UNIVERSIDADE DE CAXIAS DO SUL

ÁREA DO CONHECIMENTO DE CIÊNCIAS DA VIDA

INSTITUTO DE BIOTECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

O Projeto Genoma de *Penicillium echinulatum* 2HH e S1M29: A Genômica Viabilizando a Descoberta de Conhecimento

Alexandre Rafael Lenz

CAXIAS DO SUL

2020

Alexandre Rafael Lenz

O Projeto Genoma de *Penicillium echinulatum* 2HH e S1M29: A Genômica Viabilizando a Descoberta de Conhecimento

Tese apresentada ao Programa de Pós-graduação em Biotecnologia da Universidade de Caxias do Sul, visando a obtenção de grau de Doutor em Biotecnologia.

Orientador: Prof. Dr. Aldo José Pinheiro Dillon
Coorientador: Profa. Dra. Scheila de Avila e Silva

CAXIAS DO SUL
2020

Dados Internacionais de Catalogação na Publicação (CIP)
Universidade de Caxias do Sul
Sistema de Bibliotecas UCS - Processamento Técnico

L575p Lenz, Alexandre Rafael
O projeto genoma de *Penicillium echinulatum* 2HH e S1M29 [recurso eletrônico] : a genômica viabilizando a descoberta de conhecimento /Alexandre Rafael Lenz. – 2020.
Dados eletrônicos.

Tese (Doutorado) - Universidade de Caxias do Sul, Programa de Pós-Graduação em Biotecnologia, 2020.
Orientação: Aldo José Pinheiro Dillon.
Coorientação: Scheila de Avila e Silva.
Modo de acesso: World Wide Web
Disponível em: https://repositorio.ucs.br

1. Penicillium. 2. Genoma. 3. Celulase. 4. Regulação de expressão gênica. 5. Álcool. I. Dillon, Aldo José Pinheiro, orient. II. Silva, Scheila de Avila e, coorient. III. Título.

CDU 2. ed.: 582.282.123.2

Catalogação na fonte elaborada pela(o) bibliotecária(o)
Carolina Machado Quadros - CRB 10/2236

ALEXANDRE RAFAEL LENZ

O PROJETO GENOMA DE *Penicillium echinulatum* 2HH E S1M29: A GENÔMICA VIABILIZANDO A DESCOBERTA DE CONHECIMENTO

Tese apresentada ao Programa de Pós-graduação em Biotecnologia da Universidade de Caxias do Sul, visando à obtenção do título de Doutor em Biotecnologia.

Orientador: Prof. Dr. Aldo José Pinheiro Dillon

Co-orientadora: Profa. Dra. Scheila de Avila e Silva

TESE APROVADA EM 10 DE DEZEMBRO DE 2020.

_____

Orientador: Prof. Dr. Aldo José Pinheiro Dillon

_____

Co-orientadora: Profa. Dra. Scheila de Avila e Silva

_____

Profa. Dra. Eliane Ferreira Noronha

_____

Profa. Dra. Camille Eichelberger Granada

_____

Prof. Dr. Sergio Echeverrigaray Laguna

"A floresta está viva. Só vai morrer se os brancos insistirem em destruí-la. Se conseguirem, os rios vão desaparecer debaixo da terra, o chão vai se desfazer, as árvores vão murchar e as pedras vão rachar no calor. A terra ressecada ficará vazia e silenciosa. Então morreremos, um atrás do outro, tanto os brancos quanto nós. Todos os xamãs vão acabar morrendo. Quando não houver mais nenhum deles vivo para sustentar o céu, ele vai desabar."

Davi Kopenawa (A Queda do Céu: Palavras de um Xamã Yanomami)

"In a time like this, when a page in history is being turned and civilization is in crisis, everything depends upon how humanity assimilates all of its known experience, knowledge, and wisdom into a body of awareness that leads humanity into the future. If the future experience is sustainable, then we survive; if it isn't, we go the way of the dinossaurs and become extinct. It is up to us. We have the power to change our future now. In fact, the here and now is the only place and time possible to make this change."

Drunvalo Melchizedek (The Mayan Ouroboros: The true positive Mayan Prophecy is revealed)

Formulada nos termos de uma metafísica distante da realidade modernizante, essa concepção de unidade e equilíbrio é característica recorrente dos povos ancestrais que habitam nosso planeta.

Esta pesquisa é dedicada às pessoas que compreendem a magnitude dessa concepção da realidade e que, de alguma forma, cooperam para realizar o sonho da coerência global equilibrada, pacífica e amorosa. Primeiramente, às civilizações ancestrais que resistem bravamente ao impetuoso liquidificador modernizante. Em especial a todos os cientistas que encontraram um propósito de vida alinhado a essa concepção e, principalmente, aos pesquisadores do Instituto de Biotecnologia da UCS.

AGRADECIMENTOS

O meu retorno ao Rio Grande do Sul e à UCS surgiu da possibilidade de viver alguns anos próximo da minha família. Se não fosse pela minha família, provavelmente eu não teria retornado à UCS. Sem palavras para agradecer essa possibilidade ímpar, me considero privilegiado de poder fazer meu doutorado próximo da família, após 10 anos vivendo em outros estados. Agradeço principalmente à minha Mãe Glaci, por sua generosidade, carinho e amor incondicional. Agradeço às minhas duas irmãs Tere e Cris pelo incentivo e carinho. À Tere por me receber sempre com o maior carinho e por conservar nossas raízes ancestrais na casa onde fui criado. E especialmente à Cris pelas aulas de inglês e revisão de artigos, fundamentais para este doutorado, sem esquecer das comidas maravilhosas nos dias frios caxienses e dos almoços em família nos domingos. Ao meu cunhado Marcos Casa pela inspiração acadêmica e pelo apoio e incentivo durante toda minha trajetória acadêmica. Aos meus sobrinhos Daniel, Rodrigo e Pedro. Principalmente ao Pedro que também trabalha no Laboratório de Bioinformática, por compartilhar o chimarrão com bergamota nas manhãs geladas, por jogar frescobol e tênis comigo no parque e por compartilhar conhecimentos biológicos.

Os agradecimentos principais são direcionados aos orientadores Dra. Scheila de Avila e Silva e Dr. Aldo José Pinheiro Dillon, pelos ensinamentos, paciência, disponibilidade e amizade. Scheila, nossos caminhos científicos assustadoramente opostos só se cruzaram graças à Bioinformática, você buscando a tecnologia para solução de problemas biológicos e eu buscando uma aplicação tecnológica alinhada aos meus valores. Foi uma aventura um tanto quanto assustadora, principalmente no início tumultuado do projeto, obrigado por facilitar o percurso. Obrigado por compartilhar seu conhecimento durante esses quatro anos e principalmente por estar sempre aberta para ouvir e assimilar as minhas mudanças de ideia sobre a pesquisa. A liberdade de pesquisa, aliada à sua flexibilidade para mudanças, sem sombra de dúvidas foram fatores determinantes para chegarmos juntos a esse resultado. Aldo, você é uma personalidade ilustre, de um carisma fora de série. Jamais esquecerei daquele abraço naquela segunda-feira triste após as eleições de 2018. Toda vez que você e o Sérgio me chamam de filho, me sinto mais próximo do meu pai que faleceu em 2012, obrigado pelo carinho.

Scheila e Aldo, foi uma honra ser orientado por vocês, obrigado por me apresentarem as Ciências da Vida e, principalmente, por me mostrarem uma outra forma de trabalho, diferente do que eu estava habituado nas ciências exatas. Um Congresso Brasileiro de Micologia na Amazônia não tem preço. Certa vez li um depoimento que me marcou muito (não lembro para citar - aqui pode), ele dizia que ter um título de doutorado é relativamente fácil, no entanto ser um Doutor é uma tarefa desafiadora, pois vai muito além da pesquisa acadêmica. Esse depoimento também dizia que um doutorado deve ser prazeroso, contrastando com o que geralmente observamos no meio acadêmico. Esse depoimento dizia ainda que ser um Doutor

exige o estudo e aperfeiçoamento de uma série de aspectos pessoais. Alguns relativamente óbvios como a organização e a forma de se comunicar com as pessoas. E outros um tanto quanto complexos como a humildade e a visão holística. O suporte de vocês como orientadores me fez evoluir muito nesses quatro anos, obrigado por tudo.

Agradecimentos especiais são direcionados aos pesquisadores e amigos do Laboratório de Bioinformática e Biologia Computacional [1] da Universidade de Caxias do Sul: Eduardo Balbinot, Nikael Souza de Oliveira, Fernanda Pessi de Abreu, Pedro Lenz Casa e Marcos Rossetto. Eduardo, Nika e Fer, sem o apoio e suporte de vocês essa tese não seria possível, obrigado de coração. Os agradecimentos também são direcionados à Dra. Letícia Osório da Rosa, à Dra. Marli Camassola e à Dra. Roselei Claudete Fontana do Laboratório de Enzimas e Biomassas da Universidade de Caxias do Sul e a todos aqueles que contribuíram para que esta pesquisa fosse possível, em especial também ao Prof. Dr. Sérgio Echeverrigaray Laguna, pelos inúmeros *insights* que mudaram os rumos desta pesquisa.

Agradecimentos especiais também são dirigidos aos pesquisadores Dr. Ernesto Perez-Rueda e Dr. Edgardo Galán-Vasquez do *Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas* (IIMAS) [2] da Universidad Nacional Autónoma de México (UNAM), Cidade de México e Mérida, ao pesquisador Dr. Jos Houbraken do *Westerdijk Fungal Biodiversity Institute* [3] de Utrecht, The Netherlands, e aos pesquisadores Dr. Nelson Menolli Jr. e Me. Mariana de Paula Drewinski do Instituto de Botânica do Jardim Botânico de São Paulo [4].

Morar sozinho no exterior é uma tarefa desafiadora. Morar sozinho no exterior e terminar o doutorado é uma tarefa extremamente desafiadora. Morar sozinho no exterior e terminar o doutorado durante uma pandemia global é uma tarefa inexplicavelmente assustadora e desafiadora. Inicialmente, as três semanas na Cidade do México trabalhando com Edgardo foram extremamente produtivas e de muito aprendizado. Em seguida, os seis meses que eu deveria morar em Mérida para trabalhar com Ernesto se transformaram em incertezas. Foi possível apenas uma ida para o Parque Tecnológico de Yucatán antes do lockdown. Naturalmente busquei refúgio em um local seguro e continuamos as investigações científicas por meio on-line. Para minha sorte, a pesquisa do intercâmbio necessitava apenas de conexão com internet. Ernesto e Edgardo me apresentaram uma forma altamente dinâmica de trabalho colaborativo. Meu trabalho nunca rendeu tanto e de forma tão fluida como nessa parceria. Ernesto é uma inspiração, um pesquisador reconhecido mundialmente, mas que chama a atenção por ser simples, calmo, amistoso e impecavelmente correto. Foi uma honra trabalhar com esses pesquisadores e ter essa experiência mexicana maravilhosa, tanto em nível acadêmico quanto em nível pessoal.

