UNIVERSIDADE FEDERAL DE MINAS GERAIS

Instituto de Ciências Biológicas

Programa Interunidades de Pós-graduação em Bioinformática

Heron Oliveira Hilário

UTILIZAÇÃO DA GENÔMICA COMPARATIVA NA BUSCA POR ADAPTAÇÕES DE LEVEDURAS DO GÊNERO *METSCHNIKOWIA* À VIDA EM BAIXAS TEMPERATURAS

Belo Horizonte

2020

Heron Oliveira Hilário

UTILIZAÇÃO DA GENÔMICA COMPARATIVA NA BUSCA POR ADAPTAÇÕES DE LEVEDURAS DO GÊNERO *METSCHNIKOWIA* À VIDA EM BAIXAS TEMPERATURAS

Versão final

Tese apresentada ao Programa Interunidades de Pós-graduação em Bioinformática da Universidade Federal de Minas Gerais como requisito parcial à obtenção do título de Doutor em Bioinformática.

Orientadora: Dra. Glória Regina Franco

Coorientador: Dr. Thiago Mafra Batista

Belo Horizonte

2020

043 Hilário, Heron Oliveira.
Utilização da genômica comparativa na busca por adaptações de leveduras do gênero Metschnikowia à vida em baixas temperaturas [manuscrito] / Heron Oliveira Hilário. – 2020.
138 f. : il. ; 29,5 cm.

Orientadora: Dra. Glória Regina Franco. Coorientador: Dr. Thiago Mafra Batista.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Genômica. 3. Metschnikowia. 4. Regiões Antárticas. I. Franco, Glória Regina. II. Batista, Thiago Mafra. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Fabiane C M Reis – CRB 6 - 2680

Universidade Federal de Minas Gerais

Instituto de Ciências Biológicas

Programa Interunidades de Pós-Graduação em Bioinformática da UFMG

ATA DA DEFESA DE TESE	**119/2020**
	entrada
Heron Oliveira Hilário	**1º/2016**
	CPF: 069.766.006-04

Às nove horas do dia **25 de março de 2020**, reuniu-se, no Instituto de Ciências Biológicas da UFMG, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"Utilização da genômica comparativa na busca por adaptações de leveduras do gênero *Metschnikowia* à vida em baixas temperaturas"**, requisito para obtenção do grau de Doutor em **Bioinformática.** Abrindo a sessão, a Presidente da Comissão, **Dra. Glória Regina Franco**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Prof./Pesq.	Instituição	CPF	Indicação
Dra. Glória Regina Franco	UFMG	623387496-34	Aprovado
Dr. Thiago Mafra Batista	UFSB	060.041.026-95	Aprovado
Dr. Aristóteles Góes Neto	UFMG	544.348.825-20	Aprovado
Dr. Arthur Gruber	USP	112.681.108-41	Aprovado
Dr. João Luís Reis Cunha	UFMG	095.105.936-05	Aprovado
Dra. Priscila Grynberg	EMBRAPA/CENARGEN	013.295.566-07	Aprovado

Pelas indicações, o candidato foi considerado: Aprovado

O resultado final foi comunicado publicamente ao candidato pela Presidente da Comissão. Nada mais havendo a tratar, a Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 25 de março de 2020.

Dra. Glória Regina Franco - Orientadora Glória Regina Franco

Dr. Thiago Mafra Batista - Coorientador _____

Dr. Aristóteles Góes Neto _____

Dr. Arthur Gruber _____

Dr. João Luís Reis Cunha _____

Dra. Priscila Grynberg _____

A Maura, que vive para sempre sem nas memórias e valores que me deixou.

AGRADECIMENTOS

Agradeço à minha mãe, Maura Hollywood de Oliveira, e ao meu pai, João Augusto Hilário de Souza, por terem me criado com amor pela natureza e à luz da ciência, por terem estimulado minha curiosidade, e por terem me dado, além de amor e apoio, todas as condições para me dedicar aos estudos e à vida acadêmica. Ao meu irmão, Arthur Oliveira Hilário, eterno companheiro de descobertas e desafios, pelo apoio, estímulo e inspiração. Agradeço a toda minha família, em especial à minha Tia Cida, por sempre me tratar como um filho, e pelo suporte incondicional, especialmente nestes últimos anos.

Agradeço imensamente à Chefa, Professora Glória Franco, minha mãe científica, por todos os ensinamentos, por me guiar e me resgatar para a ciência, por acreditar em mim mesmo quando eu não acreditei, e me abrir muitas portas e caminhos para o universo do conhecimento. Agradeço também pela amizade e pelo exemplo de dedicação, profissionalismo, compromisso e cordialidade.

Agradeço ao Professor Thiaguinho Mafra, meu amigo, que me iniciou na bioinformática e tanto me ensinou sobre as ferramentas e estratégias desse universo infinito, diretamente ou me mandando RTFM. Agradeço pela paciência e pela impaciência, em doses distintas, porém necessárias.

Agradeço à Mel, minha namorada, por ser minha companheira no lazer e no estudo, por seu carinho e apoio nos momentos mais difíceis, e pelo exemplo de garra e comprometimento.

Aos meus amigos de LGB e aos Gloriosos, pela amizade e por sempre estarem disponíveis para ensinar e ajudar nos desafios da Bioinformática. Em especial, à Nayara Toledo e à Stella Soares, pelo companheirismo. Ao Lúcio Queiroz, por infinitos ensinamentos em *hardware*, *software*, abordagem científica e visualização de dados, e por aumentar minha capacidade de processamento viabilizando todos estes estudos. Ao Thomas Luscher, pelos ensinamentos no mundo das redes. Ao Agnello Picorelli, por tanto se empenhar a decifrar comigo os códigos e análises de ortólogos. Ao Dr. Bruno Carvalho, pelas ajudas na ciência molhada. À Jéssica Duarte, por me ajudar a conduzir a parte experimental e por dar continuidade a esta linha de pesquisa.

Agradeço ao Prof. Luiz Del Bem, pela amizade e pelos ensinamentos, por estimular uma visão evolutiva e holística na abordagem científica, e pelo apoio e motivação.

Agradeço ao Professor Luiz Henrique Rosa e ao Professor Carlos Augusto Rosa, pela oportunidade de trabalhar com um organismo tão interessante, e por estarem sempre disponíveis para esclarecimentos sobre a Antártica, sua diversidade microbiológica e sobre as *Metschnikowia*.

Agradeço pela ajuda e pelos ensinamentos às pessoas sem quem este trabalho não seria possível como foi realizado, Dra. Ana Raquel Santos, ao grupo MycoAntar, em especial à Dra. Mayara Ogaki, Dra. Thamar Holanda, Dra. Graciéle de Menezes, Dra. Vívian Nicolau e MsC. Bárbara Porto. Ao Dr. Rennan Garcia, pelos ensinamentos e ajudas na genômica e sequenciamentos. Ao Prof. Erich Tahara, pela ajuda com as curvas de crescimento e *spot-tests*. À Daniela de Laet, pela ajuda com as PCRs.

Agradeço aos Professores Carlos Renato, Andréia Macedo, Érich Thara e Sérgio Pena, e a todos os membros da Família LGB com quem tive a oportunidade de conviver, por construírem um ambiente saudável e colaborativo, onde o trabalho de cada um é tão importante quanto o próprio. Também pelas críticas e sugestões que tanto ajudaram no desenvolvimento deste trabalho. Agradeço também aos demais Professores, Pesquisadores, Técnicos e Secretários do ICB, pela união na construção e manutenção de um centro de excelência onde se produz e se luta pela ciência de qualidade.

Agradeço aos meus amigos da Bio, e da vida, companheiros de dia a dia, cientistas ou humanos normais: Lucas Perillo, Ivan Menezes, Beatriz Gasparini, Sabrina Lima, Matteus Carvalho, Fernando Resende, Ludmila Hufnagel, Pablo Matias, Bruno Diniz, José Luiz, Sérgio Renato, Tiê Gomes, Pedro de Filippis, Juliana Miari, Túlio Jorge, Gustavo Pucci e todos os outros que me acompanharam e acompanharão.

Obrigado!

RESUMO

O gênero *Metschnikowia* compreende espécies de leveduras ascomicetas encontradas nos mais diversos ambientes. Além das espécies florícolas, associadas principalmente a flores efêmeras da família Convolvulacea e seus polinizadores, há também espécies aquáticas, que vivem em associação com algas e eventualmente parasitam microcrustáceos. Entre as várias espécies de *Metschnikowia*, *M. bicuspidata* é a mais difundida, sendo encontrada em praticamente todos os oceanos e alguns lagos temperados. *M. australis* foi por muito tempo considerada apenas uma linhagem de *M. bicuspidata*, mas atualmente é classificada como uma nova espécie. Apesar de próximas, a taxa de fertilidade em cruzamentos entre as duas espécies é muito baixa e, além disso, *M. australis* tem sua distribuição confinada aos mares antárticos, região onde *M. bicuspidata* ainda não foi encontrada. Apesar de tanto *M. bicuspidata* quanto *M. australis* serem tolerantes ao congelamento, *M. australis* apresenta uma melhor capacidade de crescimento em baixas temperaturas, como mostramos nesse estudo. Também investigamos os genomas destas leveduras em busca dos elementos que poderiam estar associados à sobrevivência ao congelamento e a outras características para crescimento em ambientes frios. Para isso, utilizamos genomas públicos de leveduras do gênero *Metschnikowia* e também sequenciamos e montamos o genoma de *M. australis*, isolada pelo ProjetoMycoAntar. Os genomas foram submetidos a um *pipeline* de predição e anotação, e foram realizadas análises comparativas para a elucidação da relação das *Metschnikowia* marinhas e as demais representantes do gênero. Observa-se que *M. australis* e *M. bicuspidata* apresentam mais genes de tRNAs, menor conteúdo de repetições, menos CDSs preditas e menor densidade gênica que as demais leveduras do clado aquático e que a média das *Metschnikowia*. As análises filogenômicas, realizadas utilizando 1317 sequências de proteínas preditas a partir de genes ortólogos de cópia única compartilhados entre os genomas, confirmaram a proximidade entre *M. australis* e *M. bicuspidata* e a separação entre o clado aquático e o florícola. Foi também desenvolvida uma estratégia de rede para a visualização dos conjuntos de genes ortólogos compartilhados entre os genomas, que evidencia expansões parálogas exclusivas do clado aquático, compartilhadas entre *M. australis* e *M. bicuspidata*, e alguns conjuntos de parálogos exclusivos de cada um destes genomas, que podem estar associadas à tolerância ao congelamento ou à outras características dos nichos ocupados por estas leveduras. Mais de 70% destas CDSs parálogas e outras CDSs de genes cópia-única não puderam ser anotadas por não possuírem similaridade com sequências conhecidas, podendo se tratar de genes exclusivos desses organismos.

Os genes ortólogos também foram utilizados para estimar o tempo de divergência entre *M. australis* e *M. bicuspidata*, posicionando a especiação destas dentro da janela de glaciação do continente Antártico. Finalmente, algumas sequências codificadoras (CDSs) exclusivas de *M. australis*, ou parcialmente compartilhadas com *M. bicuspidata* foram submetidas a diferentes classificadores de proteínas anticongelantes para a seleção dos melhores candidatos para futuros ensaios funcionais relacionados à resistência ao frio. A expressão em baixas temperaturas desses genes candidatos a anticongelantes foi analisada por RT-PCR. Verificamos que a maioria dessas CDSs é expressa em *M. australis* na temperatura de 12˚C, mas não em *M. bicuspidata*. Análises funcionais futuras deverão comprovar o envolvimento desses genes com a capacidade de tolerância ao frio.

Palavras-chave: Leveduras, Metschnikowia, Antártica, genômica comparativa

ABSTRACT

The *Metschnikowia* genus comprises ascomycetous yeasts from diverse environments. Besides the flower-associated species there are also aquatic species, macroalgae associated and micro crustaceans' parasites. *M. bicuspidata*, the most widespread, is found on almost every ocean and many temperate lakes. Despite being freeze tolerant, *M. bicuspidata* is not present in the Antarctic Ocean. *M. australis*, a closely related species, occupies its niche on these southern waters. Also freeze tolerant, *M. australis* performs better at low temperatures, as we show in this study. We have sequenced *M. australis* genome and investigated it together with *M. bicuspidata* and other 33 publicly available *Metschnikowia* genomes searching for freeze tolerance associated elements. All genomes were predicted and annotated using the same pipeline, and 1317 common single copy orthologous genes were used to reconstruct these yeasts phylogeny. We observed that *M. australis* has a smaller genome, with less predicted coding sequences and repetitive content in comparison to *M. bicuspidata*. We have also developed a homology-based network approach to visualize and identify orthologous genes shared among genomes, which shows paralogous expansions shared by *M. australis* and *M. bicuspidata* genomes and also exclusive to each organism, which may relate to adaptions do cold environments. 247 *M. australis* exclusive CDSs were analyzed by 3 Antifreeze protein classifiers to select the 16 most prominent candidates for *in vivo* detection. We found that 15 of those are expressed at 6 and 12˚C. Most have no similarity to any known gene, and future analyses will be done to identify their influence in freeze tolerance.

Keywords: yeast, Metschnikowia, Antarctica, comparative genomics

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

µL Microlitro

AA Aminoácidos

ACT γ-actina

AFGPs *AntiFreeze GlycoProteins* – Glicoproteínas anticongelamento

AFPs AntiFreeze Proteins

AG Algoritmo Genético

CCA Corrente Circumpolar Antártica

cDNA DNA complementar

CDS Coding Sequence – Região de Codificação

D1/D2 Domínios 1 e 2 da subunidade 28S do rDNA

diAA Diaminoácidos

DNAse Desoxiribonuclease

dNTP Desoxirribonucleotídeo-fosfato

dsDNA *Double stranded DNA* – DNA fita dupla

EACF Estação Antártica Comandante Ferraz

eDNA *Environmental DNA* – DNA ambiental

EDTA Ácido Etilenodiamino Tetra-Acético

FADs *Fatty Acids Desaturase* – Dessaturases De Ácidos Graxos

GC Guanina/Citosina

gDNA DNA genômico

HMM *Hidden Markov Model* – Modelos de Markov Ocultos

HSPs Heat Shock Proteins – Proteínas de Choque Térmico

IA Ice Adhesins – Adesinas

IBPs *Ice Binding Proteins* – Proteínas de Ligação ao Gelo

ICB Instituto de Ciências Biológicas

INPs *Ice Nucleating Proteins* – Proteínas Nucleadoras de Gelo

ITS Internal Transcribed Spacer

LCA *Last Common Ancestor* – Último Ancestral Comum

LPT1 Proteína De Transferência De Lipídeo 1

LPT2 Proteína De Transferência De Lipídeo 2

MAA Milhões de Anos Atrás

mDNA DNA mensageiro

***M*EGd** *Metschnikowia* de Esporo Grande

***M*EPq** *Metschnikowia* de Esporo Pequeno

mL Mililitro

MPS Massive Parallel Sequencing

NCBI National Center for Biotechnology Information

NCBInr National Center for Biotechnology Information – non-redundant database

OD Densidade ótica

OPERANTAR Operações Antártica

PBS *Phosphate-buffered saline* – Tampão fosfato-salino

PCR Polymerase Chain Reaction – Reação de Cadeia de Polimerase

Pfam Proteins Families

PGK Fosfoglicerato Cinase

pH Potencial Hidrogeniônico (-log da concentração de H^+)

PROANTAR Programa Antártico Brasileiro

pseAA *Pseudo-amino acids* – Pseudo-aminoácidos

PSSM Position Specific Scoring Matrix – Matriz de Pontuação Posição-Específica

PUFAs *Poli Unsaturated Fatty Acids* – Ácido Graxos Poli-insaturados

rDNA DNA ribossomal

RNAse Ribonuclease

RNAseq *RNA sequencing* **–** Sequenciamento de RNA

RPB2 RNA Polimerase II

SVM *Support Vector Machine* – "Máquina de suporte de vetor"

SVMGA Support Vector Machine Genetic Algorithmic

T Tonelada

TBE Tampão Tris/Borato/EDTA

tDNA DNA transportador

TEF1α Fator de Elongação da Tradução 1-α

TEF3 Fator de Elongação da Tradução 3

TEMED Tetrametiletilenodiamina

TH Thermal Hysteresis – Histerese Térmica

TOPI Topoisomerase I

TUB2 β-tubulina II

UFMG Universidade Federal de Minas Gerais

UV Ultravioleta

YPD *Yeast extract Peptone Dextrose* – Extrato De Levedura Peptona Dextrose

SUMÁRIO

1 INTRODUÇÃO

1.1 Antártica

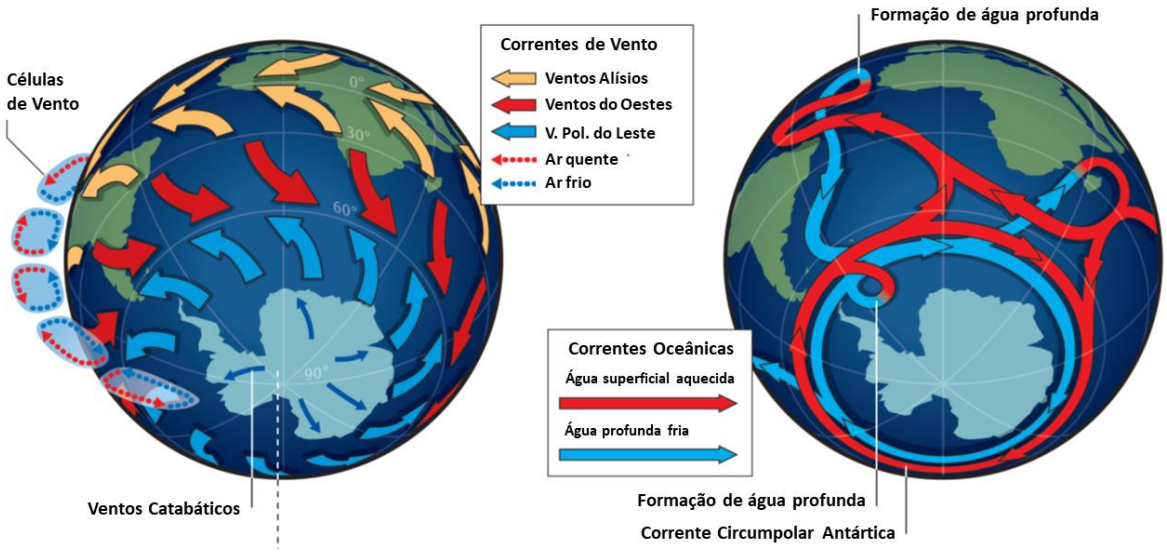
1.1.1 Um continente extremo e fundamental

As condições ambientais encontradas na Antártica são insustentáveis para maioria dos seres vivos de outras regiões do planeta. Suas baixas temperaturas aprisionam a umidade sob a forma de gelo, em uma espessa camada que recobre grande parte dos seus 14.000.000 km2 e pode atingir até 4km de profundidade (AASSP, 2011). Esta camada corresponde a 90% de todo o gelo terrestre, concentra de 60 a 70% da água doce terrestre e é altamente reflexiva à radiação solar, atuando como um atenuador de tendências de aquecimento global (Wilkins *et al.*, 2013). O continente tem influência sobre o clima do planeta, atuando diretamente no funcionamento das correntes e ecossistemas oceânicos, sendo considerado o motor da "Esteira de Distribuição Termohialina" (Thermohaline Conveyor Belt), que dispersa água com nutrientes e alto teor de oxigênio dissolvido (Barnes & Tarling, 2017) (Figura 1). Ao se aproximar do continente, a água se resfria e afunda, criando um fluxo orientado para os trópicos sobre o leito oceânico. Estas propriedades físico-químicas possibilitam altos níveis de produção primária por micro-organismos e algas que, além de corresponderem à uma majoritária parcela da teia alimentar marinha, também são responsáveis pelo sequestro de CO2 antropogênico e sua deposição na forma de sedimentos marinhos. Os mares austrais representam em torno de 30% da tomada global oceânica de CO2, apesar desse valor ser o equivalente a apenas cerca de10% em superfície total. Além disso, o continente estabiliza os níveis dos oceanos através mudanças no balanço de formação e derretimento de gelo (Barnes & Tarling, 2017).

1.1.2 A exploração da Antártica

Por ser uma das últimas fronteiras naturais e um ecossistema tão sensível e importante, a Antártica oferece um paradoxo – ao mesmo tempo que deve ser preservada de perturbações externas, precisa também ser explorada cientificamente para o conhecimento tanto de seus ambientes, recursos naturais abióticos e seus organismos endêmicos (assim como de suas capacidades), quanto para propor a melhor maneira de preservá-los.

Figura 1 — A influência da Antártica em correntes atmosféricas e oceânicas.
Movimento das correntes atmosféricas e oceânicas propelido pelo resfriamento das águas em torno do Continente Antártico.



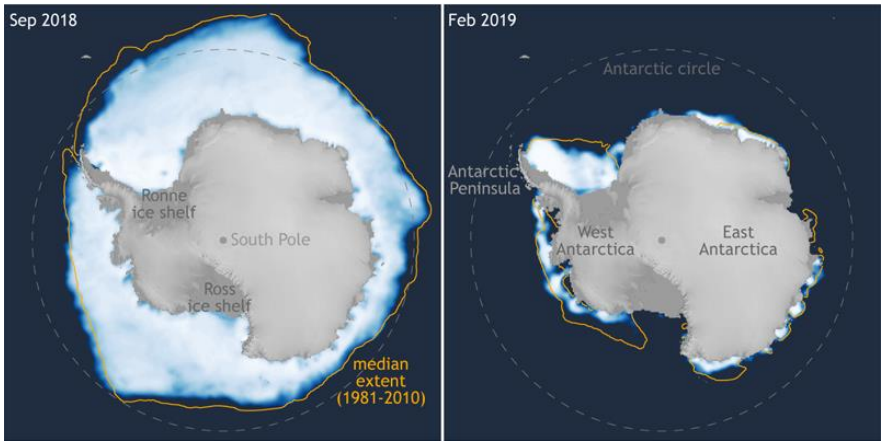
Fonte: Adaptado de Cavicchioli, 2015.

Os primeiros exploradores aportaram na Antártica em 1821 (Cowan *et al.*, 2011). Desde então, o conhecimento sobre o Continente tem crescido a cada nova expedição. Em 1959, foi assinado o Tratado da Antártica (Antarctic Treaty, 1959), determinando sua exploração para fins pacíficos e para investigação científica cooperativa. A entrada do Brasil no Tratado em 1975 abriu a oportunidade para a presença do País no Continente. O PROANTAR (Programa Antártico Brasileiro), iniciado em 1982, teve como marco a construção da Estação Antártica Comandante Ferraz (EACF), localizada na Baía do Almirantado, na Ilha Rei George, a 130km da Península Antártica. A ECAF começou a operar em fevereiro de 1984 e, desde então, possibilitou a permanência do País e a investigação do Continente pelos cientistas Brasileiros, juntamente com a frota naval e suporte indispensável da Marinha do Brasil. A estação foi destruída por um incêndio em 2012 e sua reconstrução modernizada foi inaugurada em fevereiro de 2020 (Marinha do Brasil - https://www.marinha.mil.br/secirm/proantar).

Em função de sua latitude elevada, o continente apresenta estações climáticas marcantes. Durante o inverno no Hemisfério Sul a duração dos dias é reduzida drasticamente. As regiões ao Sul do círculo polar (66˚33’47.0’’ S) não são atingidas por radiação solar direta por grande parte da estação. Como consequência desse evento, há redução na temperatura do continente, o que desencadeia a expansão da camada de gelo da calota polar sobre a água do mar em até 6.000.000

km^2 (Cavicchioli, 2015) (Figura2), recobrindo até o extremo da Península Antártica e, eventualmente, as Ilhas Shetland do Sul que, apesar do nome, constituem a região mais a norte do continente, arquipélago onde está localizada a ECAF.

Figura 2 — Variação da cobertura de gelo da Antártica por estação.
Esquerda: inverno; Direita: verão; Linha tracejada em cinza: Círculo Polar Antártico; Linha amarela: extensão média do gelo entre os anos 1981 e 2010.



Fonte: https://www.climate.gov/news-features/understanding-climate/understanding-climate-antarctic-sea-ice-extent.

Durante o verão austral, estação marcada por dias longos e temperaturas mais elevadas, a retração do gelo permite a circulação de embarcações dos diversos países que exploram o continente por razões científicas e estratégicas. As expedições brasileiras são denominadas OPERANTARs e estão atualmente na 38ª edição (Marinha do Brasil). Nelas estão envolvidos militares da Marinha do Brasil e pesquisadores brasileiros de várias áreas das ciências naturais, como Biologia, Física, Geologia, Meteorologia e Oceanografia. A investigação microbiológica de leveduras e fungos filamentosos é desenvolvida pelo *Projeto MycoAntar*, que realiza a prospecção dos ecossistemas e ambientes antárticos em busca de micro-organismos extremófilos. O projeto foi iniciado em 2005 sob coordenação do Prof. Dr. Luiz Henrique Rosa, do Departamento de Microbiologia do Instituto de Ciências Biológicas (ICB) da Universidade Federal de Minas Gerais (UFMG). Em suas expedições anuais, o *MycoAntar* tem investigado a biodiversidade associada a inúmeros tipos de amostra: neve, gelo, água e sedimento marinhos e lacustres, solo ornitogênico, rochas, líquens, plantas e algas – com foco na taxonomia, diversidade, ecologia e análise de micro-organismos extremófilos de interesse biotecnológico (Rosa, 2019).

1.1.3 A biodiversidade Antártica

Apesar dos desafios do continente, a vida persiste. A vida macroscópica terrestre é visivelmente menos diversa que em outros ambientes. São encontradas apenas duas espécies de plantas Angiospermas, *Colobanthus quitensis* e *Deschampsia antarctica*, e uma espécie, extremamente rara, de Poaceae: *Poa annua* (British Antarctic Survey, 2015). Líquens apresentam uma riqueza maior, com mais de 200 espécies, assim como musgos, com mais de 100 espécies catalogadas (Chown *et al.*, 2015). Há apenas 23 espécies de mamíferos, todas marinhas, sendo 17 cetáceos e 6 pinípedes (focas, morsas e leões-marinhos) Além destas, há 61 espécies de aves, das quais apenas 5 procriam exclusivamente no continente. Destas, o pinguim imperador, *Aptenodytes forsteri*, é único animal capaz de procriar durante o inverno (AASSP, 2011). Tanto as plantas quanto os vertebrados que frequentam o meio terrestre estão restritos a colônias esparsas, localizadas em afloramentos rochosos nas áreas costeiras. Os representantes mais difundidos da fauna são os invertebrados, como tardígrados, nematódeos, colêmbolos e ácaros (Chown *et al.*, 2015). Uma vez que os ambientes terrestres são excepcionalmente pobres em nutrientes, sua biodiversidade é sustentada pela vida marinha. Esta é muito mais distinta, apresentando mais de 40 espécies de macroalgas (Oliveira, *et al.*, 2009), e mais de 250 espécies apenas de peixes. Estão presentes também múltiplas espécies de echinodermas, brachípodas e crustáceos. Destes, destaca-se o *Krill*, principal consumidor primário das microalgas e elo importante na cadeia trófica, servindo de alimento para organismos maiores (Barnes & Clarke, 2011). Várias espécies compõe o *Krill*, das quais se destaca *Euphausia superba* em razão do seu tamanho, 65 mm, e do seu ciclo de vida que atinge até 6 anos. É considerada a espécie animal mais abundante do planeta, com biomassa estimada em 215 milhões de toneladas, sendo também amplamente explorada pela indústria pesqueira para a produção de óleo utilizado na alimentação humana e de peixes, com produção de 300.000 T em 2018 (Cavan *et al.*, 2019).

Grande parte biodiversidade da Antártica passou a ser conhecida com a investigação microbiológica. Micro-organismos são encontrados em todos os ecossistemas das Ilhas Subantárticas e também do Continente (Godinho *et al.*, 2013); nos oceanos e lagos antigos à neve recém depositada; em líquens sobre rochas; em solos ornitogênicos e *permafrost* (Duarte *et al.*, 2016); assim como também em depósitos pré-históricos de gelo e em sedimentos marinhos profundos (Vaz *et al.*, 2011). As condições climáticas extremas selecionaram organismos com

capacidades e adaptações únicas a inúmeros processos e estruturas celulares, tornando a biodiversidade Antártica um reservatório singular de organismos e genes que podem ser aproveitados pela indústria da biotecnologia (Godinho *et al.*, 2013).

A caracterização microbiológica, bioquímica e filogenética dos organismos Antárticos tem sido a principal abordagem utilizada até o momento (Wentzel *et al.*, 2018; Yajima *et al.*, 2017), seguida da investigação de genes individuais, em particular os codificadores de proteínas anticongelantes e enzimas adaptadas ao funcionamento em baixas temperaturas (Berlemont *et al.*, 2013; Hashim *et al.*, 2013; Koo *et al.*, 2016). Apesar da dificuldade de replicar os diversos ambientes e comunidades antárticas em condições de laboratório, muitos fungos foram isolados (Kochkina *et al.*, 2014; Selbmann *et al.*, 2014; Turchetti *et al.*, 2011), – com expressiva participação do *MycoAntar* e colaboradores –, nos ambientes e amostras da região da Península Antártica (de Menezes *et al.*, 2017; Duarte *et al.*, 2016; Godinho *et al.*, 2013; Godinho *et al.*, 2019; Rosa, 2019; Silva *et al.*, 2019). Alguns destes organismos têm sido selecionados para a bioprospecção dos extratos derivados de suas culturas na busca por novas moléculas candidatas a fármacos anti-patógenos e anticâncer (Godinho *et al.*, 2013).

Avanços recentes nas técnicas de análise de DNA, em especial o desenvolvimento de métodos de sequenciamento massivo em paralelo (*MPS – Massive Parallel Sequencing*) (Goodwin, McPherson & McCombie, 2016), assim como a constante evolução de *hardware* e ferramentas bioinformáticas, possibilitaram abordagens novas e complementares na investigação dos organismos de todos os ecossistemas. Estas abordagens são ainda mais importantes para a compreensão da biodiversidade quando se considera organismos extremófilos, como os antárticos (Cowan *et al.*, 2015).

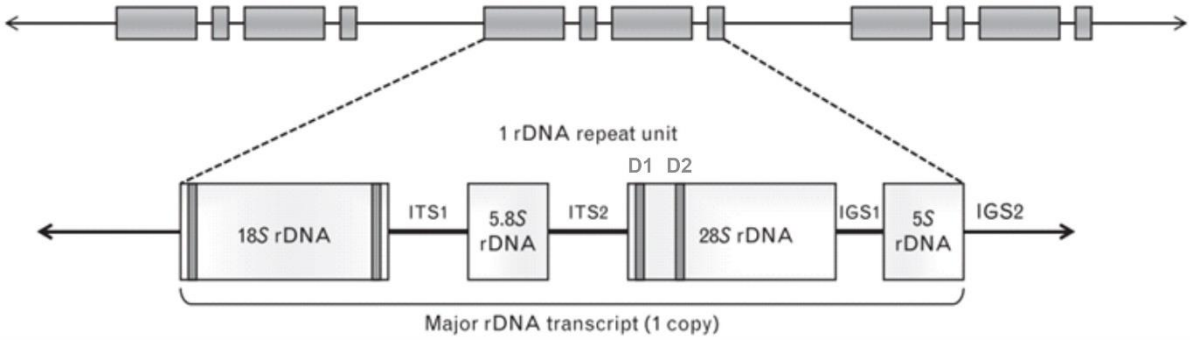
Os fungos compõem o grupo mais diverso já caracterizado nos ecossistemas antárticos (Godinho *et al.*, 2013). Mais de 1.000 espécies de fungos já foram identificadas, sendo a maior parte pertencente a grupos cosmopolitas, apesar de existirem também espécies endêmicas (Bridge & Spooner, 2012). A diversidade compartilhada provém de propágulos carreados por correntes atmosféricas e marítimas globais, e pela migração de aves, mamíferos e invertebrados (Vincent, 2000). As espécies consideradas endêmicas constituem uma reserva ainda pouco explorada, em virtude da dificuldade de acesso e, mais ainda, da complexidade de reprodução das condições ambientais antárticas em laboratório, necessárias ao cultivo de muitos organismos. Este impedimento tem sido contornado com a utilização de abordagens metagenômicas e

metaproteômicas. No entanto, a aplicabilidade dessas dois processos, atualmente, ainda se concentra no estudo da biodiversidade bacteriana (Antony *et al.*, 2016; Berlemont *et al.*, 2013; Lopatina *et al.*, 2016; Pearson *et al.*, 2015; Wilkins *et al.*, 2013) em virtude da maior capacidade de identificação destes organismos e de seus genes por técnicas de *barcoding*. Estas caracterizações se baseiam no sequenciamento de regiões de sequência hipervariável, flanqueadas por regiões conservadas. As regiões conservadas permitem a utilização de pares de *iniciadores* universais (para um dado clado), com o objetivo de amplificar sequências alvo em múltiplos organismos relacionados. A região variável, por sua vez, fornece informações para a classificação taxonômica destes organismos, sendo a região-alvo a mais utilizada na identificação de bactérias o rDNA 16S. Atualmente, as sequências mais utilizadas como padrão para a categorização de fungos compreendem as regiões ITS (*Internal Transcribed Spacer*). Esta região, comum a todos eucariotos, corresponde às sequências espaçadoras transcritas juntamente com o RNA ribossomal, na forma de um policistron (Halliday *et al.*, 2015) e suas sequências possibilitam determinar micro-organismos componentes de uma amostra complexa, através da comparação com bancos de sequências já correlacionadas. Até 2012, mais de 172.000 sequências completas de ITSs de fungos haviam sido depositadas no *GeneBank*, referentes a mais de 15.500 espécies (Schoch *et al.*, 2012). Entretanto, para algumas espécies dos gêneros *Penicillium*, *Aspergillus* e *Cladosporium*, que representam taxa abundantes, cosmopolitas, adaptados ao frio e presentes na Antártica, a região ITS é muito semelhante, sendo insuficiente para a separação destes em nível de espécie. Isto é contornado com a utilização de outras regiões, como sequências parciais dos genes da β-tubulina II (TUB2), γ-actina (ACT), fator de elongação da tradução 1-α (TEF1α), RNA polimerase II (RPB2), fator de elongação da tradução 3 (TEF3), topoisomerase I (TOPI) e fosfoglicerato cinase (PGK) (Stielow *et al.*, 2015).

Embora leveduras também sejam fungos, elas são identificadas por protocolos levemente diferentes. São inicialmente caracterizadas morfológica e fisiologicamente utilizando métodos tradicionais (Kurtzman *et al.*, 2011), seguidos do sequenciamento total ou parcial das regiões ITS, ou dos domínios D1/D2 (Rosa, 2019) – dois domínios de 500-600 nt na extremidade 5’ da Subunidade Maior (28S) do rDNA (Kurtzman & Robnett, 1998) (Figura 3). Algumas espécies, entretanto, apresentam elevado polimorfismo nessa região (Lachance *et al.*, 2003). Também é sabido que a utilização de marcadores únicos, apesar de bem estabelecidos na determinação

taxonômica, não é suficiente para reconstruções filogenéticas, tornando-as altamente enviesadas (Shen, Hittinger & Rokas, 2017).

Figura 3 — Estrutura do DNA ribossomal de eucariotos.
Na parte superior, a estrutura do rDNA, organizado em cópias múltiplas em tandem. Abaixo, uma cópia em detalhe, onde as regiões correspondentes às subunidades podem ser vistas. Bandas cinza-escuro correspondem às regiões conservadas comumente utilizadas para *barcoding*. ITS: *Internal Transcribed Spacer* (Espaçador Interno Transcrito); IGS: *InterGenic Spacer* (Espaçador Intergênico).



Fonte: Adaptado de Halliday *et al*., 2015.

As estações climáticas tão contrastantes e os demais fatores adversos presentes na Antártica selecionaram vários mecanismos nas diferentes espécies existentes (Cowan & Tow, 2004). Entretanto, as adaptações extremófilas não consistem apenas nas capacidades necessárias à vida nas baixas temperaturas do continente, apesar de este ser o desafio preponderante. A adaptabilidade dos micro-organismos também está associada à outras pressões seletivas que variam conforme o ambiente, como baixa disponibilidade de nutrientes, variação na radiação solar por longos períodos e flutuações de salinidade ocasionadas pela formação de salmouras durante o congelamento da água do mar (de Pascale *et al.*, 2012).

Nos ecossistemas costeiros, como os da Península Antártica e suas ilhas, a temperatura varia de – 35 ˚C a 5 ˚C (Tsuji, 2016). A coluna d'água permanece líquida durante o ano, apesar de recoberta por uma camada de gelo com 1 a 2 m de profundidade média (National Snow & Ice Data Center, 2019). Com a expansão desta camada de gelo, muitos organismos marinhos são expostos a condições de congelamento, em especial os fotossintetizadores, os consumidores primários e outros organismos associados, que ocupam os estratos superficiais onde a luz solar está disponível. A temperatura da água do mar no entorno do continente, a uma profundidade de até 400 m, pode variar de -1.2 a 2˚C ao longo do ano, apesar desses valores estarem em ascensão frente às mudanças climáticas, com uma elevação de 0,6˚C nas últimas duas décadas (Etourneau *et al.*, 2019).

1.2 Adaptações à vida no frio

A vida em baixas temperaturas oferece muitas restrições ao funcionamento celular, influenciando negativamente na integridade celular, na viscosidade da água e em sua disponibilidade, na difusão de solutos, na fluidez das membranas plasmática e organelar, na cinética enzimática e nas interações macromoleculares (De Maayer *et al.*, 2014). A capacidade de sobrevivência de um organismo a estas condições requer inúmeras estratégias adaptativas.

A escala de tempo de exposição ao frio é essencial para interpretar seus efeitos na fisiologia e genética microbiana. Uma exposição prolongada leva à aclimatação, processo que implica em mecanismos regulatórios para um ajuste fisiológico, sendo esta uma resposta individual de cada organismo. Exemplos de respostas de aclimatação são a modulação e a regulação da expressão gênica, através de sensores celulares e vias de transdução de sinal. Por outro lado, a seleção de alelos que aumentem o *fitness* para um dado nicho, como sobrevivência à *habitats* frios, ocorre numa escala de tempo que abrange várias gerações, e é denominada adaptação (Buzzini & Margesin, 2014). Assim sendo, a adaptação é um fenômeno populacional e de linhagem. A capacidade de aclimatação e adaptação se mesclam e possibilitam aos organismos responder a curto e a longo prazo à exposição a baixas temperaturas. Alguns organismos *mesófilos* – aqueles cujas temperaturas ótimas de crescimento são entre 15 e 40 ˚C –, são capazes de crescer a baixas temperaturas e são, por isso, denominados *psicrotolerantes*. São denominados *psicrófilos* os organismos cujas temperaturas ótimas de crescimento vão de -20 a 10 ˚C (Santiago *et al.*, 2016). Mais que resistentes às baixas temperaturas, estes organismos são irreversivelmente adaptados a ambientes frios, sendo incapazes de crescer em temperaturas mesofílicas.

1.2.1 Membrana celulares

Membranas compartimentalizam grande parte das reações bioquímicas que ocorrem na célula e são também barreiras protetoras que separam o citoplasma do meio extracelular. Sua fluidez é intrínseca ao seu funcionamento da estrutura e este sistema precisa manter seu dinamismo frente às variações ambientais. A fluidez membranar varia com a temperatura do meio e é influenciada por sua composição. As modificações mais comuns para o aumento da fluidez são o aumento no número de insaturações nos ácidos graxos e o tamanho das suas cadeias de carbono; a

redução de tamanho e carga nos grupos polares dos fosfolipídios; e a isomerização dos ácidos graxos de formas *trans* para *cis* (De Maayer *et al.*, 2014).

Quanto maior a proporção de ácido graxos poli-insaturados (*poli unsaturated fatty acids – PUFAs*), mais fluida é a membrana, de tal modo que diversas espécies de leveduras psicrófilas apresentam de 50 a 90% de PUFAs em suas membranas (Sugawara & Nikaido, 2009). Muitos estudos identificam a expressão de dessaturases de ácidos graxos (*fatty acids desaturases – FADs*) como uma resposta rápida, de até 24h, à exposição à temperaturas entre 0 e 5 ˚C (An *et al.*, 2013; Schade *et al.*, 2004). Os PUFAs são também mais abundantes em membranas de organismos *piezofílicos*, ou seja, adaptados ao crescimento em alta pressão, como bactérias do oceano profundo (Margesin, 2008).

Uma outra classe de lipídeos essenciais à estrutura da membrana celular são os esteróis, sendo o ergosterol o principal em leveduras. Os esteróis se intercalam entre as cadeias apolares dos ácidos graxos, influenciando na estabilidade, permeabilidade e fluidez da bicamada lipídica (Buzzini & Margesin, 2014) e sua proporção poder variar em até 10 vezes entre a temperatura de crescimento mínima a máxima. Apesar destas tendências, a composição da membrana depende de muitos outros fatores, como o meio de cultura, por exemplo. Mas, mesmo não parecendo haver nenhum mecanismo regulatório que seja comum a todos os micro-organismos adaptados ao frio, a diminuição na proporção de ergosterol e o aumento de PUFAs são amplamente encontrados em psicrófilos (Margesin, 2008).

1.2.2 Enzimas adaptadas ao frio

Baixas temperaturas diminuem a atividade enzimática. A influência do frio na mobilidade e flexibilidade molecular pode levar a um decaimento exponencial das taxas de catálise. Apesar disso, verifica-se que enzimas adaptadas ao frio apresentam uma atividade catalítica maior em temperaturas médias e baixas, em comparação com suas correspondentes de temperaturas moderadas (Michetti *et al.*, 2017). Diversas proteínas homólogas, principalmente bacterianas (Santiago *et al.*, 2016), oriundas de organismos adaptados a diferentes faixas de temperatura, têm sido investigadas para compreensão destas adaptações moleculares, através de expressão heteróloga e cristalização (Gerday *et al.*, 2000), sendo as alfa-amilases uma das classes de enzimas mais estudadas (EC: 3.2.1.1) (Hiteshi & Gupta, 2014). Estas pesquisas associam a atividade, em

baixas temperaturas, à frouxidão molecular, que pode ser tanto global quanto superficial (Isaksen, Åqvist, & Brandsdal, 2014), ou na região do sítio ativo, para melhor acomodação do substrato. Esta flexibilidade resulta da substituição de resíduos rígidos, como a prolina, especialmente em regiões de alças (Vester, Glaring, & Stougaard, 2015). Isto também implica menor estabilidade térmica. Proteínas adaptadas ao frio desnaturam mais facilmente. Esta característica aumenta seu potencial biotecnológico pois, além da aplicabilidade em processos industriais a baixas temperaturas e com um menor gasto energético, é possível realizar sua inativação através de choques térmicos brandos, eliminando a necessidade de inativação química (Santiago *et al.*, 2016).

1.2.3 Chaperonas

Chaperonas são proteínas ubíquas que favorecem o enovelamento adequado de proteínas novas ou desnaturadas. Presentes em praticamente todos organismos exceto vírus, foram inicialmente descobertas ao serem expressas em condições de choque térmico, e por isso são comumente denominadas HSPs (*Heat Shock Proteins*). Entretanto, estas proteínas são também expressas em resposta a outros estresses, como a irradiação UV, hipóxia, agentes químicos e choque frio. A expressão heteróloga destas últimas possibilita o crescimento de bactérias mesófilas em temperaturas baixas. Estas proteínas já têm sido utilizadas em vetores de expressão comerciais para facilitar o enovelamento na expressão a frio de proteínas termo sensíveis (Santiago *et al.*, 2016).

1.2.4 Proteínas de ligação ao gelo

O gelo é uma ameaça à integridade celular. Sua formação intracelular é letal. Ao se organizar em cristais, as moléculas de água se depositam em planos preferenciais, formando espículas cujo crescimento ocasiona a perfuração de estruturas membranares. Muitos organismos adaptados à baixas temperaturas produzem proteínas capazes de interagir com o gelo. Coletivamente chamadas de Proteínas de Ligação ao Gelo (IBPs – *Ice Binding Proteins*) (Davies, 2014), elas são capazes de induzir uma formação ordenada, interferir no crescimento dos cristais ou se adsorver aos planos cristalinos. Inúmeras IBPs foram identificadas em todos os reinos, exceto Archaea (Dolev, Braslavsky, & Davies, 2016).

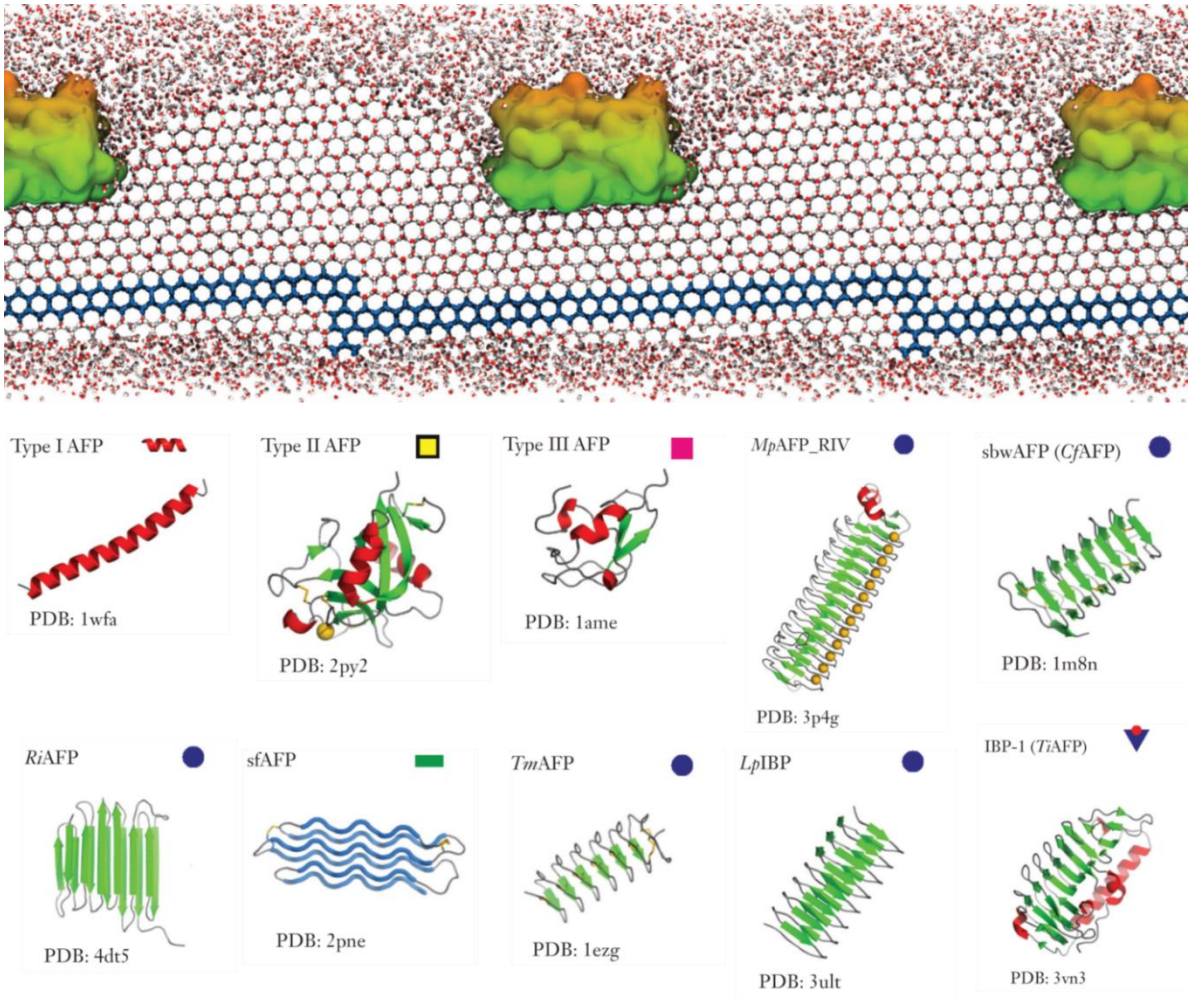
As primeiras IBPs foram descobertas no final da década de 1960 (DeVries & Wohlschlag, 1969), no sangue de peixes nototenióides antárticos e, por se tratarem de glicoproteínas, foram denominadas AFGPs (*AntiFreeze GlycoProteins*), em contraste com as demais classes de AFPs (*AntiFreeze Proteins*), que não são glicosiladas. A presença de AFPs em concentrações milimolares é suficiente para proteger do congelamento os organismos que as produzem, alcançando um efeito equivalente ao de solutos como açúcares e sais com uma eficiência até 10.000 vezes maior (Dolev *et al.*, 2016). Ao se ligarem espaçadamente aos planos do cristal de gelo em formação, causam microcurvaturas que tanto inibem a deposição de novas moléculas de água na superfície do cristal de gelo, quanto impedem a recristalização (Figura 4). A recristalização é um processo no qual cristais menores se associam para formar cristais maiores, e, portanto, mais nocivos ao organismo. Sua atividade também resulta em um fenômeno denominado Histerese Térmica *(TH – Thermal Hysteresis)*, caracterizado pelo abaixamento do ponto de congelamento da água associado a um aumento do ponto de fusão. Desta maneira, cristais já formados tendem a derreter e é mais difícil para novos cristais se desenvolverem (Celik *et al.*, 2010).

