Análise Transcritômica de *Cryptococcus Gattii*

Rodrigo Silva Araujo Streit[1]; Charley Christian Staats[12]

[1]Laboratório de Fungos de Importância Médica e Biotecnológica , Centro de Biotecnologia, UFRGS, Porto Alegre, Brasil

[2]Departamento de Biologia Molecular e Biotecnologia, Centro de Biotecnologia, UFRGS, Porto Alegre, Brasil

Introdução

A levedura basidiomicética *Cryptococcus gatti* é um dos agentes etiológicos da criptococose, doença que pode acometer pele, pulmões e sistema nervoso central (Kwon-Chung *et al*. Cold Spring Harb Perspect Med, a019760, 2014). As sequências genômica e transcritômica predita da linhagem R265 de *C. gattii* foram disponibilizadas pelo BroadInstitute (*Cryptococcus gattii* Sequencing Project, Broad Institute of Harvardand MIT) no ano de 2009, passando por suas últimas atualizações em 2015. Apesar se sua recente atualização, os métodos empregados para a obtenção de modelos gênicos mantiveram-se os mesmos ao longo das versões, utilizando softwares de anotação automática como Augustus, FGNESH e outros preditores *de novo*, metodologia à qual a literatura vem demonstrado haver erros inerentes e limitações em suas predições, especialmente em organismos onde é encontrada justaposição de genes. Ao passo que erros vêm sendo constatados na metodologia clássica, novos métodos de análise que utilizam dados de sequenciamento de última geração vêm sendo desenvolvidos e têm mostrando extrema eficiência e robustez na predição transcricional, de forma que têm sido constantes as correções e complementações de informações prévias de diversos organismos (Zhao *et al.* BMC Genomics, 14:21, 2013). Assim, o presente estudo tem por objetivo a análise e correção do transcritoma da linhagem R265 de *C. gattii*, baseando-se em metodologias de RNA-seq.

Metodologia

Alinhamento e predição gênica: Para a execução dos alinhamentos, foram utilizadas quatro bibliotecas de *reads*, cada qual referente a uma condição de cultivo diferente, provenientes do sequênciamento de RNA da linhagem R265 de *C. gattii* na plataforma IlluminaHiSeq, totalizando 160.411.087 *reads*. Os alinhamentos foram realizados utilizando o software Tophat, alinhando os *reads* contra as sequências de transcritos, genes e supercontigs disponibilidazos pelo Broad Institute, sendo a contabilização dos *reads* alinhados feita pelo próprio software. Em seguida, os dados do alinhamento contra o genoma foram utilizados para uma predição inicial dos transcritos através do software Cufflinks. Por fim, os modelos gerados pelo Cufflinks foram refinados pelo software CodingQuarry, através da associação da predição por alinhamento com algoritmos de predição automática;

Visualização e correção manual da predição gênica: Os modelos gênicos obtidos foram revisados manualmente no software IGV, utilizando o resultado do alinhamento inicial como parâmetro para a verificação de predições errôneas. Foi feita então a correção manual dos genes preditos de forma incorreta, observando regras de delimitação de íntrons, estrutura de genes ortólogos e os alinhamentos prévios. Após, foram comparados os resultados obtidos da predição com a anotação atual.

Resultados

Com a análise dos 3 alinhamentos executados contra as sequências gênicas e de regiões codificantes disponibilizadas pelo Broad Institute, foi possível verificar que uma quantidade considerável de *reads* alinhava em regiões que não codificavam exons (Tabela 1), especialmente em íntrons, indicando possíveis erros na anotação atual.

Tabela 1: Número de reads alinhados uma única vez de acordo com a estrutura gênica no genoma atual.

	Número de *reads*	(%)
Éxons	104990211	(65,45%)
Íntrons	8098125	(5,05%)
Regiões intergênicas	24802183	(15,46%)
Não alinhados	22520568	(14,04%)
Total	160411087	(100%)

Até então, 4400 dos 6456 genes presentes na atual anotação tiveram seus transcritos revisados, sendo que destes, 1818 sofreram alterações em sua fase de leitura com base nos dados do alinhamento (Figura 1). Dentro desse grupo de transcritos cuja anotação foi corrigida, destacaram-se 3 principais motivos de alteração: (i) a delimitação de íntrons e éxons, correspondendo a 1583 transcritos (figura 2A), (ii) genes unidos ou separados de forma equívoca, correspondendo a 86 transcritos, e (iii) genes que apresentaram erros em sua sequência genômica que alteravam sua fase de leitura, correspondendo a 149 transcritos (Figura 2B). Ainda, 375 foram excluídos das análises devido à quantidade insatisfatória de *reads* alinhados em suas regiões, impossibilitando a avaliação visual e reduzindo a acurácia dos preditores, sendo aceito, para esses casos, a anotação prévia.