Agradecimentos especiais também são voltados para a Universidade do Estado da Bahia (UNEB) pelo financiamento desta pesquisa em forma de licença remunerada e a todos os colegas

[1] <https://www.ucs.br/site/nucleos-de-pesquisa/bioinformatica/>
[2] <https://www.iimas.unam.mx/>
[3] <http://www.westerdijkinstitute.nl>
[4] <https://www.infraestruturameioambiente.sp.gov.br/institutodebotanica/>

professores do Departamento de Sistemas de Informação [5], em especial ao colega Dr. Eduardo Manuel de Freitas Jorge pela amizade ímpar e pela disponibilidade para me substituir como professor das minhas disciplinas.

Agradecimentos também são dirigidos aos colegas e amigos do PPGBIO por tornar essa etapa de vida mais prazerosa: Keoma da Silva, Luisa Vivian Schwarz, Letícia Osório da Rosa e Fernando Joel Scariot.

Agradecimentos também são direcionados à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) [6] pelo financiamento desta pesquisa em forma de bolsa, modalidade Programa de Suporte à Pós-Graduação de Instituições Comunitárias de Educação Superior (PROSUC).

[5] <https://portal.uneb.br/salvador/cursos/sistemas-de-informacao/>
[6] <https://www.capes.gov.br/>

“If you want to find the secrets of the universe,
think in terms of energy, frequency and vibration.”
(Nikola Tesla)

LISTA DE QUADROS

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

AAs ..... Atividades Auxiliares
ANP ..... Agência Nacional do Petróleo
CAZymes Enzimas Ativas em Carboidratos
CBMs .... Módulos de Ligação ao Carboidrato
cDNA .... DNA complementar
CDSs ..... sequências codificadoras de proteínas
CEs ..... Carboidrato Esterases
ChIP-Seq . Imunoprecipitação de Cromatina
CREA .... Creatine Sucrose agar
CTBE .... Laboratório Nacional de Ciência e Tecnologia do Bioetanol
CTC ..... Centro de Tecnologia Canavieira
CYA ..... Czapek Yeast Autolysate agar
CYAS .... Blakeslee's MEA and CYA with 5% NaCl
CZ ..... Czapek's agar
DG18 .... Dichloran 18% Glycerol agar
EC ..... *Enzyme Commission*
ESTs ..... etiquetas de sequências expressas
etanol 2G . etanol de segunda geração
GEE ..... gases do efeito estufa
GHs ..... Glicosil Hidrolases
GRN ..... rede regulatória de genes
GTs ..... Glicosil Transferases
iBGLs .... β-glicosidases intracelulares
IEA ..... Agência Internacional de Energia
IRENA ... Agência Internacional de Energia Renovável
LSF ..... fermentação em estado líquido
MEA ..... Malt Extract agar
MEAbl ... Blakeslee's Malt extract agar
MME .... Ministério de Minas e Energia
mRNA ... RNA mensageiro
ncRNA ... RNA não codificante
OA ..... Oatmeal agar
ORFs ..... fases de leitura aberta

PLs ..... Polissacarídeo Liases
rDNA .... DNA ribossômico
siRNA .... RNA de interferência
SSF ..... fermentação em estado sólido
STs ..... transportadores de açúcares
TFBSs ... sítios de ligação de fatores de transcrição
TFs ..... fatores de transcrição
TGs ..... genes-alvo
TSSs ..... sítios de início da transcrição
YES ..... Yeast Extract Sucrose agar

RESUMO

LENZ, A. R. **O Projeto Genoma de *Penicillium echinulatum* 2HH e S1M29: A Genômica Viabilizando a Descoberta de Conhecimento**. 2020. 201 p. Tese (Doutorado) – Instituto de Biotecnologia, Universidade de Caxias do Sul, Caxias do Sul – RS, 2020.

Enzimas celulolíticas produzidas por fungos são utilizadas em diversos processos industriais, como a produção de tecidos, papel, alimentos e biocombustíveis. *Penicillium echinulatum* 2HH é um ascomiceto isolado do trato digestório de larvas de um coleóptero em 1979, também conhecido por seus coquetéis enzimáticos. Para melhorar os rendimentos de sacarificação da biomassa celulósica para exploração comercial, uma estratégia é o design de cepas hipersecretoras de enzimas. No entanto, o conhecimento molecular sobre o sistema celulolítico deste fungo é bastante restrito. No ano de 2013 foi realizado o sequenciamento dos genomas do mutante S1M29 e do parental selvagem 2HH de *P. echinulatum*. O objetivo desta tese compreendeu a montagem, anotação e depósito dos dois genomas *draft* em bancos de dados públicos, viabilizando a descoberta de conhecimento a partir de dados genômicos. A descoberta de conhecimento abrangeu uma série de domínios: i) identificação molecular para reposicionamento da espécie; ii) caracterização de mutações acumuladas no mutante S1M29; iii) análises evolucionárias a partir de marcadores gerais e específicos; iv) caracterização de genes de interesse: Enzimas Ativas em Carboidratos (CAZymes), transportadores de açúcares (STs) e fatores de transcrição (TFs); v) construção da rede regulatória de genes (GRN); e vi) identificação de genes-alvo para obtenção de linhagens comerciais. A identificação molecular da linhagem selvagem 2HH e seu reposicionamento na série *Oxalica* destacam-se pela primordialidade para estudos comparativos com outros microrganismos. O depósito dos genomas *draft* da linhagem selvagem 2HH e do mutante S1M29 no GenBank possibilita a ampliação do entendimento molecular desse fungo. A análise das mutações acumuladas no mutante S1M29 destacou um amplo conjunto de mutações, evidenciando a enzima BGL2 e o fator de transcrição FlbA que provavelmente contém as principais mutações envolvidas na hiperprodução de celulases. Também identificamos que o fenótipo albino do mutante S1M29 resultou de uma mutação na enzima ALB1, pertencente à via de biossíntese de DHN-melanina. Nossos resultados relacionados às análises evolucionárias hipotetizam uma simbiose mutualística potencial a longo prazo entre *P. echinulatum* 2HH e *Anobium punctatum*, cujas interações ambiente-específicas poderiam explicar a diferença na composição gênica em relação à *Penicillium oxalicum* 114-2. Ademais, a caracterização do CAZyoma de *P. echinulatum* 2HH demonstra que os genes que constituem o sistema celulolítico são predominantemente ortólogos a *P. oxalicum* 114-2, incluindo uma monooxigenase da família AA16, descrita pela primeira vez nesses dois fungos. Em seguida, a caracterização do transportoma de açúcares demonstrou a diversidade e especificidade de STs, incluindo oito famílias com especificidade para diferentes grupos de açúcares. Finalmente, a caracterização do TFoma e a inferência das redes regulatórias de genes de *P. echinulatum* 2HH e *P. oxalicum* 114-2

compreendem interações regulatórias que abrangem diversos processos biológicos, explorando diversos módulos regulatórios, como CpcA, FF-7, COL-26, AmyR, ClrB, CreA e XlnR. Esta tese contribui fortemente para construção de uma estrutura de suporte para compreensão molecular de *P. echinulatum* 2HH, revelando características relacionadas à produção de enzimas celulolíticas, captação de açúcares, produção de melanina, composição da parede celular e regulação da expressão gênica desse fungo. O conhecimento de vias regulatórias, aliado à caracterização de CAZymes e STs fornecem instrumentos para concepção de cepas comerciais, possibilitando a utilização de *P. echinulatum* 2HH para a produção de etanol 2G em larga escala. Finalmente, destacamos *P. echinulatum* 2HH como um importante aliado biotecnológico para a produção de biocombustíveis, auxiliando na transição energética global.

Palavras-chave: *Penicillium echinulatum*, genoma, celulases, regulação gênica, rede de regulação, etanol 2G.

ABSTRACT

LENZ, A. R. **The Genome Project of *Penicillium echinulatum* 2HH and S1M29: Genomics Enabling Knowledge Discovery**. 2020. 201 p. Tese (Doutorado) – Instituto de Biotecnologia, Universidade de Caxias do Sul, Caxias do Sul – RS, 2020.

Cellulolytic enzymes produced by fungi are used in several industrial processes, such as the production of fabrics, paper, food and biofuels. *Penicillium echinulatum* 2HH is an ascomycete isolated from the digestive tract of coleoptera larvae in 1979, also known for its enzymatic cocktails. To improve the saccharification yield of cellulosic biomass for commercial exploitation, one strategy is the design of hypersecretory strains of enzymes. However, the molecular knowledge about the lignocellulolytic system of this fungus is quite limited. In 2013, both genomes of S1M29 mutant and 2HH wild-type of *P. echinulatum* were sequenced. The purpose of this thesis included the assembly, annotation and deposit of both draft genomes in public databases, enabling knowledge discovery from genomic data. The knowledge discovery covered a series of domains: i) molecular identification to reposition the species; ii) characterization of accumulated mutations in S1M29 mutant; iii) evolutionary analyzes based on general and specific markers; iv) characterization of target genes: carbohydrate-active enzymes (CAZymes), sugar transporters (STs) and transcription factors (TFs); v) construction of the gene regulatory network (GRN); and vi) identification of target genes to obtain commercial strains. The molecular identification of 2HH wild-type strain and its repositioning in the *Oxalica* series stand out for the primordiality for comparative studies with other microorganisms. The deposited draft genomes of 2HH wild-type and S1M29 mutant at GenBank makes it possible to expand the molecular understanding of this fungus. Analysis of accumulated mutations in S1M29 mutant highlighted a wide range of mutations, highlighting BGL2 enzyme and FlbA transcription factor that probably contains the main mutations involved in hyperproduction of cellulases. We also identified that the albino phenotype of S1M29 mutant resulted from a mutation in ALB1 enzyme, which belongs to the DHN-melanin biosynthesis pathway. Our results related to evolutionary analyzes hypothesize a potential long-term mutualistic symbiosis between *P. echinulatum* 2HH and *Anobium punctatum*, whose environment-specific interactions could explain the difference in gene composition in relation to *Penicillium oxalicum* 114-2. Furthermore, the CAZyome characterization of *P. echinulatum* 2HH demonstrates that the genes of the cellulolytic system are predominantly orthologous to *P. oxalicum* 114-2, including a monooxygenase of AA16 family, described for the first time in both fungi. Furthermore, the sugar transportome characterization has demonstrated the diversity and specificity of STs from *P. echinulatum* 2HH, including eight families with specificity for different groups of sugars. Finally, the TFome characterization and the GRNs of *P. echinulatum* 2HH and *P. oxalicum* 114-2 comprise regulatory interactions that cover several biological processes, exploring diverse regulatory modules, such as CpcA, FF-7, COL-26, AmyR, ClrB, CreA and XlnR. This thesis contributes strongly in building a framework

for molecular understanding of *P. echinulatum* 2HH, revealing characteristics related to the cellulolytic enzymes production, sugar uptake, melanin production, cell wall composition and regulation of gene expression of this fungus. Knowledge of regulatory pathways, coupled with the characterization of CAZymes and STs provide tools to design commercial strains, enabling the use of *P. echinulatum* 2HH for large-scale 2G ethanol production. Finally, we highlight *P. echinulatum* 2HH as an important biotechnological ally for lignocellulosic biofuels production, helping in the global energy transition.