Muitas AFPs foram caracterizadas em peixes de regiões polares ou de zonas temperadas adjacentes (Fletcher, Hew, & Davies, 2001) e em artrópodes terrestres, capazes de sobreviver à temperaturas negativas do inverno boreal (Duman, 2001). As AFPs dos artrópodes têm maior TH, podendo atingir 5˚C ou mais de diferença em ensaios *in vitro* em uma concentração até 10 vezes menor que as AFPs de peixes (Scotter *et al.*, 2006). Esta hiperatividade está relacionada com a capacidade destas AFPs se ligarem a outros planos do cristal de gelo além do plano basal (Graethert *et al.*, 2000; Pertaya *et al.*, 2008).

Outras classes de IBPs são as Proteínas Nucleadoras de Gelo (*INPs – Ice Nucleating Proteins*) e as Adesinas (*IA – Ice Adhesins*). As INPs desempenham uma função inversa às demais AFPs, estimulando o congelamento. Estas são proteínas grandes, associadas à membrana, que favorecem a formação ordenada dos cristais no ambiente extracelular em temperaturas próximas ao congelamento, protegendo o citosol. Algumas bactérias patogênicas, como a *Pseudomonas syringae*, produzem INPs para facilitar a invasão de seus hospedeiros vegetais. Os cristais de gelo nucleados pela bactéria lesionam as paredes celulares das plantas formando uma porta de entrada para a infecção (Warren & Corotto, 1989). As adesinas, por sua vez, são utilizadas pela bactéria antártica marinha *Marinomonas primoryensis* para se ligar a cristais de gelo. A bactéria, aeróbia

obrigatória, garante sua posição junto às camadas superficiais da coluna d'água ao se ligar ao gelo, que é menos denso e flutua (Guo *et al.*, 2012).

Figura 4 — Proteínas anticongelantes: mecanismo de funcionamento e variedade estrutural.
(Superior) Exemplo do funcionamento: AFPs da mariposa *Choristoneura* sp. (sbwAFP) na interface do cristal de gelo (moléculas de água em organização hexagonal) com água (moléculas dispersas aleatoriamente).
(Inferior) Exemplos de AFPs de vários organismos representando sua diversidade estrutural.
AAT:AFGPs; Espirais vermelhos: AFPs-I; Quadrados amarelos: AFPs-II; Quadrados roxos: AFPs-III; Círculos azuis: AFPs-beta-solenoide; Elipses azuis: AFPs achatadas com solenoides semelhantes à proteína da seda; Triângulos azuis com círculos vermelhos: IBPs-beta-solenoide/alfa-hélice.



Fonte: Adaptado de Dolev *et al.*, 2016.

1.2.4.1 Aplicações biotecnológicas das AFPs

As possíveis aplicações biotecnológicas de AFPs são inúmeras. Na indústria alimentícia, AFPs de centeio estão sendo utilizadas em sorvetes a fim de reduzir a recristalização e alterar a

textura do alimento (Kaleda *et al.*, 2018). Há também diversas pesquisas voltadas para o congelamento como forma de armazenamento a longo prazo. Alguns estudos com a aplicação de AFGPs de peixes antárticos em carneiros horas antes do abate demonstram redução dos cristais de gelo formados na carne e da perda de líquido, que, em amostras não tratadas, decorre da ruptura das fibras musculares pelos cristais. Estudos com carne de frango imersas em solução de AFPs obtidas de brotos de rabanete mostraram que houve redução na perda de líquido pela metade (Das *et al., 2018).* Testes com vegetais, como pepino e abobrinha, demonstraram que AFPs bacterianas protegem o tecido vegetal de lesões associadas ao congelamento (Muñoz *et al.*, 2017).

O recente caso de contaminação por um anticongelante orgânico não-biológico – o dietilenoglicol, utilizado no processo de produção de cerveja para resfriamento na pré-fermentação – evidencia uma outra aplicação patente: processos industriais associados à produção alimentícia. Ressalta-se que possíveis contaminações de alimentos por proteínas anticongelantes seriam potencialmente menos nocivas, se as propriedades imunogênicas proteínas fossem avaliadas adequadamente antes do uso.

AFPs também podem ser aplicadas à criopreservação de células, tanto para propósitos científicos quanto médicos, reduzindo a lesão celular e a toxicidade causadas pelos agentes criopreservantes comumente utilizados nas técnicas atuais de criopreservação e vitrificação. AFPs podem também possibilitar a preservação de órgãos e tecidos, contribuindo imensamente para procedimentos de transplante (Taylor *et al.*, 2019). A limitação da aplicação das AFPs se dá pela dificuldade de obtenção das mesmas. Apesar disto, já há empresas (https://www.antifreezeprotein.com/) comercializando AFPs de origem natural, extraídas de larvas de besouro, para aplicações como criopreservação de células e crioablação de tumores. A produção industrial de AFPs pode ser facilmente obtida pela clonagem dos genes destas proteínas diretamente nos alimentos ou em organismos que possam ser cultivados em larga escala. A clonagem para expressão heteróloga de AFPS já é uma realidade, mas tem sido utilizada principalmente para propósitos científicos (Kim *et al.*, 2017; Porta *et al.*, 2013).

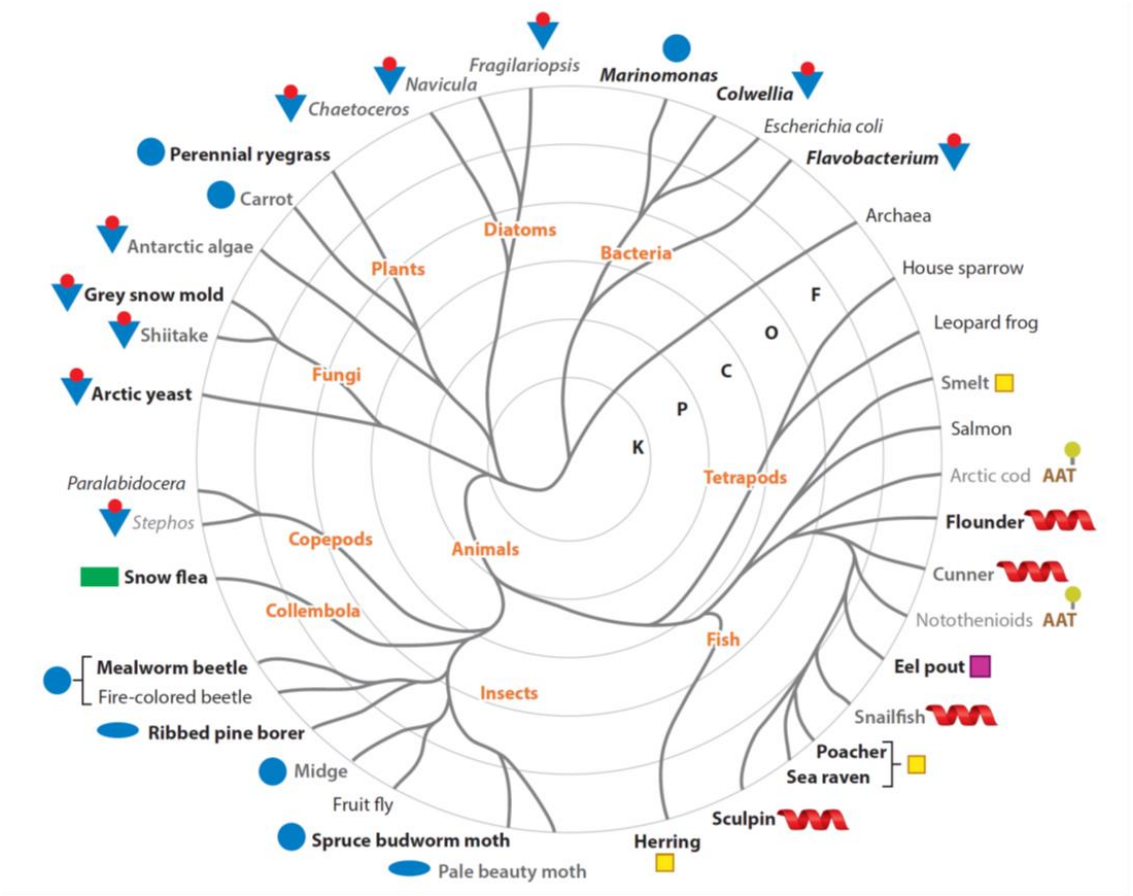
1.2.4.2 Diversidade estrutural e classificação de AFPs

Apesar de possuírem funções muito semelhantes, as IBPs são extremamente diversas estruturalmente. Isto reflete sua origem polifilética e dificulta a sumarização das características

dessa classe (Kandaswamy *et al.*, 2011) (Figura 5). Uma explicação possível para esta variedade seria o fato de o gelo possuir muitas superfícies distintas, com arranjos diferentes dos átomos de oxigênio (Davies *et al.*, 2002). Apenas as AFPs de peixes, por exemplo, são classificadas em 5 categorias diferentes, com ordenações e domínios distintos. A busca por novas IBPs e AFPs é dificultada por esta diversidade. Uma vez que estratégias baseadas em similaridade de sequência só são efetivas para a identificação de IBPs e AFPs semelhantes às já descritas, encontrá-las em táxons menos caracterizados torna-se uma tarefa difícil.

Uma alternativa é olhar diretamente para a estrutura da proteína candidata à AFP. Em 2006, Doxey e colaboradores desenvolveram um algoritmo para avaliar dados de cristalografia de Raio-X de 3 AFPs e 3196 não-AFPs não-redundantes (identidade <25%) então disponíveis no PDB, sendo estas principalmente de peixes e insetos. A partir da premissa das interações através de uma superfície relativamente plana e hidrofóbica, desenvolveram um algoritmo para avaliar coordenadas atômicas de átomos de carbono em resíduos expostos ao solvente. A predição de superfícies compatíveis em AFPs caracterizadas foi altamente concordante com os resíduos de ligação ao gelo já descritos. Esse algoritmo foi capaz de identificar a potencialidade AFP ainda não descrita em uma Proteína de Transferência de Lipídeo induzida como resposta ao frio (LPT1) no centeio, posteriormente comprovada experimentalmente. Também verificaram que em outra proteína homóloga, LPT2 (70% similaridade), a diferença de uma alanina para um ácido aspártico na superfície de interação com gelo resultava na perda de atividade.

Figura 5 — Diversidade filogenética e estrutural de proteínas de ligação ao gelo.
Árvore filogenética dos organismos que produzem IBPs. (KPCOF: Kingdom, Phylum, Class, Order, Family). Nomes em negrito: Organismos cujas IBPs foram resolvidas por ressonância magnética nuclear ou cristalografia de raios-X. AAT: AFGPs; Molas vermelhas: AFPs-I; Quadrados amarelos: AFPs-II; Quadrados roxos: AFPs-III; Círculos azuis: AFPs-beta-solenoide; Elipses azuis: AFPs achatadas com solenoides semelhantes às proteínas da seda; Triângulos azuis com círculos vermelhos: IBPs-beta-solenoide/alfa-hélice.



Fonte: Adaptado de Dolev *et al.*, 2016.

Em 2011, dois grupos apresentaram estratégias independentes para classificar, em AFPs e não-AFPs, as proteínas sem similaridade com as AFPs já conhecidas, baseando essa categorização das sequências de aminoácidos das proteínas e no uso de classificadores treinados com AFPs e não-AFPs conhecidas e caracterizadas. Yu e Lu (2011) utilizaram uma estratégia de *Support Vector Machine* (*SVM*) para identificar em *n*-peptídeos (n=1,2,3,...) características que pudessem estar associadas à atividade AFP, tais como hidrofobicidade, volume de *van der Waals*, polarizabilidade, polaridade e outras características físico-químicas. Em seguida, utilizaram um *Algoritmo Genético* (*GA*) para reduzir em até 50% o número de características utilizadas para a identificação. Seu conjunto de dados era composto por 3762 não-AFPs e 44 AFPs, com estruturas definidas por

cristalografia, disponíveis no PDB (Protein Data Bank). O **SVMGA** desenvolvido foi capaz de predizer as AFPs com uma acurácia de 99,3% e foi capaz de identificar alguns resíduos comprovadamente associados à superfície de ligação ao gelo. Entretanto, quando aplicado a outras AFPs sem dados cristalográficos, a eficiência caiu para 70% em peixes e para 20% em plantas e algas, incorrendo em *overfitting*, processo este que se caracteriza por trabalhar muito bem com dados semelhantes aos usados para treino, mas apresentar uma performance inferior quando usado com outros dados. Os autores postulam que isto poderia melhorar com o aumento de depósitos de novas estruturas de AFPs. O **SVMGA** foi disponibilizado num *webservice*, mas não se encontra mais ativo.

Kandaswamy e colaboradores (2011), por sua vez, propuseram a utilização de um classificador baseado no método de *random forest*, em que árvores de decisão são criadas partir de subamostragens de um conjunto de dados de AFPs e não-AFPs, e depois testadas com os dados restantes. O **AFP-pred** utiliza 119 características de sequências para treinamento e classificação, sendo elas relativas à frequências de 10 grupos de aminoácidos baseados nas cadeias laterais – fenil (F/W/Y), carboxila (D/E), imidazol (H), amina primária(K), guanidino (R), tiol (C), enxofre (M), amido (Q/N), hidroxila (S/T) e não-polar (A/G/I/L/V/P), assim com regiões neutras, polares e apolares, negativas e positivas, hidrofílicas e hidrofóbicas; conteúdo de hélice, fita ou *coil*; e propriedades físico-químicas. Seu conjunto de dados foi construído a partir de 221 AFPs depositadas no Pfam (este era o valor até o momento em que esta análise foi realizada), ampliado para 481 por meio do PSI-BLAST contra o NR e teve sua redundância removida com o CD-HIT. O conjunto de dados negativo, de não-APFs, continha 9493 proteínas *seed* (proteínas representativas de famílias proteicas do Pfam) não relacionadas às AFPs. De cada um desses conjuntos foram selecionadas aleatoriamente 300 proteínas para compor o conjunto de treinamento. As demais foram utilizadas como conjunto de teste. O **AFP-pred** foi capaz de classificar corretamente entre 80% (utilizando 10 características) e 83% (119 características) das 181 AFPs e 9193 não-AFPs, mostrando-se como um marco para a classificação de proteínas anticongelantes.

Em seguida, Zhao, Ma & Yin (2012) propuseram a utilização de *SVM* para avaliar quais características de sequência seriam mais informativas para classificação: composição de aminoácidos, de di-aminoacidos, de pseudo-aminoacidos (pseAA) (Chou, 2001) e de PSSM - *Matriz de Pontuação Posição-Específica* (*Position Specific Scoring Matrix*). A composição de pseAA, além da composição dos 20 aminoácidos, agrega outros **N** índices numéricos que refletem

características da homogeneidade de composição entre cada aminoácido e o aminoácido (Chou, 2009). Já a PSSM se refere a uma matriz gerada em função da composição média de cada posição do alinhamento de uma dada proteína em comparação à outras contidas em um banco de dados. Verificou-se que das quatro características de sequência, a PSSM apresentava a maior capacidade de classificar corretamente AFPs e não-AFPs. Desta maneira os autores desenvolveram o **AFP_PSSM**, utilizando *SVM* treinado com perfís de PSSM do mesmo conjunto de dados utilizado por Kandaswamy e colaboradores (2011). A acurácia atingida com este processo foi de 93%.

Vários outros classificadores de AFPs foram desenvolvidos desde então, utilizando estatégias semelhantes com pequenas variações, como o **AFP-PseAAC** (Mondal & Pai, 2014); o **AFP-Ensemble** (Yang *et al.*, 2015); o **TargetFreeze** (He *et al.*, 2015); o **iAFP-ense** (Xiao, Hui, & Liu, 2016); o **Cryoprotect** (Pratiwi *et al.*, 2017); um classificador de proteínas com estruturas conhecidas (Kozuch, Stillinger, & Debenedetti, 2018); o iAFP-gap-SMOTE (Akbar *et al.*, 2018); o **afpCOOL** (Eslami *et al.*, 2018); o **RAFP-pred** (Khan, Naseem, Togneri, & Bennamoun, 2018); e o **AFP-CKSAAP** (Usman & Lee, 2019). A maior parte destes categorizadores utiliza composições de AA, de diAA, pseudoAA, classes de AA e classificação de segmentos proteicos, se baseando em métodos diversos como *Random Forest*, Aprendizado de Máquina, Redes Neurais, *Support Vector Machine*, Algoritmos Genéticos. Entretanto, apenas o **RAFP-pred**, o **Cryoprotect** e o **iAFP-ense** estão disponibilizados como pacotes ou serviços *web*.

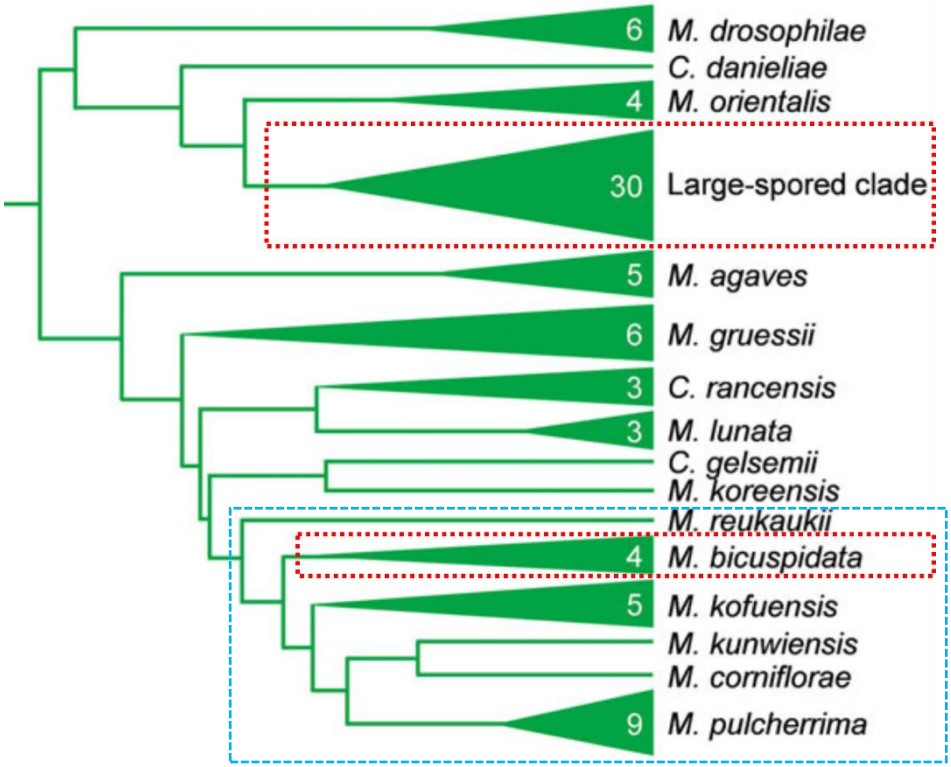
1.3 As leveduras do gênero *Metschnikowia*

Leveduras são fungos que se reproduzem assexuadamente por brotamento ou fissão, podendo ser tanto ascomicetos quanto basidiomicetos. O último levantamento sistemático realizado por Kurtzman *et al.* (2011) reconheceu 149 gêneros abrangendo 1.500 espécies catalogadas, que compreendem isolados individualizados por critérios morfológicos, bioquímicos, reprodutivos e, mais recentemente, moleculares.

A família Metschnikowiaceae é grupo irmão da família Saccharomycetaceae dentro do subfilo *Saccharomycotina* e, apesar de menos conhecido popularmente que o gênero *Saccharomyces*, *Metschnikowia* é um gênero muito comum, com representantes presentes em habitats espalhados por todo o globo (Lachance *et al.*, 2016). Kurtzman e colaboradores em 2011 reconheceram 39 espécies para o gênero. Entretanto, diversas outras novas espécies têm sido descritas (Santos *et al.*, 2020) ou reclassificadas recentemente, podendo totalizar mais de 81 espécies (Daniel *et al.*, 2014; Lachance, 2016). O fato de que há isolados de difícil manutenção em laboratório e outros de complexo cultivo ou identificação molecular fortalece a possibilidade de existirem muitas outras espécies a serem identificadas (Lachance, 2016).

Tão variado quanto ubíquo, a divergência entre os membros do clado é grande. A distância filogenética entre espécies do gênero, estimada por sequências de DNA ribossomal (rDNA), pode exceder a distância separando vários gêneros dentro de Saccharomycetaceae. O domínio D2 do rDNA é tão variável que não exibe similaridade detectável entre os subclados. A definição das "verdadeiras" *Metschnikowia* é controversa, uma vez que o gênero é estudado por grupos distintos, de vários países, que adotam perspectivas contrastantes na sua classificação (Naumov, 2012). Para facilitar a apresentação, utilizaremos definições de Lachance, que se refere às *Metschnikowia* florícolas tropicais como "*Metschnikowia* de Esporo Grande (***M*EGd**)". Analogamente, iremos nos referir ao outro clado, que compreende algumas das espécies utilizadas neste trabalho para demais comparações, como "*Metschnikowia* de Esporo Pequeno (***M*EPq**)" (Figura 6).

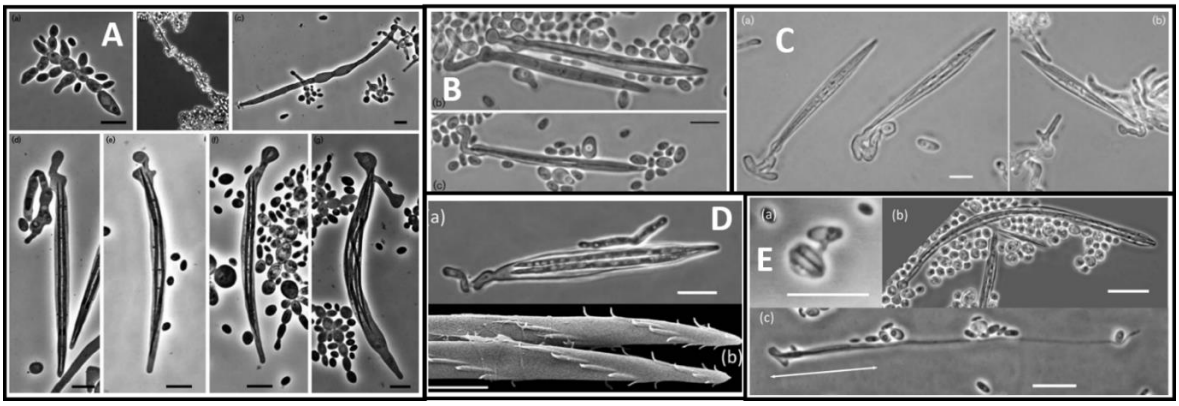
Figura 6 — Filogenia do gênero *Metschnikowia*.
Algumas das espécies utilizadas neste trabalho estão destacadas em retângulos vermelhos. O clado das *Metschnikowia* de Esporo Pequeno está destacado pelo retângulo azul. Nomes representam a espécies de referência de cada subclado. Filogenia apresentada por Lachance (2016) a partir da combinação de árvores de outros estudos, construídas tendo como base características morfológicas, bioquímicas e moleculares (D1/D2). Números representam as espécies reconhecidas para cada clado à época da publicação.



Fonte: Lachance, 2016

A primeira subdivisão, as ***M*EGd**, é composta por mais de 32 espécies terrestres de distribuição principalmente intertropical. É assim denominada pois seus ascosporos têm comprimento superior a 80 µm, sendo maiores que a forma vegetativa em até 50 vezes. Suas espécies estão principalmente associadas a flores efêmeras das Famílias Convolvulaceae e Malvaceae e a seus polinizadores, principalmente besouros do gênero *Conotelus* (Famílias Nitidulidae) e moscas (Família Drosophilidae), mas também a borboletas, abelhas, efemerópteras e outros. São haploides, heterotálicas e alógamas, ou seja, a maior parte de seu ciclo de vida é haploide, com *mating-types* complementares sendo necessários à conjugação. Esta subdivisão tem sido extensivamente caracterizada por Lachance *et al.* com representantes em todas as Américas (Lachance *et al.*, 2016) (Figura 7).

Figura 7 — Diversidade dos esporos das ***M*EGd**.
(A) *M. proteae*; (B) *M. orientalis*; (C) *M. shivogae*; (D) *M. borealis*; (E) *M. cerradonensis*



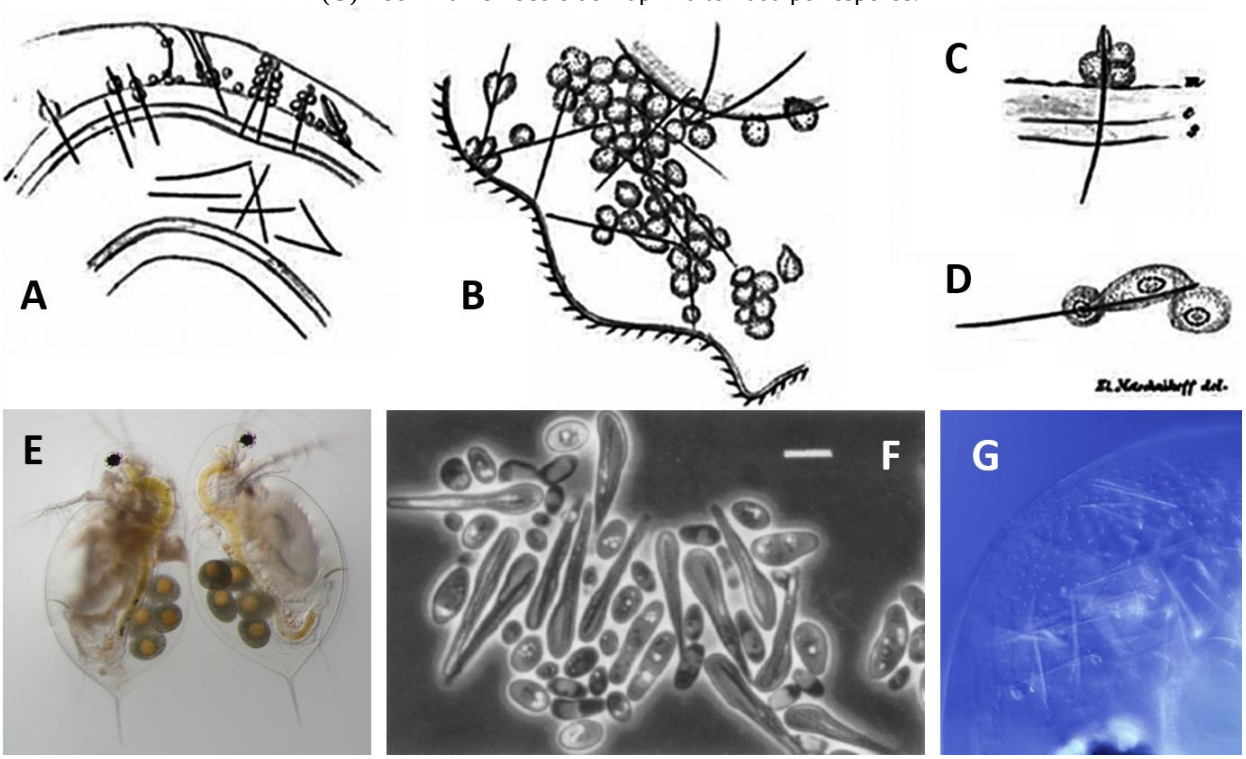
Fonte: Adaptado de Kurtzman *et al*., 2011.

A segunda subdivisão, as ***M*EPq**, apresenta mais de 40 leveduras com esporos alongados com até 60 µm de comprimento. O clado apresenta espécies tanto haploides quanto diploides, geralmente homotálicas e a formação de esporos é desencadeada por fatores ambientais (Lee *et al.*, 2018). Dentro do clado das ***M*EPq** há uma outra subdivisão entre leveduras **terrestres** e **aquáticas**. As espécies terrestres são encontradas principalmente em regiões de clima temperado como a Europa, Ásia, e América do Norte e estão associadas à seiva e flores, e aos frutos, polinizadores e diversos insetos nectarívoros. Destas, a melhor caracterizada é *M. pulcherrima*. É assim batizada em função da sua capacidade e produção de ácido pulquérrimo, molécula capaz de quelar íons de Fe^{2+}, formando a pulquerrimina. Outra representante bem estudada da divisão terrestre das ***M*EPq** é *M. reukaufii*. Também associada às flores e seus polinizadores, verifica-se que sua atividade metabólica resulta no aquecimento do botão floral, elevando sua temperatura em 6ºC graus acima do ambiente, o suficiente para facilitar a volatilização dos componentes aromáticos que atraem os polinizadores. Muitos destes compostos são também produzidos pela própria levedura (Herrera & Pozo, 2010). As ***M*EPq terrestres** são estudadas principalmente por grupos da Europa, Rússia e China.

A segunda divisão agrega as ***M*EPq aquáticas**, tipificado por *M. bicuspidata*. Esta foi a primeira espécie a ser descrita para o gênero *Metschnikowia*, em 1884, inicialmente como *Monospora bicuspidata* (Metschnikoff), sendo crucial em estudos de grande impacto na biologia celular e imunologia. Ao observar indivíduos de *Daphnia magna*, coletadas no mar da região de Odessa, Oeste da Rússia, Ellie Metschnikoff verificou que os crustáceos apresentavam seu trato

gastrointestinal parasitado pela levedura. Seus esporos rompiam a parede do intestino, ganhando acesso à hemocele, onde se proliferavam levando à morte do animal. No processo, Metschnikoff observou a existência de células tentado ativamente internalizar os esporos, e a partir deste fenômeno descreveu a fagocitose (Figura 8) (Lachance, 2016). Este trabalho levou o autor a ser laureado com o Prêmio Nobel em Fisiologia e Medicina, em 1908 (Nobel AB, 2020). Entre todo o gênero, apenas para o clado **aquático** é considerado que o formato e tamanho dos esporos tenha valor adaptativo.

Figura 8 — M. bicuspidata, a primeira Metschnikowia.
Daphnia parasitada por M. bicuspidata. (A e B) esporos perfurando a parede do epitélio intestinal.
(C e D) macrófagos da hemocele tentando fagocitar os esporos (Ilustrações originais de Metschnikoff);
(E) Daphnia parasitada por M. bicuspidata (esquerda) e saudável (direita);
(F) Esporos de M. bicuspidata após 8 dias em meio V8 (Lachance, 2011);
(G) Zoom na hemocele de Daphnia tomada por esporos.



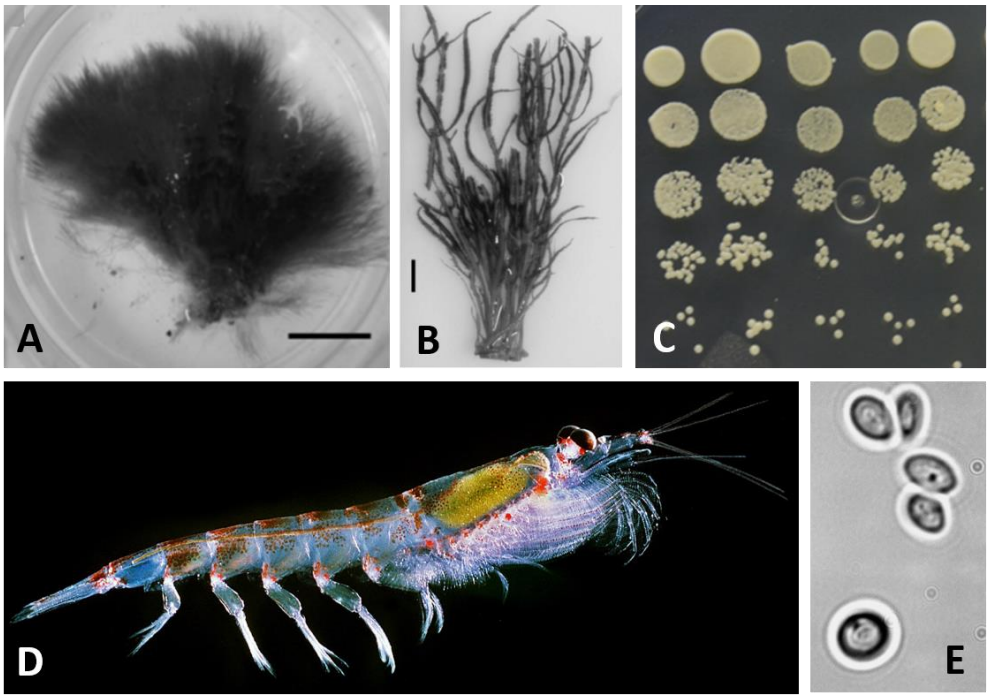
Fonte: The Duffy Lab of Aquatic Evolutionary Ecology.

A levedura *M. bicuspidata* apresenta elevada osmotolerância, suportando concentrações salinas de cerca de 10% de (Butinar *et al.*, 2005), e suas linhagens podem ser encontradas nos mais diversos ambientes aquáticos, como mares e lagos, de água doce ou hipersalinos. Sua distribuição é ampla. A levedura já foi isolada desde mares temperados, como a costa da Califórnia, até a

ambientes subglaciais, como *fjords* glaciares do arquipélago de Svalbard, na Noruega (Butinar *et al.*, 2011).

A principal levedura investigada neste estudo, *M. australis*, é endêmica dos mares no entorno do continente Antártico. É encontrada em associação com algas marinhas e em vida livre na coluna d'água. Também está relacionada com a infecção de microcrustáceos, com presença observada no conteúdo estomacal do *Krill* antártico *Euphrasia superba* (Cleary *et al.*, 2019; Donachie & Zdanowski, 1998) (Figura 9). Sua semelhança com *M. bicuspidata* fez com que ela fosse inicialmente descrita como *M. bicuspidata* var. *australis*. No entanto, em vista da baixa taxa de hibridização com DNA de *M. bicuspidata* (tipo), além de reduzida fertilidade em cruzamentos e de sua distribuição geográfica específica (Mendonça-Hagler *et al.*, 1985), ela pôde ser elevada a espécie.

Figura 9 — *M. australis* e organismos associados.
(A-B) Algas antárticas das quais foram isoladas as *M. australis* pertencentes a este trabalho.
(A) *Acrosiphonia arcta* (Dillwyn) Gain e (B) *Desmarestia menziesii*; (Godinho *et al*., 2013)
(C) Colônias de *M. australis*; (D) Krill *Euphausia superba* não infectado (Plymouth Marine Laboratory);
(E) Célula vegetativa de M. australis.



Um exemplar de *M. australis* foi isolado de algas coletadas na região costeira de Ilhas da Península Antártica pelos pesquisadores do *Mycoantar*, depositado na coleção da UFMG (UFMG-

CM-Y6158) e selecionado para sequenciamento e investigação genômica do presente estudo. A denominação *Candida* é dada a diversos isolados sem características distintivas suficientes – sem *mating-types* complementares – e para os quais esporos não podem ser obtidos, sendo impossível classificá-los por critérios morfológicos. É um agrupamento artificial, polifilético, composto de espécies que apresentam clara afinidade com ascomicetos, mas não formam ascósporos (Kurtzman *et al.*, 2011). A revisão taxonômica utilizando dados moleculares tem redesignado várias espécies descritas como *Candida* para outros gêneros. Este é o caso de *C. golubevii* e *C. torresi*, duas outras leveduras utilizadas neste trabalho atualmente classificadas como *Metschnikowia.* Apesar de *M. golubevii* ter sido descrita a partir de isolados da flores, independentemente, na Tailândia e no Pantanal (Rosa *et al.*, 2010), essa levedura apresenta elevada relação com o clado das ***M*EPq**. Essa proximidade filogenética é recapitulada tanto pela semelhança das regiões D1/D2 do rDNA quanto em análises filogenômicas (Shen *et al.*, 2018)

A levedura *M. torresii* foi isolada da água do mar no Estreito de Torres ao norte da Austrália. Esta espécie, descrita em 1978 (MycoBank), é também pouco caracterizada, assim como as outras representantes do clado das ***M*EPq aquáticas** como *M. zobellii* – descrita em 1962 (MycoBank) em amostras de água dos mares da Califórnia – e *M. krissi* – descrita em 1961 (MycoBank). A mais recente publicação do livro de referência sobre leveduras, The Yeast (Kurtzman *et al.*, 2011), traz apenas estas espécies como componentes do clado aquático, com a exceção de *M. torresii*, considerada em outros estudos posteriores como externa às três subdivisões aqui apresentadas, compondo um clado irmão ao das ***M*EPq terrestres**.

1.4 Genômica na investigação de micro-organismos

O termo *genoma* foi cunhado em 1920, muito antes do reconhecimento do DNA como molécula armazenadora da informação genética e se referia aos cromossomos e ao protoplasma (Goldman & Landweber, 2016). Com o surgimento das técnicas de sequenciamento de DNA em 1977 (Sanger, Nicklen, & Coulson, 1977) a investigação dos traços fenotípicos em nível molecular era inicialmente realizada abordando um gene ou fragmento de gene por vez. As últimas décadas testemunharam uma rápida evolução das técnicas de biologia molecular e plataformas de sequenciamento. A miniaturização aliada à paralelização massiva possibilitaram a extensão do contexto de sequenciamento de *um gene* para *todos os genes* (Heather & Chain, 2016; Quail *et al.*, 2012). A coleta de informação biológica sobre estrutura e função dos genes e o armazenamento dessa informação em grandes bancos de dados curados, associada à possibilidade de se sequenciar todo o genoma de um organismo de uma só vez, deram origem à área da *Genômica*, e também impulsionaram a *Transcriptômica e Proteômica,* além de outras ciências ômicas.

Apesar da evolução técnica das plataformas de sequenciamento de moléculas biológicas, a grande maioria utilizada hoje em dia se baseia em *reads* pequenas, de até 300nt. A reconstituição de moléculas integrais a partir destes fragmentos, sua organização e a atribuição cruzada de características e funções conhecidas por similaridade de sequência só pôde ser realizada pelo aumento da capacidade computacional e do surgimento de um ramo especializado na manipulação digital de informação biológica, a *Bioinformática* (Hogeweg, 2011).

Mais do que todos os genes de um organismo juntos, o genoma se refere a todas as moléculas de DNA presentes em uma célula ou partícula viral (*ou moléculas de RNA, no caso de vírus de RNA). É um conjunto das sequências que compreende genes codificadores de proteína e suas estruturas regulatórias, genes de RNA não-codificadores de proteína, pseudogenes, e protogenes (Dujon & Louis, 2017). O genoma também abriga sequências repetitivas e muitas outras sem função aparente (Graur, Zheng, & Azevedo, 2015), junto com regiões informativas, que carregam a história evolutiva do organismo.

O genoma é praticamente estático durante o tempo de vida de uma célula. Alguns estresses e agentes genotóxicos podem produzir mutações que, se não corrigidas pela maquinaria de reparo de DNA altamente eficiente, podem ser passadas através das gerações. Ao alterar a codificação da informação pré-existente, estas mutações podem acarretar prejuízo ao organismo portador, além

de também, eventualmente, criar novas possibilidades de genes e interações regulatórias. Apesar de ser considerada a molécula portadora da informação genética, o DNA requer proteínas acessórias e RNAs para que esta informação seja interpretada no contexto da célula ou do organismo.

A capacidade de se estudar genomas tem catalisado muitas áreas de pesquisa (Hittinger *et al.*, 2015), virtualmente todos as áreas relacionadas às ciências biológicas, e, em especial, à microbiologia (Ziemert, Alanjary & Weber, 2016). O conhecimento sobre todos os domínios da vida é beneficiado pela capacidade de se analisar qualquer organismo pelas lentes da genômica. Essa visão aprofundada tem clarificado e redefinido muitas classificações e relações taxonômicas (Choi & Kim, 2017), elucidado funções ecológicas em nível gênico (Chan *et al*., 2013) e possibilitado a investigação de organismos invisíveis às abordagens clássicas – a matéria-escura microbiana (*microbial dark matter*) –, até mesmo na Antártica (Bernard *et al.*, 2018).

Durante o desenvolvimento deste trabalho, produzimos uma revisão dos estudos ômicos voltados para fungos e leveduras do Continente Antártico e a descrição de técnicas para o sequenciamento e investigação de genomas (**ANEXO II**, CAP 15, Hilário *et al*. 2019). Este capítulo teve como intuito contextualizar pesquisadores não-familiarizados com as abordagens e fundamentações ômicas. Também foi realizada a investigação dos genomas da levedura antártica *M. austrais* e de outras do gênero *Metschnikowia*, a fim de identificar genes relacionados às capacidades que possibilitam *M. australis* lidar com baixas temperaturas.

2 OBJETIVOS

2.1 Objetivo geral

Sequenciar e investigar o genoma da levedura *M. australis* em busca dos mecanismos associados à sua capacidade de sobreviver em baixas temperaturas.

2.2 Objetivos específicos

- Avaliar o crescimento de *M. australis* e outras *Metschnikowia* em baixas temperaturas e sua sobrevivência ao congelamento;
- Sequenciar, montar e anotar o genoma da levedura *M. australis* e comparar com outros genomas de *Metschnikowia*;
- Reconstruir a filogenia dos representantes do gênero *Metschnikowia* utilizados neste trabalho;
- Identificar genes de *M. australis*, exclusivos ou compartilhados com outras *Metschnikowia*, que possam estar relacionados a adaptações à vida em baixas temperaturas, sobrevivência a condições de congelamento e endemismo na Antártica;
- Avaliar a expressão destes genes em baixas temperaturas.

3 METODOLOGIA

3.1 Obtenção dos espécimes utilizados nos experimentos

A linhagem de *M. australis* utilizada neste estudo foi isolada de algas marinhas – *Acrosiphonia arcta* (Chlorophyta) e *Desmarestia menziesii* (Ochrophyta) (Figura 9 **— *M. australis* e organismos associados.**) – coletadas na Bahia do Almirantado, na Península de Keller da Ilha Rei George, na Península Antártica, durante a expedição OPERANTAR XXVIII (2010/2011). Está depositada na Coleção de Micro-organismos e Células da Universidade Federal de Minas Gerais (UFMG-CM-Y6158). A linhagem de *M. bicuspidata* (NRRL YB-4993), por sua vez, foi cedida da coleção de leveduras da *University of West Ontario*, Canadá, pelo Prof. Marc-André Lachance; ao passo que a linhagem de *M. golubevii* (NRRL Y-48707) foi cedida da Coleção de Micro-organismos e Células da UFMG pelo Prof. Carlos Rosa. Por fim, a linhagem de *S. cerevisiae* utilizada foi a S288-C, tendo sido obtida do acervo de Células e Leveduras do Laboratório de Genética Bioquímica da UFMG.

3.2 Ensaios de crescimento e sobrevivência em baixas temperaturas

3.2.1 Ensaios de crescimento a 6 ˚C, 12 ˚C e 28 ˚C

As leveduras *M. australis*, *M. bicuspidata*, *M. golubevii* e *S. cerevisiae* foram cultivadas em meio YPD líquido (*extrato de levedura 10g/l, peptona 20g/l e dextrose 20g/l*), em Erlenmeyers com rolha de algodão possibilitando aeração, sob agitação constante a 6 ˚C, 12 ˚C e 28 ˚C (apenas *M. australis*). O inóculo utilizado foi calculado para um número de células inicial de 2×10^5 células por ml (OD = 0,05), em um volume total de 20 ml, em triplicatas, partindo de um cultivo em mesma temperatura, com OD correspondente à fase exponencial. As medições do crescimento foram realizadas por espectrofotômetro (VIS *Spectrophotometer – BEL Photonics*) utilizando cubetas de 2 ml, em média, a cada 3 horas.

Foram realizadas 3 curvas. Na primeira, a levedura *M. australis* foi cultivada a 6, 12 e 28 ˚C. Na segunda, *M. australis*, *M. bicuspidata*, *M. golubevii* e *S. cerevisiae* foram cultivadas a 12 ˚C. Na terceira, *M. australis*, *M. bicuspidata*, *M. golubevii* e *S. cerevisiae* foram cultivadas a 6 ˚C.

3.2.2 Ensaios spot-test de sobrevivência ao congelamento

Durante os ensaios de crescimento a 6 ˚C e a 12 ˚C, alíquotas de cada cultivo foram retiradas no intervalo da curva com OD ≈15 a 20 e diluídas para uma concentração equivalente à OD = 0,2. Esta diluição inicial foi utilizada para produzir dois conjuntos de diluições seriadas de 1:10, 1:100, 1:1.000 e 1:10.000, sendo um em YPD e outro em PBS 1x (*NaCl 137 mM, Fosfato 10 mM, KCl 2.7 mM, pH 7.4*).

Dez microlitros de cada uma das diluições em YPD foram aplicados ao campo correspondente de placas de Petri contendo o meio YPD sólido (*YPD e agarose 1%*), para controle. As diluições em YPD e PBS foram submetidas a congelamento (-80 ˚C) por 2h e, posteriormente, 10µl de cada diluição foram aplicados aos campos correspondentes abaixo dos *spots* controle. As placas foram mantidas a 12 ˚C até o surgimento de colônias nos campos de menor diluição. A comparação da sobrevivência foi realizada por inspeção visual do número de colônias.

3.3 Obtenção de DNA, sequenciamento e montagem do genoma de *M. australis*

M. australis foi cultivada no meio YPD sólido, a 12 ˚C por 15 dias. Uma alçada de uma colônia isolada com 0,5cm de diâmetro foi utilizada para a extração do DNA genômico da levedura, com o *kit DNAeasy Plant Mini Kit* (*Qiagen*), segundo instruções do fabricante, com acréscimo de 4 µl/ml de RNAse (6U/µl) (*Ludwig Biotech*) e 25 U de proteinase K na primeira solução.

O DNA extraído foi precipitado para limpeza em isopropanol 99%, acetato de sódio 1%, centrifugado por 5 min a 11.000 g. Em seguida, o *pellet* foi lavado com etanol 100%, centrifugado novamente por 5 min a 11.000 g. O sobrenadante foi desprezado e o *pellet* novamente lavado com etanol 70% e centrifugado a 11.000 g por 15 min. O sobrenadante foi desprezado mais uma vez e, após secagem natural, o *pellet* foi ressuspendido em 100 µl de água Milli-Q e armazenado a -20 °C até a construção das bibliotecas ou demais PCRs realizadas. A pureza da extração foi avaliada por espectrofotometria no equipamento **NanoDrop** (*Thermo Fisher Scientific*) e a integridade da banda genômica foi verificada por eletroforese em gel de agarose 1% corado com brometo de etídeo. A quantificação do DNA extraído foi feita no equipamento Qubit 2.0, utilizando o kit dsDNA HS Assay (*Thermo Fisher Scientific*).

Uma biblioteca de gDNA foi construída de acordo com as instruções do fabricante (Kit NEXTERA XT – *Illumina*). Esta biblioteca foi sequenciada no equipamento Illumina MiSeq do Laboratório Multiusuários de Genômica do Instituto de Ciências Biológicas (ICB) da Universidade Federal de Minas Gerais (UFMG) e no equipamento Illumina HiSeq 2500 do Instituto Nacional do Câncer (INCA) - Rio de Janeiro/RJ. Os sequenciamentos foram o tipo *paired-end* com *reads* de 301 nt para as bibliotecas geradas no **MiSeq** e 101 nt para bibliotecas geradas no **HiSeq**.

Os conjuntos de *reads* gerados nos sequenciadores **MiSeq** e **HiSeq** foram avaliados quanto à qualidade do sequenciamento, utilizando o *software* **FastQC** (Andrews *et al.*, 2012). *Reads* que apresentaram apenas bases com valor de qualidade igual ou superior a *phred 30* (probabilidade de um nucleotídeo ser identificado erroneamente a cada 1000 nucleotídeos sequenciados) foram consideradas de boa qualidade. As *reads* com valores de qualidade inferior a *phred* 30 e sequências referentes a adaptadores foram *trimadas* ou removidas dos conjuntos de dados utilizando o programa **TRIMOMMATIC** (Bolger, Lohse & Usadel, 2014). A montagem foi realizada utilizando o *software* **SPAdes V3.9.1** (Bankevich *et al.*, 2012). O *pipeline* e as linhas de comando e parâmetros utilizados nesta análise e em todas subsequentes se encontram disponíveis em github.com/heronoh/prj_metsh. As análises computacionais deste trabalho foram realizadas no servidor *kiko* (Laboratório de Genética Bioquímica, ICB, UFMG) e no servidor *sagarana HPC cluster* CPAD-ICB-UFMG.

3.4 Origem dos outros genomas utilizados nesse trabalho

Adicionalmente ao genoma sequenciado, trinta e um genomas de representantes do gênero *Metschnikowia* foram disponibilizados pelo Pesquisador Marc-André Lachance (*Department of Biology – University of Western Ontario – Canada*), correspondendo a isolados de 31 espécies, oriundas de diversos continentes e ambientes. Além destes, posteriormente foram incluídos quatro outros genomas disponíveis no GenBank: *M. bicuspidata* (**GCA_001664035.1**), *Metschnikowia* (*Candida*) *golubevii* (**GCA_003708755.1**), *Metschnikowia* (*Candida*) *torresii* (**GCA_002893725.1**) e *Clavispora lusitaniae* (**GCA_000003835.1**), espécie próxima utilizada como grupo externo nas análises filogenômicas de *Metschnikowia*. A partir dos genomas montados em *contigs,* os procedimentos e análises subsequentes foram realizados igualmente para todos os genomas de *Metschnikowia, Candida* e *Clavispora*. Os genomas utilizados neste estudo estão

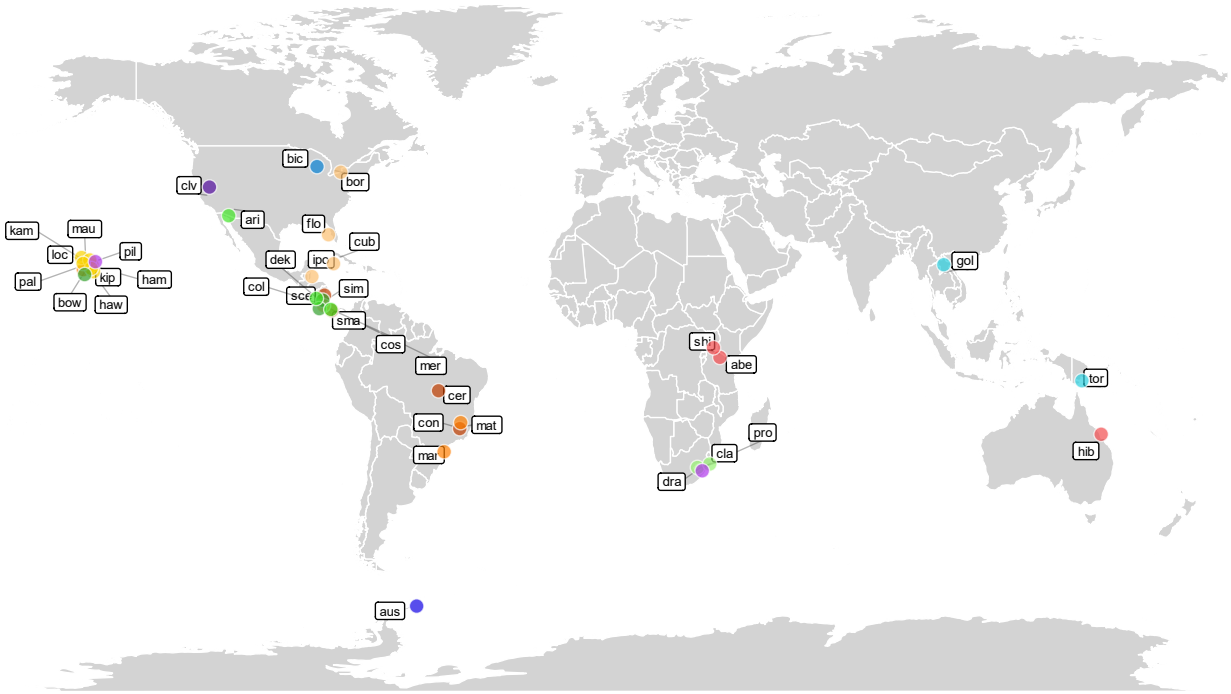
sumarizados na Tabela 1, que informa a origem, número de depósito e tombamento, assim como o número de *contigs* e a L50 (a quantidade dos maiores *contigs* que somados correspondem à metade da montagem) de cada um. A origem geográfica dos isolados de cada espécie está representada na Figura 10.

Tabela 1 — Genomas utilizados neste trabalho

Espécie	Radical	Número de *contigs*	L50	Tamanho da montagem (mb)	Completude BUSCO	NCBI genome ID	Tombamento
M. aberdeeniae	abe	70	6	10.6	95.8%	58856	UWOPS 07-202.1
M. arizonensis	ari	514	39	16.1	95.5%	58866	UWOPS 99-103.3.1
M. australis	aus	154	10	14.3	95.7%	53870	UFMG-CM-Y6158
M. bicuspidata	bic	48	3	16.0	95.1%	18263	NRRL YB-4993
M. borealis	bor	2714	95	20.5	94.7%	58871	UWOPS 96-101.1
M. bowlesiae	bow	2056	43	17.2	95.8%	58849	UWOPS 04-243x5
M. cerradonensis	cer	2263	75	20.6	96.2%	58857	UFMG 03-T67.1
M. sp. M2Y3	cla	200	4	11.5	96.8%	58855	EBD-CdV M2Y3
Clavispora lusitaniae	clv	9	3	12.1	94.8%	286	ATCC 42720
M. colocasiae	col	1206	17	14.9	96.3%	58844	UWOPS 03-202.1
M. continentalis	con	802	75	21.8	96.3%	58867	UWOPS 96-173
M. sp. 03-147.1	cos	1642	25	14.7	96.3%	58864	UWOPS 03-147.1
M. cubensis	cub	580	56	20.3	95.6%	58846	MUCL 45753
M. dekortorum	dek	1037	20	16.5	95.9%	58861	UWOPS 03-172.2
M. drakensbergensis	dra	564	10	11.8	95.9%	58853	EBD-CdVSA10-2A
M. sp. 13-106.1	flo	618	71	20.3	94.9%	58964	UWOPS 13-106.1
M. (Candida) golubevii	gol	882	30	14.7	95.7%	73597	NRRL Y-48707
M. hamakuensis	ham	2097	61	18.7	95.9%	58865	UWOPS 04-199.1
M. hawaiiensis	haw	1714	47	18.3	95.8%	58851	UWOPS 87-2203.2
M. hibisci	hib	86	4	11.3	96.1%	58852	UWOPS 95-797.2
M. ipomoeae	ipo	1014	52	18.9	96.1%	58843	UWOPS 01-141c3
M. kamakouana	kam	235	10	15.6	96.5%	58850	UWOPS 04-112.5
M. kipukae	kip	158	5	11.2	95.9%	58841	UWOPS 00-669.2
M. lochheadii	loc	2140	64	19.7	96.2%	58879	UWOPS 99-661.1
M. matae var. maris	mar	2377	47	21.1	95.9%	-	UFMG-CM-Y397
M. matae var. matae	mat	2664	69	21.3	95.6%	58859	UFMG-CM-Y395
M. mauinuiana	mau	875	26	17.2	96.3%	58862	UWOPS 04-110.4
Metschnikowia sp. 00-154.1	mer	2530	100	20.7	96.3%	58858	UWOPS 00-154.1
Metschnikowia sp. 04-218.3	pal	458	23	17.5	95.6%	58963	UWOPS 04-218.3
Metschnikowia sp. 04-226.1	pil	972	13	15.9	96.2%	58845	UWOPS 04-226.1
M. proteae	pro	731	10	12.4	96.4%	58854	EBD-T1Y1
M. santaceciliae	sce	2012	76	20.5	90.6%	58860	UWOPS 01-142b1
M. shivogae	shi	368	9	10.7	96.0%	58847	UWOPS 07-203.2
M. similis	sim	1440	49	17.2	95.9%	58863	UWOPS 03-133.4
Metschnikowia sp. 01-655c1	sma	2067	37	20.6	95.5%	58848	UWOPS 01-655c1
M. (Candida) torresii	tor	133	4	10.9	96.0%	66778	CBS 5152

Fonte: Elaborado pelo autor.

Figura 10 — Origem geográfica das leveduras utilizadas neste trabalho.
Radicais específicos de acordo com a **Tabela 1**



3.5 Avaliação da qualidade dos genomas

3.5.1 Métricas numéricas

As métricas numéricas de cada montagem foram estimadas com o *script scaffold_stats.pl* do *pipeline* de anotação desenvolvido por Kumar, S. (assemblage). Foram avaliadas características como número de *contigs* maiores que 500, tamanho da montagem, N50, L50 e conteúdo GC. Para *M. australis*, os resultados foram publicados em Batista *et al.*, 2017 (**Anexo I**), e o genoma foi depositado no NCBI DDBJ/ENA/GenBank (Genome ID: 53870, accession: **GCA_002073855.1**).

3.5.2 Métricas qualitativas

As métricas qualitativas foram avaliadas utilizando o programa **BUSCO V2.0** (Simão *et al*, 2015) e o programa **CEGMA** (Parra, Bradnam, & Korf, 2007). Ambos programas buscam por conjuntos definidos de genes ortólogos esperados para um clado (genes comuns a outros

representantes conhecidos), onde a ausência de alguns pode ser indicativa da incompletude da montagem.

3.5.2.1 CEGMA

O programa **CEGMA**, apesar de descontinuado em 2015, ainda pôde ser utilizado para estimar a completude das montagens, buscando por 248 genes ortólogos esperados para um genoma de levedura. Além disso, seu *output* no formato *gff* foi utilizado na criação de referências intrínsecas de cada genoma para a subsequente predição pelo **MAKER2** (Holt & Yandell, 2011). A partir das coordenadas do *gff* e das sequências *fasta* dos *contigs* foram construídos HMMs de perfil (*Hidden Markov Model*) que foram utilizados no treinamento do programa **SNAP** (Korf, 2004), um dos preditores gênicos *ab initio* utilizados pelo **MAKER2**. Para preparar os HMMs de perfil para o **SNAP** foram utilizados os *scripts cegma2zff*, *fathom*, *forge* e *hmm-assembler.pl* do pacote **SNAP**.

3.5.2.2 BUSCO

O programa **BUSCO 2.0** (*Benchmarking Universal Single-Copy Orthologs*) avalia a completude de um conjunto de dados (genoma, transcriptoma ou proteoma) a partir da busca por genes ortólogos esperados para linhagens representativas do clado, de maneira semelhante ao **CEGMA**. Foi utilizado um banco de referência contendo ortólogos esperados em representantes da Classe Saccharomyceta (identificador: *saccharomyceta_odb9*, com 1.759 ortólogos), na qual está inserida a Família Metschnikowiaceae. Após identificados, os ortólogos foram categorizados pelo programa em cópia única, múltiplas cópias, fragmentados e incompletos. Os genes ortólogos cópia-única foram utilizados na construção da árvore filogenômica descrita na **Seção 3.7**.

3.5.2.3 Repeat Masker

O programa Repeat Masker (Smit, Hubley, & Green, 2013) realiza a identificação, classificação e mascaramento de elementos repetitivos e sequências de baixa complexidade através da busca por similaridade dos contigs analisados com um banco de dados de repetições conhecidas

(ie.: Repbase). Este software foi utilizado na avaliação do conteúdo de repetições dos genomas e também é utilizado pelo MAKER2 no mascaramento de regiões de baixa complexidade.

3.6 Predição gênica *de novo* e anotação funcional

A predição de genes codificadores de proteínas para os genomas estudados foi realizada com o programa **MAKER2** (Holt & Yandell, 2011), que combina a predição *de novo* de três preditores externos (**GeneMark** , **AUGUSTUS** e **SNAP**). Estes preditores constroem HMMs de perfil (*Hidden Markov Model*) a partir de genes conhecidos de uma espécie próxima – **AUGUSTUS** (Stanke & Morgenstern, 2005) –, ou de genes identificados *a priori* no genoma investigado por outros *softwares* como o **CEGMA** (**SNAP**), ou ainda a partir de treinamento não-supervisionado no próprio genoma – **GeneMark-ES** (Ter-Hovhannisyan, *et al.*, 2008). Estes modelos são aplicados principalmente na discriminação de junções exon/íntron, mas também na discriminação entre regiões intergênicas e genes.

3.6.1 Treinamento do preditor AUGUSTUS

O preditor **AUGUSTUS** possui conjuntos de dados (*datasets)* treinados para diversas linhagens de referência. O *dataset* relativo à espécie filogeneticamente mais próxima de *Metschnikowia* é o de *Scheffersomyces stipitis* (no programa, referido como *Pichia stipitis*), sendo as Famílias Pichiaceae e Metschnikowiacea *taxa* irmãos. Entretanto, optamos por realizar o treinamento do **AUGUSTUS** utilizando uma referência ainda mais próxima. A espécie escolhida foi *Clavispora lusitaniae*, uma vez que *Clavispora* é gênero irmão de *Metschnikowia*, dentro da família Metschnikowiaceae e possui dados de RNAseq disponíveis. Foi comparado o número total de CDSs preditas utilizando S. *stipitis* e *C. lusitaniae*, e os maiores números obtidos para as predições com *C. lusitania* evidenciaram uma melhor capacidade de predição com esta referência. O treinamento do AUGUSTUS para *C. lusitania* foi realizado de acordo com os procedimentos descritos a seguir.

Para o treinamento foram fornecidos o genoma e o transcriptoma da espécie. A montagem do genoma utilizada foi obtida do NCBI (Assembly: **GCA_000003835.1**). O transcriptoma foi mapeado a partir de *reads* de RNAseq obtidas do SRA (SRR2141707), com o programa **STAR**

(Dobin *et al.*, 2013). O *output* do **STAR** - um arquivo *bam* - foi convertido para um arquivo de "dicas" (informações extrínsecas a serem utilizadas na predição) utilizando o *script bam2hints* (do pacote **AUGUSTUS**). Uma das etapas do treinamento consiste na avaliação da capacidade de predição para o *dataset* treinado. A montagem do genoma a ser utilizada é aleatoriamente dividida em dois grupos, um para o treinamento e outro para avaliação. A montagem do genoma de *C. lusitaniae* utilizada é composta por 9 *scaffolds.* Estes foram divididos utilizando o *script randomSplit.pl* e o arquivo *gff* da montagem como *input*. Os arquivos *gff* correspondentes aos *scaffolds* para treinamento e o genoma (*.fasta*) foram utilizados no *script autoAug.pl*. O resultado do treinamento foi associado ao identificador *clavispora_lusitaniae*.

3.6.2 Predição das CDSs pelo MAKER2

A predição das possíveis CDSs (*Coding DNA sequences* – sequências de DNA codificadoras) e proteínas (CDSs traduzidas) correspondentes foi feita com o programa **MAKER2 V2.31.8** (Holt & Yandell, 2011). Para o programa são fornecidos o arquivo *fasta* dos *contigs* do genoma e três arquivos de configuração. O programa foi rodado em duas iterações para cada genomas. Na primeira foram fornecidos os arquivos de HMMs gerados pelo **SNAP** e pelo **GeneMark-ES** para cada genoma e o *dataset* de treino gerado para o **AUGUSTUS**. Na segunda, um novo arquivo de HMM foi gerado com os *scripts* SNAP, com os mesmos procedimentos da sessão anterior, utilizando o *gff* gerado pelo **MAKER2** para cada genoma na primeira iteração e foi fornecido também o *gff* de cada genoma produzido na primeira iteração

3.6.3 Anotação funcional dos genes preditos

A anotação das CDSs preditas para os genomas foi realizada por **BLASTx** no *software* **DIAMOND 0.9.22.123** (Buchfink, Xie, & Huson, 2015), contra os bancos de dados **Uniprot/Swissprot**, e por BLASTn contra o **NCBI/NR** (v98).

Foi também utilizado o programa **INTERPROSCAN 5.39-77.0** que possibilita a análise funcional das proteínas através da busca por assinaturas de domínios de famílias proteicas conhecidas, ontologias e características estruturais. A versão utilizada foi a **v39**, que conta com as análises dos programas: TIGRFAM, SFLD, SUPERFAMILY, PANTHER, Gene3D, Hamap,

Coils, ProSiteProfiles, SMART, CDD, PRINTS, ProSitePatterns, Pfam, MobiDBLite, PIRSF, SignalP_EUK, Phobius, SignalP_GRAM_POSITIVE, SignalP_GRAM_NEGATIVE, TMHMM, ProDom.

O programa **tRNAscan-SE** (**v1.4**) (Chan & Lowe, 2019) foi utilizado para predizer genes de tRNAs nos genomas.

3.6.4 Visualização das métricas no Rstudio

Foi criado um *script* em **R** para extração das métricas dos *outputs* dos programas mencionados anterior e posteriormente, para melhor comparação da variação das métricas quantitativas e qualitativas dos genomas, em gráficos gerados no **Rstudio** (*Rstudio team*, 2015), utilizando os pacotes *dplyr*, *tibble*, *stringr* e *ggplot2*. Este *script* em **R** e outros *scripts* construídos para este *pipeline* estão disponibilizados no Git Hub (http://github.com/heronoh/prj_metsch).

3.7 Construção de árvores filogenômicas das espécies do gênero *Metschnikowia*

3.7.1 Seleção dos ortólogos completos de cópia única comuns aos genomas

O programa **BUSCO** tem como um de seus *outputs* uma pasta contendo os arquivos *fasta* correspondentes a cada um dos genes ortólogos de cópia única completos identificados para cada genoma. Cada arquivo tem como nome o identificador do ortólogo. Foram selecionados os identificadores de ortólogos de cópia única comuns a todos os genomas (ortólogos 1:1) utilizando comandos de *bash* (http://github.com/heronoh/prj_metsch) como *grep* e *awk*. A partir de uma lista de 1317 identificadores foram recuperadas as sequências de aminoácidos correspondentes utilizando o *script pegar_seqs_id.pl* (http://github.com/sujaikumar/assemblage), produzindo um arquivo *fasta* por genoma. Os arquivos *fasta* foram concatenados em um único arquivo utilizado como *input* para o *script reformat.pl*, que separa as sequências por ortólogo, produzindo, ao final, 1317 *fastas* contendo as sequências de todos os genomas para cada ortólogo.

3.7.2 Alinhamento dos ortólogos

Utilizando o **MUSCLE** (**v3.8.31**) (Edgar, 2004) (*script batch_muscle.pl*) foi realizado o alinhamento das sequências de cada ortólogo. Em seguida, as extremidades não-alinhadas de cada ortólogo foram removidas com o **trimAL** (Capella-Gutierrez, Silla-Martinez, & Gabaldon, 2009) (*script batch_trimal.pl)*. Desta maneira, todos os ortólogos alinhados passam a ter o mesmo número de posições. Todos os ortólogos foram novamente concatenados por organismo em um único arquivo *fasta* (utilizando o *script concatenar.pl*), convertido a.*phy* (*script convert_fasta_to_phy.sh*) que é o *input* do **RAxML**. Concomitantemente a este arquivo *fasta,* o *script* gera um outro arquivo de partições do *.phy*, que contém as coordenadas de início e fim de cada um dos ortólogos, para serem tratados como entidades independentes no cálculo automático de matrizes de substituição geradas pelo **RAxML**.

3.7.3 Construção da árvore filogenômica por Máxima Verossimilhança no RAxML

O *software* **RAxML** (*Randomized Axelerated Maximum Likelihood* - Stamatakis, 2014) foi utilizado para construção da árvore filogenômicas, com *bootstrap* de 100 réplicas e matriz de substituição estimada automaticamente para cada partição. O *output* no formato *newick*, correspondente à árvore, foi visualizado utilizando a ferramenta *web* **ITOL** (Letunic & Bork, 2016).

3.8 Construção de redes de similaridade de ortólogos

3.8.1 Identificação dos genes homólogos utilizando o orthoMCL

Para a identificação das relações de similaridade que refletem ortologia e paralogia entre as proteínas de todos os 35 genomas de *Metschnikowia* e *Clavispora* foi inicialmente utilizado o programa **DIAMOND**, tendo tanto como *query* quanto como *subject* todas as proteínas de todos os genomas, utilizando a *flag* **-k o**, para a recuperação de todos os *hits* possíveis. O resultado desta análise no formato tabular foi utilizado como *input* para o **orthoMCL** (Li, Stoeckert, & Roos, 2003), que relaciona os genes em grupos de ortólogos com base em *best hits* bidirecionais. O valor de *inflation* utilizado foi de 1,5.

3.8.2 Construção de uma rede de similaridade de ortólogos para os genomas

A partir do resultado do **DIAMOND** foram extraídos os valores de percentual de similaridade entre os pares de genes, através de um *script* construído em **R**, para produzir dois arquivos (arquivos de **Nós** e **Vértices**) utilizados como input para o programa **GEPHI** (Bastian, Heymann, & Jacomy, 2009). Foram adicionados vértices artificiais referentes às espécies para organizar a distribuição dos genes. Todos os genes foram associados às respectivas espécies com valor peso de aresta = 10. Foram utilizados os algoritmos de construção de grafo **FORCE ATLAS 2** (Jacomy *et al.*, 2014), **Open Ord** (Martin *et al.*, 2011) e a função "Expansão". Para melhor visualização, foram omitidos genes sem relação para os critérios de corte utilizados na construção da rede. Para simplificação da estrutura da rede foram utilizadas apenas as duas espécies mais próximas do clado de interesse, o clado das ***M*EPq aquáticas**, totalizando 6 espécies. Foram selecionados apenas pares *query*:*subject* com alinhamento superior a 50 nt, cobertura recíproca >70 %, similaridade >80 % (mas estes valores podem ser customizados no *script*). Os valores de similaridade foram utilizados como peso nas respectivas arestas. Estes valores foram escolhidos a partir de outras redes geradas com *cutoffs* diferentes, produzindo o melhor resultado visual. Esta metodologia está em desenvolvimento para integrar valores de similaridade menores, a fim de melhor representar as relações entre genes ortólogos correspondendo aos mesmos grupos encontrados pelo **orthoMCL**.

3.8.3 Identificação de grupos de parálogos com variação entre M. australis, M. bicuspidata e as demais Metschnikowia

Um *script* em **R** foi escrito para identificar, a partir dos grupos de ortólogos identificados pelo **orthoMCL**, aqueles que possuem ortólogos com variação do número de cópias entre as leveduras. Foram construídos gráficos de densidade para relacionar as distribuições destes grupos entre as espécies e com relação à média das demais *Metschnikowia*.

3.9 Estimativa do tempo de divergência de *M. australis e M. bicuspidata*

O programa **gKaKs v1.3** (Zhang *et al.*, 2013) foi utilizado para estimar a taxa se substituição sinônima e não-sinônima entre pares de ortólogos correspondentes no genoma de *M. australis* e *M. bicuspidata*. A partir deste resultado foram selecionados dois conjuntos de genes: a) os genes para os quais *Ka/Ks* > 1 são considerados sob seleção positiva, já que as mutações levam à troca de AA; b) os genes para os quais *Ka/Ks* << 1 são considerados de seleção purificadora ou negativa, pois as mutações não alteram a sequência de AA. A taxa de substituição média destes genes pode ser utilizada para estimar o tempo de divergência das duas espécies (De Mendonça Vilela *et al.*, 2017). Para tal, a taxa de substituição é combinada com a taxa de mutação estimada para leveduras e o tempo médio de duplicação celular das linhagens, segundo a equação a seguir:

$$T = \frac{d}{2r}$$

T: tempo de divergência (número de gerações)

d: número de substituições entre duas sequências (kS médio)

r: taxa de substituição/ mutação estimada para a espécie

A taxa de mutação de nucleotídeo único utilizada foi de **$1.67 \pm 0.04 \times 10^{-10}$ por base por geração** estimada por Zhu e colaboradores (2014).

3.10 Identificação de CDSs exclusivas de *M. australis* ou com baixa similaridade a *M. bicuspidata*

Para encontrar as CDSs presentes apenas no genoma de *M. australis* foi realizado um **BLASTp** das CDSs traduzidas preditas em *M. australis* contra todas as CDSs traduzidas dos demais genomas, utilizando o *evalue* de *cutoff* = **$1e^{-10}$**. A CDSs que não apresentaram *hits* foram consideradas exclusivas. Elas foram submetidas a um tBLASTn contra os *contigs* de *M. bicuspidata* para verificar a existência de CDSs compartilhadas que pudessem não ter sido preditas nos genomas de *M. bicuspidata.* Estas foram identificadas e sinalizadas. Quando parcialmente compartilhadas, a sequência das CDSs e dos *contigs* de *M. bicuspidata* correspondentes foram alinhadas com o programa **MUSCLE 3.8** (Edgar, 2004).

Todas as CDSs de todos genomas foram classificadas pelos dois classificadores de AFPs **RAFP- pred** e **CryoProtec**, mas apenas as CDSs exclusivas foram avaliadas pelo terceiro classificador, o **iAFP-ense**. As CDSs exclusivas classificadas como AFPs por todos os três programas foram selecionadas para construção de iniciadores e verificação de sua expressão pelas leveduras *M. australis* e *M. bicuspidata* quando cultivadas a 6 e a 12 ˚C.

3.11 Avaliação das CDSs preditas por classificadores de AFPs e preditores de características estruturais

Todas CDSs preditas em todos genomas foram submetidas à uma categorização pelos dois classificadores de AFPs disponíveis, o **CryoProtect** (Pratiwi *et al.*, 2017) e o **RAFP-pred** (Khan *et al.*, 2018) para avaliar a capacidade de distinção dos mesmos. A proporção das CDSs classificadas para os genomas foi relacionada para visualização utilizando um *script* em **R**. Um terceiro classificador disponível, o **iAFP-ense** (Xiao *et al.*, 2016), foi utilizado apenas na classificação das CDSs exclusivas de *M. australis* selecionadas na **Seção 3.10**, uma vez que este serviço *web* tem limite de classificação de 100 sequências. Similarmente, o serviço *web* **Protter** (Omasits *et al.*, 2014) também foi utilizado para predizer características de sequências de aminoácidos apenas para as CDSs selecionadas.

3.11.1 CryoProtect

Este programa utiliza a composição de aminoácidos (AA) e diAA de cada CDS traduzida e realiza a classificação através da estratégia de *random forest*, com árvores de decisão construídas e testadas com um *dataset* contendo 478 AFPs e 9.139 não-AFPs, sendo 300 proteínas de cada tipo selecionadas aleatoriamente para treino. Para classificação, todas as CDSs traduzidas de todos genomas foram enviadas para um servidor *web*, no endereço http://codes.bio/cryoprotect/ .

3.11.2 RAFP-pred

Este programa também utiliza composições de AA e diAA para dois segmentos de cada CDS traduzida (a CDS é dividida ao meio) e a informação destes dois segmentos é combinada na

classificação. Para a obtenção destas composições as CDS foram submetidas previamente ao programa **pseAA-builder v1.05** (Du *et al.*, 2012). O *output* no formato *.tab* foi convertido para o *input* do **RAFP-pred** (*.arff*) por meio de um *parser* escrito em **R**. O **RAFP-pred** também utiliza da estratégia de *random forest* e é treinado com o mesmo conjunto de dados do **CryoProtect**. A classificação foi realizada na plataforma de análises estatísticas **WEKA 3.9** (Hall *et al.*, 2009).

3.11.3 iAFP-ense

Este programa utiliza características PSSM (*Position-specific Scoring Matrix*) e de pseudocomposição de AA onde, além da composição de AA tradicional, adiciona-se informação de composição de pares de aminoácidos separados por uma distância **N**, incorporando-se assim uma certa relação estrutural. A partir desta informação, a estratégia de treinamento e classificação utilizada se baseia em *random forest*, em um serviço *web* (http://www.jci-bioinfo.cn/iAFP-Ense). O **iAFP-ense** também é treinado com o mesmo conjunto de dados do **RAFP-pred** e do **CryoProtect**.

3.11.4 Protter

O programa **Protter** foi utilizado na predição de características intrínsecas às sequências das proteínas preditas, tais como regiões transmembrana, regiões intra/extracelulares, sítios de N-glicosilação, presença de peptídeo sinal e sítios de clivagem proteolítica. Estas atribuições de características de sequência são realizadas através da recuperação pelo **Protter** de informações de bancos de dados como o UniProt, ou recrutamento de *web services* de terceiros, como o preditor de domínios transmembrana e peptídeo sinal **Phobius** (Käll, Krogh, & Sonnhammer, 2004). As sequências foram submetidas no formato *fasta* (aminoácidos) e o *web service* retornou diagramas das proteínas destacando os motivos e sítios preditos. Estes resultados foram considerados na seleção de CDSs na **Seção 3.10**, mas não foram determinantes para as escolhas.

3.12 Detecção da expressão das CDSs exclusivas de *M. australis* ou parcialmente compartilhadas com *M. bicuspidata*

3.12.1 Seleção de CDSs e desenho de primers

Das CDSs selecionadas de *M. australis* classificadas como AFPs pelos 3 classificadores foram escolhidas 16 para os ensaios de verificação de expressão em crescimento a baixas temperaturas. Foram também selecionadas sequências correspondentes ao RNA ribossomal 28S e aos genes codificadores das proteínas Actina e Tubulina, como controles positivos de expressão. Os iniciadores foram construídos utilizando o serviço *web* de construção de iniciadores do NCBI, o **primer-BLAST** (Ye *et al.*, 2012). Para cada par de *iniciadores* foi fornecida a CDS respectiva e os genomas de *M. australis* e *M. bicuspidata*. Foram selecionados os pares de *iniciadores* correspondentes a *amplicons* de tamanho entre 110 nt e 130 nt com alinhamento em *M. australis* e, quando possível, capazes de alinhar no genoma de *M. bicuspidata*. As sequências dos iniciadores estão relacionadas na Tabela 2.

Tabela 2 — Relação dos iniciadores construídos para as CDSs selecionadas.

CDS	Amplicon esperado em		Pares de *primers*	Sequência do *primer*
	M. bicuspidata	*M. australis*		
aus_3013	126	126	3013_fwd	ACCACCCCTAGGTGCTATAC
			3013_rev	ATGTTGAAGTCTACCCTGCG
aus_3157		128	3157_fwd	GACCTCCATGTTCCCTGCAT
			3157_rev	TCGTGGAGCTGGAGGTATATG
aus_3218	190	117	3218_fwd	CCGAGACCCGCGATACT
			3218_rev	GACCCGATGCAAATTCGAAAA
aus_3362	744	116	3362_fwd	GATCGTTGGGTATGTCGGAG
			3362_rev	TTACGGTCCCGTCGATAAC
aus_3391		119	3391_fwd	TATCTGCCTCGTTTTCGCTC
			3391_rev	ATTCGCAATACGACACGACA
aus_3484		111	3484_fwd	CGAGACGTCGTGTACTTTCA
			3484_rev	CAGTGATTGCCATATCCCGA
aus_3629		129	3629_fwd	GCCGAGGAAGAATCTACAGC
			3629_rev	CAATGGTTCCCATCCAGCAG
aus_3748	202	122	3748_fwd	TGGCTCTATTCACGGGAAAC
			3748_rev	TTGTTCTCCCATTGTTGGCT
aus_3830		128	3830_fwd	ACGGGCATCTAAAGACAGAT
			3830_rev	ACTCCTTTGTTGAGAAGTCCTG
aus_3946		130	3946_fwd	CAATGCCGCCAAATTAACCG
			3946_rev	ACGCACGTATCCATTTCCAAT
aus_3951		121	3951_fwd	CGCGGTCGCTATCCTAAAAT
			3951_rev	AATCTGGAGATTTGGCCACC
aus_3966	618,1835	120	3966_fwd	GATCTCCAACAAAAAGCCCG
			3966_rev	TTCGATATTGGTCTGCCACG
aus_4100	120	120	4100_fwd	GTCAAATGTCGCGTTCATCG
			4100_rev	GCGAGGCTAACTTTTTCATCA
aus_4351		111	4351_fwd	GACGTGCTCACCAATGACTC
			4351_rev	TGAAGCTCCACTAATTTGCGA
aus_4866	727,764	130	4866_fwd	GCATTGAATACCACGTCAGC
			4866_rev	TTATTCCTGGGCCCTCTCC
aus_5130	125	125	5130_fwd	ACGAGGACGACACTATCGAA
			5130_rev	CAGTGAGTCATTGGTGAGCA
Tubulin	130	130	Tub_fwd	ATCGAGGGTTCTGAATTGGC
			Tub_rev	GAACGAAGGCTCTCTTGGAG
Actin	119	119	Act_fwd	TGGTTATCGACAACGGTTCC
			Act_rev	CATGCCGACCATGATACCTT
28S	123,12	123	28S_fwd	GACGAAGCCTTGGATGTGAA
			28S_rev	TCTGCTGTTGACATGGAACC

Fonte: Elaborado pelo autor

3.12.2 Extração de RNA de cultivos a 12 ˚C e obtenção do cDNA

A partir de cada um dos cultivos de *M. australis* e *M. bicuspidata* a 12 ˚C em meio líquido, foi retirado 1 ml de meio (OD ~ 15, em torno de $6x10^7$ células). Em seguida, cada tubo foi agitado vigorosamente em vortex e centrifugado a 1.000g por 5 min. O sobrenadante foi descartado e ao *pellet* foi adicionado 1ml de Trizol (*Thermo Fisher Scientific*). Após incubação por 5 min à temperatura ambiente foi adicionado 100 µl de clorofórmio, seguido de agitação por 15 segundos e incubação por 3 min em temperatura ambiente. Após centrifugação a 11.000g por 15 min, o sobrenadante contendo o RNA foi transferido para um novo tubo ao qual foi adicionado 250 µl de etanol 70% gelado. Este tubo foi centrifugado a 4 ˚C por 5 min a 7500 g. Esta última etapa foi repetida mais uma vez para limpeza do RNA. Em seguida, o excesso de etanol 70% foi retirado e o tubo foi colocado a 50 ˚C até secar completamente. Posteriormente, foram adicionados 30 µL de água milli-Q e o tubo foi mantido a 50 ˚C por 40 min para solubilização completa do RNA. Para a remoção de DNA da amostra, foi adicionado 1 µl DNase I (1U/µl) (*Thermofisher*) e o tubo foi mantido a 37 ˚C por 30 min. A DNase foi removida utilizando resina inativadora (*kit DNAse I*) e centrifugação por 90 segundos a 10.000 g e o sobrenadante contendo o RNA foi transferido para um novo tubo.

O RNA extraído teve sua pureza e quantidade avaliadas por espectrofotometria no equipamento NanoDrop (*Thermo Fisher Scientific*). Para a produção de cDNA, o volume correspondente a 700 µg de RNA foi acrescido de 1 µl de oligo-dT_{20} (50mM - *iniciador*) e 1 µl de dNTP (10mM). Esta mistura foi incubada a 65 ˚C por 5 min e colocada no gelo por 1 min. Este tubo foi acrescido de uma mistura contendo 2 µl de tampão 10x, 4 µl de $MgCl_2$ (25mM), 1 µl de RNAseOUT (40U/µl - *kit Superscript III – Thermofisher*) e 1µl da polimerase SuperScript III RT (200U/ µl) (*Thermofisher* (*Invitrogen*). Esta mistura foi incubada a 50˚C por 50 min. Em seguida, a reação foi paralisada pelo aumento da temperatura para 85˚C por 5 min. Após o término, 1 µl (2 U/µl) de RNase H (*Thermofisher* (*Invitrogen*) foi adicionado e a mistura foi incubada a 37 C por 20 min. Ao final destas etapas, o cDNA foi armazenado a -20 ˚C até o momento de sua utilização nas reações de PCR.

3.12.3 PCR do gDNA e cDNA para amplificação das CDSs selecionadas

Para cada reação de 20 µl foi utilizado uma mistura contendo 14 µl de água milli-Q autoclavada, 0,2 µl (1U) de *Taq* DNA polimerase (*Phoneutria*), 1,6 µl de dNTPs (2,5 mM), 2 µl do tampão da enzima ($MgCl_2$ 25mM), 0,6 µl do iniciador *forward* (200mM) e 0,6 µl do iniciador *reverse* (200mM). A essa mistura foram adicionados 1µl do gDNA ou do cDNA (diluídos a 200ng/µl) das leveduras *M. australis* ou *M. bicuspidata*. As reações foram realizadas segundo as etapas descritas na Tabela 3:

Tabela 3 — Programa de PCR utilizado na amplificação dos gDNAs e cDNAs

Etapa	Tempo	Temperatura	Ciclos
1	10’	95 ˚C	1x
2	30’’	95 ˚C	30x
3	40’’	55 ˚C	
4	30’’	72 ˚C	
5	5’	72 ˚C	1x
6	Hold	4 ˚C	1x

Fonte: Elaborado pelo autor

Os produtos resultantes foram aplicados em géis de Acrilamida 12% (TBE 5x (10 ml), acrilamida 29:1 bis acrilamida (10 ml), água milli-Q (30 ml), TEMED (40 µl), persulfato de amônio (400 µl) e submetidos a eletroforese a 80V e 85mA e por uma hora. Em seguida os géis foram corados por Nitrato de Prata: a) Solução Fixadora (*etanol 10%v/v, ácido acético 0,5%v/v, água milli-Q q.s.p*) por 10’; b) Solução de Prata – *Solução Fixadora (30% v/v, $AgNO_3$ 1,3g/l)*, por 10’; c) Solução Reveladora (*NaOh 30g/l, formaldeído 5%v/v, água milli-Q q.s.p.*, até a visualização das bandas) e digitalizados no equipamento ***Gel Doc EZ Imager*** (Bio Rad).

4 RESULTADOS

4.1 Ensaios de crescimento

4.1.1 Curva de crescimento de M. australis a 6 ˚C, 12 ˚C e 28 ˚C

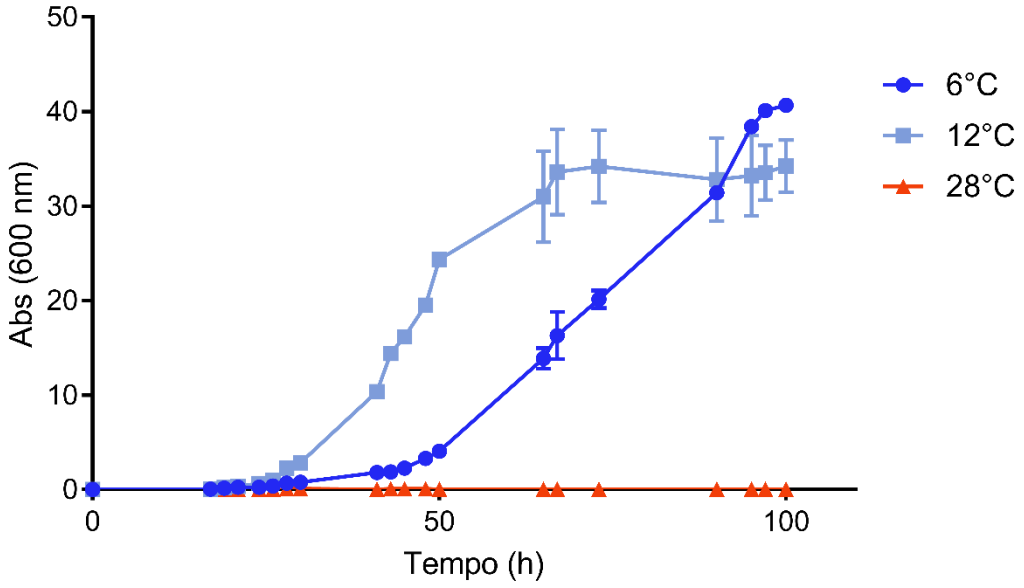
O primeiro ensaio, apenas com *M. australis* nas temperaturas de 6 ˚C, 12 ˚C e 28 ˚C, produziu os perfis de crescimento em YPD mostrados na **Figura 11**. Observa-se que a temperatura em que a levedura apresentou melhor crescimento foi a de 6 ˚C, atingindo OD^{MAX}≈40, superior aos cultivos nas demais temperaturas, mas em uma taxa de crescimento menor que a 12 ˚C. Observou-se também que a levedura não foi capaz de crescer a 28 ˚C. Entretanto, as células da cultura continuaram total ou parcialmente viáveis, uma vez que, ao término destas medições, os cultivos de 28 ˚C foram transferidos para 12 ˚C, atingindo OD^{MAX} ≈ 30 (dados não representados).

OD^{MAX} *M. australis* 6 ˚C ≈ 38,4

OD^{MAX} *M. australis* 12 ˚C ≈ 32

OD^{MAX} *M. australis* 28 ˚C ≈ 0

Figura 11 — *M. australis* cultivada a 6 ˚C, 12 ˚C e 28 ˚C.
Os pontos representam a média de três réplicas experimentais. Cultivo realizado em YPD, partindo do inóculo inicial de $2x10^5$ células. Os pré-inóculos foram cultivados por dois dias nas temperaturas respectivas, partindo de um inóculo de $1x10^6$ células, exceto o correspondente a 28 ˚C, que foi inicialmente cultivado a 20 ˚C.



Fonte: Elaborado pelo autor.

4.1.2 Curva de crescimento de M. australis, M. bicuspidata, M. golubevii e S. cerevisiae a 12 ˚C

Realizamos experimentos de crescimento das quatro leveduras em meio YPD por 72 horas, sendo isso mostrado na **Figura 12**. Observa-se que todas leveduras são capazes de crescer a 12 ˚C, sendo que *M. golubevii* apresentou o melhor desempenho, atingindo maior OD^{MAX} (38). Observa-se também que *M. australis* e *M. bicuspidata* apresentam perfil de crescimento semelhante. A levedura *S. cerevisiae* teve o pior desempenho de crescimento nesta temperatura, atingindo a $OD^{MAX} \approx 19$.

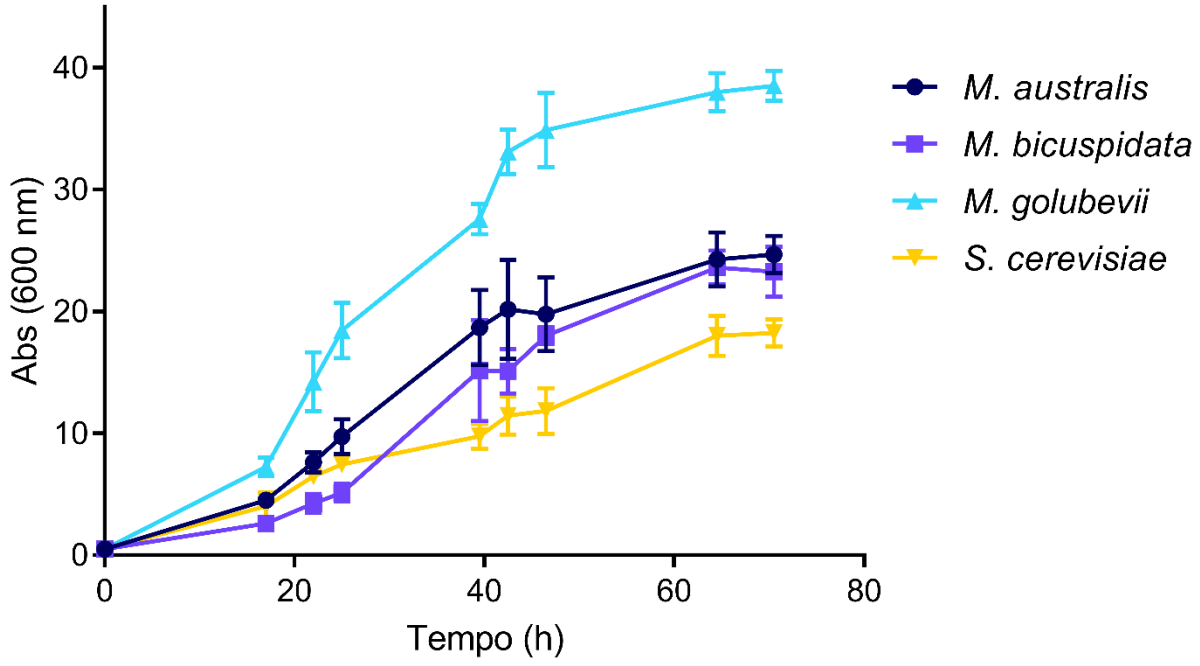
OD^{MAX} M. golubevii ≈ 38

OD^{MAX} M. australis ≈ 25

OD^{MAX} M. bicuspidata ≈ 23

OD^{MAX} S. cerevisiae ≈ 18

Figura 12 — Cultivo de *M. australis*, *M bicuspidata*, *M. golubevii* e *S. cerevisiae* a 12 ˚C.
Os pontos representam a média de três réplicas experimentais. Cultivo realizado em YPD, partindo do inóculo inicial de $2x10^5$ células. Os pré-inóculos foram cultivados a 12 ˚C por 2 dias, partindo de um inóculo de $1x10^6$ células.



Fonte: Elaborado pelo autor.

4.1.3 Curva de crescimento de M. australis, M. bicuspidata, M. golubevii e S. cerevisiae a 6 ˚C

A **Figura 13** mostra o crescimento das quatro leveduras em YPD a 6˚C. Observa-se que todas leveduras são capazes de crescer nesta temperatura, sendo que *M. golubevii* apresentou novamente o melhor desempenho, atingindo maior OD^{MAX} (29). Observa-se também que nesta temperatura há grande diferença entre os perfis de crescimento de *M. australis* e *M. bicuspidata*, sendo o crescimento desta última semelhante ao de *S. cerevisiae*. Já a levedura *M. australis* apresentou um perfil de crescimento semelhante ao de *M. golubevii*.

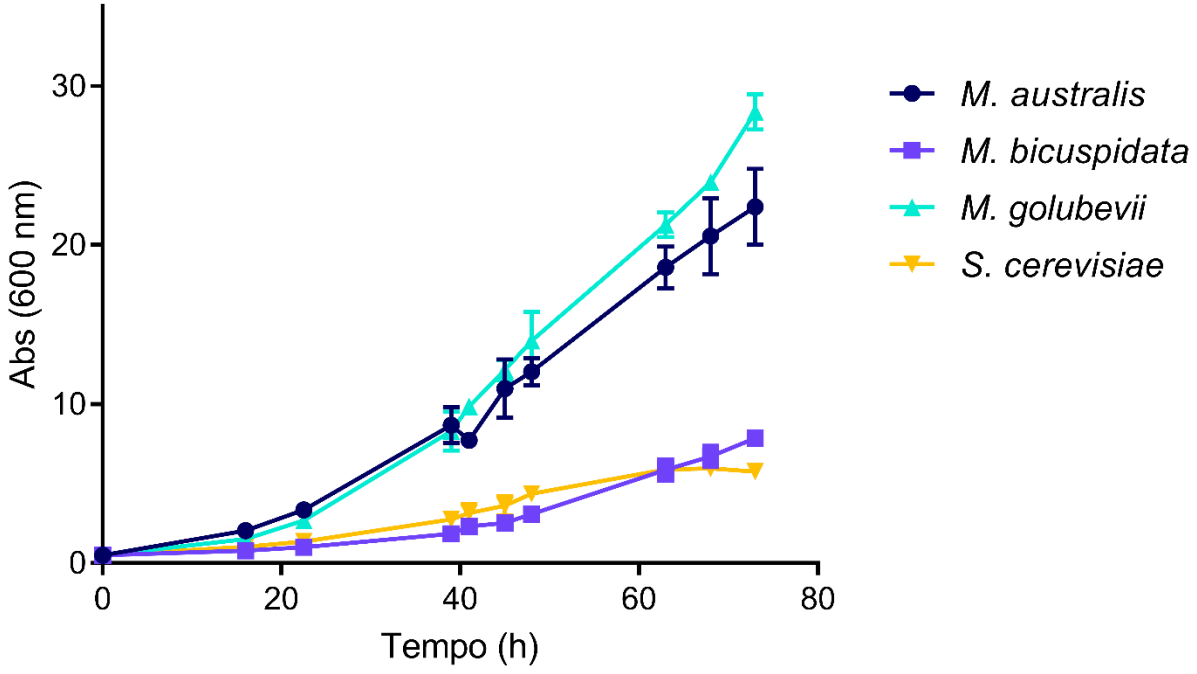
OD^{MAX} M. golubevii ≈ 29

OD^{MAX} M. australis ≈ 23

OD^{MAX} M. bicuspidata ≈ 8

OD^{MAX} *S. cerevisiae* ≈ 5,5

Figura 13 — Cultivo de *M. australis*, *M bicuspidata*, *M. golubevii* e *S. cerevisiae* a 6˚C.
Os pontos representam a média de três réplicas experimentais. Cultivo realizado em YPD, partindo do inóculo inicial de $2x10^5$ células. Os pré inóculos foram cultivados a 6˚C por 2 dias, partindo de um inóculo de $1x10^6$ células.



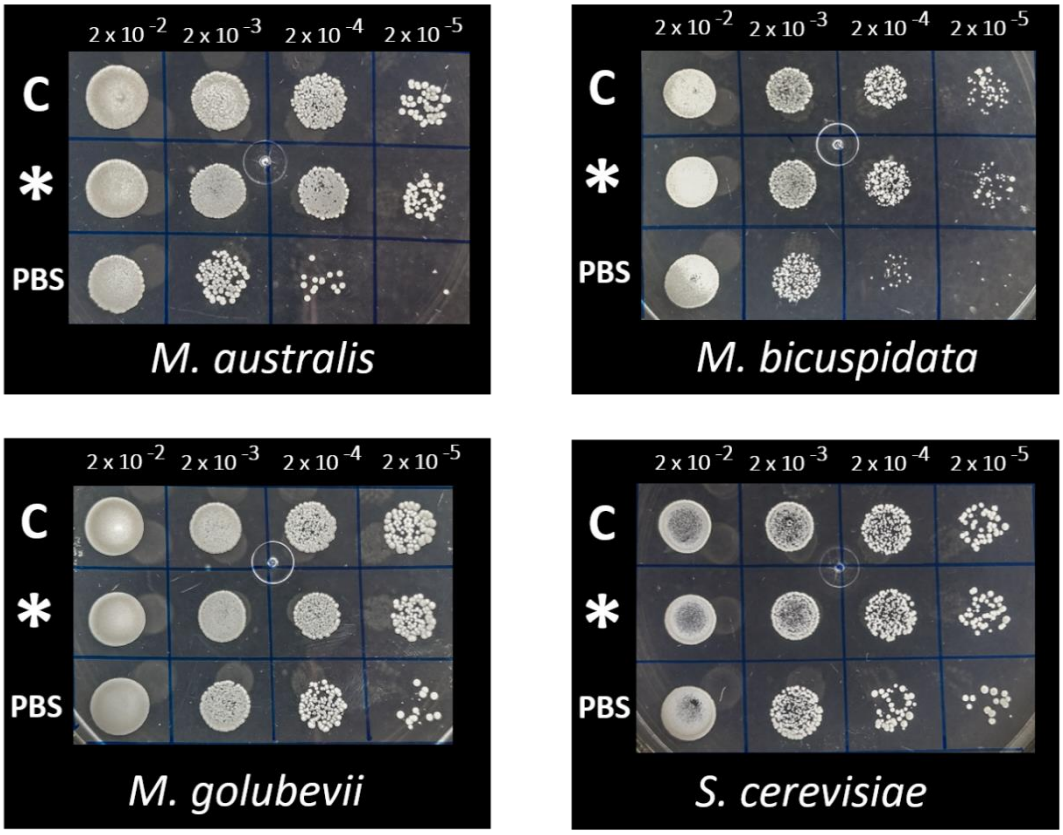
Fonte: Elaborado pelo autor.

4.2 Ensaios de sobrevivência ao congelamento a -80 ˚C

4.2.1 Sobrevivência ao congelamento a -80 ˚C

Realizamos os ensaios de sobrevivência ao congelamento com as quatro leveduras desse estudo e os resultados podem ser visualizados na Figura 14. Observa-se que todas leveduras testadas, provenientes de cultivos realizados a 6˚C, sobreviveram à exposição a - 80 ˚C por 2 horas, com crescimento de colônias correspondente aos níveis do grupo controle quando diluídas em YPD. Observa-se que as diluições das células em PBS apresentaram redução da sobrevivência ao congelamento. Os perfis de sobrevivência ao congelamento aqui apresentados são semelhantes aos observados quando as células partiram de um crescimento a 12 ˚C (dados não mostrados).

Figura 14 — *Spot-test* de sobrevivência à exposição a - 80 ˚C das leveduras *M. australis*, *M bicuspidata*, *M. golubevii* e *S. cerevisiae* a partir de inóculos a 6 ˚C.
Após plaqueamento e exposição ao congelamento, as leveduras foram cultivadas a 12 ˚C, até o aparecimento de colônias na menor diluição.



Fonte: Elaborado pelo autor.

4.3 Sequenciamento, montagem e anotação do genoma de *M. australis*

O sequenciamento do tipo *Whole Genome Sequencing* (WGS) realizado na plataforma *MiSeq* gerou 2x 792.561 *reads* de até 301 nt (*paired-ends* – PE), com um tamanho médio de inserto de 1,167 pb. O sequenciamento do tipo WGS realizado na plataforma *HiSeq* gerou 2x 51.656.292 *reads* de até 101 nt (PE), com um tamanho médio de inserto de 552 pb.

Após controle de qualidade realizado com o programa **Fastqc** e remoção total ou parcial de *reads* com valor de PHRED menor que 30 com o programa **Trimmomatic**, restaram para as *reads* de *MiSeq*, 2x 743.528 *reads* pareadas, mais 47.422 *reads* não-pareadas (totalizando 96,80% do conjunto inicial) e, para as reads de *HiSeq*, 2x 35.754.852 *reads* pareadas, mais 15.296.650 *reads* não-pareadas (totalizando 84,02% do conjunto inicial). A montagem resultante destes conjuntos de dados resultou num genoma de 14,35Mb, organizado em 154 *contigs*, cujas principais métricas estão sumarizadas na Tabela 3, a seguir. A predição gênica inicial, utilizando o **MAKER2**, encontrou 4442 CDS. Estes procedimentos foram publicados em (Batista *et al.*, 2017) (Anexo I) e o genoma foi depositado no NCBI DDBJ/ENA/GenBank (accession: GCA_002073855.1). Uma nova de predição gênica utilizando o **MAKER2** em duas iterações aumentou o número de CDSs preditas para 5440. As características dos genomas estão sumarizadas na Tabela 4.

Tabela 4 — Características da montagem do genoma de *M. australis*.

Número de *contigs* >500bp	Número de contigs >1000bp	Tamanho do genoma	Profundidade	L50	N50	CDSs preditas	%GC	Completude CEGMA	Completude BUSCO
153	94	14,35 Mb	780x	10	542.232 bp	5440	47,2	95,9%	95,7%

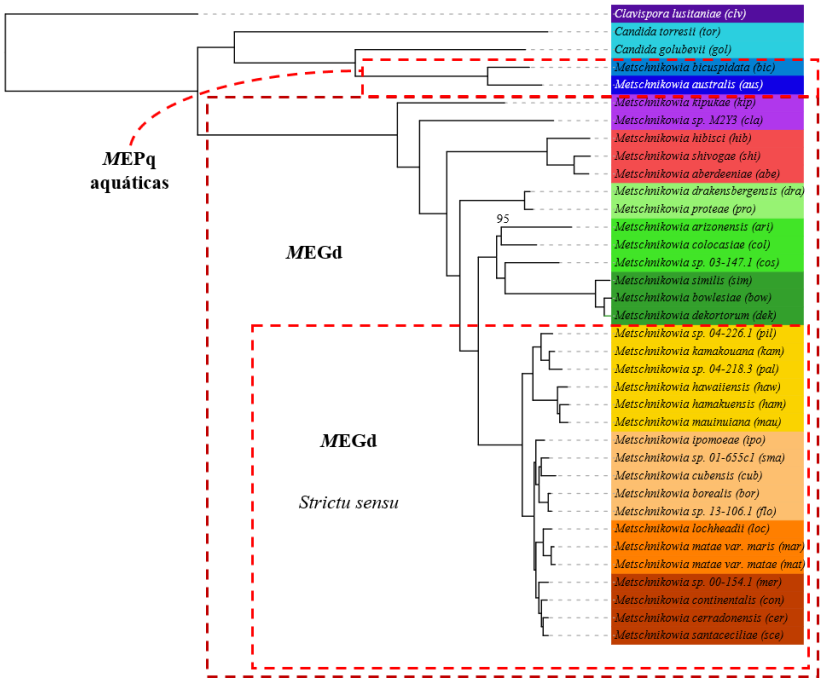
Fonte: Elaborado pelo autor.

4.4 Árvores filogenômicas das espécies do gênero *Metschnikowia*

A filogenia do clado das ***M*EGd** é bem estabelecida e foi realizada com dados genômicos (Lachance *et al.*, 2016), assim como a relação desta subdivisão com outras *Metschnikowia* (Shen *et al.*, 2018). Entretanto, como estas reconstruções filogenômicas não abrangem todas leveduras abordadas neste trabalho, foi realizada uma nova reconstrução.

O programa BUSCO foi utilizado para identificar 1759 genes ortólogos presentes nos 36 genomas. A partir destes foram selecionados os ortólogos de cópia única comuns a todos os genomas, totalizando 1317 genes. Estes genes traduzidos (aminoácidos) foram alinhados, tiveram suas regiões não-alinhadas removidas, foram concatenados e utilizados para a construção de uma filogenia de *Máxima Verossimilhaça,* com modelos de substituição calculados automaticamente, de maneira independente, para cada gene. Esta análise resultou na árvore filogenômica que pode ser visualizada na Figura 15, abaixo.

Figura 15 — Árvore filogenômica das *Metschnikowia* utilizadas neste trabalho.
Árvore construída utilizando 1317 sequências de proteínas preditas de genes ortólogos de cópia única comuns entre as espécies identificados pelo BUSCO. Árvore construída com o programa RAxML 8.2.10, utilizando 432632 posições, com 2,63% de gaps ou posições não definidas. Apenas os valores de bootstrap diferentes de 100 estão representados.
MEPq – Metschnikowia de esporo pequeno; MEGd – Metschnikowia de esporo grande.



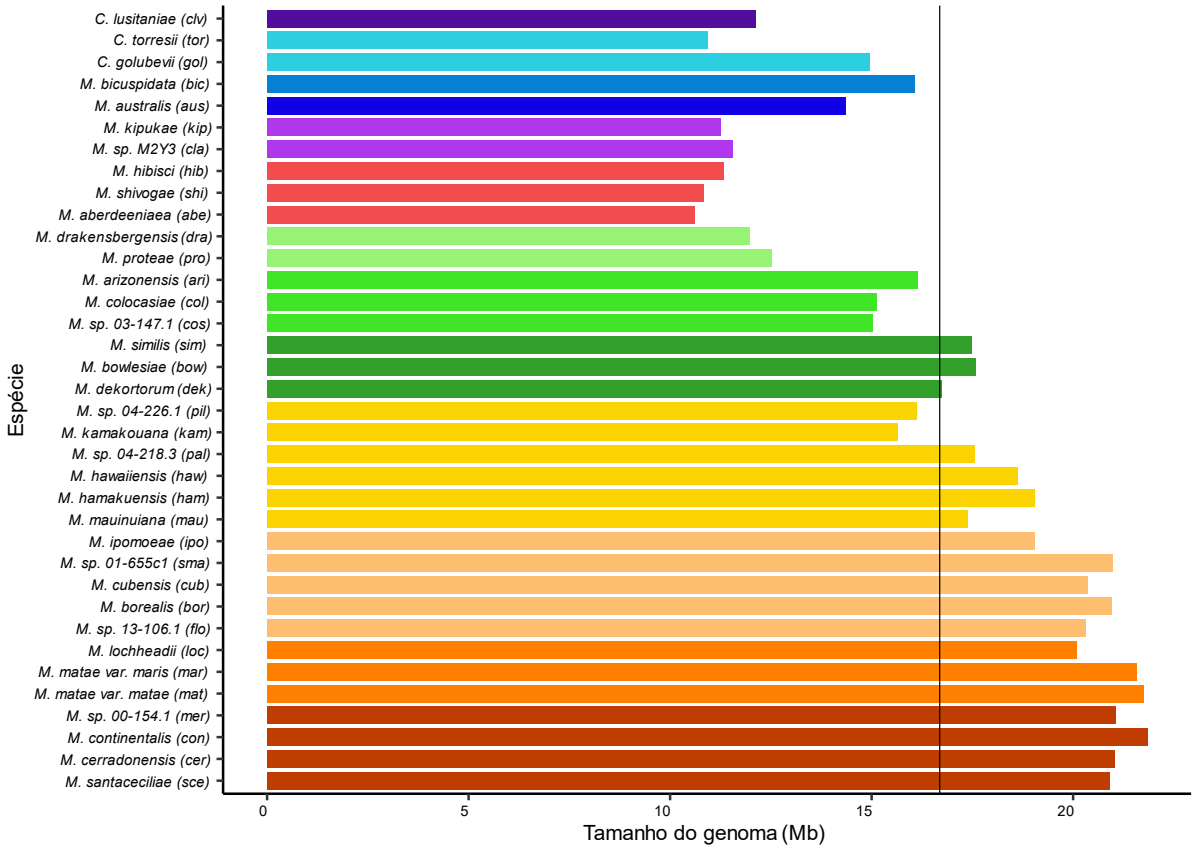
Fonte: Elaborado pelo autor.

4.5 Predição, anotação, e caracterização dos genomas utilizados neste estudo

4.5.1 Tamanho dos genomas

Na Figura 16 observa-se que clados mais derivados das ***M*EGd** (de *M. similis* até *M. santaceciliae*). e em especial as ***M*EGd *sensu strictu***, apresentam os maiores genomas do conjunto estudado. Os representantes dos clados de *M. kipukae* (*M. kipukae* e *M.* sp M2Y3) e *M. hibisci* (*M. hibisci*, *M. shivogae*, *M. aberdeeniaea*), apresentam os menores tamanhos de genomas, semelhantes ao genoma de *M. torresii*, que corresponde a um clado irmão das ***M*EPq**. *M. golubevii*, uma ***M*EPq florícola**, apresenta genoma com tamanho intermediário a *M. australis* e *M. bicuspidata*, ***M*EPq aquáticas**.

Figura 16 — Tamanho dos genomas utilizados neste estudo.
Os valores foram obtidos a partir da análise dos *contigs* com o script *scaffold_stats.pl*, do *pipeline* de anotação de Kumar. S. (https://github.com/sujaikumar/assemblage). Linha vertical: média dos genomas.
Mb: megabases.

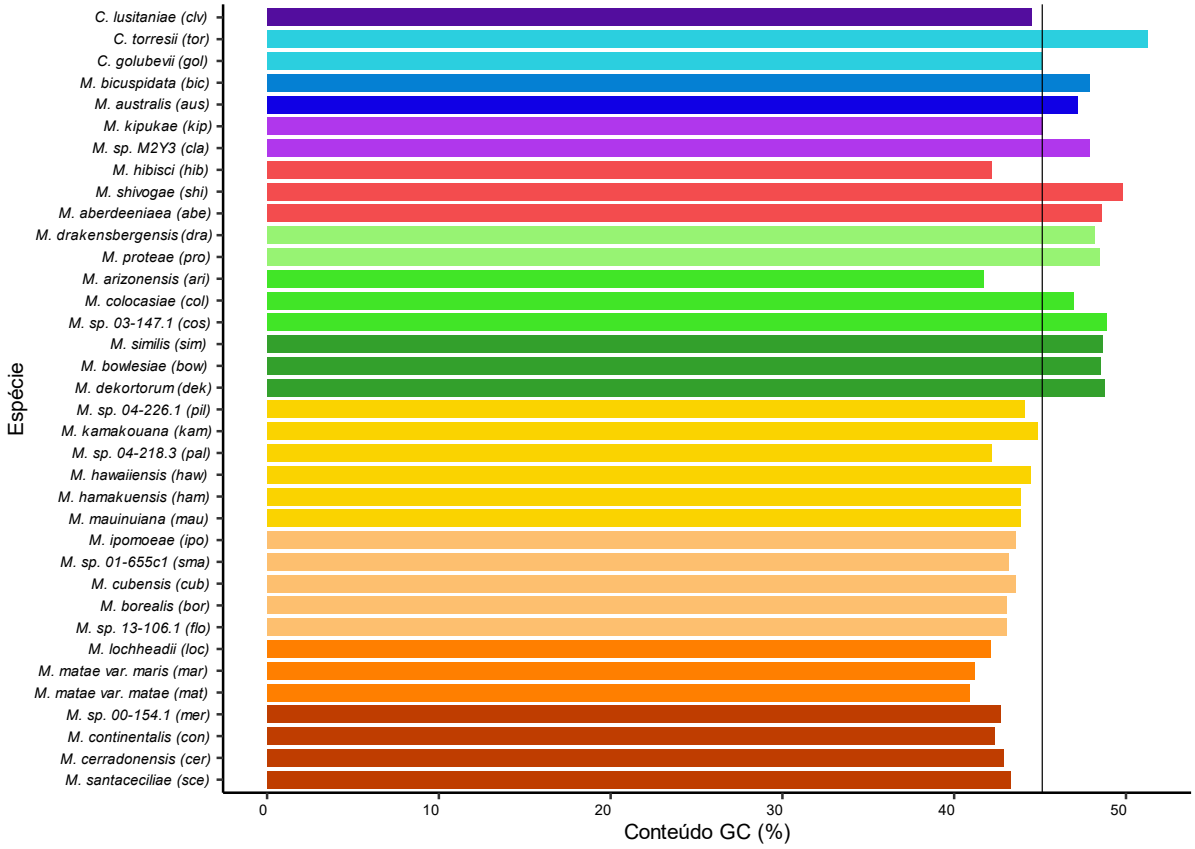


Fonte: Elaborado pelo autor.

4.5.2 Conteúdo GC dos genomas

Na Figura 17 observa-se que o conteúdo GC é bem variável entre os genomas estudados, mas mantém-se relativamente constante para todo o clado das **MEGd *sensu strictu*.** Considerando todos genomas, há pouca variação. O conteúdo GC de *M. australis* e *M. bicuspidata* são praticamente idênticos.

Figura 17 — Conteúdos GC dos genomas utilizados neste estudo
Os valores foram obtidos a partir da análise dos *contigs* com o script *scaffold_stats.pl*, do *pipeline* de anotação de Kumar. S. (https://github.com/sujaikumar/assemblage). Linha vertical: média dos genomas.

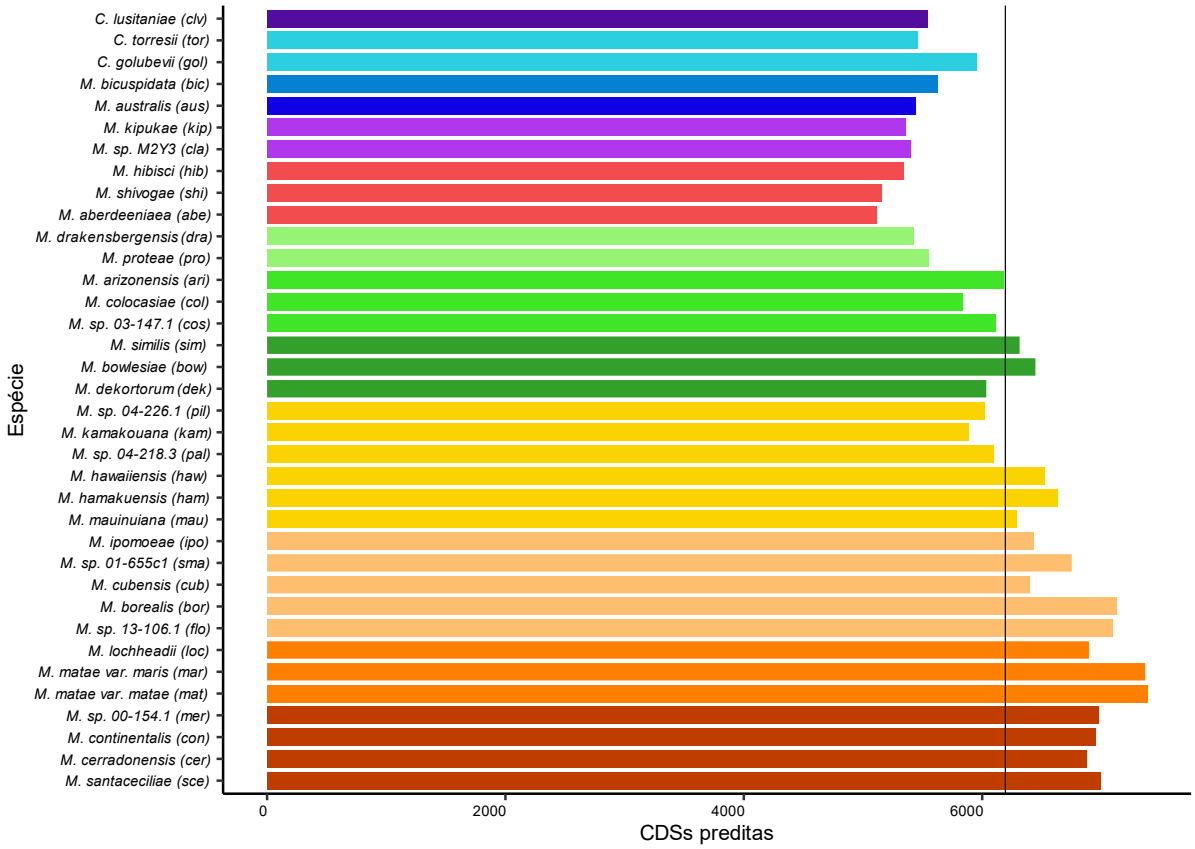


Fonte: Elaborado pelo autor.

4.5.3 Quantidade de CDSs preditas para os genomas

Na Figura 18 observa-se que os clados mais derivados das ***M*EGd** apresentam mais CDSs preditas que as demais espécies, correlacionando com o maior tamanho destes genomas. O clado das ***M*EPq aquáticas** apresenta quantidade de CDSs semelhante aos clados menos derivados das ***M*EGd**. *M golubevii*, representante das ***M*EPq florícolas**, apresenta mais CDSs que os genomas do seu clado mais próximo, as ***M*EPq aquáticas**.

Figura 18 — Número de CDSs preditas para os genomas utilizados neste estudo.
As CDSs foram preditas *de novo* a partir dos *contigs* utilizando o programa MAKER2 em duas iterações, com HMMs construídos para cada genoma com o programa GeneMark, com o SNAP utilizando ortólogos identificados pelo CEGMA, e utilizando AUGUSTUS com espécime modelo treinada *in house* a partir de dados genômicos e de RNAseq de *Clavispora lusitaniae*. CDSs: *coding sequences* (sequências codificadoras); Linha vertical: média dos genomas.



Fonte: Elaborado pelo autor.

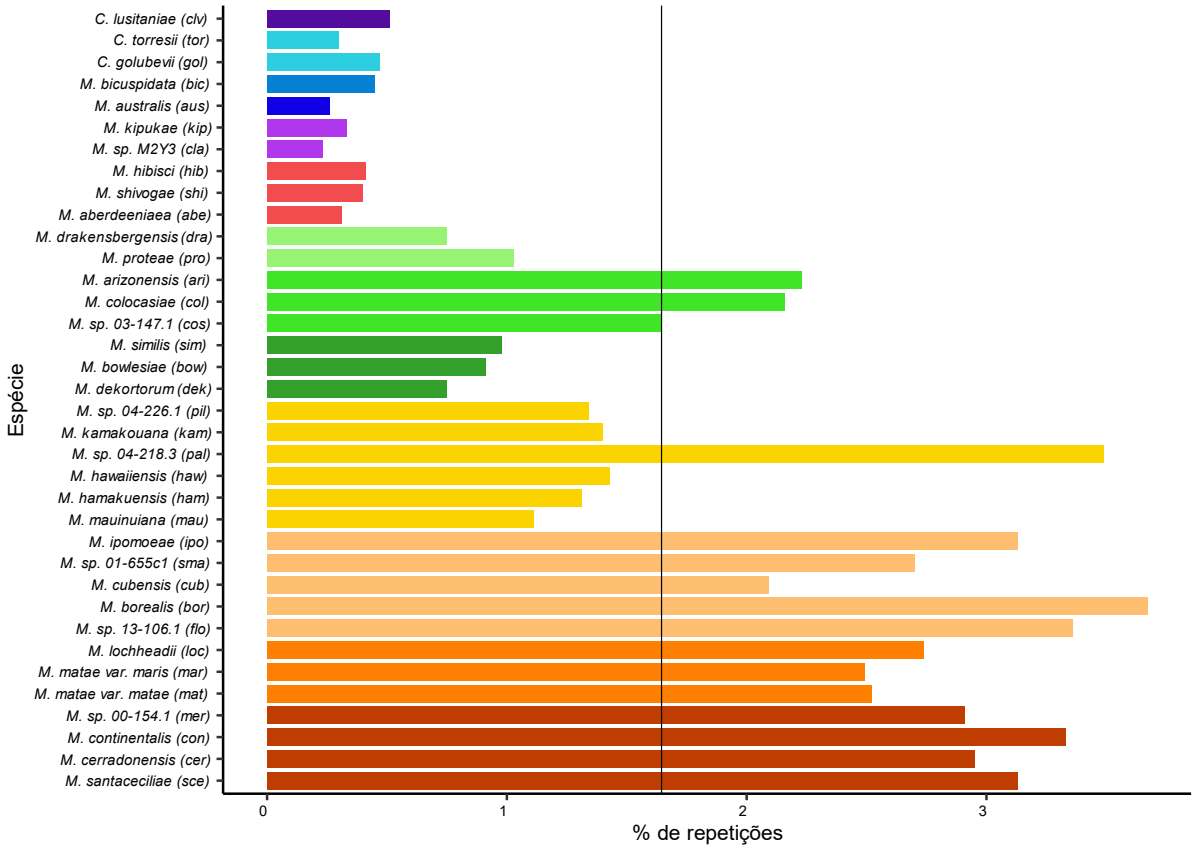
4.5.4 Conteúdo de repetições dos genomas

Na Figura 19 está representada a porcentagem de repetições por genoma. Há grande variação no conteúdo de repetições dos genomas, mesmo ao se comparar organismos próximos. Observa-se, no entanto, que as ***M*EGd,** a partir do clado contendo *M. arizonensis*, apresentam uma expansão no conteúdo de repetições. Essa expansão é ainda maior nas ***M*EGd *sensu strictu***.

4.5.5 Densidade gênica dos genomas

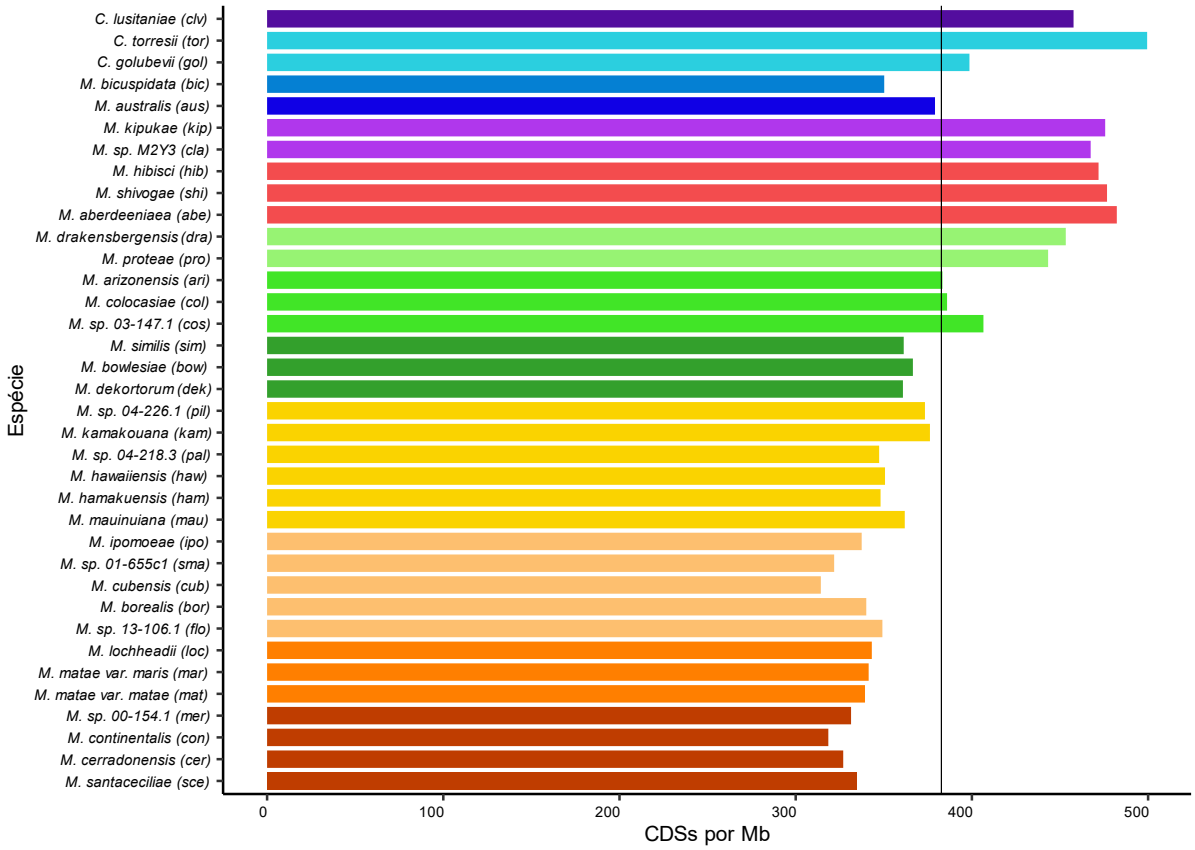
A densidade gênica possibilita uma melhor comparação dos organismos, uma vez que normaliza o conteúdo de genes pelo tamanho dos genomas. Observa-se na Figura 20 que os clados mais derivados das ***M*EGd** apresentam os genomas menos densos, possivelmente por possuírem um maior conteúdo de repetições, como podemos ver na Figura 19. Apesar de possuírem poucas repetições, *M. australis* e *M. bicuspidata* têm genomas de densidade intermediária, mas, interessantemente, *M. australis* apresenta um genoma mais compacto que o de sua espécie mais próxima.

Figura 19 — Conteúdo de Repetições para os genomas estudados.
Porcentagem de repetições, retroelementos, transposons e outras sequências identificadas pelo programa RepeatMasker a partir dos *contigs* de cada genoma, utilizando o *dataset Fungi* como referência. Linha vertical: média dos genomas.



Fonte: Elaborado pelo autor.

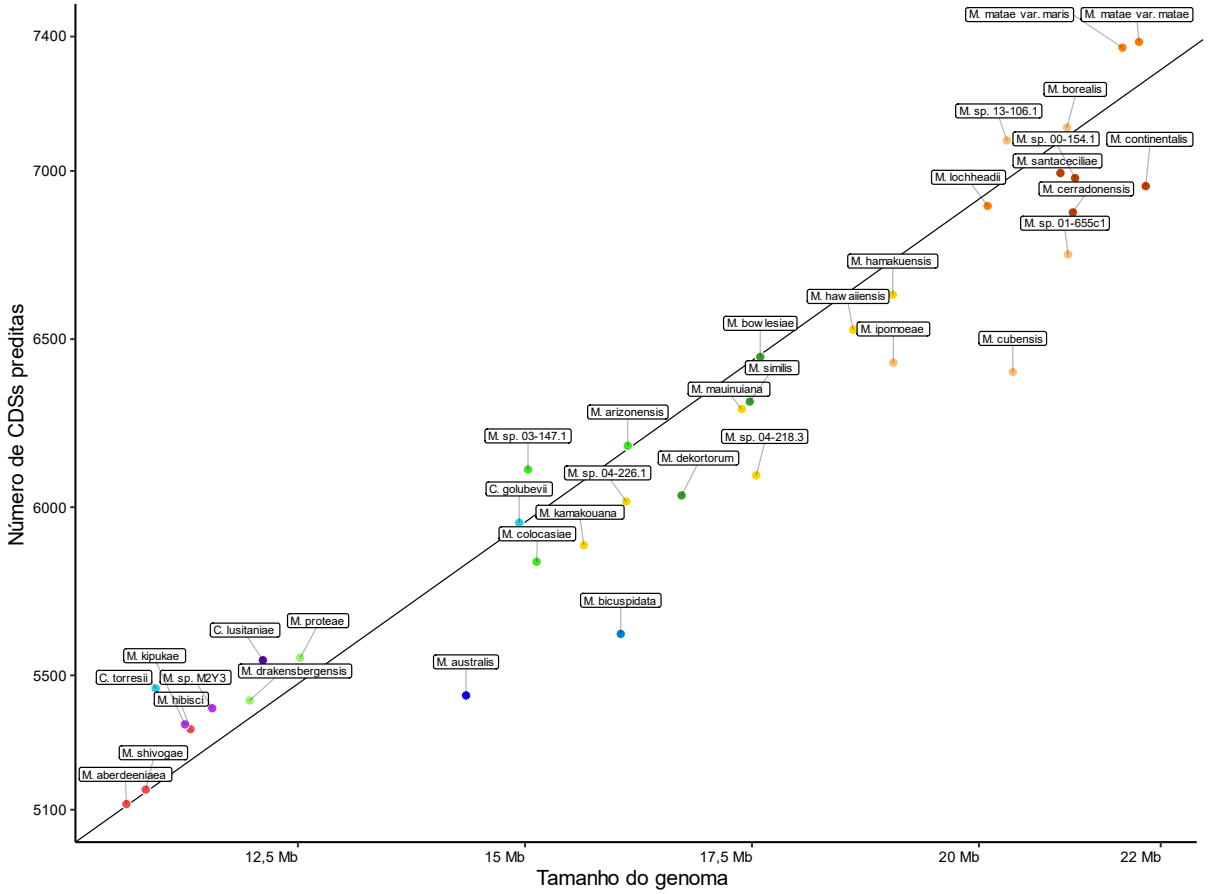
Figura 20 — Densidade Gênica dos genomas estudados.
Densidade gênica calculada com a divisão do número de CDSs preditas pelo tamanho do genoma.
CDSs: *coding sequences* (sequências codificadoras); Mb: mega bases. Linha vertical: média dos genomas.



Fonte: Elaborado pelo autor.

A Figura 21 mostra uma outra possível visualização da densidade gênica dos genomas estudados. Percebe-se que *M. australis* e *M. bicuspidata* se posicionam levemente fora da linha de tendência da dispersão, com genomas menos densos que os das demais leveduras. Além disso, mesmo possuindo um genoma maior, *M. bicuspidata* tem uma proporção de CDSs semelhante a *M. australis*. A levedura *M. cubensis* apresenta esta mesma característica.

Figura 21 — Dispersão do tamanho dos genomas e número de CDSs preditas.
CDSs: *coding sequences* (sequências codificadoras); Mb: mega bases. Linha diagonal: tendência da relação entre Número de CDSs e Tamanho do genoma.

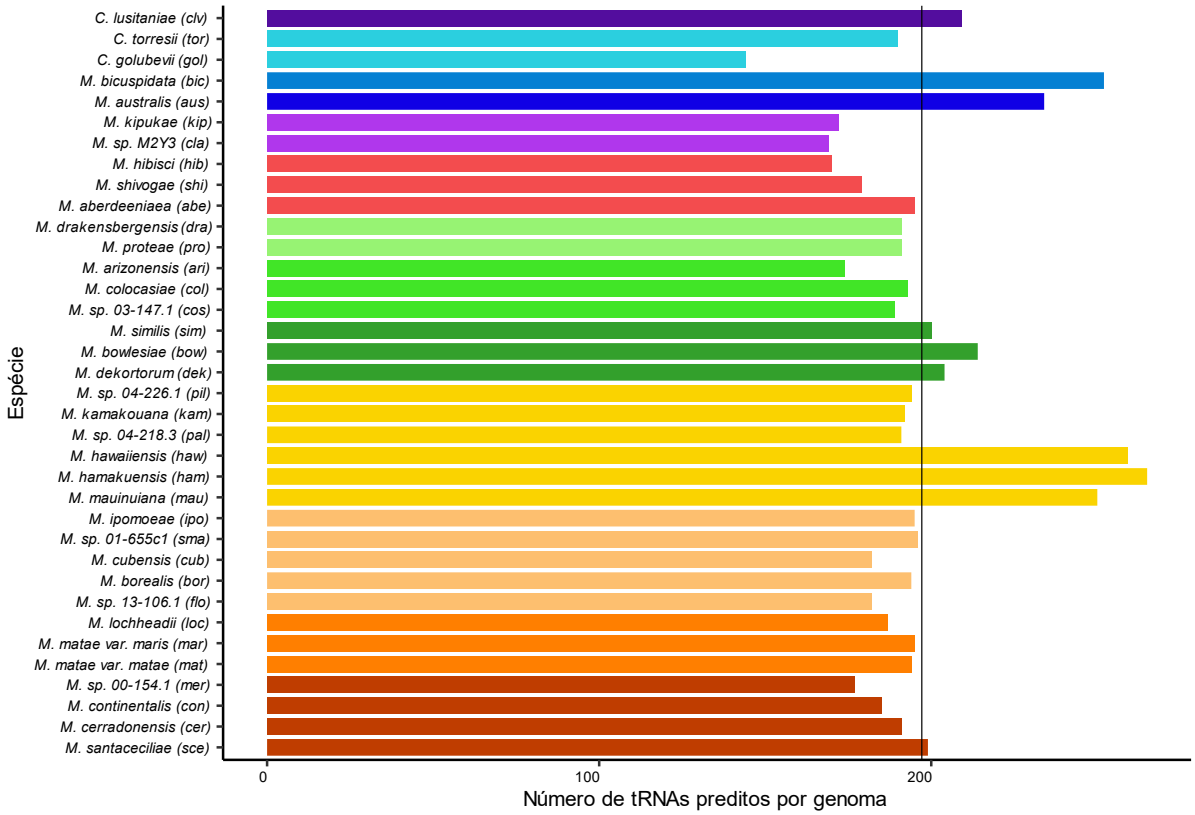


Fonte: Elaborado pelo autor.

4.5.6 Número de tRNAs preditos nos genomas

A Figura 22 mostra que o número de tRNAs preditos é semelhante na maioria das espécies, com exceção do clado composto por *M. hawaiiensis*, *M. hamakuensis* e *M. mauinuiana*, e do clado das ***M*EPq aquáticas**, composto por *M. australis* e *M. bicuspidata*. A levedura *M. golubevii* se destaca por apresentar um número de tRNAs bem inferior à média das demais, especialmente quando comparada com *M. australis* e *M. bicuspidata*, as espécies mais próximas neste conjunto de genomas.

Figura 22 — Número de tRNAs preditos para os genomas estudados
Número de tRNAs identificados nos genomas utilizando o programa tRNAscan-SE v.2.0.2, com os padrões *default* para a busca por tRNAs eucarióticos. Linha vertical: média dos genomas.



Fonte: Elaborado pelo autor.

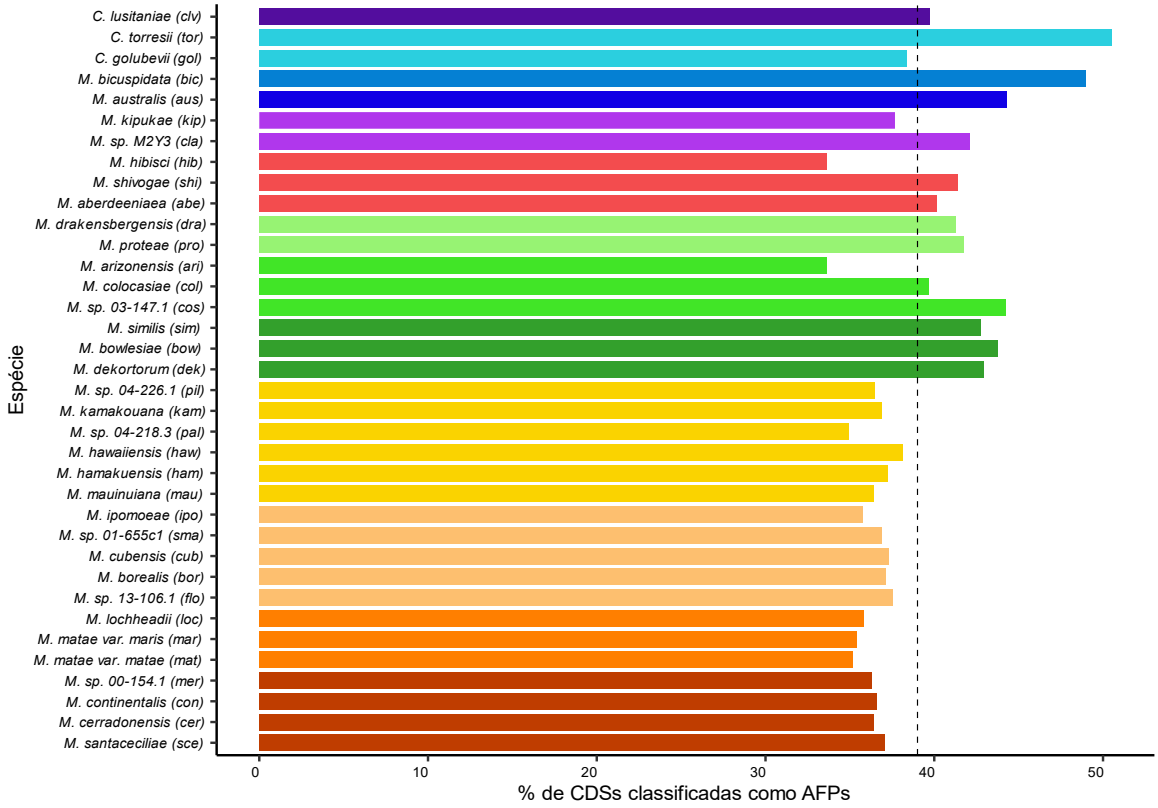
4.6 Classificação das CDSs preditas por Classificadores de AFPs

Para avaliação da eficiência dos Classificadores utilizados, todas as CDSs de todos os genomas deste estudo foram submetidas aos classificadores **RAFP-pred** e **CryoProtect**.

4.6.1 CDSs classificadas como AFPs pelo RAFP-pred

Na Figura 23, observa-se que este programa classificou em torno de 38% de todas CDSs preditas de todos os genomas como AFPs. O clado das ***M*EGd *strictu sensu*** apresentou uma proporção inferior à média. A grande proporção de classificação em todos os genomas indica uma inconsistência do classificador, uma vez que AFPs tendem a estar representadas nos genomas em baixíssimos números e, além disso, a maioria destas espécies têm distribuição tropical.

Figura 23 — Porcentagem das CDSs classificadas como AFPs nos genomas pelo RAFP-pred. Proporção obtida pela classificação de todas CDSs de todos os genomas estudados. Linha vertical: média dos genomas.

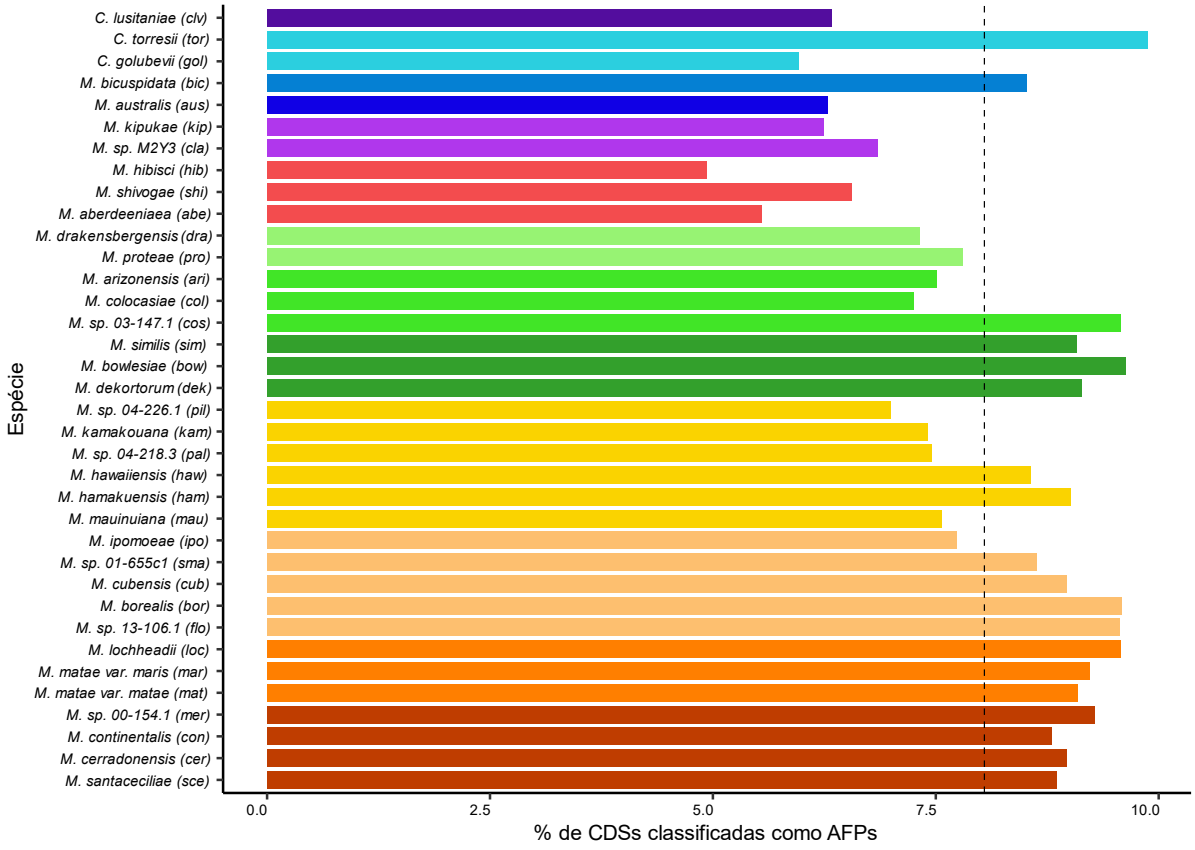


Fonte: Elaborado pelo autor.

4.6.2 CDSs classificadas como AFPs pelo CryoProtect

A Figura 24 mostra que este programa apresentou uma proporção bem menor de CDSs classificadas como AFPs, em média 8%. Estes resultados são mais consistentes com a proporção de AFPs esperada para um genoma, demonstrando que este programa é mais rigoroso. Entre as ***M*EGd** observou-se maior proporção de CDSs classificadas como AFPs, apesar destas espécies serem predominantemente de clima tropical e distribuição equatorial. Quanto às ***M*EPq**, *M. torresii* e *M. bicuspidata* apresentaram de 20 a 30 % mais CDSs classificadas como AFPs que *M. australis*.

Figura 24 — Proporção de CDSs classificadas como AFPs nos genomas pelo CryoProtect. Proporção obtida pela classificação de todas CDSs de todos os genomas estudados. Linha vertical: média dos genomas

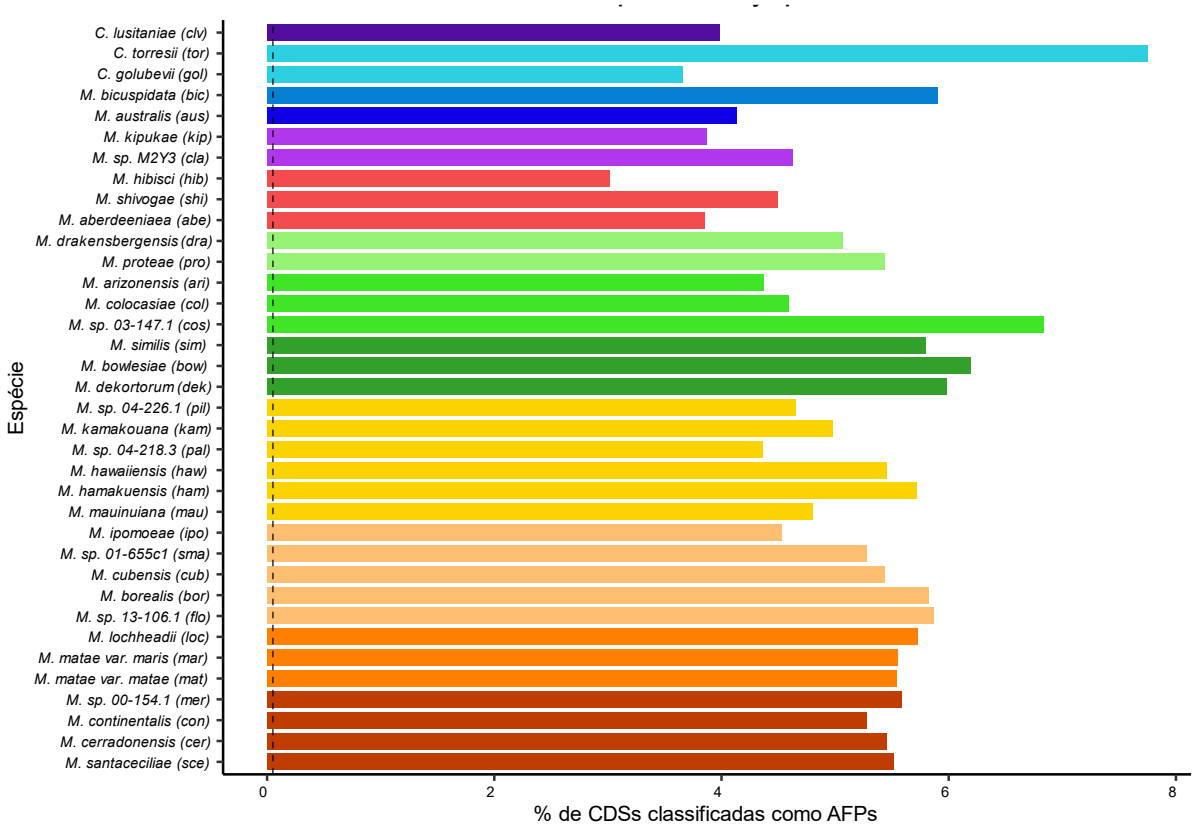


Fonte: Elaborado pelo autor.

4.6.3 CDSs classificadas como AFPs por RAFP-pred & CryoProtect

Na Figura 25, podemos observar que quando avaliamos quais CDSs são classificadas como AFPs pelos dois classificadores, uma vez observado que o *RAFP-pred* é menos rigoroso, sua classificação teve menor influência no perfil geral, o qual se assemelha mais ao perfil de predição do *CryoProtect*. Este resultado demonstra que grande parte das CDSs classificadas como AFPs pelo *CryoProtect* foram assim classificadas pelo primeiro, ou seja, cerca de 63%. Observa-se que *M. torresii* teve mais CDSs com classificação concordante pelos dois programas, apresentando-se ainda mais distante da média nesta comparação.

Figura 25 — Proporção de CDSs consensualmente classificadas como AFPs pelo RAFP-pred e pelo CryoProtect
Porcentagem das CDSs classificadas como AFPs independentemente pelos dois programas.

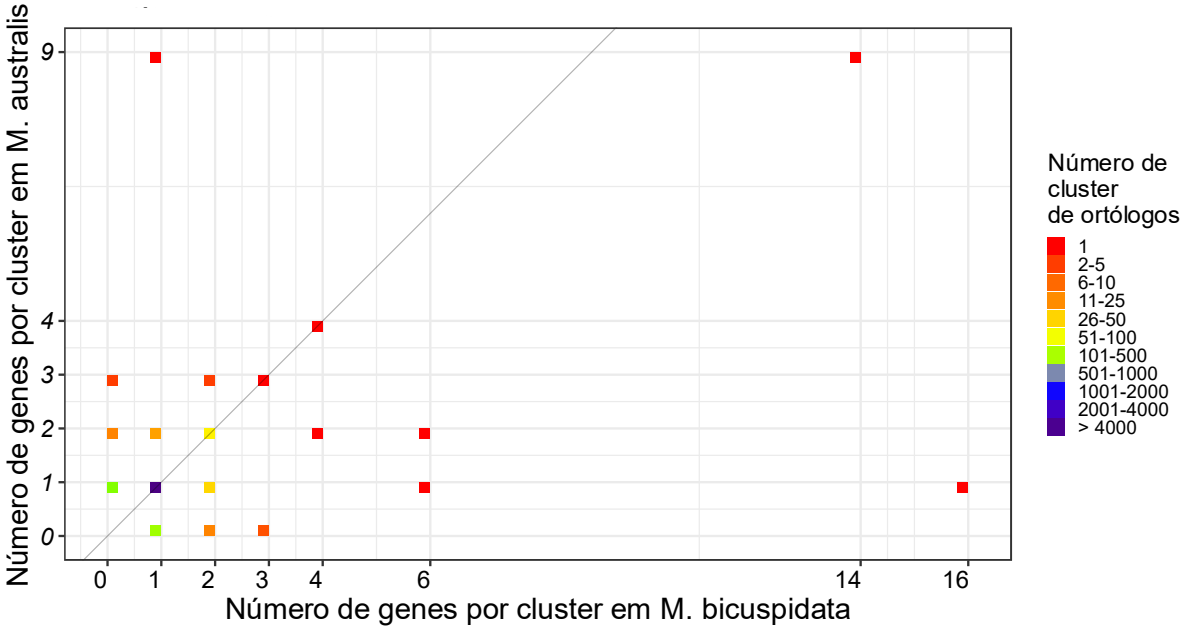


Fonte: Elaborado pelo autor.

4.7 Identificação dos homólogos utilizando o orthoMCL

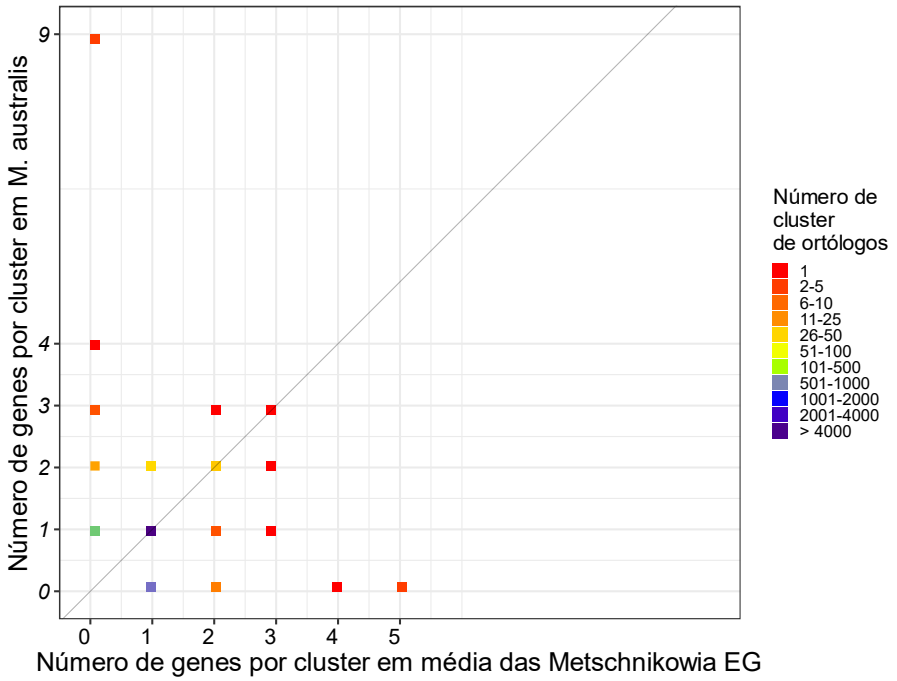
O programa **orthoMCL** encontrou 9380 *clusters* de genes entre os genomas analisados. Destes, 3297 correspondem a genes ortólogos de cópia única presentes em todos os 36 genomas. Os *clusters* com variações no número de cópias nos genomas de *M. australis*, *M. bicuspidata* e demais *Metschnikowia* estão relacionados na Figuras 26, Figura 27 e Tabela 5.

Figura 26 — Relação dos *clusters* gênicos compartilhados entre os genomas de *M. australis* e *M. bicuspidata*
Clusters de genes parálogos identificados pelo programa OrthoMCL. Os *clusters* estão plotados em função do número de cópias que apresentam nos genomas comparados.
(A-H) *Clusters* em destaque com as descrições na Tabela 5, a seguir. Linha diagonal: proporção correspondente a um número de cópias igual nos genomas comparados.



Fonte: Elaborado pelo autor.

Figura 27 — Relação do número de cópias para *clusters* compartilhados entre os genomas de *M. australis* e demais *Metschnikowia* exceto *M. bicuspidata*.
Clusters de genes parálogos identificados pelo programa OrthoMCL. Os *clusters* estão plotados em função do número de cópias que apresentam nos genomas comparados.
(A-H) *Clusters* em destaque com as descrições na Tabela 5, a seguir. Linha diagonal: proporção correspondente a um número de cópias igual nos genomas comparados.



Fonte: Elaborado pelo autor.

Tabela 5 — Anotações para *clusters* gênicos com diferença no número de cópias, destacados nas figuras 26 e 27.

Índice	cluster	Número de genes por *cluster*			Anotação	Códigos
		australis	bicuspidata	Outras *Metschnikowia*		
A	5505	9	1	1	Adenoviral fibre protein; Pectin lyase fold/virulence factor; Pectin lyase-like; consensus disorder prediction;	PF00608; IPR000939; IPR011050; SSF51126; GO:0019062; PTHR37917
B	5184	9	14	0	consensus disorder prediction;	-
C	5330	1	16	0	NO ANNOTATION	-
D	5733	2	6	0	consensus disorder prediction;	-
E	**6592**	**3**	**0**	**0**	**NO ANNOTATION**	-
F	6561	3	0	0	Hyphally regulated cell wall protein N-terminal; consensus disorder prediction;	PF11765
G	5732	4	4	0	consensus disorder prediction; coil;	-
H	7324	2	0	0	Endonuclease/exonuclease/phosphatase superfamily	IPR036691
	7357	2	0	0	NO ANNOTATION	-
	7393	2	0	0	Hyphally-regulated cell wall protein, N-terminal	IPR021031
	7404	2	0	0	NO ANNOTATION	-
	7407	2	0	0	NAD-dependent epimerase/dehydratase	PF01370
	7408	2	0	0	Hyphally-regulated cell wall protein	PF11765
	7409	2	0	0	NO ANNOTATION	-
	7427	2	0	0	consensus disorder prediction	-

Fonte: Elaborado pelo autor.

Dos *clusters* gênicos encontrados pelo **orthoMCL**, observa-se que 10 são compostos apenas por proteínas de *M. australis*. Entre todos *clusters* preditos, 118 são compartilhados exclusivamente por *M. australis* e *M. bicuspidata*, sendo 10 com múltiplas cópias e 108 com uma cópia em cada genoma. Destes, 22 não puderam ser anotados, e os 86 remanescentes foram relacionados pelo programa **InterProScan** às ontologias descritas na Tabelas 6 e Tabela 7:

Tabela 6 — Processos biológicos relacionados aos 86 *clusters* de cópia única, exclusivos de *M. australis* e *M. bicuspidata*.

GO	Processo Biológico
GO:0007618	*mating*
GO:0030001	*metal ion transport*
GO:0046294	*formaldehyde catabolic process*
GO:0006629	*lipid metabolic process*
GO:0006487	*protein N-linked glycosylation*
GO:0006694	*steroid biosynthetic process*
GO:0006508	*proteolysis*
GO:0055114	*oxidation-reduction process*
GO:0006355	*regulation of transcription, DNA-templated*
GO:0055085	*transmembrane transport*
GO:0071577	*zinc II ion transmembrane transport*
GO:0034755	*iron ion transmembrane transport*
GO:0035434	*copper ion transmembrane transport*
GO:0006865	*amino acid transport*
GO:0000750	*pheromone-dependent signal transduction involved in conjugation with cellular fusion*

Fonte: Elaborado pelo autor.

Tabela 7 — Funções moleculares relacionados aos 86 *clusters* de cópia única exclusivos de *M. australis* e *M. bicuspidata*.

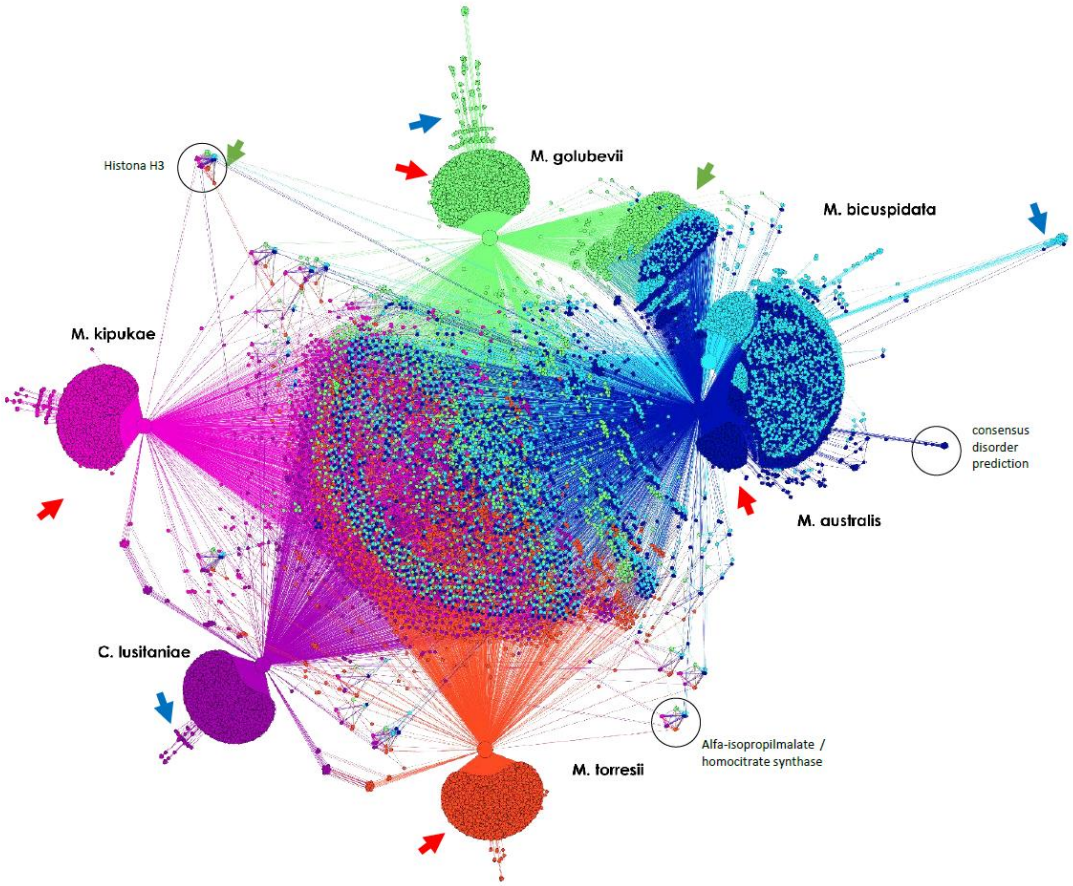
GO	Função molecular
GO:0003824	*catalytic activity*
GO:0003676	*nucleic acid binding*
GO:0016491	*oxidoreductase activity*
GO:0022857	*transmembrane transporter activity*
GO:0008270	*zinc ion binding*
GO:0050662	*coenzyme binding*
GO:0000287	*magnesium ion binding*
GO:0005506	*iron ion binding*
GO:0016616	*oxidoreductase activity, acting on the CH-OH*
GO:0020037	*heme binding*
GO:0016705	*oxidoreductase activity, acting on paired*
GO:0046873	*metal ion transmembrane transporter activity*
GO:0004497	*monooxygenase activity*
GO:0000981	*RNA polymerase II transcription factor*
GO:0016831	*carboxy-lyase activity*
GO:0004190	*aspartic-type endopeptidase activity*
GO:0030976	*thiamine pyrophosphate binding*
GO:0004185	*serine-type carboxypeptidase activity*
GO:0005381	*iron ion transmembrane transporter activity*
GO:0000030	*mannosyltransferase activity*
GO:0005375	*copper ion transmembrane transporter activity*
GO:0005385	*zinc ion transmembrane transporter activity*
GO:0003854	*3-beta-hydroxy-delta5-steroid dehydrogenase*
GO:0018738	*S-formylglutathione hydrolase activity*
GO:0000772	*mating pheromone activity*

Fonte: Elaborado pelo autor.

4.8 Construção de uma rede de similaridade para visualização das relações entre os genomas

As relações entre genes ortólogos e parálogos são de difícil visualização quando em conjunto. A figura a seguir exemplifica uma estratégia em desenvolvimento para a visualização de relações de homologia entre os genes de várias espécies simultaneamente, a partir do mesmo *input* utilizado no **orthoMCL**. No grafo a seguir estão representadas as relações de similaridade superiores a 80% entre os genes de 6 espécies.

Figura 28 — Rede baseada em similaridade para todos os genes de 6 espécies.
Rede construída a partir de relações de similaridade e cobertura recíproca extraídas do *output* em formato tabular do DIAMOND. Relações extraídas e utilizadas na construção de arquivos de *input* para o GEPHI, através de um *script* em R. Esta rede foi construída utilizando valores de similaridade superiores a 80% e cobertura recíproca superior a 70%.
Setas azuis: expansões parálogas; Setas vermelhas: genes não relacionados por apresentarem similaridade ou cobertura abaixo dos *cutoffs*; Setas verdes: conjuntos de genes compartilhados entre clados; Círculos: anotações selecionadas para exemplo.



Fonte: Elaborado pelo autor.

Nesta rede é possível observar um aglomerado central, que corresponde à grande maioria dos genes conservados entre os genomas e de cópia única. Alguns genes com maior similaridade e de múltiplas cópias correlacionadas entre os genomas se projetam para o exterior da rede, como é o caso da Histona H3, do conjunto destacado pertencente a *M. australis* anotado como *consensus disorder* prediction, e da enzima alfaisopropilmalato homocitrato sintase. Também podemos visualizar expansões parálogas, tanto exclusivas quanto compartilhadas entre dois ou mais genomas.

4.9 Estimativa do tempo de divergência das espécies *M. australis e M. bicuspidata*

O programa **gKaKs** identificou 2024 pares de genes com seleção neutra (Ka/Ks < 0). Estes genes estão sofrendo alterações na sequência nucleotídica que não interferem na sequência de aminoácidos, e por isso podem ser utilizados para acompanhar o acúmulo de mutações nas linhagens/espécies. A média destes valores de Ks (taxa de substituição sinônima) foi de 0.3841. Considerando um tempo de dobramento de 2 horas, como de acordo com as curvas a 12 ˚C, e a taxa de mutação estimada para leveduras de 1,67 ± 0.04 × 10^{-10} por base por geração,

$$T = \frac{d}{2r}$$

T: tempo de divergência (número de gerações)

d: número de substituições entre duas sequências (kS médio)

r: taxa de substituição/ mutação estimada para a espécie

$$T = \frac{0{,}3841}{2 * \ 1{,}67 * 10^{-10}}$$

$$T = \frac{0{,}3841}{3{,}34 * 10^{-10}}$$

$$T = 0{,}115 * 10^{10}\ Gerações$$

$$T = 11{,}5 * 10^{8}\ Gerações$$

O tempo de dobramento de 2h equivale a 12 gerações por dia, ou 4.380 gerações por ano. Assim:

$$T(anos) = \frac{11{,}5}{4380} * 10^8$$

$$T(anos) = \mathbf{262.557}$$

Se considerarmos um tempo de dobramento menor, como o de *M. bicuspidata* a 6 ˚C, que é de 20h, o valor de *T(anos)* deve ser multiplicado por um fator de 10x.

$$T(anos) = 262.557 * 10$$

$$T(anos) = \mathbf{2.625.570}$$

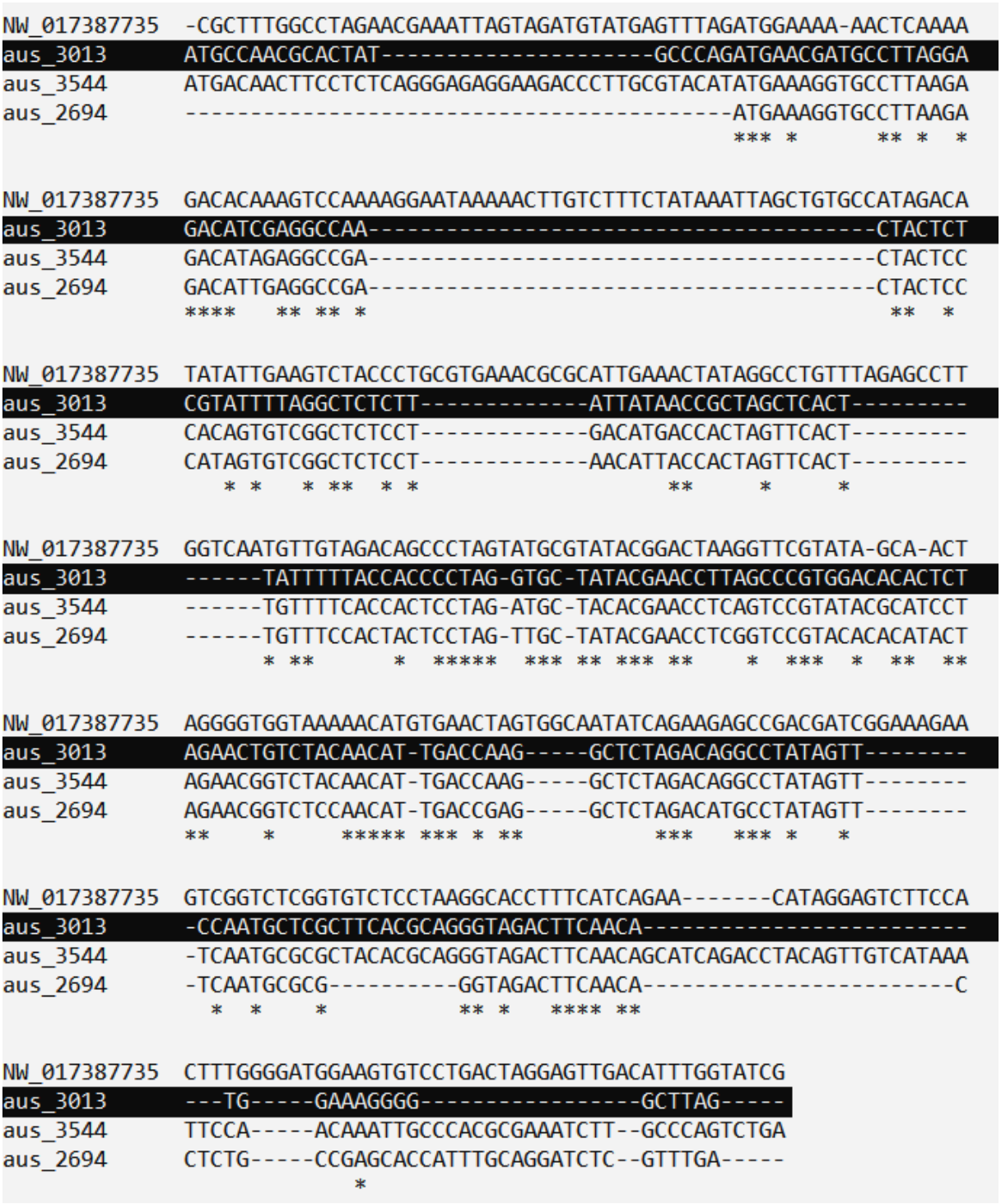
Desta maneira, o intervalo que compreende o possível tempo de divergência entre *M. australis* e *M. bicuspidata* pode ser estimado entre 260 mil e 2,6 milhões de anos.

4.10 Identificação de CDSs exclusivas de *M. australis* ou parcialmente compartilhadas com *M. bicuspidata*

4.10.1 Seleção de CDSs e confecção de iniciadores

Foram encontradas 247 proteínas exclusivas em *M. australis* com relação às demais *Metschnikowia*. Destas, 10 apresentaram alguma similaridade com *M. bicuspidata*. Entretanto, a categorização como exclusiva é subjetiva pois depende dos *cutoffs* de identidade, similaridade e cobertura. Entre estas 10, nenhuma apresentou cobertura da CDS correspondente maior que 40%. A CDS com maior semelhança com um *contig* de *M. bicuspidata* foi **aus_3013**. Esta CDS é também componente de *cluster* de parálogos identificado pelo **orthoMCL**, exclusivo de *M. australis*, junto com as CDSs **aus_3544** e **aus_2694**. O *cluster* 6592 está em destaque na Tabela 5, e não foi possível identificar nenhuma anotação para estas CDSs. Um alinhamento realizado no programa **MUSCLE** das 3 CDSs e da região correspondente do *contig* de *M. bicuspidata* pode ser visualizado na Figura 29.

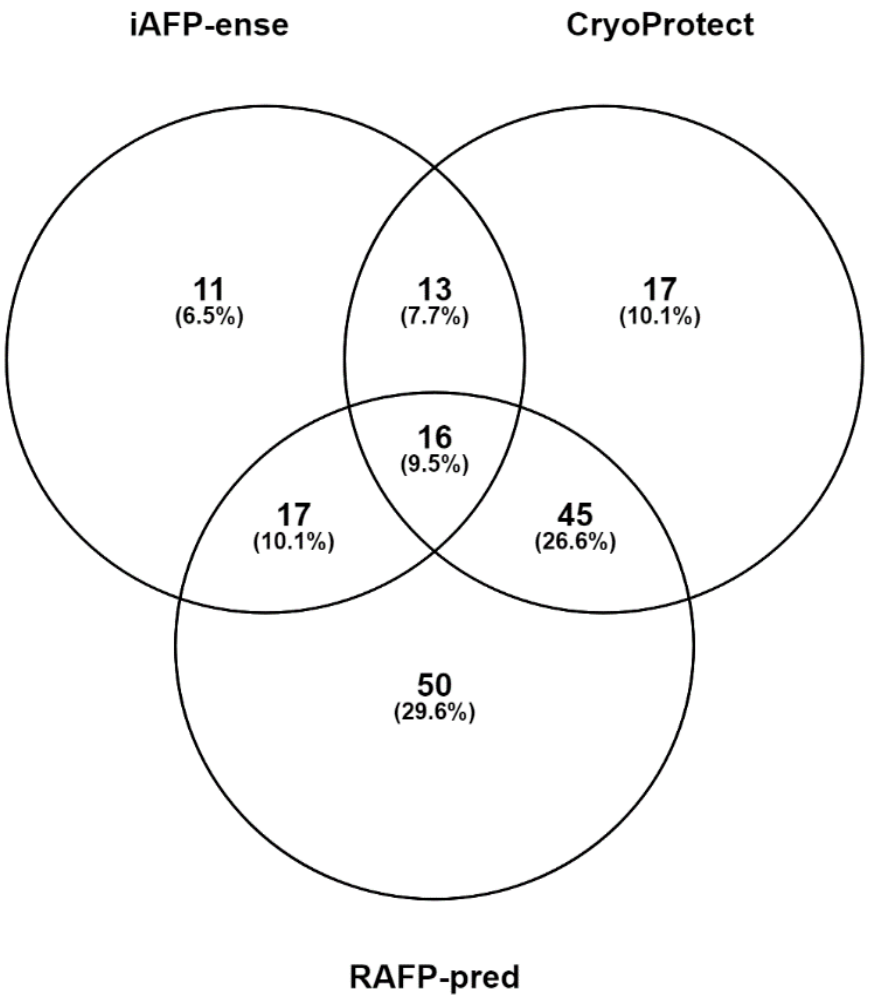
Figura 29 — Alinhamento das CDSs parálogas de *M. australis* e a região correspondente do *contig* de *M. bicuspidata*.
Alinhamento realizado no programa MUSCLE entre as CDSs identificadas como *cluster* parálogo exclusivo de *M. australis* e a região do *contig* de *M. bicuspidata* que apresentou similaridade nas análises por tBLASTn. Em destaque a CDS aus_3013.



Fonte: Elaborado pelo autor.

Após análise pelos 3 classificadores de AFPs (Seção 3.11), 169 CDSs foram classificadas como AFPs por ao menos um dos programas. A Figura 30 mostra a relação entre as classificações. As 16 CDSs que apresentaram consenso foram selecionadas para a avaliação de sua expressão em condições de congelamento.

Figura 30 — Proporção das CDSs exclusivas que foram classificadas como AFPs pelos 3 classificadores utilizados. Diagrama de Venn representando as 169 das 247 CDSs exclusivas de *M. australis* que foram classificadas como AFPs pelos programas RAFP-pred, Cryoprotect e iAFP-ense.



Fonte: Elaborado pelo autor.

Estas 16 CDSs selecionadas foram submetidas à predição de domínios transmembrana, sítios de N-glicosilação e presença de peptídeo sinal pelo programa **Protter**, e suas características estão relacionadas na Tabela 8.

Tabela 8 — Características preditas pelo programa Protter para as 16 CDSs selecionadas para análise de expressão

CDS	Tamanho		Predição pelo Protter		
	(AA)	(nt)	Peptídeo Sinal	D. Transmembrana	S. de N-glicosilação
aus_3013	81	246	0	0	2
aus_3157	77	234	0	0	1
aus_3218	59	180	0	0	0
aus_3362	59	180	0	0	0
aus_3391	89	270	0	0	1
aus_3484	59	180	0	0	1
aus_3629	120	363	0	0	1
aus_3748	210	633	0	1	3
aus_3830	79	240	0	0	0
aus_3946	66	201	0	0	0
aus_3951	100	303	0	0	0
aus_3966	105	318	0	0	0
aus_4100	96	291	0	0	0
aus_4351	71	216	0	0	1
aus_4866	189	570	0	0	1
aus_5130	61	212	0	0	1

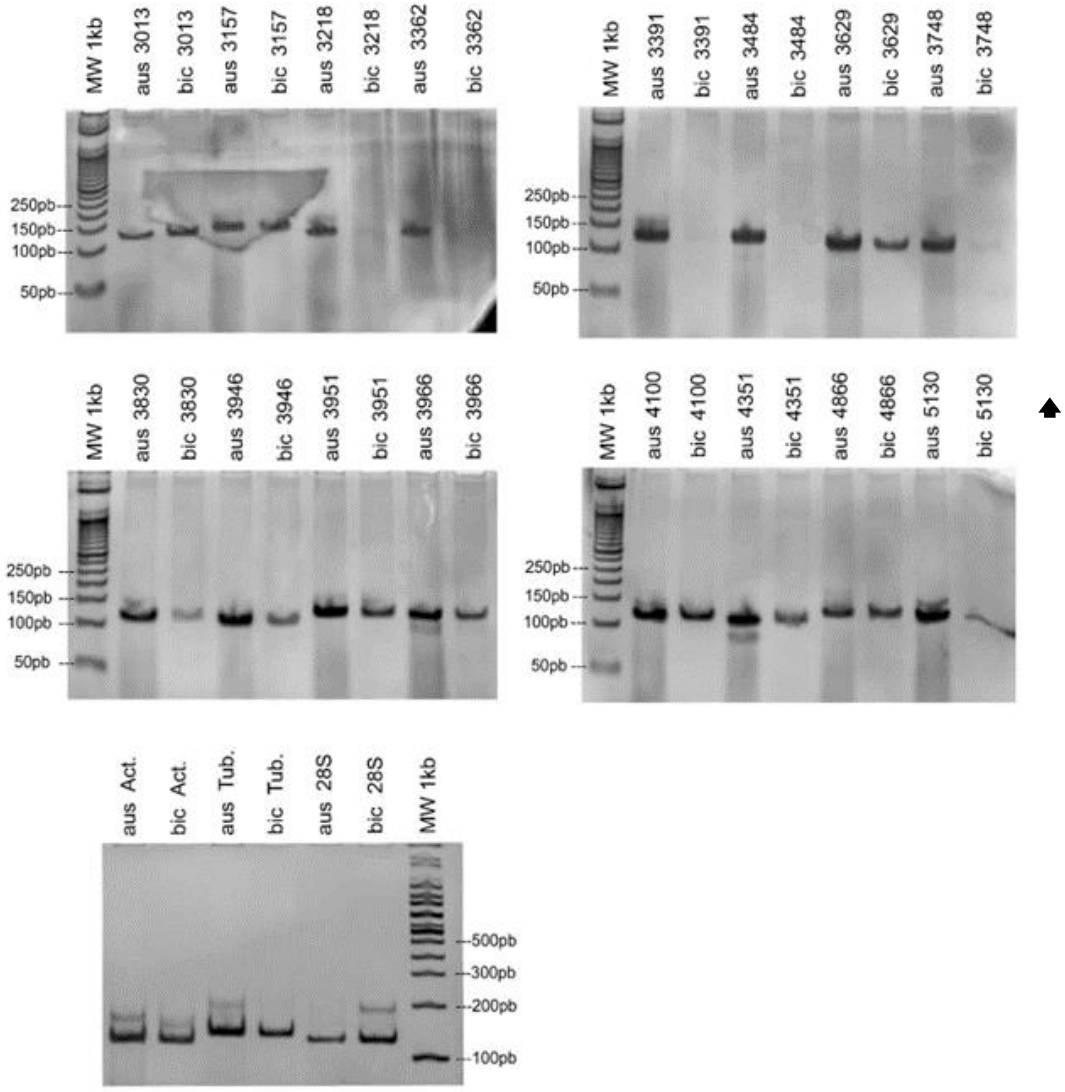
Fonte: Elaborado pelo autor.

4.10 Verificação da expressão das 16 CDSs exclusivas selecionadas

4.10.1 Amplificação controle utilizando o gDNA de M. australis e M. bicuspidata

Observa-se que todos os iniciadores amplificaram o DNA genômico de *M. australis* e 12, diferindo dos 8 esperados, amplificaram o DNA de *M. bicuspidata*. O amplicons destes apresentaram tamanho semelhantes aos de *M. australis*, em contraste com seu tamanho esperado predito pelo **primer-BLAST.**

Figura 31 — Amplificação das CDSs selecionadas a partir do gDNA de M. australis e M. bicuspidata. Géis de Acrilamida 12% corados com Nitrato de Prata com os produtos da PCR do gDNA das leveduras M. australis e M. bicuspidata utilizando os iniciadores desenhados para as 16 CDSs exclusivas selecionadas. MW: Molecular Weight – Padrão de peso molecular; Setas: amplificações fracas.

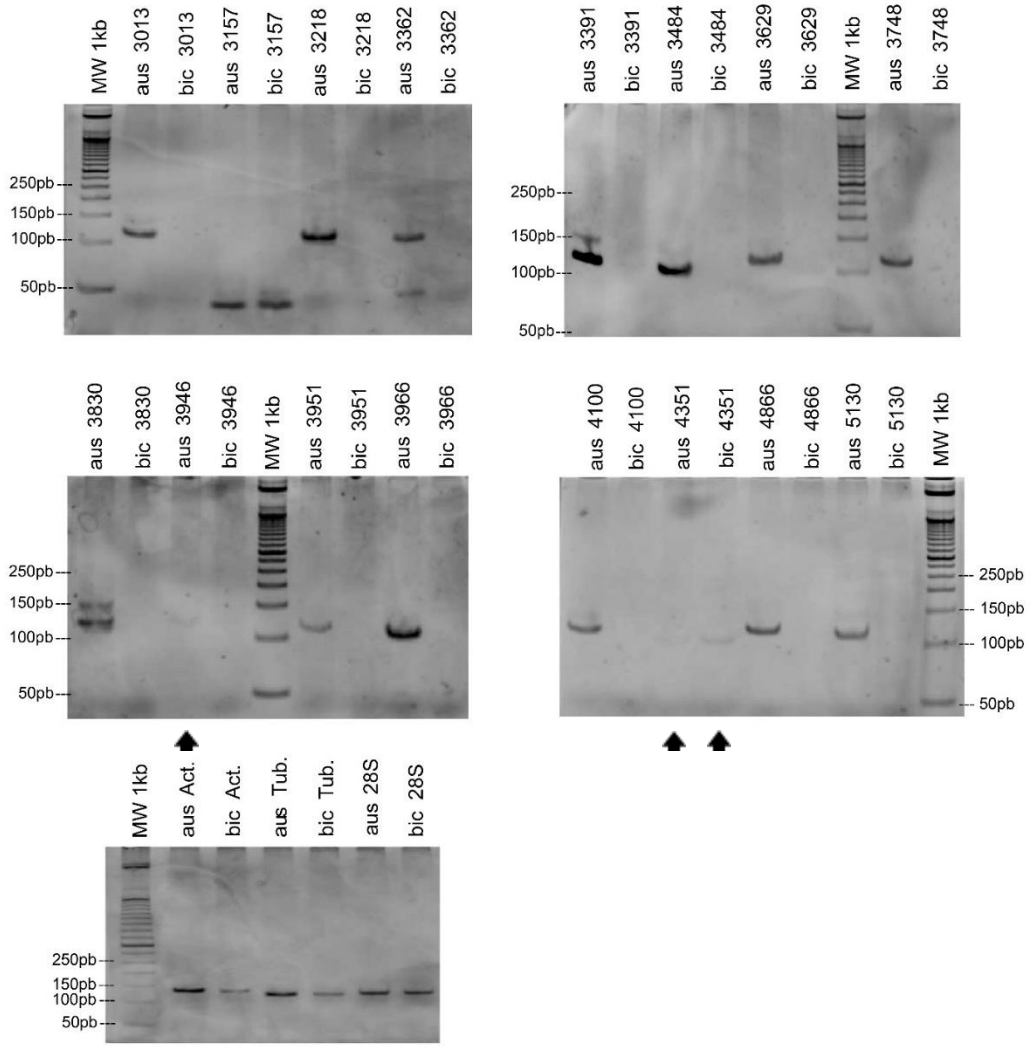


Fonte: Elaborado pelo autor.

4.10.2 Amplificação utilizando o cDNA de M. australis e M. bicuspidata cultivadas a 12˚C

Observa-se que alguns iniciadores que amplificaram no gDNA das leveduras não amplificaram o cDNA, sendo 1 para *M. australis* (aus_4351) e 14 para *M. bicuspidata*. Outros ainda produziram apenas bandas fracas. Os resultados estão sumarizados na **Tabela 9**, após os géis.

Figura 32 — Amplificação das CDSs selecionadas a partir do cDNA de *M. australis* e *M. bicuspidata* extraídos de curvas de crescimento a 12˚C.
Géis de Acrilamida 12% corados com Nitrato de Prata com os produtos da PCR do cDNA das leveduras *M. australis* e *M. bicuspidata* cultivadas a 12˚C, utilizando os iniciadores desenhados para as 16 CDSs exclusivas selecionadas. MW: *Molecular Weight* – Padrão de peso molecular; Setas: amplificações fracas.

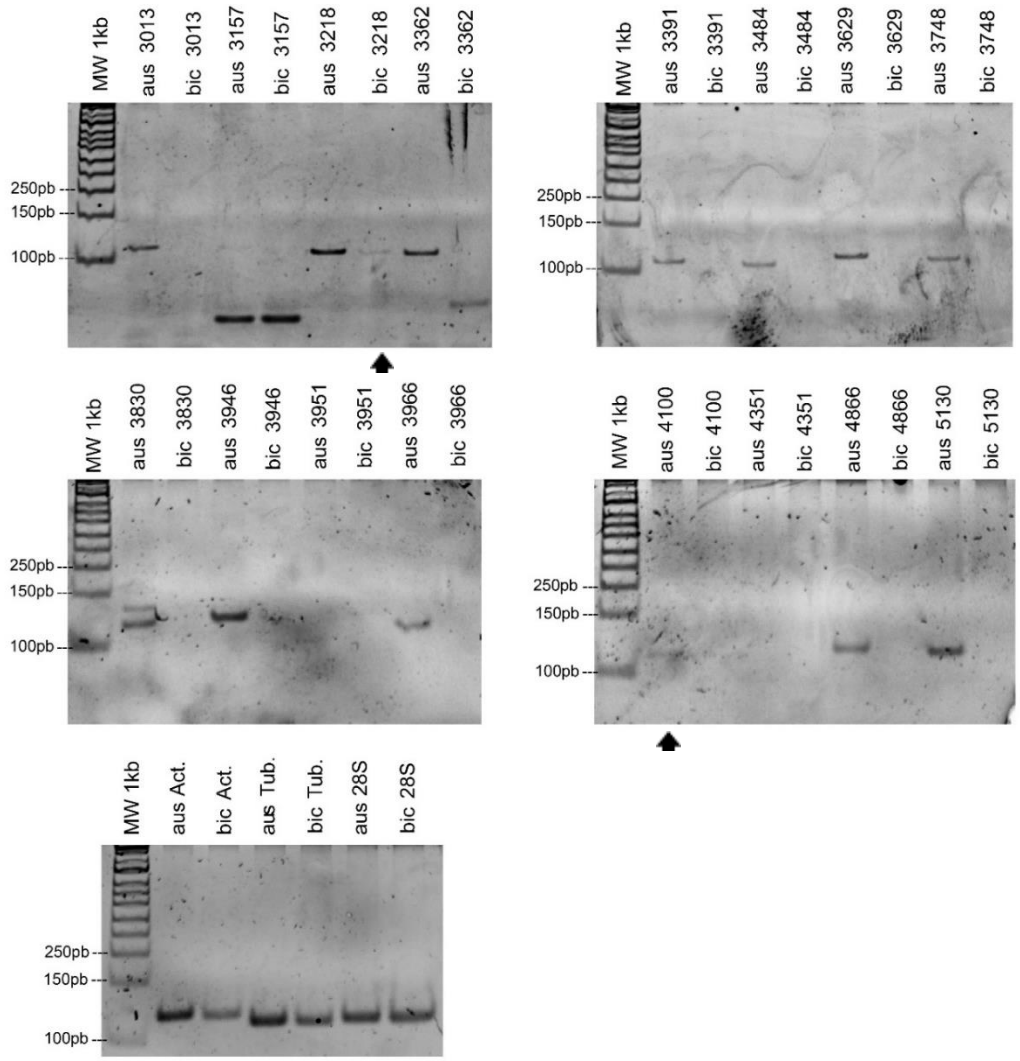


Fonte: Elaborado pelo autor.

4.10.2 Amplificação utilizando o cDNA de M. australis e M. bicuspidata cultivadas a 6˚C

Verificamos na Figura 33 que um par de iniciadores que amplificou em *M. australis* a partir de cultivos a 12˚ não amplificou a 6˚C (aus_3951). O mesmo ocorreu para o par de iniciadores **aus_4351** na amplificação com cDNA de *M. bicuspidata*. Curiosamente, nesta levedura, houveram duas amplificações no cDNA que não ocorreram no gDNA (**aus_3218 e 3362**).

Figura 33 — Amplificação das CDSs selecionadas a partir do cDNA de *M. australis* e *M. bicuspidata* extraídos de curvas de crescimento a 6˚C.
Géis de Acrilamida 12% corados com Nitrato de Prata com os produtos da PCR do cDNA das leveduras.
Setas: amplificações fracas.



Fonte: Elaborado pelo autor.

Nas Figuras 31, 32 e 33 podemos observar que todas amplificações positivas no gDNA de *M. australis* também apresentaram produtos no cDNA. Entretanto, uma delas (**aus_4351**) aparece tão fracamente no gel correspondente ao cultivo a 12˚C que não é possível afirmar que não esteja também no cDNA correspondente ao cultivo a 6 ˚C, e tenha sido amplificada abaixo do limite de detecção.

Apesar de 12 pares de iniciadores terem amplificado o gDNA de *M. bicuspidata*, apenas quatro foram produziram amplificações a partir do cDNA de cultivos a 6 e 12˚C, sendo 2 na primeira temperatura e 3 na segunda. Semelhantemente a *M. australis*, a amplificação correspondente a **aus_4351** gerou uma banda muito fraca no gel correspondente ao cultivo a 12˚C.

Tabela 9 — Amplicons obtidos nas PCRs de gDNA e cDNA de *M. australis* e *M. bicuspidata*.
Quantidade de bandas presentes em cada amplificação.
Números sublinhados: Produtos com tamanho diferente na amplificação de gDNA e cDNA.
Números em negrito: bandas de baixa intensidade.

CDS	Amplicom esperado em		*M. australis*			*M. bicuspidata*		
	M. australis	*M. bicuspidata*	gDNA	cDNA 12˚C	cDNA 6˚C	gDNA	cDNA 12˚C	cDNA 6˚C
Maus_3013	126	126	1	1	1	1	0	0
Maus_3157	128		1	1	1	1	1	1
Maus_3218	117	190	1	1	1	0	0	**1**
Maus_3362	116	744	1	2	1	0	0	1
Maus_3391	119		1	2	1	0	0	0
Maus_3484	111		1	1	1	0	0	0
Maus_3629	129		1	1	1	1	0	0
Maus_3748	122	202	1	1	1	**1**	0	0
Maus_3830	128		1	2	2	1	0	0
Maus_3946	130		1	1	1	1	0	0
Maus_3951	121		1	1	0	1	0	0
Maus_3966	120	618,1835	1	1	1	1	0	0
Maus_4100	120	120	1	1	**1**	1	0	0
Maus_4351	111		1	**1**	0	1	**1**	0
Maus_4866	130	727,764	1	1	1	1	0	0
Maus_5130	125	125	1	2	1	1	0	0
Tubulin	130	130	1	1	1	1	1	1
Actin	119	119	1	1	1	1	1	1
Ribosome 28S	123	123,12	1	1	1	1	1	1

Fonte: Elaborado pelo autor.

5 DISCUSSÃO

5.1 *Metschnikowia* e o frio

5.1.1 O genoma de M. australis

A levedura *M. australis*, isolada pelo MycoAntar, foi selecionada para investigação genômica por sua capacidade de resistir ao congelamento sem a adição de substâncias crioprotetoras e por ser a única representante do gênero *Metschnikowia* na Antártica. Após concluirmos o sequenciamento e montagem do seu genoma, publicado por Batista e colaboradores em 2017, o primeiro procedimento realizado foi a comparação deste com os genomas de outras *Metschnikowia*, disponibilizados pelo Professor Marc-André Lachance, da *University of West Ontario*, Canadá. A esse grupo foi adicionando o genoma de *M. bicuspidata*, a espécie mais próxima de *M. australis*. Nossa montagem se apresentou alta qualidade, contendo 153 *contigs* maiores que 500pb – dos quais 10 representam metade da montagem (L50) – e sem qualquer base ou região indefinida (N). Em comparação, os demais genomas apresentam, em média, 554 *contigs* maiores que 500pb (L50 = 36,8) e 76 regiões indefinidas. Apesar destas diferenças, todos os genomas apresentaram completude superior a 94,7% quando estimada pelo **BUSCO**, mostrando-se um conjunto confiável para as análises posteriores.

A levedura *M. austrais* apresenta um genoma de tamanho intermediário entre o clado menos derivado das ***M*EGd** e o clado das ***M*EGd *Strictu sensu***. Dada a diversidade representada, comparações são mais pertinentes entre espécies próximas. Com relação a *M. bicuspidata*, *M. australis* apresenta um genoma menor, com menos CDSs e tRNAs preditos. *M. australis* também apresenta um conteúdo de repetições 40% menor. Este perfil mais compacto poderia ser uma adaptação ao ambiente extremófilo, fenômeno já observado em outros organismos antárticos (Kelley *et al.*, 2014).

Sob o ponto de vista da análise do conteúdo GC, *M. australis* e *M. bicuspidata* mostram-se semelhantes com 47,2% e 47,8%, respectivamente, apresentando, portanto, uma diferença de apenas 0,6%. Entretanto, quando considerado somente o conteúdo GC das CDSs preditas, observa-se uma diferença de 2,3%, com *M. australis* apresentando GC^{CDS} = 49,6% e *M. bicuspidata*, GC^{CDS} = 51,9%. Apesar de ainda não haver nenhuma correlação comprovada entre psicrófilos e menor conteúdo GC, é tentador considerar a possibilidade de que a diminuição neste conteúdo,

especialmente em regiões de genes codificadores de proteínas, reduza a energia necessária para a abertura da dupla hélice e leve à melhor expressão desses genes em baixas temperatura.

Após as comparações quantitativas e qualitativas, o próximo procedimento realizado foi a construção de uma Árvore Filogenômica para posicionamento de *M. australis* junto às demais espécies, adicionando a este conjunto o genoma de *M. bicuspidata* e utilizando *Clavispora lusitaniae* como grupo externo. Esta árvore foi construída por Máxima Verossimilhança a partir de 1317 sequências de aminoácidos correspondentes a genes ortólogos de cópia única, e mostrou-se em concordância com outras filogenias já realizadas para estas espécies.

5.1.2 A busca por genes de adaptação ao frio em M. australis

Para investigar *M. australis*, adotamos a hipótese inicial de que as diferenças entre os genomas desta levedura e das demais *Metschnikowia* pudessem refletir adaptações de *M. australis* às condições de vida Antártica e que a possibilitam ser a única *Metschnikowia* endêmica dos mares no entorno do continente em questão. *M. bicuspidata*, apesar de presente em regiões temperadas e subpolares, não foi ainda identificada na região Antártica e, desta maneira, apresenta distribuição antagônica.

A disponibilização de uma linhagem de *M. bicuspidata*, gentilmente cedida pelo Prof. Mark André Lachance (*University of West Ontario*), possibilitou o cultivo comparativo dessas duas *Metschnikowia*. A caracterização das temperaturas de crescimento confirmou a hipótese já esperada da melhor performance de *M. australis* em temperaturas mais baixas. A partir destes resultados e de experimentos preliminares sobre a tolerância ao congelamento, iniciamos a busca por elementos gênicos exclusivos da levedura, que poderiam estar relacionados às suas características ímpares.

M. australis se mostra um organismo de difícil investigação. Apesar de ter sido possível relacionar 92% de suas CDSs preditas com anotações dos bancos de dados do **InterProScan**, estas são correspondentes a genes compartilhados com as diversas leveduras investigadas neste trabalho. Além disso, grande parte dessas anotações (mais do que 30%) corresponde a classificações pouco objetivas, como *consesus disorder prediction* ou *coil*, sendo isso observado em 86 das 247 CDSs exclusivas selecionadas nas análises de BLAST. As 161 restantes pertencem aos 8% de genes sem anotação. A busca por relações com bancos de dados mais abrangentes como o **NCBI nr** não trouxe

mais esclarecimentos. Quando há *hits*, estes são parciais e de baixa identidade, com proteínas putativas das próprias montagens depositadas ou outros organismos fracamente relacionados. Assim, com a identificação destas 247 CDSs foi preciso desenvolver uma estratégia de seleção para definir quais teriam prioridade na investigação experimental da expressão gênica nas condições de crescimento em baixas temperaturas. Muitos organismos adaptados à climas frios, incluindo diversas espécies Antárticas (Arai *et al.*, 2019; Davies, 2016; Muñoz *et al.*, 2017; Villarreal *et al.*, 2018) são produtores de AFPs. Estas proteínas apresentam alta diversidade estrutural e filogenética e são de grande interesse biotecnológico. Sendo *M. australis* endêmica da região mais fria do planeta, a possibilidade de que alguma destas CDSs pudesse corresponder a uma AFP norteou a seleção.

Como as estratégias tradicionais de anotação por similaridade não haviam encontrado nenhuma associação entre AFPs conhecidas e qualquer uma das 5.440 CDSs preditas em *M. australis*, optamos por utilizar classificadores de AFPs disponíveis na literatura. Muitos programas com estratégias de classificação diversas já haviam sido desenvolvidos. Entretanto, dos mais de dez classificadores já publicados, apenas três estavam disponíveis para uso, sendo estes o **RAFP-pred**, **CryoProtect** e **iAFP-ense**. Apesar de treinados com AFPs de organismos distantes, como peixes, plantas, insetos e bactérias, existia a possibilidade de que alguma das características de sequência exploradas por estes classificadores pudesse identificar atividade semelhante nas proteínas de *M. australis*. Por avaliarem características diferentes das sequências de aminoácidos para treinamento e predição, optamos por utilizar todos estes na busca de um consenso para o direcionamento das análises. A partir da consonância destes classificadores foram selecionadas 16 CDSs com potencial de ser AFP. Treze destas CDSs foram consideradas exclusivas de *M. australis* e três apresentaram similaridade parcial com regiões do genoma de *M. bicuspidata*.

Optamos por também adicionar a levedura *S. cerevisiae* aos experimentos de crescimento, com o intuito de utilizar um organismo de referência não-adaptado a baixas temperaturas. Nas curvas de crescimento a 12˚C, *S. cerevisiae* mostrou um crescimento inferior ao observado para *M. australis* e *M bicuspidata*. Estas duas últimas apresentaram um padrão de crescimento muito semelhante a 12˚C, apesar de *M. australis* demonstrar uma performance ligeiramente melhor. Nas curvas de crescimento a 6˚C, a diferença entre o padrão de crescimento de *M. australis* e *M. bicuspidata* aumentou drasticamente, com a segunda apresentando crescimento fraco, semelhante ao de *S. cerevisiae*.

Este é o primeiro estudo demonstrando o perfil de crescimento de *M. australis* e *M. bicuspidata* a 6 e 12 ˚C. Como já reportado, *M. australis* é capaz de crescer a 26 ˚C (Lachance *et al.*, 2011), não se encaixando na definição clássica de psicrófila. Destacamos que a levedura foi, em nossos testes, incapaz de crescer em cultivos a 28 ˚C. Entretanto, após uma semana mantida nesta temperatura, a levedura retomou seu crescimento normalmente quando colocada a 12˚C (resultados não mostrados), o que demonstra que a cultura foi capaz de resistir mesmo quando exposta à temperatura mais elevada. Novos experimentos seriam necessários para quantificar a taxa de sobrevivência.

O melhor desempenho de *M. australis* em baixas temperaturas poderia estar associado aos elementos de seu genoma. Para comprovar que as CDSs exclusivas selecionadas são de fato expressas e correspondem a genes transcritos e não a erros de predição gênica, foram desenhados iniciadores para sondar sua presença no cDNA de *M. australis* quando cultivada nessas condições. Quando possível, estes iniciadores foram desenhados com capacidade de amplificar sequências de *M. bicuspidata*, também para comprovar que estas sequências parcialmente compartilhadas não correspondiam a genes, ou a éxons de genes que haviam escapado à predição e fazem parte também do genoma de *M. bicuspidata*.

Paralelamente foram realizados os *spot-tests* de sobrevivência ao congelamento, com resultados surpreendentes. Mesmo com diferentes performances a baixas temperatura, todas as três leveduras se mostraram resistentes ao congelamento. Decidimos, então, buscar uma nova espécie para utilização como controle. Para poder associar a investigação experimental à investigação genômica seria preciso que esta levedura fosse filogeneticamente próxima das leveduras analisadas, estivesse disponível na coleção da UFMG e possuísse genoma sequenciado e depositado em repositórios públicos. Atendendo a todos estes critérios encontramos a levedura *Metschnikowia (Candida) golubevii*, cujo genoma estava também disponível. Esta espécie de levedura, pertencente às ***M*EPq florícolas,** foi isolada, independentemente, de uma flor de *Ipomoea* no Pantanal Brasileiro, às margens do Rio Paraguai, e de excrementos de insetos à beira da cachoeira Than-Thong, na Tailândia (Rosa *et al.*, 2010). Optamos por utilizar o isolado Tailandês, uma vez que este foi o utilizado para o sequenciamento do genoma. Uma outra *Metschnikowia* marinha, *M. (Candida) torresii*, também possuía genoma sequenciado e disponível, o que possibilitou sua inclusão nas análises computacionais. Assim como *M. golubevii*, esta levedura pertence ao clados das ***M*EPq**. *M. (C.) torresii* foi isolada da água do mar na região do Estreito de

Torres, ao norte da Austrália. Como o isolado da levedura não estava disponível em nossa coleção ou na de nossos colaboradores, esta foi incluída apenas nas análises computacionais.

Com a adição destas leveduras às análises, uma nova árvore filogenômica foi construída no nosso estudo. A árvore mostrou-se concordante com outras filogenias do gênero *Metschnikowia* (Lachance *et al.*, 2016; Lee *et al.*, 2018; Naumov, 2012). A árvore apresenta valores de suporte mais robustos para os clados e contém a levedura *M. golubevii*, ausente nas demais.

Podemos verificar a clara subdivisão do gênero *Metschnikowia* entre o subgrupo das ***M*EGd** e as demais, bem como a presença de dois genomas externos aos grupos, *M. kipukae* (kip) e *M.* sp. M2Y3 (cla), corroborando as filogenias anteriores. As ***M*EPq** são representadas por *M. golubevii* e *M. torresii*, *M. australis* e *M. bicuspidata*. Apesar de *M. torresii* ter sido isolada a partir da água do mar do Estreito de Torres ao norte da Austrália e ser, portanto, ecologicamente mais relacionada com *M. australis* e *M. bicuspidata*, a levedura se mostrou filogeneticamente mais distante destas, ao contrário de *M. golubevii*.

As curvas de crescimento a 12˚C e a 6˚C foram, então, realizadas incluindo as quatro leveduras. Apesar de ter sido isolada de regiões topicais, *M. golubevii* apresentou um crescimento a baixas temperaturas muito superior às leveduras sabidamente associadas a climas frios (*M. australis* e *M. bicuspidata)* e resistência aos testes de sobrevivência ao congelamento, assim como as demais. Isto reforçou a possibilidade de que a resistência ao congelamento fosse resultante de características fisiológicas e não da produção de AFPs. Mais ainda, o fato de as duas AFPs do único organismo ascomiceto produtor de proteínas anticongelantes já identificado, *Antarctomyces psicrotrophicus,* serem oriundas de transferência gênica horizontal a partir de bactérias (Arai, 2019), enfraqueceu a hipótese inicial de que *M. australis* fosse capaz de produzir alguma proteína com esta função.

5.1.2 A proteção inata das Metschnikowia

Leveduras florícolas ou associadas a frutos estão constantemente expostas a meios com alta concentração de açúcares (Dhami, Hartwig & Fukami, 2016). Estes meios apresentam baixa disponibilidade de água e podem levar à morte celular ou impedir o crescimento. Para contrabalancear estas condições, muitos organismos ascomicetos acumulam solutos intracelularmente. Estudos com *Candida sake* (Saccharomycetales) demonstraram que a levedura

apresenta aumento na concentração intracelular de açúcares (como glicose e trealose) e polióis (como glicerol, eritritol, arabitol e manitol), quando cultivadas em meios com baixa disponibilidade de água. Ressalta-se que estes solutos, mesmo em alta concentração, não prejudicam as atividades enzimáticas da célula (Teixido *et al.*, 1998). Em paralelo, um estudo sobre a recuperação de fungos filamentosos e leveduras a partir amostras da região costeira do Ártico (Butinar *et al.*, 2011) demonstrou que a diversidade de espécies de Basídiomicetos recuperados destas amostras é, em média, quatro vezes maior que a diversidade de Ascomicetos. O estudo reporta que uma maior concentração de açúcares no meio de isolamento leva a um aumento na diversidade recuperada dos dois Filos, havendo um aumento expressivo na diversidade recuperada de Ascomicetos. Além disso, a adição de NaCl de 10 a 15% inibe o crescimento de Basidiomicetos favorecendo os Ascomicetos.

Açúcares em alta concentração são utilizados como crioprotetores no congelamento de células e micro-organismos (Tsai *et al.*, 2018). Estes interferem nas propriedades coligativas da água e nos padrões de cristalização do gelo e também estabilizam a membrana plasmática (Teixido *et al.*, 1998). Assim, concluímos que a utilização de YPD, um meio rico em açúcares, poderia estar favorecendo a sobrevivência de todas as leveduras testadas em nossos experimentos. A ideia de utilizá-lo teve como intuito remover qualquer possível interferência nutricional na performance das leveduras durante os testes. Assim, decidimos por utilizar PBS para as realizar as diluições seriadas nos *spot tests*, o que levou à diminuição da sobrevivência, reforçando a hipótese da atividade crioprotetora do YPD. Não é possível afirmar, no entanto, se a diminuição da sobrevivência foi resultante da exposição ao PBS ou da ausência de atividade crioprotetora deste, uma vez que não foram feitas diluições controle para as diluições em PBS nos experimentos realizados.

5.1.3 As CDSs "exclusivas" de M. australis

Mesmo tendo direcionado as análises para a busca por AFPs possivelmente inexistentes, nossa estratégia de priorização de CDSs para investigação acabou fortuitamente encontrando CDSs com diferente expressão nas leveduras *M. australis* e *M. bicuspidata* em cultivos em baixas temperaturas. A maior parte destas CDSs se mostrou igualmente expressa por *M. australis* nos cultivos a 12˚C e a 6˚C, entretanto uma apresentou variação na expressão. A CDS **aus_3946** foi expressa apenas a 6˚C. A CDS **aus_3951** parece ter sido expressa apenas a 12˚C. No gel

correspondente há uma banda de difícil visualização. A baixa intensidade da banda pode estar relacionada a um mRNA de menor expressão, o que pode ser verificado com a aplicação de mais material amplificado em géis de experimentos futuros.

Destacamos o fato de que 6 dos 8 pares de iniciadores referentes a CDSs sem correspondência com *M. bicuspidata* foram capazes de amplificar, a partir do seu gDNA, um fragmento do mesmo tamanho daquele de *M. australis*. Dois destes, referentes às CDSs **aus_3157** e **aus_4351**, também apresentaram amplificações a partir do cDNA. O primeiro foi amplificado em ambas temperaturas, enquanto o segundo, apenas a 12˚C. Com relação aos 8 pares de iniciadores para o quais era esperada amplificação de fragmentos no gDNA de *M. bicuspidata*, todos amplificaram, sendo que 5 destes apresentaram amplicons de tamanho semelhante aos de *M. australis*, menores que o tamanho calculado pela ferramenta de construção de iniciadores, o **Primer BLAST**. Dos genomas estudados, *M. bicuspidata* é o que apresenta maior conteúdo de bases indefinidas (N) na sua montagem. Estas correspondem a 3,4% de seu genoma, ou seja, 377.466 bases. Se seus *scaffolds* forem divididos nas extensões de Ns maiores que 10, o número de *scaffolds* aumenta de 48 para 558. Em contraste, o genoma de *M. australis* montado por nós não apresenta nenhuma região indefinida e nenhum dos *contigs* foi unido artificialmente. Geralmente as informações utilizadas em uma montagem para se unir *contigs* distantes são produzidas por sequenciamento de bibliotecas do tipo *mate pair* ou por mapas óticos. Muitas vezes essas regiões de Ns apresentam extensão arbitrária, apenas para denotar contiguidade, mas também significam que a montagem pode estar incompleta. Estas regiões não definidas do genoma de *M. bicuspidata* podem corresponder a regiões homólogas em *M. australis*, levando à falsa identificação de exclusividade em *M. australis*, ou à predição de amplicons maiores do que o tamanho real em *M. bicuspidata*.

Mesmo que 11 dos 16 pares de iniciadores tenham produzido amplificações no genoma de *M. bicuspidata*, apenas 3 foram capazes de amplificar o cDNA. Em comparação, todos os 16 pares de iniciadores produziram amplicons no genoma de *M. australis*, e destes, 15 foram capazes de produzir amplificações a partir do cDNA oriundo do cultivo a 12˚C, e 13, do cultivo a 6˚C, o que representa que, de fato, mesmo que genômicamente compartilhados, a maior parte destes genes só está ativa em *M. australis* nas condições testadas. Todas as CDSs selecionadas precisam ser mais bem caracterizadas para identificação de sua possível função e a avaliação de o quão são determinantes no crescimento de *M. australis*, em especial em temperaturas baixas. Esta sondagem

superficial do transcriptoma já demonstra uma pequena diferença no perfil transcricional entre os crescimentos a 12˚C e a 6˚C, e estas temperaturas podem servir de referência para estudos posteriores. Idealmente deve ser realizado um experimento de RNAseq a fim de identificar quais são os genes de expressão constitutiva e quais estão associados à resposta à diminuição da temperatura. Dados de RNAseq também podem ser utilizados para verificar o nosso *pipeline* de anotação, mostrando se os genes estão sendo preditos corretamente, incluindo a estrutura de íntrons e éxons. Além disso, podem ajudar na identificação de RNAs não codificadores, que escapam aos preditores, voltados principalmente para genes codificadores de proteína.

5.1.4 Os genes ortólogos e parálogos

Os tratos e fenótipos que distinguem as leveduras analisadas quanto ao seu desempenho em baixas temperaturas podem estar relacionados a genes presentes em todos os genomas. Para os ortólogos de cópia única, as diferenças podem ser resultantes da regulação em cada organismo, o que não é possível verificar apenas com dados genômicos. Quanto aos parálogos, a presença de expansões no número de cópias de um determinado gene pode refletir sua importância para os processos celulares do organismo que as possui. O orthoMCL identificou diversos *clusters* de genes ortólogos, compostos tanto por ortólogos de cópia única quanto com mais de uma cópia para alguns organismos. Para as espécies *M. australis* e *M. bicuspidata* foram identificados 118 *clusters* de genes exclusivos compartilhados por estas leveduras. Dado a proximidade filogenética dessas duas espécies e à diferença de nicho das demais *Metschnikowia* estudadas (exceto *M. torresii*, também marinha), muitos desses *clusters* estão possivelmente associados com o habitat e ecologia destas leveduras. Muitos desses ortólogos são importantes por atuarem no transporte de íons e aminoácidos e na atividade proteolítica, que podem estar relacionados com a captação destes nutrientes do meio externo, seja a água do mar ou dos organismos que estas leveduras parasitam, ou aos quais estão vinculadas. Há *clusters* relacionados também com feromônios e sinalização de fatores de acasalamento (*mating*). Para estes é compreensível a exclusividade, dado que o isolamento reprodutivo de um clado é concomitante com acúmulo de variação nos genes envolvidos nos processos de acasalamento (Lee *et al.*, 2018).

Com relação à resistência ao congelamento, podemos sugerir a importância de 2 *clusters*. O primeiro corresponde a um gene com uma cópia em *M. australis* e *M. bicuspidata*. Estes genes

foram anotados pelo **InterProScan** como dessaturase de ácidos graxos. Todas as demais *Metschnikowia* estudadas apresentam 4 genes com esta função molecular, e todas estes genes foram relacionados em *clusters* de ortólogos de cópia única, comuns a todos genomas. Apenas *M. australis* e *M bicuspidata* apresentam um quinto gene correspondente a esta função. Isto poderia estar associado às alterações na composição lipídica da membrana plasmática, um fenômeno comum na resposta à exposição a baixas temperaturas (Buzzini & Margesin, 2014).

O segundo *cluster* também corresponde a um gene com uma cópia em *M. australis* e *M. bicuspidata*. Estes genes foram anotados pelo **InterProScan** como beta-manosil-transferases. Muitas APFs, por exemplo as AFP(G)s de peixes, são glicosiladas. Mesmo que *M. australis* e *M. bicuspidata* não produzam AFPs, a ligação de açúcares a proteínas pode estar analogamente associada à resistência ao frio. Além disso, a O-manosilação é bem estudada em *S. cerevisiae* e defeitos nas proteínas que realizam esta função interferem na integridade da parede celular (Loibl & Strahl, 2013).

Há ainda 9 *clusters* contendo 2genes cada, exclusivos de *M. australis*. Três destes foram parcialmente anotados como *Hyphally regulated cell wall protein*. Apesar de leveduras não produzirem hifas, esta é uma anotação relativamente comum para estes organismos. Geralmente é concernente a proteínas necessárias à atividade patogênica e que estão fixas à parede celular por âncoras de Glicosilfosfatidilinositol (GPI). Uma vez que *M. australis* é reconhecidamente uma levedura parasita do *Krill* antártico (Cleary *et al.*, 2019; Donachie & Zdanowski, 1998), estas proteínas podem atuar na interação da levedura com seu hospedeiro. *M. bicuspidata* também tem semelhante atividade parasita. Assim, a exclusividade desse *cluster* pode refletir a especialização de *M. australis* para os organismos do ecossistema antártico.

Outros dois *clusters* exclusivos foram anotados apenas como *consensus disorder prediction* (predição desordenada consensual), o que não é muito esclarecedor quanto à sua possível atividade. Quatro *clusters* não puderam ser anotados.

5.2 O gênero *Metschnikowia* e sua diversificação

5.2.1 Tanto no ar quanto no mar

De acordo com as estimativas feitas por Shen e colaboradores (2018), o Último Ancestral Comum (LCA, *Last Commom Ancestor*) do gênero *Metschnikowia* teria existido por volta de

90 MAA (Milhões de Anos Atrás). Observa-se também que o LCA de *M. bicuspidata* e *M. golubevii* teria existido por volta de 50 MAA. Apesar da árvore gerada neste trabalho não ser calibrada com um relógio molecular, é evidente que a especiação de *M. australis* e *M. bicuspidata* ocorreu posteriormente a este período.

A versatilidade do gênero *Metschnikowia* é evidenciada pela ubiquidade de seus membros, que são encontrados tanto no meio terrestre quanto aquático. Além das *Metschnikowia* estudadas neste trabalho, há ainda muitas outras representantes, coletadas em outros continentes e ambientes. Estes ambientes são semelhantes aos já aqui apresentados – como o trato digestivo de besouros da China (Chai *et al.*, 2019), ou a superfície de frutos (Xue *et al.*, 2006) – mas também inéditos, como madeira em decomposição (Hui *et al.*, 2013). Isto demonstra como a verdadeira diversidade do gênero pode ainda estar sendo subestimada. Além do pouco contato entre os grupos que estudam o gênero entre as frentes oriental e ocidental, com muitas novas espécies tendo sido descritas nos últimos anos, a procura por leveduras associadas a flores e frutos pode estar levando a um viés de representação. Além disso, as *M.* florícolas estão associadas à insetos polinizadores, principalmente besouros (Coleoptera). Sendo esta a ordem mais diversa de artrópodes, com mais de 400.000 espécies descritas (McKenna *et al.*, 2019), é possível que a coevolução orquestrada destes dois organismos tenha produzido muitas *Metschnikowia* ainda desconhecidas. Ao se levar em consideração também o fato de que muitas populações de plantas com flores efêmeras que albergam estas leveduras e são visitadas por besouros se encontrarem isoladas geograficamente (Babiychuk *et al.*, 2019), o potencial para o isolamento genético é altíssimo.

A maior abundância de espécies do clado florícola nas análises apresentadas neste estudo pode ser uma evidência do maior esforço amostral neste nicho, mas pode refletir também um menor isolamento reprodutivo das leveduras do clado aquático. Apesar de suas dimensões, o conjunto dos oceanos do planeta é comunicante e muitas correntes marítimas auxiliam na dispersão global dos organismos que estão à deriva (Cavicchioli, 2015). Estas correntes podem facilitar o fluxo gênico, reduzindo o ritmo dos processos de especiação entre os representantes do clado aquático.

A levedura *M. australis* foi por muito tempo considerada apenas uma variedade de *M. bicuspidata*. Sendo tão semelhantes, tanto em nicho quanto morfológica e geneticamente, qual fator que teria desencadeado o processo de especiação nestes organismos? A linhagem *M. bicuspidata* foi isolada do Lago Michigan (Wickerham, 1964), que é anualmente recoberto por gelo. Como também demonstrado pelos experimentos de sobrevivência ao congelamento realizados neste

trabalho, fica evidente que a pressão seletiva do congelamento não poderia ser responsável pela especiação. Mesmo tendo praticamente a mesma resistência ao congelamento que *M. australis*, *M. bicuspidata* apresenta uma menor taxa de crescimento em baixas temperaturas (6˚C), como evidenciado em nos cultivos produzidos nesse trabalho. Este melhor desempenho de *M. australis* no crescimento a temperaturas extremas é possivelmente uma consequência, e não a causa, da especiação.

5.2.2 Correntes marítimas: pontes ou barreiras?

Um recente trabalho por Clarke e colaboradores (2019) utilizando sequenciamento de 16S investigou o microbioma bacteriano do *Krill Euphausia superba*. Os autores destacam a importância dele para a manutenção da biodiversidade microbiológica, sendo cada organismo um ecossistema rico em nutrientes, em contraste com os mares que habitam. Como já mencionado, este crustáceo, que é parasitado por *M. australis*, é encontrado em abundância em todos os mares ao redor da Antártica e, inclusive, sua distribuição coincide amplamente com a extensão máxima da camada de gelo formada no inverno, sob a qual se abriga durante o período larval (Cavan *et al.*, 2019). Além disso, este organismo apresenta uniformidade genética em torno de todo o continente (Bortolotto *et al.*, 2011), mostrando que toda espécie se comporta como uma única população, integrada pelas correntes pericontinentais.

As correntes podem também atuar como barreiras, tanto para as espécies quanto para o clima. Com a abertura da passagem entre a América do Sul e a Península Antártica, estimada entre 41 e 16 MAA, não havia mais impedimento para a formação de uma corrente marítima em torno do continente. Supõe-se que Corrente Circumpolar Antártica (CCA), considerada a mais forte do planeta, aumentou o isolamento térmico do continente, contribuindo para a formação e expansão da Calota Polar (Barker *et al.*, 2007). Estima-se que a Calota Polar teve sua formação iniciada a 34 MAA e, entre períodos de expansão e retração, se estendeu gradativamente até recobrir as regiões marítimas por volta de 30 MAA, durante o Oligoceno (Pyne *et al.*, 2015). Com o resfriamento dos mares austrais, adaptações compatíveis com as condições extremas foram selecionadas nos organismos ali presentes. Mais ainda, as populações e linhagens com vantagens adaptativas para o frio foram se distanciando das linhagens ancestrais, garantindo a exclusividade e se especializando para sobrevivência neste ambiente ímpar. Análises de relógio molecular realizadas utilizando

variações na subunidade maior do ribossomo mitocondrial do *Krill E. superba*, crustáceo dominante dos mares antárticos, e *E. vallentini*, de distribuição subantártica, estimaram que a divergência entre estas espécies ocorreu há 19 MAA (Patarnello *et al.*, 1996), posteriormente à formação da CCA e ao resfriamento do continente antártico. A estimativa produzida para os crustáceos utilizou apenas um gene ribossomal mitocondrial. Apesar de genes ribossomais nucleares serem altamente conservados, os mitoribossomos apresentam uma grande tendência à diversificação (Petrov *et al.*, 2019). Além disso, o DNA mitocondrial reconhecidamente apresenta uma taxa de mutação discrepante em ralação ao DNA genômico, podendo acumular mutações numa proporção até 10 vezes maior em invertebrados (Allio *et al.*, 2017). Este maior número de mutações pode levar a uma superestimativa do tempo de divergência. Assim, existe a possibilidade que o tempo de divergência das espécies de *Krill* possa ser menor que o considerado. A estimativa da divergência de *M. australis* e *M. bicuspidata* realizada neste trabalho posiciona a separação destas espécies num intervalo entre 2,5M e 250 mil anos. Esta separação pode ter sido desencadeada pelo isolamento gradual das linhagens de *Metschnikowia* associadas ao zooplâncton da região antártica. Mesmo que as estimativas da divergência das espécies de *Krill* esteja correta, a especiação mais recente de *M. australis* e *M. bicuspidata* pode refletir o hábito destes organismos. O zooplâncton é livre natante e tende a se aglomerar em regiões ricas em produção primária (Cavan *et al.*, 2019), sendo capaz de se mover voluntariamente, além de também estar sob influência das correntes oceânicas, mesmo que em menor escala. Já as leveduras, quando não associadas a estes organismos, estão à deriva. Seus esporos podem ser carregados pelas correntes para outras regiões, possibilitando um fluxo gênico, ainda que pequeno, que retardaria o processo de especiação.

As leveduras *M. australis* e *M. bicuspidata* foram isoladas independentemente várias vezes. *M. bicuspidata*, em especial, tem presença identificada em regiões muito distantes. A classificação taxonômica destes isolados é geralmente realizada por um único marcador molecular (Kurtzman *et al.*, 2011), o que pode não refletir a verdadeira variedade existente na espécie. Apenas um exemplar de cada uma destas leveduras tem o genoma sequenciado. Os isolados utilizados para estes sequenciamentos foram coletados a uma distância de 15.000 km. Tendo uma distribuição tão abrangente, é possível que populações de *M. bicuspidata* mais próximas da Antártica apresentem maior semelhança as leveduras *M. australis*. A obtenção de genomas de outros isolados das duas leveduras possibilitará uma melhor compreensão de suas divergências.

5.2.3 Metschnikowia: leveduras anfíbias?

A levedura *M. golubevii* foi adicionada às análises com o objetivo inicial de servir de controle para os ensaios de sobrevivência ao congelamento, por sua disponibilidade na coleção da UFMG e por ter seu genoma sequenciado. Apesar de ter sido isolada independentemente de dois ambientes tropicais – flores de *Ipomoea* sp. às margens do Rio Paraguai, no Brasil, e em fezes de insetos, na Tailândia – análises filogenômicas a posicionam justaposta ao clado ***M*EPq aquáticas**. Mais surpreendente ainda foi o fato de que esta levedura apresenta maior taxa de crescimento a baixas temperaturas que *M. australis* e *M. bicuspidata*. Os dois pontos de onde a levedura foi isolada estão diretamente associados a cursos de água doce. Isto abre um precedente para a possibilidade de que alguns representantes, tanto do clado das ***M*EGd florícolas** quanto do clado das ***M*EPq aquático** compartilhem nichos. Estaria a levedura *M. golubevii* também associada a invertebrados aquáticos? Há inúmeras espécies de besouros aquáticos, tanto na fase larval quanto na adulta. Dentre estes, inclusive, é descrita uma espécie de besouro que é aquática facultativa. *Amphicrossus japonicus* é encontrado em associação com colmos de bambu em ambientes alagados, nas Filipinas. Esta é a única espécie aquática descrita pertencente à Família Nitidulidae (Kovac *et al.*, 2007), a qual estão associadas mais da metade das ***M*EGd** descritas.

Não é possível precisar se o ancestral comum das *Metschnikowia* é aquático ou terrestre. Tampouco se alguma das espécies desse gênero é anfíbia. Como a maioria das identificações parte de coletas direcionadas, é necessária uma nova abordagem para a prospecção de novas espécies. Além de direcionar novas coletas para ambientes ainda não explorados como rios e os invertebrados que os habitam, o sequenciamento de DNA ambiental (eDNA) proveniente destes ecossistemas pode ser uma estratégia promissora para se encontrar novos representantes de *Metschnikowia*. Esta técnica já tem sido utilizada para identificar diversas comunidades de organismos aquáticos. Um estudo realizado com rios a jusante de um lago habitado por *Daphnia longispina* foi capaz de detectar seu DNA a uma distância de até 25 km do lago de origem. O gênero *Daphnia* apresenta espécies altamente parasitadas por *M. bicuspidata* (Duffy & Sivars-Becker, 2007) e, se é possível detectar o hospedeiro, a investigação de amostras deste tipo com sondas desenhadas para *Metschnikowia* pode revelar uma diversidade oculta.

5.3 Classificadores de proteínas anticongelantes

Uma máxima muito conhecida na Ciência da Computação é '*garbage in, garbage out*' ('lixo entra, lixo sai'). Uma vez que os computadores são máquinas determinísticas que seguem rigorosamente instruções, a qualidade do resultado (*output*) de qualquer análise computacional depende da qualidade do da informação fornecida para esta análise (*input*). Esta máxima é bem estabelecida para a filogenética, onde é sabido que a qualidade de qualquer reconstrução é altamente influenciada pelos dados utilizados (Bininda-Emonds *et al.*, 2004).

Os programas baseados em aprendizado de máquina são grandes facilitadores dos processos computacionais e tem ganhado muito espaço nos últimos anos. Para os problemas que são muito complexos para ter suas soluções definidas por algoritmos diretos ou heurísticos é possível deixar a cargo do computador encontrar as melhores estratégias para sua resolução. Seu funcionamento ótimo muitas vezes independe da compreensão dos mecanismos que adotam para tal. Para estes, a eficiência também é altamente dependente dos conjuntos de dados utilizados em seu treinamento. Mais do que bons dados, é preciso ter bons dados bem categorizados, pois a união de entidades diferentes em uma mesma categoria pode ser um fator de confusão.

Os diversos classificadores de AFPs existentes utilizam diferentes informações de sequência e estratégias de classificação. Entre os doze já publicados, o **SVMGMA** utiliza SVM e Algoritmos Genéticos e realiza treinamento e teste com um *dataset* composto por AFPs com estrutura resolvida disponível no PDB (Yu & Lu, 2011). Um segundo, sem nome, utiliza dinâmica molecular para encontrar superfícies com potencial AFPs, também em proteínas com estrutura resolvida (Kozuch *et al.*, 2018). Os outros classificadores de AFPs desenvolvidos são: **AFP_PSSM** (Zhao *et al.*, 2012); **AFP-PseAAC** (Mondal & Pai, 2014); **AFP-Ensemble** (Yang *et al.*, 2015); **TargetFreeze** (He *et al.*, 2015); **iAFP-ense** (Xiao *et al.*, 2016); **Cryoprotect** (Pratiwi *et al.*, 2017); **iAFP-gap-SMOTE** (Akbar *et al.*, 2018); **RAFP-pred** (Khan *et al.*, 2018); **afpCOOL** (Eslami *et al.*, 2018); **AFP-CKSAAP** (Usman & Lee, 2019). Todos estes 10 classificadores, apesar de utilizarem as mais variadas estratégias e características de sequência para a classificação, utilizam o mesmo *dataset* selecionado inicialmente por Kandaswamy e colaboradores em 2011 para o desenvolvimento do **AFP-pred**. Este *dataset* é composto por 481 AFPs de plantas, peixes, insetos e bactérias e 9493 proteínas não-AFPs de origem variada.

A utilização deste mesmo *dataset* por todos classificadores é justificada pois possibilita a comparação de sua eficiência. Cada uma das estratégias apresenta ganhos e perdas em especificidade e sensibilidade. Entretanto, mais importante que o desenvolvimento de novas abordagens para o desafio da classificação de AFPs, é necessária uma reavaliação dos dados utilizados. Desde a construção deste *dataset* em 2011, muitas novas AFPs foram descritas (Arai *et al.*, 2019; Davies, 2016; Firdaus-Raih *et al.*, 2018; Hashim *et al.*, 2013). Estas novas AFPs podem contribuir para o treinamento dos classificadores já existentes, melhorando sua eficiência de classificação para organismos que não possuíam AFPs representadas no *dataset* de Kandaswamy, como fungos.

As AFPs são um agrupamento artificial de proteínas capazes de realizar uma mesma função, se ligar aos planos do cristal de gelo em formação interferindo na sua dinâmica de crescimento e associação (Dolev, Braslavsky, & Davies, 2016). Sua origem é polifilética e mesmo organismos próximos produzem AFPs sem relação estrutural. Apenas em peixes encontramos 4 tipos distintos de AFPs. Assim, a união de entidades tão diferentes em uma única categoria pode ser um fator de confusão para o treinamento dos classificadores, diluindo informações preciosas, intrínsecas a cada um dos subtipos de AFPs. Desta maneira, a separação destas proteínas em grupos discretos que reflitam suas características estruturais e filogenéticas pode fazer com que os classificadores já existentes tenham uma melhora significativa na sua capacidade de discriminação.

6 CONCLUSÃO

Neste trabalho concluímos que:

- A reconstituição filogenômica realizada para as *Metschnikowia* estudadas é concordante com outras filogenias já construídas e confirma a relação das *Metschnikowia* do clado aquático com as demais *Metschnikowia* de esporo pequeno;
- A levedura *M. australis* possui um genoma mais compacto, com menos CDSs e tRNAS, e menor conteúdo de repetições que sua espécie mais próxima, *M. bicuspidata*;
- *M. australis* apresenta melhor capacidade de crescimento em baixas temperaturas que *M. bicuspidata;*
- Fomos capazes de identificar 16 genes exclusivos que podem estar relacionados com a melhor capacidade de crescimento da levedura e seu endemismo na Antártica;
- A capacidade de crescimento a baixas temperaturas não é exclusiva das *Metschnikowias* presentes em ambientes temperados e polares.
- A divergência entre as espécies *M. australis e M. bicuspidata* pode ser mais recente que a dos organismos aos quais estas leveduras estão ecologicamente relacionadas.

7 PERSPECTIVAS

Percebemos que as estratégias de cultivo podem ter influenciado fortemente a performance das leveduras nos experimentos realizados. Assim, novos cultivos utilizando meios com menor concentração de açúcares podem realçar as capacidades das leveduras adaptadas a climas frios e temperados.

A investigação dos outros genes exclusivos selecionados é também uma possibilidade. A busca por enzimas adaptadas a baixas temperaturas é um dos objetivos da prospecção biotecnológica de muitos organismos antárticos. Seria algum destes genes codificador de uma enzima adaptada ao funcionamento em baixas temperaturas?

Há também possibilidade de investigação do transcriptoma de *M. australis*. Nosso estudo preliminar já demonstrou que há diferença no perfil trasncricional da levedura quando cultivada a 6 e a 12 ˚C. Além de possibilitar a compreensão da resposta da levedura à diminuição da temperatura, estes dados também serão de grande valor para a validação de nosso *pipeline* de predição e anotação.

Com relação aos Classificadores de AFPs, sugerimos que a atualização do principal conjunto de dado utilizado possa melhorar sua capacidade de discriminação entre AFPs e não-AFPs, ou mesmo viabilizar a discriminação entre diferentes tipos de AFPs. Um deles, o RAFP-pred, têm modelo estatístico de classificação em uma plataforma aberta, o que possibilita a realização de novos treinamentos com um conjunto de dados atualizado. Além disso, muitas etapas de obtenção das características das proteínas a serem avaliadas são manuais. Parte destas etapas foi otimizada com *scripts* em R, e ainda há possibilidade de otimização de outras.

REFERÊNCIAS

AASSP. Australian Antarctic science strategic plan. **Antarctic Science**, p. 1–85, 2011.

AB, Nobel Media. **Ilya Mechnikov – Facts**. Disponível em: <https://www.nobelprize.org/prizes/medicine/1908/mechnikov/facts/>.

AKBAR, Shahid; HAYAT, Maqsood; KABIR, Muhammad; *et al*. iAFP-gap-SMOTE: An Efficient Feature Extraction Scheme Gapped Dipeptide Composition is Coupled with an Oversampling Technique for Identification of Antifreeze Proteins. **Letters in Organic Chemistry**, v. 16, n. 4, p. 294–302, 2018.

ALLIO, Remi; DONEGA, Stefano; GALTIER, Nicolas; *et al*. Large Variation in the Ratio of Mitochondrial to Nuclear Mutation Rate across Animals: Implications for Genetic Diversity and the Use of Mitochondrial DNA as a Molecular Marker. **Molecular Biology and Evolution**, v. 34, n. 11, p. 2762–2772, 2017. Disponível em: <http://academic.oup.com/mbe/article/34/11/2762/3976052>.

AN, Meiling; MOU, Shanli; ZHANG, Xiaowen; *et al*. Temperature regulates fatty acid desaturases at a transcriptional level and modulates the fatty acid profile in the Antarctic microalga *Chlamydomonas* sp. ICE-L. **Bioresource Technology**, v. 134, p. 151–157, 2013. Disponível em: <http://dx.doi.org/10.1016/j.biortech.2013.01.142>.

ANDREWS, Simon; KRUEGER, Felix; SEGONDS-PICHON, Anne; *et al*. FastQC: a quality control tool for high throughput sequence data. 2012. Disponível em: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.

ANTONY, Runa; SANYAL, Aritri; KAPSE, Neelam; *et al*. Microbial communities associated with Antarctic snow pack and their biogeochemical implications. **Microbiological Research**, v. 192, p. 192–202, 2016. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0944501316304682?via%3Dihub>.

ARAI, Tatsuya; FUKAMI, Daichi; HOSHINO, Tamotsu; *et al*. Ice-binding proteins from the fungus *Antarctomyces psychrotrophicus* possibly originate from two different bacteria through horizontal gene transfer. **FEBS Journal**, v. 286, n. 5, p. 946–962, 2019. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/febs.14725>.

ARGENTINA; AUSTRÁLIA; BÉLGICA; *et al*. Antarctic Treaty. 1959.

BABIYCHUK, Elena; TEIXEIRA, Juliana Galaschi; TYSKI, Lourival; *et al*. Geography is essential for reproductive isolation between florally diversified morning glory species from Amazon canga savannahs. **Scientific Reports**, v. 9, n. 1, p. 1–18, 2019.

BARKER, Peter F.; FILIPPELLI, Gabriel M.; FLORINDO, Fabio; *et al*. Onset and role of the Antarctic Circumpolar Current. **Deep-Sea Research Part II: Topical Studies in Oceanography**, v. 54, n. 21–22, p. 2388–2398, 2007.

BARNES, David K.A.; CLARKE, Andrew. Antarctic marine biology. **Current Biology**, v. 21, n. 12, p. R451–R457, 2011. Disponível em: <http://dx.doi.org/10.1016/j.cub.2011.04.012>.

BARNES, David K.A.; TARLING, Geraint A. Polar oceans in a changing climate. **Current Biology**, v. 27, n. 11, p. R454–R460, 2017. Disponível em: <http://dx.doi.org/10.1016/j.cub.2017.01.045>.

BASTIAN, M; HEYMANN, S; JACOMY, M. Gephi: an open source software for exploring and manipulating networks. BT - International AAAI Conference on Weblogs and Social. **International AAAI Conference on Weblogs and Social Media**, p. 361–362, 2009. Disponível em: <https://dx.plos.org/10.1371/journal.pone.0098679>.

BATISTA, Thiago M; HILÁRIO, Heron O; MOREIRA, Rennan G; *et al*. Draft Genome Sequence of *Metschnikowia australis* Strain UFMG-CM-Y6158, an Extremophile Marine Yeast

Endemic to Antarctica. **Genome Announcements**, v. 5, n. 20, p. e00328-17, 2017. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/28522704>.

BERLEMONT, Renaud; JACQUIN, Olivier; DELSAUTE, Maud; *et al*. Novel Cold-Adapted Esterase MHlip from an Antarctic Soil Metagenome. **Biology**, v. 2, n. 1, p. 177–188, 2013. Disponível em: <http://www.mdpi.com/2079-7737/2/1/177>.

BERNARD, Guillaume; PATHMANATHAN, Jananan S; LANNES, Romain; *et al*. Microbial darkmatter investigations: howmicrobial studies transform biological knowledge and empirically sketch a logic of scientific discovery. **Genome Biology and Evolution**, v. 10, n. 3, p. 707–715, 2018. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/29420719>.

BORTOLOTTO, Erica; BUCKLIN, Ann; MEZZAVILLA, Massimo; *et al*. Gone with the currents: lack of genetic differentiation at the circum-continental scale in the antarctic krill *Euphausia superba*. **BMC Genetics**, v. 12, 2011.

BRIDGE, P. D.; SPOONER, B. M. Non-lichenized antarctic fungi: transient visitors or members of a cryptic ecosystem? **Fungal Ecology**, v. 5, n. 4, p. 381–394, 2012. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1754504812000086>.

BRITISH ANTARCTIC SURVEY. **Antarctic Wildlife**. Disponível em: <https://www.bas.ac.uk/about/antarctica/wildlife/plants/>.

BUCHFINK, Benjamin; XIE, Chao; HUSON, Daniel H. Fast and sensitive protein alignment using DIAMOND. **Nature Methods**, v. 12, n. 1, p. 59–60, 2015. Disponível em: <http://www.nature.com/articles/nmeth.3176>.

BUTINAR, L.; SANTOS, S.; SPENCER-MARTINS, I.; *et al*. Yeast diversity in hypersaline habitats. **FEMS Microbiology Letters**, v. 244, n. 2, p. 229–234, 2005. Disponível em: <https://academic.oup.com/femsle/article-lookup/doi/10.1016/j.femsle.2005.01.043>.

BUTINAR, Lorena; STRMOLE, Tadeja; GUNDE-CIMERMAN, Nina. Relative incidence of Ascomycetous yeasts in Arctic Coastal environments. **Microbial Ecology**, v. 61, n. 4, p. 832–843, 2011. Disponível em: <http://link.springer.com/10.1007/s00248-010-9794-3>.

BUZZINI, Pietro; MARGESIN, Rosa. Cold-adapted Yeasts. Berlin, Heidelberg. **Springer Berlin Heidelberg**, 2014. Disponível em: <http://link.springer.com/10.1007/978-3-642-39681-6>.

CAPELLA-GUTIERREZ, S.; SILLA-MARTINEZ, J. M.; GABALDON, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. **Bioinformatics**, v. 25, n. 15, p. 1972–1973, 2009. Disponível em: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp348>.

CAVAN, E. L.; BELCHER, A.; ATKINSON, A.; *et al*. The importance of antarctic krill in biogeochemical cycles. **Nature Communications**, v. 10, n. 1, p. 4742, 2019. Disponível em: <http://dx.doi.org/10.1038/s41467-019-12668-7>.

CAVICCHIOLI, Ricardo. Microbial ecology of antarctic aquatic systems. **Nature Reviews Microbiology**, v. 13, n. 11, p. 691–706, 2015.

CELIK, Yeliz; GRAHAM, Laurie A.; MOK, Y.-F.; *et al*. Superheating of ice crystals in antifreeze protein solutions. **Proceedings of the National Academy of Sciences**, v. 107, n. 12, p. 5423–5428, 2010. Disponível em: <http://www.pnas.org/cgi/doi/10.1073/pnas.0909456107>.

CHAI, Chun-Yue; HUANG, Lin-Na; CHENG, Han; *et al*. *Metschnikowia baotianmanensis* f.a., sp. nov., a new yeast species isolated from the gut of the rhinoceros beetle *Allomyrina dichotoma*. **International Journal of Systematic and Evolutionary Microbiology**, v. 69, n. 10, p. 3087–3092, 2019. Disponível em: <https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijsem.0.003593>.

CHAN, Patricia P.; LOWE, Todd M. tRNAscan-SE: searching for tRNA genes in genomic sequences. **Methods in Molecular Biology**, v. 1962, n. 5, p. 1–14, 2019. (Methods in Molecular

Biology). Disponível em: <http://www.springer.com/series/7651%0Ahttp://link.springer.com/10.1007/978-1-4939-9173-0>.

CHAN, Yuki; VAN NOSTRAND, Joy D; ZHOU, Jizhong; *et al*. Functional ecology of an Antarctic Dry Valley. **Proceedings of the National Academy of Sciences**, v. 110, n. 22, p. 8990–8995, 2013. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/23671121>.

CHOI, JaeJin; KIM, Sung-Hou. A genome tree of life for the Fungi kingdom. **Proceedings of the National Academy of Sciences of the United States of America**, v. 114, n. 35, p. 9391–9396, 2017. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/28808018>.

CHOU, Kuo-Chen. Prediction of Protein Cellular Attributes Using Pseudo-Amino Acid Composition. **Proteins: Struct., Funct., Genet.**, v. 255, p. 246–255, 2001.

CHOU, Kuo-Chen. Pseudo Amino Acid Composition and its Applications in Bioinformatics, Proteomics and System Biology. **Current Proteomics**, v. 6, n. 4, p. 262–274, 2009. Disponível em: <http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1570-1646&volume=6&issue=4&spage=262>.

CHOWN, Steven L.; CLARKE, Andrew; FRASER, Ceridwen I.; *et al*. The changing form of antarctic biodiversity. **Nature**, v. 522, n. 7557, p. 431–438, 2015. Disponível em: <http://www.nature.com/articles/nature14505>.

CLARKE, Laurence J.; SUTER, Léonie; KING, Robert; *et al*. Antarctic krill are reservoirs for distinct southern ocean microbial communities. **Frontiers in Microbiology**, v. 9, p. 1–9, 2019. Disponível em: <https://www.frontiersin.org/article/10.3389/fmicb.2018.03226/full>.

CLEARY, Alison C.; CASAS, Maria C.; DURBIN, Edward G.; *et al*. Parasites in antarctic krill guts inferred from DNA sequences. **Antarctic Science**, v. 31, n. 1, p. 16–22, 2019. Disponível

em:
<https://www.cambridge.org/core/product/identifier/S0954102018000469/type/journal_article>.

COWAN, DA; RAMOND, J-B; MAKHALANYANE, TP; *et al*. Metagenomics of extreme environments. **Current Opinion in Microbiology**, v. 25, p. 97–102, 2015. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1369527415000569>.

COWAN, Don A.; CHOWN, Steven L.; CONVEY, Peter; *et al*. Non-indigenous microorganisms in the Antarctic: assessing the risks. **Trends in Microbiology**, v. 19, n. 11, p. 540–548, 2011. Disponível em: <http://dx.doi.org/10.1016/j.tim.2011.07.008>.

COWAN, Don A.; TOW, Lemese Ah. Endangered antarctic environments. **Annual Review of Microbiology**, v. 58, n. 1, p. 649–690, 2004.

DANIEL, Heide-Marie; LACHANCE, Marc-André; KURTZMAN, Cletus P. On the reclassification of species assigned to *Candida* and other anamorphic ascomycetous yeast genera based on phylogenetic circumscription. **Antonie van Leeuwenhoek**, v. 106, n. 1, p. 67–84, 2014. Disponível em: <http://link.springer.com/10.1007/s10482-014-0170-z>.

DAS, Annada; CHAUHAN, Geeta; SATYAPRAKASH, Kaushik; *et al*. Application of antifreeze proteins in foods of animal origin – a review. **International Journal of Livestock Research**, v. 8, n. 5, p. 70, 2018. Disponível em: <https://www.ejmanager.com/fulltextpdf.php?mno=291813>.

DAVIES, Peter L. Antarctic moss is home to many epiphytic bacteria that secrete antifreeze proteins. **Environmental Microbiology Reports**, v. 8, n. 1, p. 1–2, 2016. Disponível em: <http://doi.wiley.com/10.1111/1758-2229.12360>.

DAVIES, Peter L. Ice-binding proteins: a remarkable diversity of structures for stopping and starting ice growth. **Trends in Biochemical Sciences**, v. 39, n. 11, p. 548–555, 2014. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0968000414001716>.

DAVIES, Peter L; BAARDSNES, Jason; KUIPER, Michael J; *et al*. Structure and function of antifreeze proteins. *In*: **Philosophical Transactions of the Royal Society B: Biological Sciences**. [s.l.: s.n.], 2002, v. 357, p. 927–935. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1692999/pdf/12171656.pdf>.

DE MAAYER, P.; ANDERSON, D.; CARY, C.; *et al*. Some like it cold: understanding the survival strategies of psychrophiles. **EMBO reports**, v. 15, n. 5, p. 508–517, 2014. Disponível em: <http://embor.embopress.org/cgi/doi/10.1002/embr.201338170>.

DE MENDONÇA VILELA, Mariane; DEL BEM, Luiz Eduardo; VAN SLUYS, Marie Anne; *et al*. Analysis of Three Sugarcane Homo/Homeologous Regions Suggests Independent Polyploidization Events of *Saccharum officinarum* and *Saccharum spontaneum*. **Genome Biology and Evolution**, v. 9, n. 2, p. 266–278, 2017.

DE MENEZES, Graciéle C.A. A.; GODINHO, Valéria M.; PORTO, Bárbara A.; *et al*. *Antarctomyces pellizariae* sp. nov., a new, endemic, blue, snow resident psychrophilic ascomycete fungus from Antarctica. **Extremophiles**, v. 21, n. 2, p. 259–269, 2017. Disponível em: <http://link.springer.com/10.1007/s00792-016-0895-x>.

DE PASCALE, Donatella; DE SANTI, Concetta; FU, Juan; *et al*. The microbial diversity of polar environments is a fertile ground for bioprospecting. **Marine Genomics**, v. 8, p. 15–22, 2012. Disponível em: <http://dx.doi.org/10.1016/j.margen.2012.04.004>.

DEVRIES, Arthur L.; WOHLSCHLAG, Donald E. Freezing resistance in some antarctic fishes. **Science**, v. 163, n. 3871, p. 1073–1075, 1969. Disponível em: <http://www.sciencemag.org/cgi/doi/10.1126/science.163.3871.1073>.

DHAMI, Manpreet K.; HARTWIG, Thomas; FUKAMI, Tadashi. Genetic basis of priority effects: insights from nectar yeast. **Proceedings of the Royal Society B: Biological Sciences**, v. 283, n. 1840, 2016.

DOBIN, Alexander; DAVIS, Carrie A.; SCHLESINGER, Felix; *et al*. STAR: ultrafast universal RNA-seq aligner. **Bioinformatics**, v. 29, n. 1, p. 15–21, 2013. Disponível em: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts635>.

DOLEV, Maya Bar; BRASLAVSKY, Ido; DAVIES, Peter L. Ice-binding proteins and their function. 2016. Disp. em: http://www.annualreviews.org/doi/pdf/10.1146/annurev-biochem-060815-014546.

DONACHIE, SP; ZDANOWSKI, MK. Potential digestive function of bacteria in krill *Euphausia superba* stomachs. **Aquatic Microbial Ecology**, v. 14, n. 2, p. 129–136, 1998. Disponível em: <http://www.int-res.com/abstracts/ame/v14/n2/p129-136/>.

DOXEY, Andrew C.; YAISH, Mahmoud W.; GRIFFITH, Marilyn; *et al*. Ordered surface carbons distinguish antifreeze proteins and their ice-binding regions. **Nature Biotechnology**, v. 24, n. 7, p. 852–855, 2006.

DU, Pufeng; WANG, Xin; XU, Chao; *et al*. PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. **Analytical Biochemistry**, v. 425, n. 2, p. 117–119, 2012. Disponível em: <https://www-sciencedirect-com.ez27.periodicos.capes.gov.br/science/article/pii/S0003269712001819?via=ihub>.

DUARTE, Alysson Wagner Fernandes; PASSARINI, Michel Rodrigo Zambrano; DELFORNO, Tiago Palladino; *et al*. Yeasts from macroalgae and lichens that inhabit the South Shetland Islands, Antarctica. **Environmental Microbiology Reports**, v. 8, n. 5, p. 874–885, 2016. Disponível em: <http://doi.wiley.com/10.1111/1758-2229.12452>.

DUFFY, Meghan A.; SIVARS-BECKER, Lena. Rapid evolution and ecological host-parasite dynamics. **Ecology Letters**, v. 10, n. 1, p. 44–53, 2007. Disponível em: <http://doi.wiley.com/10.1111/j.1461-0248.2006.00995.x>.

DUJON, Bernard A; LOUIS, Edward J. Genome diversity and evolution in the budding yeasts (Saccharomycotina). **Genetics**, v. 206, n. 2, p. 717–750, 2017. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/28592505>.

DUMAN, John G. Antifreeze and ice nucleator proteins in terrestrial arthropods. **Annual Review of Physiology**, v. 63, n. 1, p. 327–357, 2001. Disponível em: <http://www.annualreviews.org/doi/10.1146/annurev.physiol.63.1.327>.

EDGAR, Robert C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. **Nucleic Acids Research**, v. 32, n. 5, p. 1792–1797, 2004. Disponível em: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkh340>.

ESLAMI, Morteza; SHIRALI HOSSEIN ZADE, Ramin; TAKALLOO, Zeinab; *et al*. afpCOOL: a tool for antifreeze protein prediction. **Heliyon**, v. 4, n. 7, p. e00705, 2018. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2405844018329451>.

ETOURNEAU, Johan; SGUBIN, Giovanni; CROSTA, Xavier; *et al*. Ocean temperature impact on ice shelf extent in the eastern Antarctic Peninsula. **Nature Communications**, v. 10, n. 1, p. 304, 2019. Disponível em: <http://www.nature.com/articles/s41467-018-08195-6>.

FIRDAUS-RAIH, Mohd; HASHIM, Noor Haza Fazlin; BHARUDIN, Izwan; *et al*. The *Glaciozyma antarctica* genome reveals an array of systems that provide sustained responses towards temperature variations in a persistently cold habitat. **PLOS ONE**, v. 13, n. 1, p. e0189947, 2018. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/29385175>.

FLETCHER, Garth L; HEW, Choy L; DAVIES, Peter L. Antifreeze proteins of teleost fishes. **Annual Review of Physiology**, v. 63, n. 1, p. 359–390, 2001. Disponível em: <http://www.annualreviews.org/doi/10.1146/annurev.physiol.63.1.359>.

GERDAY, Charles; AITTALEB, Mohamed; BENTAHIR, Mostafa; *et al*. Cold-adapted enzymes: from fundamentals to biotechnology. **Trends in Biotechnology**, v. 18, n. 3, p. 103–107, 2000. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0167779999014134>.

GODINHO, Valéria M; FURBINO, Laura E; SANTIAGO, Iara F; *et al*. Diversity and bioprospecting of fungal communities associated with endemic and cold-adapted macroalgae in Antarctica. **The ISME Journal**, v. 7, n. 7, p. 1434–1451, 2013. Disponível em: <http://www.nature.com/articles/ismej201377>.

GODINHO, Valéria Martins; DE PAULA, Maria Theresa Rafaela; SILVA, Débora Amorim Saraiva; *et al*. Diversity and distribution of hidden cultivable fungi associated with marine animals of Antarctica. **Fungal Biology**, v. 123, n. 7, p. 507–516, 2019. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1878614618304057?via%3Dihub>.

GOLDMAN, Aaron David; LANDWEBER, Laura F. What Is a Genome? **PLOS Genetics**, v. 12, n. 7, p. e1006181, 2016. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/27442251>.

GOODWIN, Sara; MCPHERSON, John D; MCCOMBIE, W. Richard. Coming of age: ten years of next-generation sequencing technologies. **Nature Reviews Genetics**, v. 17, n. 6, p. 333–351, 2016.

GRAETHER, Steffen P.; KUIPER, Michael J.; GAGNÉ, Stéphane M.; *et al*. β-Helix structure and ice-binding properties of a hyperactive antifreeze protein from an insect. **Nature**, v. 406, n. 6793, p. 325–328, 2000. Disponível em: <http://www.nature.com/articles/35018610>.

GRAUR, Dan; ZHENG, Yichen; AZEVEDO, Ricardo B R. An evolutionary classification of genomic function. **Genome biology and evolution**, v. 7, n. 3, p. 642–5, 2015. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/25635041>.

GUO, Shuaiqi; GARNHAM, Christopher P.; WHITNEY, John C.; *et al*. Re-evaluation of a bacterial antifreeze protein as an adhesin with ice-binding activity. **PLoS ONE**, v. 7, n. 11, p. e48805, 2012. Disponível em: <http://dx.plos.org/10.1371/journal.pone.0048805>.

HALL, Mark; FRANK, Eibe; HOLMES, Geoffrey; *et al*. The WEKA data mining software. **ACM SIGKDD Explorations Newsletter**, v. 11, n. 1, p. 10, 2009. Disponível em: <http://portal.acm.org/citation.cfm?doid=1656274.1656278>.

HALLIDAY, C. L.; KIDD, S. E.; SORRELL, T. C.; *et al*. Molecular diagnostic methods for invasive fungal disease: the horizon draws nearer? **Pathology**, v. 47, n. 3, p. 257–269, 2015.

HASHIM, Noor Haza Fazlin; BHARUDIN, Izwan; NGUONG, Douglas Law Sie; *et al*. Characterization of Afp1, an antifreeze protein from the psychrophilic yeast *Glaciozyma antarctica* PI12. **Extremophiles**, v. 17, n. 1, p. 63–73, 2013. Disponível em: <http://link.springer.com/10.1007/s00792-012-0494-4>.

HE, Xue; HAN, Ke; HU, Jun; *et al*. TargetFreeze: Identifying Antifreeze Proteins via a Combination of Weights using Sequence Evolutionary Information and Pseudo Amino Acid Composition. **Journal of Membrane Biology**, v. 248, n. 6, p. 1005–1014, 2015. Disponível em: <http://link.springer.com/10.1007/s00232-015-9811-z>.

HEATHER, James M.; CHAIN, Benjamin. The sequence of sequencers: the history of sequencing DNA. **Genomics**, v. 107, n. 1, p. 1–8, 2016. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0888754315300410?via%3Dihub>.

HERRERA, Carlos M.; POZO, María I. Nectar yeasts warm the flowers of a winter-blooming plant. **Proceedings of the Royal Society B: Biological Sciences**, v. 277, n. 1689, p. 1827–1834, 2010. Disponível em: <https://royalsocietypublishing.org/doi/10.1098/rspb.2009.2252>.

HITESHI, Kalpana; GUPTA, Reena. Thermal adaptation of α-amylases: a review. **Extremophiles**, v. 18, n. 6, p. 937–944, 2014. Disp. em: http://link.springer.com/10.1007/s00792-014-0674-5.

HITTINGER, Chris Todd; ROKAS, Antonis; BAI, Feng-Yan; *et al*. Genomics and the making of yeast biodiversity. **Current opinion in genetics & development**, v. 35, p. 100–9, 2015. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/26649756>.

HOGEWEG, Paulien. The roots of bioinformatics in theoretical biology. **PLoS Computational Biology**, v. 7, n. 3, p. e1002021, 2011. Disponível em: <https://dx.plos.org/10.1371/journal.pcbi.1002021>.

HOLT, Carson; YANDELL, Mark. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. **BMC bioinformatics**, v. 12, p. 491, 2011. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/22192575>.

HUI, Feng Li; CHEN, Liang; LI, Zhi Hui; *et al*. *Metschnikowia henanensis* sp. nov., a new anamorphic yeast species isolated from rotten wood in China. **Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology**, v. 103, n. 4, p. 899–904, 2013.

ISAKSEN, Geir Villy; ÅQVIST, Johan; BRANDSDAL, Bjørn Olav. Protein Surface Softness Is the Origin of Enzyme Cold-Adaptation of Trypsin. **PLoS Computational Biology**, v. 10, n. 8, p. e1003813, 2014. Disponível em: <https://dx.plos.org/10.1371/journal.pcbi.1003813>.

JACOMY, Mathieu; VENTURINI, Tommaso; HEYMANN, Sebastien; *et al*. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. **PLoS ONE**, v. 9, n. 6, p. e98679, 2014. Disponível em: <https://dx.plos.org/10.1371/journal.pone.0098679>.

KALEDA, Aleksei; TSANEV, Robert; KLESMENT, Tiina; *et al*. Ice cream structure modification by ice-binding proteins. **Food Chemistry**, v. 246, p. 164–171, 2018. Disponível em: <https://doi.org/10.1016/j.foodchem.2017.10.152>.

KÄLL, Lukas; KROGH, Anders; SONNHAMMER, Erik L.L. A combined transmembrane topology and signal peptide prediction method. **Journal of Molecular Biology**, v. 338, n. 5, p. 1027–1036, 2004.

KANDASWAMY, Krishna Kumar; CHOU, Kuo-Chen; MARTINETZ, Thomas; *et al*. AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties. **Journal of Theoretical Biology**, v. 270, n. 1, p. 56–62, 2011. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0022519310005849?via%3Dihub>.

KELLEY, Joanna L.; PEYTON, Justin T.; FISTON-LAVIER, Anna-Sophie; *et al*. Compact genome of the Antarctic midge is likely an adaptation to an extreme environment. **Nature Communications**, v. 5, n. 1, p. 4611, 2014. Disponível em: <http://www.nature.com/articles/ncomms5611>.

KHAN, Shujaat; NASEEM, Imran; TOGNERI, Roberto; *et al*. RAFP-Pred: Robust Prediction of Antifreeze Proteins Using Localized Analysis of n-Peptide Compositions. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, 2018.

KIM, Minjae; GWAK, Yunho; JUNG, Woongsic; *et al*. Identification and characterization of an isoform antifreeze protein from the antarctic marine diatom, *Chaetoceros neogracile,* and suggestion of the core region. **Marine drugs**, v. 15, n. 10, 2017. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/29057803>.

KOCHKINA, G. A.; OZERSKAIA, S. M.; IVANUSHKINA, N. E.; *et al*. Fungal diversity in the Antarctic active layer. **Mikrobiologiia**, v. 83, n. 2, p. 236–244, 2014. Disponível em: <http://link.springer.com/10.1134/S002626171402012X>.

KOO, Hyunmin; HAKIM, Joseph A.; FISHER, Phillip R.E.; *et al*. Distribution of cold adaptation proteins in microbial mats in Lake Joyce, Antarctica: analysis of metagenomic data by using two bioinformatics tools. **Journal of Microbiological Methods**, v. 120, p. 23–28, 2016. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167701215301111?via%3Dihub>.

KORF, Ian. Gene finding in novel genomes. **BMC Bioinformatics**, v. 5, p. 1–9, 2004.

KOVAC, Damir; JELÍNEK, Josef; HASHIM, Rosli; *et al*. Transition from bamboo sap to water: aquatic habits in the sap beetle *Amphicrossus japonicu*s (Coleoptera: Cucujoidea: Nitidulidae). **European Journal of Entomology**, v. 104, n. 3, p. 635–638, 2007.

KOZUCH, Daniel J.; STILLINGER, Frank H.; DEBENEDETTI, Pablo G. Combined molecular dynamics and neural network method for predicting protein antifreeze activity. **Proceedings of the National Academy of Sciences**, v. 115, n. 52, p. 13252–13257, 2018. Disponível em: <http://www.pnas.org/lookup/doi/10.1073/pnas.1814945115>.

KURTZMAN, Cletus P.; FELL, Jack W.; BOEKHOUT, Teun. **The Yeast I**. [s.l.: s.n.], 2011.

KURTZMAN, Cletus P.; ROBNETT, Christie J. Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences. **Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology**, v. 73, n. 4, p. 331–371, 1998.

LACHANCE, M; DANIEL, H; MEYER, W; *et al*. The D1/D2 domain of the large-subunit rDNA of the yeast species is unusually polymorphic. **FEMS Yeast Research**, v. 4, n. 3, p. 253–258, 2003. Disponível em: <https://academic.oup.com/femsyr/article-lookup/doi/10.1016/S1567-1356(03)00113-2>.

LACHANCE, Marc-André. *Metschnikowia*: half tetrads, a regicide and the fountain of youth. **Yeast**, v. 33, n. 11, p. 563–574, 2016. Disponível em: <http://doi.wiley.com/10.1002/yea.3208>.

LACHANCE, Marc-Andre; BOEKHOUT, Teun; SCORZETTI, Gloria; *et al*. **The Yeast II**. [s.l.: s.n.], 2011.

LACHANCE, Marc André; HURTADO, Emilia; HSIANG, Tom. A stable phylogeny of the large-spored *Metschnikowia* clade. **Yeast**, v. 33, n. 7, p. 261–275, 2016.

LEE, Dong Kyung; HSIANG, Tom; LACHANCE, Marc-André. *Metschnikowia* mating genomics. **Antonie van Leeuwenhoek**, v. 111, n. 10, p. 1935–1953, 2018. Disponível em: <http://link.springer.com/10.1007/s10482-018-1084-y>.

LETUNIC, Ivica; BORK, Peer. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. **Nucleic Acids Research**, v. 44, n. W1, p. W242–W245, 2016. Disponível em: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw290>.

LI, Li; STOECKERT, Christian J Jr; ROOS, David S. OrthoMCL: identification of ortholog groups for eukaryotic genomes -- Li *et al*. 13 (9): 2178 -- Genome Research. **Genome Research**, v. 13, n. 9, p. 2178–2189, 2003. Disponível em: <http://genome.cshlp.org/cgi/content/full/13/9/2178>.

LOIBL, Martin; STRAHL, Sabine. Protein O-mannosylation: what we have learned from baker's yeast. **Biochimica et Biophysica Acta (BBA) - Molecular Cell Research**, v. 1833, n. 11, p. 2438–2446, 2013. Disponível em: <http://dx.doi.org/10.1016/j.bbamcr.2013.02.008>.

LOPATINA, Anna; MEDVEDEVA, Sofia; SHMAKOV, Sergey; *et al*. Metagenomic analysis of bacterial communities of antarctic surface snow. **Frontiers in Microbiology**, v. 7, 2016.

MARGESIN, Rosa. Psychrophiles: from Biodiversity to Biotechnology. Berlin, Heidelberg. **Springer Berlin Heidelberg**, 2008. Disponível em: <http://link.springer.com/10.1007/978-3-540-74335-4>.

MARINHA DO BRASIL. **Programa Antártico Brasileiro**. Disponível em: <https://www.marinha.mil.br/secirm/proantar>.

MARTIN, Shawn; BROWN, W. Michael; KLAVANS, Richard; *et al*. OpenOrd: an open-source toolbox for large graph layout. *In*: **Visualization and Data Analysis 2011**. [s.l.: s.n.], 2011, v. 7868, p. 786806. Disponível em: <http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.871402>.

MCKENNA, Duane D.; SHIN, Seunggwan; AHRENS, Dirk; *et al*. The evolution and genomic basis of beetle diversity. **Proceedings of the National Academy of Sciences of the United States of America**, v. 116, n. 49, p. 24729–24737, 2019.

MENDONÇA-HAGLER, L. C.; HAGLER, A. N.; PHAFF, H. J.; *et al*. DNA relatedness among aquatic yeasts of the genus *Metschnikowia* and proposal of the species *Metschnikowia australis* comb. nov. **Canadian Journal of Microbiology**, v. 31, n. 10, p. 905–909, 1985.

MICHETTI, Davide; BRANDSDAL, Bjørn Olav; BON, Davide; *et al*. A comparative study of cold and warm adapted Endonucleases – a using sequence analyses and molecular dynamics simulations. **PLOS ONE**, v. 12, n. 2, p. e0169586, 2017. Disponível em: <http://dx.plos.org/10.1371/journal.pone.0169586>.

MONDAL, Sukanta; PAI, Priyadarshini P. Chou′s pseudo amino acid composition improves sequence-based antifreeze protein prediction. **Journal of Theoretical Biology**, v. 356, p. 30–35, 2014. Disponível em: <https://www.sciencedirect.com/science/article/pii/S002251931400215X?via%3Dihub>.

MUÑOZ, Patricio A; MÁRQUEZ, Sebastián L; GONZÁLEZ-NILO, Fernando D; *et al*. Structure and application of antifreeze proteins from antarctic bacteria. **Microbial cell factories**, v. 16, n. 1, p. 138, 2017. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/28784139>.

NATIONAL SNOW & ICE DATA CENTER. **All about sea ice**. Disponível em: <https://nsidc.org/cryosphere/seaice/characteristics/difference.html>.

NAUMOV, G. I. Genus assignment of small-spored aquatic and terrestrial species of the *Metschnikowia* yeasts. **Microbiology**, v. 81, n. 2, p. 263–265, 2012. Disponível em: <http://link.springer.com/10.1134/S0026261712020075>.

OLIVEIRA, Eurico C.; ABSHER, Theresinha M.; PELLIZZARI, Franciane M.; *et al*. The seaweed flora of Admiralty Bay, King George Island, Antarctic. **Polar Biology**, v. 32, n. 11, p. 1639–1647, 2009. Disponível em: <http://link.springer.com/10.1007/s00300-009-0663-9>.

OMASITS, Ulrich; AHRENS, Christian H.; MÜLLER, Sebastian; *et al*. Protter: interactive protein feature visualization and integration with experimental proteomic data. **Bioinformatics**, v. 30, n. 6, p. 884–886, 2014. Disponível em: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt607>.

PARRA, G.; BRADNAM, K.; KORF, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. **Bioinformatics**, v. 23, n. 9, p. 1061–1067, 2007. Disponível em: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btm071>.

PATARNELLO, T.; BARGELLONI, L.; VAROTTO, V.; *et al*. Krill evolution and the Antarctic ocean currents: Evidence of vicariant speciation as inferred by molecular data. **Marine Biology**, v. 126, n. 4, p. 603–608, 1996.

PEARSON, Gareth A.; LAGO-LESTON, Asuncion; CÁNOVAS, Fernando; *et al*. Metatranscriptomes reveal functional variation in diatom communities from the Antarctic Peninsula. **ISME Journal**, v. 9, n. 10, p. 2275–2289, 2015.

PERTAYA, Natalya; MARSHALL, Christopher B.; CELIK, Yeliz; *et al*. Direct visualization of spruce budworm antifreeze protein interacting with ice crystals: Basal plane affinity confers hyperactivity. **Biophysical Journal**, v. 95, n. 1, p. 333–341, 2008.

PETROV, Anton S.; WOOD, Elizabeth C.; BERNIER, Chad R.; *et al*. Structural patching fosters divergence of mitochondrial ribosomes. **Molecular Biology and Evolution**, v. 36, n. 2, p. 207–219, 2019.

PORTA, Amalia; FORTINO, Vittorio; ARMENANTE, Annunziata; *et al*. Cloning and characterization of a Δ9-desaturase gene of the Antarctic fish *Chionodraco hamatus* and *Trematomus bernacchii*. **Journal of Comparative Physiology B**, v. 183, n. 3, p. 379–392, 2013. Disponível em: <http://link.springer.com/10.1007/s00360-012-0702-7>.

PRATIWI, Reny; MALIK, Aijaz Ahmad; SCHADUANGRAT, Nalini; *et al*. CryoProtect: a web server for classifying antifreeze proteins from nonantifreeze proteins. **Journal of Chemistry**, v. 2017, p. 1–15, 2017. Disponível em: <https://www.hindawi.com/journals/jchem/2017/9861752/>.

PYNE, A.; GOLLEDGE, N. R.; NAISH, T. R.; *et al*. Antarctic Cenozoic climate history from sedimentary records: ANDRILL and beyond. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, v. 374, n. 2059, p. 20140301, 2015.

ROSA, Carlos A.; JINDAMORAKOT, Sasitorn; LIMTONG, Savitree; *et al*. *Candida golubevii* sp. nov., an asexual yeast related to *Metschnikowia lunata*. **INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY**, v. 60, n. 3, p. 704–706, 2010. Disponível em: <https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.014050-0>.

ROSA, Luiz Henrique. Fungi of Antarctica. **Springer International Publishing**, 2019. Disponível em: <http://link.springer.com/10.1007/978-3-030-18367-7>.

SANGER, F; NICKLEN, S; COULSON, A R. DNA sequencing with chain-terminating inhibitors. **Proceedings of the National Academy of Sciences of the United States of America**, v. 74, n. 12, p. 5463–7, 1977. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/271968>.

SANTIAGO, Margarita; RAMÍREZ-SARMIENTO, César A.; ZAMORA, Ricardo A.; *et al*. Discovery, molecular mechanisms, and industrial applications of cold-active enzymes. **Frontiers in Microbiology**, v. 7, 2016. Disponível em: <http://journal.frontiersin.org/Article/10.3389/fmicb.2016.01408/abstract>.

SANTOS, Ana Raquel de Oliveira; LEE, Dong Kyung; FERREIRA, Andressa Graebin; *et al*. The yeast community of *Conotelus* sp. (Coleoptera: Nitidulidae) in Brazilian passionfruit flowers (*Passiflora edulis*) and description of *Metschnikowia amazonensis* sp. nov., a large-spored clade yeast. **Yeast**, 2020. Disponível em: <http://doi.wiley.com/10.1002/yea.3453>.

SCHADE, Babette; JANSEN, Gregor; WHITEWAY, Malcolm; *et al*. Cold Adaptation in Budding Yeast. **Molecular Biology of the Cell**, v. 15, n. 12, p. 5492–5502, 2004. Disponível em: <https://www.molbiolcell.org/doi/10.1091/mbc.e04-03-0167>.

SCHOCH, Conrad L; SEIFERT, Keith A; HUHNDORF, Sabine; *et al*. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. **Proceedings of the National Academy of Sciences of the United States of America**, v. 109, n. 16, p. 6241–6, 2012. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/22454494>.

SCOTTER, Andrew J.; MARSHALL, Christopher B.; GRAHAM, Laurie A.; *et al*. The basis for hyperactivity of antifreeze proteins. **Cryobiology**, v. 53, n. 2, p. 229–239, 2006. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0011224006001106>.

SELBMANN, Laura; TURCHETTI, Benedetta; YURKOV, Andrey; *et al*. Description of *Taphrina antarctica* f.a. sp. nov., a new anamorphic ascomycetous yeast species associated with Antarctic endolithic microbial communities and transfer of four *Lalaria* species in the genus *Taphrina*. **Extremophiles**, v. 18, n. 4, p. 707–721, 2014. Disponível em: <http://link.springer.com/10.1007/s00792-014-0651-z>.

SHEN, Xing-Xing; HITTINGER, Chris Todd; ROKAS, Antonis. Contentious relationships in phylogenomic studies can be driven by a handful of genes. **Nature Ecology & Evolution**, v. 1, n. 5, p. 0126, 2017. Disponível em: <http://www.nature.com/articles/s41559-017-0126>.

SHEN, Xing-Xing; OPULENTE, Dana A.; KOMINEK, Jacek; *et al*. Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. **Cell**, v. 175, n. 6, p. 1533-1545.e20, 2018. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0092867418313321?via%3Dihub>.

SILVA, Tiago R.; TAVARES, Renata S. N.; CANELA-GARAYOA, Ramon; *et al*. Chemical characterization and biotechnological applicability of pigments isolated from antarctic bacteria. **Marine Biotechnology**, v. 21, n. 3, p. 416–429, 2019. Disponível em: <http://link.springer.com/10.1007/s10126-019-09892-z>.

SIMÃO, Felipe A.; WATERHOUSE, Robert M.; IOANNIDIS, Panagiotis; *et al*. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. **Bioinformatics**, v. 31, n. 19, p. 3210–3212, 2015.

SMIT, AFA, HUBLEY, R & GREEN, P. RepeatMasker Open-4.0. 2013. Disponível em: <http://www.repeatmasker.org>.

STAMATAKIS, Alexandros. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. **Bioinformatics**, v. 30, n. 9, p. 1312–1313, 2014.

STANKE, Mario; MORGENSTERN, Burkhard. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. **Nucleic Acids Research**, v. 33, p. W465–W467, 2005. Disponível em: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gki458>.

STIELOW, J. B.; LÉVESQUE, C. A.; SEIFERT, K. A.; *et al*. One fungus, which genes? Development and assessment of universal primers for potential secondary fungal DNA barcodes. **Persoonia: Molecular Phylogeny and Evolution of Fungi**, v. 35, n. 1, p. 242–263, 2015.

SUGAWARA, Etsuko; NIKAIDO, Hiroshi. Yeast biotechnology: diversity and applications. **Dordrecht: Springer Netherlands**, 2009. Disponível em: <http://aac.asm.org/lookup/doi/10.1128/AAC.03728-14>.

TAYLOR, Michael J.; WEEGMAN, Bradley P.; BAICU, Simona C.; *et al*. New approaches to cryopreservation of cells, tissues, and organs. **Transfusion Medicine and Hemotherapy**, v. 46, n. 3, p. 197–215, 2019.

TEIXIDO, N.; VIÑAS, I.; USALL, J.; *et al*. Improving ecological fitness and environmental stress tolerance of the biocontrol yeast *Candida* sake by manipulation of intracellular sugar alcohol and sugar content. **Mycological Research**, v. 102, n. 11, p. 1409–1417, 1998. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0953756208610394>.

TER-HOVHANNISYAN, Vardges; LOMSADZE, Alexandre; CHERNOFF, Yury O.; *et al*. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. **Genome Research**, v. 18, n. 12, p. 1979–1990, 2008.

TSAI, Sujune; CHONG, Gabriella; MENG, Pei-Jie; *et al*. Sugars as supplemental cryoprotectants for marine organisms. **Reviews in Aquaculture**, v. 10, n. 3, p. 703–715, 2018. Disponível em: <http://doi.wiley.com/10.1111/raq.12195>.

TSUJI, Masaharu. Cold-stress responses in the antarctic basidiomycetous yeast *Mrakia blollopis*. **Royal Society Open Science**, v. 3, n. 7, 2016.

TURCHETTI, Benedetta; THOMAS HALL, Skye R.; CONNELL, Laurie B.; *et al*. Psychrophilic yeasts from Antarctica and European glaciers: description of *Glaciozyma* gen. nov., *Glaciozyma*

martinii sp. nov. and *Glaciozyma watsonii* sp. nov. **Extremophiles**, v. 15, n. 5, p. 573–586, 2011. Disponível em: <http://link.springer.com/10.1007/s00792-011-0388-x>.

USMAN, Muhammad; LEE, Jeong A. AFP-CKSAAP: Prediction of antifreeze proteins using composition of k-Spaced amino acid pairs with deep neural network. **arXiv**, p. 1–6, 2019. Disponível em: <http://arxiv.org/abs/1910.06392>.

VAZ, Aline B. M; ROSA, Luiz H; VIEIRA, Mariana L. A; *et al*. The diversity, extracellular enzymatic activities and photoprotective compounds of yeasts isolated in Antarctica. **Brazilian Journal of Microbiology**, v. 42, n. 3, p. 937–947, 2011. Disponível em: <http://www.scielo.br/pdf/bjm/v42n3/v42n3a12.pdf>.

VESTER, Jan Kjølhede; GLARING, Mikkel Andreas; STOUGAARD, Peter. An exceptionally cold-adapted alpha-amylase from a metagenomic library of a cold and alkaline environment. **Applied Microbiology and Biotechnology**, v. 99, n. 2, p. 717–727, 2015.

VILLARREAL, Pablo; CARRASCO, Mario; BARAHONA, Salvador; *et al*. Antarctic yeasts: Analysis of their freeze-thaw tolerance and production of antifreeze proteins, fatty acids and ergosterol. **BMC Microbiology**, v. 18, n. 66, p. 1–10, 2018. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/29976143>.

VINCENT, Warwick F. Evolutionary origins of antarctic microbiota: invasion, selection and endemism. **Antarctic Science**, v. 12, n. 3, p. 374–385, 2000.

WARREN, Gareth; COROTTO, Loren. The consensus sequence of ice nucleation proteins from *Erwinia herbicola*, *Pseudomonas fluorescens* and *Pseudomonas syringae*. **Gene**, v. 85, n. 1, p. 239–242, 1989. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/0378111989904885>.

WENTZEL, Lia Costa Pinto; INFORSATO, Fábio José; MONTOYA, Quimi Vidaurre; *et al*. Fungi from Admiralty Bay (King George Island, Antarctica) Soils and Marine Sediments. **Microbial Ecology**, p. 1–13, 2018.

WICKERHAM, Lynferd J. A preliminary report on a Perfect family of exclusively protosexual yeasts. **Mycologia**, v. 56, n. 2, p. 253–266, 1964. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/00275514.1964.12018107>.

WILKINS, David; YAU, Sheree; WILLIAMS, Timothy J.; *et al*. Key microbial drivers in Antarctic aquatic environments. **FEMS Microbiology Reviews**, v. 37, n. 3, p. 303–335, 2013. Disponível em: <https://academic.oup.com/femsre/article-lookup/doi/10.1111/1574-6976.12007>.

XIAO, Xuan; HUI, Mengjuan; LIU, Zi. iAFP-Ense: An Ensemble Classifier for Identifying Antifreeze Protein by Incorporating Grey Model and PSSM into PseAAC. **Journal of Membrane Biology**, v. 249, n. 6, p. 845–854, 2016. Disponível em: <http://link.springer.com/10.1007/s00232-016-9935-9>.

XUE, Meng Lin; ZHANG, Li Qun; WANG, Qi Ming; *et al*. *Metschnikowia sinensis* sp. nov., *Metschnikowia zizyphicola* sp. nov. and *Metschnikowia shanxiensis* sp. nov., novel yeast species from jujube fruit. **International Journal of Systematic and Evolutionary Microbiology**, v. 56, n. 9, p. 2245–2250, 2006.

YAJIMA, Yuka; TOJO, Motoaki; CHEN, Bo; *et al*. *Typhula* cf. *subvariabilis*, new snow mould in Antarctica. **Mycology**, v. 8, n. 3, p. 147–152, 2017. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/21501203.2017.1343753>.

YANG, Runtao; ZHANG, Chengjin; GAO, Rui; *et al*. An Effective Antifreeze Protein Predictor with Ensemble Classifiers and Comprehensive Sequence Descriptors. **International Journal of Molecular Sciences**, v. 16, n. 9, p. 21191–21214, 2015. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/26370959>.

YE, Jian; COULOURIS, George; ZARETSKAYA, Irena; *et al*. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. **BMC Bioinformatics**, v. 13, n. 1, p. 134, 2012. Disponível em: <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-134>.

YU, Chin-Sheng; LU, Chih-Hao. Identification of Antifreeze Proteins and Their Functional Residues by Support Vector Machine and Genetic Algorithms based on n-Peptide Compositions. **PLoS ONE**, v. 6, n. 5, p. e20445, 2011. Disponível em: <https://dx.plos.org/10.1371/journal.pone.0020445>.

ZHANG, Chengjun; WANG, Jun; LONG, Manyuan; *et al*. GKaKs: The pipeline for genome-level Ka/Ks calculation. **Bioinformatics**, v. 29, n. 5, p. 645–646, 2013. Disponível em: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt009>.

ZHAO, Xiaowei; MA, Zhiqiang; YIN, Minghao. Using Support Vector Machine and Evolutionary Profiles to Predict Antifreeze Protein Sequences. **International Journal of Molecular Sciences**, v. 13, n. 2, p. 2196–2207, 2012. Disponível em: <www.mdpi.com/journal/ijms>.

ZHU, Yuan O; SIEGAL, Mark L; HALL, David W; *et al*. Precise estimates of mutation rate and spectrum in yeast. **Proceedings of the National Academy of Sciences of the United States of America**, v. 111, n. 22, p. E2310-8, 2014. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/24847077>.

ZIEMERT, Nadine; ALANJARY, Mohammad; WEBER, Tilmann. The evolution of genome mining in microbes – a review. **Natural Product Reports**, v. 33, n. 8, p. 988–1005, 2016. Disponível em: <http://xlink.rsc.org/?DOI=C6NP00025H>.

ANEXO I – Artigo Científico

Draft Genome Sequence of *Metschnikowia australis* Strain UFMG-CM-Y6158, an Extremophile Marine Yeast Endemic to Antarctica

AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements™

EUKARYOTES

Draft Genome Sequence of *Metschnikowia australis* Strain UFMG-CM-Y6158, an Extremophile Marine Yeast Endemic to Antarctica

Thiago M. Batista,[a] Heron O. Hilário,[a] Rennan G. Moreira,[c] Carolina Furtado,[d] Valéria M. Godinho,[b] Luiz H. Rosa,[b] Glória R. Franco,[a] Carlos A. Rosa[b]

Departamento de Bioquímica e Imunologia, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil[a]; Departamento de Microbiologia, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil[b]; Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil[c]; Instituto Nacional do Câncer, Rio de Janeiro, Brazil[d]

ABSTRACT Here we report the draft genome sequence of *Metschnikowia australis* strain UFMG-CM-Y6158, a yeast endemic to Antarctica. We isolated the strain from the marine seaweed *Acrosiphonia arcta* (*Chlorophyta*). The genome is 14.3 Mb long and contains 4,442 predicted protein-coding genes.

Received 17 March 2017 **Accepted** 30 March 2017 **Published** 18 May 2017

Citation Batista TM, Hilário HO, Moreira RG, Furtado C, Godinho VM, Rosa LH, Franco GR, Rosa CA. 2017. Draft genome sequence of *Metschnikowia australis* strain UFMG-CM-Y6158, an extremophile marine yeast endemic to Antarctica. Genome Announc 5:e00328-17. https://doi.org/10.1128/genomeA.00328-17.

Address correspondence to Carlos A. Rosa, carlrosa@icb.ufmg.br.

The genus *Metschnikowia* comprises a clade consisting of approximately 81 species. The sexual life cycles of the members of this clade involve the formation of elongated asci containing two, often needle-shaped, spores (1). *M. australis* is a species endemic to Antarctica, and has been isolated from seawater, marine invertebrates, sponges, and macroalgae (2–6). Owing to the extremely cold environment of Antarctica, *M. australis* may have unique metabolic traits enabling it to survive under such stressful conditions; exploring these can help identify potential antifreeze compounds for biotechnological use.

We isolated *M. australis* strain UFMG-CM-Y6158 from a marine macroalgae, *Acrosiphonia arcta* (*Chlorophyta*), collected in Admiralty Bay of King George Island in Keller Peninsula, Antarctica (5). We cultivated the strain on marine agar (Himedia, India) at 10°C for 15 days, and the genomic DNA was isolated by phenol:chloroform (1:1) extraction. We assessed DNA quality by gel electrophoresis and determined its purity and quantity using both the NanoDrop 1000 UV-Vis spectrophotometer and the Qubit version 2.0 fluorometer with the Qubit dsDNA HS assay kit (Thermo Fisher Scientific). We used the Nextera XT DNA kit (Illumina) to construct paired-end libraries and assessed their quality using Bioanalyzer HS Assay (Agilent Technologies). Generated fragments with a mean length of 1,167 bp were sequenced using the Illumina MiSeq sequencer, whereas those with a mean length of 550 bp were sequenced using the Illumina HiSeq 2500 sequencer. The former generated 1,585,122 reads (2 × 301) with 35× coverage, while the latter generated 103,312,458 reads (2 × 101) with 745× coverage. We assembled the genome using SPAdes version 3.9.1 (7). The assembled draft genome consisted of 14,356,710 bp over 160 contigs (>505 bp) with a G+C content of 47.2%. The longest contig was 1,116,518 bp long, and the N_{50} contig length was 542,232 bp. CEGMA (8) analysis showed that the assembly was 95.9% complete, whereas analysis with BUSCO version 2 (9) using the *Saccharomycetales* lineage data set indicated 90.2% completeness based on the presence of conserved orthologous genes among species of the genus. We identified 4,442 protein-coding genes using MAKER2 (10). A sequence similarity search using the BLASTx tool in BLAST version 2.2.31+ (11) returned 4,348 protein matches (97.8%), with e-value ≤1e^{-6}, against NCBI's nonredundant database. We identified 249 tRNAs using tRNAscan-SE (12).

Using the OrthoVenn web platform (13), we compared *M. australis* protein-coding genes with those of two previously sequenced *Metschnikowia* genomes—*M. fructicola* and *M. bicuspidata*. The analysis showed that *M. australis* has a much shorter predicted proteome than that of *M. fructicola* (5,851 protein-coding genes) and *M. bicuspidata* (6,028 protein-coding genes). Additionally, we found six exclusive clusters of paralogous genes, of which four did not match any protein in the NCBI and UniProt-Swissprot databases. These results highlight the importance of investigating yeast endemic to Antarctica, such as *M. australis*, not only to identify novel genes associated with adaptation to extreme environments, but also for potential application in biotechnology.

Accession number(s). Data related to this whole-genome shotgun project have been deposited at DDBJ/ENA/GenBank under the accession number MVNQ00000000. The version described in this paper is the first version, MVNQ01000000.

ACKNOWLEDGMENTS

Laboratório Multiusuário de Bioinformática—EMBRAPA Informática Agropecuária, Campinas, Brazil, provided access to genome annotation. This work was supported by Conselho Nacional de Desenvolvimento Cientifico e Tecnológico (CNPq-PROANTAR 407230/2013-0, INCT Criosfera, 457499/2014-1); Fundação de Apoio à Pesquisa do Estado de Minas Gerais (FAPEMIG-0050-13, APQ-01525-14); and CAPES (23038.003478/2013-92).

REFERENCES

1. Lachance MA, Hurtado E, Hsiang T. 2016. A stable phylogeny of the large-spored *Metschnikowia* clade. Yeast 33:261–275. https://doi.org/10.1002/yea.3163.
2. Loque CP, Medeiros AO, Pellizzari FM, Oliveira EC, Rosa CA, Rosa LH. 2010. Fungal community associated with marine macroalgae from Antarctica. Polar Biol 33:641–648. https://doi.org/10.1007/s00300-009-0740-0.
3. Lachance M-A. 2011. *Metschnikowia* Kamienski (1899), p 575–620. *In* Kurtzman CP, Fell JW, Boekhout T (ed), The yeasts: a taxonomic study, 5th ed, vol 1. Elsevier, London.
4. Vaca I, Faúndez C, Maza F, Paillavil B, Hernández V, Acosta F, Levicán G, Martínez C, Chávez R. 2013. Cultivable psychrotolerant yeasts associated with Antarctic marine sponges. World J Microbiol Biotechnol 29:183–189. https://doi.org/10.1007/s11274-012-1159-2.
5. Godinho VM, Furbino LE, Santiago IF, Pellizzari FM, Yokoya NS, Pupo D, Alves TM, Junior PA, Romanha AJ, Zani CL, Cantrell CL, Rosa CA, Rosa LH. 2013. Diversity and bioprospecting of fungal communities associated with endemic and cold-adapted macroalgae in Antarctica. ISME J 7:1434–1451. https://doi.org/10.1038/ismej.2013.77.
6. Furbino LE, Godinho VM, Santiago IF, Pellizari FM, Alves TMa, Zani CL, Junior PA, Romanha AJ, Carvalho AG, Gil LH, Rosa CA, Minnis AM, Rosa LH. 2014. Diversity patterns, ecology and biological activities of fungal communities associated with the endemic macroalgae across the Antarctic peninsula. Microb Ecol 67:775–787. https://doi.org/10.1007/s00248-014-0374-9.
7. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.
8. Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23:1061–1067. https://doi.org/10.1093/bioinformatics/btm071.
9. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351.
10. Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12:491. https://doi.org/10.1186/1471-2105-12-491.
11. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. Blast+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421.
12. Lowe TM, Eddy SR. 1997. TRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964. https://doi.org/10.1093/nar/25.5.0955.
13. Wang Y, Coleman-Derr D, Chen G, Gu YQ. 2015. OrthoVenn: a web server for genome wide comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Res 43:W78–W84. https://doi.org/10.1093/nar/gkv487.

ANEXO II – Capítulo de livro

Fungi of Antarctica

Genomics of Antarctic Fungi: A New Frontier

Chapter 15
Genomics of Antarctic Fungi: A New Frontier

Heron Oliveira Hilário, Thiago Mafra Batista, and Glória Regina Franco

15.1 Introduction

Despite the extreme conditions of the Antarctic continent, life has managed to survive. Microorganisms are found in every Antarctic ecosystem (Godinho et al. 2013) from the ocean and ancient lakes to fresh snow, from lichens on rocks to the permafrost and ornithogenic soils (Duarte et al. 2016), and from prehistoric ice deposits to deep sea marine sediments (Vaz et al. 2011). The extreme Antarctic conditions – freezing cycles, low liquid water availability, and high UV incidence – have selected species with diverse unique features and adaptations in all cellular processes and structures, making the Antarctic biodiversity a reservoir of unique organisms and genes that could be exploited by the biotechnology industry (Godinho et al. 2013).

The microbiological, biochemical, and phylogenetic characterization of these Antarctic organisms and their features has been the main approach used to date (Wentzel et al. 2018; Yajima et al. 2017), followed by investigation of individual genes, in particular, those encoding antifreeze proteins and cold-adapted enzymes (Hashim et al. 2018). Further, bioprospection of extracts derived from cultivated fungi seems to be promising in the search for novel drugs with anti-pathogen and anticancer activities (Godinho et al. 2013).

H. O. Hilário (✉) · G. R. Franco
Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brasil
e-mail: heron-oh@ufmg.br

T. M. Batista (✉)
Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brasil

Centro de Formação em Ciências Ambientais, Universidade Federal do Sul da Bahia, Porto Seguro, Bahia, Brasil
e-mail: thiagomafra@ufsb.edu.br

L. H. Rosa (ed.), *Fungi of Antarctica*,
https://doi.org/10.1007/978-3-030-18367-7_15

Recent advances in the techniques for DNA analysis, especially the advent of next-generation sequencing (NGS) (Goodwin et al. 2016) and improvements in bioinformatics tools, have enabled a new and complementary approach for the characterization of Antarctic organisms. The rationale is to investigate the genomes and transcriptomes of these organisms in the search for genes responsible for specific phenotypic traits (Firdaus-Raih et al. 2018) and also to peek into the molecular universe of the cell to determine unexpected features that are only accessible at the nucleic acid sequence level. This chapter is not intended as a detailed manual for genome analysis. Instead, its goal is to familiarize the reader with the concepts in the field and the basis for genome interrogation.

15.2 The Omics Era: Genomics and Other Omics

The term genome was coined in 1920, much before the recognition of DNA as the genetic information storage molecule, and was meant to designate the chromosomes and protoplasm (Goldman and Landweber 2016). With the rise of DNA sequencing in 1977 (Sanger et al. 1977), investigation of phenotypic traits at the molecular level was primarily performed on one or a few genes at a time. The last decades have witnessed rapid development of molecular biology and DNA sequencing platforms. Miniaturization allied to massive parallelization in high-throughput strategies have facilitated the extension of the context of sequencing from one gene to all genes (Heather and Chain 2016; Quail et al. 2012). Gathering of knowledge on biological gene structures and functions organized in large curated databases and the possibility to sequence the whole genome of an organism at once have given rise to the field of *genomics*.

More than all the genes of an organism put together, the genome is the set of all DNA molecules present in a cell (except for RNA viruses) organized as chromosomes. It is an ensemble of sequences comprising protein-coding genes and their regulatory sequences, non-coding RNA genes, pseudogenes, and protogenes (Dujon and Louis 2017). The genome also harbours repetitive sequences and many, only apparently, functionless stretches of DNA (Graur et al. 2015) along with informative regions, which carry the organism's evolutionary history. The genome is, in most cases, static during the cell lifetime. Some stresses or genotoxic agents can produce mutations that, if not corrected by the highly efficient DNA repair machinery, might be passed on through generations. Besides being considered as the molecule carrying the genetic information, DNA needs accessory proteins and RNAs to get its information interpreted in the context of an organism or cell.

The ability to study genomes is catalysing many research fields (Hittinger et al. 2015), virtually all fields related to biological sciences and, in particular, microbiology (Ziemert et al. 2016). The knowledge on all life domains is benefited by the capability of looking at any organism through the genomic lenses. This deeper vision has clarified and redefined taxonomic relationships and classifications (Choi and Kim 2017), elucidated ecological functions at the genetic

level (Chan et al. 2013), and made it possible to observe organisms that are invisible to classical approaches – the microbial dark matter – even in Antarctica (Bernard et al. 2018).

Derived from *genomics*, the field of *metagenomics* uses modern sequencing techniques to address the diversity of an ecosystem. The prefix "meta" refers to something beyond. Thus, metagenomics aims to analyse more than the genome of a given species, the complete set of genetic material present in a sample. By sequencing DNA from environmental or other complex samples and comparing it with known sequences from curated databases, it is possible to infer the composition and abundance of distinct taxa.

Composite samples can be assessed using two main strategies. The first is called *amplicon metagenomics*. In a similar manner as performed in *barcoding* analysis, PCR primers are designed to amplify the informative regions expected to be present in the investigated clades. For fungi, the most common regions are D1/D2 from the 18S ribosomal DNA (Baeza et al. 2017). Baeza et al. (2017) used these markers to perform amplicon metagenomic analysis in order to estimate the fungal diversity of Antarctica's terrestrial habitats, and identified 87 known genera and 123 species of 37 unknown genera. The most represented fungal classes were *Lecaronomycetes* and *Eurotiomycetes*.

The difference between the classical amplicon analysis with isolated organisms and *metagenomics* is the ability of the latter approach to process complex samples and infer the abundance of identified organisms from their proportion in the obtained sequencing reads. The drawback is that, for ecosystems where the biodiversity is little known, as for some Antarctic habitats, many sequences will not be identified, being only assigned to higher taxonomic ranks. Yet, with the constant advances in the gathering of biological information from Antarctica, this obstacle could be overcome in the near future.

The second strategy is called *shotgun metagenomics*. Its name is derived from the analogy to the genome sequencing strategy devised by Craig Venter in the race for the Human Genome Project (Venter et al. 2001), i.e. *shotgun sequencing*. Prior to its development, the current approach used to assemble genomes was the "contig strategy". To sequence long DNA stretches, one would work in an ordered manner by fragmenting the genome to different sizes and cloning its pieces into vectors with distinct insert size capabilities to produce libraries that were later organized by sequencing their insert ends. The strategy to reconstruct the genome was to use overlapping clones to produce a physical map, guiding the genome assembly, in a top-down approach.

The *shotgun* strategy, supported by Venter, was based on cutting the genome in a random manner using physical or enzymatic methods, as if shattering it with a shotgun shot, sequencing all the pieces and using computational power to determine overlapping parts in order to guide the contigs assembly, in a bottom-up approach. This strategy later proved to be the more efficient, especially when allied with modern sequencing capabilities. So, in *shotgun metagenomics*, all the DNA retrieved from a given sample is randomly sequenced, without any previous amplification step. The output data is a mixture of diverse fragments, corresponding to pieces of

the genomes present in the sample. More abundant organisms yield more DNA, generating more reads. For some organisms with small genomes, like viruses and some bacteria, it is sometimes possible to assemble entire genomes from these data. For others, with bigger genomes or that are proportionally less abundant in the sample, less data is retrieved, hindering full genome assembly. The advantage of *shotgun metagenomics* is that phylogenetically informative regions can be assessed along with virtually any part of the genome. From the pieces of the reconstituted genome, it is possible to assign taxonomic classification and also to infer an organism's role in the environment, as the presence of genes that perform specific molecular tasks denotes the ecological niche and functionality of the organism (Sharpton 2014; Calderoli et al. 2018).

The promising frontier of environmental DNA analysis has been largely explored recently, boosting the knowledge on Antarctic bacterial and viral diversity (Lloyd et al. 2018). As an example of the importance of such investigation, Koo et al. (2016) analysed metagenomic samples of six Antarctic microbial mats searching for the presence and distribution of bacterial cold adaptation proteins: antifreeze proteins (AFPs), ice nucleating proteins (INPs), cold shock proteins (CSPs), trehalose synthase (TA), and fatty acid desaturases (FADs). They found many CSPs, TAs, and FADs in all collected samples. However, the INPs and AFPs were less abundant, corroborating the fact that ice was not a constant selective pressure compared to low temperature, because the lakes from where the microbial mats were derived experience only mild freezing at localized sites.

Fungi, as organisms with genomes up to hundred times larger than those of bacteria, archaea, and virus, are harder to study and are usually not at the focus of most shotgun metagenomic studies conducted nowadays. However, an interesting feature of metagenomics is that as all the DNA of a sample is sequenced, information is gathered for all the constituent organisms, regardless of the groups that are the aim of the study. This facilitates investigation of sequences produced for a particular study with a different perspective compared to the original. As many metagenomic raw sequences are being deposited to public databases, a wealth of information is waiting to be mined with different optics. This richness was exploited by Donovan et al. (2018), who developed a pipeline to investigate public metagenomic data on fungal diversity. Thirteen metagenomic datasets of Antarctic soil, derived from studies on bacterial diesel degradation, were analysed. From those, 4.91% of the reads were assigned to the genus *Pseudogymnoascus*. They also postulated that some part of the oil-degrading capability was featured by these fungi, as one species found in two datasets, *P. pannorum*, had already been linked to diesel oil biodegradation in the Amazon.

Another facet of metagenomics is called *functional metagenomics*. In this approach, pieces of environmental DNA are randomly introduced into bacteria or yeast for heterologous expression. Colonies are later screened for a particular capability of interest, allowing identification of molecular activities without even knowing the donor organism. This strategy was used by Berlemont et al. (2013) to produce and screen a metagenomic library from Antarctic soil for lipase activity. A novel enzyme, Mhlip, was identified, characterized, and shown to be adapted to the

cold Antarctic conditions. Ferrés et al. (2015) also used this strategy to search metagenomic samples of glacial melt water for lipase/esterase, cellulase, and manganese oxidase activities. Additionally, an alkalophilic esterase was identified from Antarctic desert soil samples (Hu et al. 2012). These classes of enzymes are easily identified by functional metagenomic analyses because the activity assays are simple and low cost.

To be functional, any gene, protein-coding or not, must be transcribed from the genome to RNA. So, *transcriptomics* refers to the study of the transcriptome, the set of transcripts produced by a cell in a given moment. Its composition is directly influenced by the cell cycle, developmental stage, environmental conditions, and stresses, among other factors. It is extremely plastic and changes according to different circumstances. RNAs perform diverse roles in the cell: they are structural molecules, code for proteins, regulate chromatin modification, transcription, stability of transcripts, translation, and many other cellular processes. RNA molecules range from tens to thousands of bases, and are organized in classes related to their coding potential, size, and functionality (Wang et al. 2009). Investigation of an organism's transcriptome is paramount to identify the genetic agents underlying a response by comparing the gene expression profile before and after a specific stimulus.

Further, in an analogous way, *metatranscriptomics* refers to strategies where the RNA is assessed directly from environmental samples. It is similar to metagenomics, but this approach pictures the expressed genes involved in processes that are actually being carried out by the microbial community, whereas *metagenomics* pictures any parts of the composing genomes, regardless of their gene content or transcriptional state. From the assembled transcripts is also possible to identify the organisms of origin and reconstitute the microbial community diversity, based on sequence comparison (Bashiardes et al. 2016). This approach was used to compare three diatom communities of the Antarctic Peninsula for functional differences by Pearson et al. (2015), but its application in Antarctic fungi remains to be demonstrated.

15.3 Sequencing: From Molecules to Data

To investigate and study the genome or transcriptome of any given organism, it is necessary to sequence its DNA or RNA. Sequencing is the process by which DNA/RNA molecules have their information decoded and digitalized, allowing their computational manipulation. The last decades have witnessed the rise and improvement of many different sequencing strategies, each with its pros and cons, but all participating in the field's revolution.

The first sequencing platforms were based on the chain termination method, also called dideoxy method, developed by Frederick Sanger in 1977 (Sanger et al. 1977). Dideoxynucleotide triphosphate molecules (ddNTPs) are nucleotides lacking the hydroxyl group to which the next base to be incorporated in the elongating strand

would attach. In this method, the DNA to be sequenced is amplified in four separate PCR reactions, each one with conventional nucleotides (dNTPs) plus a small fraction of one of the distinct ddNTPs – ddATP, ddCTP, ddTTP, or ddGTP – which, when added, terminate the elongation of the newly synthesized strand. As each ddNTP incorporation happens only once on each newly produced molecule, at the end of the reaction, many DNA strands with different lengths are generated, each corresponding to every possible strand size. The amplification products are then resolved by electrophoresis in a denaturing acrylamide gel where the material of each of the four reaction tubes is run separately. The nucleotide sequence is inferred from the position of the bands in each of the four lanes. The platforms based on this method were automated to detect the DNA fragments with lasers, as long as the products composed of one fluorescent nucleotide migrated through the gel. Moreover, the labelling of each ddNTP with a distinct fluorescence tag made it possible to amplify and run the four reactions together in a single lane (Smith et al. 1986). These automated sequencers were capable of generating intermediate size reads from 300 to 1000 bases long, with relatively low error rates and are still used nowadays in cases where few sequences are to be investigated, or to aid in short reads assembly.

The Sanger sequencing method opened the doors for sequence-level genetic analysis but was not powerful enough for applications such as eukaryotic genome sequencing, making it extremely labour-intensive and expensive. The Human Genome Project triggered the race for the development of other derivate techniques devised to overcome sensitivity, throughput, and cost limits, aiming at lowering the price to sequence a full human genome at USD 1000 or less. The new techniques and platforms are together referred to as *next-generation sequencing* or, shortly, *NGS* (Heather and Chain 2016).

In NGS technologies, more powerful sensors allowed working with much lower sample input, but great improvement was obtained with highly massive parallelization. The sequencing process is now based on nucleotide incorporation detection rather than on chain termination. At each new incorporation cycle, the added nucleotides are detected directly, by fluorescence labelling (Illumina platforms), detection of pyrophosphates released on base incorporation (Pyrosequencing, Roche), or by proton release and pH variation sensors (Ion platforms, Thermo Fisher Scientific). Moreover, by the molecular addition of index sequences in the sample preparation process, many samples can be multiplexed to be analysed in a single run. However, most NGS platforms still output short reads when compared to Sanger sequencing, ranging from 35 (the first ones) to a little more than 1000 bases long, with the current technologies. This size limitation is bypassed with computational and bioinformatics tools, which have evolved side by side with the sequencing platforms. Yet, computational power sets the limits nowadays. As a wide variety of sequencers have become popular equipment in every field of life sciences and adjoining areas, the amount of generated raw sequencing data greatly surpasses the ability to process and interpret it. Currently, the most used NGS platforms are produced by Thermo Fisher, with Ion PGM, Ion Proton, and S5 sequencers, and by Illumina, with the

benchtop sequencers iSeq, MiniSeq, and MiSeq, and the production-scale sequencers HiSeq, NextSeq, and NovaSeq.

In the sequencing process, the same fragment can be sequenced in different manners. It is possible to sequence one end, which generates single end reads (SE), or both ends, which generates paired end reads (PE). In the latter case, if shorter than the sum of two reads size, the sequenced ends overlap in the middle of the fragment sequence, representing it all. The PE reads bear information of synteny, even when they do not overlap. The reads might belong to the ends of a fragment of up to 800 bases. Sequence assembler softwares use this valuable information to connect regions and extend the assembly. A derivation of PE sequencing is mate pairs (MP). In this technique, longer fragments from 2 kb up to 5 kb are circularized by biotinylation, fragmented, and sequenced to produce associate reads belonging to the ends of the original fragment. They are also used to infer contiguity, but at greater distances, and are fundamental to assemble DNA species like chromosomes, from short reads data.

The more recent achievements of the sequencing revolution are altogether called *third-generation sequencing* (*TGS*) techniques (van Dijk et al. 2018). These techniques have in common the ability to sequence very long fragments of single molecules, reaching read lengths longer than 2,000,000 base pairs (Payne et al. 2018), with no need of amplification (Heather and Chain 2016). Some, like SMRT (single-molecule real-time) sequencing developed by PacBio (Pacific Biosciences) are based on incorporation detection with high sensitivity for a fluorescent nucleotide in a very small volume – zeptoliters ($10e^{-21}$) – in the bottom of a well where a DNA polymerase is attached. Others, like the MinIon, GridION, and PromethION (Oxford Nanopore), work on direct detection of sequence composition by reading electric variations in a membrane pore when each base of the sequence fragment passes through, also allowing detection of base chemical modifications, such as methylation. At first these platforms had high error rates up to 13%, but with the advances in sequence chemistry and the ability to re-sequence the same molecule, this drawback is being put aside. In particular, the Oxford Nanopore platforms are already able to sequence RNA directly (Garalde et al. 2018).

NGS strategies can be used alone or in combination, complementing each other in a hybrid approach. Inaccurate long reads can aid the scaffolding of accurate contigs generated from short reads (Sohn and Nam 2016) or can aid in determining the length of repetitive regions. Repetitive regions were regarded for a long time as useless products of expansion of repetitive elements, transposons, and viral remains and remained ignored on genetic analyses. Although some of these elements have no function at all, others are essential. They have been domesticated to play important roles in cellular processes such as formation of the kinetochore and other nucleoprotein complexes (Shapiro and von Sternberg 2005), and being part of the bacterial and archaeal adaptative immune system (Kunin et al. 2007). Concerning Antarctic biotechnological prospection, repetitive sequences should not be ignored. Many AFPs have active sites composed of repetitive blocks of amino acids, which can be encoded by repetitive nucleotides (Davies 2014; Baalsrud et al. 2017).

15.4 Assembling the Genome

Any reads generated on sequencing, big or small, are usually fragments of a given genome (or transcriptome). To reconstitute the original full genome sequence, it is necessary to use computational strategies in a process analogous to assemble a jigsaw puzzle but with millions of pieces. The information needed to accomplish this task is contained in the reads to be assembled, as it is in the pieces of the puzzle. In the latter, adjacent pieces are connected based on their complementary shapes and patterns. For a genome or transcriptome, the assembly is done by finding overlapping parts in read pairs. When a pair of reads shares a common sequence, they are connected, generating a continuous sequence called contig (Fig. 15.1). This process is repeated progressively until no more shared sequences are found in the reads universe.

There are two main assembling strategies, de novo assembly and referenced assembly. The de novo assembly (also called ab initio assembly) uses only the information contained in the reads to reconstitute the original sequence. It is the preferred strategy and usually the first to be attempted. The outcome depends on the read quality, coverage, and the complexity of the sequence to be assembled.

The sequence quality refers to the probability of each position of a read to actually represent the original nucleotide from which the sequence is derived. It is measured at the sequencing process and relates to the intensity of the signal for each base sequenced. The quality can vary drastically along the read, and checking it is the first step before the start of any assembly. Full reads or read pieces of low quality must be removed from the set to assure that only quality reads are used in the assembly, in a process known as trimming.

The coverage refers to the percent of the sequenced fragment that is covered by the generated reads. The coverage is complemented by another concept, sequencing depth, which refers to the number of times each position of the sequence is represented in the reads set. Due to biases inherent to the sample nucleotide composition and size, and to the library construction and sequencing process, some parts of the genome will be better represented than others. Despite being fully covered in the sequencing output, a sequenced fragment might have inner regions that vary in depth. Prior to assembly, the average expected depth can be estimated by multiplying the number of output reads by the read size and dividing by the approximate size of the genome or the sequence to be assembled. Both concepts are also used at the end of the assembly process as quality metrics. Higher depth values for a given position mean that there is more evidence in the reads set to guarantee that this position truly represents the actual base in the original sequence. In this manner, coverage and depth are measures of assembly reliability.

The sequence complexity corresponds to the content of the original sequence. Sequences of high complexity are easier to assemble as there is low probability that a read aligns to more than one position. Low-complexity sequences such as repetitive regions are harder to assemble, especially if they are longer than the read size. They make it difficult for the assembly software to precisely define the connection

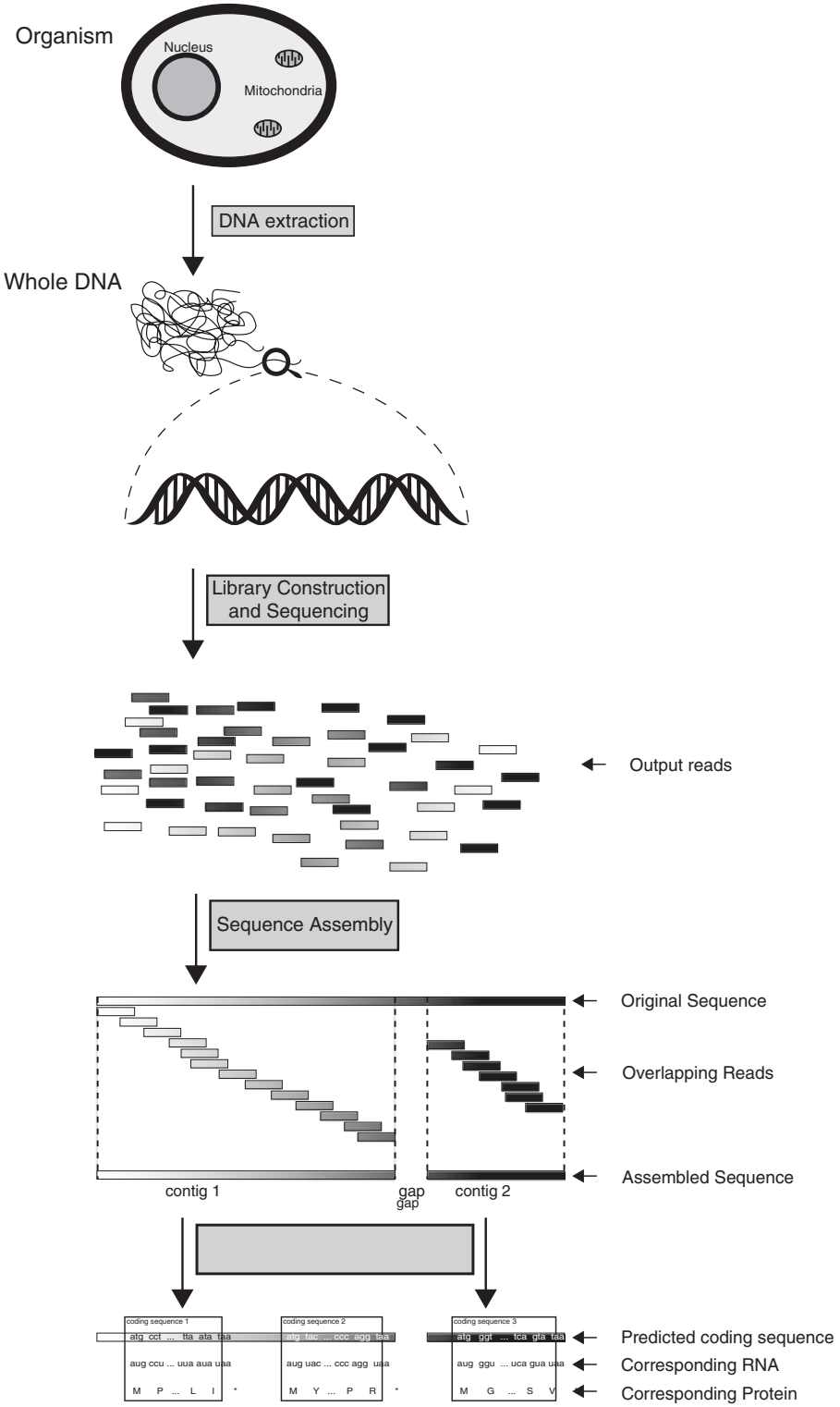


Fig. 15.1 Summarized workflow of the genome sequencing, assembly, and prediction process

between the reads in the read set, as there are greater possibilities of read pairs to concatenate.

With appropriate coverage and high-quality reads, it possible to assemble an entire genome or, more commonly, its regions of high complexity. As low complexity regions do not contain genes at most times, it is not unusual to proceed to the next analytic steps, namely gene prediction, annotation, and genetic comparison, before fully assembling the genome. Moreover, some genomic analyses require full genome reconstruction, which makes it necessary to connect the contigs to structures that are closer to the genomic structures of chromosomes. This process is called scaffolding. It can be approached by sequencing the ends of long fragments that will be later used to connect and position the contigs relatively to each other, the formerly referred *mate pairs*. As these fragments are much longer than the reads and as it is not possible most of the times to precisely determine their length, the resulting scaffolds will be contigs connected by stretches of a sequence of arbitrary size and unknown composition. These connecting regions are represented by the letter N, meaning any nucleotide. Despite being a very elegant strategy, mate pair reads may become obsolete with the popularization of long read platforms.

An alternative way to perform scaffolding is to use another assembled genome of the same species, when available, or a genome of a close species as a reference. This process is called *referenced assembly*. This would be the least desired alternative as even close species have their own genomic particularities that may introduce biases. Further, for using genomes of the same species, it is necessary to be sure of the reference genome assembly quality and reliability, as errors in the reference genome will be propagated to the new assembly.

With improvements in sequencing technologies yielding longer and more accurate reads, the *referenced assembly* strategy may become obsolete. Despite applying the referenced, de novo, or mixed strategies, many softwares are available to perform the assembling task, each with its own particularities, most appropriated for different types of organisms and their genomic structures. Each assembler also presents many possibilities of configuration, which drastically affects the outcome. It is usual to try different softwares and configurations and compare the output assemblies based on their quality metrics to determine the best suited assembly to each case. Assemblies with a longer span, structured in fewer contigs are preferred. Two metrics to be considered are the *L50* and *N50*. The first refers to the number of the largest contigs that are, together, equivalent to half of the assembly size. The lower this value, the less fragmented is the assembly. From the L50 contigs, the N50 corresponds to the size (in base pairs) of the smallest contig, thus allowing comparison of two assemblies with the same L50 value.

In addition to numerical metrics, the quality of an assembly can be evaluated in terms of its completeness. It refers to the percentage of expected genes for that species, which are present in the assembly. As genomes vary in content even between close species, this measure is an estimate and depends on the previous knowledge on the close species genomes. The key genes used to calculate the completeness are single-copies orthologous genes, which are a special type of homologous genes.

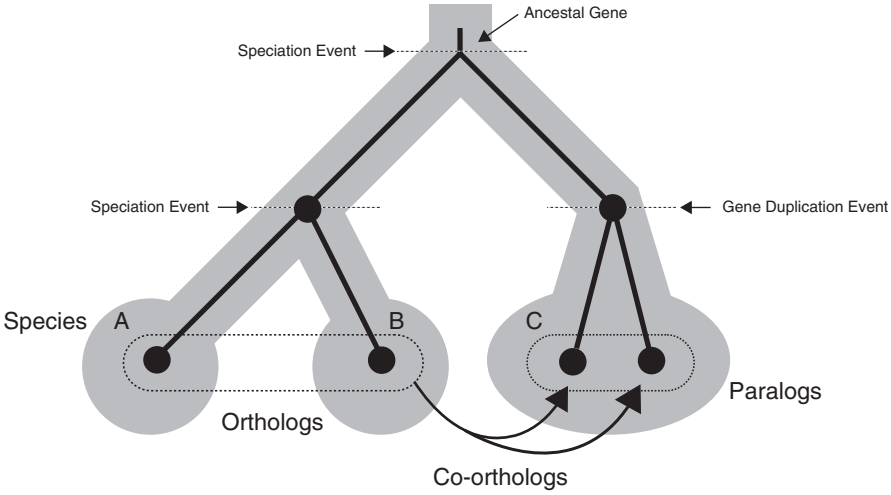


Fig. 15.2 Summary of homology relationships: genes related in terms of ancestry (homologous) that retained the same function in distinct species are called orthologous genes (orthologs). Genes duplicated along the evolutionary history of a species, whether they have retained the same function or not, are coined as paralogous genes (paralogs)

Homology is the existence of shared ancestry between a pair of structures or genes (Fig. 15.2). A gene that is common to two (or more taxa) is said to be homologous. Eventually, through the course of generations, a gene might be duplicated. The organism's genome will now carry two or more copies of this sequence. As the original molecular function of the gene product is maintained by one of the copies, the others are free to mutate. These new copies might assume other functions that, if beneficial to the organism, will therefore be maintained. This is one of the ways in which new genes arise (Wu and Knudsona 2018). Homology relations are subdivided into two main types. If two genes (or DNA sequences) share ancestry because of a speciation event, they are said to be *orthologous*. If they share ancestry because of a duplication event in a species, they are said to be *paralogous*.

By analysing high-quality genome assemblies of related species, it is possible to identify orthologous genes that are always present on the genomes of clade members. These are called the *core orthologous genes* (*COGs*) for that clade. Some of these genes are virtually vital for the organisms, and thus always tend to be present in any related genome. The closer the species are, the greater will be the number of COGs. So, a new genome assembly of another member of the clade is expected to contain almost all of these genes. The absence of expected COGs might indicate that this assembly is not truly representing the actual genome, as some parts – *contigs* containing these genes – might be missing. With a good assembly in hand, one can proceed to a deeper investigation for understanding the organisms' biology from a genetic and molecular perspective.

15.5 Investigation: From Bases to Genes, from Genes to Function

The choice to bioprospect an organism usually comes from an interest to understand a remarkable capability, but any organism has more to tell than the phenotypic traits that catch one's attention at first. Identifying genomic traits related to the particular phenotype of interest is not a simple task, and resembles the search for a needle in a haystack. However, it can be immensely facilitated depending on the availability of close species genomes, transcriptomes, and genetic data or previous knowledge on similar gene functionalities. In the *comparative genomics* field, the genomic information of different species is compared to highlight the differences or similarities that can be accounted for particular traits. For example, AFPs are a class of proteins that has evolved independently many times in different taxa. Some AFP classes are well understood and characterized, but some are yet unknown, with evidence of their existence coming only from the observation that the organism can survive freezing (Bar Dolev et al. 2016). If one aims to identify AFPs in novel organisms, comparison with sequences of other AFPs from closely related organisms might indicate the most probable candidates. Conversely, if one is working with lineages from which AFPs have not been identified yet, a subtractive approach can be applied. By comparing the genome of an AFP-producing organism to a close one that does not have this ability, the exclusive regions might contain the AFP-coding genes that are sought for. This approach greatly reduces the regions of the genome to be mined.

Once assembled, the genome is purely a set of strings carrying letters with no evident significance. To describe it, it is necessary to identify where all the genetic features are and to assign their functions. This can be done by two main strategies. The first one is by searching for similarities with sequences from previously studied organisms. Genetic information is available in many specialized curated sequence databases, where sequences are stored together with their attributes, obtained by experimental, or theoretical approaches. The second is to work with information contained in the genome itself. This strategy is called ab initio gene prediction and involves identifying the regions of interest in the genome with the aid of computational tools trained to identify specific features. Locating these regions and assigning their composing elements, such as protein-coding genes, regulatory sequences, intron-exon boundaries, repetitive elements, as well as inserting additional information and comments, is a process called *genome annotation*.

Despite differences, all known life forms work on an almost identical coding system (Koonin and Novozhilov 2017). Their protein-coding genes are all virtually structured as *open reading frames* (*ORFs*): a start codon in frame with amino acid-coding codons, ending in a stop codon. Gene prediction softwares and algorithms use this principle to identify every piece of the genome that matches these characteristics. This does not imply that the region is actually a gene, but it is an indication. Since the probability of a stop codon occurrence is approximately 1 to every 20 codons, ORFs with 60 nt or smaller tend to be disregarded. Classical ORF predictors

usually consider only ORFs longer than 150 nt (50 amino acids) by default (Rombel et al. 2002), but there is growing evidence that small ORFs also code active genes (Couso and Patraquim 2017). The longer the ORF, the lower is its probability to originate and be maintained in the genome if not under selective pressure. Further, similar to eukaryotes, fungi have some of their coding sequences (CDSs) interrupted by introns, that must be spliced out after transcription to generate a mature RNA. In the annotation process, the intronic sequences must be considered, and this is achieved by using an intron-aware annotation software.

Further, identification of sequences with self-complementarity and palindromic sequences indicate possible ncRNAs. There are many gene predictors and pipelines available to accomplish the annotation task, most in constant improvement. All follow the same logic but are better suited to specific situations, such as organisms that have an exon/intron structure, gene editing processes, or those that are derived from metagenomic samples, among other particularities.

Proof of a gene's existence, protein coding or not, is the presence of its corresponding RNA in the cell. However, as many genes are expressed only in particular situations, it is not always easy to provide the right stimuli, especially when the cultivating conditions are so extreme and hard to replicate, as for most Antarctic microorganisms.

Once we know the possible genes, the next step is to assign their function. The best way to do so is to compare each gene's sequence to the characterized sequences deposited in databases. The UniProt/Swiss-Prot is the most suited database for this purpose, as it is composed of curated information, which is sometimes experimentally validated. Many other databases are available, dedicated to specific taxa and organisms, with or without experimental evidences.

Pairwise comparison between sequences is performed with alignment tools like *BLAST* (Basic Local Alignment Search Tool) (Altschul et al. 1990) or *DIAMOND* (Buchfink et al. 2015). These tools look for sequence similarity that ultimately suggests a taxonomic relationship and make it possible to extend information from known sequences to the query sequence. The search must be performed by choosing extremely low e-value cutoffs and selecting only highly similar alignments. These characteristics might indicate same ancestor origin and homology relations. It is also common to look for characterized protein domains/signatures using protein domain databases as a reference. *InterProScan* is the most commonly used software for this, and Pfam is the most used reference database.

These strategies are efficient at finding the core metabolism and structural genes, as most of them are extremely conserved throughout life domains. As the Antarctic biodiversity is yet poorly characterized and has thousands of years of divergence, many of their organisms' new genes do not share nucleotide or amino acid sequence similarities with any known sequences. This is the case for most antifreeze proteins, as most of them are unique to each organism; having evolved independently many times (Davies 2014).

15.6 Omics Studies in Antarctica

The omics strategies have been largely used in the study of bacteria and archaea, as their genomes are usually much smaller than those of eukaryotes, making their investigation considerably cheaper and more feasible. Some studies have characterized archaea (Anderson et al. 2016) and bacteria (Dsouza et al. 2015; Han et al. 2016) endemic to the Antarctic region. The advantages of using omics to investigate Antarctic fungi can be exemplified in a recent study on a psychrophilic basidiomycete yeast isolated in the Antarctic region, *Glaciozyma antarctica*. Firdaus-Raih et al. (2018) have characterized this yeast. It has a 20 Mb genome, composed of 7857 predicted CDSs, of which 67% were confirmed by EST (Expressed Sequence Tag) sequencing. Nine genes were found to have similarity with known AFP-coding genes. These genes had their expression patterns characterized after sub-zero temperature exposure and were also heterologously expressed in bacteria to characterize their antifreeze properties. Other genes related to organism freezing tolerance were investigated. FADs are enzymes responsible for the introduction of double bonds in fatty acids that compose the lipid membrane. The double bond content is directly related to membrane fluidity, and the expression of these FADs was found to increase after sub-zero temperature exposure. The purified proteins were also submitted to thermal hysteresis assays to determine their influence on the freezing/melting point of water, and were found capable of lowering the water freezing point by 0.04–0.08 °C.

Another example of Antarctic fungi genomic investigation is the study on the black cryptoendolithic *Cryomyces antarcticus* (Sterflinger et al. 2014). The genome of 24.3 Mb was assembled in 12,492 contigs larger than 300 bp, with an N50 of 4.72 Mb. Gene prediction identified 10,731 putative protein-coding genes and the genome was compared to those of five other close species of melanized fungi from non-extreme habitats that were available, but no gene related to freezing tolerance was found.

Genomic analyses are being carried out by our group on the genomes of Antarctic fungi. *Metschnikowia australis* is a marine yeast that lives in close association with diverse Antarctic marine algae. It was selected for genomic analysis due to its capability to withstand freezing down to −80 °C, without the addition of any cryoprotectant (i.e. glycerol). The yeast's genome was sequenced, assembled, and annotated (Batista et al. 2017), totalling 4442 protein-coding genes along its 14.3 Mb genome. Since no similarity was found with any known antifreeze protein gene, a different approach to investigate this phenotypic trait is necessary. One possible strategy is to identify genes directly associated with this feature using comparative genomics. The genome of the organism in study is compared to other genomes from close species that lack the freeze tolerance capability. Exclusive genes present only in the interrogated species can be identified, thus reducing the candidates to be screened for the antifreeze property.

15.7 Ontology and Metabolic Analyses

In some cases, the same molecular activity can be performed by different and unrelated genes. These situations require a classification system not based on homology. Further, the grouping of an organism's genes into functional classes allows rapid, albeit superficial, assessment of the processes that are enriched in a genome or transcriptome. Similar and related functions are organized in hierarchies that ultimately encompass all cellular processes, with specific associations to a unified, computable terminology, in what is called *ontology*. One widely used ontology classifier is *gene ontology* (Carbon et al. 2017), which describes genes by three major classifications: cellular component, biological process, and molecular function. Another one is the KEGG BRITE, a highly organized classification of hierarchical functions and networks of molecular interaction for genes constituting the KEGG Orthology database (Kanehisa and Goto 2000). These initiatives are fundamental to unify biological knowledge under a common vocabulary and provide important resources to investigate any genome functionality.

Secondary metabolites (*SMs*) are molecules produced by an organism that are not necessary for normal development or growth (Fox and Howlett 2008). Fungi are molecular factories capable of producing many SM compounds with various functions. These compounds act as antibiotics, virulence factors or pigments, and are related to processes like signalling, among other functions (Macheleidt et al. 2016). These molecules are incredibly diverse and have immense biotechnological applications, and fungal extracts are commonly prospected for activities against pathogens and cancer (Godinho et al. 2013; Spiteller 2015).

Genomic investigation of fungi shows that there is still a lot to comprehend. It is now easier to identify many gene clusters related to SM production by sequence similarity, than it is to identify their final products. For example, sixty-eight gene clusters are known in the model ascomycete *Aspergillus nidulans*, but the products for only 20 of those are known (Macheleidt et al. 2016). There are databases dedicated to storing information on secondary metabolites and their related genes. These databases also provide computational tools to assess genomes in pursuit of SM producing genes. Once a specific activity of interest is detected by extract screening, one important starting point for the identification of the responsible pathway is to seek genes possibly involved in secondary metabolite production. By revealing these genes, one can devise strategies to boost pathway activity or even proceed to cloning its component genes for heterologous expression (Macheleidt et al. 2016). One example is the work performed by Hossain et al. (2016), in which the three genes of the itaconic acid (IA) production pathway from *Aspergillus terreus* were cloned into *A. niger* and overexpressed to increase the IA production fivefold.

It is documented that many Antarctic yeasts are capable of producing compounds like pigments, microsporines, and other UV-protective compounds, as screened by biochemical analysis (Vaz et al. 2011). Yet, most of these molecules and their biosynthetic pathways remain to be characterized. This reinforces the potential of

Antarctic microorganisms in the production of new biomolecules. As the genetic knowledge on Antarctic organisms increases, bioinformatic tools will aid in comprehending and exploiting this potential.

15.8 Phylogenomic Analysis

An organism's capabilities are as important as its relations to other known species. Genetic marker-based phylogeny is widely employed, especially in microbiology. Fungi, as aforementioned, are commonly classified based on the sequence of ITS regions in the ribosomal RNA gene (Schoch et al. 2012). Despite being powerful, easy to scale up, and widely used, ITS phylogeny is based on one or a few markers that might not be robust enough in some cases. The possibility to use hundreds or thousands of genes in this analysis has brought phylogeny to the omics era, as *phylogenomics*.

With many accessible genomes, there is no need to limit their comparison to short stretches of sequences. These comparisons can be performed using all correspondent sequences among the analysed organisms. As genetic similarity weakens in distantly related individuals, more informative shared regions exist in closer individuals to support their phylogenetical relationship. In phylogenomics there are no preferred genes. All that is common might be compared. Yet, some organisms have gene duplications and gene deletion events that would skew the comparison. A good strategy to overcome this problem is to use only common single-copy orthologous genes.

For phylogenomic analysis, despite having a complete or partial genome, it is important to compare only sequences that have equivalents in all the genomes to be analysed, the aforementioned orthologous genes. It is also important that these sequences have only one copy in each genome – paralogs must be avoided. This guarantees that the information used is unique (Simão et al. 2015). Genes to be compared must be aligned and have the same size. In the alignment process, inner regions that are not shared on the sequences are extended as gaps. After that, the unshared outer regions are trimmed off, leaving all sequences with the same number of positions.

You can either use amino acid or nucleotide sequences for phylogenomic analysis. As many of the investigated sequences correspond to proteins that are under strong selective pressure to maintain their functions and, as amino acids can be coded by more than one codon, close species may bear features almost identical at the protein level, but with DNA sequences being less conserved. In this sense, phylogenomic trees made from nucleotides tend to be better suited to resolve close species relations. Conversely, when studying distantly related species, phylogenomic trees generated from proteins should be used. There is no ultimate rule and both strategies must be tested and compared to choose the most robust and adequate solution. The reliability of output trees can be evaluated by analysing the terminal branches bootstrap – a measure of how frequent a branch configuration is maintained

after a random subsampling of the data. The higher the bootstrap values, the more reliable is the tree.

15.9 Conclusions and Perspectives

The advances in nucleic acid sequencing and interpretation are revolutionizing many fields of biological sciences. These powerful omics tools are available to every microbiologist and give uncountable perspectives on the investigation of microbial diversity, genetic capabilities, and environment deciphering. To use these tools, it is crucial for a microbiologist to be familiar with the fast-evolving approaches, strategies, logic, and vocabulary of the omics field. Nonetheless, their use has unprecedented potential to help understand the immense microbiological diversity that remains to be investigated, especially in the Antarctic continent.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Anderson IJ, DasSarma P, Lucas S, Copeland A, Lapidus A, Del Rio TG, Kyrpides NC (2016) Complete genome sequence of the Antarctic *Halorubrum lacusprofundi* type strain ACAM 34. Stand Genomic Sci 11:70

Baalsrud HT, Tørresen OK, Solbakken MH, Salzburger W, Hanel R, Jakobsen KS, Jentoft S (2017) De novo gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data. Mol Biol Evol 35:593–606

Baeza M, Barahona S, Alcaíno J, Cifuentes V (2017) Amplicon-metagenomic analysis of fungi from antarctic terrestrial habitats. Front Microbiol 8:2235

Bar Dolev M, Braslavsky I, Davies PL (2016) Ice-binding proteins and their function. Annu Rev Biochem 85:515–542

Bashiardes S, Zilberman-Schapira G, Elinav E (2016) Use of metatranscriptomics in microbiome research. Bioinform Biol Insights 10:19–25

Batista TM, Hilário HO, Moreira RG, Furtado C, Godinho VM, Rosa LH, Franco GR, Rosa CA (2017) Draft genome sequence of *Metschnikowia australis* strain UFMG-CM-Y6158, an extremophile marine yeast endemic to Antarctica. Genome Announc 5(20) e00328-17. doi: 10.1128/genomeA.00328-17

Berlemont R, Jacquin O, Delsaute M, La Salla M, Georis J, Verté F, Galleni M, Power P (2013) Novel cold-adapted esterase mhlip from an Antarctic soil metagenome. Biol 2:177–188

Bernard G, Pathmanathan JS, Lannes R, Lopez P, Bapteste E (2018) Microbial darkmatter investigations: how microbial studies transform biological knowledge and empirically sketch a logic of scientific discovery. Genome Biol Evol 10:707–715

Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60

Calderoli PA, Espínola FJ, Dionisi HM, Gil MN, Jansson JK, Lozada M (2018) Predominance and high diversity of genes associated to denitrification in metagenomes of subantarctic coastal sediments exposed to urban pollution. PLoS One 13:e0207606

Carbon S, Dietze H, Lewis SE, Mungall CJ, Munoz-Torres MC, Basu S, Westerfield M (2017) Expansion of the gene ontology knowledgebase and resources: the gene ontology consortium. Nucleic Acids Res 45:D331–D338

Chan Y, Van Nostrand JD, Zhou J, Pointing SB, Farrell RL (2013) Functional ecology of an Antarctic Dry Valley. Proc Natl Acad Sci U S A 110:8990–8995

Choi J, Kim SH (2017) A genome tree of life for the fungi kingdom. Proc Natl Acad Sci U S A 114:9391–9396

Couso JP, Patraquim P (2017) Classification and function of small open reading frames. Nat Rev Mol Cell Biol 18:575–589

Davies PL (2014) Ice-binding proteins: a remarkable diversity of structures for stopping and starting ice growth. Trends Biochem Sci 39:548–555

Donovan PD, Gonzalez G, Higgins DG, Butler G, Ito K (2018) Identification of fungi in shotgun metagenomics datasets. PLoS One 13:e0192898

Dsouza M, Taylor MW, Turner SJ, Aislabie J (2015) Genomic and phenotypic insights into the ecology of *Arthrobacter* from Antarctic soils. BMC Genomics 16:36

Duarte AWF, Passarini MRZ, Delforno TP, Pellizzari FM, Cipro CVZ, Montone RC, Petry MV, Putzke J, Rosa LH, Sette LD (2016) Yeasts from macroalgae and lichens that inhabit the South Shetland Islands, Antarctica. Environ Microbiol Rep 8:874–885

Dujon BA, Louis EJ (2017) Genome diversity and evolution in the budding yeasts (Saccharomycotina). Genetics 206:717–750

Ferrés I, Amarelle V, Noya F, Fabiano E (2015) Construction and screening of a functional metagenomic library to identify novel enzymes produced by Antarctic bacteria. Adv Polar Sci 26:96–101

Firdaus-Raih M, Hashim NHF, Bharudin I, Abu Bakar MF, Huang KK, Alias H, Lee BKB, Mat Isa MN, Mat-Sharani S, Sulaiman S, Tay LJ, Zolkefli R, Muhammad Noor Y, Law DSN, Abdul Rahman SH, Md-Illias R, Abu Bakar FD, Najimudin N, Abdul Murad AM, Mahadi NM (2018) The *Glaciozyma antarctica* genome reveals an array of systems that provide sustained responses towards temperature variations in a persistently cold habitat. PLoS One 13:1–18

Fox EM, Howlett BJ (2008) Secondary metabolism: regulation and role in fungal biology. Curr Opin Microbiol 11:481–487

Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, Jordan M, Ciccone J, Serra S, Keenan J, Martin S, McNeill L, Wallace EJ, Jayasinghe L, Wright C, Blasco J, Young S, Brocklebank D, Juul S, Clarke J, Heron AJ, Turner DJ (2018) Highly parallel direct RNA sequencing on an array of nanopores. Nat Methods 15:201–206

Godinho VM, Furbino LE, Santiago IF, Pellizzari FM, Yokoya NS, Pupo D, Alves TM, Junior PA, Romanha AJ, Zani CL, Cantrell CL, Rosa CA, Rosa LH (2013) Diversity and bioprospecting of fungal communities associated with endemic and cold-adapted macroalgae in Antarctica. ISME J 7:1434–1451

Goldman AD, Landweber LF (2016) What Is a Genome? PLoS Genet 12:e1006181

Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17:333–351

Graur D, Zheng Y, Azevedo RBR (2015) An evolutionary classification of genomic function. Genome Biol Evol 7:642–645

Han SR, Kim KH, Ahn DH, Park H, Oh TJ (2016) Complete genome sequence of carotenoid-producing *Microbacterium* sp. strain PAMC28756 isolated from an Antarctic lichen. J Biotechnol 226:18–19

Hashim NHF, Mahadi NM, Illias RM, Feroz SR, Abu Bakar FD, Murad AMA (2018) Biochemical and structural characterization of a novel cold-active esterase-like protein from the psychrophilic yeast *Glaciozyma antarctica*. Extremophiles 22:607–616

Heather JM, Chain B (2016) The sequence of sequencers: the history of sequencing DNA. Genomics 107:1–8

Hittinger CT, Rokas A, Bai FY, Boekhout T, Gonçalves P, Jeffries TW, Kominek J, Lachance MA, Libkind D, Rosa CA, Sampaio JP, Kurtzman CP (2015) Genomics and the making of yeast biodiversity. Curr Opin Genet Dev 35:100–109

Hossain AH, Li A, Brickwedde A, Wilms L, Caspers M, Overkamp K, Punt PJ (2016) Rewiring a secondary metabolite pathway towards itaconic acid production in *Aspergillus niger*. Microb Cell Factories 15:130

Hu XP, Heath C, Taylor MP, Tuffin M, Cowan D (2012) A novel, extremely alkaliphilic and cold-active esterase from Antarctic desert soil. Extremophiles 16:79–86

Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30

Koo H, Hakim JA, Fisher PRE, Grueneberg A, Andersen DT, Bej AK (2016) Distribution of cold adaptation proteins in microbial mats in Lake Joyce, Antarctica: analysis of metagenomic data by using two bioinformatics tools. J Microbiol Methods 120:23–28

Koonin EV, Novozhilov AS (2017) Origin and evolution of the universal genetic code. Annu Rev Genet 51:45–62

Kunin V, Sorek R, Hugenholtz P (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome Biol 8:R61

Lloyd KG, Steen AD, Ladau J, Yin J, Crosby L (2018) Phylogenetically novel uncultured microbial cells dominate earth microbiomes. mSystems 3:e00055–e00018

Macheleidt J, Mattern DJ, Fischer J, Netzker T, Weber J, Schroeckh V, Valiante V, Brakhage AA (2016) Regulation and role of fungal secondary metabolites. Annu Rev Genet 50:371–392

Payne A, Holmes N, Rakyan V, Loose M (2018) BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. Bioinformatics https://doi.org/10.1093/bioinformatics/bty841

Pearson GA, Lago-Leston A, Cánovas F, Cox CJ, Verret F, Lasternas S, Duarte CM, Agusti S, Serrão EA (2015) Metatranscriptomes reveal functional variation in diatom communities from the Antarctic Peninsula. ISME J 9:2275–2289

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13:341

Rombel IT, Sykes KF, Rayner S, Johnston SA (2002) ORF-FINDER: a vector for high-throughput gene identification. Gene 282:33–41

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74:5463–5467

Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, Fungal Barcoding Consortium Author List (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc Natl Acad Sci U S A 109:6241–6246

Shapiro JA, von Sternberg R (2005) Why repetitive DNA is essential to genome function. Biol Rev Camb Philos Soc 80:227–250

Sharpton TJ (2014) An introduction to the analysis of shotgun metagenomic data. Front Plant Sci 5:1–14

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE (1986) Fluorescence detection in automated DNA sequence analysis. Nature 324:674–679

Sohn J, Nam JW (2016) The present and future of de novo whole-genome assembly. Brief Bioinform 19:bbw096

Spiteller P (2015) Chemical ecology of fungi. Nat Prod Rep 32:971–993

Sterflinger K, Lopandic K, Pandey RV, Blasi B, Kriegner A (2014) Nothing special in the specialist? Draft genome sequence of *Cryomyces antarctic*us, the most extremophilic fungus from Antarctica. PLoS One 9:e109908

van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C (2018) The third revolution in sequencing technology. Trends Genet 34:666–681

Vaz ABM, Rosa LH, Vieira MLA, de Garcia V, Brandão LR, Teixeira LCRS, Moliné M, Libkind D, van Broock M, Rosa CA (2011) The diversity, extracellular enzymatic activities and photoprotective compounds of yeasts isolated in Antarctica. Braz J Microbiol 42:937–947

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Miklos GLG, Nelson C, Broder S, Clark AG, Nadeau J, Mckusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Rosane C, Kabir C, Zuoming D, Francesco VD, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji R, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, ZYuan W, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu SC, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, Mccawley S, Mcintosh T, Mcmullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers Y, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-stine J, Caulk P, Chiang Y, Coyne M, Dahlke C, Mays AD, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, Mcdaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, David W, Wu M, Xia A, Zandieh A, Zhu X (2001) The sequence of the human genome. Science 291:1304–1351

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63

Wentzel LCP, Inforsato FJ, Montoya QV, Rossin BG, Nascimento NR, Rodrigues A, Sette LD (2018) Fungi from Admiralty Bay (King George Island, Antarctica) Soils and Marine Sediments. Microb Ecol 77:12–24

Wu B, Knudsona A (2018) Tracing the de novo origin of protein-coding genes in yeast. MBio 9:1–11

Yajima Y, Tojo M, Chen B, Hoshino T (2017) *Typhula* cf. *subvariabilis*, new snow mould in Antarctica. Mycology 8:147–152

Ziemert N, Alanjary M, Weber T (2016) The evolution of genome mining in microbes-a review. Nat Prod Rep 33:988–1005

ANEXO III

Todo o *pipeline*, as configurações utilizadas nos programas e os *scripts* gerados e utilizados neste trabalho se encontram disponibilizados no GitHub, e podem ser acessados através do *link* https://github.com/heronoh/prj_metsh/.