UNIVERSIDADE DE MOGI DAS CRUZES

DAVID ACIOLE BARBOSA

REANOTAÇÃO FUNCIONAL DOS GENOMAS DOS PRINCIPAIS REPRESENTANTES DO GÊNERO *Paracoccidioides*

Mogi das Cruzes, SP

2017

UNIVERSIDADE DE MOGI DAS CRUZES

DAVID ACIOLE BARBOSA

REANOTAÇÃO FUNCIONAL DOS GENOMAS DOS PRINCIPAIS REPRESENTANTES DO GÊNERO *Paracoccidioides*

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Biotecnologia da Universidade de Mogi das Cruzes como parte dos requisitos para obtenção do título de Mestre em Biotecnologia.
Área de Concentração: Biotecnologia Aplicada à Saúde

Orientadora: Profa Dra Daniela Leite Jabes

Co-orientador: Prof Dr Luiz R. Nunes

Mogi das Cruzes, SP

2017

FINANCIAMENTO

FICHA CATALOGRÁFICA
Universidade de Mogi das Cruzes - Biblioteca Central

Barbosa, David Aciole

Reanotação funcional dos genomas dos principais representantes do gênero *paracoccidioides* / David Aciole Barbosa. – 2017.

78 f.

Dissertação (Mestrado em Biotecnologia) - Universidade de Mogi das Cruzes, 2017
Área de concentração: Biotecnologia aplicada à Saúde
Orientador: Prof.ª Dr.ª Daniela Leite Jabes

1. Paracoccidioides 2. Genoma 3. Reanotação funcional 4. Database I. Jabes, Daniela Leite

CDD 660.6

UMC
universidade de mogi das cruzes

Tel.: (011) 4798-7000
Fax: (011) 4799-5233
http://www.umc.br

ATAS

ATA DA SESSÃO PÚBLICA DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO EM BIOTECNOLOGIA DA UNIVERSIDADE DE MOGI DAS CRUZES

Às quatorze horas do dia vinte e oito de novembro de dois mil e dezessete, na Universidade de Mogi das Cruzes, realizou-se a defesa de dissertação "Reanotação funcional dos genomas dos principais representantes do gênero *Paracoccidioides*" para obtenção do grau de Mestre pelo(a) candidato(a) **David Aciole Barbosa**. Tendo sido o número de créditos alcançados pelo(a) mesmo(a) no total de 50 (cinquenta), a saber: 26 unidades de crédito em disciplinas de pós-graduação e 24 unidades de crédito no preparo da dissertação, o(a) aluno(a) perfaz assim os requisitos para obtenção do grau de Mestre. A Comissão Examinadora estava constituída dos Senhores Professores Doutores Daniela Leite Jabes, Regina Lúcia Batista da Costa de Oliveira e Denise Costa Arruda da Universidade de Mogi das Cruzes, sob a presidência da primeira, como orientadora da dissertação. A Sessão Pública da defesa de dissertação foi aberta pela Senhora Presidente da Comissão que apresentou o(a) candidato(a). Em seguida o(a) candidato(a) realizou uma apresentação oral da dissertação. Ao final da apresentação da dissertação, seguiram-se as arguições pelos Membros da Comissão Examinadora. A seguir a Comissão, em Sessão Secreta, conforme julgamento discriminado por cada membro, considerou o(a) candidato(a)

aprovado por unanimidade.
(aprovado(a) /reprovado(a)) (unanimidade/maioria)

Mogi das Cruzes, 28 de novembro de 2017.

Comissão Examinadora	Julgamento
Regina L. B. C. Oliveira	aprovado (aprovado(a)/reprovado(a))
Profª Drª Denise Costa Arruda	APROVADO (aprovado(a)/reprovado(a))
Profª Drª Daniela Leite Jabes	aprovado (aprovado(a)/reprovado(a))

Este trabalho é dedicado

Aos meus pais

Marlene Florêncio da Silva

e

Gaspar Aciole Barbosa

À minha noiva

Nathalia Sampaio da Silva

Meu Reverso e Complementar

AGRADECIMENTOS

À Profa Dra Regina Costa de Oliveira agradeço por ter me recebido no Núcleo Integrado de Biotecnologia como aluno de iniciação científica e pela atenção que sempre teve desde os tempos das aulas da graduação.

Ao Prof Doutor Luiz R. Nunes agradeço pela paciência e habilidade raras com que explica todas as questões, reações, metodologias e vias metabólicas do nosso dia-a-dia no laboratório e pelas excelentes sugestões para contornar os desafios que encontramos no meio do caminho da ciência. Não posso deixar de agradecer por todas as broncas e puxões de orelha que tenho certeza que me ajudaram a crescer.

À minha orientadora Profa Doutora Daniela Leite Jabes pela disposição e alegria com que chega todo dia para acompanhar e ajudar no andamento dos trabalhos, mesmo com tantas aulas para ministrar. Agradeço pela paciência infinita em tentar entender minhas dúvidas e os meus raciocínios que ainda nem tinham chegado a ser dúvidas propriamente ditas. Agradeço por todas as formas de colaboração que a minha orientadora me deu. Elas foram inestimáveis.

Ao Prof Dr Moacir Wuo, que colaborou muito na resolução das dúvidas sobre a estrutura da dissertação.

Ao Mestre Fabiano Menegidio, que se tornou meu novo colega de análises e vem me ajudando imensamente no aprendizado da Bioinformática.

À Profa Dra Valquíria Campos Alencar, que sempre está disposta a falar dos experimentos mais interessantes e ajudar a cuidar do laboratório sempre com bom humor.

Aos meus colegas de laboratório, as nibetes Renata Ozelami Vilas Boas, Tabata Hiromi Usuginu (Tati), Juliana de Fátima, Yara Natércia e ao nibete Yan Teixeira de Oliveira, amigos e parceiros que sempre me ajudaram muitíssimo nos experimentos e nos planejamentos.

Agradeço a minha família, meus pais, meus irmãos e minha noiva, Nathalia Sampaio da Silva, que sempre ajudaram-me a contornar dificuldades.

Aos órgãos de fomento CAPES, FAEP e FAPESP, agradeço pela contribuição direta ou indireta de recursos financeiros.

Agradeço à Universidade de Mogi das Cruzes pelo espaço e serviços fornecidos durante a realização do projeto.

"'O universo (que outros chamam a Biblioteca) é composto de um número indefinido e talvez infinito de galerias hexagonais ... Existem cinco prateleiras para cada uma das paredes do hexágono; cada estante contém trinta e cinco livros de formato uniforme; cada livro é de quatrocentos e dez páginas, cada página, de quarenta linhas, cada linha, de algumas oitenta letras que são de cor preta.'

Essa biblioteca infinita conteria qualquer livro possível, mas também infinitas combinações de letras sem sentido. Encontrar os livros reais entre a pilha de textos sem sentido é uma excelente metáfora do desafio que constitui a informação extraída do fluxo de dados na era pós-genômica".

Retirado do site Babelomics, uma plataforma computacional de análises de bioinformática, da qual o nome foi inspirado na obra "A biblioteca de Babel", do escritor argentino Jorge Luis Borges.

RESUMO

A paracoccidioidomicose (PCM) é uma micose sistêmica causada pelos fungos termodimórficos *Paracoccidioides brasiliensis* e *Paracoccidioides lutzii*. A PCM é prevalente na América Latina e apresenta distribuição geográfica heterogênea, com áreas endêmicas localizadas principalmente na Argentina, Brasil, Colômbia, Equador e Venezuela, embora já haja confirmação de pelo menos 60 casos de paracoccidioidomicose na América do Norte, Europa, Oriente Médio, Ásia e África. O Brasil representa 80% do total de casos reportados no mundo, e ocupa o primeiro lugar em mortes causadas por micoses sistêmicas. A mortalidade por PCM varia de 2 a 23%, mas quando associada à AIDS, pode chegar a 30%. Em 2011, o genoma completo de três isolados (Pb03, Pb18 e Pb01) representando diferentes grupos filogenéticos de *Paracoccidioides* foi publicado, por meio da tecnologia Sanger. Em 2014, houve uma atualização dos genomas utilizando tecnologias de sequenciamento de nova geração (NGS). Em 2016, outros dois genomas de *Paracoccidioides* foram divulgados para os isolados Pb300 e PbCnh. Dessa forma, foram disponibilizados genomas de representantes dos cinco grupos filogenéticos de *Paracoccidioides* a saber, Pb18 (S1), Pb3 (PS2), Pb300 (PS4), PbCnh (PS3) e Pb01 (Pl). Os dados do sequenciamento e anotação das linhagens-referência de *P. brasiliensis* (Pb18, Pb03) e *P. lutzii* (Pb01) eram centralizados em um banco de dados gerado pela equipe do *Broad Institute*, fornecendo ferramentas como o BLAST, mapas de sintenia e busca de genes por palavra-chave. Entretanto, nenhum *database* exclusivo de *Paracocicidioides* está disponível atualmente, uma vez que o banco mantido pelo Broad Institute foi descontinuado em 2016. Atualmente, dados referentes a esses sequenciamentos encontram-se depositados em bancos públicos, como NCBI RefSeq/GenBank, Ensembl, MycoCosm e FungiDB. No entanto, anotações funcionais a que estes genomas foram submetidos basearam-se em critérios excessivamente restritivos e, como consequência, não apresentam descrição de função para 70 - 90% das proteínas codificadas por membros do gênero *Paracoccidioides*. Nesse sentido, o presente trabalho utilizou-se do *software Blast2GO* e o *database* DAVID como ferramentas para reanotação funcional dos genomas de *Paracoccidioides*. Estes programas possuem recursos integrados das principais bases de dados de informações estruturais e evolutivas de genes e proteínas, como InterPro, Pfam e Gene Onthology. Foi possível alcançar aumento de anotação funcional entre 34,7 a 61% do total de proteínas nos genomas das linhagens-referência já sequenciadas. Esses dados foram organizados em tabelas no formato Excel, como também arquivos multi-fasta e GFF3. Os dados organizados foram disponibilizados em um banco de dados relacional construído com o intuito de centralizar, armazenar e disponibilizar *online* os dados gerados. Nesse sentido, apresentamos o ParaDB, acessado pelo endereço **http://paracoccidioides.com**, um portal genômico dedicado ao gênero *Paracoccidioides*, que conta com reanotação acurada e atualizada das informações gênicas dos cinco isolados sequenciados, além de, ferramentas de BLAST e busca de genes por palavras-chave e/ou códigos de identificação. Portanto, as informações atualizadas sobre as funções dos genes presentes no genoma de *Paracoccidioides* permitem melhoria de qualidade nas análise de dados para este gênero, especialmente as de larga escala, colaborando junto à comunidade científica para melhor acesso aos dados, que estarão centralizados em um único repositório.

Palavras-chave: *Paracoccidioides*. Genoma. Reanotação funcional. *Database*.

ABSTRACT

Paracoccidioidomycosis (PCM) is a systemic mycosis caused by the thermodymorphic fungi *Paracoccidioides brasiliensis* and *Paracoccidioides lutzii*. PCM is prevalent in Latin America and has a heterogeneous geographical distribution, with endemic areas located mainly in Argentina, Brazil, Colombia, Ecuador and Venezuela. Moreover, there are at least 60 confirmed cases of paracoccidioidomycosis in North America, Europe, Middle East, Asia and Africa. Brazil represents 80% of the total cases reported in the world, where the disease occupies the first place in deaths caused by systemic mycoses. Mortality by PCM ranges from 2 to 23%, but may reach 30%, when associated with AIDS. The complete genome of three *Paracoccidioides* isolates (Pb03, Pb18 and Pb01) representing different phylogenetic groups of these fungi was published in 2011, using Sanger sequencing technology. In 2014, there was an update in these genomes, using Next Generation Sequencing (NGS) technologies. In 2016, two additional *Paracoccidioides* genomes were reported. Thus, genomes of representatives for the five main phylogenetic groups of *Paracoccidioides*, namely, Pb18 (S1), Pb3 (PS2), Pb300 (PS4), PbCnh (PS3) and Pb01 (Pl) are currently available. A *database* generated by the Broad Institute centralized information regarding sequencing and annotation data from *P. brasiliensis* (Pb18, Pb03) and *P. lutzii* (Pb01) reference lineages, providing tools such as BLAST, synteny maps and keyword search for genes. However, this *database* was discontinued in 2016 and no other *Paracocicidioides database* is currently available. The only repositories to contain genomic information regarding these fungi are public banks such as NCBI's RefSeq/GenBank, Ensembl, MycoCosm and FungiDB. However, the automated computational tools used for functional annotation are excessively restrictive and do not present function description for 70 – 90% of all protein-coding genes identified in *Paracoccidioides* spp. In this sense, the present study used Blast2GO *software* and the DAVID *database* as tools for functional reannotation of the currently available *Paracoccidioides* genomes. These programs have built-in capabilities to scan key structural and evolutionary *database*s of genes and proteins, such as InterPro, Pfam and Gene Ontology. It was possible to achieve gain of functional annotation for 34.7 to 61% of the total proteins present in the genomes of the sequenced reference lineages. These data were organized into tables in Excel format, as well as multi-fasta and GFF3 files. The organized data was uploaded into a relational *database* built in order to centralize, store and make it available online. Thus, this work presents ParaDB, a genomic portal dedicated to the genus *Paracoccidioides* (available at the url **http://paracoccidioides.com**), which provides accurate and updated reannotations, along with search engines based on BLAST, keywords and identification gene codes. We expect that the, up-to-date information regarding the genomes of *Paracoccidioides* spp will assist in further analyses regarding the general biology of these important fungi , especially in the case of large-scale experimental evaluations, collaborating with the scientific community for better access to the data, centralized in exclusive dedicated repository.

Keywords: *Paracoccidioides*. Genome. Functional reanotation. Database.

LISTA DE FIGURAS

Figura 1 – Pacientes com lesões causadas pela PCM..... 17

Figura 2 – Áreas geográficas de endemismo da paracoccidioidomicose na América Latina..... 18

Figura 3 – Transição morfológica induzida pela variação de temperatura em *P. brasiliensis* Pb18..... 20

Figura 4 – Filogenia atual e distribuição dos grupos de *Paracoccidioides*..... 21

Figura 5 – Fontes de dados que compõem as análises por *Blast2GO* e o *database* DAVID..... 31

Figura 6 – Representação dos principais Modelos de Sistemas de Bancos de Dados - MSBDs..... 32

Figura 7 – Esquema utilizado para a reanotação funcional dos genomas publicamente disponíveis para os cinco grupos filogenéticos do gênero *Paracoccidioides*..... 38

Figura 8 – Exemplo do arquivo multi-fasta gerado para Pb18 mostrando três sequências..... 56

Figura 9 – Exemplo do arquivo GFF3 gerado para Pb18 mostrando as principais características das sequências..... 57

Figura 10 – Imagem de teste da página inicial do banco de dados ParaDB..... 58

Figura 11 – Imagem de teste de busca de palavras na página de *P. brasiliensis* isolado Pb18 no banco de dados ParaDB..... 58

Figura 12 – Imagens da versão para dispositivos móveis do banco de dados ParaDB..... 59

Figura 13 – Imagens das análises de BLAST realizadas no banco de dados ParaDB..... 60

LISTA DE GRÁFICOS

Gráfico 1 – Fontes de anotação funcional de *Paracoccidioides brasiliensis* isolado Pb18..... 47

Gráfico 2 – Anotação funcional do genoma CDS de *Paracoccidioides brasiliensis* isolado Pb18..... 47

Gráfico 3 – Fontes de anotação funcional para *Paracoccidioides lutzii* isolado Pb01..... 48

Gráfico 4 – Anotação funcional do genoma CDS de *Paracoccidioides lutzii* isolado Pb01.....49

Gráfico 5 – Fontes de anotação funcional para *Paracoccidioides brasiliensis* isolado Pb03..... 50

Gráfico 6 – Anotação funcional do genoma CDS de *Paracoccidioides brasiliensis* isolado Pb03..... 50

Gráfico 7 – Fontes de anotação funcional para *Paracoccidioides brasiliensis* isolado Pb300..... 51

Gráfico 8 – Anotação funcional do genoma CDS de *Paracoccidioides brasiliensis* isolado Pb300..... 52

Gráfico 9 – Fontes de anotação funcional para *Paracoccidioides brasiliensis* isolado PbCnh..... 53

Gráfico 10 – Anotação funcional do genoma CDS de *Paracoccidioides brasiliensis* isolado PbCnh.....53

LISTA DE QUADROS

Quadro 1 – Parâmetros usados para a etapa de BLAST no *software Blast2GO*..... 39

Quadro 2 – Parâmetros usados para a etapa de *Gene Onthology Annotation* no *software Blast2GO*…..... 40

Quadro 3 – Parâmetros usados para a etapa de *InterProScan* no *software Blast2GO*..... 41

Quadro 4 – Categorias Gene Onthology e domínios de proteína utilizados na anotação funcional de *Paracoccidioides* utilizando a plataforma DAVID..... 42

Quadro 5 – Colunas da tabela construída para organização os dados de anotação funcional obtidos para Pb18 e descrições do conteúdo de cada coluna..... 55

LISTA DE ABREVIATURAS E SIGLAS

ANSI = *American National Standards Institute* (tradução: Instituto Nacional Americano de Padronizações)

AspGD = *Aspergillus* Genome *Database*

BLAST = *Basic Local Alignment Search Tool* (tradução: Ferramenta de Busca de Alinhamento Local Básico)

BLASTn = BLAST de nucleotídeos

BLASTp = BLAST de aminoácidos

CATH = *Class/Architecture/Topology/Homologous superfamily* (tradução: Classe/Arquitetura/Topologia/superfamílias Homólogas)

CAZY = Carbohydrate-Active enZYmes *Database* (tradução: *Database* de enZImas Carboidrato-Ativas)

CDD = *Conserved Domain Database* (tradução: *Database* de Domínios Conservados)

CDS = *Coding DNA Sequence* (tradução: Sequência de DNA Codificante)

CGD = *Candida* Genome *Database*

CMS = *Content Manager System* (tradução: Sistema de Gerenciamento de Conteúdo)

DAVID = *Database* for Annotation, Visualization and Integrated Discovery (tradução: *Database* para Anotação, Visualização e Descoberta Integrada)

DDBJ = DNA *DataBank of Japan* (tradução: Banco de Dados DNA do Japão)

EC-*number* = *Enzyme Comission number* (tradução: número da Comissão de Enzimas)

ENA = *European Nucleotide Archive* (tradução: Depósito de Nucleotídeos da Europa)

EXP = Inferido por Experimento

FTP = *File Transfer Protocol* (tradução: Protocolo de Transferência de Arquivo)

GFF = *General Feature Format* (tradução: Formato de Características Gerais)

GFF3 = *General Feature Format version 3* (tradução: Formato de Características Gerais versão 3)

GO = *Gene Ontology* (tradução: Ontologia Gênica)

GTF = *General Transfer Format* (tradução: Formato de Transferência Geral)

HAMAP = *High-quality Automated and Manual Annotation of microbial Proteomes* (tradução: Anotação Manual e Automatizada de Alta-qualidade de Proteomas microbianos)

HGNC = HUGO *Gene Nomenclature Committee* (tradução: Comitê de Nomenclatura de Genes HUGO)

HistoBase = *Histoplasma* Base

HMMs = *Hidden Markov Models* (tradução: Modelos Ocultos de Markov)

HSP = *High Scoring Pairs* (tradução: Pares de Alto Escore)

IBA = Inferido por aspecto Biológico do Ancestral

IBD = Inferido por aspecto Biológico do Descendente

IC = Inferido pelo curador

IDA = Inferido por Ensaio Direto

IDs = Identificadores

IEA = Inferido de Anotação Eletrônica

IEP = Inferido por Padrão de Expressão

IGC = Inferido por Contexto Genômico

IGI = Inferido por Interação Genética

IKR = Inferido por Resíduos Chave

IMP = Inferido por Fenótipo Mutante

INSDC = *International Nucleotide Sequence Database Collaboration* (tradução: Colaboração Internacional de *Database* de Sequência de Nucleotídeos)

IPI = Inferido por Interação Física

ISA = Inferido por Alinhamento da Sequência

ISM = Inferido por Modelo da Sequência

ISO = Inferido por Ontologia da Sequência

ISS = Inferido por Análise Computacional Revisada

ISS = Inferido por Divergência Rápida

ISS = Inferido por Sequência ou Similaridade estrutural

Mb = Mega pares de bases

MBD = Modelos de Bancos de Dados

MIT = *Massachusetts Institute of Technology* (tradução: Instituto de Tecnologia de Massachussets)

MSBDs = Modelos de Sistema de Bancos de Dados

NAS = Declaração de autor Não Rastreável

NCBI = National Center for Biological Information

ND = Nenhum código de evidência de dados biológicos disponível

NGS = sequenciamento de nova geração

NR = Não Registrado

PANTHER = *Protein ANalysis THrough Evolutionary Relationships* (tradução: Análises de Proteínas Através de Relacionamentos Evolutivos)

ParaDB = *Paracoccidioides Database*

Pb01 = *Paracoccidioides lutzii* isolado 01

Pb03 = *Paracoccidioides brasiliensis* isolado 03

Pb18 = *Paracoccidioides brasiliensis* isolado 18

Pb300 = *Paracoccidioides brasiliensis* isolado 300

PbCnh = *Paracoccidioides brasiliensis* isolado Cnh

PCM = Paracoccidioidomicose

PDF = *Portable Document Format* (tradução: Formato de Documento Portátil)

PGAP = *Eukaryotic Genome Annotation Pipeline* (tradução: *Pipeline* de Anotação de Genoma Eucariótico)

PGAP = *Prokaryotic Genome Annotation Pipeline* (tradução: *Pipeline* de Anotação de Genoma Procariótico)

PHP = *Hypertext Processor* (tradução: Processador de Hipertexto)

PIR/PSD = *Protein Information Resource/Protein Sequence Database* (tradução: Recurso de Informação de Proteína/ *Database* de Sequência de Proteína)

RNA-seq = Sequenciamento de RNA

SFLD = Struc*ture-Function Linkage Database* (tradução: *Database* de Ligação Estrutura-Função)

SGBD **= S**istemas de Gerenciamento de Banco de Dados

SGBDR = Sistemas de Gerenciamento de Banco de Dados Relacional

SGD = *Saccharomyces* Genome *Database*

SMART = *Simple Modular Architecture Research Tool* (tradução: Ferramenta de Pesquisa de Arquitetura Modular Simples)

SNPs = *Single Nucleotide Polymorphisms* (tradução: Polimorfismos de Nucleotídeo Único)

SQL = *Structured Query Language* (tradução: Linguagem de Consulta Estruturada)

TAS = Declaração de Autor Rastreável

WPCLI = *Wordpress Command List Interface* (tradução: Interface de Lista de Comando do *Wordpress*)

SUMÁRIO

1 INTRODUÇÃO 15
1.1 Considerações Gerais 15
1.2 Paracoccidioidomicose 16
1.3 *Paracoccidioides* spp: Agentes Causadores Da Paracoccidioidomicose 19
1.4 Projeto Genoma De *Paracoccidioides* 22
1.5 Anotação Funcional 26
1.6 Bancos De Dados Biológicos 30
2 OBJETIVOS 36
2.1 Objetivos Gerais 36
2.2 Objetivos Específicos 36
3 MÉTODO 37
3.1 Obtenção E Manipulação Dos Genomas 37
3.2 Processo De Reanotação Funcional 37
3.3 Curadoria Manual 43
3.4 Estruturação Dos Dados Gerados 43
3.5 Construção Do Banco De Dados Relacional 44
4 RESULTADOS 46
4.1 Identificação Das Funções Dos Genes 46
4.1.1 *Paracoccidioides brasiliensis* Isolado Pb18 46
4.1.2 *Paracoccidioides lutzii* isolado Pb01 48
4.1.3 *Paracoccidioides brasiliensis* Isolado Pb03 49
4.1.4 *Paracoccidioides brasiliensis* Isolado Pb300 50
4.1.5 *Paracoccidioides brasiliensis* Isolado PbCnh 52
4.2 Estruturação Dos Dados Obtidos 54
4.3 Avaliação E Descrição Do *Database* 57
5 DISCUSSÃO 60
6 CONCLUSÕES E PERSPECTIVAS 68
REFERÊNCIAS 69

1 INTRODUÇÃO

1.1 CONSIDERAÇÕES GERAIS

As infecções fúngicas sistêmicas são infecções causadas por fungos que comprometem vários órgãos sistemas do hospedeiro e são vistas como um importante problema para a saúde pública devido ao grande número de indivíduos por elas afetados, bem como a severidade do quadro apresentado pelos pacientes. Grande parte destas infecções é causada por fungos dimórficos, como *Blastomyces dermatitidis*, *Coccidioides* spp, *Histoplasma capsulatum* e *Paracoccidioides* spp. (MARESCA e KOBAYASHI, 1989; ROONEY e KLEIN, 2002; LUPETTI *et al.,* 2002; KLEIN e TEBBETS, 2007).

Fungos dimórficos são aqueles que podem assumir duas formas distintas, micélio ou levedura, dependendo das condições a que são submetidos. Diversos fatores podem estar associados à interconversão da morfologia fúngica, como por exemplo, temperatura, pH, níveis de glicose, fonte de nitrogênio, presença de metais de transição e agentes quelantes, entre outros (HORNBY *et al.,* 2004).

Estima-se que em torno de 1,2 bilhão de pessoas em todo o mundo sofram de doenças causadas por fungos, sendo que algumas delas são invasivas e/ou crônicas, além de difíceis para diagnosticar e tratar. Nesse contexto, estima-se que 1,5 a 2 milhões de pessoas morram a cada ano devido a infecções fúngicas (BROWN *et al*., 2012; DENNING e BROMLEY, 2015). O clima e a riqueza da biodiversidade da América Latina, contribuem para a ocorrência de diversos fungos patogênicos, responsáveis por micoses endêmicas, que desempenham importante papel na saúde pública da região, com destaque para histoplasmose, coccidioidomicose e paracoccidioidomicose (PCM) (COLOMBO *et al.,* 2011).

Este trabalho se concentra nas espécies *Paracoccidioides brasiliensis e Paracoccidioides lutzii,* ambos agentes etiológicos da micose sistêmica granulomatosa denominada paracoccidioidomicose, que acomete principalmente trabalhadores de atividades rurais imunossuprimidos em diversos países da América Latina, com maior prevalência no Brasil, Colômbia, Venezuela e Argentina (TEIXEIRA *et al.,* 2014), embora também, haja registros de casos confirmados na América do Norte, Europa, Oriente Médio e Ásia (BOCCA *et al*., 2013; MARTINEZ, 2015).

1.2 PARACOCCIDIOIDOMICOSE

O primeiro caso de PCM foi descrito pelo médico brasileiro Adolpho Lutz em 1908, ao verificar graves lesões na mucosa oral de pacientes no Instituto Bacteriológico de São Paulo (atualmente Instituto Adolfo Lutz, em homenagem ao pesquisador). Em 1912, o médico bacteriologista italiano Alfonso Splendore classificou o fungo como *Zymonema brasiliensis*. Somente no ano de 1930, o micologista paulista Floriano Paulo de Almeida definiu o nome aceito atualmente para designar o fungo. Por isso, a micose também foi chamada de doença de Lutz-Splendore-Almeida antes de ser finalmente definida sua nomenclatura em uma reunião de micologistas em Medelín, na Colômbia, em 1971 (MOREIRA *et al.*, 2008; MARQUES, 2012).

A PCM é uma micose sistêmica causada pelos fungos termodimórficos *Paracoccidioides brasiliensis* e *P. lutzii* e tem início com a inalação de propágulos miceliais como hifas e conídios (LACAZ *et al.*, 1991). Após entrar em contato com o epitélio pulmonar, o fungo passa por transição para a forma de levedura e em seguida pode ocorrer sua multiplicação no interior do organismo hospedeiro. Do trato respiratório, as células fúngicas podem disseminar-se para outros órgãos e sistemas através da corrente sanguínea, nos quais podem desencadear o surgimento de lesões secundárias nodulares e/ou ulcerativas (Figura 1) (BRUMMER *et al.*,1993; RAMOS E SILVA e SARAIVA, 2008).

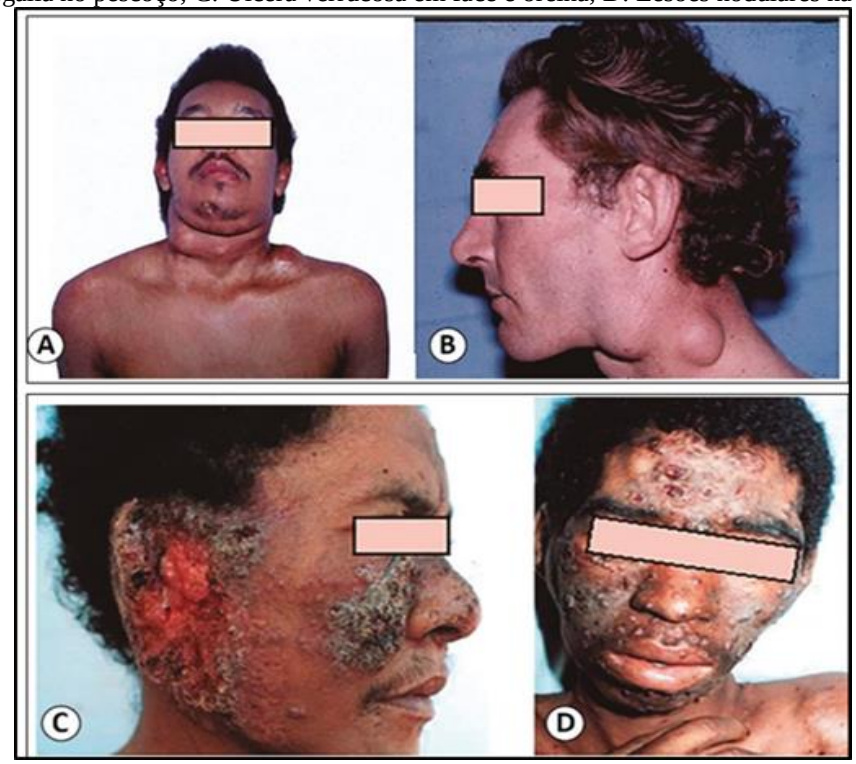
Entre as principais drogas de escolha para o tratamento da PCM em humanos estão sulfonamidas (sulfametoxazol-trimetoprim), derivados imidazólicos (itraconazol, cetoconazol, fluconazol) e polienos (anfotericina B). Estas substâncias podem ser utilizadas isoladamente ou em associações, dependendo da necessidade e evolução clínica dos pacientes (VISBAL *et al.*, 2005). Contudo, independentemente da classe farmacológica da substância escolhida para o tratamento, o período de terapia necessário para que critérios de cura sejam atingidos é extenso e está diretamente relacionado à severidade dos efeitos colaterais observados, especialmente no que se refere a lesões renais e hepáticas, visto o elevado potencial nefrotóxico e hepatotóxico destas substâncias (SHIKANAI-YASUDA *et al.*, 2006; TUON *et al.*, 2013).