Keywords: *Penicillium echinulatum*, genome, cellulases, gene regulation, regulatory network, ethanol 2G.

SUMÁRIO

1 INTRODUÇÃO

As energias renováveis modernas incluem vários tipos de bioenergia e biocombustíveis sustentáveis e de baixo custo, que compreendem elementos cruciais para a transição energética. Apesar das dificuldades técnicas e do atraso na expansão comercial, as tecnologias para produção de etanol de segunda geração (etanol 2G) em larga escala, obtido a partir de biomassa lignocelulósica, tem evoluído bastante nos últimos anos. Como consequência do desenvolvimento dessas tecnologias de produção, é possível obter um aumento na produção deste biocombustível, sem aumentar a área de cultivo, dando um passo importante para tentar alcançar a independência em relação aos combustíveis fósseis. Além disso, pesquisas recentes sugerem a utilização de etanol para alimentar células combustíveis. Estas, por sua vez, viabilizam o fornecimento de energia elétrica para inúmeras aplicações, de *drones* até automóveis.

O etanol 2G pode ser produzido a partir da fermentação de açúcares gerados através da hidrólise enzimática de diferentes tipos de resíduos lignocelulósicos, os quais podem ser obtidos da palha e do bagaço da cana-de-açúcar, de resíduos do milho, de resíduos de madeira, do capim-elefante, da palha de arroz, etc. As enzimas necessárias para degradação da biomassa lignocelulósica compreendem celulases e hemicelulases, as quais geram glicose e pentoses que, por sua vez, são fermentadas para a produção de etanol 2G ou outros produtos biotecnológicos. No entanto, um dos desafios para este processo são os elevados custos das enzimas utilizadas na sacarificação do componente celulósico. Todavia, o progressivo interesse nesse processo apoia a prospecção e o estudo de novas espécies de fungos filamentosos especializados na produção deste tipo de sistema enzimático. E, consequentemente, o conhecimento da expressão gênica desses microrganismos permite o melhoramento de linhagens para viabilizar a produção comercial de enzimas.

Estudos prospectivos realizados no final da década de 70 buscavam por fungos filamentosos naturalmente especializados em degradar matéria vegetal celulósica. Um dos isolados do gênero *Penicillium*, denominado 2HH, mostrou-se apto para estudos biotecnológicos subsequentes devido à produção de celulases. Posteriormente classificado como *Penicillium echinulatum*, o isolado selvagem 2HH foi alvo de um programa a longo prazo, envolvendo mutagênese e seleção de mutantes. O mutante S1M29, gerado a partir do programa citado, apresenta potencial biotecnológico para a secreção de enzimas e conversão enzimática eficiente de bagaço de cana-de-açúcar em etanol 2G. A secreção de enzimas obtidas em distintos processos e períodos de cultivo são resultados da expressão de um conjunto de genes e os efeitos dos seus respectivos mecanismos reguladores. Apesar do alto potencial do mutante S1M29 para hidrolisar material lignocelulósico, a otimização do processo de produção de enzimas por esse fungo é entravada pela carência de conhecimentos fisiológicos, metabólicos, genômicos e, principalmente, regulatórios.

A transcrição é um passo fundamental para a decodificação de informações codificadas no DNA. Portanto, o conhecimento da regulação da transcrição é essencial para a compreensão da variação natural da expressão gênica. Dessa forma, a identificação das interações entre fatores de transcrição (TFs) e seus genes-alvo torna-se fundamental para compreensão da expressão gênica e para a engenharia de sistemas biológicos para os mais variados fins. A maioria dos genes que codificam enzimas que degradam a biomassa lignocelulósica estão sob o controle de diversos reguladores transcricionais. Em diferentes fungos, foram identificados reguladores transcricionais que ativam ou reprimem a expressão de genes envolvidos na degradação da biomassa lignocelulósica. Esses reguladores são muito diversos, compreendendo TFs, elementos cis-reguladores (sítios de ligação de fatores de transcrição (TFBSs)), RNAs, enzimas, transportadores de açúcares (STs), sinalizadores, entre outros.

Os avanços da genômica, pós-genômica e principalmente da biologia computacional abrangem uma série de métodos e ferramentas que permitem a compreensão da evolução natural de espécies e de mecanismos de regulação transcricional. Esses avanços constituem um passo importante para a análise dinâmica de regulação transcricional e para construção da rede de regulação de processos biológicos. Os sequenciamentos dos genomas do mutante S1M29 e do parental 2HH, realizados no ano de 2013, propiciaram novas direções para a construção de linhagens aprimoradas para aplicações industriais, lançando luz sobre aspectos que ajudam a identificar redes de regulação e novos alvos para melhorar a expressão do sistema celulolítico. Dessa forma, viabiliza-se a engenharia desse fungo, a fim de obter linhagens hiperprodutoras de misturas enzimáticas.

A disponibilidade dessas sequências genômicas do mutante e do parental selvagem estimula a busca por respostas às diversas perguntas em aberto sobre esse fungo filamentoso: i) A classificação morfológica realizada na década de 90 e a nomenclatura desta espécie estão corretas? ii) Quais são as características gerais dos genomas sequenciados? iii) Quais são as possíveis mutações acumuladas na linhagem S1M29 que causam a hiperprodução de celulases? iv) Quais são as possíveis mutações na linhagem S1M29 que causam o albinismo? v) Quais são os genes codificadores de CAZymes (Enzimas Ativas em Carboidratos) e quais são as enzimas que formam o sistema celulolítico deste fungo? vi) Como o nicho ecológico e a disponibilidade nutricional afetaram a evolução desta espécie em relação aos seus parentes? vii) Quais genes influenciam a produção de enzimas celulolíticas a partir das rotas de metabolismo de açúcares? viii) Quais TFs influenciam a produção de enzimas celulolíticas a partir de interações regulatórias com seus genes-alvo? ix) É possível obter interações regulatórias, a partir de biologia computacional, para inferência de uma GRN (rede regulatória de genes) global? x) Quais são os principais genes-alvo para melhoramento genético, pretendendo o incremento da produção de enzimas celulolíticas?

A busca por respostas para essas questões motivam a presente tese através dos objetivos apresentados na sequência.

1.1 Objetivos

O objetivo geral desta tese compreende a descoberta de conhecimento por meio da análise dos dados genômicos do sequenciamento das linhagens 2HH e S1M29 do gênero *Penicillium*, devidamente montados, anotados e depositados em bases de dados públicas.

Para alcançar o objetivo geral, foram traçados os seguintes objetivos específicos:

1. montar, anotar e depositar os genomas das linhagens 2HH e S1M29 nas bases de dados públicas DDBJ, ENA e GenBank;
2. efetuar a identificação molecular do isolado selvagem 2HH;
3. identificar as mutações que resultaram na hiperprodução de enzimas celulolíticas pelo mutante S1M29;
4. identificar as mutações que resultaram no albinismo do mutante S1M29;
5. efetuar análises comparativas e evolutivas a partir de marcadores genômicos gerais e específicos;
6. caracterizar funcionalmente os genes que expressam CAZymes;
7. caracterizar funcionalmente os genes que expressam transportadores de açúcares;
8. caracterizar funcionalmente os genes que expressam fatores de transcrição;
9. inferir a rede regulatória de genes global a partir de biologia computacional;
10. identificar genes-alvo para obtenção de linhagens visando a hiperprodução de enzimas celulolíticas.

Os capítulos subsequentes abordam a revisão bibliográfica (Capítulo 2) e o material e métodos (Capítulo 3). Em seguida, são apresentados os três manuscritos de artigos científicos que compõem os resultados alcançados nesta tese (Capítulo 4), bem como a discussão geral dos resultados (Capítulo 5). Por fim, são apresentadas as conclusões (Capítulo 6) e perspectivas futuras baseadas nas limitações desta pesquisa (Capítulo 7), bem como as referências utilizadas nesta tese.

2 REVISÃO BIBLIOGRÁFICA

Este capítulo apresenta uma revisão bibliográfica da literatura, embasando a metodologia para alcançar os objetivos traçados na presente tese, permitindo também a discussão dos resultados obtidos. São abordados os seguintes assuntos: i) levantamento cronológico de estudos realizados desde o isolamento da linhagem selvagem 2HH de *Penicillium*; ii) taxonomia do gênero *Penicillium*; iii) conceitos e métodos para montagem e anotação de genomas e para descoberta de conhecimento, abordando análises filogenéticas e redes regulatórias de genes; e iv) aplicações biotecnológicas com ênfase no uso de celulases para produção de biocombustíveis avançados.

2.1 Do isolamento à genômica: os quarenta anos de estudos da linhagem selvagem 2HH e de seus mutantes

O interesse crescente no processo de degradação de matéria vegetal lignocelulósica apoia a prospecção e o estudo da expressão gênica em microrganismos produtores de sistemas enzimáticos. Assim, os fungos são evidenciados, devido ao seu papel central como degradadores de macromoléculas na natureza. Estes microrganismos têm sido amplamente estudados por sua produção de sistemas enzimáticos que degradam macromoléculas de plantas, como a celulose e a hemicelulose (HYDE *et al.*, 2019).

A diversidade bioquímica e a adaptabilidade a diferentes ambientes são características biológicas bem conhecidas de fungos filamentosos. Dessa maneira, esses microrganismos são amplamente distribuídos por toda a biosfera, integrando estilos de vida saprófitos (essenciais para a ciclagem de nutrientes), patogênicos e simbióticos (com vários animais e plantas)(MUSZEWSKA *et al.*, 2017). Sua versatilidade biológica está alinhada com a capacidade de secretar uma grande variedade de enzimas que degradam as macromoléculas disponíveis no ambiente de crescimento. Destaca-se a capacidade de produção de enzimas lignocelulolíticas desses microrganismos, devido à sua co-evolução com as plantas (SILVA *et al.*, 2014).

A degradação enzimática de polissacarídeos vegetais por fungos apresenta várias aplicações industriais, como papel, tecidos, alimentos, ração animal, produtos químicos e biocombustíveis (HYDE *et al.*, 2019). Os fungos filamentosos dos gêneros *Aspergillus*, *Rhizopus*, *Trichoderma*, *Neurospora*, *Penicillium*, etc., são organismos naturalmente especializados na desconstrução da biomassa lignocelulósica. Esse recurso representa um potencial para a produção de biocombustíveis a partir de fontes renováveis (DALENA *et al.*, 2019).

Em alguns casos, insetos que ingerem madeira se beneficiam de substratos pré-digeridos

por microrganismos, incluindo fungos que habitam o seu trato digestório, degradando polissacarídeos em uma simbiose mutualística. Esse conhecimento biológico, aliado ao potencial das enzimas lignocelulolíticas, motivou a investigação e triagem de fungos simbiontes por várias décadas. Esperava-se que os fungos simbiontes, encontrados no trato digestório de larvas de broca de madeira, fossem naturalmente especializados na degradação da celulose, o principal polissacarídeo da biomassa vegetal. Tendo em vista que a desconstrução da celulose é alcançada através da atuação sinérgica de várias enzimas digestivas de insetos, juntamente com potenciais enzimas de bactérias e fungos simbiontes (PARKIN, 1940).

No ano de 1979, em um mural de madeira localizado nas dependências do Bloco A, edifício da Reitoria da Universidade de Caxias do Sul, Rio Grande do Sul, Brasil, foi isolado um exemplar do gênero *Penicillium*, denominado 2HH. Este fungo filamentoso foi encontrado no trato digestório de larvas do coleóptero *Anobium punctatum*. Esse coleóptero é comumente conhecido como besouro dos móveis, alimentando-se de madeira e reduzindo objetos de madeira a pó fino. O isolamento desse fungo foi publicado no Simpósio Internacional de Engenharia Genética que ocorreu na cidade de São Paulo no ano de 1981 (CARRAU *et al.*, 1981).

Em posterior classificação morfológica (não publicada) realizada na década de 90, esse fungo foi identificado como *Penicillium echinulatum*. Todos os trabalhos publicados antes da presente tese utilizam essa classificação. O uso de marcadores moleculares facilitou a taxonomia e identificação de espécies do gênero *Penicillium* a partir dos anos 90. Atualmente utilizam-se métodos padronizados para identificação e caracterização, os quais definem de forma precisa a estrutura taxonômica do gênero (VISAGIE *et al.*, 2014).

As larvas de *A. punctatum* normalmente vivem em madeira deteriorada que geralmente é enfraquecida e pré-digerida, permitindo que a larva force um caminho a seguir pela madeira (WHEELER; CROWSON, 1982). Considerando que lignina, celulose e hemicelulose fornecem uma função estrutural nas plantas e que a lignina é a responsável pela rigidez e sustentação (GLASS *et al.*, 2013), pode-se afirmar que a dieta das larvas é basicamente composta de celulose e hemicelulose, contendo apenas resíduos de lignina. Nesse sentido, especulações evolutivas que remetem aos primeiros experimentos de produção enzimática pela linhagem 2HH, sugerem uma possível simbiose mutualística a longo prazo entre a linhagem 2HH e as larvas de *A. punctatum* (CARRAU *et al.*, 1981). Essas hipóteses também incluem uma potencial adaptação natural para secreção de enzimas celulolíticas, como uma possível adaptação à dieta das larvas como única condição de crescimento disponível para o fungo. Atualmente essas hipóteses evolutivas são sustentadas apenas pela mistura de celulases secretada pelo isolado 2HH, a qual fornece uma formulação enzimática eficaz para sacarificação completa dos resíduos vegetais ricos em celulose e hemicellulose (SCHNEIDER *et al.*, 2016).

Geralmente, associações estáveis entre fungos e insetos em uma simbiose mutualística procedem de várias adaptações ambientais e respostas ao estresse. Mudanças no conteúdo do genoma fornecem informações sobre as adaptações que os organismos podem sofrer devido a

mudanças de nichos ecológicos ou associações com plantas ou animais hospedeiros (STAJICH, 2017). Os microrganismos que vivem dentro de um inseto possuem vantagens sobre os de vida livre, uma vez que no intestino os fungos são banhados por um suprimento regular de nutrientes. Outro benefício pode ser a dispersão direta. Os fungos de vida livre geralmente esgotam seu substrato e a dispersão para um novo substrato ocorre pelo vento, pela água ou por animais. No caso de organismos que vivem dentro de insetos, no entanto, a dispersão muda para se tornar altamente dependente dos insetos (VEGA; BLACKWELL, 2005). Esse modo de transmissão pode resultar em linhagens simbióticas que permanecem vinculadas a uma linhagem hospedeira a partir de uma transmissão vertical (WOOLFIT; BROMHAM, 2003).

Penicillium é um dos gêneros mais utilizados na biotecnologia e é conhecido por produzir uma vasta gama de CAZymes, incluindo o sistema celulolítico que atua na degradação da celulose. *Aspergillus* spp. e *Trichoderma reesei* tem sido os fungos mais estudados e utilizados comercialmente para produção de enzimas celulolíticas (AMORE; GIACOBBE; FARACO, 2013). No entanto, estudos anteriores mostraram que *Penicillium* spp. podem produzir sistemas enzimáticos com desempenho superior a *T. reesei* e *Aspergillus niger*. Assim, são justificados a identificação de novas espécies e os estudos para melhoramento genético deste gênero, a fim de incrementar a produção de celulases e propiciar a sua aplicação para produção de biocombustíveis de segunda geração (MARTINS *et al.*, 2008; GUSAKOV, 2011; GUSAKOV; SINITSYN, 2012; SINGHANIA *et al.*, 2013; LIU *et al.*, 2013; MÄKELÄ *et al.*, 2016; VAISHNAV *et al.*, 2018).

O potencial para secreção de celulases do isolado 2HH motivou uma série de estudos e publicações relevantes nos anos decorrentes do seu isolamento até a atualidade. A seguir, são apresentados em ordem cronológica, os principais estudos que foram conduzidos a partir do isolamento da linhagem selvagem 2HH até a obtenção de mutantes hiperprodutores de enzimas. Ressalta-se a obtenção de mutantes, a produção enzimática e a aplicação biotecnológica como produtor de enzimas para etanol 2G. Todas as cepas mutantes usadas para a produção de enzimas são derivadas do isolado selvagem 2HH, evidenciando essa espécie pelo potencial biotecnológico para a secreção de enzimas em relação às demais espécies do gênero *Penicillium*.

Dillon et al. (1992) obtiveram os primeiros mutantes a partir da exposição da linhagem selvagem 2HH à luz ultravioleta. Neste trabalho foi isolada uma colônia mutante de coloração rosa, denominada 3MUV243, que posteriormente deu origem aos mutantes albinos. É importante salientar que até a publicação deste trabalho não havia sido realizada a identificação da espécie, ou seja, o isolado selvagem 2HH era classificado como *Penicillium* sp.

Camassola et al. (2004) caracterizaram as atividades enzimáticas para celulases e β-glicosidases do mutante 9A02S1. Como resultado, foi identificado que a estabilidade térmica ocorre até 55 °C e a atividade ótima ocorre sob uma escala de pH entre 4 a 5. O mutante 9A02S1 foi obtido a partir de sucessivas etapas de mutagênese com exposição à luz ultravioleta e peróxido de hidrogênio, seguido por seleção de mutantes em meio suplementado com 2-desoxiglicose. Esse mutante é caracterizado por ser um mutante parcialmente desreprimido à glicose (DILLON

et al., 2006). Camassola & Dillon (2007) destacam o potencial do mutante 9A02S1 para produção de celulases e hemicelulases utilizando bagaço de cana-de-açúcar pré-tratado em fermentação em estado líquido (LSF) e farelo de trigo em fermentação em estado sólido (SSF).

As celulases também são utilizadas na indústria têxtil, atuando nas fibrilas de celulose da superfície do tecido. Em lavanderias de jeans, essas enzimas são utilizadas para clarear e obter o efeito de desgaste, produzindo a cor de peças jeans índigo (*stone washed*). No ano de 2008 foi realizada a avaliação de utilização do sistema celulolítico do mutante 9A02S1 nesse processo e os resultados indicam que as enzimas do fungo removem mais cor dos tecidos denim e produzem menos redeposição de corante (RAU *et al.*, 2008). Atualmente, a indústria têxtil compreende a única aplicação em escala industrial do sistema celulolítico deste fungo.

Rubini et al. (2010) realizaram a clonagem da principal endo-1,4-β-glicanase da família GH5-5 secretada pelo fungo, denominada EGL1. Este estudo realizou a expressão heteróloga da enzima em *Pichia pastoris* permitindo a sua caracterização. Sua temperatura ótima é 60 °C e a atividade ótima ocorre sob uma ampla escala de pH que varia de 5 a 9.

Diversos ciclos de mutagênese a longo prazo resultaram no mutante S1M29, o qual foi obtido a partir do mutante 9A02S1, empregando peróxido de hidrogênio e seleção de mutantes em meio suplementado com 2-desoxiglicose (DILLON *et al.*, 2011). As melhorias compreendem mutações no mutante S1M29, diferenciando-o de seu parental selvagem. Atualmente, este é o melhor mutante obtido, pois fornece uma melhor hidrólise de biomassa, devido a um aumento significativo nos títulos enzimáticos, se tornando o primeiro mutante com possibilidade de exploração comercial (SCHNEIDER *et al.*, 2016; SCHNEIDER *et al.*, 2018).

Ribeiro et al. (2012) utilizaram o mutante 9A02S1 para conduzir o estudo de 20 secretomas obtidos a partir de cinco diferentes condições de cultivo baseadas em bagaço de cana-de-açúcar. Foram identificadas 99 proteínas secretadas, em sua grande maioria (74%) foram CAZymes, dentre as quais mais de 80% atuam no complexo lignocelulose, reforçando o potencial do fungo como produtor de enzimas para degradação de celulose e hemicelulose. Destaca-se a produção das glicosil hidrolases (GH) identificadas, incluíndo enzimas como as endoglicanases e celobiohidrolases das famílias GH5, GH6, GH7, GH12, β-glicosidases da família GH3, xilanases das famílias GH10 e GH11, e hemicelulases desramificadoras das famílias GH43 e GH62.

No ano de 2013 foi realizado o sequenciamento completo dos genomas da linhagem selvagem 2HH e do mutante S1M29. Apesar das características gerais dos genomas terem sido publicadas (SCHNEIDER *et al.*, 2016), as respectivas sequências, montagens e anotações genômicas não foram disponibilizadas para comunidade científica. As possibilidades de estudos são amplamente enriquecidas com o sequenciamento desses genomas, motivando esta tese que contempla a descoberta de conhecimento ao explorar os genomas e publicá-los para comunidade científica. Esses dados biológicos permitem a diferenciação do mutante S1M29 e seu parental 2HH, a caracterização de conjuntos de genes de interesse, a construção da rede regulatória de genes, a identificação de genes-alvo para obter linhagens comerciais, além de possibilitar o

esclarecimento de hipóteses evolutivas. Todas esses campos de estudo são contemplados pela presente tese, salientando que a publicação dos genomas para comunidade científica permite uma vasta série de estudos subsequentes por diferentes grupos de pesquisa.

Novello et al. (2014) realizaram a avaliação da indução de enzimas com xilose em cultivos com a linhagem selvagem 2HH e os mutantes 9A02S1 e S1M29, resultando em maiores títulos enzimáticos utilizando o mutante S1M29. Também foi observado que a xilose atua como um indutor para a produção de xilanases e celulases, especialmente endoglicanases. Schneider et al. (2014) estudaram a influência das diferentes fontes de carbono na morfologia do fungo. Como resultado, foi observado que no meio formulado com celulose ou bagaço de cana-de-açúcar, o micélio cresceu mais disperso, o que favoreceu a secreção enzimática.