Além das correções dos modelos gênicos já existentes, foram identificados até então 39 possíveis novos genes, sendo esses mesmos confirmados através de buscas por genes ortólogos .

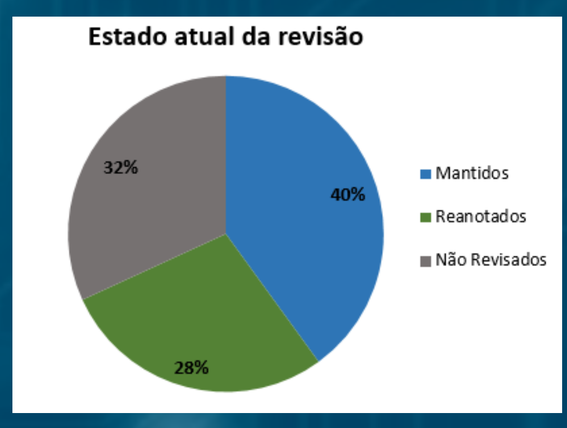

Figura 1: Representação gráfica do estado atual da revisão do transcritoma da linhagem R265 de C. *gattii*. Dos 4400 genes até então revisados (68%), 1818 genes sofreram reanotação (28%), enquanto que 2582 genes mantiveram sua anotação inalterada (40%).

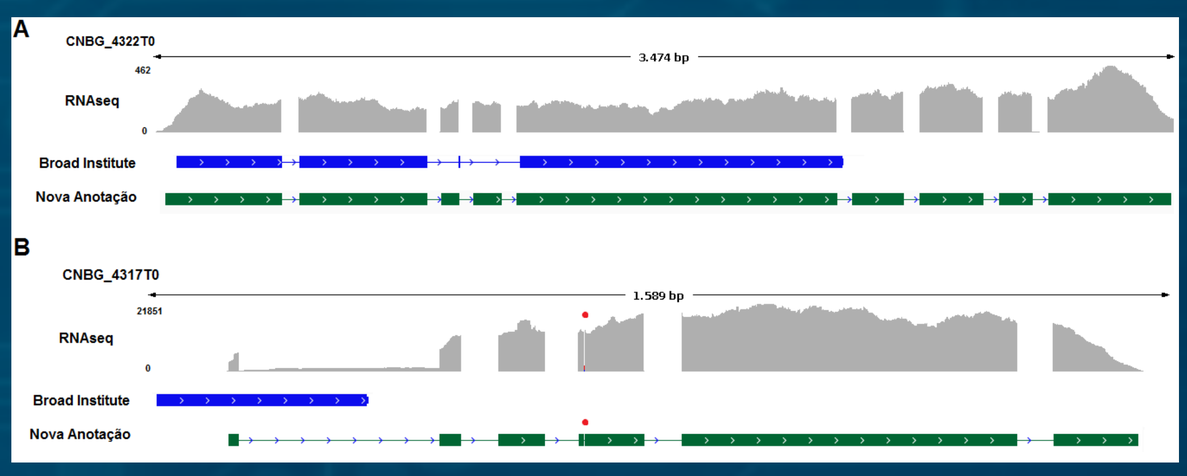

Figura 2: Visualização no software IGV de dois genes reanotados, CNBG_4322T0(A) e CNBG_4317T0 (B). Reads alinhados representados acima em cinza, predição automática representada em azul, predição por RNAseq representada em verde e pontos com erro de anotação no genoma apontados com ponto vermelho.

Perspectivas

1- Término do refinamento manual da estrutura dos transcritos;

2- Confirmar as estruturas preditas *in silico* através de RT-PCR de transcritos aleatoriamente selecionados.

APOIO FINANCEIRO

CAPES

CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico