cufflinks parameters 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 the application of C3Q pipeline with the best selected parameters (“Q” combination) in the five RNA-Seq libraries from *C. neoformans* H99 and the 4 RNA-Seq libraries from *C. neoformans* JEC21 (Table S1). The generated predictions were then compared with reference annotations of both strains in GFFcompare program.

Effect of different conditions in the predictions

To evaluate the effect of the conditions in the predicted annotation, were performed combinations between 2 and 4 RNA-Seq condition libraries for *C. neoformans* H99 and between 2 and 3 RNA-Seq condition libraries for *C. neoformans* JEC21. The predictions for each combination were performed according the C3Q

pipeline and “Q” Cufflinks parameters, with the merge of the selected conditions before Coding Quarry processing. The prediction combination values were generated by comparison with reference annotations of both species in GFFcompare program.

Evaluation of BAM sampled files in gene prediction quality

The downsampling of the replicates from EXPO30 condition was obtained with the tool *PositionBasedDownsampleSam* (default parameters) from Picard package (https://broadinstitute.github.io/picard/), which applies a random algorithm to downsampling the BAM files. Thus, were defined fractions of 1, 5, 7.5, 10, 15, 20, 30 and 40 million reads for each replicate. Posteriorly, the predictions were performed according C3Q pipeline (with the merging of predicted GTFs from each replicate), Cufflinks “Q” parameters combination and these downsampled BAM files. The analysis of prediction values was performed by comparison with reference annotations for *C. neoformans H99* and *C. neoformans* JEC21 in GFFcompare program.

Gene predictions with downsampled H99 and JEC21 RNA-Seq libraries

The gene predictions with downsampled BAM files for the RNA-Seq conditions was performed according the model of C3Q pipeline, “Q” Cufflinks parameters and downsampled RNA-Seq alignment files for all H99 and JEC21 sequenced conditions (Table S1). The downsampling of each condition replicate to 7.5 million reads each, was performed with Picard package, as previously described. The analysis of prediction

values was performed by comparison with reference annotations for *C. neoformans H99* and *C. neoformans* JEC21 in GFFcompare program.

Gene filtering

The selection of small coding sequences, as well as intronless sequences and genome-predicted sequences was performed by a home AWK-based script. The data from gene lengths and exons numbers was obtained in output files of transcripts analysis from GFFcompare program. The evaluation of reads count and FPKM to selection of genome-predicted sequences with no reads was performed with HTSeq-count (ANDERS; PYL; HUBER, 2015).

Analysis of missed and novel genes

The analysis of missed genes and novel genes was performed with the data (number of exons, length) obtained by GFF compare program and the FPKM calculated from HTSeq read counts with reference and predicted annotations. The evaluation of conserved domains was performed with the functional annotation server Pannzer2 (http://ekhidna2.biocenter.helsinki.fi/sanspanz/).

Gene prediction with *C. gattii* R265

The prediction with *C. gattii* R265 was performed by the application of C3Q pipeline with the best selected parameters (“Q” combination). For this, five RNA-Seq triplicate libraries from *C. gattii* R265 (Table S1) were previously subsampled to 7.5 million reads each with Piccard tools and separately assembled in transcripts by Cufflinks. The GTF files resulting from the replicate assemblies were then merged by Cuffmerge and processed in CodingQuarry predictor with the R265 genome (NCBI assembly GCA_002954075.1) and default parameters. The filtering of small sequences up to 150 nt, intronless sequences up to 300 nt and genome-predicted sequences with no reads was performed according previously described.

The retrieve of non-predicted genes was performed with Exonerate v2.2.0 (SLATER; BIRNEY, 2005) by the alignment of *C. neoformans* H99 and JEC21 protein sequences to R265 genome (protein2genome --percent 50 --bestn --minintron 30 –maxintron 4000)**.**

The evaluation of the predicted annotation was performed by comparison with manually corrected genes from chromosomes 9 and 14, as well as, by comparison with annotations from Broad (NCBI assembly GCA_000149475.3) and Ferrareze et al., 2017. The identification of CDS gene coordinates from old annotations in the new sequenced genome was performed with Exonerate aligner (coding2genome). The statistics of the gene annotations was generated by GFFcompare program.

The manual corrections of genes from chromosomes 9 and 14 were performed with the software Artemis (CARVER et al., 2012), the new sequenced R265 genome

(NCBI assembly GCA_002954075.1) and the five triplicate conditions of stranded paired-end RNA-Seq data from *C. gattii* R265 (Table S1).

The evaluation of orthogroups and unique genes to *C. neoformans* H99, JEC21 and *C. gattii* R265 was performed with Orthofinder v2.3.3 (EMMS; KELLY, 2015). The analysis of conserved domains in unique sequences, as well as, the functional annotation of *C. gattii* R265 were performed with Pannzer2 server.

RESULTS AND DISCUSSION

Pipeline selection and optimization

We used and compared the performances of two previously published annotations pipelines in coding genes *de novo* annotation in intron-rich fungi using the same RNA-Seq set of data. We more specifically compared the unsupervised pipeline BRAKER1, which combines the softwares Gene Mark-ET and Augustus and is already optimized with the best prediction parameters with a pipeline composed by the softwares Cufflinks and CodingQuarry. We tested them on the *C. neoformans* strain H99 genome first because its annotation is most probably one of most complete and detailed among the intron-rich fungi (JANBON et al., 2014; WALLACE et al., 2019). Second, with 99.5% of 6795 annotated coding genes containing introns, five to six introns per coding gene and 37832 introns overall, an automatic annotation of this genome can be considered as challenging (WALLACE et al., 2019). Third, large sets of RNA-Seq data has been produced by our laboratory from cells grown in different

conditions. Lastly, because we also produced a high quality genome annotation and RNA-Seq datasets from cells grown on the same conditions of a close relative strain (i.e. *C. neoformans* JEC21) which could be used eventually to validate our conclusions (GONZALEZ-HILARION et al., 2016; WALLACE et al., 2019).

We used RNA-Seq data obtained in biological triplicate from cells grown in exponential phase at 30°C in complete medium (YPD) (EXPO30) (WALLACE et al., 2019). Reads were trimmed and aligned to the *C. neoformans* genome and, the resulting bam files were used to feed the two pipelines. For the Cufflink-CodingQuarry we used two alternative strategies. In one case, we first merged the three bam files whereas (Cuff – CQ protocol), in the second case, each replicate bam file was use to generate a different GTF prediction file which were merged using Cuffmerge and used by CodingQuarry as a single transcript source (C3Q protocol) (Figure 1).

In order to compare the quality of these predictions, we considered first the coding genes and calculated the sensitivity of the prediction (percentage of annotated features with overlap in the predictions) and the specificity of the prediction (the percentage of predicted features with overlap in the annotation). As shown on the Figure 2 and Table S2, BRAKER1 is by far the most sensitive pipeline, missing only 91 coding genes in the *C. neoformans* H99 genome and 54 in *C. neoformans* JEC21. However, it is also the less specific, predicting “novel” 622 coding genes for H99 and 1065 “novel” coding genes for JEC21 (absent in the reference annotations) (Figure 3 and Table S2). In contrast, both Cufflinks – CodingQuarry protocols miss more coding genes (783 and 447 for H99; and 625 and 318 for JEC21, in Cuff-CQ and C3Q,

respectively), but C3Q reach a specificity of approximately 95%, making these predictions less noisy.

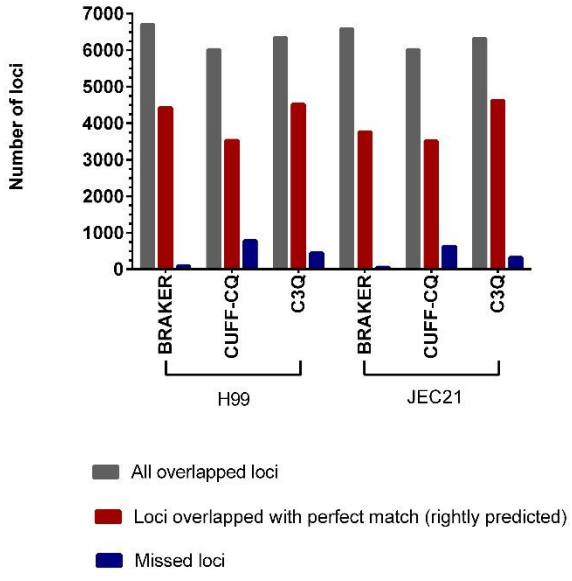

Figure 2. Numbers of genes from H99 and JEC21 reference annotations that are overlapped, missed and rightly predicted in gene predictions from BRAKER1, Cuff-CQ and C3Q.

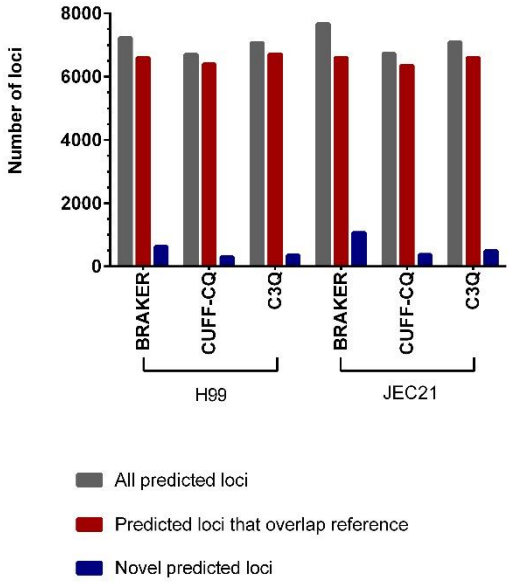

Figure 3. Numbers of predicted, predicted loci overlapping reference, and novel loci in gene predictions from BRAKER1, Cuff-CQ and C3Q. Values obtained by comparison with the reference annotations from *C. neoformans* H99 and JEC21.

At the level of the introns and exons within identified reference genes (Figure 4 and table S2), the pattern is similar. BRAKER1 pipeline is very sensitive with only 0.4% missed introns (n= 157) and 0.4% missed exons (n=164) in H99 prediction, but poorly specific with 4471 novel introns and 3065 novel exons not present in the H99 reference annotation. In the other hand, both Cufflink – CodingQuarry protocols (Cuff – CQ and C3Q) miss more introns (n = 4944 and 3238 respectively) and more exons (n= 4281 and n= 2777, respectively) in H99 gene predictions, but are both more specific

predicting less than 200 introns or exons not present in the reference annotation (Table S2).

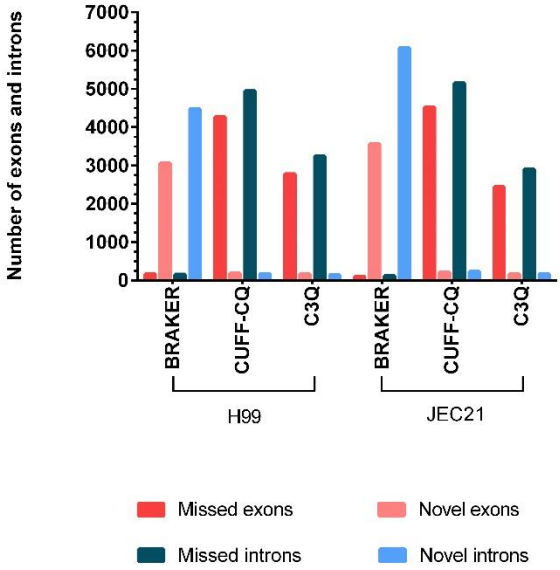

Figure 4. Numbers of missed exons, missed introns, novel exons and novel introns in gene predictions from BRAKER1, Cuff-CQ and C3Q. Relative to predicted genes that overlap reference annotations from *C. neoformans* H99 and JEC21.

Overall, the Cufflink – CodingQuary protocols use a more conservative model. They are more specific in the predicted structures with a lower number of new insertions (novel exons/introns) or new genes at the cost of missing information with higher number of missed features. Accordingly, if one considers the structure of the gene, the C3Q protocol is the most sensitive with 66.5% (n= 4516) of the H99 genes with a perfectly predicted exon-intron layout (65.2% and 51.9% for BRAKER1 and the CUFF - CQ protocol, respectively) (Figure 2 and Table S2). For JEC21, C3Q pipeline reach 69.5% (n= 4616) of JEC21 reference genes with rightly predicted structures, while 56.6% of BRAKER1 predicted JEC21 genes have perfect match with reference. It is

here also the most specific with 63.9% and 65.2% of the predicted genes perfectly fitting the H99 and JEC21 reference gene structures, respectively (Table S2). The achievement of similar results of specificity and prediction quality for H99 and JEC21 confirm that the C3Q protocol was the most promising one to be test for optimization.

Cufflinks parameters optimization

In order to improve both the number of perfectly predicted gene structures and the percentage of predicted loci perfectly fitting the reference coding gene structures, we altered the Cufflinks parameters in 17 combinations (see material and methods). We here tested the effect of the minimum distance between transfrags to merge, the trimming of 3' ends of reads, the filtering of alignments that lie within the intronic intervals, the filtering of suspicious spliced reads and the minimum RNA-Seq fragments allowed to assemble transfrags. These modifications of the Cufflinks parameters all reduce the number of missed genes and improved the prediction quality by increasing of rightly predicted structures. However, despite P and Q combinations present the best results applicable for both *C. neoformans* strains, the Q combination reached best general values of specificity and sensibility. The difference between these combinations is related to the filter of alignments that lie within intronic intervals; P have a filter value of 100% and Q present a filter value of 85%. As observed in supplementary analysis with gene predictions generated by each replicate – without merging – and predictions with all conditions (Table S2), the optimal filter value may vary between replicates and

conditions, possibly reflecting RNA-Seq library differences. Nevertheless, for the utilization of all RNA-Seq conditions, Q combination offers the best results.

Thus, Q combination presented rightly predicted structures for 74.4% and 71.5% of the H99 and JEC21 overlapped reference genes. This means that 70.8% of the H99 and 68% of the JEC21 reference genes were perfectly predicted by 66.8% and 64.2% of the predicted genes, respectively. Besides, 95.1% and 95.2% of H99 and JEC21 reference genes were reached by these parameters in EXPO30 condition (Table S2).

Effect of the RNA-seq data set

In *Cryptococcus* samples, the differences in growth cultures produced differences in RNA-Seq libraries, and, as consequence, in gene prediction. This can be seen by perfect match values found between conditions, as well as the number of missed and predicted loci (Figure 5 and Table S2). Each RNA-Seq condition generated a different number of predicted loci, ranging between 7049 loci (STAT37) up to 7199 loci (EXPO30) in H99; and 6670 (STAT37) up to 7187 (STAT30) in JEC21; of these, about 93% are specific to reference predictions. Moreover, between 65.2 (MATING, n= 4427) and 71.5% (EXPO37, n= 4858) of H99 reference annotation was rightly predicted. The ratio of missed loci also varies from 4.9 (EXPO30, n= 333) to 6% (STAT37/MATING, n= 410). JEC21, as well, obtained perfect match between 57.6 (STAT37, n= 3821) and 69.6% (STAT30, n= 4624) of the JEC21 reference loci; with missed loci numbers between 4.2 (STAT30, n= 278) and 8.9% (STAT37, n= 592).