Estudos de hidrólise enzimática de capim-elefante empregando o mutante 9A02S1 verificaram que o pré-tratamento com hidróxido de sódio foi o que proporcionou as maiores liberações de açúcares redutores (MENEGOL *et al.*, 2014a; MENEGOL *et al.*, 2014b). Também foi observado que quando o capim-elefante foi hidrolisado no reator de hidrólise por rotação, a produção de etanol 2G foi aproximadamente o dobro do que foi produzido quando a biomassa foi hidrolisada em um reator estático (SR). Estes dados indicam que é possível produzir etanol 2G a partir de capim-elefante quando o tratamento da moagem e a hidrólise enzimática são executados ao mesmo tempo (MENEGOL *et al.*, 2016).

Zampieri et al. (2014) avaliaram 23 genes candidatos do mutante S1M29 para permitir estudos de expressão gênica usando PCR quantitativo em tempo real (qRT-PCR). O resultado desta avaliação indicou que β-actina (*actb*) foi o gene de referência mais estável expresso no mutante S1M29, o qual foi recomendado como um controle endógeno para estudos de expressão gênica de celulases e hemicelulases.

Zampieri (2015) conduziu o primeiro estudo de expressão gênica do sistema celulolítico do mutante S1M29. Esse estudo contemplou a avaliação da atividade enzimática, bem como a análise da expressão gênica das celulases, β-glicosidases, xilanases e swolenina em condições de indução em cultivos submersos. O estudo de qRT-PCR revelou a presença de quatro genes para endoglicanases com padrão de expressão distintos, um gene para celobiohidrolase, um gene para β-glicosidase, um gene para xilanase e um gene para swolenina. Foi observada ainda uma expressão coordenada dos genes codificadores de endoglicanases, celobiohidrolase, β-glicosidase e swolenina.

Schneider et al. (2016) realizaram a análise secretômica da linhagem selvagem 2HH e do mutante S1M29, identificando 165 proteínas. O secretoma obtido foi constituído principalmente por enzimas CAZy, sendo majoritariamente: glicosil hidrolases, carboidrato esterases, pectina liases, enzimas auxiliares e swolenina. No total, 36 proteínas foram exclusivas da linhagem selvagem, 18 proteínas foram exclusivas do mutante S1M29 e 111 foram comuns em ambas. A partir das 147 proteínas encontradas na linhagem selvagem, 63 proteínas (aprox. 43%) são CAZymes. Enquanto que, 57 proteínas (aprox. 44%) são CAZymes, a partir das 129 proteínas

encontradas na linhagem mutante. Esse estudo também sugere que a hipersecreção de enzimas do mutante S1M29 ocorre devido a mudanças, possivelmente em nível da regulação da expressão gênica, aumentando a capacidade de produção de enzimas extracelulares. Assim, o mutante S1M29 poderia ser empregado comercialmente para hidrólise de lignocelulósicos.

Posteriormente, foi conduzido um estudo do perfil enzimático do mutante S1M29 e do seu parental selvagem 2HH, cultivados em diferentes fontes de carbono (bagaço de cana-de-açúcar pré-tratado por explosão à vapor, celulose, glicose e glicerol) em diferentes tempos de cultivo. Para determinar as atividades enzimáticas, 23 substratos foram utilizados. O mutante S1M29 apresentou maiores títulos enzimáticos para a maioria das enzimas consideradas (celulases, hemicelulases, esterases e, em menor proporção, pectinases) e seu caldo enzimático foi utilizado para hidrólise enzimática da biomassa vegetal (SCHNEIDER *et al.*, 2018).

O estudo mais recente apresenta a produção de etanol à partir de madeira de *Eucalyptus globulus* utilizando primeiramente lacases produzidas por *Marasmiellus palmivorus* para deslignificar a biomassa lignocelulósica e posteriormente a utilização do sistema enzimático do mutante S1M29 para hidrólise enzimática da biomassa pré-tratada. A deslignificação da madeira resultou em uma diminuição de 31% no teor de lignina e um aumento de 10% no rendimento de etanol (SCHNEIDER *et al.*, 2020).

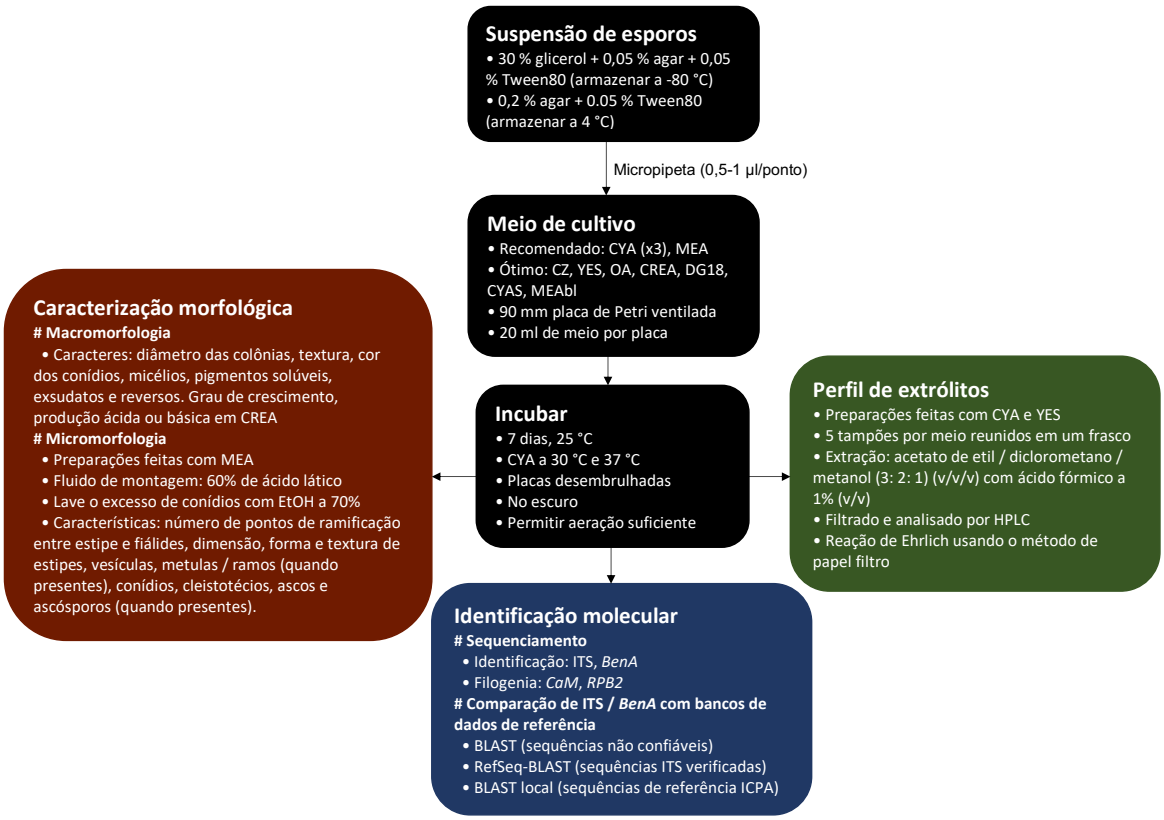
Apesar dos diversos estudos realizados nos últimos quarenta anos, conhece-se pouco sobre os mecanismos moleculares deste fungo. Conforme citado anteriormente, o sequenciamento completo dos genomas da linhagem 2HH e do mutante S1M29 viabilizam a busca por respostas a inúmeras questões que permanecem em aberto. Dentre as quais, algumas são respondidas na presente tese. Como, por exemplo, as duas principais características que diferenciam o mutante S1M29 da linhagem selvagem 2HH: a hiperprodução de celulases e o fenótipo albino.

2.2 Taxonomia do gênero *Penicillium*

O gênero *Penicillium* ocorre em diversos habitats, do solo à vegetação, ao ar, ambientes internos e em diversos produtos alimentares (HOUBRAKEN; VRIES; SAMSON, 2014). Antes do advento do sequenciamento de DNA nos anos 90, a taxonomia deste gênero era baseada somente na classificação e identificação morfológica. Embora as espécies de *Penicillium* sejam muito comuns e a estrutura taxonômica do gênero esteja bem definida, a identificação das espécies ainda é problemática. Os principais problemas incluem uma lista desatualizada de espécies aceitas e a falta de um banco de dados de sequências completas e verificadas para os marcadores moleculares.

Visagie et al. (2014) definem como consenso a utilização de uma abordagem polifásica para identificação e nomenclatura do gênero *Penicillium*: i) caracterização morfológica; ii) perfil de extrólitos; e iii) identificação molecular. Para tanto, as culturas fúngicas são realizadas a

Figura 1 – Fluxograma resumo dos métodos recomendados para a identificação e caracterização do gênero *Penicillium*.



Fonte: Visagie *et al.* (2014).

partir de diversos meios de cultivo [1] [2] [3] [4] [5] [6] [7] [8] [9]. A Figura 1 apresenta o fluxograma que resume os métodos recomendados para a identificação e caracterização do gênero *Penicillium*. No entanto, na taxonomia moderna os dados moleculares possuem mais peso do que os dados morfológicos ou extrólitos. ITS é o principal marcador molecular e compreende fragmentos de DNA ribossomal (rDNA) contendo os espaçadores internos transcritos (ITS1 e ITS2), a subunidade 5.8S e a região D1/D2 da subunidade 28S. Já o marcador BenA compreende um fragmento do gene da β-tubulina, CaM compreende um fragmento do gene da calmodulina e RPB2 é composto por um fragmento do gene da subunidade II da RNA polimerase.

[1] Czapek Yeast Autolysate agar (CYA)
[2] Blakeslee's MEA and CYA with 5% NaCl (CYAS)
[3] Malt Extract agar (MEA)
[4] Blakeslee's Malt extract agar (MEAbl)
[5] Czapek's agar (CZ)
[6] Yeast Extract Sucrose agar (YES)
[7] Oatmeal agar (OA)
[8] Creatine Sucrose agar (CREA)
[9] Dichloran 18% Glycerol agar (DG18)

2.3 Genômica e descoberta de conhecimento

A genômica compreende a análise *in silico* da sequência completa de nucleotídeos de um dado organismo, ou simplesmente genoma. Essa ciência pode se dedicar a determinar a sequência completa do DNA de organismos ou, em menor escala, pode se limitar à uma porção do genoma que seja de interesse. Sua relevância tem se expandido desde o primeiro mapeamento dos genes relacionados ao fenótipo da doença de Huntington no cromossomo 4 em humanos (GUSELLA *et al.*, 1983).

Os bioinformatas desempenham um papel chave estabelecendo a conexão entre os biólogos e os especialistas em sistemas computacionais (LAMPA *et al.*, 2013). A colaboração entre biólogos, bioinformatas e especialistas em sistemas computacionais deve ser estabelecida já na fase de planejamento de qualquer projeto de sequenciamento de genoma (EKBLOM; WOLF, 2014).