Além disso, o isolamento de fungos resistentes tem se tornado cada vez mais frequente nos últimos anos, sendo possível encontrar na literatura diversas descrições de resistência para isolados de diferentes fungos patogênicos, incluindo *Candida* spp, *Cryptococcus neoformans*, *Aspergillus* spp, *Histoplasma capsulatum* e *P. brasiliensis*, que apresentou resistência simultânea a drogas de algumas classes farmacológicas diferentes como cetoconazol, associação sulfametoxazol-trimetoprim e anfotericina B, utilizadas no tratamento

medicamentoso da PCM (ALEXANDER e PERFECT, 1997; HAHN *et al.*, 2003; KONTOYIANNIS e LEWIS, 2002; BALTAZAR *et al.*, 2015; SANGLARD, 2016).

A resistência em fungos é caracterizada pela falência da terapia medicamentosa, que pode ser mensurada pelo aumento da Concentração Inibitória Mínima (MIC) comparada a valores obtidos para organismos susceptíveis tomados como referência (ALEXANDER e PERFECT, 1997; SANGLARD e ODDS, 2002; SANGLARD, 2002). Os mecanismos mais frequentemente envolvidos na resistência são: diminuição da afinidade dos alvos celulares pelas drogas, degradação ou inativação enzimática das substâncias e, principalmente, a superexpressão de genes codificadores de proteínas que controlam o efluxo de drogas (WHITE *et al.*, 1998). O surgimento de fenótipos de resistência a drogas em fungos é um obstáculo para o tratamento farmacológico das micoses sistêmicas e incentiva a investigação da atividade antimicrobiana de novos compostos, principalmente quando estes possuem mecanismos de ação distintos dos antifúngicos convencionais (VITALE *et al.*, 2007). Substâncias com tais características possuem potencial farmacológico isoladamente ou associadas a drogas já empregadas no tratamento de infecções fúngicas como a PCM.

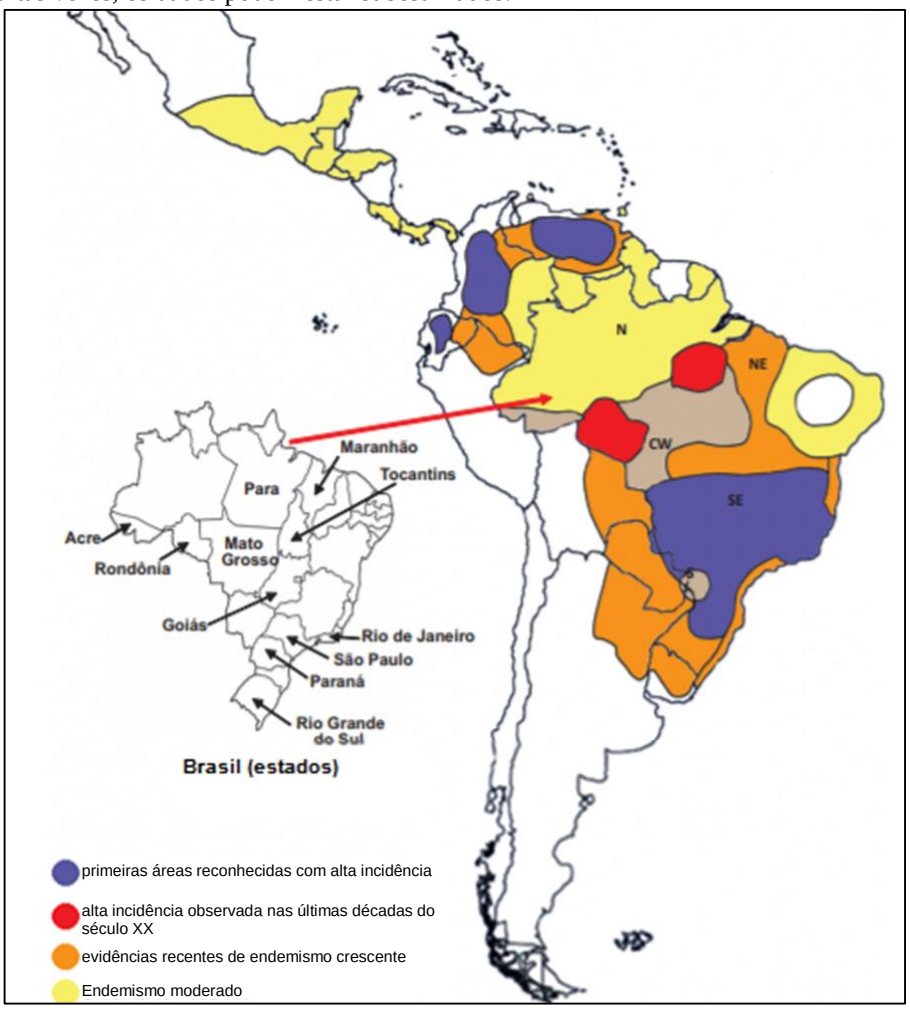
Figura 1 – Pacientes com lesões causadas pela PCM. A: Massa ganglionar em pescoço, mandíbula e clavícula; B: Linfadenomegalia no pescoço; C: Úlcera verrucosa em face e orelha; D: Lesões nodulares na face.



FONTE: Adaptado de SHIKANAI-YASUDA *et al.*, 2017.

A PCM é prevalente na América Latina e apresenta distribuição geográfica heterogênea, com áreas endêmicas localizadas principalmente na Argentina, Brasil, Colômbia, Equador e Venezuela (Figuras 2) (SHIKANAI-YASUDA *et al.*, 2006). O Brasil representa 80% do total de casos reportados no mundo, onde a doença ocupa o primeiro lugar em mortes causadas por micoses sistêmicas, e a oitava posição entre as doenças predominantemente crônicas, considerando males infecciosos e parasitários. A mortalidade por PCM varia de 2 a 23% em regiões endêmicas, embora, quando associada à casos de imunossupressão, como em pacientes portadores de AIDS ou com outros quadros de imunossupressão, o índice de mortes possa chegar a 30% (COUTINHO *et al.*, 2002; BELLISSIMO-RODRIGUES *et al.*, 2011).

Figura 2 – Áreas geográficas de endemismo da paracoccidioidomicose na América Latina. A paracoccidioidomicose, ou PCM, é endêmica na América Latina e apresenta incidência heterogênea ao longo desta área, onde o Brasil soma cerca de 80% do total de casos notificados. As regiões em azul indicam as primeiras áreas de com alta incidência de PCM e as regiões em vermelho correspondem aos focos de alta incidência da micose observados no final do século XX. Já os locais em laranja e amarelo representam áreas de endemismo crescente e moderado, respectivamente. No entanto, a PCM é uma doença negligenciada e as notificações de casos não são precisas, e, muitas vezes, os dados podem estar subestimados.



FONTE: Modificado de SHIKANAI-YASUDA *et al*., 2017.

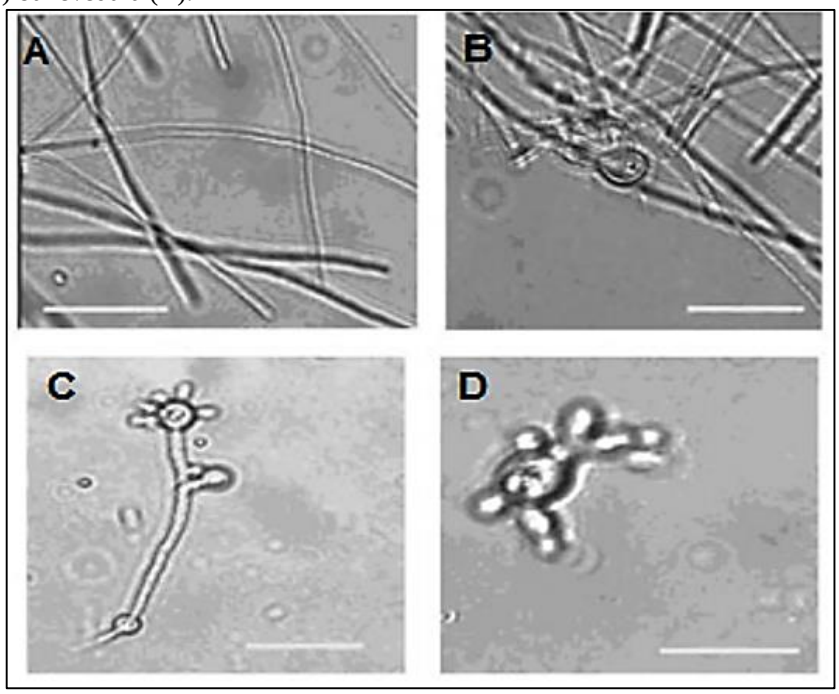
Além das áreas endêmicas, já há confirmação de pelo menos 60 casos de paracoccidioidomicose nos Estados Unidos, Canadá, Espanha, Alemanha, Inglaterra, França, Holanda, Oriente Médio, Japão e África. Essas ocorrências estão relacionadas ao turismo, uma vez que os pacientes com PCM diagnosticada nesses países estiveram em zonas endêmicas como Brasil, Venezuela, Bolívia, Equador, Argentina e Paraguai (MARTINEZ, 2015), o que gera preocupação também para os órgãos de saúde de demais locais do mundo.

1.3 *Paracoccidioides* spp: AGENTES CAUSADORES DA PARACOCCIDIOIDOMICOSE

Paracoccidioides é um gênero de fungos ascomicetos taxonomicamente representado na Ordem Onygenales, família Onygenaceae. Esta família inclui importantes gêneros de fungos patogênicos como *Blastomyces*, *Coccidioides, Histoplasma*, e *Lacazia* (BAGAGLI *et al.*, 2008). *Paracoccidioides* é composto pelas espécies *P. brasiliensis* e *P. lutzii*, que apresentam termodimorfismo, ou seja, transição dimórfica devida à variação de temperatura. Em temperaturas próximas a 37º C, o fungo é caracterizado como uma levedura de 5 a 25 µm de diâmetro e parede dupla, com brotamentos (comumente chamado de "roda-de-leme") e em meio de cultura sólido seu crescimento apresenta coloração creme e aspecto cerebriforme, já o micélio ocorre em temperaturas próximas a 26º C, com hifas septadas ramificadas com clamidósporos, conídios (estruturas relacionadas à reprodução assexuada) e sem corpo de frutificação (PALMEIRO *et al.*, 2005; MOREIRA *et al.*, 2008). Em temperaturas intermediárias ocorrem formas transitórias, como pseudo-hifas. O dimorfismo também é relacionado à virulência, uma vez que os conídios presentes nas hifas são a forma infectante e, após estabelecerem-se no hospedeiro, a 37º C, existe a conversão para a estrutura leveduriforme (Figura 3) (NUNES *et al.*, 2005; KLEIN e TEBBETS, 2007).

Paracoccidioides é tido como sendo anamórfico ou mitospórico, o que diz respeito a uma condição de reprodução assexual por meio de brotamentos e conídios. Entretanto, estudos recentes propõe que, na realidade, diferentes estratégias de reprodução podem ocorrer nos representantes do gênero. O principal *locus* envolvendo reprodução sexuada de fungos é o *mating-type*, com os idiomorfos MAT1-1 e MAT1-2. Eventos de recombinação foram detectados em *P. lutzii* e *P. brasiliensis* e o *locus* MAT (*mating type* -MAT1-1 ou MAT1-2), foi identificado nos genomas sequenciados, sugerindo a presença de um ciclo sexual em *Paracoccidioides* (TEIXEIRA *et al.*, 2014).

Figura 3 – Transição morfológica induzida pela variação de temperatura em *P. brasiliensis* Pb18. Micélios cultivados em meio líquido, durante a fase exponencial de crescimento, foram induzidos a sofrer transição morfológica, alterando a temperatura de incubação de 26º C para 37º C. Ao longo da transição, as unidades morfológicas foram classificadas como hifas (A), hifas em diferenciação (caracterizadas pelo desenvolvimento de células tipo clamidósporos) (B), levedura em transformação (caracterizado pela produção de múltiplos brotos pelo clamidósporo) (C) ou levedura (D).

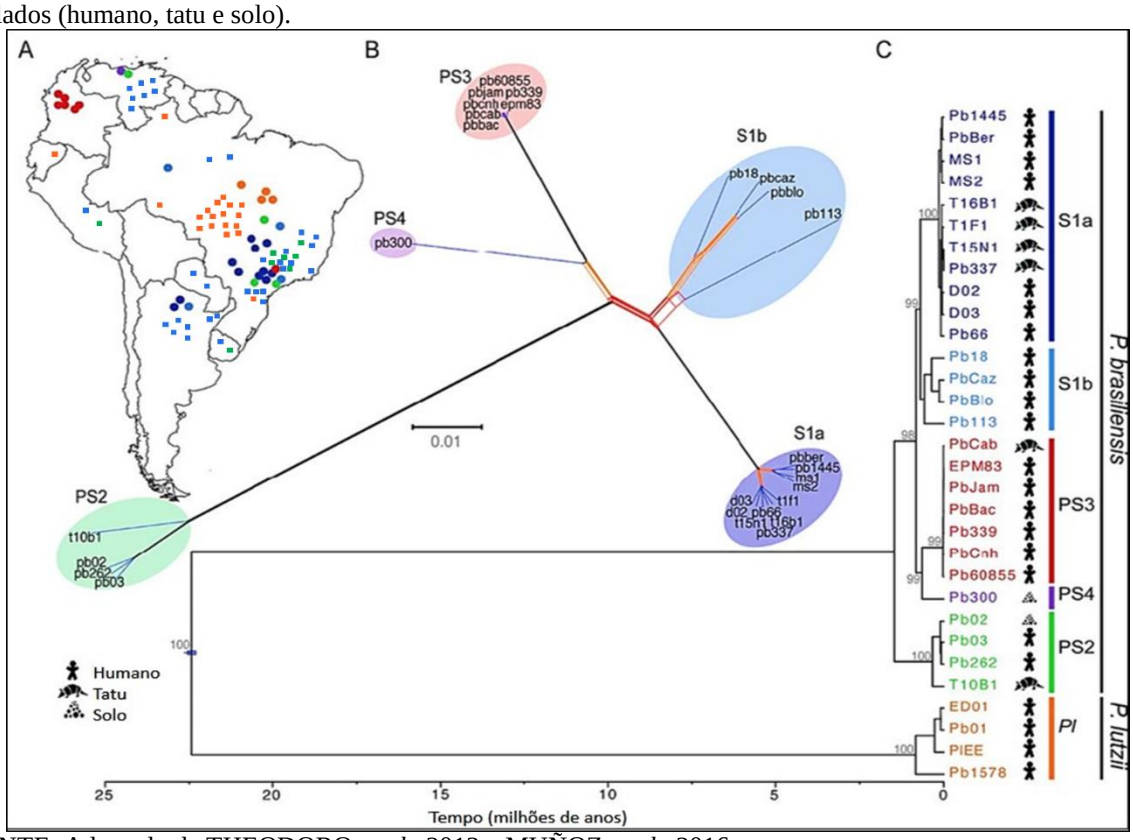


FONTE: Modificado de Nunes *et al*., 2005.

Desde a sua descrição por Adolfo Lutz (1908), *Paracoccidioides* era caracterizado como constituído por uma só espécie: *P. brasiliensis*, embora grande diversidade genética possa ser verificada entre os isolados. Além disto, existe grande variação nos níveis de virulência desses isolados, como verificado em modelos experimentais estabelecidos (SINGER-VERMES *et al*., 1989). Estes isolados foram distribuídos em cinco grupos distintos (Figura 4A) que eram tratados de maneira extra-oficial como espécies crípticas de *P. brasiliensis* (revisto em MUÑOZ *et al.*, 2016). A partir de 2009, no entanto, o grupo P1 (que tem como principal representante o isolado Pb01) passou a ser considerado como uma espécie distinta, denominada *Paracoccidioides lutzii*, devido a diferenças genéticas significativas quando comparado com os demais grupos (TEIXEIRA *et al.,* 2009). Estes, por sua vez, continuam sendo tratados como uma mesma espécie, *P. brasiliensis*, ainda que significativa divergência evolutiva seja verificada entre eles (Figura 4B). O grupo Pl contem isolados distribuídos no Brasil e Equador. O grupo S1 é o maior e abrange isolados oriundos de Brasil, Argentina, Venezuela, Peru e Paraguai. Estudos recentes chegam, inclusive, a sugerir sua divisão em dois subgrupos, denominados S1a e S1b, haja vista a grande divergência observada entre seus isolados

(MUÑOZ *et al.*, 2016). Já o grupo PS3, ocorre em sua maioria na Colômbia, com apenas um isolado (Pb339) no sudeste do Brasil. PS2 é composto por seis isolados oriundos de Brasil e Venezuela (THEODORO *et al.*, 2012). Finalmente, o grupo PS4, recentemente proposto, é constituído por um único isolado altamente divergente, identificado em uma região da Venezuela (MUÑOZ *et al.*, 2016).

Figura 4 – **Filogenia atual e distribuição dos grupos de *Paracoccidioides***. A: Distribuição geográfica dos isolados de cada grupos filogenético de *Paracoccidioides*. Os círculos destacam os isolados sequenciados; B: Rede filogenética mostrando isolados de cada grupo filogenético – PS1 dividido em S1a destacado em azul escuro e S1b em azul claro, segundo recente proposta; PS2 em verde; PS3 em vermelho, PS4 em roxo e Pl em laranja – C: árvore filogenética indicando o tempo de divergência em milhões de anos e os nichos de onde foram obtidos os isolados (humano, tatu e solo).



FONTE: Adaptado de THEODORO *et al.*, 2012 e MUÑOZ *et al.*, 2016.

Embora o habitat natural de *Paracoccidioides* não tenha sido unanimemente confirmado, seus isolados já foram encontrados em uma vasta gama de ambientes, tais como solo, ração para cachorro, fezes de morcegos da espécie *Artibeus lituratus* e fezes de pinguins da espécie *Pygoscelis adeliae*. Existe alta incidência de isolamento do fungo em tatus *Dasypus novemcinctus* e *Cabassous centralis*, os quais se constituem como hospedeiros sintomáticos e não apenas reservatório natural do patógeno, uma vez que apresentam lesões provenientes da

infecção (BAGAGLI *et al.,* 2003; CORREDOR *et al.*, 2005). Estudos apontam que a forma infecção de tatus é provavelmente similar à de humanos, via inalação de propágulos no ar, e que a transmissão de tatus para humanos é improvável visto que a forma do fungo encontrada nos tecidos destes animais apresenta baixa infectividade. Entretanto, a infecção de tatus é considerada um fator altamente associado com as características epidemiológicas da PCM, assim como locais de florestas úmidas e sombreadas com distúrbio antrópico (BAGAGLI *et al*., 1998; BAGAGLI *et al.,* 2003)

1.4 PROJETO GENOMA DE *Paracoccidioides*

No ano de 2007, Cardoso e colaboradores (2007) publicaram o genoma mitocondrial do isolado 18 de *P. brasiliensis* (Pb18), que é o principal representante do grupo S1. Este genoma foi caracterizado como uma molécula circular de 71.334 pares de bases, dentre os quais foram identificados 25 tRNAs (RNA transportador), subunidades maior e menor de rRNA (RNA ribossomal), uma proteína ribossomal, 11 genes codificadores de proteínas de cadeia respiratória, 3 subunidades de ATP sintases e 3 maturases intrônicas (proteínas relacionadas a *splicing*). Os dados referentes a este genoma foram publicados no GenBank sob o número de acesso AY955840 (CARDOSO *et al*., 2007).

Quatro anos mais tarde, em 2011, o DNA (ácido desoxirribonucleico) genômico de dois isolados de *Paracoccidioides brasiliensis*, Pb18 e Pb03 (principal representante do grupo PS2), juntamente com o isolado Pb01 (hoje *Paracoccidioides lutzii*), foram sequenciados pelo *Broad Institute* do *Massachusetts Institute of Technology* (MIT), e pela universidade de *Harvard*, com participação de laboratórios do Brasil, Colômbia, Venezuela e Polônia. Usando a tecnologia Sanger de sequenciamento de DNA, foram geradas sequências com cobertura total de cerca de 8-10 vezes o tamanho de cada genoma, fornecendo um total de 30 Mb de dados. A predição de genes para Pb18 identificou 8042 genes com alta confiabilidade e um adicional de 699 genes de confiabilidade duvidosa, distribuídos em 669 *contigs* (conjunto de sequências consenso), sendo 8741 destes genes caracterizados como codificadores de proteínas. As informações obtidas através desse trabalho foram depositados sob o número de acesso ABKI00000000 no GenBank. Para Pb03, foram preditos 7875 genes codificadores de proteínas em 552 *contigs*, sendo 7610 de alta confiabilidade e outros 206 cuja predição não apresentava alta confiabilidade. O acesso a esses dados foi disponibilizado pelo código ABHV00000000. Finalmente, um total de 9132 genes codificadores de proteínas em 885 *contigs* foram preditos para Pb01 e, destes, 8130 foram apresentados como sendo de alta confiabilidade e 1002 como

genes de baixa confiabilidade. Os dados estão depositados sob o número de acesso ABKH00000000 (DESJARDINS *et al.*, 2011).

Uma classificação funcional dos genes preditos foi feita utilizando *softwares* e bancos de dados como *Blast2GO*, CAZY e MEROPS, além de mapeamento de vias metabólicas pelo *software Pathway*. Portanto, o sequenciamento do genoma permitiu a identificação dos genes presentes no gênero *Paracoccidioides* e os produtos que são codificados por eles. Essas informações possibilitaram o mapeamento das principais vias metabólicas e processos biológicos que este fungo é capaz de realizar (DESJARDINS *et al.*, 2011).

Em 2014 uma atualização do genoma foi realizada para as mesmas cepas-referência sequenciadas por Desjardins e colaboradores (2011). Dessa vez, no entanto, para o genoma v2 (versão 2) o projeto de sequenciamento foi baseado na tecnologia NGS (sequenciamento de nova geração) utilizando a plataforma *Illumina HiSeq2000*, sendo produzidos 93,6 milhões de *reads* (sequências brutas geradas no sequenciamento) com quase 200 vezes de cobertura para cada genoma. Os autores apontaram, dentre as melhorias mais significativas em relação ao genoma publicado em 2011, a correção de diversos SNPs (*Single Nucleotide Polymorphisms*), inserções e deleções presentes nas versões originais, assim como o fechamento de diversos *gaps* (espaços de sequências desconhecidas). A anotação destas novas montagens levou à descrição de novos genes, previamente não encontrados nos genomas originais. Especificamente, foram descritos 840 novos genes para o isolado Pb18, 933 para o isolado Pb03 e 936 para o isolado Pb01 (*P. lutzii*). Paralelamente, diversos genes presentes nas anotações originais foram removidos (1187 em Pb18, 490 em Pb03 e 1265 em Pb01), resultando em um total de 8390 genes codificadores de proteínas identificados em Pb18, 8427 em Pb03 e 8826 em Pb01, Em resumo, apenas 23% dos genes foram preservados na segunda predição em relação à primeira. As novas montagens genômicas, bem como suas respectivas anotações também foram submetidas ao GenBank, sob os números ABKI00000000.2 (Pb18), ABHV00000000.2 (Pb03) e ABKH00000000.2 (Pb01) (MUÑOZ, *et al.*, 2014).

Outros dois genomas de *Paracoccidioides brasiliensis* foram publicados pelo grupo do *Broad Institute*, em 2016, com colaboração de institutos de pesquisa dos Estados Unidos, Colômbia e Brasil. Também nestes casos, foi utilizada tecnologia *Illumina HiSeq*, gerando cobertura de 129 vezes para o genoma do isolado Cnh (PbCnh – representante do grupo PS3) e 158 vezes para o genoma do isolado 300 (Pb300 – representante do grupo PS4). Ambos genomas têm tamanho de ~29,4 Mb e a predição de genes resultou em 8324 genes codificadores de proteínas para PbCnh e 8070 para Pb300. Dessa forma, foram disponibilizados genomas de

representantes de todas as quatro linhagens filogenéticas de *P. brasiliensis*, além do isolado Pb01 (MUÑOZ *et al*., 2016).

Cabe destacar que os dados referentes aos sequenciamento e anotação das três primeiras linhagens-referência de *Paracoccidioides* (Pb18 e Pb03) e *P. lutzii* (Pb01) eram centralizados em um banco de dados gerado e mantido pela equipe do *Broad Institute*, fornecendo ferramentas como o BLAST (*Basic Local Alignment Search Tool* – Ferramenta de Busca de Alinhamento Local Básico), mapas de sintenia e busca de genes por palavra-chave. Esse conjunto de dados centralizados em uma plataforma única era extremamente útil para análises de genes individuais, além de estudos em larga escala, como aqueles obtidos através de microarranjos de DNA e sequenciamento de RNA, uma vez que permitia uma busca rápida, eficiente e centralizada dos dados das sequências gênicas e de suas anotações funcionais.

Entretanto, nenhum banco de dados dedicado e exclusivo de *Paracocicidioides* spp. está disponível atualmente, uma vez que o banco de dados mantido pelo *Broad Institute* foi descontinuado em 2016. Dessa maneira, as sequências geradas pelo projeto genoma *Paracoccidioides* encontram-se disponibilizadas apenas em bancos de dados gerais, como GenBank, Ensembl e RefSeq, além de bancos de dados fúngicos como *MycoCosm* e *EuPathDB/FungiDB.* No entanto, esses recursos não oferecem uma anotação consolidada e manualmente curada das proteínas preditas para cada linhagem, dificultando a busca e localização de genes, além de análises de experimentos de expressão gênica e proteômica, pois, ou não possuem ferramentas que possibilitem a análise de dados em larga escala (arquivos multi-fasta), ou a anotação funcional existente nestes bancos é muito reduzida, imprecisa e/ou desatualizada.

Por exemplo, ao analisar a versão 2 do genoma de Pb18 junto à base de dados RefSeq e GenBank do NCBI (*National Center for Biological Information*), são encontrados 7395 genes descritos apenas com o termo "*hypothetical protein*" (proteína hipotética), o que representa ~88% dos 8390 genes codificadores de proteínas originalmente preditos para este microrganismo. Portanto, pouco mais de 10% das proteínas do Pb18 apresentam descrição de função. Situações semelhantes podem ser vistas com os isolados Pb03 e Pb01, que apresentam descrição funcional para apenas ~12% (990/8427) e 37% (3330/8826) de seus genes codificadores de proteínas, respectivamente.

Ao analisar outros bancos de dados, situação assemelhada é encontrada na maioria dos casos. O Ensembl (http://fungi.ensembl.org/index.html), por exemplo, possui os mesmos dados de anotação encontrados no RefSeq, assim como o *EuPathDB*/*FungiDB*, que é um banco especializado em fungos (http://fungidb.org/fungidb/). O único banco que apresenta uma

anotação mais abrangente no que se refere à identificação funcional de genes de *Paracoccidioides* é o *MycoCosm* (https://genome.jgi.doe.gov/programs/fungi/index.jsf), onde um número bem menor de genes codificadores de proteínas (~32%) são descritos como responsáveis por proteínas hipotéticas no genoma de Pb18 (versão 2). Porém, o MycoCosm não apresenta informações para os demais genomas de *Paracoccidioides* spp. descritos acima, à exceção de Pb03. No entanto, as informações para o Pb03 estão desatualizadas neste repositório, uma vez que referem-se à primeira versão do genoma, descrita por DESJARDINS *et al*., em 2011.

Além da questão relacionada à escassez de anotação funcional nos genomas das cepas-referência do gênero *Paracoccidioides* spp., bancos como GenBank, Ensembl, RefSeq, MycoCosm e EuPathDB/FungiDB compreendem diversos organismos, ou seja, não são *databases* exclusivos para um gênero, ou para um grupo de microrganismos evolutivamente próximos, o que facilitaria estudos entre microrganismos biologicamente assemelhados, sobretudo para análises comparativas que visem à identificação de variações em seus mecanismos de patogenicidade, ou a adaptação a diferentes hospedeiros e/ou sub-regiões geográficas. Além disso, um banco contendo um número mais reduzido de espécies exige menor uso de recursos de informática, garantindo dessa forma uma possibilidade de maior performance durante o desenvolvimento de diversos tipos de análises computacionais e consulta as informações do banco.

De fato, diversos grupos de fungos contam com seus *databases* próprios, inteiramente dedicados à análise de isolados de uma mesma espécie e/ou gênero, como o *database* de *Saccharomyces* (*Saccharomyces* Genome *Database* - SGD) (https://www.yeastgenome.org/) (CHERRY *et al.*, 2011), o *Candida* Genome *Database* (http://www.candidagenome.org/) (SKRZYPEK *et al.*, 2017), o AspGD (http://www.aspgd.org/), para *Aspergilllus* (CERQUEIRA *et al.*, 2014) e a HistoBase (http://histo.ucsf.edu/), da espécie *Histoplasma capsulatum* (SILLAB, 2017), entre outros. É importante lembrar que, mesmo existindo bases de dados exclusivas, dados de sequenciamento/anotação para estes fungos também estão presentes nos bancos de dados gerais como EuPathDB/FungiDB e MycoCosm. Nesse sentido, considerando a importância médica e epidemiológica da Paracoccidioidomicose e de seus agentes causadores, associada à ineficiente anotação atualmente disponível para os principais isolados de referência do grupo, seria interessante contar com um banco de dados centralizado, que oferecesse informações mais precisas acerca de anotações funcionais dos genomas de representantes do gênero *Paracoccidioides*.