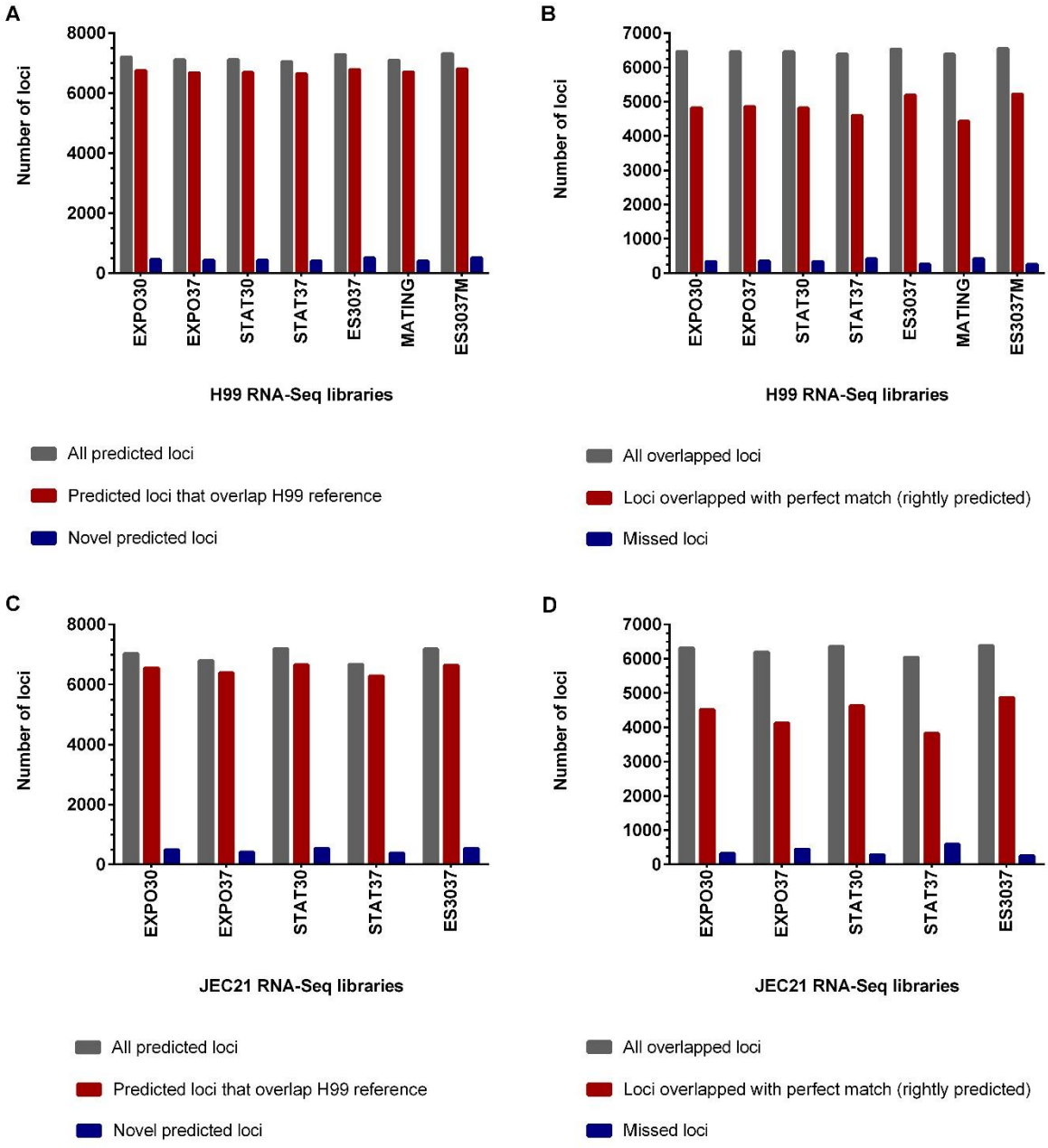

Figure 5. Prediction results obtained with C3Q pipeline and Cufflinks parameters combination Q applied to RNA-Seq libraries from conditions of *C. neoformans* H99 and JEC21. Figures A and C: relative values to the entire predicted annotations, including hypothetical novel loci in H99 and JEC21. Figures B and D: number of genes from H99 and JEC21 reference annotations that are overlapped, missed and rightly predicted. ES3037: EXPO30 + EXPO37 + STAT30 + STAT37. ES3037M: EXPO30 + EXPO37 + STAT30 + STAT37 + MATING.

Nevertheless, the merging of the Cufflinks GTF files from all condition libraries compensates these small differences; consequently, the process led to the increase of the number of rightly predicted loci (n= 5215 for H99 and 4869 for JEC21) and the decrease of the value of missed loci in the final annotation (254 for the both strains).

Thus, 79.7% of the predicted loci that overlap H99 reference loci had rightly predicted structures (or 76.4% of H99 reference annotation) with the merging of EXPO30, EXPO37, STAT30 and STAT37 and MATING. In JEC21, similarly, the merge of Cufflinks GTFs from all four available conditions (EXPO30, EXPO37, STAT30 and STAT37) improves the value of loci with perfect match 76.3% that overlap H99 reference (73.3% of the JEC21 reference annotation). All values of missed loci decreased, reaching 3.7 and 3.8% in H99 and JEC21 annotations. Lastly, decreased numbers of novel and missed introns and exons are similarly achieved by the merge of gene predictions from all conditions in C3Q pipeline (Figure 6 and Table S2).

Evaluation of the effect of different combinations of RNA-Seq libraries in the quality of prediction

The evaluation of the effect of different libraries in the quality (perfect match ratio) of prediction results was performed with combinatorial analysis (Table S2). Combinations of 2, 3 and 4 elements (conditions) were generated for H99, and combinations of 2 and 3 elements were generated for JEC21. Each combination was created according C3Q pipeline, with the merging of GTF files of Cufflinks transcripts for the selected conditions, with posterior gene prediction by Coding Quarry.

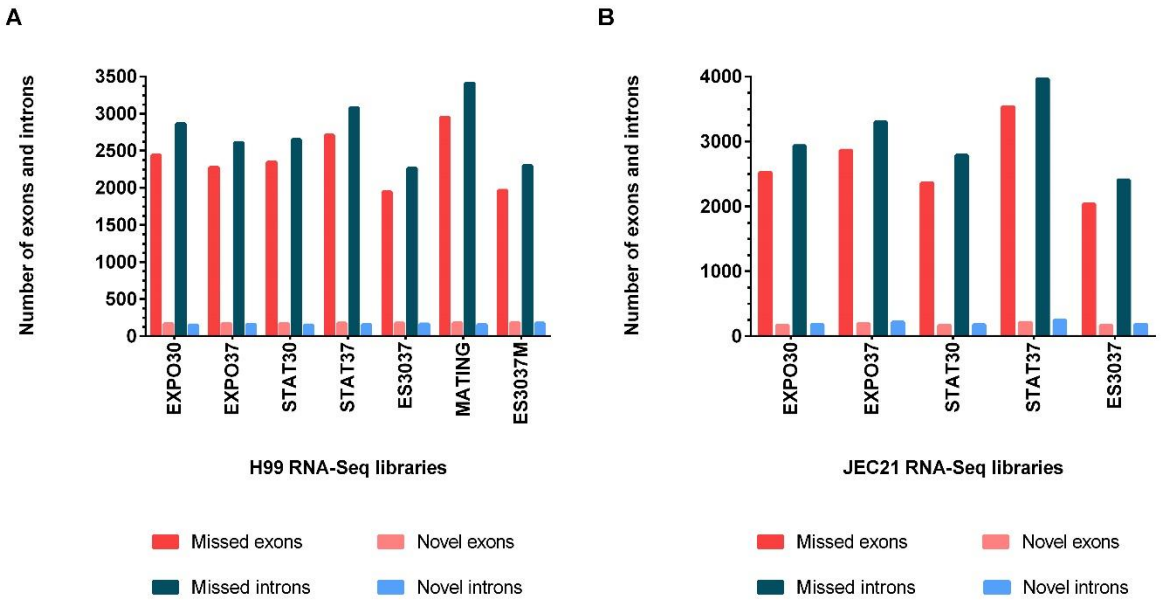

Figure 6. Specific prediction results obtained with C3Q pipeline and Cufflinks parameters combination Q applied to RNA-Seq libraries from conditions of *C. neoformans* H99 (A) and JEC21 (B). Relative values of exons and introns for loci overlapped in reference annotations. ES3037: EXPO30 + EXPO37 + STAT30 + STAT37. ES3037M: EXPO30 + EXPO37 + STAT30 + STAT37 + MATING.

As previously observed for H99, despite STAT37 and MATING libraries have the lowest perfect match values (70% of the reference loci), when they are combined with other RNA-Seq libraries, the results are reimbursed. JEC21, as previously seen by individual results of libraries, had lower values than H99. However, the combinations successfully got perfect match values of at least 65.5% (or 69.7% of the loci that overlapped H99 annotation) for EXPO37+STAT37 combination; being restored with the inclusion of EXPO30/STAT30 libraries (Table S2). Still, it is possible to observe that the merge of both GTF files before prediction with Coding Quarry improves the individual results obtained by each condition.

As expected, including more samples improved the annotation quality. The usage of the 5 conditions improve the C3Q annotation quality due the fact that the more

data is used the more predicted gene not present in the reference genome are identified (Table S2).

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

Our comparison of pipeline performances reveals the usage of a single BAM file of larger size decreases the efficiency of the Cufflinks – CodingQuarry pipeline in transcript assembly and in de novo annotation the *C. neoformans* genome. Similarly, the whole set of RNA-Seq replicates (EXPO30, EXPO37, STAT30 and STAT37) was used to feed the BRAKER1 pipelines the resulting quality of the predicted annotation was worse than when an any of the single conditions was used (Table S2). Therefore, for all pipelines, the amount of information seems to alter the quality of gene prediction.

We looked at the quality of the annotation depending on the size of the initial file used comparing the results obtained using the C3Q pipeline using individual replicates of the EXPO30 RNA-Seq data and Cuff-CQ pipeline with the merged RNA-Seq libraries. As shown on the Figure 2, the larger the dataset is the poorest the annotation. Accordingly, the same analysis performed in *C. neoformans* JEC21, which gave a similar result, therefore, suggesting that the sequencing depth may affect deeply the quality of predictions and, should considered as a possible parameter of optimization for gene prediction pipelines. In order to improve the analysis of the effect of size of the data set on the quality of gene prediction, the C3Q pipeline was tested with different representative fractions of reads from each EXPO30 replicates. Thus, replicate samples with 1, 5, 7.5, 10, 15, 20, 30 and 40 million of reads were used to de novo

annotation the *C. neoformans* H99 genome and the quality of the gene predictions was compared. We performed the same analysis using the same strategy for *C. neoformans* JEC21. As shown on the Figure 7 and Table S2, for both species the quality of the gene prediction is highly dependent on the size of the RNA-Seq initial data set. Interestingly, the quality of the gene prediction is strongly anti-correlated the number of Cufflinks assembled transcripts. As seen in Figure 7, the best perfect gene predictions were obtained with BAM replicates between 5 and 7.5 million of reads. Using this adjusted read depth of the alignment files, the predicted gene structure was improved for nearly all parameters including missed genes, missed exons and missed introns thus improving the annotation (Table S2).

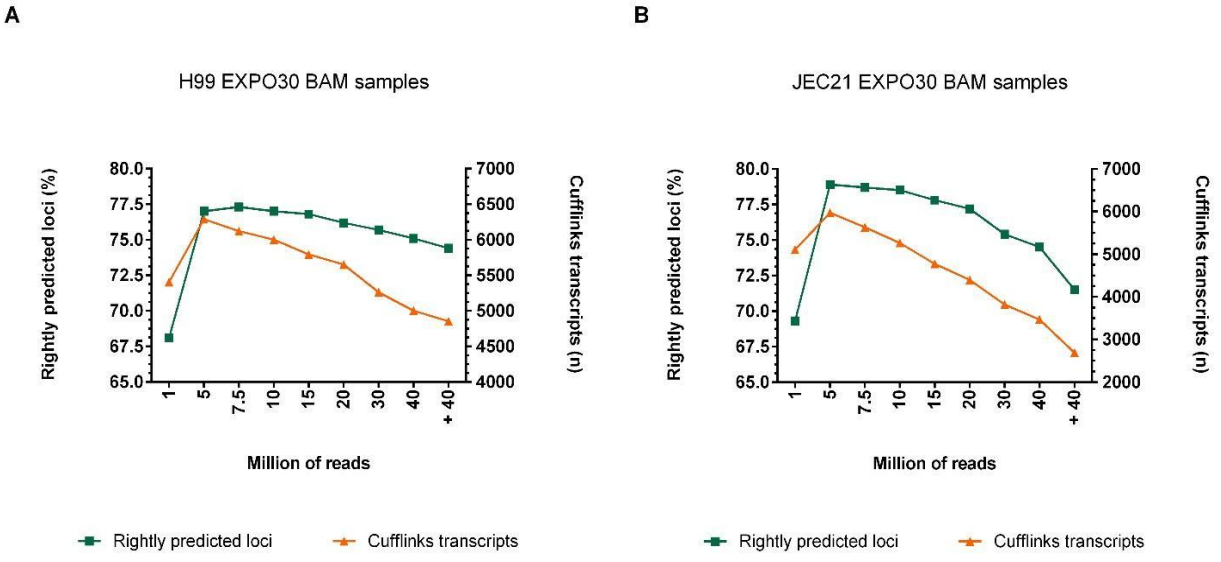

Figure 7. Prediction results obtained with C3Q pipeline and Cufflinks parameters combination Q applied to different EXPO30 downsampled BAM files from *C. neoformans* H99 and *C. deneoformans* JEC21. +40: prediction performed with the entire BAM replicates from EXPO30 condition (number of reads available in Table S1).

In fact, contrary to the idea that more information (in this case, a higher number of reads) leads to a better transcript assembly and gene prediction, these results also shown that despite relatively low number of reads is necessary to keep the high prediction accuracy. Thus, we adjusted the number of reads to 7.5 million for each replicate of each condition and used this adjusted RNA-Seq data set to de novo annotated, the *C. neoformans* H99 and JEC21 genomes. For each data set, we observed an improvement of the numbers of rightly predicted genes and a decrease in the number of the missed genes and of the overall quality of the gene prediction (Figures 8, 9 and Table S2).

The percent of loci that overlap H99 reference annotation with a rightly predicted structure range between 72.8 (MATING) and 79.2% (EXPO37). The ratio of missed loci decreased, reaching values between 3.2 (EXPO37) and 3.9% (MATING). JEC21, as well, obtained values ranging between 74 (STAT37) and 78.7% (EXPO30) of loci that overlap JEC21 reference with perfect match; with missed loci numbers between 2.2 (EXPO30) and 3.4% (EXPO37/STAT37).

Still, the merging of the Cufflinks GTF files from all condition libraries also maintain a stronger effect in gene prediction values when applied with this optimized pipeline. In this context, 83.5% of the predicted loci that overlap H99 reference loci had rightly predicted structures (81.9% of H99 reference annotation), whereas the ratio of missed loci is retained in 1.9%. In JEC21, similarly, the merge of Cufflinks GTFs from all four conditions improved the value of loci with perfect match in 83.2% of the predicted loci that overlap H99 reference (81.8% of JEC21 reference annotation). Jointly, the value of missed loci decreased, reaching 1.7%.