Um projeto de sequenciamento de genoma completo pode ser divido em três fases: sequenciamento, montagem e anotação (LANTZ *et al.*, 2018). Esta seção revisa o fluxo de trabalho envolvido nas fases de montagem e anotação genômica, com referência particular a genomas de fungos. Em seguida conceitos e métodos para análises genômicas *in silico* são apresentados, com intuito de fundamentar a descoberta de conhecimento a partir dos genomas anotados.

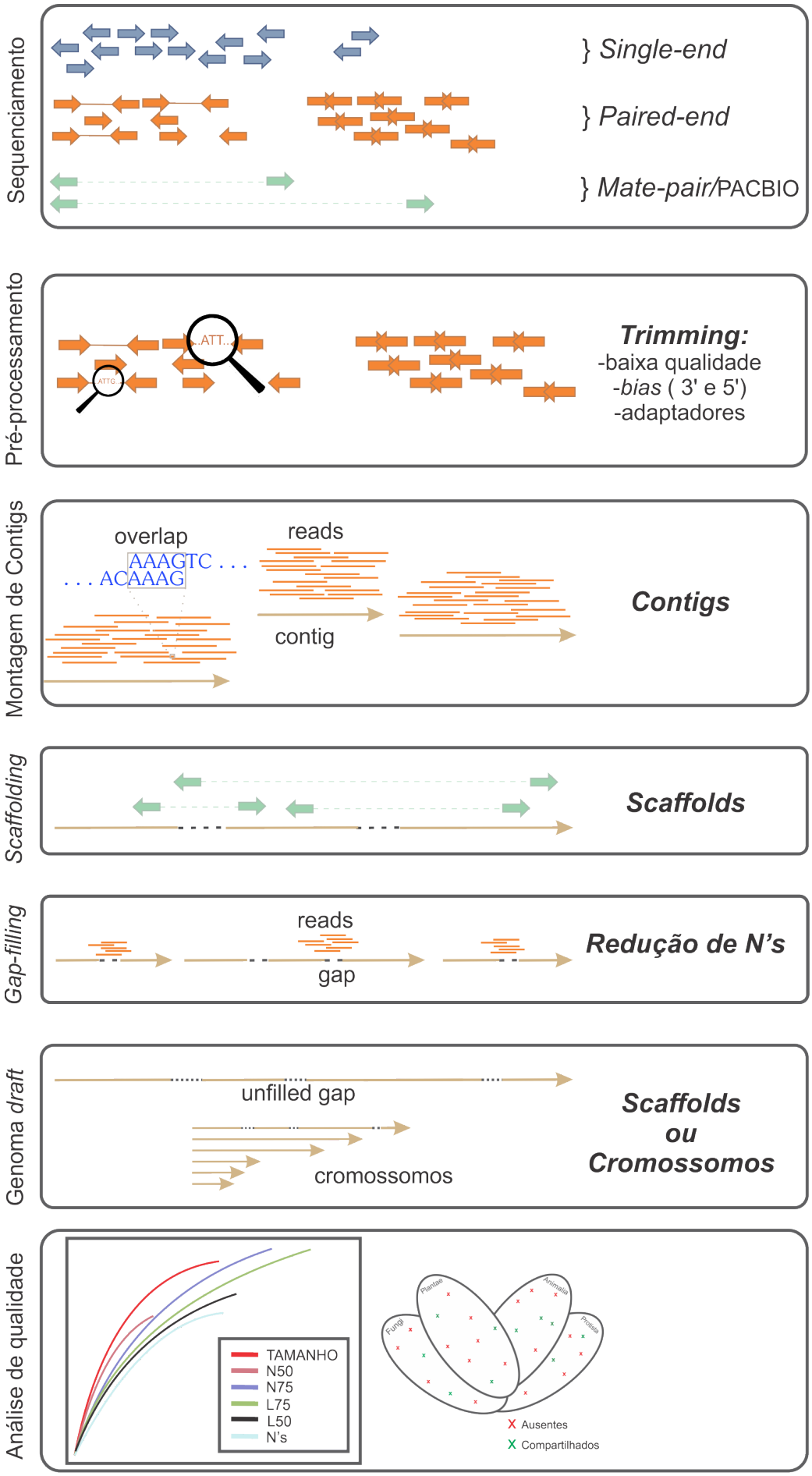
2.3.1 Montagem de Genoma

Basicamente existem duas abordagens para montagem de genoma: montagem guiada por referência e montagem *de novo*. A escolha da abordagem adequada ocorre a partir da existência ou não de um genoma montado anteriormente para ser usado como referência. No contexto desta tese não existe um genoma de referência para esta espécie, portanto somente a montagem *de novo* é abordada nesta seção.

Independentemente da tecnologia de sequenciamento escolhida, pode-se utilizar o mesmo processo para montagem *de novo*, conforme ilustrado na Figura 2. Atualmente, a maioria dos projetos de genoma usam estratégias de sequenciamento *shotgun* para sequenciamento de genoma. Em um primeiro passo, o DNA genômico é cortado em pequenos fragmentos aleatórios. Dependendo da tecnologia, os fragmentos são sequenciados independentemente com um determinado comprimento. Por exemplo, as tecnologias de leitura curta, como o Illumina HiSeq, tipicamente geram leituras de tamanho entre 50-300 pb. Outras tecnologias, como a Pacific Biosciences, produzem leituras mais longas (até 5 kb) (EKBLOM; WOLF, 2014).

Geralmente, é aconselhável a utilização de 45-50x de cobertura para bibliotecas de leituras curtas *paired-end*, complementadas com 45-50x de cobertura para bibliotecas de tamanho médio (3–10 kb) e 1–5x para bibliotecas de tamanho longo (10–40 kb). Tamanhos de leituras > 20 kb fazem uma grande diferença para a contiguidade final do genoma e para a etapa de

Figura 2 – Ilustração das etapas do processo de montagem *de novo* de genoma.



Fonte: Adaptada de Sohn e Nam (2018), Ekblom e Wolf (2014).

scaffolding. Ao usar apenas bibliotecas de leituras curtas, é necessária uma cobertura total de leitura alta (>100x) para obter um genoma com contiguidade. Uma cobertura baixa pode resultar em uma montagem altamente fragmentada e em problemas graves para anotação e identificação de variantes (EKBLOM; WOLF, 2014).

Antes da montagem propriamente dita, é necessária uma etapa de pré-processamento, responsável por avaliar a qualidade dos dados de sequenciamento. São avaliados diversos parâmetros como o conteúdo geral de GC, a abundância de repetições, a proporção de leituras duplicadas e a existência de contaminantes. Nesta etapa, também são removidos adaptadores utilizados para o sequenciamento, leituras de baixa qualidade, bem como *bias* localizados nas extremidades 3' e 5' das leituras (LANTZ *et al.*, 2018). Diversos softwares podem ser utilizados para avaliar a qualidade dos dados do sequenciamento e para executar o pré-processamento, tais como FastQC (ANDREWS *et al.*, 2015) e Trim Galore (KRUEGER, 2015), respectivamente.

As etapas de montagem de *contigs*, *scaffolding*, *gap-filling* e geração do genoma *draft* geralmente são executadas pelos softwares montadores. Esses softwares implementam algoritmos que são utilizados para reunir as leituras em longas sequências de consenso sem *gaps*, chamadas *contigs*. Para uma montagem correta, é importante que exista sobreposição suficiente entre as leituras em cada posição no genoma, o que requer alta cobertura de sequenciamento. Presumivelmente, para leituras mais longas, pode-se esperar mais sobreposição, reduzindo a cobertura total necessária. Nota-se que existe uma infinidade de ferramentas relacionadas ao processo de montagem. Escolher o conjunto de ferramentas mais adequado para os dados disponíveis nem sempre é uma tarefa fácil, portanto, na etapa de montagem, várias montadores devem ser testados em paralelo e os resultados são comparados na etapa de análise de qualidade (EKBLOM; WOLF, 2014; LANTZ *et al.*, 2018).

Em seguida, na etapa de *scaffolding*, os *contigs* são conectados por leituras longas (*mate-pair* / PACBIO), que geralmente se originam de grandes fragmentos de DNA de várias kilobases de comprimento. O conjunto ordenado de *contigs* conectados é definido como um *scaffold*. Uma vez que os *contigs* são unidos em *scaffolds*, se não houver sobreposição entre os *contigs*, espaços chamados *gaps* permanecem entre os *contigs*. Para esses gaps, bases indefinidas (N) e distâncias aproximadas são estimadas. Posteriormente, os *gaps* são cuidadosamente preenchidos usando outras leituras independentes (compreendendo a etapa de *gap-filling*) para concluir a montagem do genoma *draft*. Denomina-se genoma *draft*, aquele que não está completo em nível de cromossomos, contendo *gaps* e bases indefinidas (SOHN; NAM, 2018).

Na análise de qualidade, frequentemente ocorrem ajustes dos parâmetros e os montadores são executados novamente para obter montagens melhores. O objetivo geralmente é criar uma montagem de genoma com as sequências montadas mais longas possíveis (montagem menos fragmentada) e com o menor número de bases indefinidas. Dessa forma, são executados softwares estatísticos, como o Quast (GUREVICH *et al.*, 2013), que avaliam todas as montagens geradas na fase anterior para que a montagem que contenha as melhores métricas baseadas em tamanho

seja escolhida (LANTZ *et al.*, 2018).

Além das métricas baseadas em tamanho, é importante avaliar as montagens em relação à completude, identificando os genes ortólogos conservados. Essa avaliação é mais acurada que as métricas baseadas em tamanho e fundamental para avaliar o genoma *draft* antes da anotação. Softwares como BUSCO (WATERHOUSE *et al.*, 2018) podem ser usados para avaliar as montagens, fornecendo medidas quantitativas baseadas em expectativas evolutivas do conteúdo de genes ortólogos universais de cópia única, selecionados do banco de dados OrthoDB (KRIVENTSEVA *et al.*, 2015). Por exemplo, para fungos filamentosos do gênero *Penicillium*, é recomendado o uso do conjunto de dados *Eurotiomycetes*, contemplando um total de 4046 genes, que podem ser utilizados para realizar esta a avaliação (AGUILAR-PONTES *et al.*, 2018).

2.3.2 Anotação Genômica

O termo "Anotação Genômica" inclui a identificação de sequências que codificam proteínas e sequências não codificadoras (e.g., sequências repetitivas, DNA ribossômico (rDNA), e RNA não codificante (ncRNA)) em genomas, atribuindo informações biológicas (metadados) a esses elementos genômicos identificados (HARIDAS; SALAMOV; GRIGORIEV, 2018).

Embora a anotação do genoma envolva a caracterização de uma infinidade de elementos biologicamente significativos em uma sequência genômica, na prática, o esforço despendido para anotação genômica concentra-se na predição correta de sequências codificadoras de proteínas (CDSs) e na atribuição de nomes e de funções com significado biológico para esses genes (EKBLOM; WOLF, 2014).

Contudo, isso não menospreza o papel essencial desempenhado por sequências não codificantes na regulação transcricional, mas principalmente porque as abordagens para caracterizá-las são razoavelmente diretas (e.g., detecção de ncRNA), ou são o foco de análises muito especializadas (e.g., TFBSs e elementos promotores) (LANTZ *et al.*, 2018).

Esse processo de anotação de sequências de DNA consiste em várias etapas sucessivas, sendo tipicamente complicado por envolver uma grande quantidade de ferramentas computacionais e muitos arquivos de entrada e saída (EKBLOM; WOLF, 2014). Esses conjuntos de ferramentas de anotação são geralmente chamados de *pipelines* de anotação. A qualidade da anotação genômica é fortemente dependente da qualidade da montagem e da disponibilidade de dados associados, tais como sequências de RNA e proteínas do organismo em questão ou de algum organismo intimamente relacionado (LANTZ *et al.*, 2018; HARIDAS; SALAMOV; GRIGORIEV, 2018).

Embora os *pipelines* de anotação genômica costumam ter detalhes diferentes, eles compartilham um conjunto principal de recursos. Geralmente, a anotação de estruturas gênicas é dividida em duas fases distintas. Na primeira fase, a fase computacional, ocorre a identificação de sequências que codificam proteínas no genoma também conhecida como predição de genes.

Na segunda fase, a fase de anotação, ocorre a atribuição de informações biológicas aos elementos genômicos, sendo comumente chamada de anotação funcional (YANDELL; ENCE, 2012).

2.3.2.1 Predição de genes

A predição de genes é o processo de determinar corretamente a localização e a estrutura dos genes codificadores de proteínas em um genoma. Esse processo está bem estabelecido e conta com o suporte de muitos algoritmos de sucesso desenvolvidos nas últimas décadas. Esta fase inicia-se com a identificação de sequências repetitivas, etapa fundamental para garantir o funcionamento adequado dos algoritmos de predição.

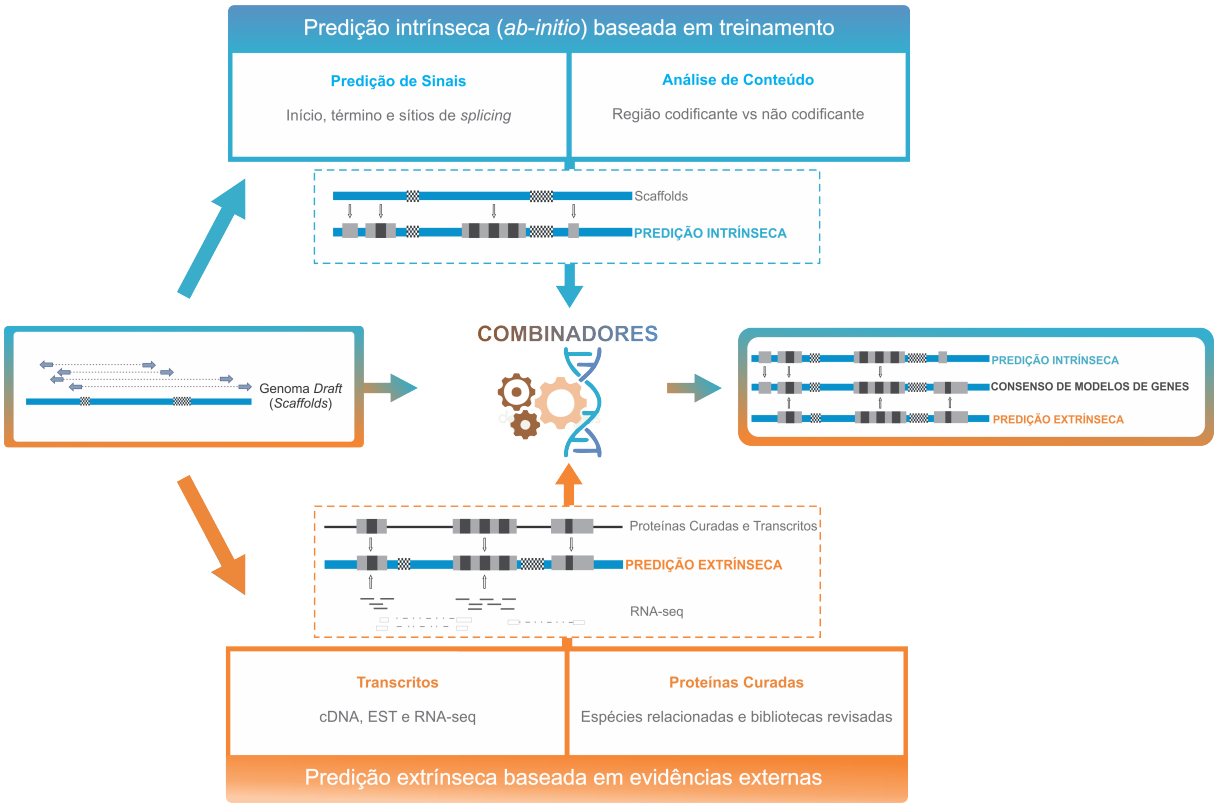
A maioria dos genomas de organismos mais complexos abriga uma abundância de sequências repetitivas que precisam ser excluídas dos passos subsequentes. Existem vários tipos de repetições, desde casos simples que compreendem dezenas a várias centenas de repetições de um mesmo motivo longo de nucleotídeos a genes móveis, elementos transponíveis ou fragmentos de genes de origem viral (BAO; KOJIMA; KOHANY, 2015). Portanto, este é o primeiro passo da anotação automática de genes, sendo realizado por ferramentas computacionais especializadas. Após o mascaramento de repetições, o próximo passo envolve a predição de genes propriamente dita (YANDELL; ENCE, 2012; EKBLOM; WOLF, 2014).

Em geral, como pode ser observado na Figura 3, existem três abordagens principais para predição de genes em um genoma: intrínseca (ou *ab-initio*), extrínseca e combinada. Enquanto que a abordagem intrínseca se concentra apenas na informação que pode ser extraída da própria sequência genômica, como a probabilidade da mesma ser uma sequência codificadora de proteína e o reconhecimento de sítios de *splicing*, a abordagem extrínseca usa a similaridade a outros tipos de sequência como informação (e.g., transcritos e/ou proteínas). Existem vantagens e desvantagens inerentes a cada uma das abordagens, assim, as duas abordagens podem ser combinadas computacionalmente para melhorar a acurácia das predições, dando origem à abordagem combinada (LANTZ *et al.*, 2018).

A predição intrínseca de genes tenta identificar genes usando modelos estatísticos, os quais necessitam de treinamento e otimização. Esta categoria de algoritmos realiza a busca sistemática na sequência de DNA por certos sinais indicadores de genes codificadores de proteínas. Os modelos estatísticos dependem de características genômicas específicas do organismo, como frequências de códons e distribuições de comprimentos de íntron-éxon, para distinguir genes de regiões intergênicas e para determinar estruturas íntron-éxon. Dessa forma, o objetivo dos preditores *ab-initio* é a busca por fases de leitura aberta (ORFs) (de *open reading frame* em inglês, trechos de sequência sem códons de parada e potencialmente codificando uma proteína).

Um bom conjunto de treinamento é primordial para esta abordagem, ou seja, um conjunto de genes estruturalmente bem anotados, usados para construir modelos e para treinar as ferramentas de predição de genes. Como cada genoma é diferente, esses modelos e parâmetros devem ser específicos para cada genoma e, portanto, precisam ser reconstruídos e retreinados

Figura 3 – Ilustração das abordagens para predição de genes.



Fonte: Adaptada de Lantz *et al.* (2018).

para cada nova espécie. Esta é, no entanto, também a grande vantagem desta abordagem, já que é capaz de realizar a predição de genes em rápida evolução e genes específicos de uma espécie (LANTZ *et al.*, 2018).

A maioria dos preditores de genes *ab-initio* vem com arquivos de parâmetros pré-calculados para alguns genomas clássicos. No entanto, a menos que o genoma a ser anotado esteja intimamente relacionado a um organismo modelo para o qual os arquivos de parâmetros pré-compilados estejam disponíveis, o preditor de genes precisa ser treinado para realizar a predição no genoma que está sendo estudado, pois mesmo organismos intimamente relacionados podem diferir em relação ao comprimento de íntrons e conteúdo GC (YANDELL; ENCE, 2012).

Outra vantagem dos preditores *ab-initio* é que, em princípio, eles não precisam de evidências externas para identificar um gene ou para determinar sua estrutura íntron-éxon. No entanto, esta categoria de algoritmos geralmente limita-se a encontrar CDSs e não contemplam regiões não traduzidas (UTRs) ou transcritos de *splicing* alternativo. Essa abordagem também tende a falhar na precisão, embora a maioria das ferramentas computacionais possam ser treinadas para ajustar seus parâmetros de predição às características específicas do organismo, melhorando assim a precisão (YANDELL; ENCE, 2012).

A predição extrínseca, por outro lado, baseia-se em evidências externas que podem ser alinhadas sobre o genoma de interesse. Uma das abordagens compreende a predição de genes

baseada em homologia de proteínas, sendo realizada a partir do mapeamento de proteínas de outros organismos sobre o genoma de interesse. Esta abordagem de predição de genes permite explorar um vasto número de sequências proteicas revisadas que encontram-se disponíveis em bancos de dados públicos (e.g., NCBI/RefSeq, UniProtKB/Swiss-Prot) (LANTZ *et al.*, 2018).

Conforme Gabaldón e Koonin (2013), ortólogos são tipicamente os genes mais similares nas respectivas espécies em termos de sequência, estrutura, arquitetura de domínio e função. Assim, sequências de proteínas revisadas de outras espécies fornecem uma boa indicação sobre a presença e localização de genes. Como as sequências polipeptídicas geralmente são mais conservadas do que as sequências de nucleotídeos, elas ainda podem ser alinhadas mesmo a partir de espécies relacionadas mais distantes.

Embora seja muito útil para determinar a presença de genes, esta abordagem nem sempre fornece informações precisas sobre a estrutura exata de um gene. Portanto, sugere-se a utilização de proteínas bem anotadas e revisadas de espécies estreitamente relacionadas para elevar a acurácia dos modelos de genes preditos por essa abordagem (LANTZ *et al.*, 2018; HARIDAS; SALAMOV; GRIGORIEV, 2018).

Outra abordagem extrínseca compreende a predição de genes baseada em homologia de transcriptoma. As informações de transcriptoma, sejam etiquetas de sequências expressas (ESTs) (de *expressed sequence tags* em inglês), DNA complementar (cDNA), RNA-Seq ou outros tipos de transcritos disponíveis, desempenham um papel ainda mais importante na predição extrínseca, fornecendo informações muito precisas para a predição correta da estrutura dos genes (LANTZ *et al.*, 2018).