1.5 ANOTAÇÃO FUNCIONAL

A anotação de um genoma inclui dois processos distintos: a anotação estrutural (predição de genes) e a anotação funcional. A anotação estrutural é conduzida de maneira a identificar e localizar genes e estruturas como íntrons (regiões não informacionais) e éxons (regiões informacionais) (YANDELL e ENCE, 2012). Normalmente, essa análise é realizada após o sequenciamento por ferramentas baseadas em similaridade, que partem do suposto que exons são mais conservados do que regiões não funcionais (regiões intergênicas ou intrônicas), como BLAST (*Basic Local Alignment Search Tool* = Ferramenta de Busca de Alinhamento Local Básico). O *software* AUGUSTUS representa outro método de predição, usando elementos estruturais como modelo de detecção, chamado de predição *ab initio*, que depende de dois tipos de informações: sensores de sinal (*motifs* de sequência curta, junções de *splicing*, pontos de ramificação e códons de iniciação e de parada) e sensores de conteúdo (detecção de exon e padrões de uso de códons exclusivos de espécies) (WANG *et al*., 2004).

Já a anotação funcional é o procedimento por meio do qual uma função biológica é atribuída à sequência de resíduos de aminoácidos, no caso das proteínas, ou à sequência de nucleotídeos, no caso dos genes, que são preditos na anotação estrutural (YANDELL e ENCE, 2012). Este procedimento é fundamental para prover significado aos dados iniciais de um genoma e permite o entendimento biológico, estrutural e funcional das informações moleculares para a comunidade científica (KOONIN e GALPERIN, 2003).

Existem vários métodos computacionais para a anotação funcional, com destaque para: *Rosetta-Stone* (baseado na análise de redes de interações entre proteínas), *Phylogenetic Profiling* (baseado em genômica filogenética comparativa), *Genome Context* (baseado em interações, reguladores e processos celulares) e Co-expressão (baseado na similaridade de perfis de expressão) (SIVASHANKARI e SHANMUGHAVEL, 2006). Há ainda métodos baseados em homologia de sequências como *InterProScan*, que integra predições de bases de dados de domínios proteicos (JONES *et al*., 2014), e *Gene Ontology*, que usa vocabulários hierárquicos para classificação funcional (LOEWENSTEIN *et al*., 2009).

Anotações devem ser ainda manualmente curadas por pesquisadores. Esta pós-análise pode ser feita *in silico*, por métodos computacionais para comparação de dados biológicos diversos (GOPAL *et al*., 2014) ou *in situ*, mediante experimentação (PFEIFFER *et al.*, 2015). Os mais importantes grupos com iniciativas de anotação realizam curadoria manual, como o grupo BioCyc, que disponibiliza uma coleção online de *databases* de genomas e vias metabólicas (CASPI *et al*., 2016), como também o grupo do *Saccharomyces* Genome *Database*

(SGD), o banco de dados do genoma do fungo *Saccharomyces cerevisiae* (CHERRY *et al*., 2012).

Dentre as plataformas mais intuitivas de anotação funcional estão o programa *Blast2GO* (CONESA e GÖTZ, 2008) e o banco de dados DAVID (*Database for Annotation, Visualization and Integrated Discovery*) (HUANG *et al*., 2009), que serão descritos em maior detalhe adiante. Estas ferramentas são muito utilizadas, pois possuem recursos integrados (e gratuitos, em sua maioria) de várias bases de dados biológicos, além de atualizações recentes. A base de artigos científicos especializada PubMed (NCBI, 2017), reúne mais de 3800 citações para *Blast2GO* e o DAVID é citado mais de 8100 vezes.

O *Blast2GO* é um *software* de para anotação funcional e análise de dados genômicos que apresenta interface gráfica interativa, assim como ferramentas para busca de informação biológica em uma série de bancos de dados especializados brevemente descritos a seguir:

- *Gene Ontology/GO* - O Consórcio de Ontologia Gênica (*Gene Ontology Consortium*) é uma iniciativa de padronização de informação funcional que foi idealizada com vistas a identificar as principais funções biológicas de eucariotos e procariotos, usando uma classificação estruturada em três vocabulários hierárquicos: Componente Celular (*Cellular Component*, ou CC), Função Molecular (*Molecular Function*, ou MF) e Processo Biológico (*Biological Process*, ou BP) (ASHBURNER *et al*., 2000; THE GENE ONTOLOGY, 2017).
- *EC-number* - *Enzyme Comission number* (Número do Comitê de Enzimas). São códigos identificadores que classificam enzimas hierarquicamente, de acordo com a reação que catalisam e os grupos químicos envolvidos (TIPTON e BOYCE, 2000; NC-IUBMB, 2017).
- *InterProScan* - pacote de *software* que permite busca por assinaturas *InterPro*, que por sua vez, é um *database* de modelos preditivos fornecidos por vários outros bancos de dados que compõem o *Consórcio InterPro* (citados abaixo) (QUEVILLON *et al*., 2005).
- ProDom - *Database* de famílias de domínios de proteínas (SERVANT *et al*., 2002) gerado por SWISS-PROT (uma base de dados de proteínas anotadas e manualmente curadas) e TrEMBL, com anotação exclusivamente computacional (BOUTET *et al.*, 2007).
- PRINTS - Base de "impressões digitais de proteínas", ou grupos de *motifs* (motivos) conservados usados para caracterizar famílias de proteínas (ATTWOOD *et al*., 2003).
- PIR/PSD - *Protein Information Resource* (Recurso de Informação de Proteínas). Ferramenta pública integrada de informática de proteínas e mantém a PSD (*Protein Sequence Database*), uma base de proteínas abrangendo todo a variedade taxonômica (WU *et al*., 2003).

- Pfam - Banco de dados amplamente utilizado para a identificação de famílias de proteínas, contendo também entradas com curadoria manual (FINN *et al*., 2014).
- SMART - *Simple Modular Architecture Research Tool* (Ferramenta de Pesquisa de Arquitetura Modular Simples), contém ferramentas para a identificação e anotação de domínios de proteínas e análise de arquiteturas de domínios proteicos (LETUNIC e BORK, 2017).
- TIGRFAMs - É um *database* com recursos para alinhamentos de sequências múltiplas com curadoria, HMMs (*Hidden Markov Models*, Modelos Ocultos de Markov) para classificação de sequências de proteínas e informações associadas para suportar anotação automatizada de proteínas (HAFT *et al*., 2003).
- PROSITE - Este banco de dados consiste em uma grande coleção de assinaturas biologicamente significativas, descritas como padrões ou perfis. Cada assinatura é ligada a uma documentação que fornece informações biológicas úteis de família de proteínas, domínio ou sitio funcional identificado pela assinatura (HULO *et al*., 2006).
- HAMAP – *High-quality Automated and Manual Annotation of microbial Proteomes* (Anotação Automatizada e Manual de Alta qualidade de Proteomas microbianos), usa modelos de anotações construídas manualmente para famílias de proteínas, usando critérios rígidos (PEDRUZZI *et al*., 2015).
- SUPERFAMILY - Fornece dados estruturais, funcionais e evolutivos para proteínas de todos os genomas completamente sequenciados e grande coleções de sequência. Domínio de proteínas e HMMs com base em SCOP (*Structural Classification of Proteins*, Classificação Estrutural de Proteínas) no nível da superfamília são usados para fornecer anotação (WILSON *et al*., 2009).
- SignalP - SignalP Prediz presença e localização de sítios de clivagem de peptídios sinal em sequências de aminoácidos tanto de procariotos como eucariotos, incorporando predição de sitios de clivagem e uma predição com base em uma combinação de várias redes neurais artificiais (PETERSEN *et al*., 2011).
- PANTHER - *Protein ANalysis THrough Evolutionary Relationships* (Análise de Proteínas Através de Relacionamentos Evolutivos), faz inferência de função de proteínas por árvores filogenéticas, com foco em representações mais detalhadas de eventos evolutivos na história de famílias de genes (MI *et al*., 2012).
- Gene3D - É um banco de anotações de domínio globular (assinatura de proteínas com estrutura 3D globular) para milhões de sequências de proteínas disponíveis (LEWIS *et al*., 2017), usando perfis HMM do *database* CATH, que fornece relações evolutivas de domínios

de proteínas em quatro níveis: Classe/Arquitetura/Topologia/superfamílias Homólogas (*Class/Architecture/Topology/Homologous superfamily*) (SILLITOE *et al*., 2015)

- Phobius - Phobius usa um modelo de HMMs combinando topologia de proteínas transmembrana e peptídios sinal para predição de anotação (KÄLL *et al*., 2007).
- CDD - *Conserved Domain Database* (*Database* de Domínios Conservados), é uma base do NCBI que realiza anotação funcional de proteínas pela localização de "pegadas" (*footprints*) de domínio conservadas, incluindo modelos com curadoria manual que usam estrutura 3D para refinar relacionamentos, estrutura e função (MARCHLER-BAUER *et al*., 2010).
- SFLD - *Database* de Ligação Estrutura-Função (Stru*cture-Function Linkage Database*) é uma base que contém ferramentas de classificação com curadoria manual e descrição de relações estrutura-função para superfamílias de enzimas (AKIVA *et al*., 2014).
- MobiDB - *Database* para anotação de proteínas intrinsecamente desordenadas (não possuem estruturas 3D ordenadas ou fixamente imóveis) (TOMPA, 2012), diversidade conformacional e interações de proteínas (PIOVESAN *et al*., 2017).

A *pipeline* (conjunto de etapas) de anotação do *Blast2GO* é composta por etapas de *Load*, *Blast*, *Mapping*, *Annotation*, *InterProScan* e *Export*:

- *Load* - Carregamento. Primeiramente são carregados os arquivos fasta, por exemplo, do genoma que se deseja anotar.
- BLAST - As sequências presentes no arquivo do genoma são analisados por BLAST para a busca de descrições (*blast hits*) obtidas de sequencias similares àquelas carregadas (*query*). Este passo é feito *online* utilizando *databases* do NCBI ou localmente (*local blast*), a partir de bases para BLAST customizadas que podem ser criadas no próprio programa e armazenadas no computador.
- *Mapping* - Mapeamento GO. Este mapeamento é o processo de obtenção direta via *database Gene Ontology* de termos GO associados aos *blast hits*. Também é possível aqui buscar códigos EC para os resultados de termos GO encontrados.
- *Annotation* – Anotação GO. Este processo seleciona os termos GO obtidos pelo *Mapping* e atribui regras de anotação, para encontrar anotações mais específicas e com níveis de confiabilidade.
- *InterProScan* – Aqui são buscadas anotações do *InterPro* (citado acima), permitindo recuperar informações de domínio e motivo de proteínas e também termos GO.
- *Export* – Exportação de dados. *Blast2GO* também dispõe de funções para exportação das anotações obtidas em arquivos fasta, annot, GFF e GTF entre outros, que são de grande

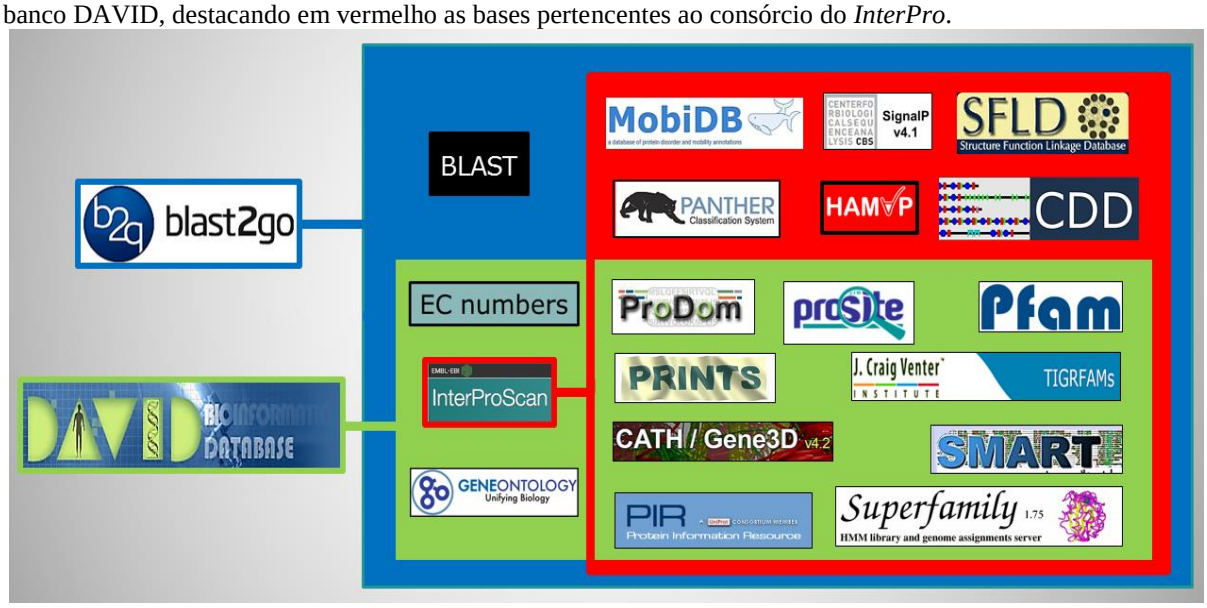
utilidade para estudos genômicos e podem ser compartilhados entre pesquisadores. Estas utilidades podem auxiliar ao diminuir o tempo total de anotação de genomas completos, o que pode durar semanas para um eucarioto complexo (CONESA e GÖTZ, 2008).

Já o *database* DAVID disponibiliza conjuntos de dados de anotação funcional para entendimento do sentido biológico de grandes listas de genes em experimentos de genômica. Diferentemente do *Blast2GO*, o DAVID armazena dados já prontos e organizados para mais de 90.000 espécies, além de ferramentas para conversão de IDs e enriquecimento de genes (HUANG *et al*., 2009). O banco de dados do DAVID apresenta não só organismos modelo, mas também organismos não-modelo, fazendo com que seu uso se estenda de forma muito mais ampla se comparado as principais ferramentas publicamente disponíveis para anotação funcional, como por exemplo, AmiGO (CARBON *et al*., 2009), Babelomics (ALONSO *et al*., 2015) e Mercator (LOHSE *et al*., 2014).

Para a anotação através do banco DAVID, basta inserir uma lista de IDs de genes de um organismo (os códigos identificadores, ou simplesmente "IDs" de genes, são sequências de letras e números únicos para representar um gene para uma espécie ou isolado), escolher as opções de ferramentas de anotação funcional e descarregar as informações obtidas em arquivos de texto simples no formato txt, que contêm as descrições de anotação e dados estatísticos para a lista de genes fornecida. Dentre a rede de informações biológicas contidas no DAVID estão dados de *Gene Ontology* e domínios de proteínas. Os termos GO nesta base estão disponíveis nos cinco níveis hierárquicos dos três vocabulários GO, além da opção *GOTERM_DIRECT*, que reduz a redundância ao anotar termos diretamente das bases de dados. Os dados de domínios proteicos do DAVID são oriundos das bases Gene3D, InterPro, Pfam, PIR, PRINTS, ProDom, PROSITE, SMART, SUPFAM e TIGRFAMs, descritas anteriormente, e que também fazem parte das bases de domínio de proteína do *Blast2GO* (CONESA e GÖTZ, 2008; HUANG *et al*., 2009).

Dessa forma, o uso de ferramentas computacionais de anotação funcional automatizadas, compreendidas pelo *software Blast2GO* e pelo *database* DAVID (Figura 5) pode fornecer uma anotação mais robusta e ampla para as linhagens-referência de *Paracoccidioides* spp. já sequenciadas. Esta integração de múltiplas fontes heterogêneas de dados é conhecida como fusão de dados biológicos *in silico* (*in silico biological data fusion*) e é muito útil para reanotação funcional de genomas (POTOK *et al*., 2003; GOPAL *et al*., 2014).

Figura 5 – Fontes de dados que compõem as análises por *Blast2GO* e o *database* DAVID. As fontes de dados estão apresentadas pelos seus logotipos ou nomes e as cores mostram onde estão contidas: estão englobadas em azul todas as fontes de dados disponíveis no *Blast2GO* e em verde aquelas presentes tanto no *Blast2Go* quanto no banco DAVID, destacando em vermelho as bases pertencentes ao consórcio do *InterPro*.



1.6 BANCOS DE DADOS BIOLÓGICOS

A partir da década de 1990, o status da informação passou a ser considerada pelos profissionais e pensadores da área de Tecnologia da Informação como "Tempo de Conhecimento Interativo", o qual se caracteriza pela possibilidade de extrema velocidade na troca de informação entre quem a fornece e quem a recebe (BARRETO, 2002). Neste sentido, bancos de dados são um dos recursos mais necessários atualmente, uma vez que a informação digital é transmitida rapidamente, de maneira banal e por diversos meios eletrônicos.

Um banco de dados é um conjunto de informações manipuláveis e de mesma natureza inseridas em um mesmo local, obedecendo a um padrão de armazenamento. Existem variados modelos de sistema de bancos de dados (MSBDs), que são a descrição formal da estrutura de um banco de dados, seus registros, relacionamentos e regras. Os MSBDs podem ser classificados em dois tipos principais, não-relacionais e relacionais (MANNINO, 2008).

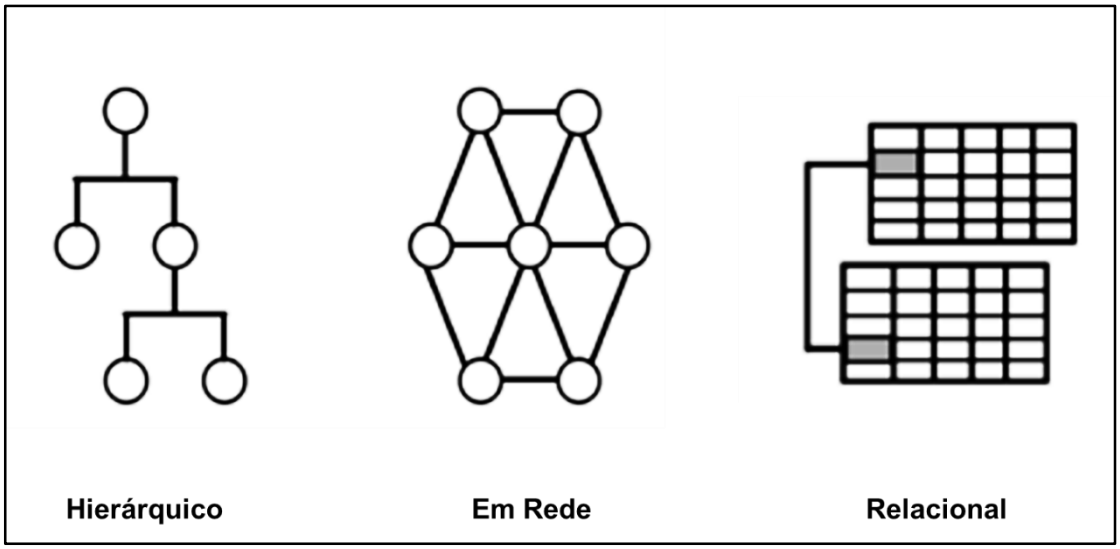
O primeiro modelo a ser utilizado foi o modelo hierárquico, um modelo não-relacional onde os dados são apresentados como conjuntos de registros e as ligações que existem entre estes registros obedecem a uma hierarquia de forma a compor uma estrutura de árvore (Figura 6). O registro da hierarquia que precede a outros é o "registro-pai", os outros são chamados de "registros-filhos". Este modelo surgiu na década de 1950 e se caracteriza por um alta

redundância de informações e baixa eficiência para estruturas de bancos complexos com relacionamentos múltiplos entre os registros (SILVA, 2001).

Na década seguinte foi desenvolvido o modelo em rede, outro tipo de modelo não-relacional e também apresenta os dados como coleções dos registros e as ligações que se formam entre eles, porém sem que exista uma definição clara de hierarquia, podendo possuir uma complexidade bem maior que o MBD anteriormente citado (Figura 6) (CARNEIRO, 2004).

Porém, o modelo mais utilizado a partir de então foi o modelo relacional, desenvolvido por volta da década de 1970 e que ficou consagrado por suas características versáteis e eficientes, visto que, neste modelo, os dados são representados por relações definidas em linhas e colunas que podem ser entendidos de forma simples como uma série de tabelas, oferecendo, ainda, operadores relacionais para extrair novas relações a partir das primeiras (Figura 6). Além disso, o modelo relacional de dados é considerado o único "real", pois sua consistência pode ser demonstrada matematicamente pela álgebra relacional, que compreende uma reunião de operações utilizadas para manipular relações (BERGERON, 2003; HEUSER, 2010).

Figura 6 – Representação dos principais Modelos de Sistemas de Bancos de Dados - MSBDs. Os MSBDs podem ser classificados em dois tipos principais, não-relacionais e relacionais. O modelo hierárquico é um modelo não-relacional ineficiente para bancos complexos. O modelo em rede, sem hierarquias, é mais eficiente que o primeiro, embora sua natureza não-relacional possa torná-lo mais complexo. O modelo relacional é versátil e eficiente, com relações em linhas e colunas, como um conjunto de tabelas, possuindo operadores para extrair novas relações a partir das primeiras. É considerado o único modelo "real", pois sua consistência pode ser demonstrada matematicamente.



Fonte: Adaptado de BERGERON, 2003

Para criar, utilizar e gerir um banco de dados são necessárias ferramentas computacionais que dão suporte a essas ações. Estes são os **S**istemas de Gerenciamento de Banco de Dados (SGBD) (MANNINO, 2008). Um Sistemas de Gerenciamento de Banco de Dados Relacional (SGBDR) permite obter informações de um banco de dados relacional de maneira conveniente e eficiente, envolvendo a definição das estruturas de armazenamento, os mecanismos de consulta e manipulação de dados, a qualidade dos resultados de consulta e se responsabiliza pela segurança dos dados, para prevenir falhas e casos de acessos não autorizados (SILBERSCHATZ *et al*., 1999). Comumente, a maior parte dos SGBD atualmente usam a linguagem SQL (*Structured Query Language*, ou Linguagem de Consulta Estruturada), que é uma linguagem computacional de consulta baseada em álgebra relacional e que possui uma padronização para bancos de dados relacional realizada pelo Instituto de Padronização Nacional Americano (ANSI). Alguns dos SGBD mais populares são o Oracle, o PostgreSQL, o MySQL e o MariaDB (FRANCO, 2016; SILBERSCHATZ *et al*., 1999).

Oracle é um SGBD com alto padrão de qualidade industrial, com capacidade extremamente elevada. As maiores características do Oracle e que chamam a atenção na área comercial são versatilidade, estabilidade, segurança, compatibilidade (SUEHRING, 2002).

O PostgreSQL é um SGBD que suporta tipos de dados definidos por usuário, além de um amplo conjunto de funções e tipos de SQL. É um projeto de código aberto (*open source*) (o código-fonte pode ser baixado em http://www.postgresql.org) com extensa documentação disponível *online*. PostgreSQL é relativamente limitado ao lidar com um grande número de operações, em relação a Oracle e outros SGBD comerciais, mas para bancos de dados de médio porte é uma ótima opção (GIBAS e JAMBECK, 2001).

O MySQL é um SGBD relacional também de código aberto. É relativamente fácil de configurar e usar. Possui um rico e complexo conjunto de recursos, e é um tanto diferente do PostgreSQL e Oracle. O MySQL define limites menores no número de operações permitidas, em comparação com PostgreSQL e Oracle sendo por isso considerado adequado para aplicações de banco de dados de pequenas e médias empresas, em vez de projetos de banco de dados pesados, mas isto não é uma regra (FRANCO, 2016).

O MariaDB é um dos servidores de banco de dados mais populares do mundo. É feito pelos desenvolvedores originais do MySQL e tem garantias para permanecer como código aberto. Usuários notáveis incluem Wikipedia, WordPress.com e Google. O MariaDB transforma dados em informações estruturadas em uma ampla gama de aplicativos, que vão desde bancos a sites (MORGENSTERN, 2016; MARIADB FOUNDATION, 2017).

Concomitante ao desenvolvimento de modelos e sistemas de gerenciamento de bancos de dados, por volta da década de 1980, necessidades de recursos computacionais para manipulação e armazenamento de dados de biologia molecular foram surgindo à medida que verdadeiras “avalanches de dados” eram produzidas a partir dos primeiros projetos de sequenciamento de genomas. Assim, foram criados os primeiros bancos de dados biológicos, por grupos de pesquisadores nos Estados Unidos e na Europa (BERGERON, 2003).

As principais bases de dados e ferramentas disponíveis para armazenamento de dados biológicos são representadas pelo GenBank (https://www.ncbi.nlm.nih.gov/genbank/), RefSeq (*Reference Sequence*) do NCBI (https://www.ncbi.nlm.nih.gov/refseq/) do NCBI, e Ensembl (http://www.ensembl.org/index.html).

A coleção GenBank é um banco com anotação de todas as sequências de DNA disponíveis publicamente. O GenBank faz parte da Colaboração Internacional de Bancos de Dados de sequências de nucleotídeos (*International Nucleotide Sequence Database Collaboration* - INSDC), que compreende também a DNA *DataBank* do Japão (DDBJ) e o Depósito de Nucleotídeos da Europa (*European Nucleotide Archive* - ENA). A última versão GenBank até o momento é a 223 (Outubro de 2017) que possui 234.997.362.623 bases e 201.663.568 sequências (BENSON *et al*., 2013; NCBI-GENBANK, 2017).

A coleção RefSeq fornece um conjunto extenso, integrado e não redundante de sequências de DNA genômico, transcritos e proteínas de acesso público. Difere bastante do GenBank, pois possui um registro único e exclusivo para cada tipo de biomolécula, e também por ser relativamente mais restritiva. Essa restritividade vem de fatores como disponibilidade e qualidade da montagem do genoma, qualidade da anotação da montagem, abundância de sequências de cDNA no INSDC, a relevância do organismo para as comunidades médicas e de pesquisa e as contribuições das comunidades de pesquisa. Como toda base de dados, o RefSeq já passou por uma série de atualizações e está atualmente na versão 85 (13 de Novembro de 2017 - *RefSeq Release 85*) e contém indexados 146710309 de registros, incluindo 1204502588476 nucleotídeos, 100,043,962 proteínas, 20,905,608 transcritos e 73996 organismos (PRUITT *et al*., 2012; O'LEARY *et al*., 2016; NCBI-REFSEQ, 2017).

GenBank e RefSeq são mantidos pelo NCBI, e as anotações dos genomas depositados são realizadas pelo NCBI *Prokaryotic Genome Annotation Pipeline* (Pipeline de Anotação de Genoma Procarioto, PGAP), para procariotos, e do NCBI *Eukaryotic Genome Annotation Pipeline* (*Pipeline* de Anotação de Genoma Eucarioto, EGAP) para eucariotos. A anotação funcional da PGAP é feita através de BLAST contra todos os produtos proteicos identificados contra uma base de dados revisada, composta por representantes de todos os conjuntos de

proteínas procarióticas provenientes da base *UniProt-SwissProt* e todas as proteínas de bacteriófagos da coleção RefSeq, submetidas a curadoria (TATUSOVA *et al.*, 2016).

Com EGAP, os nomes de proteínas são atribuídos com base no tipo de locus, homologia de proteínas e informações de ortologia, e dados do banco de dados de Gene, que podem, por sua vez, ser baseados em nomenclatura de um grupo externo, como o HUGO *Gene Nomenclature Committee* (HGNC). Os genes previstos são avaliados para ortologia, que é a característica de sequências homólogas descendentes da mesma sequência ancestral, (HOLDING *et al*., 2002) em genes de espécies de referência usando um processo de comparação baseado em alinhamentos de proteínas e informações de locais de sintenia, ou seja, co-localização física de *loci* no mesmo cromossomo de um indivíduo, ou entre espécies (MCCOUCH, 2001). Se um ortólogo pode ser determinado, um *gene symbol* (símbolo de gene) e nome do gene são transferidos das espécies de referência. Se um ortólogo não pode ser determinado, genes previstos são nomeados com base no nome da proteína *SwissProt* mais assemelhada, adicionando o sufixo "*like*" ("parecido") indicando incerteza. Genes previstos para os quais nenhum nome pode ser determinado recebem um nome de gene e proteína genérico como "*uncharacterized*" ("não-caracterizado") (THIBAUD-NISSEN, 2013).

O Ensembl é um *database* desenvolvido com intuito de oferecer anotação automática do genoma humano, além de integrar esta anotação com outros dados biológicos e disponibilizar as informações publicamente na internet. Vários outros genomas foram indexados na base atualmente, que oferece páginas específicas como EnsemblBacteria, EnsemblPlants, EnsemblProtists, EnsemblMetazoa e EnsemblFungi. A versão atual de Ensembl é a *Ensembl90* (Agosto de 2017). As anotações funcionais disponíveis no Ensembl são fornecidas através da curadoria direta, importadas da base UniProt, ou inseridas a partir de análises por *InterProScan* (YATES, 2016; KERSEY *et al*., 2017).

Embora estes bancos sejam atualizados frequentemente, organismos não-modelo podem permanecer com grande carência de anotação funcional de seus genes. Como visto para *Paracoccidioides* spp., as anotações funcionais disponibilizadas nos principais bancos de dados de informação biológica e molecular são carentes e apresentam uma quantidade alta de proteínas anotadas como hipotéticas. Nesse sentido, além de serem desenvolvidas e disponibilizadas novas anotações do genoma dos fungos do gênero *Paracoccidioides*, seria interessante a organização dessas informações em uma base de dados exclusiva para o fungo, de forma a garantir: (i) disponibilização centralizada da anotação funcional e (ii) menor uso de recursos de infraestrutura computacional, quando comparado aos bancos gerais de fungos, com menos dados nas tabelas e pesquisas direcionadas ao SGDB.

2 OBJETIVOS

2.1 OBJETIVOS GERAIS

Realizar uma reanotação funcional dos genes presentes nos genomas das espécies-referência representando os 5 grupos filogenéticos de *Paracoccidioides*, a saber, *Paracoccidioides brasiliensis* isolado 18 (Pb18, pertencente ao grupo S1), *P. brasiliensis* isolado 03 (Pb3, pertencente ao grupo PS2), *P. brasiliensis* isolado 300 (Pb300, pertencente ao grupo PS4), *P. brasiliensis* isolado Cnh (PbCnh, pertencente ao grupo PS3) e *Paracoccidioides lutzii* isolado 01 (historicamente referido como Pb01, pertencente ao antigo grupo Pl).

2.2 OBJETIVOS ESPECÍFICOS

- Reanotar os genes presentes nos genomas das espécies-referência representando os grupos filogenéticos de *Paracoccidioides*;
- Organizar os dados gerados pela reanotação em arquivos que serão utilizados para montagem de um banco de dados contendo os genomas dos fungos do gênero *Paracoccidioides;*
- Criar um banco de dados relacional para armazenamento e consulta dos dados genômicos de *Paracoccidioides* a partir da reanotação funcional.

3 MÉTODO

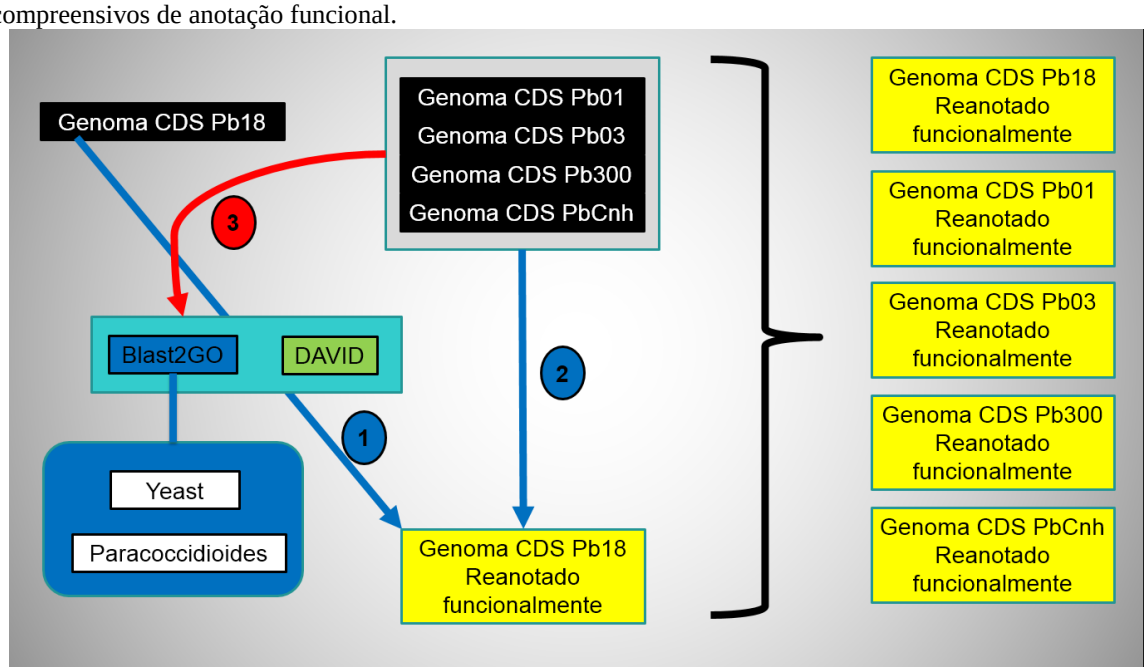
3.1 OBTENÇÃO E MANIPULAÇÃO DOS GENOMAS

Os arquivos fasta (formato de arquivo padrão de sequências de nucleotídeos, ou aminoácidos) dos genomas CDS (*Coding DNA Sequence*) dos isolados de *P. brasiliensis* Pb18 (grupo S1), Pb03 (grupo PS2), Pb300 (grupo PS4) e PbCnh (grupo PS3), bem como do isolado de *P. lutzii* Pb01 (grupo Pl) foram baixados dos diretórios FTP (*File Transfer Protocol*) do NCBI, nas suas respectivas páginas, em formato *.fna.gzip (compactado). Os genomas do isolado Pb18 e Pb01 foram obtidos dos diretórios RefSeq, enquanto os demais foram obtidos dos respectivos diretórios GenBank, já que não estão contidos nos diretórios RefSeq. Também foram obtidas tabelas no formato tabular, contendo nomes e identificadores dos seus genes (IDs), a partir das respectivas páginas no NCBI. Os arquivos foram previamente descompactados para serem usados nos pacotes de *software* utilizados nos processos de anotação funcional.

3.2 PROCESSO DE REANOTAÇÃO FUNCIONAL

O esquema de anotação funcional dos grupos filogenéticos de *Paracoccidioides* spp. foi realizado, primeiramente, submetendo o genoma do isolado Pb18 a etapas computacionais de anotação funcional utilizando o *software Blast2GO* e o *database* DAVID, seguido de curadoria manual das descrições obtidas. Em seguida, os outros quatro isolados foram analisados por BLAST contra Pb18, de forma a garantir anotações preliminares e consistentes para todos os grupos *Paracoccidioides* spp de maneira preliminar, como mostrado na Figura 7.

Figura 7 – Esquema utilizado para a reanotação funcional dos genomas publicamente disponíveis para os cinco grupos filogenéticos do gênero *Paracoccidioides*. 1: O genoma de Pb18 foi primeiramente reanotado utilizando o *software Blast2GO* e dados do arquivo Yeast, que um compilado de sequências de fungos do NCBI, como também um arquivo de sequências dos cinco genomas de *Paracoccidioides*; 2: Em seguida, os demais genomas de *Paracoccidioides* foram reanotados a partir da reanotação do isolado Pb18; 3: Os genomas dos demais isolados foram também submetidos ao *Blast2GO*. *Blast2GO* é uma plataforma de bioinformática para anotação funcional de alta qualidade e análise de dados genômicos e o DAVID disponibiliza conjuntos de dados organizados e compreensivos de anotação funcional.



O arquivo do genoma CDS de Pb18 foi carregado no *software Blast2GO* versão 4.0 (CONESA e GÖTZ, 2008) pela opção *Load* > *Load Sequences* > *Load Fasta File (.fasta)*. No programa, foi utilizado o comando *BLAST* > *Make BLAST Database* para que fossem criadas *databases* para *BLAST* local utilizando um arquivo fasta de sequências de aminoácidos de todos os grupos de *Paracoccidioides* e também da sequência multi-fasta *Yeast* (arquivo contendo todas as sequências de aminoácidos de fungos) depositada no NCBI. Os demais isolados de *Paracoccidioides* também foram blastados (submetidos ao BLAST) contra *Yeast*. O processo de *BLAST* foi realizado através do comando *BLAST* > *Run BLAST* > *Local BLAST*. As opções e parâmetros para o BLAST estão descritos no Quadro 1, de acordo com a documentação do *software*. Na etapa de BLAST, as sequências presentes no arquivo do genoma são analisadas para a busca de descrições (*blast hits*) obtidas de sequencias similares àquelas carregadas (*query*) (CONESA e GÖTZ, 2008).

Quadro 1 – **Parâmetros usados para a etapa de BLAST no *software Blast2GO*.** As descrições dos parâmetros desta etapa foram obtidos do manual do *software*. Na etapa de BLAST as sequências presentes no arquivo do genoma são analisados para a busca de descrições (*blast hits*) obtidas de sequencias similares àquelas carregadas (*query*).

Parâmetro	Descrição
Local BLAST Configuration	Configuração de BLAST Local
BLAST program = BLASTx.	Este parâmetro configura para comparação de uma sequência de nucleotídeos contra uma base de proteínas
BLAST Expectation Value (E-Value) = 1,0 x (10^{-10}).	Este ajuste é o limite de confiança estatística para os resultados de BLAST. Valores menores de E-Value são mais estringentes. O programa apresenta opções de E-Value entre 1 x (10^{1}) até 1 x (10^{-100})
Number of BLAST Hits = 20 (padrão).	Número de alinhamentos alcançados
BLAST Description Annotator = Sim.	Escolha para achar a melhor descrição para um resultado de BLAST (BLAST Hit)
Advanced Configuration –	Configuração Avançada:
BLAST parameters:	Parâmetros de BLAST
Word Size = 6.	Tamanho da palavra inicial de alinhamento. Este tamanho varia de acordo com BLAST Program. O valor 6 é indicado para BLASTx.
Low Complexity Filter = Sim (padrão).	Para filtrar regiões de pouca complexidade usando o algoritmo SEG.
Run parameters –	Parâmetros de Corrida:
Number of Threads = 2.	Número de CPUs (processadores) utilizados durante o BLAST.
Filter Options –	Opções de Filtragem:
HSP Length Cutoff = 33 (padrão).	Corte para o comprimento mínimo de HSP (High Scoring Pair – Pares de Pontuação Alta: corresponde à região correspondente ao par formado pela sequência submetida e a sequência resultante do banco de dados)
Filter by Description = Não (padrão).	Este ajuste permite remover resultados que contenham as palavras fornecidas, neste caso não foi utilizado.

Fonte: CONESA e GÖTZ, 2008.

Além dos processos de BLAST, foram realizadas demais etapas de anotação funcional que são possíveis na *pipeline* do *Blast2GO*:

Gene Onthology Mapping – Mapeamento Gene Onthology = Mapeia cada resultado de BLAST com o Gene Onthology *database*. Esta etapa não necessita de ajustes de parâmetros. O mapeamento foi feito usando os comandos Mapping > Run Mapping.

Gene Onthology Annotation – Anotação Gene Onthology = Este passo avalia os vários termos GO (termos Gene Onthology) assinalados no processo de GO Mapping, e tenta achar anotações mais específicas e com alto grau de confiabilidade. Para esta etapa os comandos usados são Annot > Run Annotation, e, os parâmetros para esta etapa foram ajustados como descrito no Quadro 2, de acordo com a documentação do *software*.

Quadro 2 – **Parâmetros usados para a etapa de *Gene Onthology Annotation* no software *Blast2GO*.** As descrições dos parâmetros desta etapa foram obtidos do manual do programa e também do seu site. Este passo avalia os vários termos GO (termos Gene Onthology) assinalados no processo de GO *Mapping*, e tenta achar anotações mais específicas e com alto grau de confiabilidade.

Parâmetro	Descrição
Annotation Configuration	Configuração de Anotação:
Annotation CutOff = 55 (padrão).	Esta regra de anotação seleciona o menor termo por ramo que fica acima desse limite.
GO Weight = 5 (padrão).	Este é o peso dado à contribuição dos termos "filhos" mapeados para a anotação de um termo "pai".
Filter GO by taxonomy = Não (padrão).	Filtra os resultados por taxonomia, removendo aqueles que não se encontram dentro do táxon definido.
E-Value-Hit-Filter = 1 x 10^{-6} (padrão).	Esse valor é como um pré-filtro; somente os termos GO obtidos de hits com E-value maior do que o dado serão usados.
Hsp-HitCoverage CutOff = 0 (padrão).	Define a cobertura mínima necessária entre um Hit e seu HSP.
Hit Filter = 500 (padrão).	Esta opção permite que considerar somente os primeiros "N" hits durante a anotação.
Only Hits with GO = Sim (padrão).	Esta opção, juntamente com a opção "Hit Filter" (anterior), permite anotar somente hits que tenham um termo GO candidato.
Evidence Code Weights	Peso de Códigos de Evidência:
Computational Analysis Evidence Codes	Códigos de Evidência de Análise Computacional (GO inclui código para indicar como a anotação é suportada):
ISS = 0.8 (padrão).	Inferido por Sequência ou Similaridade estrutural.
ISO = 0.8 (padrão).	Inferido por Ontologia da Sequência.
ISA = 0.8 (padrão).	Inferido por Alinhamento da Sequência.
ISM = 0.8 (padrão).	Inferido por Modelo da Sequência.
IGC = 0.7 (padrão).	Inferido por Contexto Genômico.
IBA = 0.8 (padrão).	Inferido por aspecto Biológico do Ancestral.
IBD = 0.8 (padrão).	Inferido por aspecto Biológico do Descendente.
IKR = 0.8 (padrão).	Inferido por Resíduos Chave
ISS = 0.8 (padrão).	Inferido por Divergência Rápida.
ISS = 0.8 (padrão).	Inferido por Análise Computacional Revisada.
Experimental Evidence Codes	Códigos de Evidência Experimental (Códigos de evidência que indicam grau variável de curadoria por experimentos).
IDA = 1 (padrão).	Inferido por Ensaio Direto.
IPI = 1 (padrão).	Inferido por Interação Física.
IMP = 1 (padrão).	Inferido por Fenótipo Mutante.
IGI = 1 (padrão).	Inferido por Interação Genética.
IEP = 1 (padrão).	Inferido por Padrão de Expressão.
EXP = 1 (padrão).	Inferido por Experimento.
Author Statement Codes	Códigos de Declaração de Autor (indicam que a anotação foi feita com base em uma declaração feita pelo autor na referência citada)
TAS = 0,9 (padrão).	Declaração de Autor Rastreável.
NAS = 0,8 (padrão).	Declaração de autor Não Rastreável.
Curatorial Statement Codes	Códigos de Declaração Curatorial (indica anotação feita com base em julgamento curatorial que não se encaixa em uma das outras classificações de código de evidência):
IC = 0,9 (padrão).	Inferido pelo curador
ND = 0,5 (padrão).	Nenhum código de evidência de dados biológicos disponível.
Automatically-Assigned Evidence Code	Código de Evidência atribuído Automaticamente (usa códigos de evidência atribuídos por métodos automatizados, sem julgamento curatorial):
IEA = 0,7 (padrão).	Inferido de Anotação Eletrônica.
Obsolete Evidence Codes	Códigos De Evidência Obsoletos (anotações feitas antes de os curadores começassem a rastrear os tipos de evidências):
NR = 0 (padrão).	Não Registrado.

Fonte: CONESA e GÖTZ, 2008

A próxima etapa realizada através do *Blast2GO* foi o mapeamento de *Enzyme Codes* (EC numbers – Códigos de Enzima, ou códigos EC), seguindo os comandos *Analysis > Enzyme Code and KEGG > Run GO-Enzyme Code Mapping*. Este passo permite mapear os códigos EC existentes a partir dos termos GO mapeados.

A última anotação feita pelo *Blast2GO* foi via *InterProScan*, que identifica domínios e famílias de proteínas, e que permite também anotar termos GO e mesclar os termos GO mapeados nas etapas citadas acima. Este processo é realizado com os *comandos InterPro > Run InterProScan* (*online*), com os parâmetros apresentados no Quadro 3.

Quadro 3 – Parâmetros usados para a etapa de *InterProScan* no *software Blast2GO*. As descrições dos parâmetros desta etapa foram obtidos do manual do programa e também do seu site. A etapa InterProScan identifica domínios e famílias de proteínas, e permite também anotar termos GO e mesclar aos termos GO mapeados nas etapas de mapeamento e anotação GO.

Parâmetro	Descrição
InterProScan Configuration	Configuração InterProScan:
Choose applications to run	Escolha das aplicações para a operação:
BLASTProDom = Sim (padrão).	Escaneia a base ProDom de famílias de domínios de proteínas.
FPrintScan = Sim (padrão).	Escaneia "impressões digitais" do *database* PRINTS
HMMPIR = Sim (padrão).	Escaneia Hidden Markov Models (HMMs) do *database* de sequencias proteicas Pir (PIR-PSD).
HMMPfam = Sim (padrão).	Escaneia HMMs na Pfam – *Database* de Famílias de proteínas.
HMMSmart = Sim (padrão).	Escaneia HMMs na Smart – *Database* de Domínios Famílias de Domínios de proteínas.
HMMTigr = Sim (padrão).	Escaneia HMMs na TIGRFAMs – *Database* de famílias de proteínas.
ProfileScan = Sim (padrão).	Escaneia perfis PROSITE.
HAMAP = Sim (padrão).	Escaneia perfis HAMAP.
PatternScan = Sim (padrão).	Escaneia PatternScan, uma nova versão de busca de padrões PROSITE.
SuperFamily = Sim (padrão).	Escaneia SUPERFAMILY, uma biblioteca de perfis de HMMs de todas as proteínas de estrutura conhecida.
SignalPHMM = Sim (padrão).	Escaneia SignalP, que prediz a presença e a localização de sítios de clivagem de peptídios sinal.
TMHMM = Sim (padrão).	Escaneia a predição de hélices transmembrana em proteínas.
HMMPanther = Sim (padrão).	Escaneia HMMs PANTHER, que define famílias de proteínas e subfamílias modeladas em divergência de funções.
Gene3D = Sim (padrão).	Escaneia a *database* Gene3D, que é uma coleção de domínios CATH
Phobius = Sim (padrão).	Escaneia pelo programa Phobius de predição de topologia transmembrana e peptídios sinal.
Coils = Sim (padrão).	Escaneia fazendo predição de regiões de hélice super-enrolada (coiled-coil).
CDD = Sim (padrão).	Escaneia a *Database* de Domínios Conservados do InterPro/NCBI.
SFLD = Sim (padrão).	Escaneia a *Database* de Ligação Estrutura-Função.
MobiDBLite = Sim (padrão).	Escaneia MobiDB, uma *database* que centraliza informações de anotação de desordem intrinca e mobilidade de proteínas.

Fonte: CONESA e GÖTZ, 2008.

Em seguida, foi realizada a combinação de termos GO obtidos pelo *InterProScan* aos termos GO obtidos das etapas de mapeamento de *Gene Onthology* pelo comando *InterPro > Merge InterProScan GOs to Annotation.*

Os resultados das análises realizadas com auxílio do *software Blast2GO* foram exportados como planilhas contendo todos os dados provenientes de cada etapa da *pipeline* do programa, pelo comando *File > Export > Export Table*. Esta planilha foi utilizada, posteriormente, para comparar com os resultados obtidos pelo *database* DAVID.

Portanto, os dados depositados no *database* DAVID (*Database for Annotation, Visualization and Integrated Discovery)* versão 6,8, também foram utilizados para a reanotação dos genomas (HUANG *et al*., 2009). Para Pb18 e Pb01, as anotações funcionais foram obtidas usando como entrada para busca uma lista contendo os códigos identificadores *Locus_tag* destes organismos, e escolhido como tipo de identificador, o *Locus_tag*. Cabe destacar que Pb18 e Pb01 são os únicos isolados representantes do gênero *Paracoccidioides* que se encontram disponíveis no DAVID até o momento. Então, aplicou-se a ferramenta *Functional Annotation tool*, sendo escolhidas para busca, visualização e download as categorias descritas no Quadro 4.

Quadro 4 – Categorias Gene Onthology e domínios de proteína utilizados na anotação funcional de *Paracoccidioides* utilizando a plataforma DAVID. Dentre a rede de informações biológicas contidas na DAVID, estão dados de *Gene Ontology* e domínios de proteínas. A opção *GOTERM_DIRECT* reduz a redundância dos termos GO dos cinco níveis hierárquicos. Os dados de domínios proteicos do DAVID são oriundos das bases Gene3D, InterPro, Pfam, PIR, PRINTS, ProDom, PROSITE, SMART, SUPFAM e TIGRFAMs, descritas acima e que também fazem parte das bases de domínio de proteína do *Blast2GO*.

Categorias	Descrição
GENE ONTHOLOGY:	Categorias Gene Onthology
GOTERM_BP_DIRECT (Biological process)	Processo Biológico
GOTERM_CC_DIRECT (Cellular component)	Componente Celular
GOTERM_MF_DIRECT (Molecular Function)	Função Molecular
PROTEIN DOMAINS:	Domínios de Proteína
INTERPRO	InterPro
PFAM	Pfam
PRINTS	Prints
PROSITE	Prosite
SMART	Smart
SUPFAM	Supfam
TIGRFAMS	Tigrfams

Fonte: HUANG *et al*., 2009

Os arquivos de anotação via DAVID foram baixados como formato texto separados por tabulação, e salvos como planilha no formato tabular, sendo usados para gerar uma nova planilha contendo também os dados obtidos do processo de anotação funcional pelo *software Blast2GO*. Após a confecção desta planilha final de Pb18, foi possível anotar os genomas dos demais representantes do gênero *Paracoccidioides*, Pb03 (grupo PS2), Pb300 (grupo PS4), PbCnh (grupo PS3) e isolado Pb01 de *P. lutzii* (grupo Pl), que foram submetidos a BLAST contra o Pb18 pelo *Blast2GO*.

3.3 CURADORIA MANUAL

Os processos computacionais automatizados de anotação funcional são rápidos e práticos para grandes volumes de informação, como genomas completos. Entretanto essas anotações podem apresentar inconsistências, além de serem muitas vezes incompletas. Portanto, faz-se necessário inspecionar, individualmente, as predições automáticas geradas.

Portanto, de maneira a atribuir aos genes antes classificados como hipotéticos uma descrição mais pertinente e concisa, realizou-se, manualmente, a comparação das informações existentes para cada gene oriundas das diversas bases de dados que constituíram as fontes de anotação utilizadas pelo *Blast2GO* e pela base DAVID, sendo definido como "consenso de anotação" aquele termo que retornava com maior frequência e era mais informativo, representando a descrição de anotação funcional final para o gene. Esta abordagem por métodos computacionais para comparação de dados biológicos diversos é conhecida como fusão de dados biológicos *in silico* (GOPAL *et al*., 2014).

Além disso, foi necessário rever e alterar manualmente as descrições de alguns genes, visto que estes apresentavam informação insuficiente e/ou termos que poderiam dificultar ou confundir o entendimento da função predita.

3.4 ESTRUTURAÇÃO DOS DADOS GERADOS

Para cada um dos genomas de *Paracoccidioides* spp. funcionalmente anotados, foram geradas planilhas nos formatos *.xlsx (planilha EXCEL) e *txt (texto separado por tabulação). Estas planilhas contém a anotação funcional obtida em cada uma das fontes de dados, separadas em colunas. Também foi construída uma coluna que recebe a descrição "consenso de anotação", ou seja, a descrição de anotação funcional final para cada gene, obtida após o processo de anotação e curadoria manual. Foram construídos arquivos fasta para os genomas dos grupos de

Paracoccidioides baseados na reanotação funcional realizada, contendo um cabeçalho (*string*) que traz dados referentes a ID *locus_tag*, ID de produto proteico, descrição consenso, e a descrição de domínio proteico. Cada linha da tabela contém ainda a sequência de aminoácidos ou de nucleotídeos para cada gene, que foram inseridos com o auxílio da ferramenta "*Tabular-to-Fasta*" de conversão de formato tabular para fasta. A criação do arquivo fasta é muito útil, pois o fasta é um formato simples que se tornou padrão para sequências e é um dos mais utilizados nas análises de bioinformática (LEON e MARKEL, 2003), assim como o formato GFF3 (*General Feature Format version 3*, Formato Geral de Característica versão 3), um formato padronizado e também muito utilizado que armazena características das informações genômicas como nome do cromossomo, nome do programa que gerou a informação, ou (banco de dados ou nome do projeto), nome do tipo de característica como gene, cDNA, mRNA, exon, entre outros (GMOD, 2017). Foram criados arquivos GFF3 a partir dos arquivos GFF3 disponíveis nas bases RefSeq no caso de Pb18 e Pb01, e GenBank no caso de Pb03, Pb300 e PbCnh. Com auxílio do pacote de planilhas Excel, a descrição da anotação funcional proveniente de RefSeq e GenBank foi substituída pelas descrições obtidas na reanotação.

3.5 CONSTRUÇÃO DO BANCO DE DADOS RELACIONAL

O banco de dados dos genomas de *Paracoccidioides* foi denominado **P**ara**DB (*P*aracoccidioides *Database*)** e foi dividido em três ambientes. O ambiente do usuário consiste em uma interface de cliente e uma interface de administração, que é usada para controlar o ParaDB. Tanto as ferramentas de administração e do ambiente de usuário são implementadas em linguagem PHP (*Hypertext Processor*, Processador de hipertexto), através do sistema de gerenciamento de conteúdo open *source* mais utilizado no mundo, o *Wordpress* (BUILTWITH PTY LTD, 2017), com o acréscimo de uma interface de linha de comando (*Wordpress Command List Interface*, WPCLI), que permite uma completa configuração da aplicação pelo administrador. A interface do usuário é estruturada usando PHP e *javascript*, outra linguagem de programação muito utilizada para internet (FLANAGAN, 2011), baseado na *Foundation*, uma *Front-end Framework* (Estrutura Frontal) que oferece opções eficientes de componentes como modelos, botões e elementos de navegação (ZURB, 2017). As opções de interatividade da interface de usuário e de administração são implementadas através de *plugins* (componentes que adicionam funções extras) e *widgets* (componentes de interface gráfica que fornecem funcionalidades específicas) do *Wordpress*, que minimizam a necessidade de programação e desenvolvimento para a implementação do banco de dados. As tabelas exibidas

são criadas através de uma implementação baseada no *plugin wpDataTables*, que permite a construção de tabelas pesquisáveis e de fácil manipulação pelo administrador e usuário. Funcionalidades como a impressão da tabela e conversão para diferentes formatos (como PDF, Excel e tabular) também foram implementadas na interface de usuário (WORDPRESS.ORG, 2017).

Todo o ambiente do ParaDB encontra-se implementado em um container *Docker* (CHAMBERLAIN, 2014), uma tecnologia de *software* que fornece virtualização em nível de sistema operacional, permitindo a abstração do servidor onde se encontra instalado. A imagem utilizada nesse container foi implementada utilizando-se o sistema operacional *Linux*, distribuição *Ubuntu* (CANONICAL LTD, 2017), com todas as dependências necessárias para seu pleno funcionamento. O servidor (*webserver) Nginx* (NGINX.ORG, 2017) foi escolhido como a aplicação responsável por prover as páginas estáticas e dinâmicas do banco de dados e o SGDB MariaDB (MARIADB FOUNDATION, 2017) foi escolhido para prover as bases de dados relacionais. Ambos os serviços contam com a ferramenta de administração *EasyEngine* (RTCAMP SOLUTIONS PRIVATE LIMITED, 2017) que permite a implementação de forma rápida e prática de diferentes ambientes de desenvolvimento PHP + MySQL, com a possibilidade de instalação automatizada do *Wordpress*. A utilização do container *Docker* também possibilita o *backup* (cópia de segurança) e migração automática de todo o ambiente do banco de dados.

A criação das tabelas do banco de dados relacional teve como base o projeto BioSQL (BioSQL, 2017). O campo *Gene Symbol/Locus_tag* (símbolo de gene/locus_tag) foi definido como chave primária das tabelas, ou seja, o campo que contém um identificador único e exclusivo, e não recebe valores nulos ou repetidos (MANNINO, 2008) foi o campo responsável por permitir o relacionamento das mesmas. Decidiu-se manter a estrutura de tabelas do banco de dados em conjunto da base de dados original do *Wordpress*, permitindo futuramente uma correlação entre todas as tabelas e uma expansão do *Wordpress* em uma ferramenta plenamente adaptada para a bioinformática, assim como foi realizado com o CMS *Drupal*, com o desenvolvimento do *Trupal* (DRIES BUYTAERT, 2017). As tabelas foram populadas através de arquivos .csv (texto separado por virgula) com tabulação.

4 RESULTADOS

4.1 IDENTIFICAÇÃO DAS FUNÇÕES DOS GENES

4.1.1 *Paracoccidioides brasiliensis* isolado Pb18

O primeiro passo para a anotação funcional dos genomas de *Paracoccidiodes* foi realizado submetendo o genoma CDS (genoma apenas de DNA codificador de proteínas) do isolado Pb18 as etapas de anotação por análises computacionais automatizadas, realizadas pelo *software Blast2GO,* e também utilizando dados disponíveis no *database DAVID*. Posteriormente, os dados foram submetidos a curadoria manual, como descrito no Método. Ao todo, esta curadoria compreendeu 170 horas de trabalho e foi executada por 5 pessoas, que atuaram de maneira seriada neste processo, o que incluiu a reanotação dos 5 isolados de *Paracoccidioides*, produzindo sucessivas revisões na anotação final.

O Gráfico 1 apresenta os resultados obtidos após a anotação automática e curadoria manual. Cabe destacar que as fontes de anotação representam a origem do "consenso final", estabelecido após a curadoria manual, ou seja, de qual banco de dados o termo definido como "termo final de descrição" foi obtido. Então, é possível verificar que descrição dos 995 genes (12% do genoma CDS) anotados pelo RefSeq foram mantidos sem alteração. No entanto, 3241 genes (39% do genoma CDS) puderam ser anotados com a descrição proveniente do processo de anotação via *Blast2GO*, enquanto a anotação via DAVID permitiu caracterizar 1860 genes (22% do genoma CDS).

Genes cuja descrição de anotação apresentavam muitas divergências entre as fontes de anotação, ou traziam descrições não informativas foram mantidas como proteínas hipotéticas, restando assim, 2294 genes anotados com essa descrição (27%).

Portanto, alcançou-se anotação funcional para um total de 6096 genes codificadores de proteínas em Pb18, o que equivale a 72,7% do genoma CDS deste isolado, restando apenas 2294 proteínas (27,3% do genoma CDS) sem descrição de função. Portanto, anteriormente havia 11,9% de proteínas anotadas para Pb18 via RefSeq e, após a reanotação funcional via DAVID e *Blast2GO*, contamos com 72,7% de proteínas anotadas (Gráfico 2).

Gráfico 1 – Fontes de anotação funcional de *Paracoccidioides brasiliensis* isolado Pb18. O genoma CDS do fungo foi reanotado através de BLAST utilizando o *software Blast2GO* contra a base *Yeast* do NCBI e genomas dos demais representantes de *Paracoccidioides*. Para a caracterização final de função, também foram usados dados de anotação funcional vindos do *database* DAVID. Os valores no gráfico indicam a porcentagem de genes anotados por cada fonte utilizada.

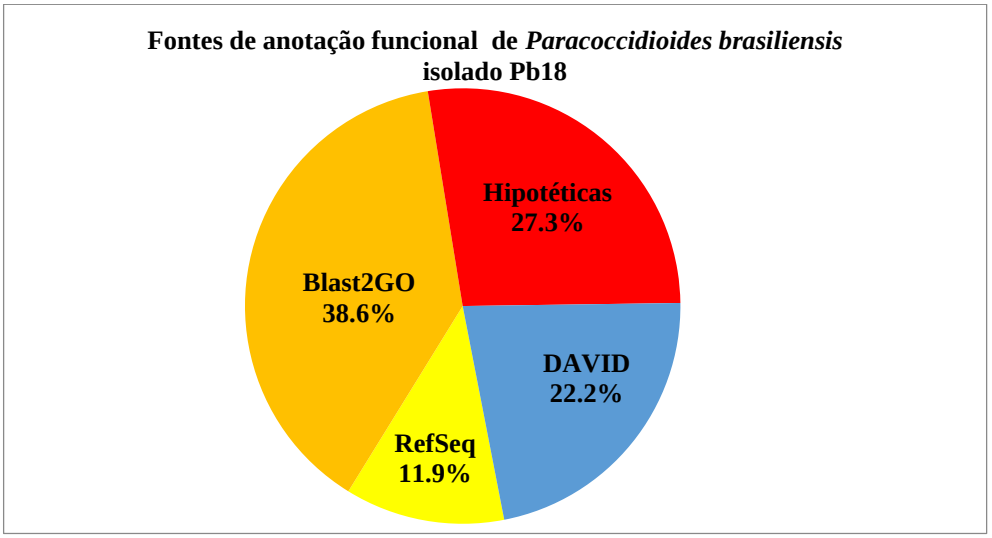
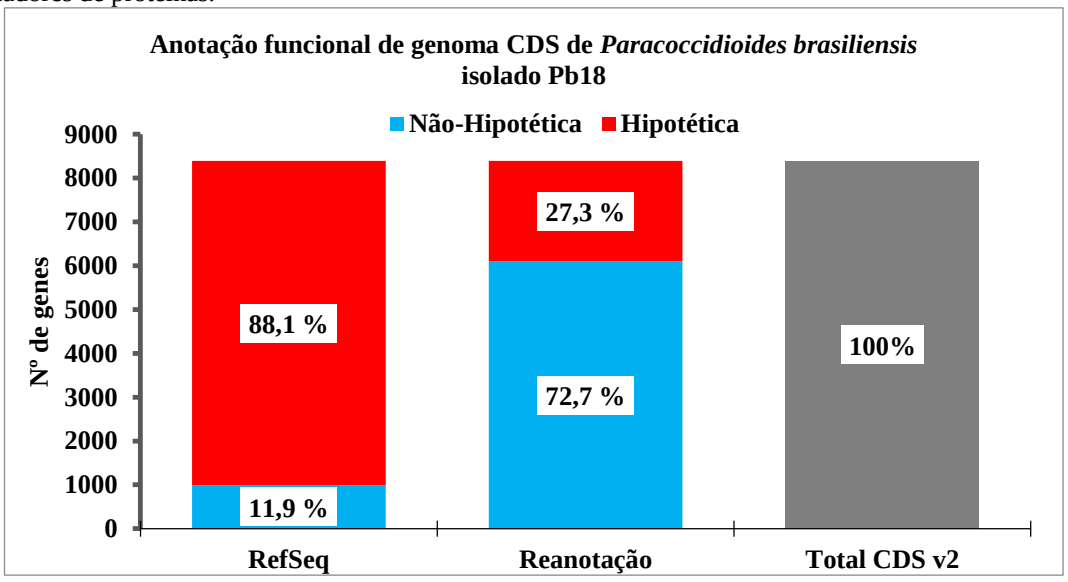


Gráfico 2 – Anotação funcional do genoma CDS de *Paracoccidioides brasiliensis* isolado Pb18 O genoma CDS do fungo foi reanotado através de BLAST pelo *software Blast2GO* contra a base *Yeast* do NCBI e contra os genomas dos demais representantes de *Paracoccidioides*. *Para a caracterização final de função*, também foram usados dados de anotação funcional vindos do *database* DAVID. Os valores no gráfico indicam a porcentagem de proteínas hipotéticas e não hipotéticas, em comparação ao RefSeq. Total CDS = número total de genes codificadores de proteínas.



4.1.2 *Paracoccidioides lutzii* isolado Pb01

Utilizando os dados de genes de *P. brasiliensis* 18 anotados funcionalmente de maneira automática e curados manualmente, foi realizada a anotação dos demais isolados, representantes dos diferentes grupos de *Paracoccidioides spp*.

Para o isolado Pb01, representante da espécie *Paracoccidioides lutzii* (Gráfico 3) foram mantidas anotações de 3287 proteínas (37% do genoma CDS) das 3330 já anotadas pelo RefSeq. No entanto, foram modificadas as anotações de 43 genes, pois a descrição destes era de "domínios não caracterizados" (DUF = *Domain of Unknown Function*), substituídas pela descrição consenso.

Ao todo, 2890 genes (32,7% do genoma CDS) receberam nova anotação funcional através do uso do *Blast2GO* e 213 genes foram reanotados através do reconhecimento de domínios de proteína através do DAVID (2,41% do genoma CDS). Ao término do processo, restaram 2436 genes (28% do genoma CDS) descritos como proteínas hipotéticas.

Dessa forma, têm-se agora anotação funcional para 6390 genes codificadores de proteínas para Pb01, produzindo uma reanotação que reconhece funções e/ou domínios funcionais para 72,4% do genoma CDS, deixando apenas 27,6% (2436) genes sem descrição de função. A anotação de Pb01 via *RefSeq* disponibilizava anotação para apenas 37,7% do genoma CDS (Gráfico 4).

Gráfico 3 – Fontes de anotação funcional para *Paracoccidioides lutzii* isolado Pb01. O genoma CDS do fungo foi reanotado através de BLAST pelo *software Blast2GO* contra a base *Yeast* do NCBI e contra o genoma reanotado de Pb18, descrito acima. Também foram usados dados de anotação funcional vindos do *database* DAVID. Os valores no gráfico indicam a porcentagem de genes anotados por cada anotador. Total CDS = número total de genes codificadores de proteínas.

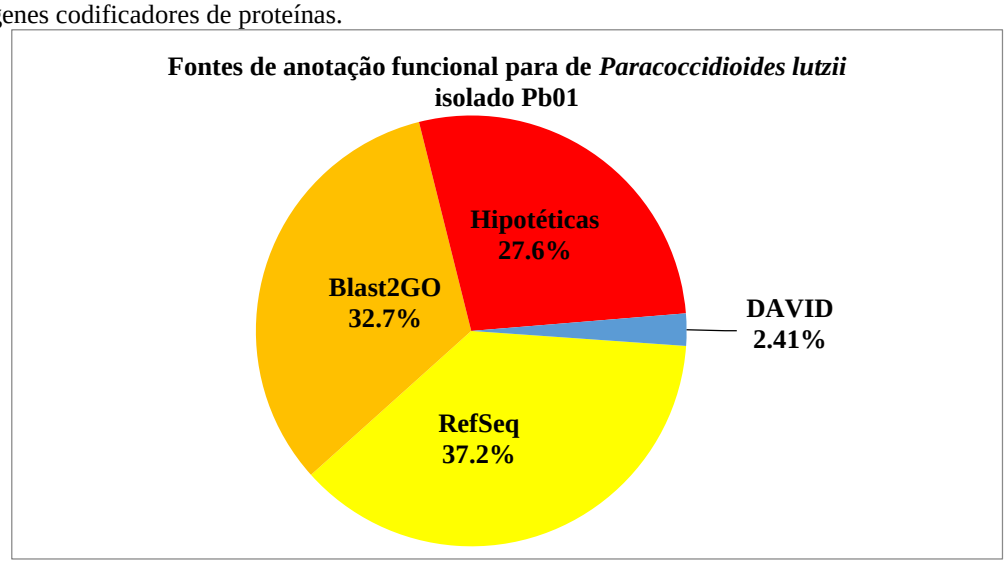
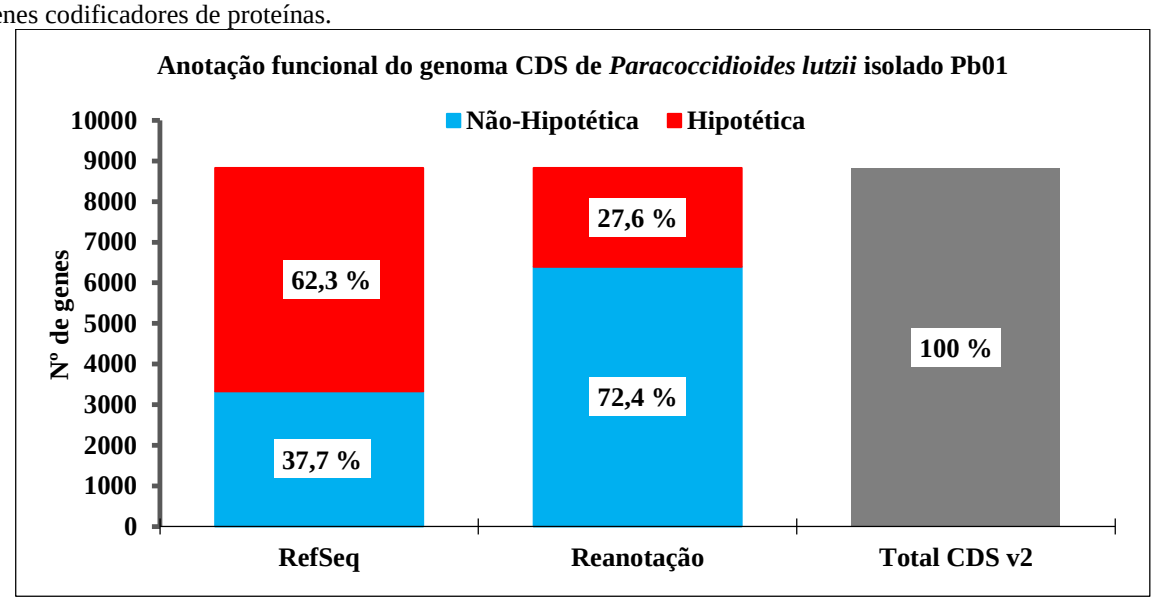


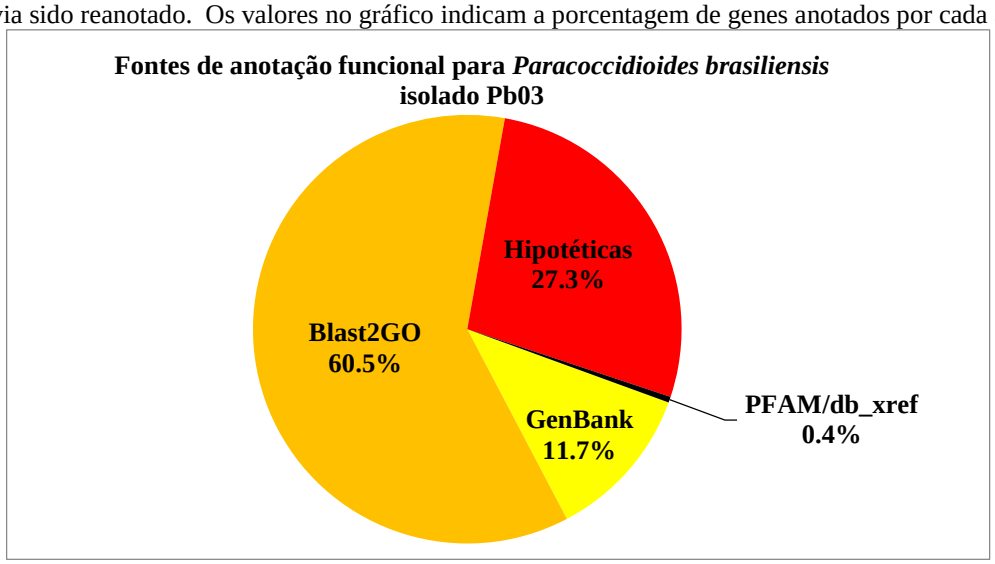
Gráfico 4 – **Anotação funcional do genoma CDS de *Paracoccidioides lutzii* isolado Pb01.** O genoma CDS do fungo foi reanotado através de BLAST no *software Blast2GO* contra a base *Yeast* do NCBI e o genoma de Pb18, e também foram usados dados de anotação funcional vindos do *database* DAVID. Os valores no gráfico indicam a porcentagem de proteínas hipotéticas e não hipotéticas em comparação ao RefSeq. Total CDS = número total de genes codificadores de proteínas.



4.1.3 *Paracoccidioides brasiliensis* isolado Pb03

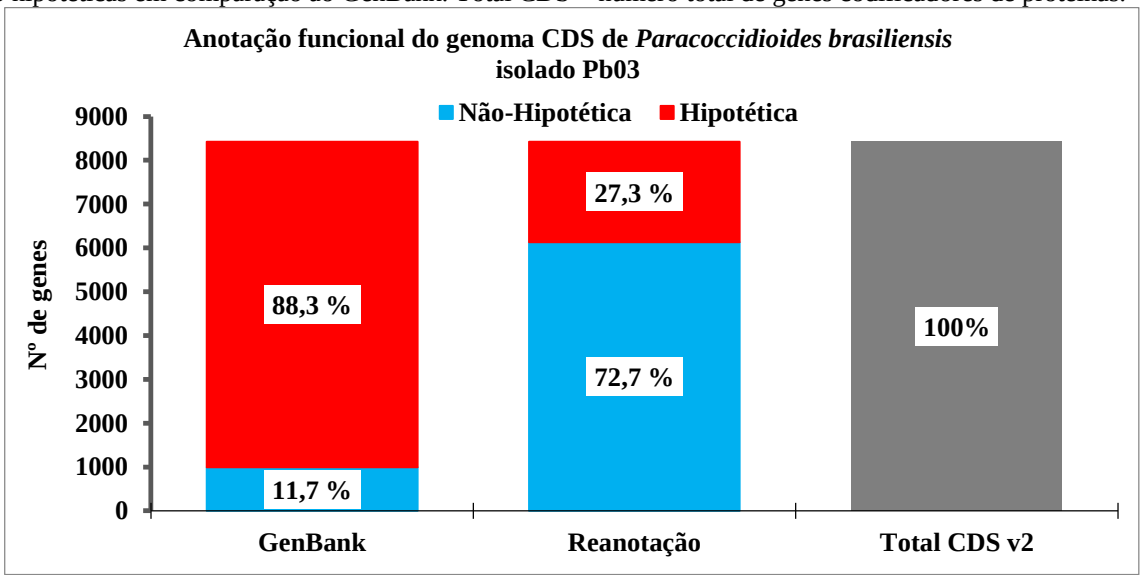
Quanto ao isolado 03 de *Paracoccidioides brasiliensis*, foram mantidas as descrições de anotação de 990 genes (11,7% do genoma CDS) disponível no GenBank. No entanto, as anotações pelo *software Blast2GO* renderam anotação para 5099 genes (60,5% do genoma CDS) e por meio de curadoria manual foi possível aprimorar a anotação (com domínios proteicos PFAM) de 37 genes (0,4% do genoma CDS) através de consulta de referência cruzada (*db_xref*) contra informações contidas na base de proteínas do NCBI. Apenas 2301 genes (27,3% do genoma CDS) permaneceram com a descrição de proteína hipotética. Lembrando que a partir daqui não é mostrado nos gráficos o *database* DAVID, já que neste existem até o momento apenas dados referentes aos isolados Pb01 e Pb18, cujos genomas são contidos no RefSeq do NCBI e representam os isolados de referência para as espécies *P. brasiliensis* e *P.lutzii*, respectivamente (Gráfico 5).

Gráfico 5 – Fontes de anotação funcional para *Paracoccidioides brasiliensis* isolado Pb03. O genoma CDS do fungo foi reanotado através de BLAST no *software Blast2GO* contra a base *Yeast* do NCBI e o genoma de Pb18, que já havia sido reanotado. Os valores no gráfico indicam a porcentagem de genes anotados por cada anotador.



O isolado Pb03 recebeu, ao final da reanotação funcional automática e manual, um total de 6126 genes (72,7% do genoma CDS) com descrição de função e apenas 2301 genes (27,3%) continuaram apresentando a descrição de proteína hipotética. Dados anteriores de anotação, via GenBank, apresentavam descrição de função para apenas 996 genes (11,7%) (Gráfico 6).

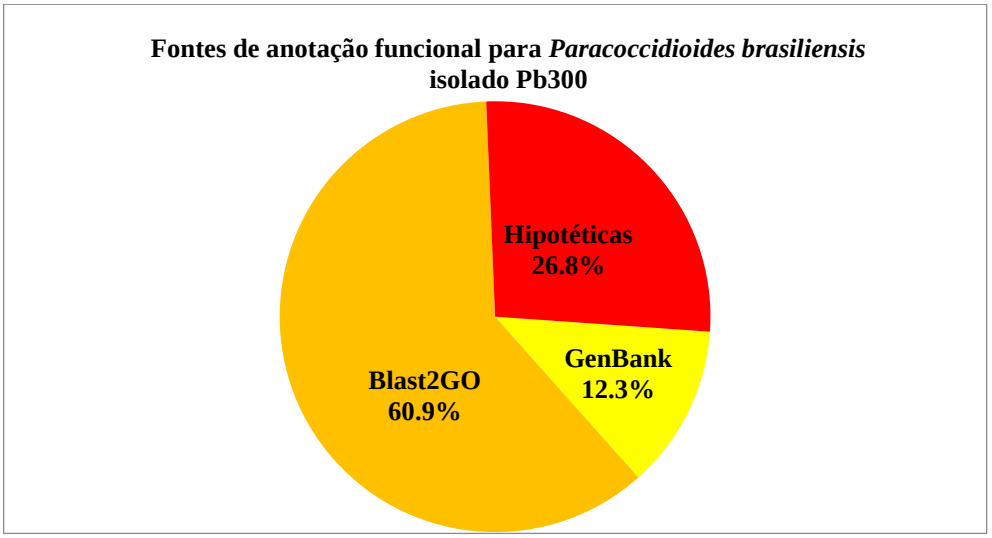
Gráfico 6 – Anotação funcional do genoma CDS de *Paracoccidioides brasiliensis* isolado Pb03. O genoma CDS do fungo foi reanotado através de BLAST no *software Blast2GO* contra a base *Yeast* do NCBI e o genoma de Pb18 que foi previamente reanotado. Os valores no gráfico indicam a porcentagem de proteínas hipotéticas e não hipotéticas em comparação ao GenBank. Total CDS = número total de genes codificadores de proteínas.



4.1.4 *Paracoccidioides brasiliensis* isolado Pb300

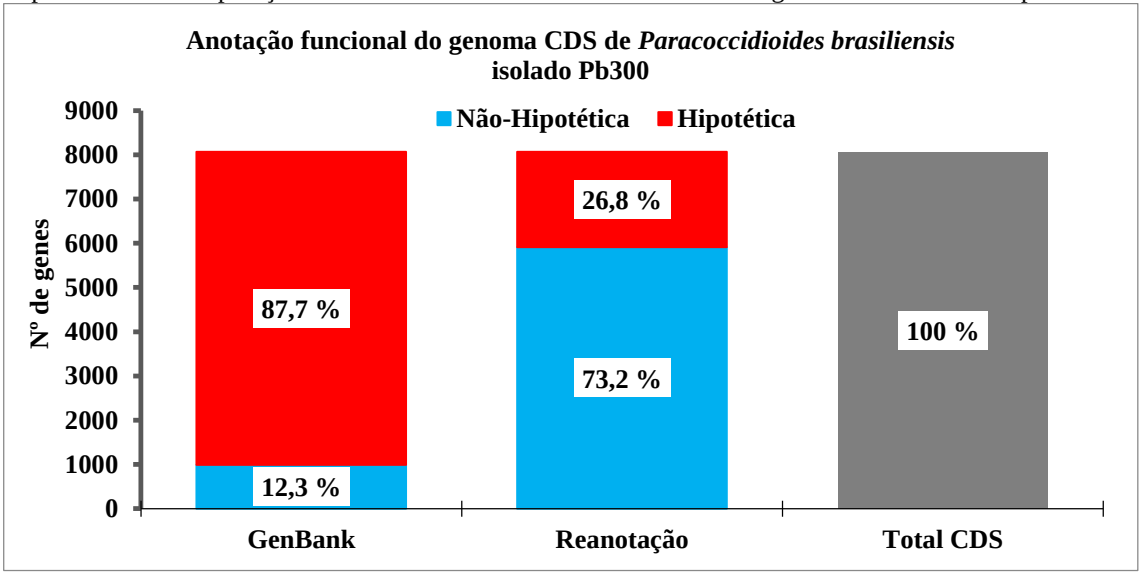
O genoma do isolado 300 *Paracoccidioides brasiliensis* foi sequenciado recentemente (MUÑOZ *et al*., 2016) e, por isso, apresenta as menores porcentagens de genes com função descrita, visto que as informações do seu sequenciamento encontram-se apenas no *database* do GenBank. Na anotação funcional realizada no presente trabalho, manteve-se a anotação de 993 genes (12,3% do genoma CDS) segundo GenBank. No entanto, a anotação via *Blast2GO* possibilitou a anotação de 4917 genes (60,9% genoma CDS) deste isolado. Somente 2160 (26,8% do genoma CDS) possuem a descrição de proteínas hipotéticas após a anotação funcional (Gráfico 7).

Gráfico 7 – Fontes de anotação funcional para *Paracoccidioides brasiliensis* isolado Pb300. O genoma CDS do fungo foi reanotado através de BLAST no *software Blast2GO* contra a base *Yeast* do NCBI e o genoma de Pb18 que foi previamente reanotado. Os valores no gráfico indicam a porcentagem de genes anotados por cada anotador.



Portanto, para Pb300, foi possível obter anotação funcional para 5911 genes (73,2% do genoma CDS), com apenas 2160 genes (26,8% do genoma CDS) anotados como proteínas hipotéticas. Dados anteriores de anotação via GenBank continham informações funcionais para apenas ~993 genes (apenas 12,3% do genoma CDS), como pode ser verificado no Gráfico 8.

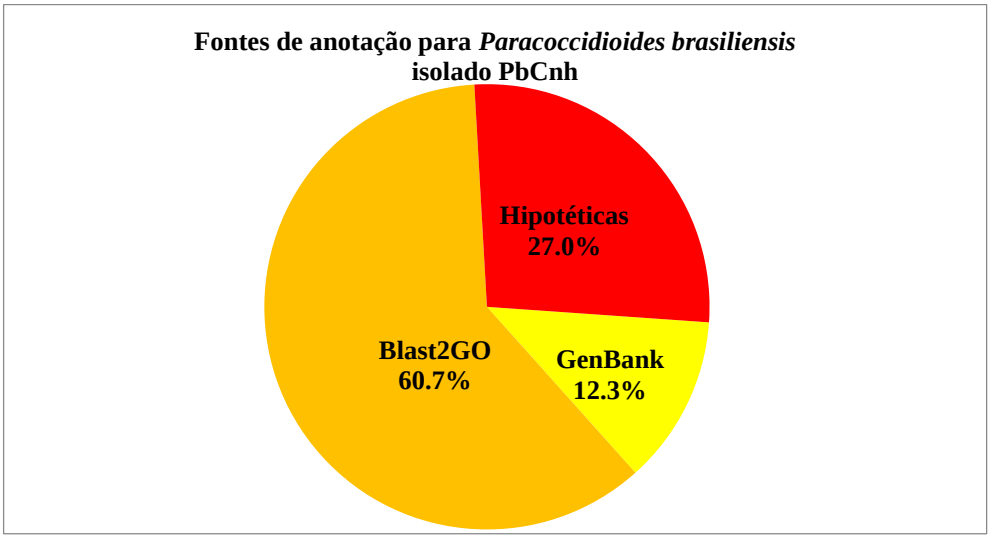
Gráfico 8 – Anotação funcional do genoma CDS de *Paracoccidioides brasiliensis* isolado Pb300. O genoma CDS do fungo foi reanotado através de BLAST no *software Blast2GO* contra a base *Yeast* do NCBI e o genoma de Pb18 que foi previamente reanotado. Os valores no gráfico indicam a porcentagem de proteínas hipotéticas e não hipotéticas em comparação ao GenBank. Total CDS = número total de genes codificadores de proteínas.



4.1.5 *Paracoccidioides brasiliensis* isolado PbCnh

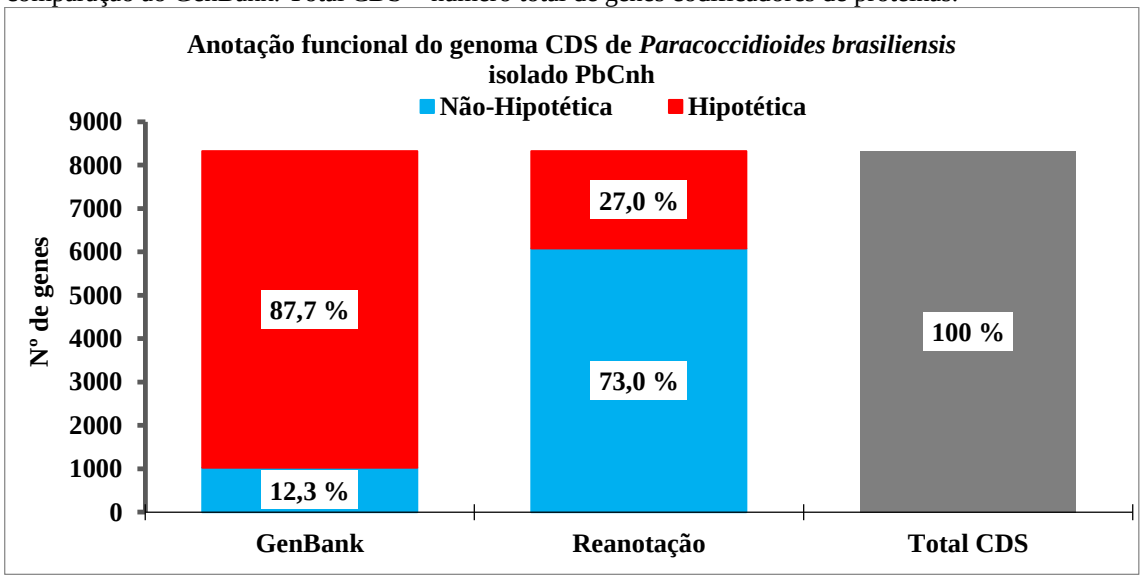
Assim como o isolado Pb300, o isolado PbCnh de *P. brasiliensis* foi recentemente sequenciado (MUÑOZ *et al*., 2016) e apresenta as mesmas características de disponibilização dos dados, apenas depositados no GenBank e 7303 genes (87,7% do genoma CDS) estavam anotados como proteínas hipotéticas. A reanotação manteve a anotação dos 1021 genes (12,3% do genoma CDS) presentes no GenBank para este isolado. No entanto, houve a adição de anotação para 5053 genes (60,7% do genoma CDS) a partir do *software Blast2GO*, restando apenas 2248 genes (27,0% do genoma CDS) representados como proteínas hipotéticas (Gráfico 9).

Gráfico 9 – Fontes de anotação funcional para *Paracoccidioides brasiliensis* isolado PbCnh. O genoma CDS do fungo foi reanotado através de BLAST no *software Blast2GO* contra a base *Yeast* do NCBI e o genoma de *P. brasiliensis* 18 que já havia sido reanotado. Os valores no gráfico indicam a porcentagem de genes anotados por cada anotador.



O sucesso na reanotação funcional para o genoma do PbCnh em relação à anotação do genoma GenBank fica evidente nos dados apresentados no Gráfico 10, onde 6076 genes (73% do genoma CDS) foram reanotados com descrição de função e somente 2248 genes (27% do genoma CDS) permaneceram sem descrição de função. Dados do GenBank apresentavam somente 1013 genes anotados (12,3% do genoma CDS).

Gráfico 10 – Anotação funcional do genoma CDS de *Paracoccidioides brasiliensis* isolado PbCnh. O genoma do fungo foi reanotado por BLAST no *software Blast2GO* contra a base *Yeast* do NCBI e o genoma de Pb18 que já havia sido reanotado. Os valores no gráfico indicam a porcentagem de proteínas hipotéticas e não hipotéticas em comparação ao GenBank. Total CDS = número total de genes codificadores de proteínas.



4.2 ESTRUTURAÇÃO DOS DADOS OBTIDOS

Para organizar os dados da anotação funcional gerada para os genomas públicos dos cinco grupos de *Paracoccidioides* spp., foram criadas planilhas nos formatos *.xlsx e *txt (texto separado por tabulação). Estas planilhas apresentam colunas provenientes das fontes de dados de cada processo de anotação. A planilha gerada para Pb18, o primeiro isolado a passar pelos processos de anotação, foi estruturada com 26 colunas, descritas no Quadro 5. As colunas B2G-merge-description, #GO, GO IDs, GO Names, Enzyme Codes, Enzyme Names, InterPro IDs, InterPro GO IDs, InterPro GO Names, INTERPRO, PFAM, PIR_SUPERFAMILY, PRINTS, PRODOM, PROSITE, SMART, SUPFAM e TIGRFAMS são produtos das fontes de anotação do *software Blast2GO*. Já as colunas DAVID-PD, DAVID-GENE-NAME, DAVID-BP, DAVID-CC e DAVID-MF são provenientes das opções de categorias de anotação do *database* DAVID. A coluna Consenso apresenta o consenso de descrições entre todas as fontes de anotação utilizadas na reanotação. A escolha dessas colunas se fez para que, ao ser consultada a descrição consenso, seja possível também verificar a anotação que veio de cada fonte de anotação e a sua consistência, ao invés de disponibilizar de forma restritiva apenas a descrição final de um determinado gene, como acontece nos bancos do NCBI.

Quadro 5 – Colunas da tabela construída para organização os dados de anotação funcional obtidos para Pb18 e descrições do conteúdo de cada coluna. O genoma CDS do fungo foi reanotado através de BLAST pelo *software Blast2GO* contra a base *Yeast* do NCBI e genomas dos demais representantes de *Paracoccidioides*, e também foram usados dados de anotação funcional obtidos através do *database* DAVID.