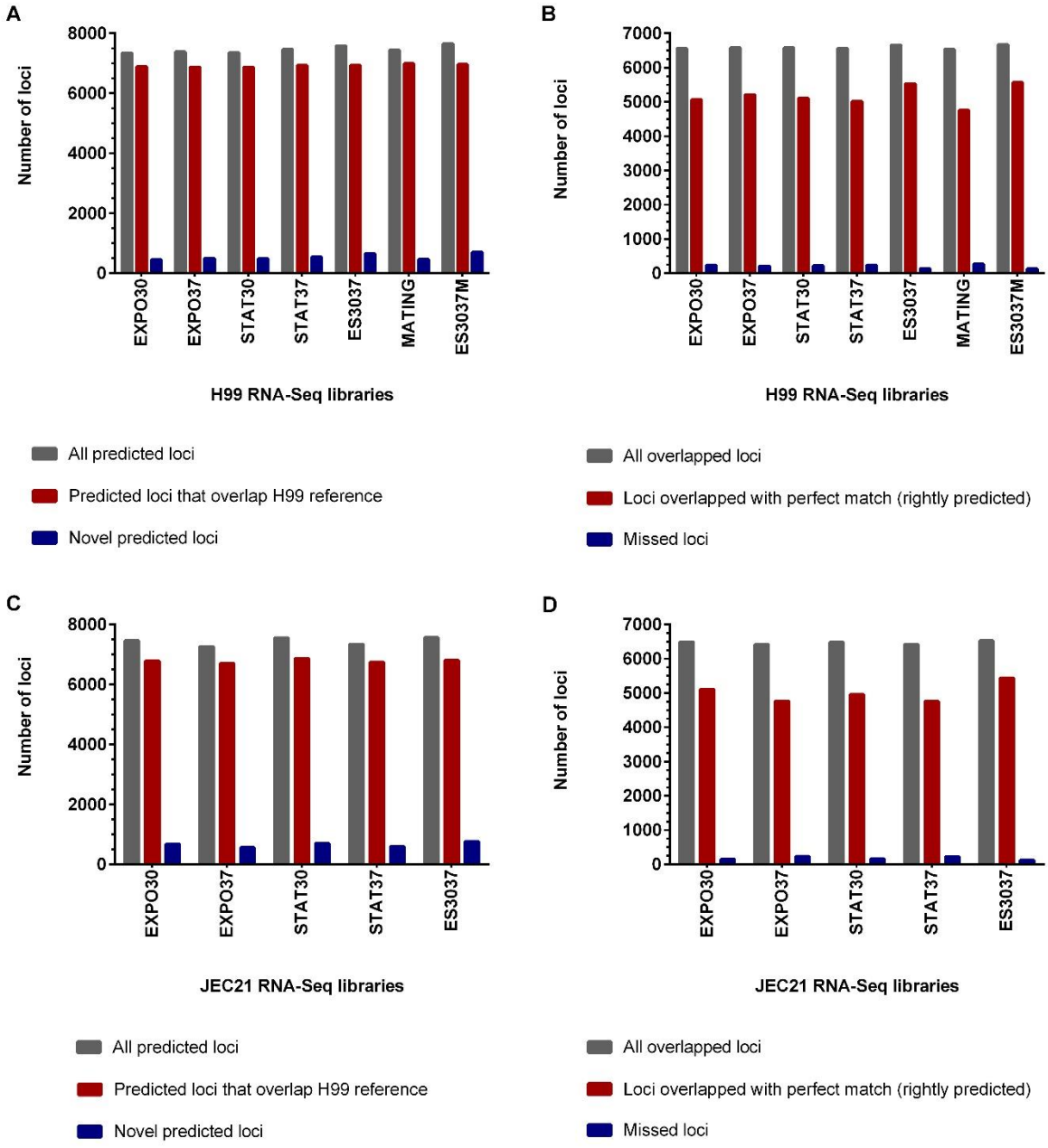

Figure 8. Prediction results obtained with C3Q pipeline and Cufflinks parameters combination Q applied to subsampled RNA-Seq libraries (7.5 million reads) from conditions of *C. neoformans* H99 and JEC21. Figures A and C: relative values to the entire predicted annotations, including hypothetical novel loci in H99 and JEC21. Figures B and D: number of genes from H99 and JEC21 reference annotations that are overlapped, missed and rightly predicted. ES3037: EXPO30 + EXPO37 + STAT30 + STAT37. ES3037M: EXPO30 + EXPO37 + STAT30 + STAT37 + MATING.

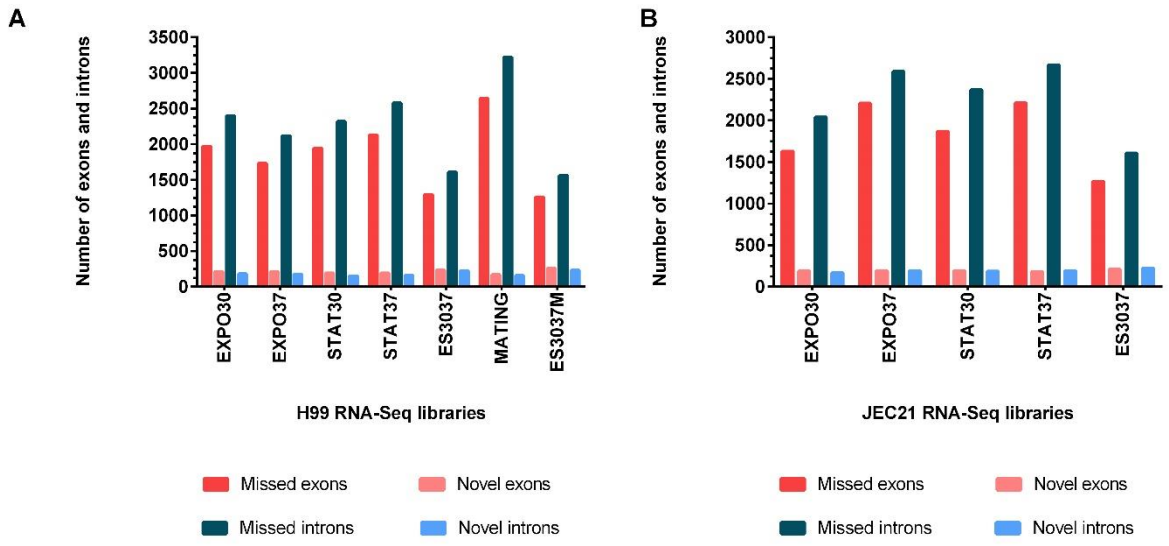

Figure 9. Specific prediction results obtained with C3Q pipeline and Cufflinks parameters combination Q applied to subsampled RNA-Seq libraries (7.5 million reads) from conditions of *C. neoformans* H99 (A) and JEC21 (B). Comparison with *C. neoformans* H99 and JEC21 reference annotations. Relative values of exons and introns for loci overlapped in reference annotations. ES3037: EXPO30 + EXPO37 + STAT30 + STAT37. ES3037M: EXPO30 + EXPO37 + STAT30 + STAT37 + MATING.

Interestingly, when separately evaluated the loci predicted by the RNA-Seq data and genome predicted loci, the effect of Cufflinks assembled transcripts in prediction accuracy is strongly seen. From 5336 RNA-Seq-predicted loci, only 5.2% is classified as novel; the other 94.8% (5060) overlap H99 reference annotation, reaching 94,4% of rightly predicted loci. For the genome-predicted loci, however, from 2843 predicted loci, 15.4% are identified as novel; and 2408 loci overlap H99 reference with a perfect match ratio of only 37.8%. As observed for H99, in JEC21 the difference between RNA-Seq and genome-predicted loci is still remarkable. From 5205 RNA-Seq predicted loci, 94.4% overlap JEC21 reference annotation, with a perfect match ratio of 93.9%. About the genome-predicted, in 2856 loci, 83.1% overlap JEC21 reference annotation, with a value of 38.9% of righty predicted loci.

The analysis of novel loci structures, especially those generated as genome-predicted, however, allow us to notice that due to absence of extrinsic reliable data, there is not a strict control about genome-based predictions.

Missed loci characterization

The improvement of the pipelines is associated with an increase of the sensitivity of gene identification. Whereas, in the initial C3Q setting, 447 reference genes were completely missed, only 159 of them are missed after the final data filtering step (Table S2). Despite the missed references have lower FPKM values in average than the predicted one (Figure S1), is not clear if the expression is the only reason for the absence of these sequences, since some very poorly expressed genes are correctly predicted and since CodingQuarry also generates genome-predicted sequences. The analysis of sequence and exons number, in turn, shows that missed genes are usually smaller and with less introns (Figure 10). Interestingly, a subset of genes (n=25) are missed in both strains (Table S3). Finally, the Pannzer2 analysis of the protein sequences encoded by the missed genes identified 16 proteins conserved domains, 7 of them being common to both.

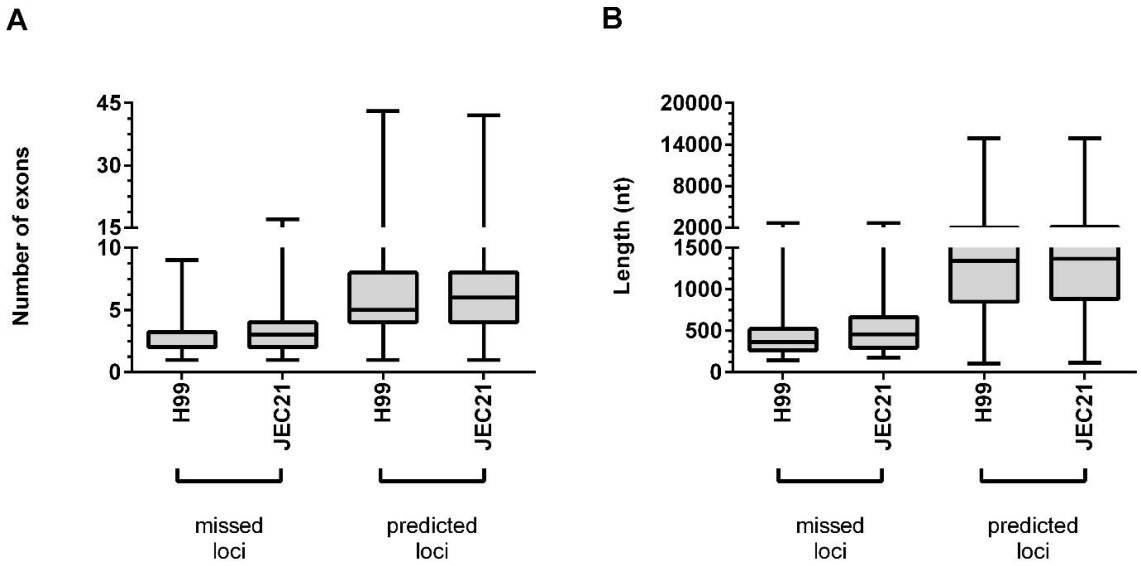

Figure 10. Gene length and number of exons comparison between missed and predicted loci from H99 and JEC21 reference annotations. Prediction results obtained with C3Q pipeline and Cufflinks parameters combination Q applied to RNA-Seq libraries from all conditions of *C. neoformans* H99 and JEC21 downsampled to 7.5 million of reads per replicate.

Genes filtering

As previously shown, each step improving the quality of the gene prediction was also associated with an increase of the number of predicted genes no present in the reference genome. After BAM subsampling strategy, the number of predicted genes not present in the reference genome was 703 and 762 for *C. neoformans* H99 and JEC21, respectively. Most of these genes can be considered as doubtful. In *C. neoformans* H99, only 61 of them has a domain signature as analyzed by Pannzer2 (Table S3). One can find 17 pseudogenes, and 30 genes bearing transposable elements sequences suggesting that they correspond to fragments of transposons or retrotransposons unannotated in the reference H99 genome. Similar results were

obtained in *C. neoformans* JEC21. In order to try to filter out some of these genes we looked at their structure and coverage. As Coding Quarry uses both the RNA-Seq data set and the sequence of the genome as information to make is gene structure prediction we considered here the genome predicted and RNA-Seq predicted gene independently. We compared these characteristics of this false positive genes to the ones of the reference genes. As shown on the Figure 11, genome predicted spliced sequences are usually longer (50% of them up to 459 nt in H99 and 580 nt in JEC21) than those from RNA evidence (50% of the transcripts up to 249 nt in H99 and 262 in JEC21).

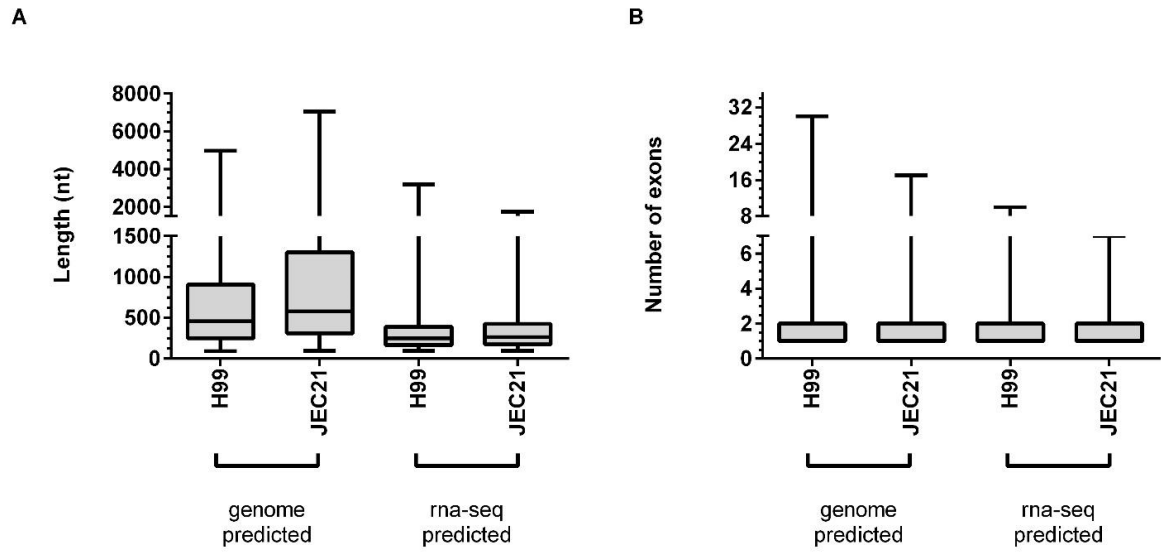

Figure 11. Gene length and number of exons comparison between genome-predicted and RNA-Seq predicted novel transcripts. Prediction results obtained with C3Q pipeline and Cufflinks parameters combination Q applied to RNA-Seq libraries from all conditions of *C. neoformans* H99 and JEC21 downsampled to 7.5 million of reads per replicate.

These predicted artefact genes are also less expressed than the reference genes (Figure S2). When evaluated the maximum FPKM value between replicates,

50% of these false positive genome predicted genes from H99 are up to 0.53; and JEC21 sequences up to 0.15, meaning that the most of them have few or no reads associated. The FPKM values from false positive RNA-Seq supported genes, as expected, are higher; H99 have expression values of 17.69 whereas JEC21 transcripts are up to 30.77. These predicted genes not present in the references genomes has also a different exon-intron structure. Whereas 224 (3.3%) genes in *C. neoformans* H99 and 191 (2.9%) genes in *C. neoformans* JEC21 may have a transcript that do not contain introns in their coding sequence, more than 50% of these predicted genes are intronless. Overall, these predicted genes not present in the annotation were short, poorly expressed and for a large part intronless. We used these three criteria alone or in combination to try to sort them out specifically as much as possible (Figures 12, 13 and Table S2). In both strains, the best combination of filter eliminates all the spliced coding region smaller than 150 nt, all the intronless genes smaller than 300 nt and all the genome-predicted genes not supported with any reads. With this setting, the number of predicted genes not present in the reference annotation was decreased from 703 (9.2%) to 408 (5.6%) in H99 (columns 1 and 8 from Figure 12); and, from 762 (10.1%) to 427 (5.9%) in JEC21.

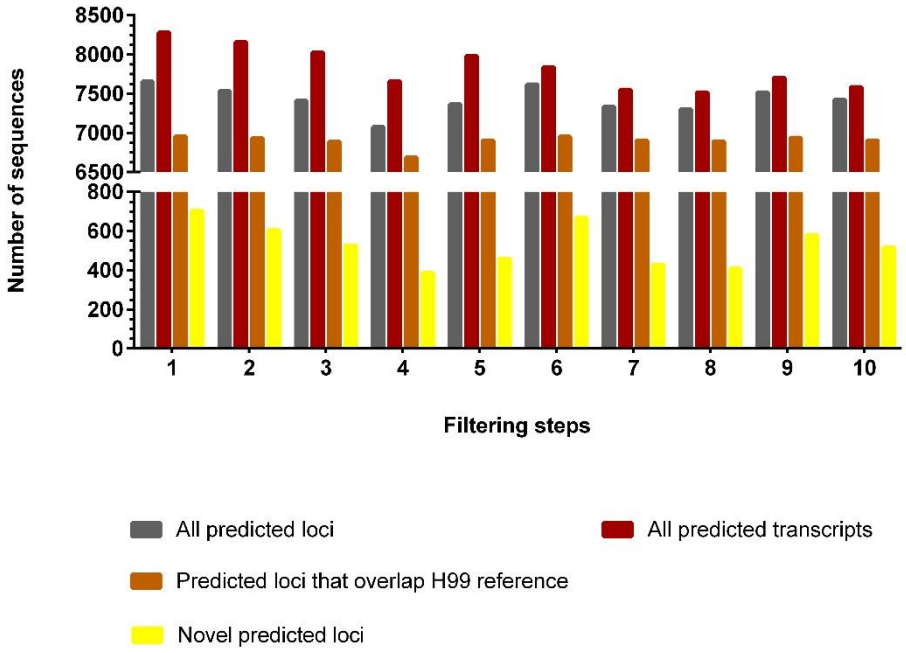

Figure 12. Number of genes and transcripts obtained after different filtering steps in C3Q gene prediction with *C. neoformans* H99. Comparison with *C. neoformans* H99 reference annotation. (1) Control (C3Q prediction with no sequence filter); (2) Filter of spliced sequences <150nt; (3) Filter of spliced sequences <200nt; (4) Filter of spliced sequences <300nt; (5) Filter of intronless sequences <300nt; (6) Filter of genome-predicted sequences with no reads; (7) Filter of genome-predicted with no reads and intronless sequences <300nt; (8) Filter of genome-predicted sequences with no reads, intronless sequences <300nt and spliced sequences <150nt; (9) Filter of genome-predicted sequences with no reads from EXPO30 condition; (10) Filter of genome-predicted sequences with FPKM <1.

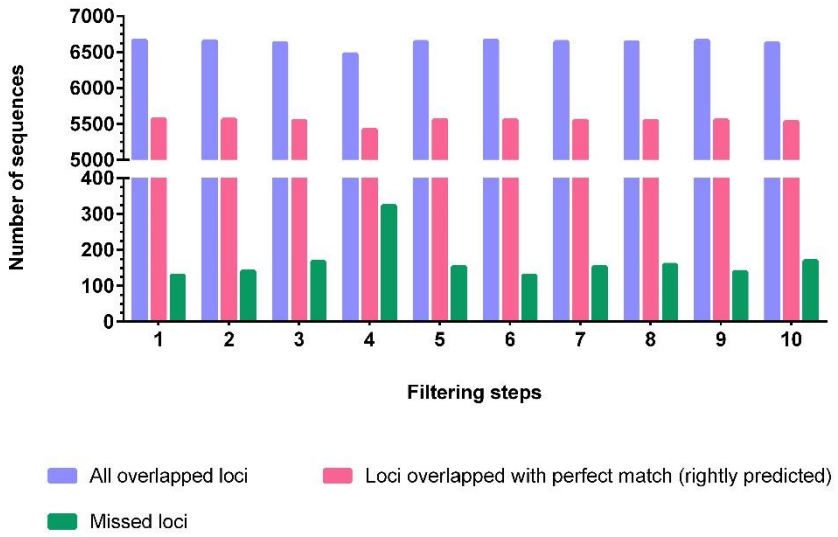

Figure 13. Number of genes from *C. neoformans* H99 reference annotation that are overlapped, missed and rightly predicted after different filtering steps in C3Q gene prediction. (1) Control (C3Q prediction with no sequence filter); (2) Filter of spliced sequences <150nt; (3) Filter of spliced sequences <200nt; (4) Filter of spliced sequences <300nt; (5) Filter of intronless sequences <300nt; (6) Filter of genome-predicted sequences with no reads; (7) Filter of genome-predicted with no reads and intronless sequences <300nt; (8) Filter of genome-predicted sequences with no reads, intronless sequences <300nt and spliced sequences <150nt; (9) Filter of genome-predicted sequences with no reads from EXPO30 condition; (10) Filter of genome-predicted sequences with FPKM <1.

Genome annotation with *C. gattii* R265

The genome annotation of *C. gattii* R265 was performed according established for C3Q pipeline: (i) Cufflinks transcript assembly with downsampled BAM file (7.5 million reads) for each replicate and “Q” parameters combination; (ii) merging of GTF files from each prediction with Cuffmerge; (iii) Coding Quarry annotation of Cufflinks assembled transcripts and ab initio prediction; (iv) filter of all transcript smaller than 150

nt; (v) filter of all intronless transcripts smaller than 300 nt; and (vi) filter of genome-predicted transcripts with no reads.

Additionally, to what we have done for the pipeline optimization, we kept only on transcript per gene and took the longest Cufflinks-predicted structure (or the longest genome-predicted structure for those that were not predicted by RNA-Seq). This strategy was applied due to the higher probability of RNA-Seq predicted genes reach right structures than genome-based predictions.

The comparison of gene predictions in different steps of the pipeline with the manually corrected genes from chromosomes 9 and 14 showed that after the filtering of sequences (without manual correction), 70.1% of the loci were rightly predicted and only 2.6% missed (Figure 14 and Table S4).

The great improvement of this strategy, however, was the deletion of suspicious transcripts. In this way, the 13.9% of novel loci presented in the first prediction was replaced by the value of 7.6% (Figure 15 and Table S4). It is important to notice that the comparison of genes from chromosomes 9 and 14 annotated by Broad, and Ferrareze and colleagues (2017), obtained 43.4% and 64.1% of perfect match with these same corrected structures (Figure 14 and Table S4).

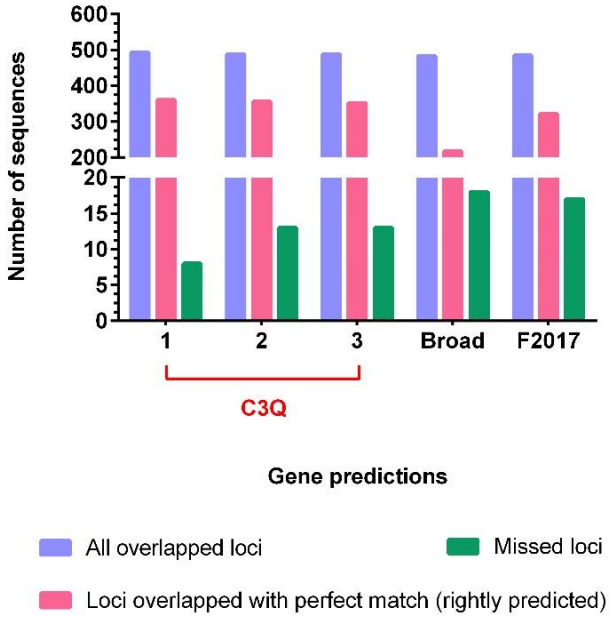

Figure 14. Number of genes from corrected annotation of *C. gattii* R265 (chromosomes 9 and 14) that are overlapped, missed and rightly predicted after different filtering steps in C3Q gene prediction (1, 2 and 3), and Broad and F2017 annotations. (1) Control (C3Q prediction with no sequence filter); (2) Prediction filtered for spliced sequences <150nt, intronless sequences <300 nt and genome-predicted sequences with no reads; (3) Prediction filtered according step 2 and multi-transcripts loci. Broad: reference genome annotation for *C. gattii* R265, updated to new genome coordinates with Exonerate. F2017: genome annotation published for *C. gattii* R265 published by Ferrareze and colleagues (2017), updated to new genome coordinates with Exonerate.

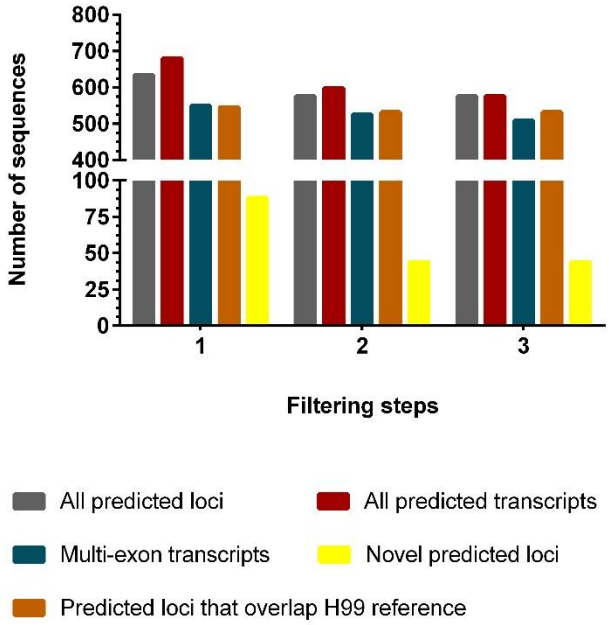

Figure 15. Number of genes and transcripts obtained after different filtering steps in C3Q gene prediction with *C. gattii* R265. Comparison with *C. gattii* corrected annotation (chromosomes 9 and 14). (1) Control (C3Q prediction with no sequence filter); (2) Prediction filtered for spliced sequences <150nt, intronless sequences <300 nt and genome-predicted sequences with no reads; (3) Prediction filtered according step 2 and multi-transcripts loci.

After, protein sequences from *C. neoformans* H99 and JEC21 were aligned against *C. gattii* R265 genome with Exonerate to retrieve exon-intron coordinates of homologous non-predicted loci. The comparison of coordinates with the predicted annotation identified 142 possible genes of non-predicted H99/JEC21 orthologs in R265 genome. Sequences in chromosomes 9 and 14 were not retrieved due to previous manual correction of these regions, posteriorly incorporated in gene prediction. Thus, 105 homologous sequences with functional ORFs were kept and annotated using this comparative strategy; of which 98 were already identified by Broad annotation.

Still, when evaluated all the prediction steps with reference annotations from Broad and Ferrareze and colleagues (2017); it is possible to observe the decrease of suspicious novel loci. Thereby, from 8057 transcripts predicted by C3Q pipeline in 7433 loci, 43.9% presented perfect match with Broad reference loci, 2.0% of the loci were missed and 7.8% was characterized as novel loci. When this same prediction was compared with the annotation of Ferrareze and colleagues (2017); this value was increased to 61.2% of perfect match and decreased to 7.1% of novel loci. After the filtering of transcripts, as expected, the 6917 loci presented 43.1% and 60.8% of perfect match with the annotations from Broad and Ferrareze et al., 2017. The decrease in novel loci values reached 4.2% and 3.7%, respectively, improving the prediction quality. The 3.3% of missed loci, in turn, were decreased to 1.8% by the recovery of H99/JEC21 homolog proteins in R265 genome (Table S4). So, the final annotation with transcripts filtering, non-predicted loci recovering and manual correction of chromosomes 9 and 14 (besides the merged genes), reached 43.9% and 62.4% of perfect match in the reference annotations from Broad and Ferrareze et al., 2017; moreover, 1.9% and 1.8% of missed loci, as well as, 3.9% and 3.4% of novel loci were reached with this pipeline.

A detailed comparison with Broad annotation revealed that 135 genes were predicted in 66 merged loci, of which, 43 (89 genes) were already merged by the previous annotation (Ferrareze et al., 2017). Two previously splitted genes (Ferrareze et al., 2017) were predicted and kept as different genes. The analysis of predicted merged genes led to the deletion of 11 fragmented loci and addition of 3 novel loci from 3 genes wrongly predicted as merged, reaching 6909 loci and 70.6% of perfect match (in comparison with corrected loci from chromosomes 9 and 14). From 151 analyzed

genes, 22 had its structure manually corrected. The addition of manually corrected sequences from chromosomes 9 and 14 led to the exclusion of false coding transcripts, of which, eight sequences were reclassified as lncRNAs (CNBG_5925, CNBG_5928, CNBG_5945, CNBG_5793, CNBG_9465) and pseudogenes (CNBG_5609, CNBG_5743, CNBG_6026).

Overall, this pipeline annotated 6941 coding genes in the *C. gattii* genome, with a number of 38968 exons and 32027 introns (Table S4). Of these, 269 were completely absent in the Broad annotation (not included the 2 novel coding sequences from the 2 splitted genes). Other 33 genes were already predicted by Ferrareze and coleagues (2017) and had CNBG Ids between 9680 and 9717; the hypothetical novel loci had their CNBG Ids above 10000. About 5811 genes (identified CNBG genes) were annotated with a single locus. A number of 147 genes were merged in 72 loci (as previously described); and 373 genes were wrongly predicted by 787 loci, due to transcript fragmentation; these loci were not merged (Table S4).

The orthology analysis between *C. neoformans* H99 and JEC21 and *C. gattii* R265 assigned 19361 genes (95.0% of total) to 6381 orthogroups (Table S5). There were 5960 orthogroups with all species present and 5768 of these consisted entirely of single-copy genes (the fragmented genes from R265 may underestimate this value).

From 6795 H99 genes, 6544 (96.3%) were assigned to orthogroups, and 251 (3.7%) are unique to *C. neoformans* H99. JEC21 had 6483 genes assigned to orthogroups (97.7%) and 156 unique genes (2.3%). R265, in turn, had 6334 (91.3%) genes in orthogroups and 607 (8.7%) unassigned genes. H99 and JEC21 share 6194 orthogrous, of which 238 were exclusive for *C. neoformans* species. R265 shares 6089

orthogroups with H99 and 6018 orthogroups with JEC21. In the three genomes, the most of the unassigned genes had no known function, being classified as uncharacterized proteins. However, the orthogroup exclusively shared between H99 and JEC21 is enriched with functions related to transmembrane transport, regulation of transcription, proteolysis and oxidation-reduction process; presenting specific gene ontology terms as gene silencing by RNA.

Lastly, the comparison between the gene length of *C. gattii* R265 and *C. neoformans* H99/JEC21 orthologous genes shows that the most of them were predicted with similar gene lengths. Thus, 5938 single-copy genes and 306 multiple-copies genes were analyzed (Table S5). About the single-copy genes, 5185 sequences presented a ratio ≥0.8 and ≤1.2, when compared with H99 ortholog sequence lengths (Figure 16 and Table S6). Multiple-copies genes, in turn, presented 216 sequences with similar length to orthologs (Figure 17 and Table S6). The presence of a moderate number of “small” sequences (ratio <0.8) (Figure 18 and Table S6), however, may be due the wrong prediction of partial gene structures, as well as, an intrinsic characteristic of the R265 ortholog or a predicted pseudogene. Nevertheless, all these data about R265 prediction indicates the quality of the prediction pipeline to this species.

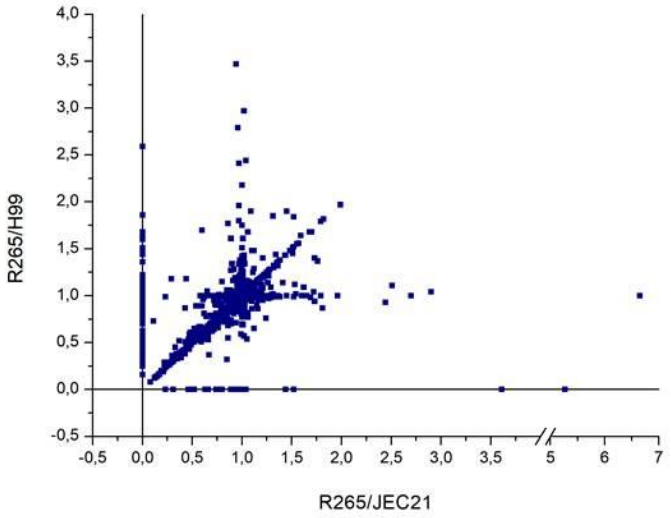

Figure 16. The ratios of the single copy orthologues (5938). Values next to 1 means a similar length of the predicted R265 genes and H99 or JEC21 orthologues. Values in the X 0.0 axis are from R265 genes without H99 orthologues. Values in Y 0.0 axis are from R265 genes without JEC21 orthologues.