Coluna	Conteúdo
ID	Identificador locus_tag
RefSeq-Desc	Descrição de anotação presente no genoma *RefSeq*
B2G-merge-description	Consenso de descrições de anotação obtidas via *Blast2GO*
#GO	Número de termos *Gene Onthology* mapeados para os resultados de BLAST
GO IDs	Identificadores *Gene Onthology* mapeados para os resultados de BLAST
GO Names	Termos *Gene Onthology* mapeados para os resultados de BLAST
Enzyme Codes	Códigos EC mapeados para os resultados de BLAST
Enzyme Names	Termos EC mapeados para os resultados de BLAST
InterPro IDs	Identificadores dos termos obtidos pelo *InterProScan* via *Blast2GO*
InterPro GO IDs	Identificadores Gene Onthology obtidos pelo InterProScan via *Blast2GO*
InterPro GO Names	Termos Gene Onthology obtidos pelo InterProScan via *Blast2GO*
INTERPRO	Termos InterPro obtidos via DAVID *database*
PFAM	Termos PFAM obtidos via DAVID *database*
PIR_SUPERFAMILY	Termos PIR obtidos via DAVID *database*
PRINTS	Termos PRINTS obtidos via DAVID *database*
PRODOM	Termos PRODOM obtidos via DAVID *database*
PROSITE	Termos PROSITE obtidos via DAVID *database*
SMART	Termos SMART obtidos via DAVID *database*
SUPFAM	Termos SUPFAM obtidos via DAVID *database*
TIGRFAMS	Termos TIGRFAMS obtidos via DAVID *database*
DAVID-PD	Consenso de descrições de domínios de proteínas via DAVID *database*
DAVID-GENE-NAME	Nomes de genes obtidos via DAVID *database*
DAVID-BP	Termos *Gene Onthology* Processo Biológico obtidos via DAVID *database*
DAVID-CC	Termos *Gene Onthology* Componente Celular obtidos via DAVID *database*
DAVID-MF	Termos *Gene Onthology* Função Molecular obtidos via DAVID *database*
Consenso	Consenso de descrições entre todas as fontes de anotação

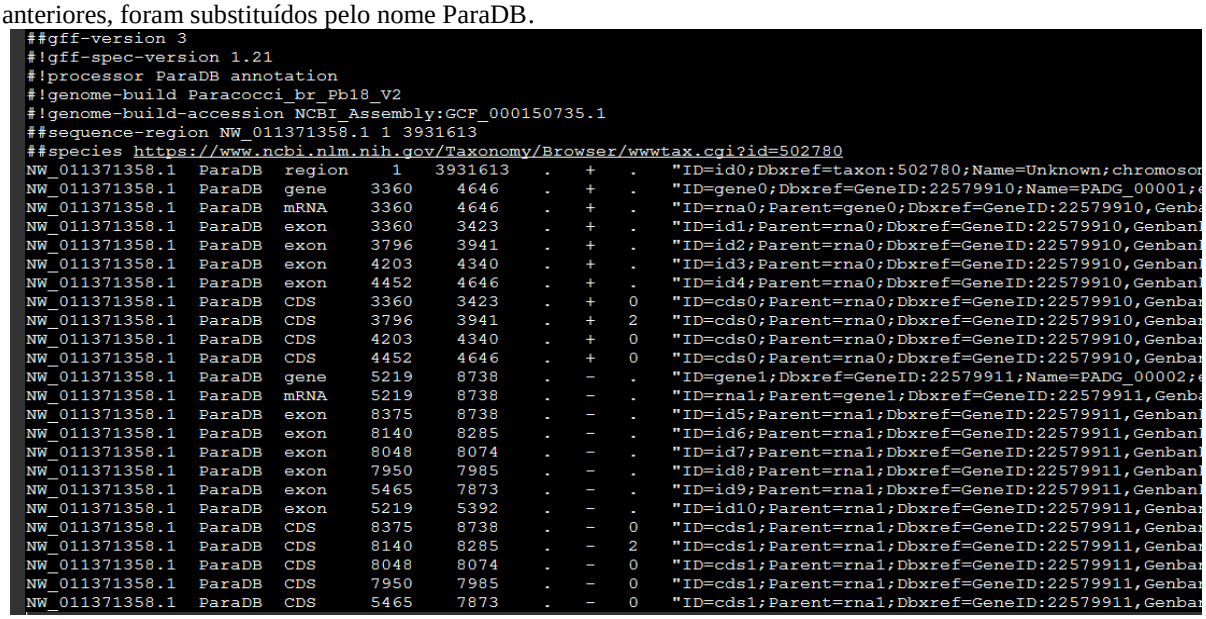
Os dados das planilhas finais de anotação foram usados para construir arquivos multi-fasta contendo dados de ID *locus_tag*, ID de produto proteico, descrição consenso, descrição de domínio proteico, formando o *string* (texto do cabeçalho de arquivos fasta) e a sequência de aminoácidos ou nucleotídeos. A Figura 8 mostra um exemplo de três sequências do arquivo multi-fasta de Pb18. Os arquivos multi-fasta, diferentemente das planilhas, apresentam menor quantidade de fontes de informação funcional, para garantir que o tamanho do arquivo não seja muito grande. Estes arquivos, disponibilizados no *database* ParaDB, serão úteis na transmissão da reanotação funcional realizada, uma vez que este formato é padronizado e um dos mais usados na bioinformática como fonte primaria de descrição dos genes (LEON e MARKEL, 2003). Arquivos no formato GFF3, outro formato padronizado e também muito utilizado em análises genômicas, foram construídos. Estes arquivos são necessários, pois armazenam, além

do nome dos genes, características como nome do cromossomo, nome do tipo de característica como gene, cDNA, mRNA, éxon, entre outros (GMOD, 2016). Foram criados arquivos GFF3 atualizados a partir dos arquivos GFF3 disponíveis nas bases RefSeq no caso de Pb18 e Pb01 e GenBank no caso de Pb03, Pb300 e PbCnh. Com auxílio de pacote Excel, a descrição da anotação funcional proveniente de RefSeq e GenBank foi substituída pelas descrições obtidas na reanotação. Os nomes RefSeq/GenBank/NCBI, que indicavam as fontes de anotação anteriores, foram substituídos pelo nome ParaDB, como pode ser visto no exemplo do arquivo criado para *P. brasiliensis* isolado Pb18, apresentado na Figura 9.

Figura 8 – **Exemplo do arquivo multi-fasta gerado para Pb18 mostrando três sequências**. O arquivo foi construído formado por um cabeçalho contendo ID *locus_tag*, ID de produto proteico e a descrição consenso final após os processos de anotação funcional e de curadoria manual.

```
>PADG_00001 | XM_010757860  | XP_010756162 | Peptidyl-prolyl cis-trans isomerase H |
peptidyl-prolyl cis-trans isomerase H(PADG_00001)|
MVARPRRNPDNPVVFFDITLGGQELGRIKMELFADVTPRTAENFRQFCTGEAKNARGKSQGYKGSKFHRVIKEFMIQGGDFINGDGTG
SASIYGSGKFADENFKISHDGPGLLSMANSGPNTNGCQFFITTTATPFLNNKHVVFGQVIEDKDNVVRRIENTNTKRDKPNQDVVIAQ
CGQI
>PADG_00002 | XM_010757348  | XP_010755650 | Alanyl-tRNA synthetase | alanyl-tRNA
synthetase(PADG_00002)|
MGVGGERVSGVVLHQGLSLTRYCSRSVCRNLVEVVGVNVGGSVPVPLPPSRRFFAWHFFSPAITNTFSTTPLFFRRHNSTLPRSQAAF
NMTSTQMVEQPEWPALRVRNAFLDFFKDNGHTFVPSSSVVPLSDPTLLFANAGMNQYKAIFLGTVEPNSDFAQLKRAHNTQKCIRAGG
KHNDLDDVGKDSYHHTFFEMLGNWSFGDYFKKEAIRYSWDLLTKVYGLDPDRLYVTYFEGNPDAGIEPDLEARDLWLSVGVAEDHLLS
GNMKDNFWEMGEQGPCGPCSEIHYDRIGGRNASHLVNQDDPNVLEIWNNVFIQYNREQDKSLRPLPNKHVDTGMGYERLVSILQNKSS
NYDTDVFTPLFEKIREITGSRPYTGKFHEEDVDGVDTAYRVVADHIRTLTFAISDGAVPNNEGRGYVVRRVLRRGARYARKYLNVEIG
GFFSQIVPTLVEQMGDMFPEIKRKATDVMEILDEEEISFAKTLDRGERLFEEYALQAKQKGLDRLHGADVWRLYDTFGFPVDLTQLMA
EERGLKIDNGEFEEARLRAKEASKGQGKQTKDVVKLTVHDLGLLETMDDVPKTDDSAKFGRGNITSQIKAIYHNKKFVDSSKDIPEGE
QFGIILDRTNFYAEQGGQEYDTGKIIIDGKAELAVEDVQVYGGYVLHTGFMKYGSFNINDSVIAEYDELRRWPIRNNHTGTHILNFAL
RKVLGDGVEQKGSLVAAEKLRFDFSHKSAVSDSDLEEIERISTDYIRQNCAVYGKDVPLSIAREITGVRAVFGETYPDPVRVVSVGVE
LEEILKDVHDPRWKEISIEFCGGTHVQKTGDIKDLVILEESGIAKGIRRIIAVTGEEAHEVQRIAKEFGERLKRFEKMELGPKKEMEA
KLLQVDLNQLTISAVEKAQFREKFTQIHKKVLEGQKAAQKLESKKALDAIMGYFQAPENKDAPHLVLQLPISANAKAVSDSLNFVKTK
MQDKSLYVFAADKEAGKIVHGCHVAENLSQQGASPNDWANSVANVVGGRAGGKGPISIGNGTDAEKLDDAIKAATEYLEKFKL
>PADG_00003 | XM_010757861  | XP_010756163 | Hypothetical protein PADG_00003 |
no-hit-All|
MAPVSGVSLASEVRGKKRHAAEELEGEQRLTKKFGLLHIGRIGQSYPPSILDNACTIKKTPTPGFGAEQPKPPAAVAAAATNESMQID
ETKDRIYIHDLESEFAKIEAEENNVAILPEIEKKLMSVPKSVLSTKPSARNELVLYRLPSSLSVPEPQDSVRKAIIEARERAREKAIA
DQMEKEKVEGEGEEKIITQHFLSHEMNGSLMDLLPPSSSALLPSPESLDLNSQDDPDAMDIDSA
```

Figura 9 – Exemplo do arquivo GFF3 gerado para Pb18 mostrando as principais características das sequências. O a partir dos arquivos GFF3 disponíveis nas bases RefSeq no caso de Pb18 e Pb01 e GenBank no caso de Pb03, Pb300 e PbCnh. A descrição da anotação funcional proveniente das bases do NCBI foi substituída pelas descrições obtidas na reanotação. Os nomes RefSeq/GenBank/NCBI, que indicavam as fontes de anotação anteriores, foram substituídos pelo nome ParaDB.



```
##gff-version 3
#!gff-spec-version 1.21
#!processor ParaDB annotation
#!genome-build Paracocci_br_Pb18_V2
#!genome-build-accession NCBI_Assembly:GCF_000150735.1
##sequence-region NW_011371358.1 1 3931613
##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=502780
NW_011371358.1  ParaDB  region    1     3931613  .  +  .  "ID=id0;Dbxref=taxon:502780;Name=Unknown;chromoso
NW_011371358.1  ParaDB  gene      3360  4646     .  +  .  "ID=gene0;Dbxref=GeneID:22579910;Name=PADG_00001;
NW_011371358.1  ParaDB  mRNA      3360  4646     .  +  .  "ID=rna0;Parent=gene0;Dbxref=GeneID:22579910,Genb
NW_011371358.1  ParaDB  exon      3360  3423     .  +  .  "ID=id1;Parent=rna0;Dbxref=GeneID:22579910,Genban
NW_011371358.1  ParaDB  exon      3796  3941     .  +  .  "ID=id2;Parent=rna0;Dbxref=GeneID:22579910,Genban
NW_011371358.1  ParaDB  exon      4203  4340     .  +  .  "ID=id3;Parent=rna0;Dbxref=GeneID:22579910,Genban
NW_011371358.1  ParaDB  exon      4452  4646     .  +  .  "ID=id4;Parent=rna0;Dbxref=GeneID:22579910,Genban
NW_011371358.1  ParaDB  CDS       3360  3423     .  +  0  "ID=cds0;Parent=rna0;Dbxref=GeneID:22579910,Genba
NW_011371358.1  ParaDB  CDS       3796  3941     .  +  2  "ID=cds0;Parent=rna0;Dbxref=GeneID:22579910,Genba
NW_011371358.1  ParaDB  CDS       4203  4340     .  +  0  "ID=cds0;Parent=rna0;Dbxref=GeneID:22579910,Genba
NW_011371358.1  ParaDB  CDS       4452  4646     .  +  0  "ID=cds0;Parent=rna0;Dbxref=GeneID:22579910,Genba
NW_011371358.1  ParaDB  gene      5219  8738     .  -  .  "ID=gene1;Dbxref=GeneID:22579911;Name=PADG_00002;
NW_011371358.1  ParaDB  mRNA      5219  8738     .  -  .  "ID=rna1;Parent=gene1;Dbxref=GeneID:22579911,Genb
NW_011371358.1  ParaDB  exon      8375  8738     .  -  .  "ID=id5;Parent=rna1;Dbxref=GeneID:22579911,Genban
NW_011371358.1  ParaDB  exon      8140  8285     .  -  .  "ID=id6;Parent=rna1;Dbxref=GeneID:22579911,Genban
NW_011371358.1  ParaDB  exon      8048  8074     .  -  .  "ID=id7;Parent=rna1;Dbxref=GeneID:22579911,Genban
NW_011371358.1  ParaDB  exon      7950  7985     .  -  .  "ID=id8;Parent=rna1;Dbxref=GeneID:22579911,Genban
NW_011371358.1  ParaDB  exon      5465  7873     .  -  .  "ID=id9;Parent=rna1;Dbxref=GeneID:22579911,Genban
NW_011371358.1  ParaDB  exon      5219  5392     .  -  .  "ID=id10;Parent=rna1;Dbxref=GeneID:22579911,Genba
NW_011371358.1  ParaDB  CDS       8375  8738     .  -  0  "ID=cds1;Parent=rna1;Dbxref=GeneID:22579911,Genba
NW_011371358.1  ParaDB  CDS       8140  8285     .  -  2  "ID=cds1;Parent=rna1;Dbxref=GeneID:22579911,Genba
NW_011371358.1  ParaDB  CDS       8048  8074     .  -  0  "ID=cds1;Parent=rna1;Dbxref=GeneID:22579911,Genba
NW_011371358.1  ParaDB  CDS       7950  7985     .  -  0  "ID=cds1;Parent=rna1;Dbxref=GeneID:22579911,Genba
NW_011371358.1  ParaDB  CDS       5465  7873     .  -  0  "ID=cds1;Parent=rna1;Dbxref=GeneID:22579911,Genba
```

4.3 AVALIAÇÃO E DESCRIÇÃO DO *DATABASE*

Como pode ser visto na Figura 10, o banco de dados ParaDB, acessível pelo endereço **http://paracoccidioides.com**, apresenta interface simples e intuitiva, e as informações genômicas do gênero *Paracoccidioides* estão organizados com as espécies *P. brasiliensis* e *P.lutzii* separadamente e podem ser acessados por botões na região superior da página. Também há botões na região central da página principal para acessar os dados de cada isolado de *Paracoccidioides*.

Na Figura 11 é possível verificar a disposição das descrições de anotação funcional obtidas na reanotação realizada, onde foi realizado um teste de busca pela palavra "*chitin*" (quitina) e os resultados para essa busca retornam todos os genes cuja descrição de anotação funcional contém o termo buscado. Nessa página também são visualizados os botões "*Columns*" (colunas), para alterar e escolher quais colunas o usuário deseja que apareçam nos resultados da busca por palavra, além dos botões para descarregar os conjuntos de dados acessados em formatos Excel, csv e PDF, como também há a opção "*copy*" (copiar), para copiar o texto visualizado como resultado da pesquisa.

Figura 10 – **Imagem de teste da página inicial do banco de dados ParaDB.** O banco de dados ParaDB apresenta interface simples e intuitiva, e as informações genômicas do gênero *Paracoccidioides* estão organizadas com as espécies *P. brasiliensis* e *P.lutzii* separadamente e podem ser acessadas por botões na região superior da página. Também há botões na página principal para acessar os dados de cada isolado de *Paracoccidioides*.

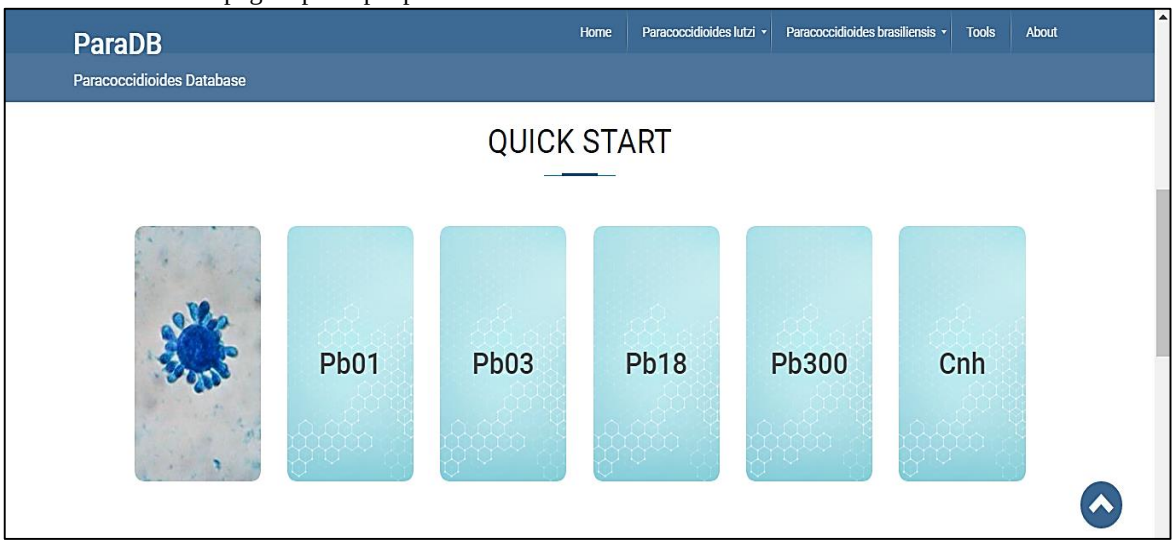
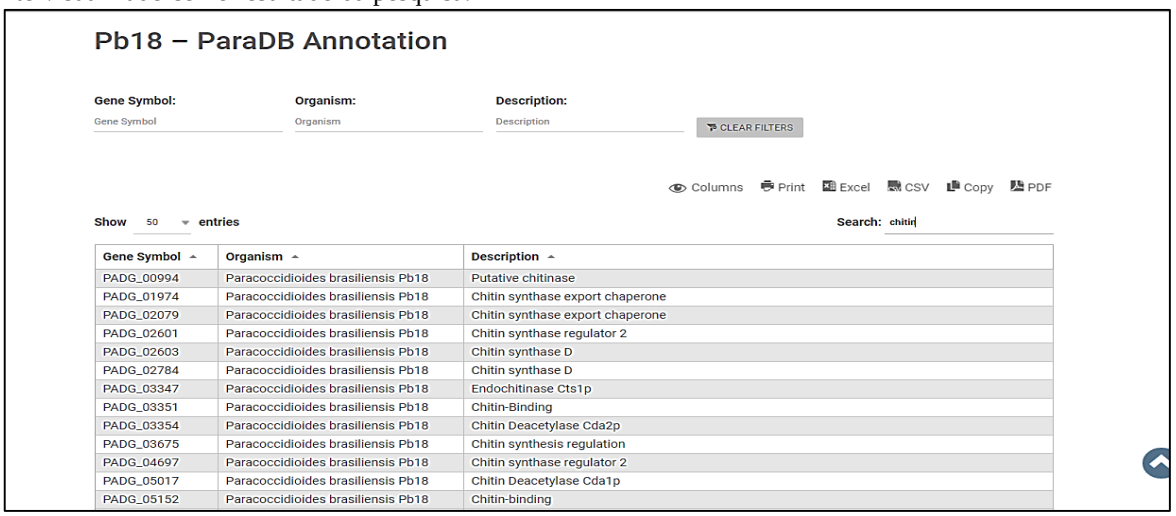


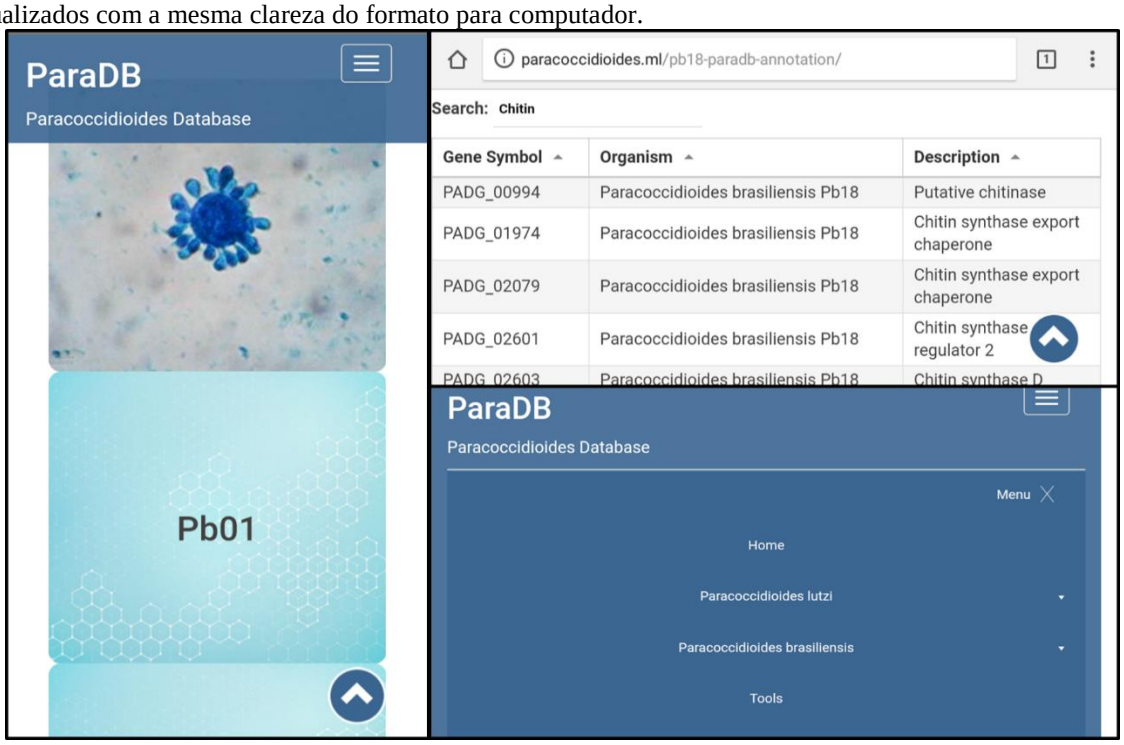
Figura 11 – **Imagem de teste de busca de palavras na página de *P. brasiliensis* isolado Pb18 no banco de dados ParaDB.** É possível ver a disposição das descrições de anotação funcional obtidas na reanotação realizada, onde foi realizado um teste de busca pela palavra "*chitin*" (quitina) e os resultados para essa busca retornam todos os genes que contém, na sua descrição, o termo buscado. Também são visualizados os botões "*Columns*" (colunas), para escolher quais colunas o usuário deseja ver nos resultados da busca, além dos botões para descarregar os conjuntos de dados acessados em formatos Excel, csv e PDF. Também há a opção "*copy*" (copiar), para copiar o texto visualizado como resultado da pesquisa.



Gene Symbol	Organism	Description
PADG_00994	Paracoccidioides brasiliensis Pb18	Putative chitinase
PADG_01974	Paracoccidioides brasiliensis Pb18	Chitin synthase export chaperone
PADG_02079	Paracoccidioides brasiliensis Pb18	Chitin synthase export chaperone
PADG_02601	Paracoccidioides brasiliensis Pb18	Chitin synthase regulator 2
PADG_02603	Paracoccidioides brasiliensis Pb18	Chitin synthase D
PADG_02784	Paracoccidioides brasiliensis Pb18	Chitin synthase D
PADG_03347	Paracoccidioides brasiliensis Pb18	Endochitinase Cts1p
PADG_03351	Paracoccidioides brasiliensis Pb18	Chitin-Binding
PADG_03354	Paracoccidioides brasiliensis Pb18	Chitin Deacetylase Cda2p
PADG_03675	Paracoccidioides brasiliensis Pb18	Chitin synthesis regulation
PADG_04697	Paracoccidioides brasiliensis Pb18	Chitin synthase regulator 2
PADG_05017	Paracoccidioides brasiliensis Pb18	Chitin Deacetylase Cda1p
PADG_05152	Paracoccidioides brasiliensis Pb18	Chitin-binding

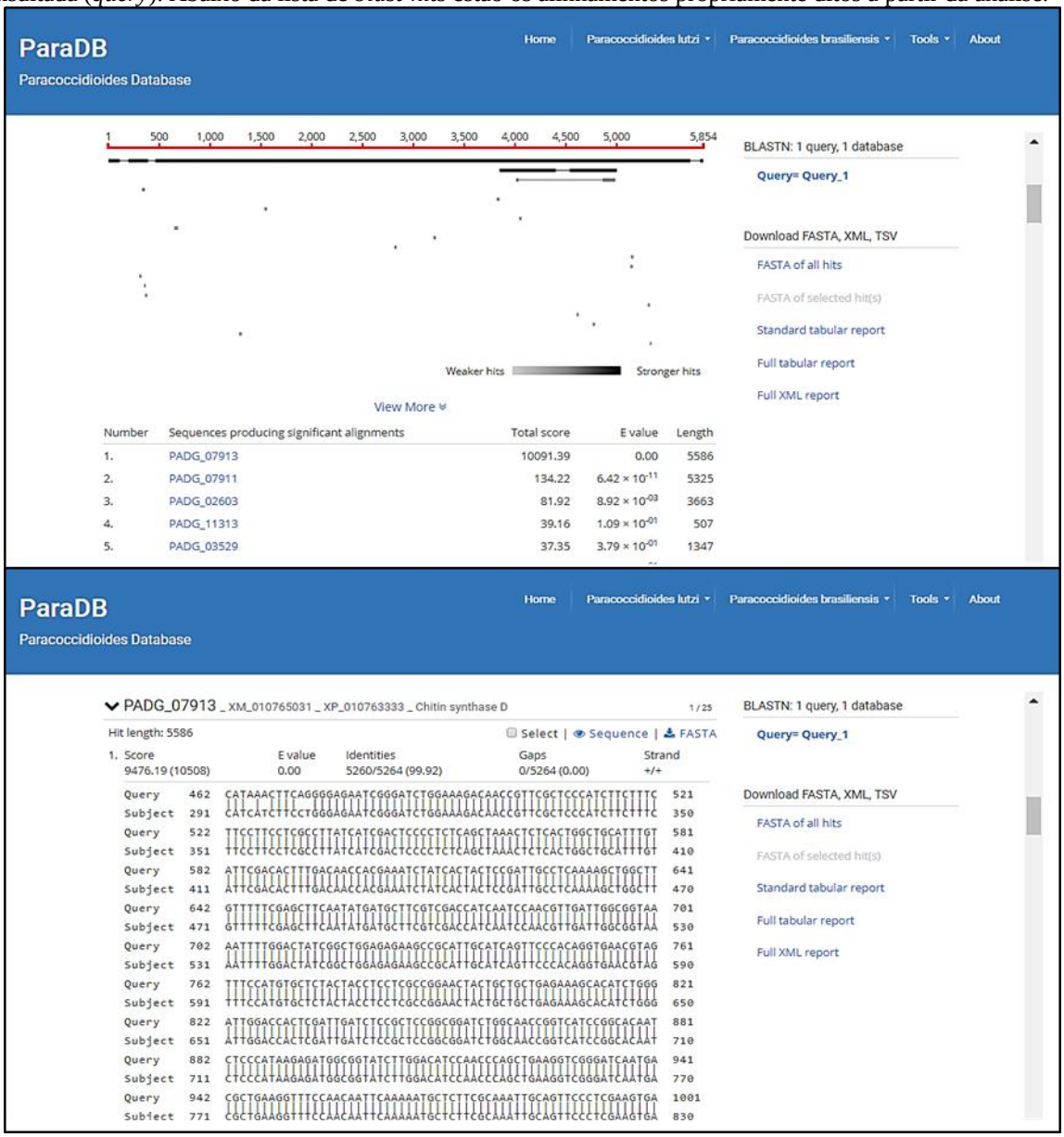
No ParaDB é possível, inclusive, acessar os dados por meio de dispositivos móveis como celulares e *tablets*, com sistema operacional Android ou IOs, com todas as funcionalidades da versão para computador do banco de dados, como mostrado na Figura 12, onde foi realizada a mesma busca feita anteriormente na versão para computador, utilizando o termo "*chitin*".

Figura 12 – **Imagens da versão para dispositivos móveis do banco de dados ParaDB.** A interface da versão mobile se adequa perfeitamente as proporções da tela dos dispositivos, tanto na posição retrato (esquerda) quanto na posição paisagem (direita). As funcionalidades da versão para computador do banco de dados são acessíveis por celular ou *tablet*. Foi feita a mesma busca anterior pelo termo "*chitin*", mostrando que os resultados são visualizados com a mesma clareza do formato para computador.



Uma das ferramentas principais disponibilizadas no ParaDB é o recurso BLAST, que pode ser realizado contra os genomas de *Paracoccidioides* depositados contendo as anotações funcionais atualizadas. Como demonstrado na Figura 13, a sequência do gene PADG_07913 (*Chitin Synthase* D) foi submetida a um BLASTn, ou seja, um BLAST contra o genoma de nucleotídeos de Pb18. A representação dos alinhamentos resultantes dessa análise são mostrados na parte superior da página, e logo abaixo, são exibidos os genes que tiveram alinhamento significativo (*blast-hits*) para a sequência consultada (*query*). Abaixo da lista de *blast-hits* encontram-se os alinhamentos propriamente ditos a partir da análise. Adicionalmente, no ParaDB, também é possível realizar BLAST contra os dados de proteínas, o BLASTp. A ferramenta de BLAST do ParaDB é capaz de reconhecer automaticamente o tipo de material genético inserido para consulta, não sendo necessário escolher entre BLASTn ou BLASTp.

Figura 13 – **Imagens das análises de BLAST realizadas no banco de dados ParaDB.** A sequência do gene PADG_07913 (*Chitin Synthase* D) foi submetida a um BLASTn, ou seja, um BLAST contra o genoma de nucleotídeos de Pb18. A representação alinhamentos resultantes da análise são mostrados na parte superior da página e, logo abaixo, são exibidos os genes que tiveram alinhamento significativo (*blast-hits*) para a sequência consultada (*query*). Abaixo da lista de *blast-hits* estão os alinhamentos propriamente ditos a partir da análise.