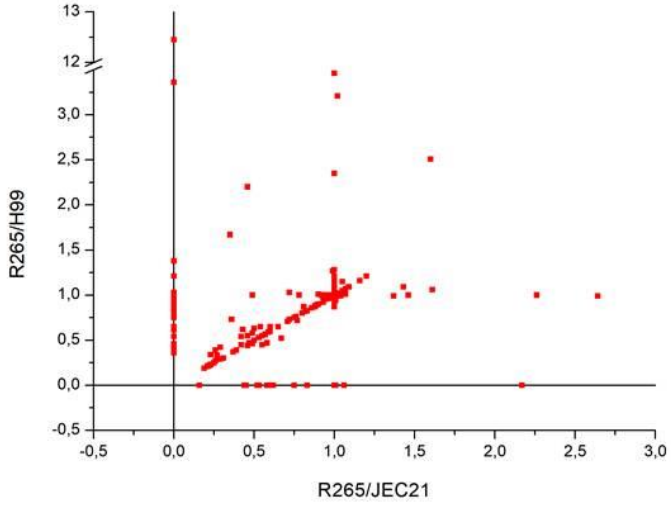

Figure 17. The ratios of the multiple copy orthologues (306; paralogues excluded and fragmented genes kept). Values next to 1 means a similar length of the predicted R265 genes and H99 or JEC21 orthologues. Values in the X 0.0 axis are from R265 genes without H99 orthologues. Values in Y 0.0 axis are from R265 genes without JEC21 orthologues.

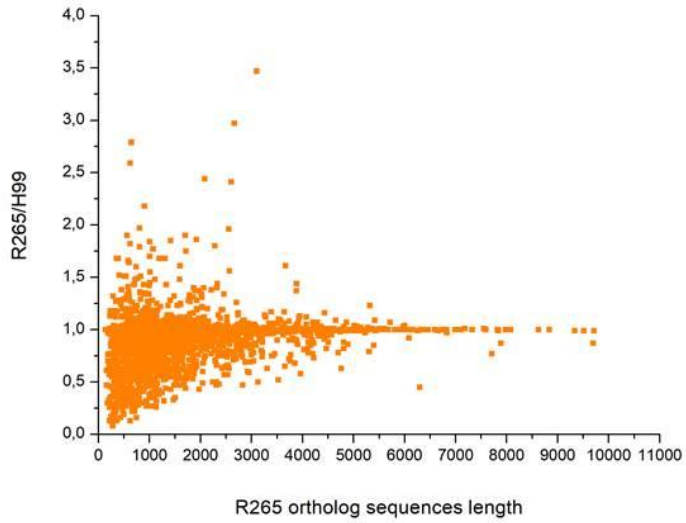

Figure 18. The relation of ratios of the single copy orthologues (R265/H99) and the length of R265 sequences. Values next to 1 means a similar length of the predicted R265 genes and H99 orthologues.

CONCLUSION

The implementation of an automated pipeline for gene prediction in intron-rich and high gene density genomes is a real challenge for the most of the prediction systems. However, the optimization of prediction parameters, the evaluation of RNA-Seq conditions and the library sizes may facilitate this task, in order to identify *Cryptocooccus* genomic characteristics applicable to all species; as well as, the best utilization of the transcriptomic data. Despite *C. neoformans* and *C. gattii* presented about 10% of difference in prediction accuracy; the utilization of *C. neoformans*, the sibling species of *C. gattii*, still represents the best way of optimization and validation, without the exhaustive work demanded by manual annotations.

REFERENCES

CARVER, Tim et al. Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. **Bioinformatics**, [s. l.], v. 28, n. 4, p. 464–469, 2012.

D'SOUZA, C. A. et al. Genome Variation in Cryptococcus gattii , an Emerging Pathogen of Immunocompetent Hosts. **mBi**o, [s. l.], v. 2, n. 1, p. 1–11, 2011.

EMMS, David M.; KELLY, Steven. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. **Genome Biology**, [s. l.], v. 16, n. 1, p. 1–14, 2015.

FARRER, Rhys A. et al. Genome evolution and innovation across the four major lineages of cryptococcus gattii. **mBio**, [s. l.], v. 6, n. 5, p. 1–12, 2015.

GONZALEZ-HILARION, Sara et al. Intron retention-dependent gene regulation in Cryptococcus neoformans. **Scientific Reports**, [s. l.], v. 6, p. 1–13, 2016.

HOFF, Katharina J. et al. BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. **Bioinformatics**, [s. l.], v. 32, n. 5, p. 767–769, 2016.

JANBON, Guilhem et al. Analysis of the Genome and Transcriptome of Cryptococcus neoformans var. grubii Reveals Complex RNA Expression and Microevolution Leading to Virulence Attenuation. **PLoS Genetics**, [s. l.], v. 10, n. 4, 2014.

KIM, Daehwan et al. TopHat2 accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. **Genome Biol**, [s. l.], v. 14, p. 14, 2013.

LANGMEAD, Ben; SALZBERG, Steven L. Fast gapped-read alignment with Bowtie 2. **Nature Methods**, [s. l.], v. 9, n. 4, p. 357–359, 2012.

LOMSADZE, Alexandre; BURNS, Paul D.; BORODOVSKY, Mark. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. **Nucleic Acids Research**, [s. l.], v. 42, n. 15, p. 1–8, 2014.

MARTIN, Marcel. Cutadapt removes adapter sequences from high-throughput sequencing reads. **EMBnet.journal; Vol 17, No 1: Next Generation Sequencing Data AnalysisDO - 10.14806/ej.17.1.200** , [s. l.], 2011.

MCGETTIGAN, Paul A. Transcriptomics in the RNA-seq era. **Current Opinion in Chemical Biology**, [s. l.], v. 17, n. 1, p. 4–11, 2013.

SCHNEIDER, Rafael de Oliveira et al. Zap1 regulates zinc homeostasis and modulates virulence in Cryptococcus gattii. **PLoS ONE**, [s. l.], v. 7, n. 8, 2012.

SLATER, Guy St C.; BIRNEY, Ewan. Automated generation of heuristics for biological sequence comparison. **BMC Bioinformatics**, [s. l.], v. 6, p. 1–11, 2005.

STANKE, Mario et al. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. **Bioinformatics**, [s. l.], v. 24, n. 5, p. 637–644, 2008.

TESTA, Alison C. et al. CodingQuarry: Highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. **BMC Genomics**, [s. l.], v. 16, n. 1, p. 1–12, 2015.

TRAPNELL, Cole et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. **Nature Biotechnology**, [s. l.], v. 28, n. 5, p. 511–515, 2010.

WALLACE, Edward et al. Start codon context controls translation initiation in the fungal kingdom. **bioRxiv**, [s. l.], p. 654046, 2019.

4 DISCUSSÃO GERAL

Diferentemente de *C. neoformans* e dos genótipos VGI, VGIII e VGIV de *C. gattii*, as linhagens VGII perderam numerosas famílias gênicas durante o processo evolutivo, dentre as quais, componentes da maquinaria do RNA de interferência. A ausência de RNAi funcional, em diversos organismos, tem sido associada a fenótipos de virulência e patogenicidade. Em *C. gattii* VGII, especificamente, a prevalência de linhagens deste genótipo, associada ao surto do Noroeste do Pacífico, levanta questões a respeito da relação entre estabilidade genômica e adaptação. Neste contexto, a ausência de clusters e a baixa prevalência de sequências completas de retrotransposons, encontrados por este estudo, especulam o mecanismo regulatório que controla a propagação destes elementos em *C. gattii* R265. Corroborando estes achados, o estudo de Yadav e colaboradores (2018) identifica a presença e impacto de retrotransposons Tcn para a evolução das regiões centroméricas dos genomas de *Cryptococcus*. Assim, presença de fragmentos de retrotransposons, bem como a maior expressão destes elementos *em C. gattii* R265, sugere eventos passados de propagação, recombinação e consequente fragmentação (inativação); de forma que retrotransposons truncados representam a consequência das taxas aumentadas de recombinação não homóloga entre tais sequências.

Por outro lado, a redução do tamanho genômico de *C. gattii* (em aproximadamente 1.3 Mb), decorrente dos impactos da ausência de um sistema de defesa, pode indicar um processo recente de estabilização do genoma pela inativação destes elementos de transposição. Aliado a isso, encontra-se a manutenção das

relações genéticas mais próximas entre linhagens de mesmo genótipo, as quais não parecem ter sido efetivamente afetadas após a divergência. Embora VGII seja considerado o genótipo basal de *C. gattii*, acumulando mutações e recombinações não presentes em outros genótipos, suas subpopulações VGIIa, VGIIb, VGIIc e VGIIx são altamente clonais e não muito diversas (BIELSKA; MAY, 2015). Além disso, tanto análises de composição genômica quanto filogenias baseadas em ortologia mostram que os quatro genótipos de *C. gattii* estão relativamente isolados desde a divergência (cerca de 12 milhões de anos) (FARRER et al., 2015).

Outros fatores como a ausência de uma metiltransferase DNMT5, também hipotetiza as consequências da perda de mecanismos de regulação para a plasticidade genômica e virulência, visto que estudos anteriores sugerem a relação entre a metilação do DNA e a repressão da recombinação e propagação de retrotransposons em diversas espécies (YADAV et al., 2018).

Todavia, conforme explicado anteriormente, a ausência de dados completos e específicos para uma anotação gênica ampla e voltada às especificidades do genoma de *C. gattii* R265, levam à produção de análises transcritômicas parciais. Com o intuito de produzir uma predição gênica atualizada e corrigida, foi realizada a anotação manual do genoma de *C. gattii* R265, com auxílio dos programas de montagem e predição, e bibliotecas *unstranded single-end* de RNA-Seq em 3 condições (WT e mutante nulo ZIP1 em YPD; e células recuperadas de lavado brônquioalveolar).

Através da avaliação do transcritoma de *C. gattii* pode-se identificar o enriquecimento funcional associado ao metabolismo primário. Assim, a predição de vias metabólicas expressas em *C. gattii* durante a infecção, pela visualização ordenada

de interações proteicas e valores de expressão, constituiu uma importante ferramenta de análise. Como resultado, vias metabólicas já reconhecidamente associadas à interação patógeno-hospedeiro puderam ser observadas, como as vias de metabolismo de carboidratos, lipídeos e aminoácidos, bem como rotas de transporte e catabolismo. Entretanto, a ausência de informação sobre interações proteicas para mais da metade dos genes de *C. gattii* R265 impede a avaliação completa do perfil metabólico deste organismo durante a patogênese.

A análise de fatores de virulência, por sua vez, permitiu a identificação de genes altamente transcritos, dentre os quais, uma metaloproteinase elastinolítica extracelular e a proteína antioxidante tiol-específica 1. Comumente, proteinases são secretadas para a clivagem de proteínas e captação das mesmas pelas células fúngicas. Em *C. neoformans*, especificamente, a atividade de proteinases secretadas leva à degradação da matriz extracelular e colágeno; processo que estabiliza a infecção no sistema nervoso central pela permeabilização das barreiras epiteliais (GARBE; VYLKOVA, 2019).

A comparação dos valores de FPKM com a mediana obtida pelo transcritoma (98.96), possibilitou, também, uma avaliação geral de genes altamente expressos, considerando que 50% dos genes analisados obtiveram valores de expressão entre 10 e 100. Assim, quando quantificados os transcritos abundantes, pode-se observar a presença de transportadores, como o transportador de ribonucleotídeos CNBG_4811, e o transportador de glicose CNBG_0170; além de diversos transportadores de aminoácidos e permeases. Semelhantemente, o estudo da resposta transcricional de *C. neoformans* à fagocitose por macrófagos e por *Amoebae castellanii*, identificou o

gene CNAG_03772 (ortólogo de CNBG_0170), com modulação similar em ambas as condições, quando comparadas ao controle; sendo em macrófagos, contudo, quase duas vezes maior (DERENGOWSKI et al., 2013). Tal diferença pode evidenciar as especificidades da imunidade nutricional conferida por macrófagos e organismos predadores.

Conforme observado em outras espécies, a fagocitose leva a importantes alterações metabólicas na célula fúngica. O fator de transcrição Gat201, por exemplo, atua na inibição da fagocitose por macrófagos devido a sua influência na produção da cápsula. Gat201 controla a expressão de mais de 1000 genes em *C. neoformans*, dentre os quais, Blp1, Gat204 e Cir1 (CHUN; BROWN; MADHANI, 2011). Em *C. gattii*, Blp1 (CNBG_5109) apresenta alta expressão; Cir1, contudo, não sinaliza por si só a disponibilidade de ferro durante a patogênese, porquanto os níveis de mRNAs associados não são influenciados pela concentração de ferro, mas a estabilidade proteica do fator de transcrição (JUNG; KRONSTAD, 2011). Ausente na anotação anterior, um possível ortólogo de Gat204 foi predito pela pipeline automática de anotação (CNBG_10571).

Igualmente, a avaliação do gene da isocitrato liase (ICL1) de *C. gattii* (CNBG_5726) e *C. neoformans* demonstra a expressão aumentada desta enzima em *Cryptococcus* fagocitados por macrófagos. Nas duas espécies, ICL1 apresenta um valor aproximadamente cinco vezes maior que a malato sintase, enzima da mesma via metabólica, que atua no passo seguinte de conversão do glioxilato em malato (DERENGOWSKI et al., 2013). Estudos transcritômicos de *C. neoformans in vivo* e *ex vivo* identificaram um *fold change* de até quarenta vezes, se comparados a condições

in vitro (CHEN et al., 2014b). A função principal do ciclo do glioxilato é a utilização de compostos C_2 como única fonte de carbono; e a mesma é induzida após a fagocitose, possivelmente devido à restrição nutricional dentro do fagossomo (LORENZ; FINK, 2002), como reportado para *C. albicans* e *A. fumigatus* (CHEAH; LIM; SANDAI, 2014).

Embora células de *C. neoformans* fagocitadas por macrófagos apresentem superexpressão de genes do ciclo do glioxilato e de transportadores de açúcares, o que sugere que este organismo possua mecanismos alternativos de aquisição de carbono, como a gliconeogênese e a utilização de lipídeos como fonte de carbono, mutantes da ICL1 não apresentam alteração na virulência, tal qual os patógenos *C. albicans* e *Mycobacterium tuberculosis* (CHEAH; LIM; SANDAI, 2014). Entretanto, diferentemente de outras espécies até então estudadas, em *C. neoformans*, ICL1 parece desempenhar uma importante função adicional durante a restrição de nitrogênio.

Apesar da quantificação de expressão do fator de transcrição Gat1 e das permeases de amônia não indicar restrição de nitrogênio durante a infecção, a alta expressão da permease de GABA (CNBG_1602), bem como a baixa expressão de Ure2p (CNBG_2927) podem sugerir a restrição de fontes preferenciais de nitrogênio, como glutamina e amônia. Como discutido no capítulo 3.2, Amt2 responde a limitação de amônia ou de fontes alternativas de nitrogênio, como aminoácidos; Ure2p é induzido pela concentração de nitrogênio, e GABA é considerada uma fonte pobre de nitrogênio, cujas permeases não são expressas na presença de amônia. Como a utilização de fontes alternativas de nitrogênio, como nitrato, nitrito, ureia ou aminoácidos resulta na formação de amônia (a qual é convertida em glutamato e,

posteriormente, glutamina); é a glutamina, o efetor da NCR, para garantir a utilização das fontes preferenciais de nitrogênio (TUDZYNSKI, 2014).

Ademais, a elevada expressão dos transportadores de aminoácidos da superfamília APC (AAP1, AAP2 e AAP4), corrobora a hipótese de limitação de fontes primárias de nitrogênio. Em *C. neoformans*, AAP2, AAP4/5 e AAP8 são altamente induzidos na presença de aminoácidos como única fonte de nitrogênio; sendo os mesmos (com exceção de AAP4/5) reprimidos pela adição de sulfato de amônia (GARBE; VYLKOVA, 2019). A regulação positiva de transportadores de aminoácidos durante a infecção pode indicar, dessa forma, que tal mecanismo é possível durante a permanência no fagossomo; ou que *C. neoformans* não sofre restrição nutricional prolongada devido à ruptura da membrana fagossomal (COELHO; BOCCA; CASADEVALL, 2014).