5 DISCUSSÃO

O gênero *Paracoccidioides* é composto por fungos termodimórficos ascomicetos, agentes etiológicos da Paracoccidioidomicose (PCM), micose sistêmica prevalente na América Latina. O Brasil soma 80% dos casos descritos e, no país, a doença está em primeiro lugar em mortes por doenças infecciosas (COUTINHO *et al*., 2002; BELLISSIMO-RODRIGUES *et al*., 2011). A doença é adquirida através da inalação de propágulos miceliais do fungo, que sofrem mudança para a forma de levedura no hospedeiro, a 37°C. A forma crônica da doença é mais frequente e de disseminação multifocal, com envolvimento dos pulmões, linfonodos, pele e mucosas. Essa forma tem evolução crônica com diagnóstico tardio (WANKE e AIDÊ, 2009). O tratamento farmacológico tradicional da PCM é longo e inclui o uso de antifúngicos com notável toxicidade e efeitos adversos frequentes, levando à frequente descontinuação do tratamento por parte dos pacientes e consequentemente reincidência dos sintomas e surgimento de sequelas (SHIKANAI-YASUDA *et al*., 2006; TUON *et al*., 2013).

Diante da importância da PCM para a América Latina, visto que a doença é endêmica nesse local, no ano de 2007, Cardoso e colaboradores (2007) publicaram o genoma mitocondrial de Pb18. Anos mais tarde, avanços foram alcançados com a publicação de genomas de três isolados, a saber, Pb18, Pb03 e *P. lutzii* (Pb01) (DESJARDINS *et al*., 2011). Já em 2014, tecnologias mais modernas de sequenciamento e análises de bioinformática foram utilizadas para atualizar os dados de sequenciamento (MUÑOZ *et al*., 2014). Mais recentemente, em 2016, foram publicados mais dois genomas de *Paracoccidioides*, representando os isolados Pb300 e PbCnh, concluindo assim o sequenciamento dos principais representantes de todos os cinco principais grupos filogenéticos do gênero (MUÑOZ *et al*., 2016). Os projetos de sequenciamento deste patógeno humano contribuíram na compreensão sobre seus principais mecanismos biológicos e também acerca dos processos relacionados a patogenicidade e interação entre o fungo e o hospedeiro.

Entretanto, as informações a respeito da função da maior parte dos cerca de 9000 genes presentes nos diferentes grupos filogenéticos do fungo permaneciam ambíguas. Esta escassez de dados referente à anotação funcional dos genomas é representada pela grande quantidade de genes anotados como proteínas hipotéticas, que são proteínas cuja função é desconhecida ou incerta. Esse tipo de anotação traz pouca informação sobre a função real dessas moléculas e apenas uma minoria delas apresenta os termos GO associados, mesmo que apenas "potencialmente associados", ou seja, agregados na informação geral disponível nos bancos, que só pode ser acessada de maneira individual, ou seja, gene-a-gene.

Neste sentido, o alvo do presente estudo foi aplicar análises computacionais automatizadas de anotação funcional aos genomas atualmente disponíveis do gênero *Paracoccidioides* através de *softwares* especializados para este tipo de abordagem, além de realizar curadoria manual após os processos automáticos, de maneira a garantir maior confiabilidade e informação a respeito das funções dos genes deste fungo. Adicionalmente, almejou-se disponibilizar a anotação em um banco de dados *online* que centralizasse e permitisse acesso rápido e eficaz aos genomas do gênero *Paracoccidioides*.

É interessante atentar-se ao fato de que este problema relacionado à falta de informações funcionais nos bancos de dados de sequências não é algo recente no que diz respeito ao gênero *Paracoccidioides*. Desde a versão anterior do genoma do fungo (versão v1) era evidente a quantidade defasada de informações acerca da função dos genes do patógeno. Na primeira versão do genoma, era possível encontrar 61,7% dos genes anotados como responsáveis por proteínas hipotéticas em Pb01, (9132 de 5636 proteínas), 60,8% em Pb18 (5314 de 8741 proteínas) e 60,2% no isolado Pb03 (4742 de 7876 proteínas).

Em 2015, Silva e colaboradores usaram abordagens *in silico* na tentativa de reanotar o genoma de Pb01. Inclusive, para tanto, utilizaram a ferramenta de bioinformática Blast2GO e, ainda, utilizando dados obtidos de um *database* de sequências expressas do NCBI, investigaram a possível expressão das proteínas hipotéticas. Como resultado, reanotaram 3044 (33%) proteínas das 5636 (62%) proteínas hipotéticas e verificaram que 2364 (26%) pareciam estar expressas em diferentes situações como nas formas miceliais e leveduriformes, durante a transição micélio-levedura e sob condições que mimetizam a infecção (SILVA *et al*., 2015).

Embora seja conveniente visualizar trabalhos similares ao presente estudo, se torna importante refletir que mesmo com tecnologias mais modernas de sequenciamento e montagem dos genomas, a incerteza da importância de grande parte dos genes se repetiu na atualização dos genomas dos organismos causadores da micose sistêmica mais proeminente da América Latina e que, apesar de ser negligenciada, já é "exportada" para outros locais do mundo (MARTINEZ, 2015).

Pode-se estender o olhar a respeito deste problema em outros fungos. No genoma de referência da levedura *Saccharomyces cerevisiae*, por exemplo, 719 genes recebem como anotação a descrição proteína hipotética no banco RefSeq, o que equivale a aproximadamente 12% do genoma do microrganismo (6002 genes codificadores de proteínas, cepa S288C). Verificando os dados do mesmo genoma disponíveis na *Saccharomyces Genome Database*, ou SGD (https://downloads.yeastgenome.org/sequence/S288C_reference/genome_releases/), a base de dados própria do organismo esse número é reduzido para 363 genes codificadores

(aproximadamente 6%), incluindo os genes que apresentam a descrição *protein of unknown function*, "proteína de função desconhecida". Entretanto, apenas 1% destes 363 genes não contém, no genoma disponível no SGD, informações acerca do domínio proteico, localização celular, e co-expressão. Dessa forma, denota-se a enorme contribuição dos desenvolvedores desta base de dados na providência de anotação funcional, mesmo para proteínas com função não elucidada e destaca-se a importância na concentração de esforços para a criação, manutenção e desenvolvimento de plataformas dedicadas a espécies/gêneros.

Com relação ao genoma de referência de *Candida albicans* (MUZZEY *et al*., 2013), outra levedura de grande importância médica, são 2330 genes codificadores com descrição de proteína hipotética no banco RefSeq, isto é, aproximadamente 39% de todas as 6030 proteínas na cepa SC5314. Consultando os arquivos de proteínas do *Candida* Genome *Database*, ou CGD (http://www.candidagenome.org/), somente 583 proteínas apresentam função desconhecida (*protein unknown function*). Dessa forma, o número de proteínas desconhecidas cai para menos de 10% do total de proteínas para esta cepa de *C. albicans* no banco CGD. Portanto, assim como descrito para *Saccharomyces*, o banco dedicado e reanotado para *Candida albicans* apresenta informações mais completas do que a anotação RefSeq.

Quando se verifica a situação dos bancos de dados gerais para as doenças negligenciadas, como é o caso da PCM, a situação da anotação dos genomas se apresenta ainda menos informativa. A micose conhecida por aspergilose é causada pelo fungo *Aspergillus fumigatus* e o isolado *A. fumigatus* Af293 possui 12260 genes que codificam proteínas em seu genoma. Cerca de 43% destes genes (5305) são identificados como proteínas hipotéticas segundo anotação encontrada no RefSeq. No entanto, em sua base centralizada, a AspGD (*Aspergillus* Genome *Database*, http://www.aspergillusgenome.org/), apenas 398 genes, ou seja, 3,2% do genoma codificador não possui função definida. Como as outras espécies de fungos citadas acima, os arquivos de sequências de material genético de *A. fumigatus* Af293 disponibilizados em sua base de dados genômicos oficial (http://www.aspergillusgenome.org/download/sequence/A_fumigatus_Af293/current/) são mais informativos que os dados contidos no RefSeq e ainda apresenta informações de co-expressão, localização celular e domínios proteicos associadas às descrições dos genes sem função proteica definida.

A base de dados HistoBase (http://histo.ucsf.edu/) contém arquivos de genoma e anotações para os isolados G217B, G186AR, H88 e H143 do fungo *Histoplasma capsulatum*, agente etiológico da histoplasmose, micose sistêmica negligenciada em muitos países. Infelizmente, os dados de anotações não estão acessíveis, possivelmente por uma falha nos

servidores do seu sistema em Janeiro de 2017, como informado na sua página. Da mesma forma do que pode ser verificado com a anotação RefSeq para a PCM e aspergilose, o isolado G186AR de *H. capsulatum* apresenta 3046 proteínas descritas como hipotéticas, dentre suas 9254 proteínas preditas, sendo assim, aproximadamente 33% do genoma consta com anotação de função incerta. Para o isolado H88 do patógeno, com 9445 genes codificadores de proteína em seu genoma, cerca de 2907 (31%) são hipotéticas. O fato do *database* dedicado à espécie estar fora de atividade, impossibilita, no momento, uma comparação entre as anotações presentes no RefSeq. No entanto, destaca-se a importância e o compromisso em, a partir da criação de uma ferramenta *online* e *open source*, que sejam mantidos o acesso e a manutenção do site, de maneira a não prejudicar a comunidade científica que trabalha com o modelo.

A blastomicose é outra micose negligenciada, causada pelo fungo *Blastomyces dermatidis*. O genoma da cepa ER-3, composto de 11539 proteínas, apresenta 5507 proteínas hipotéticas segundo GenBank. Praticamente 50% do total de genes codificadores deste fungo patogênico não apresentam descrição de função. Interessantemente, este microrganismo não apresenta um banco de dados individual até o momento, o que deve dificultar o estudo funcional das proteínas envolvidas nos mecanismos biológicos e de patogenicidade em que esse fungo possa estar envolvido, assim como ocorria nos estudos envolvendo o gênero *Paracoccidioides* até o momento. O mesmo pode ser verificado para os agente etiológicos das micoses negligenciadas coccidioidomicose e criptococose, causadas por *Coccidioides immitis* e *Cryptococcus*, respectivamente. Não existem *databases* dedicados aos genomas de ambas espécies e sua anotação disponibilizada atualmente é deficitária. Por exemplo, para o isolado RS de *Coccidioides immitis*, aproximadamente 49% dos genes descritos como proteínas hipotéticas e, para *Cryptococcus neoformans* var. *grubii* H99, aproximadamente 47% não possui descrição de função no banco de dados GenBank.

Em suma, pode-se notar que o número de proteínas sem funcionalidade atribuída nos bancos universais do NCBI pode chegar a até 50%, destacando apenas os organismos comentados aqui. Este número é ainda mais preocupante quando se trata dos genomas de *Paracoccidioides*, tanto na primeira versão, quanto na segunda. No entanto, cabe chamar a atenção para os bancos centralizados dos genomas, uma vez que apresentam menor quantidade de proteínas com função desconhecida e ainda carregam informações complementares que podem contribuir nos estudos genômicos destes importantes microrganismos causadores de doenças importantes.

Atualmente, as sequências geradas pelo projeto genoma *Paracoccidioides* estão disponibilizadas nos bancos de dados GenBank (https://www.ncbi.nlm.nih.gov/genbank/),

RefSeq (https://www.ncbi.nlm.nih.gov/refseq/), Ensembl (https://www.ensembl.org/i) (KERSEY *et al*., 2017), MycoCosm (https://genome.jgi.doe.gov/) (GRIGORIEV *et al*., 2012) e EuPathDB/FungiDB (http://fungidb.org) (STAJICH *et al*., 2012). Porém, nesses *databases* não há uma anotação consolidada e curada das proteínas preditas, dificultando a busca e localização de genes. Nesses bancos, cerca de 80% das sequências codificadoras encontra-se categorizadas como hipotéticas.

A melhoria das tecnologias para os métodos automatizados de anotação funcional tem contribuído no aumento do número de sequências analisadas paralelamente nas *pipelines* do NCBI, que anteriormente requeria mais suporte manual. Embora existam planos de longo-prazo por parte dos pesquisadores responsáveis pelos dados nestes bancos, eles admitem que a quantidade de dados imposta pela grande diversidade taxonômica e uma disponibilidade limitada de recursos para treinar as *pipelines* impossibilita o processamento de muitos dos genomas lá depositados (O'LEARY *et al*., 2016).

É sabido que abordagens genômicas e proteômicas de alto rendimento, como análises de expressão gênica por *microarrays* e experimentos de RNA-seq fornecem dados que permitem estudar uma grande variedade de mecanismos biológicos, incluindo associações com doenças. Essas tecnologias geralmente resultam em uma grande lista de genes de interesse (variando em tamanho de centenas para milhares de genes) envolvidos com as condições biológicas estudadas. A análise de conjuntos de dados tão complexos e de grande volume é uma tarefa desafiadora, que requer apoio de pacotes de *softwares* de bioinformática para aumentar o conhecimento biológico acerca dos genes estudados (HUANG *et al*., 2009), especialmente se o modelo de estudo contar com uma anotação funcional insuficiente, como é o caso dos genomas de *Paracoccidioides*. Portanto, uma reanotação funcional do gênero *Paracoccidioides* poderá ser de grande relevância nessas análises de larga escala, especialmente se forem disponibilizadas *online* para uso da comunidade científica.

Como já descrito, a anotação funcional de genes é uma etapa essencial para estudos de biologia molecular, já que uma anotação compreensível e acurada é fundamental para modelar dados genômicos funcionais para que derivem em conhecimento biológico. Nesse sentido, a reanotação de genomas passa a ser um diferencial nos processos de análise de dados, especialmente de larga escala, de maneira a permitir uma maior visibilidade global em termos de proteínas com função predita (VAN DEN BERG *et al*., 2010). Essas análises se fazem ainda mais necessárias quando se trabalha com modelos causadores de doenças negligenciadas, como a PCM e, por sua vez, menos reportados na literatura, como é o caso do gênero *Paracoccidioides* (SHIKANAI-YASUDA *et al*., 2017).

As primeiras tentativas de anotação funcional de genes eram baseadas em linguagem natural, ou seja, em algoritmos de indexação e pesquisa de texto simples, realizadas em *databases*. No entanto, logo no início, percebeu-se que este processo era vago e inespecífico para obter com precisão a função dos genes (YANDELL e MAJOROS, 2002). Os primeiros esforços para uma anotação estruturada e controlada foram apresentados com base na classificação enzimática (*Enzyme Classification* - EC), representada por um esquema de classificação numérica para as enzimas, baseado nas reações químicas que catalisam. Essa classificação ainda é amplamente utilizada, mas também se mostrou insuficiente para descrever com precisão a função dos genes. Isso motivou a introdução do *Gene Ontology* (GO), que se configura atualmente como o maior recurso desse tipo, baseado em ontologias biológicas para realizar a anotação funcional estruturada e padronizada, o que facilita a aplicação de ferramentas automatizadas de anotação de genomas (DU PLESSIS *et al*., 2011).

Neste trabalho, realizou-se a anotação funcional de todos os genomas publicamente disponibilizados do gênero *Paracoccidioides* através de estratégias computacionais automatizadas utilizando o *software Blast2GO* e dados obtidos do *database* DAVID, que utilizam, além de outras ferramentas, a busca por *Gene Ontology*. Estes programas possuem recursos integrados das principais bases de dados de informação estruturais e evolutivas de genes e proteínas, como InterPro, Pfam, além do GO, de maneira que foi possível realizar uma extensiva reanotação funcional dos genomas completos de *Paracoccidioides* disponibilizados nos bancos de dados RefSeq e GenBank, do NCBI. Através da "Fusão de Dados Biológicos *in silico*" (*in silico Biological Data Fusion*) (GOPAL *et al*., 2014) foi possível realizar a avaliação e comparação entre múltiplas fontes de dados biológicos, gerando um anotação final, denominada consenso.

Adicionalmente, os dados oriundos das diversas bases de dados consultadas foram organizados em tabelas no formato Excel, como também em arquivos multi-fasta e GFF3, que são muito úteis em estudos de bioinformática, destacando-se como as principais fontes de descrição de informação a respeito dos genes de um organismo (GMOD, 2017).

Além disso, os dados de reanotação e os arquivos gerados foram disponibilizados em um banco de dados relacional construído com o intuito de centralização, armazenamento e consulta dos dados gerados pelo presente trabalho. Nesse sentido, toda a informação gerada com a reanotação encontra-se arquivada nesse repositório público para informação genômica do gênero *Paracoccidioides* - o **ParaDB**, que pode ser acessado pelo endereço eletrônico **http://paracoccidioides.com.** No ParaDB, os dados disponíveis concentram informações precisas e informativas sobre o genoma de *P. brasiliensis* [isolados Pb18 (grupo S1), Pb03

(grupo PS2), Pb300 (grupo PS4) e PbCnh (grupo PS3)] e *P. lutzii*, [isolado Pb01 (grupo Pl)], possibilitando rápida e eficiente recuperação de sequências e suas respectivas anotações. Os dados podem ser localizados e recuperados utilizando diferentes métodos de busca, de acordo com os critérios definidos pelo usuário: (i) por meio de busca simples, com palavras-chave; (ii) usando coordenadas genômicas de uma região particular de interesse através dos números de acesso, ou (iii) por meio de buscas por similaridade de sequências, usando a ferramenta BLAST. Além disso, o código fonte do ParaDB está disponível gratuitamente e pode ser usado em qualquer projeto de anotação de genomas ou análise de transcriptomas.

Essa atualização na informação é fundamental para o melhor entendimento da complexidade do genoma deste fungo, incluindo sua organização, biologia e mecanismos relacionados à interação com o hospedeiro. Em particular, pode ser usada em análises de dados de transcriptoma, proteoma e metaboloma, servindo como ferramenta adicional em *pipelines* de anotação genômica. E de forma mais específica, nota-se a relevância destas informações levantadas para a comunidade científica que trabalha buscando novas alternativas para o tratamento farmacológico da PCM, bem como de inibidores específicos de vias metabólicas do fungo e, também, para uma melhor compreensão geral dos principais mecanismos da infecção e da doença causada por este patógeno.

Além disso, pode-se extrapolar o impacto das melhorias na anotação dos genomas de *Paracoccidioides*, visto que as deficiências em dados deste tipo podem facilmente se perpetuar devido à natureza de livre acesso dos dados nos bancos públicos, e à medida que podem ser utilizados como fonte de anotação para outros organismos, como acontece nas *pipelines* de anotação do NCBI (THIBAUD-NISSEN, 2013; TATUSOVA *et al*., 2016).

Apesar do constante aumento de dados moleculares, ainda é marcante a carência de informações biológicas para os genomas sequenciados, em especial para os fungos causadores de doenças. Dessa forma, espera-se que a busca por respostas sobre a função desconhecida de genes e proteínas possa solucionar questões ainda não respondidas sobre outras proteínas em diferentes sistemas biológicos, revelando novos participantes de processos em nível molecular (DHANYALAKSHMI *et al*., 2016). Portanto, o ParaDB representa uma melhoria nos dados públicos relacionados à anotação genômica do gênero *Paracoccidioides* que, além de centralizar as informações geradas pelo processo de reanotação, pode se tornar um portal de referência para o gênero, auxiliando a comunidade acadêmica que utiliza esse modelo em suas linhas de pesquisa.

6 CONCLUSÕES E PERSPECTIVAS

Os resultados aqui descritos apresentam a reanotação funcional dos genes presentes nos genomas das espécies-referência representando os 5 grupos de *Paracoccidioides* spp, a saber, *P. brasiliensis* isolado Pb18 (grupo S1), *P. brasiliensis* isolado Pb03 (grupo PS2), *P. brasiliensis* isolado Pb300 (grupo PS4), *P. brasiliensis* isolado PbCnh (grupo PS3) e *P. lutzii* isolado Pb01 (grupo Pl). Após extensiva análise através de ferramentas computacionais e curadoria manual, obtivemos o seguinte perfil de anotação:

- 6095 genes anotados para *P. brasiliensis* Pb18, equivalente a 72,7% do genoma CDS;
- 6390 genes anotados para *P. lutzii isolado* Pb01, equivalente a 72,4% do genoma CDS;
- 6126 genes anotados para *P. brasiliensis* Pb03, equivalente a 72,7% do genoma CDS;
- 5911 genes anotados para *P. brasiliensis* Pb300, equivalente a 73,2% do genoma CDS;
- 6076 genes anotados para *P. brasiliensis* PbCnh, equivalente a 73,0% do genoma CDS.

Portanto, alcançou-se uma média de aumento de anotação para o gênero *Paracoccidioides* de entre 34,7 a 61%, quando comparados com os dados disponibilizados nos bancos de dados RefSeq e Genbank do NCBI para esses genomas até o momento. Diante disso, os dados gerados pela reanotação foram organizados em tabelas que foram utilizadas para montagem de um banco de dados relacional contendo os genomas dos fungos do gênero *Paracoccidioides*.

Assim, apresentamos neste trabalho o banco de dados de acesso livre **ParaDB** (**http://paracoccidioides.com**), que visa centralizar dados genômicos atualizados do gênero *Paracoccidioides*, gerando um fluxo de informação mais rápido e fácil entre os pesquisadores da área. Além disso, o ParaDB proporciona meios convenientes para baixar os arquivos sequências e fornece uma interface BLAST para pesquisar as sequências hospedadas. Através da interface, vários conjuntos de dados podem ser consultados simultaneamente, permitindo a recuperação de sequências correspondentes aos organismos de interesse.

Após a criação do ParaDB, há a necessidade de serem continuamente realizadas atualizações no banco de dados e, sempre que possível, a implementação de novas ferramentas para análise de dados genômicos, bem como eventuais melhorias na forma de disponibilização dos recursos. Além disso, seria muito interessante e enriquecedor a integração dos nossos dados com outros *databases*, pois a conexão com os demais bancos de dados se mostra muito eficiente na busca de informações biológicas, assim como se percebe nas ferramentas *Blast2GO* e DAVID, utilizadas aqui.

REFERÊNCIAS

AKIVA, E.; BROWN, S.; ALMONACID, D. E.; BARBER, A. E. 2ND; CUSTER, A. F.; HICKS, M. A.; HUANG, C. C.; LAUCK, F.; MASHIYAMA, S. T.; MENG, E. C.; MISCHEL, D.; MORRIS, J. H.; OJHA, S.; SCHNOES, A. M.; STRYKE, D.; YUNES, J. M.; FERRIN, T.E.; HOLLIDAY, G. L.; BABBITT, P. C. The Structure–Function Linkage *Database*. **Nucleic Acids Res**, v. 42, n. *Database* issue, p. D521, 2014.

ALEXANDER, B. D.; PERFECT, J. R. Antifungal resistance trends towards the year 2000. **Drugs**, v. 54, p. 657-678, 1997.

ALONSO, R.; SALAVERT, F.; GARCIA-GARCIA, F.; CARBONELL-CABALLERO, J.; BLEDA, M.; GARCIA-ALONSO, L.; SANCHIS-JUAN, A.; PEREZ-GIL, D.; MARIN-GARCIA, P.; SANCHEZ, R.; CUBUK, C.; HIDALGO, M. R.; AMADOZ, A.; HERNANSAIZ-BALLESTEROS, R. D.; ALEMÁN, A.; TARRAGA, J.; MONTANER, D.; MEDINA, I.; DOPAZO, J. Babelomics 5.0: functional interpretation for new generations of genomic data. **Nucleic Acids Res**, v. 43, n. W1, p. W117-W121, 2015.

ASHBURNER, M.; BALL, C. A.; BLAKE, J. A.; BOTSTEIN, D.; BUTLER, H.; CHERRY, J. M.; DAVIS, A. P.; DOLINSKI, K.; DWIGHT, S. S.; EPPIG, J. T.; HARRIS, M. A.; HILL, D. P.; ISSEL-TARVER, L.; KASARSKIS, A.; LEWIS, S.; MATESE, J. C.; RICHARDSON, J. E.; RINGWALD, M.; RUBIN, G. M.; SHERLOCK, G. Gene Ontology: tool for the unification of biology. **Nat. Genet**, v. 25, n. 1, p. 25-29, 2000.

ATTWOOD, T. K.; BRADLEY, P.; FLOWER, D. R.; GAULTON, A.; MAUDLING, N.; MITCHELL, A. L.; MOULTON, G.; NORDLE, A.; PAINE, K.; TAYLOR, P.; UDDIN, A.; ZYGOURI, C. PRINTS and its automatic supplement, prePRINTS. **Nucleic Acids Res**, v. 31, n. 1, p. 400-402, 2003.

BAGAGLI, E.; FRANCO, M.; BOSCO, S.; HEBELER-BARBOSA, F.; TRINCA, L. A.; MONTENEGRO, M. R. High frequency of *Paracoccidioides brasiliensis* infection in armadillos (*Dasypus novemcinctus*): an ecological study. **Med. Mycol**, v. 41, n. 3, p. 217-223, 2003.

BAGAGLI, E.; SANO, A.; COELHO, K. I.; ALQUATI, S.; MIYAJI, M.; DE CAMARGO, Z. P.; GOMES, G. M.; FRANCO, M.; MONTENEGRO, M. R. Isolation of *Paracoccidioides brasiliensis* from armadillos (*Dasypus noveminctus*) captured in an endemic area of paracoccidioidomycosis. **Am J Trop Med Hyg** , v. 58, n. 4, p. 505-512, 1998.

BAGAGLI, E.; THEODORO, R. C.; BOSCO, S. M.; MCEWEN, J. G. *Paracoccidioides brasiliensis*: phylogenetic and ecological aspects. **Mycopathologia**, v. 165, p. 197-207, 2008.

BALTAZAR, L. M.; WERNECK, S. M. C.; SOARES, B. M.; FERREIRA, M. V. L.; SOUZA, D. G.; PINOTTI, M.; SANTOS, D. A.; CISALPINO P. S. Melanin protects *Paracoccidioides brasiliensis* from the effects of antimicrobial photodynamic inhibition and antifungal drugs. **Antimicrob. Agents Chemother.**, v. 59, n. 7, p. 4003-4011, 2015.

BARRETO, A. A. A condição da informação. **São Paulo Perspec**, v. 16, n. 3, p. 67-74, 2002.

BELLISSIMO-RODRIGUES, F.; MACHADO, A. A.; MARTINEZ, R. Paracoccidioidomycosis epidemiological features of a 1,000-cases series from a hyperendemic area on the southeast of Brazil. **Am J Trop Med Hyg**, v. 85, n. 3, p. 546-550, 2011.

BENSON, D. A.; CAVANAUGH, M.; CLARK, K.; KARSCH-MIZRACHI, I.; LIPMAN, D. J.; OSTELL, J.; SAYERS, E. W. GenBank. **Nucleic Acids Res**, v. 41, n. *Database* issue, p. D36, 2013.

BERGERON, B. P. ***DATABASE*S.** In**:** Bioinformatics computing. Prentice Hall Professional, 2003.

BioSQL. **BioSQL.** Disponível em: http://biosql.org/. Acesso em: 20 Nov. 2017.

BOCCA, A. L.; AMARAL, A. C.; TEIXEIRA, M. M.; SATO, P. K.; SHIKANAI-YASUDA, M. A.; SOARES FELIPE, M. S. Paracoccidioidomycosis: eco-epidemiology, taxonomy and clinical and therapeutic issues. **Future Microbiol**, v. 8, n. 9, p. 1177-1191, 2013.

BOUTET, E.; LIEBERHERR, D.; TOGNOLLI, M.; SCHNEIDER, M.; BAIROCH, A. UniProtKB/Swiss-Prot: the manually annotated section of the UniProt KnowledgeBase. **Plant Bioinformatics: Methods And Protocols**, p. 89-112, 2007.

BROWN, G. D.; DENNING, D. W.; GOW, N. A.; LEVITZ, S. M.; NETEA, M. G.; WHITE, T. C. Hidden killers: human fungal infections. **Sci. Transl. Med**, v. 4, n. 165, p. 165rv13-165rv13, 2012.

BRUMMER, E.; CASTANEDA, E.; RESTREPO, A. Paracoccidioidomycosis: an update. **Clin. Microbiol**. Rev., v. 6, n. 2, p. 89-117, 1993.

BUILTWITH PTY LTD. **CMS Usage Statistics**. Disponível em: https://trends.builtwith.com/cms. Acesso em: 20 Nov. 2017.

CANONICAL LTD. **Ubuntu**. Disponível em: https://www.ubuntu.com/. Acesso em: 20 Nov. 2017.

CARBON, S.; IRELAND, A.; MUNGALL, C. J.; SHU, S.; MARSHALL, B.; LEWIS, S. AmiGO: online access to ontology and annotation data. **Bioinformatics.** v. 25, n. 2, p. 288-289, 2009. http://biopublisher.ca/index.php/cmb/article/view/1371

CARDOSO, M. A. G.; TAMBOR, J. H. M.; NOBREGA, F. G. The mitochondrial genome from the thermal dimorphic fungus *Paracoccidioides brasiliensis*. **Yeast**, v. 24, n. 7, p. 607-616, 2007.

CARNEIRO, J. L. **Introdução a banco de dados**. Salvador, 2004. Disponível em: http://www.edilms.eti.br/uploads/file/bd/apostila_conceitos_banco_dados.pdf. Acesso em: 20 Nov. 2017.

CERQUEIRA, G. C.; ARNAUD, M. B.; INGLIS, D. O.; SKRZYPEK, M. S.; BINKLEY, G.; SIMISON, M.; MIYASATO, S. R.; BINKLEY, J.; ORVIS, J.; SHAH, P.; WYMORE, F.; SHERLOCK, G.; WORTMAN, J. R. The *Aspergillus* Genome *Database*: multispecies curation and incorporation of Rna-seq data to improve structural gene annotations. **Nucleic Acids Res** v. 42, n. *Database* issue, p. D705-10, 2014.

CHAMBERLAIN, R.; SCHOMMER, J.. **Using Docker to support reproducible research**. 2014. Disponível em: https://ndownloader.figshare.com/files/1590657. Acesso em: 20 Nov. 2017.

CHERRY, J. M.; HONG, E. L.; AMUNDSEN, C.; BALAKRISHNAN, R.; BINKLEY, G.; CHAN, E. T.; CHRISTIE, K. R.; COSTANZO, M. C.; DWIGHT, S. S.; ENGEL, S.

R.; FISK, D. G.; HIRSCHMAN, J. E.; HITZ, B. C.; KARRA, K.; KRIEGER, C. J.; MIYASATO, S. R.; NASH, R. S.; PARK, J.; SKRZYPEK, M. S.; SIMISON, M.; WENG, S.; WONG, E. D. *Saccharomyces* Genome *Database*: the genomics resource of budding yeast. **Nucleic Acids Res**, v. 40, n. D1, p. D700-D705, 2011.

COLOMBO, A. L.; TOBÓN, A.; RESTREPO, A.; QUEIROZ-TELLES, F.; NUCCI, M. Epidemiology of endemic systemic fungal infections in Latin America. **Med. Mycol**, v. 49, n. 8, p. 785-798, 2011.