O enriquecimento das vias de biossíntese de aminoácidos (fenilalanina, tirosina e triptofano; valina, leucina e isoleucina) em *C. gattii*, por sua vez, pode estar associado à incapacidade do sistema de aquisição durante a restrição nutricional; visto que as variações de expressão de permeases não são suficientes para manutenção dos níveis destes aminoácidos. A aquisição de triptofano, especialmente, é sujeita à ausência de fontes preferenciais de nitrogênio, devido à ausência de permeases de alta especificidade em *Cryptococcus*, tornando esta via biosintética essencial. Juntamente com a via de biossíntese da treonina, mutações na via biosintética do triptofano são letais à célula fúngica. Adicionalmente, efeitos de hipersensibilidade ao soro, suscetibilidade a drogas antifúngicas, menor tolerância ao estresse e atenuação da virulência, são desencadeados pela deleção de genes das vias biosintéticas da

lisina, isoleucina, valina e metionina; indicando a importância destes aminoácidos para a sobrevivência de *Cryptococcus* (FERNANDES et al., 2015).

Portanto, a maior expressão de genes associados à conversão de glutamato e liberação de amônia, bem como o enriquecimento de processos associados a produção de intermediários do metabolismo de carboidratos, através do catabolismo de aminoácidos, podem sugerir que processos de degradação de aminoácidos captados, para posterior biossíntese de novas moléculas, seja um dos mecanismos chave para suplantar a limitação nutricional.

Embora a correção manual tenha aperfeiçoado a predição gênica de *C. gattii* R265, possibilitando uma análise mais confiável do perfil transcritômico desta linhagem, o novo sequenciamento e montagem do genoma, juntamente à possibilidade de utilização de um conjunto mais amplo de dados de RNA-Seq *stranded* e *paired-end*, levaram à idealização de uma pipeline automatizada e otimizada para *Cryptococcus*. Neste contexto, a seleção inicial dos preditores levou em consideração os dados de validação disponíveis na literatura e que melhor se aplicassem às características gênicas de *Cryptococcus*. Posteriormente, *C. neoformans* H99 e JEC21 foram utilizados como referências para a validação e otimização da pipeline de predição.

Assim, conforme avaliação dos dados disponíveis nos artigos de publicação dos preditores CodingQuarry e BRAKER1, a comparação entre as predições geradas pelos preditores CodingQuarry, AUGUSTUS e TransDecoder, para as espécies fúngicas *S. cerevisiae* e *S. pombe* (contra um conjunto de genes de referência), mostra que CodingQuarry apresenta maior sensibilidade em todos os parâmetros da predição;

bem como maior precisão em nível de gene e introns. Diferentemente de *S. cerevisiae* e *S. pombe*, os quais apresentam entre 0,04 e 0,66 introns por quilobase de CDS (média de 1,08 e 2,09 introns por gene, respectivamente), *Cryptococcus* apresenta alta frequência e densidade de introns (STAJICH; DIETRICH; ROY, 2007). Como descrito na anotação de *C. neoformans* H99 e JEC21 publicada por Wallace e colaboradores (2019), 6795 genes codificantes foram preditos para H99 e 6639 para JEC21, os quais ocupam 85% do genoma e 99.5% possuem ao menos um intron; além de uma densidade intrônica de 3,35 introns/kb CDS e 5,7 introns por gene.

Ademais, dados da literatura observam que em genomas mais densos em genes (como *Cryptococcus*), as curtas regiões intergênicas levam a anotação de falsos introns, e a consequente fusão errônea de sequências pelo preditor AUGUSTUS (TESTA et al., 2015). Semelhantemente, CodingQuarry apresenta maior sensibilidade na predição gênica de *S. pombe*, em comparação com BRAKER1 e MAKER (HOFF et al., 2016). Com base nestes dados, BRAKER1 e CodingQuarry (associado a Cufflinks) foram selecionados para avaliação e validação com *C. neoformans* H99 e JEC21. Todavia, embora *C. neoformans* e *C. gattii* apresentem características semelhantes, em relação a densidade gênica e frequência e tamanho de regiões intrônicas e exônicas, uma diferença de aproximadamente 10% na acurácia da predição gênica, foi mensurada. A utilização de bibliotecas de RNA-Seq em triplicata, oriundas de cinco diferentes condições (assim como H99), obteve cerca de 71,7% dos genes de referência corretamente anotados (comparação com os genes preditos e manualmente corrigidos dos cromossomos 9 e 14). Embora a filtragem de sequências tenha diminuído este valor para 70,1%, a comparação com os dados disponíveis, de

espécies com moderada ou baixa frequência de introns, corrobora a otimização dos parâmetros como ferramenta para a aplicação deste sistema em transcritomas mais complexos. Outrossim, a comparação parcial com estruturas gênicas de apenas dois cromossomos, os quais representam apenas 8% do genoma de *C. gattii* R265 pode gerar valores de validação específicos a características destas sequências, não abrangendo os valores reais de uma comparação com todo o genoma anotado.

Adicionalmente, a predição de CDS de tamanho similar aos ortólogos de *C. neoformans*, para 87% dos genes de ortogrupos de cópia única, possibilita uma segunda validação deste protocolo; embora o tamanho da sequência não garanta a predição correta das junções introns-exons. Por fim, como esperado, *C. gattii* R265 não apresenta diversos genes; contudo, a predição de novos genes hipotéticos originou 167 (de 268) sequências que posteriormente foram identificadas como única cópia em grupos ortólogos com *C. neoformans* (além dos 33 novos previamente descritos durante a anotação manual do genoma e análise transcritômica).

5 CONCLUSÃO

Como conclusão, o presente estudo demonstrou que o menor número de retrotransposons em genomas do genótipo VGII de *C. gattii*, bem como a maior fragmentação dos mesmos e presença de regiões de quebra de sintenia, podem estar associados ao aumentado número de eventos de transposição, decorrentes da ausência de RNAi funcional. No entanto, a presença de sRNAs associados a regiões intergênicas e LTRs pode sugerir um mecanismo não canônico de regulação destas sequências.

A alta expressão de transcritos de transportadores de glicose e aminoácidos, por sua vez, surge como resposta à limitação nutricional de *C. gattii* R265 durante a infecção pulmonar em modelo murino. Ainda, a expressão aumentada de genes relacionados à degradação de moléculas mais complexas para a obtenção de intermediários do metabolismo de carboidratos e fontes de nitrogênio, denota meios alternativos de aquisição de nutrientes essenciais.

Por fim, a aplicação de uma pipeline de predição gênica otimizada para *Cryptococcus* levou à obtenção de uma anotação gênica de *C. gattii* R265, com mais de 70% de acurácia das estruturas gênicas; as quais, mantém tamanho predito semelhante (razão de 0.8 a 1.2), para aproximadamente 87% dos genes com ortólogos em *C. neoformans* (91.3%). Outrossim, *C. gattii* R265 apresenta um grande número de genes únicos e reduzido número de parálogos, os quais podem fornecer dados importantes sobre a virulência e a importância da estabilidade genômica e RNAi para a evolução e adaptação da espécie.

6 REFERÊNCIAS

ALZOHAIRY, Ahmed M. et al. Transposable elements domesticated and neofunctionalized by eukaryotic genomes. **Plasmid**, [s. l.], v. 69, n. 1, p. 1–15, 2013. Disponível em: <http://dx.doi.org/10.1016/j.plasmid.2012.08.001>

ANDERS, Simon; PYL, Paul Theodor; HUBER, Wolfgang. HTSeq-A Python framework to work with high-throughput sequencing data. **Bioinformatics**, [s. l.], v. 31, n. 2, p. 166–169, 2015.

ARKHIPOVA, Irina R. Distribution and phylogeny of penelope-like elements in eukaryotes. **Systematic Biology**, [s. l.], v. 55, n. 6, p. 875–885, 2006.

ARMAS-TIZAPANTZI, Anahí; MONTIEL-GONZÁLEZ, Alba Mónica. RNAi silencing: A tool for functional genomics research on fungi. **Fungal Biology Reviews**, [s. l.], v. 30, n. 3, p. 91–100, 2016.

AYLWARD, Janneke et al. A plant pathology perspective of fungal genome sequencing. **IMA Fungus**, [s. l.], v. 8, n. 1, p. 1–45, 2017.

BHADAURIA, Vijai et al. Fungal transcriptomics. **Methods in Molecular Biology**, [s. l.], v. 1775, p. 83–92, 2007.

BIELSKA, Ewa; MAY, Robin C. What makes Cryptococcus gattii a pathogen? **FEMS Yeast Research**, [s. l.], v. 16, n. 1, p. 1–12, 2015.

BYRNES, Edmond J. et al. Emergence and pathogenicity of highly virulent Cryptococcus gattii genotypes in the Northwest United States. **PLoS Pathogens**, [s. l.], v. 6, n. 4, p. 1–16, 2010.

CARRERAS-VILLASEÑOR, Nohemi et al. The RNAi machinery regulates growth and development in the filamentous fungus Trichoderma atroviride. **Molecular Microbiology**, [s. l.], v. 89, n. 1, p. 96–112, 2013.

CARVER, Tim et al. Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. **Bioinformatics**, [s. l.], v. 28, n. 4, p. 464–469, 2012.

CASTANERA, Raúl et al. Transposable Elements versus the Fungal Genome: Impact on Whole-Genome Architecture and Transcriptional Profiles. **PLoS Genetics**, [s. l.], v. 12, n. 6, p. 1–27, 2016.

CHANG, Zanetta et al. Broad antifungal resistance mediated by RNAi-dependent epimutation in the basal human fungal pathogen Mucor circinelloides. **PLoS Genetics**, [s. l.], v. 15, n. 2, p. 1–17, 2019.

CHEAH, Hong Leong; LIM, Vuanghao; SANDAI, Doblin. Inhibitors of the glyoxylate cycle enzyme ICL1 in Candida albicans for potential use as antifungal agents. **PLoS ONE**, [s. l.], v. 9, n. 4, 2014.

CHEN, Rui et al. Exploring microRNA-like small RNAs in the filamentous fungus Fusarium oxysporum. **PLoS ONE**, [s. l.], v. 9, n. 8, 2014. a.

CHEN, Yuan et al. The Cryptococcus neoformans Transcriptome at the Site of Human Meningitis. **mBio**, [s. l.], v. 5, n. 1, 2014. b. Disponível em: <http://go-galegroup.ez67.periodicos.capes.gov.br/ps/i.do?id=GALE%7CA426369118&v=2.1&u=capes&it=r&p=AONE&sw=w>

CHEN, Yun et al. Characterization of RNA silencing components in the plant

pathogenic fungus Fusarium graminearum. **Scientific Reports**, [s. l.], v. 5, n. June, p. 1–13, 2015. Disponível em: <http://dx.doi.org/10.1038/srep12500>
CHÉNAIS, Benoît et al. The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. **Gene**, [s. l.], v. 509, n. 1, p. 7–15, 2012. Disponível em: <http://dx.doi.org/10.1016/j.gene.2012.07.042>
CHOI, JaeJin; KIM, Sung-Hou. A genome Tree of Life for the Fungi kingdom. **Proceedings of the National Academy of Sciences**, [s. l.], v. 114, n. 35, p. 9391–9396, 2017.
CHUJO, Tetsuya; SCOTT, Barry. Histone H3K9 and H3K27 methylation regulates fungal alkaloid biosynthesis in a fungal endophyte-plant symbiosis. **Molecular Microbiology**, [s. l.], v. 92, n. 2, p. 413–434, 2014.
CHUN, Cheryl D.; BROWN, Jessica C. S.; MADHANI, Hiten D. A major role for capsule-independent phagocytosis-inhibitory mechanisms in mammalian infection by Cryptococcus neoformans. **Cell Host and Microbe**, [s. l.], v. 9, n. 3, p. 243–251, 2011.
COELHO, Carolina; BOCCA, Anamelia Lorenzetti; CASADEVALL, Arturo. **The Tools for Virulence of Cryptococcus neoformans**. 1. ed. [s.l.] : Elsevier Inc., 2014. v. 87 Disponível em: <http://dx.doi.org/10.1016/B978-0-12-800261-2.00001-3>
COOK, David E. et al. Long-Read Annotation: Automated Eukaryotic Genome Annotation Based on Long-Read cDNA Sequencing. **Plant Physiology**, [s. l.], v. 179, n. 1, p. 38–54, 2019.
D'SOUZA, C. A. et al. Genome Variation in Cryptococcus gattii , an Emerging Pathogen of Immunocompetent Hosts. **mBio**, [s. l.], v. 2, n. 1, p. 1–11, 2011.
DANG, Yunkun et al. RNA interference in fungi: Pathways, functions, and applications. **Eukaryotic Cell**, [s. l.], v. 10, n. 9, p. 1148–1155, 2011.
DEAN, Ralph A. et al. The genome sequence of the rice blast fungus Magnaporthe grisea. **Nature**, [s. l.], v. 434, n. 7036, p. 980–986, 2005.
DELEON-RODRIGUEZ, Carlos M.; CASADEVALL, Arturo. Cryptococcus neoformans: Tripping on acid in the phagolysosome. **Frontiers in Microbiology**, [s. l.], v. 7, n. FEB, p. 1–9, 2016.
DERBYSHIRE, Mark Charles et al. Small RNAs from the plant pathogenic fungus Sclerotinia sclerotiorum highlight candidate host target genes associated with quantitative disease resistance. **bioRxiv**, [s. l.], n. 7, p. 354076, 2018. Disponível em: <https://www.biorxiv.org/content/early/2018/06/22/354076>
DERENGOWSKI, Lorena da S. et al. The transcriptional response of cryptococcus neoformans to ingestion by Acanthamoeba castellanii and macrophages provides insights into the evolutionary adaptation to the mammalian host. **Eukaryotic Cell**, [s. l.], v. 12, n. 5, p. 761–774, 2013.
DESJARDINS, Christopher A. et al. Comparative genomic analysis of human fungal pathogens causing paracoccidioidomycosis. **PLoS Genetics**, [s. l.], v. 7, n. 10, 2011.
DODONOVA, Svetlana O. et al. Structure of the Ty3/Gypsy retrotransposon capsid and the evolution of retroviruses. **Proceedings of the National Academy of Sciences**, [s. l.], p. 201900931, 2019.
DONALDSON, Michael E.; SAVILLE, Barry J. Natural antisense transcripts in fungi. **Molecular Microbiology**, [s. l.], v. 85, n. 3, p. 405–417, 2012.

DONALDSON, Michael E.; SAVILLE, Barry J. Ustilago maydis natural antisense transcript expression alters mRNA stability and pathogenesis. **Molecular Microbiology**, [s. l.], v. 89, n. 1, p. 29–51, 2013.
DONNART, Tifenn et al. Filamentous ascomycete genomes provide insights into Copia retrotransposon diversity in fungi. **BMC Genomics**, [s. l.], v. 18, n. 1, p. 1–15, 2017.
DRINNENBERG, Ines a et al. RNAi in budding yeast. **Science (New York, N.Y.)**, [s. l.], v. 326, n. 5952, p. 544–50, 2009. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/19745116>
DUTREUX, Fabien et al. Data descriptor: De novo assembly and annotation of three leptosphaeria genomes using oxford nanopore minion sequencing. **Scientific Data**, [s. l.], v. 5, n. April, p. 1–11, 2018. Disponível em: <http://dx.doi.org/10.1038/sdata.2018.235>
EDWARDS, Jessica A. et al. Histoplasma yeast and mycelial transcriptomes reveal pathogenic-phase and lineage-specific gene expression profiles. **BMC Genomics**, [s. l.], v. 14, n. 1, 2013.
EMMS, David M.; KELLY, Steven. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. **Genome Biology**, [s. l.], v. 16, n. 1, p. 1–14, 2015. Disponível em: <http://dx.doi.org/10.1186/s13059-015-0721-2>
FARRER, Rhys A. et al. Genome evolution and innovation across the four major lineages of cryptococcus gattii. **mBio**, [s. l.], v. 6, n. 5, p. 1–12, 2015.
FARRER, Rhys A. et al. Microevolutionary traits and comparative population genomics of the emerging pathogenic fungus Cryptococcus gattii. **Philosophical Transactions of the Royal Society B: Biological Sciences**, [s. l.], v. 371, n. 1709, 2016.
FERETZAKI, Marianna et al. Gene Network Polymorphism Illuminates Loss and Retention of Novel RNAi Silencing Components in the Cryptococcus Pathogenic Species Complex. **PLoS Genetics**, [s. l.], v. 12, n. 3, p. 1–25, 2016.
FERNANDES, João Daniel Santos et al. The role of amino acid permeases and tryptophan biosynthesis in Cryptococcus neoformans survival. **PLoS ONE**, [s. l.], v. 10, n. 7, p. 1–22, 2015.
FIRACATIVE, Carolina et al. The status of cryptococcosis in latin America. **Memorias do Instituto Oswaldo Cruz**, [s. l.], v. 113, n. 7, p. 1–23, 2018.
GARBE, Enrico; VYLKOVA, Slavena. Role of Amino Acid Metabolism in the Virulence of Human Pathogenic Fungi. **Vascular Pharmacology**, [s. l.], v. 6, p. 108–119, 2019.
GONZALEZ-HILARION, Sara et al. Intron retention-dependent gene regulation in Cryptococcus neoformans. **Scientific Reports**, [s. l.], v. 6, p. 1–13, 2016. Disponível em: <http://dx.doi.org/10.1038/srep32252>
GOODWIN, Timothy J. D.; POULTER, Russell T. M. The diversity of retrotransposons in the yeast Cryptococcus neoformans. **Yeast**, [s. l.], v. 18, n. 9, p. 865–880, 2001.
GORDON, Sean P. et al. Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. **PLoS ONE**, [s. l.], v. 10, n. 7, p. 1–15, 2015.
HAAS, Brian J. et al. Approaches to fungal genome annotation. **Mycology**, [s. l.], v. 2, n. 3, p. 118–141, 2011.
HAGEN, Ferry et al. Recognition of seven species in the Cryptococcus gattii/Cryptococcus neoformans species complex. **Fungal Genetics and Biology**, [s.

l.], v. 78, p. 16–48, 2015. Disponível em: <http://dx.doi.org/10.1016/j.fgb.2015.02.009>
HAGIWARA, Daisuke et al. Whole-genome comparison of aspergillus fumigatus strains serially isolated from patients with aspergillosis. **Journal of Clinical Microbiology**, [s. l.], v. 52, n. 12, p. 4202–4209, 2014.
HALIC, Mario; MOAZED, Danesh. Dicer-Independent Primal RNAs Trigger RNAi and Heterochromatin Formation. **Cell**, [s. l.], v. 140, n. 4, p. 504–516, 2010.
HAMMOND, T. M.; KELLER, N. P. RNA silencing in Aspergillus nidulans is independent of RNA-dependent RNA polymerases. **Genetics**, [s. l.], v. 169, n. 2, p. 607–617, 2005.
HAVECKER, Ericka R.; GAO, Xiang; VOYTAS, Daniel F. The diversity of LTR retrotransposons. **Genome Biology**, [s. l.], v. 5, n. 6, 2004.
HEATHER, James M.; CHAIN, Benjamin. The sequence of sequencers: The history of sequencing DNA. **Genomics**, [s. l.], v. 107, n. 1, p. 1–8, 2016. Disponível em: <http://dx.doi.org/10.1016/j.ygeno.2015.11.003>
HOFF, Katharina J. et al. BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. **Bioinformatics**, [s. l.], v. 32, n. 5, p. 767–769, 2016.
JANBON, Guilhem et al. Characterizing the role of RNA silencing components in Cryptococcus neoformans. **Fungal Genetics and Biology**, [s. l.], v. 47, n. 12, p. 1070–1080, 2010. Disponível em: <http://dx.doi.org/10.1016/j.fgb.2010.10.005>
JANBON, Guilhem et al. Analysis of the Genome and Transcriptome of Cryptococcus neoformans var. grubii Reveals Complex RNA Expression and Microevolution Leading to Virulence Attenuation. **PLoS Genetics**, [s. l.], v. 10, n. 4, 2014.
JOHNSTON, Simon A.; MAY, Robin C. Cryptococcus interactions with macrophages: Evasion and manipulation of the phagosome by a fungal pathogen. **Cellular Microbiology**, [s. l.], v. 15, n. 3, p. 403–411, 2013.
JONES, T. et al. The diploid genome sequence of Candida albicans. **Proceedings of the National Academy of Sciences**, [s. l.], v. 101, n. 19, p. 7329–7334, 2004.
JUNG, Won Hee; KRONSTAD, James W. Iron influences the abundance of the iron regulatory protein Cir1 in the fungal pathogen Cryptococcus neoformans. **FEBS Letters**, [s. l.], v. 585, n. 20, p. 3342–3347, 2011. Disponível em: <http://dx.doi.org/10.1016/j.febslet.2011.09.025>
KAN, Jan A. L. Van et al. A gapless genome sequence of the fungus Botrytis cinerea. **Molecular Plant Pathology**, [s. l.], v. 18, n. 1, p. 75–89, 2017.
KIM, Daehwan et al. TopHat2 accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. **Genome Biol**, [s. l.], v. 14, p. 14, 2013.
KIRKLAND, Theo; MUSZEWSKA, Anna; STAJICH, Jason. Analysis of Transposable Elements in Coccidioides Species. **Journal of Fungi**, [s. l.], v. 4, n. 1, p. 13, 2018. Disponível em: <http://www.mdpi.com/2309-608X/4/1/13>. Acesso em: 12 jul. 2019.
KJÆRBØLLING, Inge et al. Linking secondary metabolites to gene clusters through genome sequencing of six diverse Aspergillus species . **Proceedings of the National Academy of Sciences**, [s. l.], v. 115, n. 4, p. E753–E761, 2018.
KUO, Alan; BUSHNELL, Brian; GRIGORIEV, Igor V. **Fungal genomics: Sequencing and annotation**. 1. ed. [s.l.] : Elsevier Ltd., 2014. v. 70 Disponível em: <http://dx.doi.org/10.1016/B978-0-12-397940-7.00001-X>

LANGMEAD, Ben; SALZBERG, Steven L. Fast gapped-read alignment with Bowtie 2. **Nature Methods**, [s. l.], v. 9, n. 4, p. 357–359, 2012.
LANVER, Daniel et al. The Biotrophic Development of Ustilago maydis Studied by RNA-Seq Analysis . **The Plant Cell**, [s. l.], v. 30, n. 2, p. 300–323, 2018.
LAU, Susanna K. P. et al. Identification of MicroRNA-Like RNAs in Mycelial and Yeast Phases of the Thermal Dimorphic Fungus Penicillium marneffei. **PLoS Neglected Tropical Diseases**, [s. l.], v. 7, n. 8, 2013.
LEE, Heng Chi et al. Diverse Pathways Generate MicroRNA-like RNAs and Dicer-Independent Small Interfering RNAs in Fungi. **Molecular Cell**, [s. l.], v. 38, n. 6, p. 803–814, 2010. Disponível em: <http://dx.doi.org/10.1016/j.molcel.2010.04.005>
LEEM, Young Eun et al. Retrotransposon Tf1 Is Targeted to Pol II Promoters by Transcription Activators. **Molecular Cell**, [s. l.], v. 30, n. 1, p. 98–107, 2008.
LEOPOLD WAGER, Chrissy M. et al. Cryptococcus and phagocytes: Complex interactions that influence disease outcome. **Frontiers in Microbiology**, [s. l.], v. 7, n. FEB, p. 1–16, 2016.
LOFTUS, Brendan J. et al. The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans. **Science**, [s. l.], v. 307, n. 5713, p. 1321–1324, 2005.
LORENZ, Michael C.; FINK, Gerald R. Life and Death in a Macrophage: Role of the Glyoxylate Cycle in Virulence. **Society**, [s. l.], v. 1, n. 5, p. 657–662, 2002.
MACHIDA, Masayuki et al. Genome sequencing and analysis of Aspergillus oryzae. **Nature**, [s. l.], v. 438, n. 7071, p. 1157–1161, 2005.
MARTIN, Marcel. Cutadapt removes adapter sequences from high-throughput sequencing reads. **EMBnet.journal; Vol 17, No 1: Next Generation Sequencing Data AnalysisDO - 10.14806/ej.17.1.200** , [s. l.], 2011. Disponível em: <http://journal.embnet.org/index.php/embnetjournal/article/view/200>
MCGETTIGAN, Paul A. Transcriptomics in the RNA-seq era. **Current Opinion in Chemical Biology**, [s. l.], v. 17, n. 1, p. 4–11, 2013.
MENG, Huimin et al. Dicer and Argonaute Genes Involved in RNA Interference in the Entomopathogenic Fungus Metarhizium robertsii. **Applied and Environmental Microbiology**, [s. l.], v. 83, n. 7, p. 1–15, 2017.
MEYER, Wieland et al. Molecular Typing of the Cryptococcus neoformans/Cryptococcus gattii Species Complex. In: **Cryptococcus: from human pathogen to model yeast**. [s.l.] : American Society for Microbiology, 2011. p. 327–358.
MEYER, Wieland. Cryptococcus gattii in the age of whole-genome sequencing. **mBio**, [s. l.], v. 6, n. 6, p. 6–9, 2015.
MOAZED, Danesh. Small RNAs in transcriptional gene silencing and genome defence. **Nature**, [s. l.], v. 457, n. 7228, p. 413–420, 2009.
MOHANTA, Tapan Kumar; BAE, Hanhong. The diversity of fungal genome. **Biological Procedures Online**, [s. l.], v. 17, n. 1, p. 1–9, 2015.
MUÑOZ, José F. et al. Genome Update of the Dimorphic Human Pathogenic Fungi Causing Paracoccidioidomycosis. **PLoS Neglected Tropical Diseases**, [s. l.], v. 8, n. 12, 2014.
MUÑOZ, José F. et al. Coordinated host-pathogen transcriptional dynamics revealed

using sorted subpopulations and single macrophages infected with Candida albicans. **Nature Communications**, [s. l.], v. 10, n. 1, 2019.
MUSZEWSKA, Anna et al. Cut-and-paste transposons in fungi with diverse lifestyles. **Genome Biology and Evolution**, [s. l.], v. 9, n. 12, p. 3463–3477, 2017.
MUSZEWSKA, Anna et al. Transposable elements contribute to fungal genes and impact fungal lifestyle. **Scientific Reports**, [s. l.], v. 9, n. 1, p. 1–10, 2019.
MUSZEWSKA, Anna; HOFFMAN-SOMMER, Marta; GRYNBERG, Marcin. LTR retrotransposons in fungi. **PLoS ONE**, [s. l.], v. 6, n. 12, 2011.
MUSZEWSKA, Anna; STECZKIEWICZ, Kamil; GINALSKI, Krzysztof. DIRS and Ngaro Retrotransposons in Fungi. **PLoS ONE**, [s. l.], v. 8, n. 9, p. 1–10, 2013.
NEUVÉGLISE, Cécile et al. Genomic evolution of the long terminal repeat retrotransposons in hemiascomycetous yeasts. **Genome Research**, [s. l.], v. 12, n. 6, p. 930–943, 2002.
NGUYEN, Quyet et al. A fungal Argonaute interferes with RNA interference. **Nucleic Acids Research**, [s. l.], v. 46, n. 5, p. 2495–2508, 2018.
NICOLÁS, Francisco E. et al. The RNAi machinery controls distinct responses to environmental signals in the basal fungus Mucor circinelloides. **BMC Genomics**, [s. l.], v. 16, n. 1, p. 1–14, 2015. Disponível em: <???>
NICOLÁS, Francisco E.; RUIZ-VÁZQUEZ, Rosa M. Functional diversity of RNAi-associated sRNAs in fungi. **International Journal of Molecular Sciences**, [s. l.], v. 14, n. 8, p. 15348–15360, 2013.
NICOLAS, Francisco Esteban et al. Endogenous short RNAs generated by Dicer 2 and RNA-dependent RNA polymerase 1 regulate mRNAs in the basal fungus Mucor circinelloides. **Nucleic Acids Research**, [s. l.], v. 38, n. 16, p. 5535–5541, 2010.
NOLAN, Tony et al. The post-transcriptional gene silencing machinery functions independently of DNA methylation to repress a LINE1-like retrotransposon in Neurospora crassa. **Nucleic Acids Research**, [s. l.], v. 33, n. 5, p. 1564–1573, 2005.
NOVIKOVA, Olga; FET, Victor; BLINOV, Alexander. Non-LTR retrotransposons in fungi. **Functional and Integrative Genomics**, [s. l.], v. 9, n. 1, p. 27–42, 2009.
NUNES, Cristiano C. et al. Diverse and tissue-enriched small RNAs in the plant pathogenic fungus, Magnaporthe oryzae. **BMC Genomics**, [s. l.], v. 12, p. 1–20, 2011.
OKUBO, Yoichiro et al. Histopathological study of murine pulmonary cryptococcosis Induced by Cryptococcus gattii and Cryptococcus neoformans. **Japanese Journal of Infectious Diseases**, [s. l.], v. 66, n. 3, p. 216–221, 2013.
OZATA, Deniz M. et al. PIWI-interacting RNAs: small RNAs with big functions. **Nature Reviews Genetics**, [s. l.], v. 20, n. 2, p. 89–108, 2019.
ÖZKAN, Selin et al. Profile and functional analysis of small RNAs derived from Aspergillus fumigatus infected with double-stranded RNA mycoviruses. **BMC Genomics**, [s. l.], v. 18, n. 1, p. 1–13, 2017.
PAUN, Linda; KEMPKEN, Frank. Genetic Transformation Systems in Fungi, Volume 2. [s. l.], v. 2, n. 1, p. 79–96, 2015. Disponível em: <http://link.springer.com/10.1007/978-3-319-10503-1>
PROUX-WÉRA, Estelle et al. A pipeline for automated annotation of yeast genome sequences by a conserved-synteny approach. **BMC Bioinformatics**, [s. l.], v. 13, n. 1, 2012.

RAMAN, Vidhyavathi et al. Small RNA Functions Are Required for Growth and Development of Magnaporthe oryzae . **Molecular Plant-Microbe Interactions**, [s. l.], v. 30, n. 7, p. 517–530, 2017.
RAO, Soumya et al. The landscape of repetitive elements in the refined genome of chilli anthracnose fungus colletotrichum truncatum. **Frontiers in Microbiology**, [s. l.], v. 9, n. OCT, p. 1–19, 2018.
SCHNEIDER, Rafael de Oliveira et al. Zap1 regulates zinc homeostasis and modulates virulence in Cryptococcus gattii. **PLoS ONE**, [s. l.], v. 7, n. 8, 2012.
SEPÚLVEDA, Victoria E. et al. Genome Sequences Reveal Cryptic Speciation in the Human Pathogen Histoplasma capsulatum . **mBio**, [s. l.], v. 8, n. 6, p. 1–23, 2017.
SHAABAN, Mona et al. Involvement of transposon-like elements in penicillin gene cluster regulation. **Fungal Genetics and Biology**, [s. l.], v. 47, n. 5, p. 423–432, 2010. Disponível em: <http://dx.doi.org/10.1016/j.fgb.2010.02.006>
SHARMA, Krishna Kant. Fungal genome sequencing: basic biology to biotechnology. **Critical Reviews in Biotechnology**, [s. l.], v. 36, n. 4, p. 743–759, 2015. Disponível em: <http://dx.doi.org/10.3109/07388551.2015.1015959>
SHARPTON, Thomas J. et al. Comparative genomic analyses of the human fungal pathogens Coccidioides and their relatives. **Genome Research**, [s. l.], v. 19, n. 10, p. 1722–1731, 2009.
SIEBER, Patricia et al. Comparative study on alternative splicing in human fungal pathogens suggests its involvement during host invasion. **Frontiers in Microbiology**, [s. l.], v. 9, n. OCT, p. 1–13, 2018.
SILVESTRI, Alessandro et al. In silico analysis of fungal small RNA accumulation reveals putative plant mRNA targets in the symbiosis between an arbuscular mycorrhizal fungus and its host plant. **BMC Genomics**, [s. l.], v. 20, n. 1, p. 1–18, 2019.
SLATER, Guy St C.; BIRNEY, Ewan. Automated generation of heuristics for biological sequence comparison. **BMC Bioinformatics**, [s. l.], v. 6, p. 1–11, 2005.
SRIVASTAVA, Anuj; GEORGE, Joshy; KARUTURI, Radha KM. Transcriptome analysis. **Advances in Biochemical Engineering/Biotechnology**, [s. l.], v. 127, p. 1–25, 2018.
STAJICH, Jason E. Fungal Genomes and Insights into the Evolution of the Kingdom. **The Fungal Kingdom**, [s. l.], p. 619–633, 2017.
STAJICH, Jason E.; DIETRICH, Fred S.; ROY, Scott W. Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. **Genome Biology**, [s. l.], v. 8, n. 10, 2007.
TESTA, Alison C. et al. CodingQuarry: Highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. **BMC Genomics**, [s. l.], v. 16, n. 1, p. 1–12, 2015.
TRAPNELL, Cole et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. **Nature Biotechnology**, [s. l.], v. 28, n. 5, p. 511–515, 2010. Disponível em: <http://dx.doi.org/10.1038/nbt.1621>
TRIEU, Trung Anh et al. A Non-canonical RNA Silencing Pathway Promotes mRNA Degradation in Basal Fungi. **PLoS Genetics**, [s. l.], v. 11, n. 4, p. 1–32, 2015.
TUDZYNSKI, Bettina. Nitrogen regulation of fungal secondary metabolism in fungi.

Frontiers in Microbiology, [s. l.], v. 5, n. NOV, p. 1–15, 2014.
WALLACE, Edward et al. Start codon context controls translation initiation in the fungal kingdom. **bioRxiv**, [s. l.], p. 654046, 2019. Disponível em: <https://www.biorxiv.org/content/10.1101/654046v1.full>
WANG, Lingling et al. Integrated microRNA and mRNA analysis in the pathogenic filamentous fungus Trichophyton rubrum. **BMC Genomics**, [s. l.], v. 19, n. 1, p. 1–14, 2018.
WANG, Xuying et al. Sex-induced silencing defends the genome of Cryptococcus neoformans via RNAi. **Genes and Development**, [s. l.], v. 24, n. 22, p. 2566–2582, 2010.
WANG, Xuying et al. Transgene Induced Co-Suppression during Vegetative Growth in Cryptococcus neoformans. **PLoS Genetics**, [s. l.], v. 8, n. 8, p. 1–13, 2012.
WANG, Xuying; DARWICHE, Sabrina; HEITMAN, Joseph. Sex-induced silencing operates during opposite-sex and unisexual reproduction in Cryptococcus neoformans. **Genetics**, [s. l.], v. 193, n. 4, p. 1163–1174, 2013.
WANG, Zhong; GERSTEIN, Mark; SNYDER, Michael. RNA-Seq: a revolutionary tool for transcriptomics. **Nature Reviews Genetics**, [s. l.], v. 10, n. 1, p. 57, 2009.
WEIBERG, Arne et al. Fungal small RNAs suppress plant immunity by hijacking host RNA interference pathways. **Science**, [s. l.], v. 342, n. 6154, p. 118–123, 2013.
WU, Yuan et al. A genome-wide transcriptional analysis of yeast-hyphal transition in Candida tropicalis by RNA-Seq. **PLoS ONE**, [s. l.], v. 11, n. 11, p. 1–16, 2016.
YADAV, Vikas et al. RNAi is a critical determinant of centromere evolution in closely related fungi. **Proceedings of the National Academy of Sciences**, [s. l.], v. 115, n. 12, p. 3108–3113, 2018.
YANDELL, Mark; ENCE, Daniel. A beginner's guide to eukaryotic genome annotation. **Nature Reviews Genetics**, [s. l.], v. 13, n. 5, p. 329–342, 2012. Disponível em: <http://dx.doi.org/10.1038/nrg3174>
YANG, Fen. Genome-wide analysis of small RNAs in the wheat pathogenic fungus Zymoseptoria tritici. **Fungal Biology**, [s. l.], v. 119, n. 7, p. 631–640, 2015. Disponível em: <http://dx.doi.org/10.1016/j.funbio.2015.03.008>
YANG, Ying et al. Genome sequencing and comparative genomics analysis revealed pathogenic potential in Penicillium capsulatum as a novel fungal pathogen belonging to Eurotiales. **Frontiers in Microbiology**, [s. l.], v. 7, n. OCT, p. 1–14, 2016.
ZHANG, Mingshun; SUN, Donglei; SHI, Meiqing. Dancing cheek to cheek: Cryptococcus neoformans and phagocytes. **SpringerPlus**, [s. l.], v. 4, n. 1, 2015.

7 CURRICULUM VITAE resumido

FERRAREZE, P.A.G.;

DADOS PESSOAIS

Nome: Patrícia Aline Gröhs Ferrareze

Local e data de nascimento: Portão, Rio Grande do Sul, Brasil, 20 de maio de 1991.

Endereço profissional:

Av. Bento Gonçalves, 9500; Setor 4; Prédio 43421; Lab. 219; Porto

Alegre, RS; CEP: 91501-970

Telefone profissional: +55 51 33087767

E-mail:

p.ferrareze@gmail.com

FORMAÇÃO

Curso de graduação: Ciências Biológicas, Universidade do Vale do Rio dos Sinos (UNISINOS). Período: 2009 – 2012.

Título: Identificação molecular e bioquímica de bactéria queratinolítica antártica e estudo de atividade proteolítica.

Orientador: Luís Fernando da Costa Medina.

Curso de Pós-Graduação: Mestrado *strictu sensu* em Microbiologia Agrícola e do Ambiente, Universidade Federal do Rio Grande do Sul (UFRGS). Período: 2013 – 2015.

Título: Purificação e caracterização de protease queratinolítica produzida por linhagem probiótica de *Bacillus subtilis*.

Orientador: Adriano Brandelli.

Curso de Pós-Graduação: Doutorado *strictu sensu* em Biologia Celular e Molecular, Universidade Federal do Rio Grande do Sul (UFRGS). Período: 2015 – 2019.

Título: Análises transcritômicas em *Cryptococcus gattii* e seu impacto na definição de genes e outros elementos genômicos.

Orientador: Charley Christian Staats.

FORMAÇÃO COMPLEMENTAR

Imunologia básica aplicada a Clínica. Carga horária: 252h. Período: 2015 – 2015. Universidade Federal do Rio Grande do Sul, UFRGS, Brasil.

Introdução a PCR em tempo real. Carga horária: 15h. Período: 2014 – 2014. Universidade Federal do Rio Grande do Sul, UFRGS, Brasil.

Ferramentas Moleculares para o Estudo de Proteínas. Carga horária: 70h. Período: 2012 – 2012. Pontifícia Universidade Católica do Rio Grande do Sul, PUCRS, Brasil.

Biotecnologia: Agricultura e Alimentos. Carga horária: 30h. Período: 2010 – 2010. Universidade do Vale do Rio dos Sinos, UNISINOS, Brasil.

Métodos e Aplicações Laboratoriais para o Diagnóstico Clínico. Carga horária: 15h. Período: 2010 – 2010. Universidade do Vale do Rio dos Sinos, UNISINOS, Brasil.

ESTÁGIOS

Vínculo: Monitoria de atividade acadêmica. Instituição: Universidade do Vale do Rio dos Sinos (UNISINOS). Período: 2011 – 2011. Atividade: Biologia Celular e Molecular. Responsável: Victor Hugo Valiati.

Vínculo: Monitoria de atividade acadêmica. Instituição: Universidade do Vale do Rio dos Sinos (UNISINOS). Período: 2012 - 2012. Atividade: Técnicas para Diagnóstico Molecular. Responsável: Luís Fernando da Costa Medina.

Vínculo: Estágio (Doutorado sanduíche). Instituição: Institut Pasteur. Período: 2017 – 2018. Atividade: Pesquisa (Biologia Molecular e Bioinformática). Responsável: Guilhem Janbon.

ARTIGOS COMPLETOS PUBLICADOS

REUWSAAT, JULIA CATARINA VIEIRA; MOTTA, HERYK; GARCIA, ANE WICHINE ACOSTA; VASCONCELOS, CAROLINA BETTKER; MARQUES, BÁRBARA MACHADO; OLIVEIRA, NATÁLIA KRONBAUER; RODRIGUES, JÉSSICA; **FERRAREZE, PATRÍCIA ALINE GRÖHS**; FRASES, SUSANA; LOPES, WILLIAM; BARCELLOS, VANESSA ABREU; SQUIZANI, EAMIM DAIDRÊ; HORTA, JORGE ANDRÉ; SCHRANK, AUGUSTO; RODRIGUES, MARCIO LOURENÇO; STAATS, CHARLEY CHRISTIAN; VAINSTEIN, MARILENE HENNING; KMETZSCH, LÍVIA. A Predicted Mannoprotein Participates in *Cryptococcus gattii* Capsular Structure. *mSphere*, v. 3, p. e00023-18, 2018.

FERRAREZE, PATRÍCIA ALINE GRÖHS; STREIT, RODRIGO SILVA ARAUJO; DOS SANTOS, FRANCINE MELISE; SCHRANK, AUGUSTO; KMETZSCH, LIVIA; VAINSTEIN, MARILENE HENNING; STAATS, CHARLEY CHRISTIAN. sRNAs as possible regulators of retrotransposon activity in *Cryptococcus gattii* VGII*. BMC Genomics*, v. 18, p. 294, 2017.

FERRAREZE, PATRÍCIA ALINE GRÖHS; STREIT, RODRIGO SILVA ARAUJO; SANTOS, PATRICIA RIBEIRO DOS; SANTOS, FRANCINE MELISE DOS; ALMEIDA, RITA MARIA CUNHA DE; SCHRANK, AUGUSTO; KMETZSCH, LIVIA; VAINSTEIN, MARILENE HENNING; STAATS, CHARLEY CHRISTIAN. Transcriptional Analysis

Allows Genome Reannotation and Reveals that *Cryptococcus gattii* VGII Undergoes Nutrient Restriction during Infection. *Microorganisms*, v. 5, p. 49, 2017.

RIBEIRO, NICOLE S.; DOS SANTOS, FRANCINE M.; GARCIA, ANE W. A.; **FERRAREZE, PATRÍCIA A. G.**; FABRES, LAURA F.; SCHRANK, AUGUSTO; KMETZSCH, LIVIA; ROTT, MARILISE B.; VAINSTEIN, MARILENE H.; STAATS, CHARLEY C. Modulation of Zinc Homeostasis in *Acanthamoeba castellanii* as a Possible Antifungal Strategy against *Cryptococcus gattii*. *Frontiers in Microbiology*, v. 8, p. 1626, 2017.

FERRAREZE, PATRÍCIA ALINE GRÖHS; CORREA, ANA PAULA FOLMER; BRANDELLI, ADRIANO. Purification and characterization of a keratinolytic protease produced by probiotic *Bacillus subtilis*. *Biocatalysis and Agricultural Biotechnology*, v. 7, p. 102-109, 2016.

FERRAREZE, P. A. G.; PETRY, M. V.; BRANDELLI, A.; VALIATI, V. H.; MEDINA, L. F. C. Characterization of Antartic Keratinolytic *Arthrobacter* sp. INCT-APA Annual Activity Report, p. 68-71, 2015.

RESUMOS E TRABALHOS APRESENTADOS EM CONGRESSOS

FERRAREZE, PATRÍCIA ALINE GRÖHS; VAINSTEIN, MARILENE HENNING; STAATS, CHARLEY CHRISTIAN. The impact of RNAi pathway absence in

retrotransposon activity in *Cryptococcus gattii* VGII genotype. In: 10th International Conference of Cryptococcus and Cryptococcosis, 2017, Foz do Iguaçu. 10th International Conference of Cryptococcus and Cryptococcosis, 2017.

FERRAREZE, P. A. G.; BRANDELLI, A. . Análise molecular e caracterização enzimática de proteases queratinolíticas para a degradação de queratina residual da agroindústria. In: VII Simpósio Brasileiro de Microbiologia Aplicada, 2014, Porto Alegre. Anais do VII Simpósio Brasileiro de Microbiologia Aplicada, 2014.

GONCALVES, R. M.; **FERRAREZE, P. A. G.**; BRANDELLI, A.; MEDINA, L. F. C . Caracterização da atividade queratinolítica de *Arthrobacter* sp. PF01. In: VII Simpósio Brasileiro de Microbiologia Aplicada, 2014, Porto Alegre. Anais do VII Simpósio Brasileiro de Microbiologia Aplicada, 2014.

GONCALVES, R. M.; **FERRAREZE, P. A. G.**; BRANDELLI, A.; MEDINA, L. F. C . Caracterização da atividade queratinolítica de *Arthrobacter* sp. PF01. In: III Encontro Latino Americano de Microbiologia Aplicada, 2014, Porto Alegre. Anais do III Encontro Latino Americano de Microbiologia Aplicada, 2014.

GONCALVES, R. M.; **FERRAREZE, P. A. G.**; BRANDELLI, A.; MEDINA, L. F. C . Caracterização da atividade queratinolítica de *Arthrobacter* sp. PF01. In: IX Encontro Nacional de Estudantes de Pós-graduação em Microbiologia da Área Agrícola, 2014,

Porto Alegre. Anais do IX Encontro Nacional de Estudantes de Pós-graduação em Microbiologia da Área Agrícola, 2014.

FERRAREZE, P. A. G.; BRANDELLI, A. Análise molecular e caracterização enzimática de proteases queratinolíticas para a degradação de queratina residual da agroindústria.. In: IX Encontro Nacional de Estudantes de Pós-graduação em Microbiologia da Área Agrícola, 2014, Porto Alegre. Anais do IX Encontro Nacional de Estudantes de Pós-graduação em Microbiologia da Área Agrícola, 2014.

FERRAREZE, P. A. G.; BRANDELLI, A. . Análise molecular e caracterização enzimática de proteases queratinolíticas para a degradação de queratina residual da agroindústria. In: III Encontro Latino Americano de Microbiologia Aplicada, 2014, Porto Alegre. Anais do III Encontro Latino Americano de Microbiologia Aplicada, 2014.

FERRAREZE, P. A. G.; MEDINA, L. F. C; VALIATI, V. H.; BRANDELLI, A.; PETRY, M. V. Produção de enzimas ativas a baixas temperaturas pela bactéria antártica *Arthrobacter gangotriensis* PF01. In: II Congresso de Iniciação Científica e Pós-Graduação (CICPG), 2012, São Leopoldo. Anais do II Congresso de Iniciação Científica e Pós-Graduação (CICPG), 2012.

FERRAREZE, P. A. G.; MEDINA, L. F. C . Atividade queratinolítica em bactéria psicrófila antártica. In: Salão de Iniciação Científica (2010, UFRGS, Porto Alegre, RS),

2010, Porto Alegre. Salão de Iniciação Científica (2010). Livro de resumos. Porto Alegre: UFRGS, 2010.