CONESA, A.; GÖTZ, S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. **Int J Plant Genomics**, v. 2008, p. 1-13, 2008.

CORREDOR, G. G.; PERALTA, L. A.; CASTANO, J. H.; ZULUAGA, J. S.; HENAO, B.; RESTREPO, A. The naked-tailed armadillo *Cabassous centralis* (Miller 1899): a new host to *Paracoccidioides brasiliensis*. Molecular identification of the isolate. **Med. Mycol**, v. 43, n. 3, p. 275-280, 2005.

COUTINHO, Z. F.; SILVA, D.; LAZÉRA, M.; PETRI, PETRI, V.; OLIVEIRA, R.M.; SABROZA, P.C.; WANKE, B. Paracoccidioidomycosis mortality in Brazil (1980-1995), **Cad. Saúde Pública**, v. 18, n. 5, p. 1441-1454, 2002.

DENNING, D. W.; BROMLEY, M. J. How to bolster the antifungal pipeline. **Science**, v. 347, n. 6229, p. 1414-1416, 2015.

DESJARDINS, C. A.; CHAMPION, M. D.; HOLDER, J. W.; MUSZEWSKA, A.; GOLDBERG, J.; BAILÃO, A. M.; BRIGIDO, M. MACEDO; FERREIRA, M. E. S.; GARCIA, A. M.; GRYNBERG, M.; GUJJA, S.; HEIMAN, D. I.; HENN, M. R.; KODIRA, C. D.; LEÓN-NARVÁEZ, H.; LONGO, L. V. G.; MA, L.; MALAVAZI, I.; MATSUO, A. L.; MORAIS, F. V.; PEREIRA, M.; RODRÍGUEZ-BRITO, S.; SAKTHIKUMAR, S.; SALEM-IZACC, S. M.; SYKES, S. M.; TEIXEIRA, M. M.; VALLEJO, M. C.; WALTER, M. E. M. T.; YANDAVA, C.; YOUNG, S.; ZENG, Q.; ZUCKER, J.; FELIPE, M. S.; GOLDMAN, G. H.; HAAS, B. J.; MCEWEN, J. G.; NINO-VEGA, G.; PUCCIA, R.; SAN-BLAS, G.; SOARES, C. M. A.; BIRREN, B. W.; CUOMO, C. A. Comparative genomic analysis of human fungal pathogens causing paracoccidioidomycosis. **PLoS Genet**, v. 7, n. 10, p. e1002345, 2011.

DHANYALAKSHMI, K. H.; NAIKA, M. B.; SAJEEVAN, R. S.; MATHEW, O. K.; SHAFI, K. M.; SOWDHAMINI, R.; NATARAJA, K. N. An approach to function annotation for Proteins of Unknown Function (PUFs) in the transcriptome of Indian mulberry. **PloS one**, v. 11, n. 3, p. e0151323, 2016.

DRIES BUYTAERT. **Drupal**. Disponível em: https://www.drupal.org/. Acesso em: 20 Nov. 2017.

DU PLESSIS, L.; ŠKUNCA, N.; DESSIMOZ, C. The what, where, how and why of gene ontology – a primer for bioinformaticians. **Brief. Bioinform**, v. 12, n. 6, p. 723-735, 2011.

FINN, R. D.; BATEMAN, A.; CLEMENTS, J.; COGGILL, P.; EBERHARDT, R. Y.; EDDY, S. R.; HEGER, A.; HETHERINGTON, K.; HOLM, L.; MISTRY, J.; SONNHAMMER, E. L.; TATE, J.; PUNTA, M. Pfam: the protein families *database*. **Nucleic Acids Res**, v. 42, n. *Database* issue, p. D222, 2014.

FLANAGAN, D. **JavaScript: The definitive guide: Activate your web pages**. O'Reilly Media, 2011. Disponível em: https://laptrinhx.com/topic/7355/pdf-javascript-the-definitive-guide-6th-edition. Acesso em: 20 Nov 2017.

FRANCO, M. **Sistemas de Gerenciamento de Banco de Dados**. São João da Boa Vista, Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, 2016. Disponível em: http://proedu.ifce.edu.br/handle/123456789/354. Acesso em: 20 Nov. 2017.

GIBAS, C.; JAMBECK, P. **Developing bioinformatics computer skills**. O'Reilly Media, Inc., 2001.

GMOD. **GFF**. GMOD - Generic Model Organism *Database* project. Disponível em: http://gmod.org/wiki/Main_Page. Acesso em: 20 Nov. 2017.

GOPAL, R. K.; KOSALAI, S. T.; PERUMAL, R. C.; KANDAVEL, P. K. *In silico* proteomic functional re-annotation of Escherichia coli K-12 using dynamic biological data fusion strategy. **Computat Mol Biol**, v. 4, n. 4, 2014. Disponível em: http://biopublisher.ca/index.php/cmb/article/view/1371. Acesso em: 20 Nov. 2017.

GRIGORIEV, I. V.; NIKITIN, R.; HARIDAS, S.; KUO, A.; OHM, R.; OTILLAR, R.; RILEY, R.; SALAMOV, A.; ZHAO, X.; KORZENIEWSKI, F.; SMIRNOVA, T.; NORDBERG, H.; DUBCHAK., I; SHABALOV, I. MycoCosm portal: gearing up for 1000 fungal genomes. **Nucleic Acids Res**, v. 42, n. D1, p. D699-D704, 2013.

HAFT, D. H.; SELENGUT, J. D.; WHITE, O. The TIGRFAMs *database* of protein families. **Nucleic Acids Res**, v. 31, n. 1, p. 371-373, 2003.

HAHN, R.C.; MORATO, C.Y.T; SANTOS, N. L.; FERREIRA, J. F.; HAMDAN, J. S. Disseminated paracoccidioidomycosis: correlation between clinical and in vitro resistance to ketoconazole and trimethoprim sulphamethoxazole. **Mycoses,** v. 46, n. 8, p. 342-347, 2003.

HEUSER, C. A. **Projeto de banco de dados.** 6. ed. Porto Alegre: Bookman, 2010. xii, 282 p. (Série livros didáticos informática UFRGS ; 4).

HOLDING, C.; JOHNSON, N. A.; LIGOXYGAKIS, P.; MORGAN, R. Orthology, paralogy and proposed classification for paralog subtypes. **Trends Genet**, v. 18, n. 12, 2002.

HORNBY, J. M.; JACOBITZ-KIZZIER, S.M.; McNEEL, D. J.; JENSEN, E. C.; TREVES, D. S.; NICKERSON, K. W. Inoculum Size Effect in Dimorphic Fungi: Extracellular Control of Yeast-Mycelium Dimorphism in *Ceratocystis ulmi*, **Appl Environ Microbiol**, v. 70, n. 3, p. 1356-1359, 2004.

HUANG, D. W.; SHERMAN, B. T.; LEMPICKI, Richard A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. **Nat protoc**, v. 4, n. 1, p. 44-57, 2009.

HULO, N.; BAIROCH, A,; BULLIARD, V,; CERUTTI, L,; DE CASTRO, E,; LANGENDIJK-GENEVAUX, P, S,; PAGNI, M,; SIGRIST, C, J. The PROSITE *database*. **Nucleic Acids Res**, v. 34, n. *Database* issue, p. D227, 2006.

KALL, L.; KROGH, A.; SONNHAMMER, E. L. L. Advantages of combined transmembrane topology and signal peptide prediction-the Phobius web server. **Nucleic Acids Res**, v. 35, n. Web Server issue, 1, p. W429-W432, 2007.

KERSEY, P. J.; ALLEN, J. E.; ALLOT, A.; BARBA, M.; BODDU, S.; BOLT, B. J.; CARVALHO-SILVA, D.; CHRISTENSEN, M.; DAVIS, P.; GRABMUELLER, C.; KUMAR, N.; LIU, Z.; MAUREL, T.; MOORE, B.; MCDOWALL, M. D.; MAHESWARI, U.; NAAMATI, G.; NEWMAN, V.; ONG, C. K.; PAULINI, M.; PEDRO, H.; PERRY, E.; RUSSELL, M.; SPARROW, H.; TAPANARI, E.; TAYLOR, K.; VULLO, A.; WILLIAMS, G.; ZADISSIA, A.; OLSON, A.; STEIN, J.; WEI, S.; TELLO-RUIZ, M.; WARE, D.; LUCIANI, A.; POTTER, S.; FINN, R. D.; URBAN, M.; HAMMOND-KOSACK, K. E.; BOLSER, D. M.; DE SILVA, N.; HOWE, K. L.; LANGRIDGE, N.; MASLEN, G.; STAINES, D. M.; YATES, A. Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. Nucleic Acids Res, 2017.

KLEIN, B. S.; TEBBETS, B. Dimorphism and virulence in fungi, **Curr. Opin. Microbiol.**, v. 10, p. 314-319, 2007.

KONTOYIANNIS, D, P.; LEWIS, R. E. Antifungal drug resistance of pathogenic fungi. **The Lancet**, v. 359, n. 9312, p. 1135-1144, 2002.

KOONIN, E. V.; GALPERIN, M.Y. **Genome Annotation and Analysis**. In: Sequence - Evolution - Function: Computational Approaches in Comparative Genomics**.** Boston: Kluwer Academic; 2003. Chapter 5. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK20253/. Acesso em: 20 Nov. 2017.

LACAZ, C. S.; PORTO, E.; MARTINS, J. E. C. Micologia Médica e Fungos, Morfologia e Biologia dos Fungos de Interesse Médico e Paracoccidioidomicose, *In*: **Micologia Médica**, 8ª ed., São Paulo: Sarvier, p. 1-13, 31-57, 248-297, 1991.

LEON, D.; MARKEL, S. **Sequence Analysis in a Nutshell**. O'Reilly, 2003. Disponível em: http://ommolketab.ir/aaf-lib/d5qzewcba1wb4sk6u293rv2y15u9oa.pdf. Acesso em: 20 Nov. 2017.

LETUNIC, I.; BORK, P. 20 years of the SMART protein domain annotation resource. **Nucleic Acids Res**, 2017.

LEWIS, T.; SILLITOE, I.; DAWSON, N.; LAM, S. D.; CLARKE, T.; LEE, D.; ORENGO, C.; LEES, J. Gene3D: Extensive prediction of globular domains in proteins. **Nucleic Acids Res**, 2017.

LOHSE, M.; NAGEL, A.; HERTER, T.; MAY, P.; SCHRODA, M.; ZRENNER, R.; TOHGE, T.; FERNIE, A. R.; STITT, M.; USADEL, B. Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. **Plant Cell Environ**, v. 37, n. 5, p. 1250-1258, 2014.

LUPETTI, A.; DANESI, R.; CAMPA, M.; TACCA, M. D.; KELLY, S. Molecular basis of resistance to azole antifungals. **Trends Mol Med**, v. 8, n. 2, p. 76-81, 2002.

MANNINO, M. V. **Projeto, desenvolvimento de aplicações & administração de banco de dados.** 3. ed. São Paulo: McGraw-Hill, 2008, 712 p.

MARCHLER-BAUER, A.; LU, S.; ANDERSON, J. B.; CHITSAZ, F.; DERBYSHIRE, M. K.; DEWEESE-SCOTT, C.; FONG, J. H.; GEER, L. Y.; GEER, R. C.; GONZALES, N. R.; GWADZ, M.; HURWITZ, D. I.; JACKSON, J. D.; KE, Z.; LANCZYCKI, C. J.; LU, F.; MARCHLER, G. H.; MULLOKANDOV, M.; OMELCHENKO, M. V.; ROBERTSON, C. L.; SONG, J. S.; THANKI, N.; YAMASHITA, R. A.; ZHANG, D.; ZHANG, N.; ZHENG,

C.; BRYANT, S. H. CDD: a Conserved Domain *Database* for the functional annotation of proteins. **Nucleic Acids Res**, v. 39, n. suppl_1, p. D225-D229, 2010.

MARESCA, B.; KOBAYASHI, G. S. Dimorphism in *Histoplasma capsulatum*: a Model for the Study of Cell Differentiation in Pathogenic Fungi. **Microbiol Rev**, v. 53, n. 2, p. 186-209, 1989.

MARIADB FOUNDATION. **The MariaDB Foundation – Supporting continuity and open collaboration in the MariaDB ecosystem**. Disponível em: https://mariadb.org/. Acesso em: 20 Nov. 2017.

MARQUES, S. A. Paracoccidioidomycosis. **Clin. Dermatol**, v. 30, n. 6, p. 610-615, 2012.

MARTINEZ, R. Epidemiology of paracoccidioidomycosis. **Rev. Inst. Med. Trop. São Paulo**, v. 57, p. 11-20, 2015.

MCCOUCH, S. R. Genomics and synteny. **Plant Physiol**, v. 125, n. 1, p. 152-155, 2001.

MI, H.; MURUGANUJAN, A.; THOMAS, P. D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. **Nucleic Acids Res**, v. 41, n. D1, p. D377-D386, 2012.

MOREIRA, A. P. V. Paracoccidioidomycosis: historical, etiologic agent, epidemiology, pathogenesis, clinical forms, laboratory diagnosis and antigens. **Bol. Epidemiol. Paul.**, v. 5, n. 51, p. 11-24, 2008.

MORGENSTERN, M. S. et al. UMA ANÁLISE DE DESEMPENHO UTILIZANDO O BANCO DE DADOS MARIADB. **Salão do Conhecimento**, v. 2, n. 2, 2016. Disponível em: https://www.publicacoeseventos.unijui.edu.br/index.php/salaoconhecimento/article/download/6831/5598. Acesso em: 20 Nov. 2017.

MUÑOZ, J. F.; FARRER, R. A.; DESJARDINS, C. A.; GALLO, J. E.; SYKES, S.; SAKTHIKUMAR, S.; MISAS, E.; WHISTON, E. A.; BAGAGLI, E.; SOARES, C. M.; TEIXEIRA, M. M.; TAYLOR, J. W.; CLAY, O. K.; MCEWEN, J. G.; CUOMO, C. A. Genome diversity, recombination, and virulence across the major lineages of *Paracoccidioides*. **MSphere**, v. 1, n. 5, p. e00213-16, 2016.

MUÑOZ, J. F.; GALLO, J. E.; MISAS, E.; PRIEST, M.; IMAMOVIC, A.; YOUNG, S.; ZENG, Q.; CLAY, O. K.; MCEWEN, J. G.; CUOMO, C. A. Genome update of the dimorphic human pathogenic fungi causing paracoccidioidomycosis. **PLoS Negl Trop Dis**, v. 8, n. 12, p. e3348, 2014.

MUZZEY, D.; SCHWARTZ, K.; WEISSMAN, J. S.; SHERLOCK, G. Assembly of a phased diploid *Candida albicans* genome facilitates allele-specific measurements and provides a simple model for repeat and indel structure. **Genome Biol**, v. 14, n. 9, p. R97, 2013.

NCBI. **PUBMED**. National Center for Biotechnology Information – NCBI. Disponível em: https://www.ncbi.nlm.nih.gov/pubmed/. Acesso em 20 Nov 2011.

NCBI-GENBANK. **Genbank Overview**. Disponível em: https://www.ncbi.nlm.nih.gov/genbank/. Acesso em: 20 Nov. 2017.

NCBI-REFSEQ. **RefSeq: NCBI Reference Sequence *Database***. Disponível em: https://www.ncbi.nlm.nih.gov/refseq/. Acesso em: 20 Nov. 2017.

NC-IUBMB. **Enzyme Nomenclature** - Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes by the Reactions they Catalyse. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology - NC-IUBMB. Disponível em: http://www.sbcs.qmul.ac.uk/iubmb/enzyme/. Acesso em: 20 Nov. 2017.

NGINX.ORG. **NGINX**. Disponível em: https://nginx.org/. Acesso em: 20 Nov. 2017.

NUNES, L. R.; COSTA DE OLIVEIRA, R.; LEITE, D. B.; DA SILVA, V. S.; DOS REIS MARQUES, E; DA SILVA FERREIRA, M. E.; RIBEIRO, D. C.; DE SOUZA BERNARDES, L. A.; GOLDMAN, M. H.; PUCCIA R.; TRAVASSOS, L. R.; BATISTA, W. L.; NÓBREGA, M. P.; NOBREGA, F. G.; YANG, D. Y.; DE BRAGANÇA PEREIRA, C. A.; GOLDMAN, G. H. Transcriptome analysis of *Paracoccidioides brasiliensis* cells undergoing mycelium-to-yeast transition. **Eukaryot Cell**, v. 4, n. 12, p. 2115-2128, 2005.

O'LEARY, N. A.; WRIGHT, M. W.; BRISTER, J. R.; CIUFO, S.; HADDAD, D.; MCVEIGH, R.; RAJPUT, B.; ROBBERTSE, B.; SMITH-WHITE, B.; AKO-ADJEI, D.; ASTASHYN, A.; BADRETDIN, A.; BAO, Y.; BLINKOVA, O.; BROVER, V.; CHETVERNIN, V.; CHOI, J.; COX, E.; ERMOLAEVA, O.; FARRELL, C. M.; GOLDFARB, T.; GUPTA, T.; HAFT, D.; HATCHER, E.; HLAVINA, W.; JOARDAR, V. S.; KODALI, V. K.; LI, W.; MAGLOTT, D.; MASTERSON, P.; MCGARVEY, K. M.; MURPHY, M. R.; O'NEILL, K.; PUJAR, S.; RANGWALA, S. H.; RAUSCH, D.; RIDDICK, L. D.; SCHOCH, C.; SHKEDA, A.; STORZ, S. S.; SUN, H.; THIBAUD-NISSEN, F.; TOLSTOY. I.; TULLY, R. E.; VATSAN, A. R.; WALLIN, C.; WEBB, D.; WU, W.; LANDRUM, M. J.; KIMCHI, A.; TATUSOVA, T.; DICUCCIO, M.; KITTS, P.; MURPHY, T. D.; PRUITT, K. D. Reference sequence (RefSeq) *database* at NCBI: current status, taxonomic expansion, and functional annotation. **Nucleic Acids Res**, v. 44, n. D1, p. D733-D745, 2015.

PALMEIRO, M.; CHERUBINI, K.; YURGEL, L. S. Paracoccidioidomicose: revisão da literatura. **Sci Med**, v. 15, n. 4, p. 274-278, 2005.

PEDRUZZI, I.; RIVOIRE, C.; AUCHINCLOSS, A. H.; COUDERT, E.; KELLER, G.; DE CASTRO, E.; BARATIN, D.; CUCHE, B. A.; BOUGUELERET, L.; POUX, S.; REDASCHI, N.; XENARIOS, I.; BRIDGE, A. HAMAP in 2015: updates to the protein family classification and annotation system. **Nucleic Acids Res**, v. 43, n. D1, p. D1064-D1070, 2014.

PETERSEN, T. N.; BRUNAK, S.; VON HEIJNE, G., NIELSEN, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. **Nat methods**, v. 8, n. 10, p. 785-786, 2011.

PFEIFFER, F.; OESTERHELT, D. A manual curation strategy to improve genome annotation: application to a set of haloarchael genomes. **Life**, v. 5, n. 2, p. 1427-1444, 2015.

PIOVESAN, D.; TABARO, F.; PALADIN, L.; NECCI, M.; MICETIC, I.; CAMILLONI, C.; DAVEY, N.; DOSZTÁNYI, Z.; MÉSZÁROS, B.; MONZON, A. M.; PARISI, G.; SCHAD, E.; SORMANNI, P.; TOMPA, P.; VENDRUSCOLO, M.; VRANKEN, W. F.; TOSATTO, S. C.E. MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. **Nucleic Acids Res**, 2017.

POTOK, M, T.; ELMORE T. E.; SHELDON, F. T. Dynamic data fusion using an ontology-based *software* agent system. **Proceedings of the IIIS Agent Based Computing, Orlando**, v. 7, 2003.

PRUITT, K. D.; TATUSOVA, T.; BROWN, G. R.; MAGLOTT, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. **Nucleic Acids Res**, v. 40, n. D1, p. D130-D135, 2012.

QUEVILLON, E.; SILVENTOINEN, V.; PILLAI, S.; HARTE, N.; MULDER, N.; APWEILER, R.; LOPEZ, R. InterProScan: protein domains identifier. **Nucleic Acids Res**, v. 33, n. suppl_2, p. W116-W120, 2005.

RAMOS E SILVA, M.; SARAIVA, L. do E. Paracoccidioidomycosis. **Dermatol. Clin**., v. 26, n. 2, p. 257-269, 2008.

RTCAMP SOLUTIONS PRIVATE LIMITED. **EasyEngine**. Disponível em: https://easyengine.io/. Acesso em: 20 Nov. 2017.

SANGLARD, D. Emerging threats in antifungal-resistant fungal pathogens. **Frontiers**, v. 3, n. 11, p. 1, 2016.

SANGLARD, D. Resistance of human fungal pathogens to antifungal drugs. **Curr. Opin. Microbiol.** v. 5, p. 379-385, 2002.

SANGLARD, D.; ODDS, F. C. Resistance of *Candida* species to antifungal agents: molecular mechanisms and clinical consequences. **Lancet Infect. Dis.** v. 2, p. 73-85, 2002.

SERVANT, F.; BRU, C.; CARRÈRE, S.; COURCELLE, E.; GOUZY, J.; PEYRUC, D.; KAHN, D. ProDom: automated clustering of homologous domains. **Brief. Bioinform**, v. 3, n. 3, p. 246-251, 2002.

SHIKANAI-YASUDA, M. A.; TELLES FILHO, F. Q.; MENDES, R. P.; COLOMBO, A. L.; MORETTI, M. L. Consenso em paracoccidioidomicose. **Rev. Soc. Bras. Méd. Trop**., v. 39, n. 3, p. 297-310, 2006.

SILBERSCHATZ, A.; KORTH, H. F.; SUDARSHAN, S. **Sistema de banco de dados.** 3. ed. São Paulo: Makron Books, 1999-2005 778 p.

SILLAB. **HistoBase.** Sil Lab. University of California San Francisco. Disponível em: http://histo.ucsf.edu/. Acesso em: 20 Nov. 2017.

SILLITOE, I.; LEWIS, T.; ORENGO, C. Using CATH-Gene3D to Analyze the Sequence, Structure, and Function of Proteins. **Curr Protoc Bioinformatics**, p. 1.28. 1-1.28. 21, 2015.

SILVA, L. C. **Banco de dados para web: do planejamento à implementação**. São Paulo: Érica, 2001. 242 p.

SILVA, P. F. F.; NOVAES, E., PEREIRA, M.; SOARES, C. M. A.; BORGES, C. L.; SALEM-IZACC, S. M. *In silico* characterization of hypothetical proteins from *Paracoccidioides lutzii*. **Genet Mol Res**, v. 14, n. 4, p. 17416-17425, 2015.

SINGER-VERMES, L. M.; BURGER, E.; FRANCO, M. F.; MOSCAR DI-BACCHI, M.; MENDES-GIANNINI, M. J. S.; CALICH, V. L. G. Evaluation of the pathogenicity and immunogenicity of seven *Paracoccidioides brasiliensis* isolates in susceptible inbred mice. **J Med Vet Mycol**, v. 27, n. 2, p. 71-82, 1989.

SIVASHANKARI, S.; SHANMUGHAVEL, P. Functional annotation of hypothetical proteins–A review. **Bioinformation**, v. 1, n. 8, p. 335, 2006.

SKRZYPEK, M. S.; BINKLEY, J.; BINKLEY, G.; MIYASATO, S. R.; SIMISON, M.; SHERLOCK, G. The *Candida* Genome *Database* (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. **Nucleic Acids Res**, v. 45, n. D1, p. D592-D596, 2017.

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.; STOECKERT, C. J. JR; ROOS, D. S. FungiDB: an integrated functional genomics *database* for fungi. **Nucleic Acids Res**, v. 40, n. D1, p. D675-D681, 2011.

SUEHRING, S. **MySQL bible**. John Wiley & Sons, Inc., 2002.

TATUSOVA, T.; DICUCCIO, M.; BADRETDIN, A.; CHETVERNIN, V.; NAWROCKI, E. P.; ZASLAVSKY, L.; LOMSADZE, A.; PRUITT, K. D.; BORODOVSKY, M.; OSTELL, J. NCBI prokaryotic genome annotation pipeline. **Nucleic Acids Res**, v. 44, n. 14, p. 6614-6624, 2016.

TEIXEIRA, M. M.; THEODORO, R. C.; CARVALHO, M. J.; FERNANDES, L.; PAES, H. C.; HAHN, R. C.; MENDOZA, L.; BAGAGLI, E.; SAN-BLAS, G.; FELIPE, M. S. S. Phylogenetic analysis reveals a high level of speciation in the *Paracoccidioides* genus. **Mol Phylogenet Evol**. 2009.

TEIXEIRA, M. M.; THEODORO, R. C.; NINO-VEGA, G.; BAGAGLI, E.; FELIPE, M. S. *Paracoccidioides* species complex: ecology, phylogeny, sexual reproduction, and virulence. **PLoS Pathog**, v. 10, n. 10, p. e1004397, 2014.

THE GENE ONTOLOGY. **Gene Ontology Consortium**. Disponível em: http://geneontology.org/. Acesso em: 20 Nov. 2017.

THEODORO, R. C.; TEIXEIRA, M. M.; FELIPE, M. S. S.; PADUAN, K. S; RIBOLLA, P. M.; SAN-BLAS, G.; BAGAGLI, E. Genus *Paracoccidioides*: species recognition and biogeographic aspects. **PLoS One**. 2012.

THIBAUD-NISSEN, F.; SOUVOROV, A.; MURPHY, T.; DICUCCIO, M.; KITTS, P. **Eukaryotic genome annotation pipeline**. 2013. In: The NCBI Handbook. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK169439/. Acesso em: 20 Nov. 2017.

TIPTON, K.; BOYCE, S. History of the enzyme nomenclature system. **Bioinformatics**, v. 16, n. 1, p. 34-40, 2000.

TOMPA, P. Intrinsically disordered proteins: a 10-year recap. **Trends Biochem Sci**, v. 37, n. 12, p. 509-516, 2012.

TUON, F. F.; KOENIG, F.; JACOMETTO, D.; ROCHA, J. L. Are there risk factors for acute renal failure in adult patients using deoxycholate amphotericin B? **Rev Iberoam Micol**, v. 30, n. 1, p. 21-24, 2013.

VAN DEN BERG, B. H.; MCCARTHY, F. M.; LAMONT, S. J.; BURGESS, S. C. Re-annotation is an essential step in systems biology modeling of functional genomics data. **PLoS One**, v. 5, n. 5, p. e10642, 2010.

VISBAL, G.; SAN-BLAS, G.; MURGICH, J.; FRANCO, H. *Paracoccidioides brasiliensis*, paracoccidioidomycosis, and antifungal antibiotics, **Curr. Drug Targets Infect. Disord.,** v. 5, n. 3, p. 211-226, 2005.

VITALE, R. G.; AFELTRA, J.; MEIS, J. F. G.; VERWEIJ, P. E. Activity and post antifungal effect of chlorpromazine and trifluopherazine against *Aspergillus*, *Scedosporium* and zygomycetes, M**ycoses**, v. 50, p. 270-276, 2007.

WANG, Z.; CHEN, Y.; LI, Y. A brief review of computational gene prediction methods. **Genomics Proteomics Bioinformatics**, v. 2, n. 4, p. 216-221, 2004.

WANKE, B.; AIDÊ, M. A.Paracoccidioidomycosis. **J. Bras. Pneumol.**, v. 35, n. 12, p. 1245-1249, 2009.

WHITE, T. C.; MARR, K. A.; BOWDEN, R. A. Clinical, Cellular, and Molecular Factors That Contribute to Antifungal Drug Resistance. **Clin Microbiol Rev**, v. 11, n. 2, p. 382-402, 1998.

WILSON, D.; PETHICA, R.; ZHOU, Y.; TALBOT, C.; VOGEL, C.; MADERA, M.; CHOTHIA, C.; GOUGH, J. SUPERFAMILY-sophisticated comparative genomics, data mining, visualization and phylogeny. **Nucleic Acids Res**, v. 37, n. *Database* issue, p. D380, 2009.

WORDPRESS.ORG. **About WordPress**. Disponível em: https://wordpress.org/. Acesso em: 20 Nov. 2017.

WU, C. H.; YEH, L. S.; HUANG, H.; ARMINSKI, L.; CASTRO-ALVEAR, J.; CHEN, Y.; HU, Z.; KOURTESIS, P.; LEDLEY, R. S.; SUZEK, B. E.; VINAYAKA, C. R.; ZHANG, J.; BARKER, W. C. The protein information resource. **Nucleic Acids Res**, v. 31, n. 1, p. 345-347, 2003.

YANDELL, M. D.; MAJOROS, W. H. Genomics and natural language processing. **Nat. Rev. Genet**, v. 3, n. 8, p. 601-610, 2002.

YANDELL, M.; ENCE, D. A beginner's guide to eukaryotic genome annotation. **Nat. Rev. Genet**, v. 13, n. 5, 2012.

YATES, A.; AKANNI, W.; AMODE, M. R.; BARRELL, D.; BILLIS, K.; CARVALHO-SILVA, D.; CUMMINS, C.; CLAPHAM, P.; FITZGERALD, S.; GIL, L.; GIRÓN, C. G.; GORDON, L.; HOURLIER, T.; HUNT, S. E.; JANACEK, S. H.; JOHNSON, N.; JUETTEMANN, T.; KEENAN, S.; LAVIDAS, I.; MARTIN, F. J.; MAUREL, T.; MCLAREN, W.; MURPHY, D. N.; NAG, R.; NUHN, M.; PARKER, A.; PATRICIO, M.; PIGNATELLI, M.; RAHTZ, M.; RIAT, H. S.; SHEPPARD, D.; TAYLOR, K.; THORMANN, A.; VULLO, A.; WILDER, S. P.; ZADISSA, A.; BIRNEY, E.; HARROW, J.; MUFFATO, M.; PERRY, E.; RUFFIER, M.; SPUDICH, G.; TREVANION, S. J.; CUNNINGHAM, F.; AKEN, B. L.; ZERBINO, D. R.; FLICEK, P. Ensembl 2016. **Nucleic Acids Res**, v. 44, n. D1, p. D710-D716, 2015.

ZURB. **ZURB Foundation**. Disponível em: https://foundation.zurb.com/. Acesso em: 20 Nov. 2017.