O preditor de genes baseado em homologia de transcriptoma utiliza o conjunto completo de transcritos disponíveis para construir modelos de genes a partir dos transcritos alinhados. Os dados de RNA-Seq, de preferência fita-específica, podem ser usados de duas maneiras diferentes (HAAS *et al.*, 2011; ZHAO *et al.*, 2011; CONESA *et al.*, 2016): (i) as sequências geradas pelo RNA-Seq são mapeadas diretamente sobre o genoma de interesse para realizar a identificação de transcritos; (ii) a montagem do transcriptoma ocorre sem genoma de referência, gerando o conjunto completo de transcritos a partir do RNA-Seq.

A predição extrínseca utiliza diferentes fontes de evidências como ESTs, cDNA e proteínas de espécies intimamente relacionadas. Estas evidências de espécies próximas podem ser obtidas em bancos de dados específicos (e.g., para fungos: MycoCosm). Também é comum o uso de bibliotecas curadas de proteínas (e.g., UniProtKB/Swiss-Prot e NCBI/RefSeq). Devem ser evitadas as proteínas preditas que não tenham sido revisadas e curadas, isto porque modelos não validados podem piorar a precisão dos preditores. Todas as evidências selecionadas devem ser alinhadas ao genoma e, em seguida, os dados de RNA-Seq (transcritos) também devem ser alinhados, quando disponíveis. Os sítios de *splicing* devem então ser identificados, e as evidências devem ser pós-processadas e agrupadas antes de serem enviadas para a ferramenta computacional inferir o conjunto final de modelos de genes.

De acordo com Yandell e Ence (2012), a predição de genes orientada por evidências tem um grande potencial para melhorar a qualidade da predição de genes em genomas recém-sequenciados, mas na prática pode ser difícil de usar. Esse processo é oneroso e exige o conhecimento aprofundado de diversas ferramentas computacionais especializadas, sendo um dos principais obstáculos que os *pipelines* de anotação tentam superar.

Devido à estrutura íntron-éxon dos genes de organismos eucariotos, a predição de genes é uma das partes mais desafiadoras da anotação genômica. Tendo em vista que uma das principais dificuldades na anotação genômica é a distinção entre genes codificadores de proteínas, transposons e pseudogenes. Enquanto que os preditores *ab-initio* são precisamente caracterizados como preditores de novos genes, os preditores baseados em evidências extrínsecas são geralmente necessários para estabelecer conclusivamente que um gene predito é funcional (EKBLOM; WOLF, 2014; KEILWAGEN *et al.*, 2018).

Geralmente, os autores recomendam o uso de várias abordagens de predição de genes para combinar diferentes tipos de evidências para a anotação: *ab-initio*, baseada em homologia de proteínas e baseada em homologia de transcriptoma. Nos últimos anos, várias abordagens computacionais foram desenvolvidas com intuito de combinar múltiplas fontes, permitindo um incremento significativo na acurácia da predição de genes codificadores de proteínas (HAAS *et al.*, 2011; EKBLOM; WOLF, 2014; LANTZ *et al.*, 2018; KEILWAGEN *et al.*, 2018).

Ferramentas computacionais que implementam a abordagem combinada são chamadas de Combinadores, elas tomam como entrada uma sequência genômica e implementam um método computacional para construir modelos de genes a partir de evidências geradas a partir de um conjunto diversificado de fontes (ALLEN; PERTEA; SALZBERG, 2004).

Os Combinadores são provavelmente as ferramentas de predição de genes mais populares e amplamente utilizadas. No entanto, nem todos esses Combinadores são iguais. Enquanto alguns simplesmente escolhem o modelo mais apropriado ou constroem um consenso a partir das evidências de entrada fornecidas para um determinado *locus*, outros têm uma abordagem mais integrada na qual a predição intrínseca pode ser modificada pelos dados extrínsecos, resultando em uma predição mais consistente (LANTZ *et al.*, 2018).

Ao executar a anotação genômica, é preciso fazer escolhas, não apenas em relação às ferramentas que serão utilizadas, mas também em relação às fontes e tipos de evidências que serão utilizadas em cada etapa. Obviamente a escolha deve ir em direção às fontes de dados mais confiáveis, implicando muitas vezes em evidências menos abrangentes. Por outro lado, o uso de informações de qualidade inferior levará inevitavelmente a um resultado de predição de genes inferior (LANTZ *et al.*, 2018).

Embora as ferramentas de predição geralmente forneçam bons resultados, elas continuam sendo propensas a erros, a validação qualitativa é importante (e.g., avaliando o comprimento das ORFs). A inspeção visual da anotação é outro componente para detectar problemas sistemáticos

como genes faltantes, predições falsas, vazamento de íntron (íntrons sendo anotados como éxons devido à presença de pré-mRNA), e modelos de genes desmembrados ou agrupados, os quais levam erros ao conjunto final de genes. Embora a revisão manual dos modelos de genes despenda muito tempo, é uma etapa extremamente necessária para gerar um conjunto de genes preciso e confiável. Algumas ferramentas são particularmente úteis, pois permitem que o usuário edite a estrutura do gene predito diretamente por meio de uma interface visual (HAAS *et al.*, 2011; EKBLOM; WOLF, 2014; MCDONNELL; STRASSER; TSANG, 2018).

Assim como ocorre na avaliação das montagens, a predição de genes também deve ser avaliada em relação à completude, para que os genes ortólogos conservados sejam identificados. Pode-se utilizar os mesmos conjuntos de genes ortólogos universais obtidos de OrthoDB, executando o software BUSCO em modo de avaliação de proteínas (WATERHOUSE *et al.*, 2018).

2.3.2.2 Anotação funcional

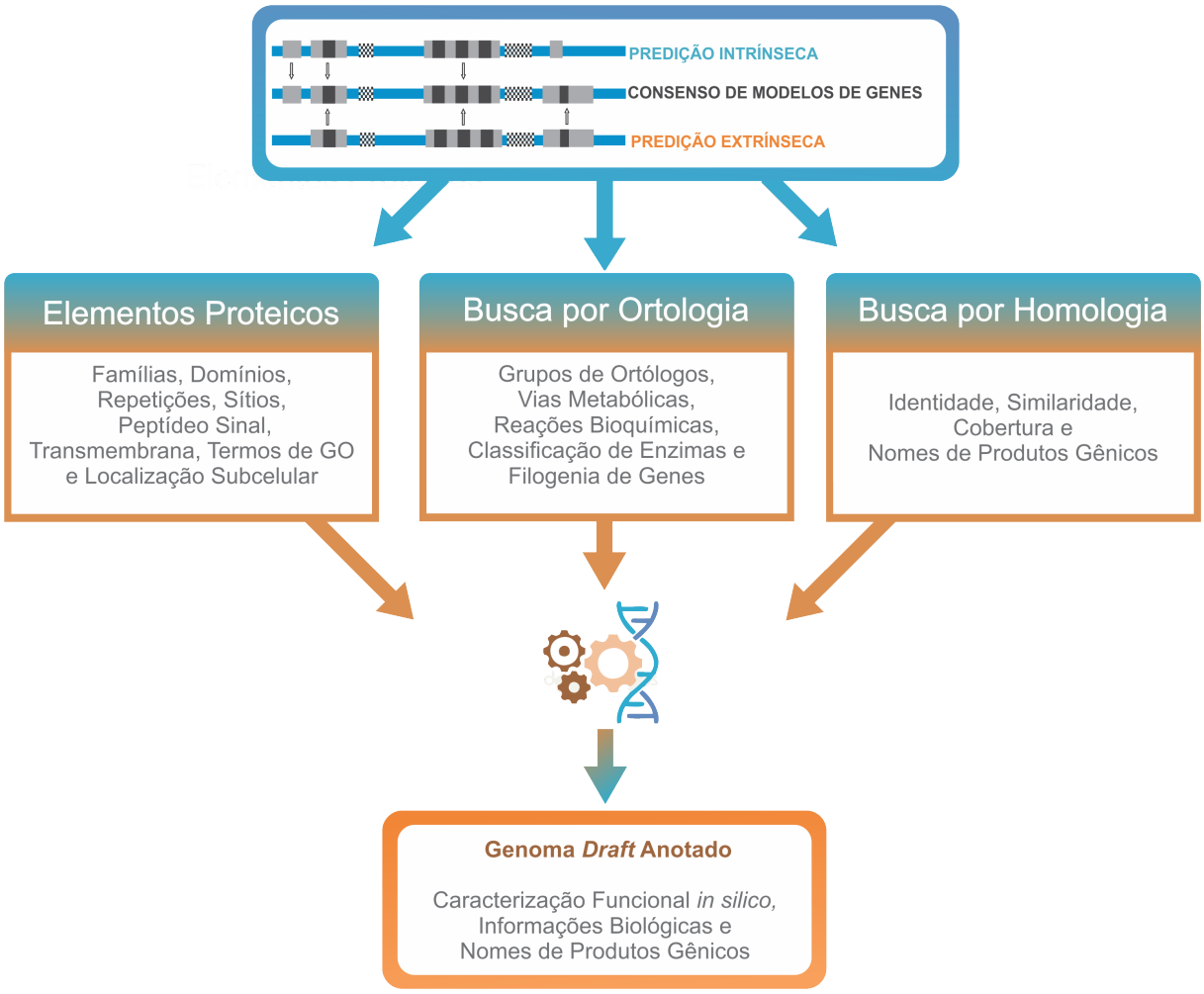
A anotação funcional consiste em atribuir informações biológicas significativas às sequências codificadoras de proteínas e aos seus elementos derivados (e.g. gene, RNA mensageiro (mRNA)), analisando a estrutura e a composição da sequência, bem como considerando o que se sabe à respeito de espécies relacionadas, que podem ser usadas como referência (LANTZ *et al.*, 2018).

O objetivo principal desta fase é a atribuição de nomes de produtos gênicos, geralmente baseados na caracterização funcional *in silico* dos genes preditos. A caracterização funcional dos elementos genômicos compreende diversas informações biológicas como função bioquímica, função biológica, interações proteicas e mecanismos de regulação e expressão. A caracterização destes elementos permite uma melhor compreensão das propriedades gênicas específicas, como as vias metabólicas e as semelhanças em comparação com espécies intimamente relacionadas (HAAS *et al.*, 2011; LANTZ *et al.*, 2018).

Existem várias ferramentas disponíveis para anotação funcional que permitem que os usuários obtenham anotações para seu conjunto de genes de interesse por meio de bancos de dados públicos. As ferramentas podem ser executadas individualmente e, em seguida, os resultados são combinados. No entanto, existem fluxos de trabalho disponíveis que fornecem todo o processo de anotação funcional de forma automatizada. Esses *pipelines* podem incluir a instalação das ferramentas necessárias e os bancos de dados correspondentes, ou os usuários podem fazer essa instalação por conta própria e o *pipeline* apenas fornece um fluxo estruturado para a análise (LANTZ *et al.*, 2018).

Em um *pipeline* típico de anotação funcional, conforme ilustrado na Figura 4, informações funcionais são atribuídas às proteínas preditas. O processo implementa três rotas paralelas para a caracterização funcional. A primeira refere-se aos domínios, motivos e famílias de proteínas, a segunda corresponde à busca de ortólogos e, por fim, a terceira compreende à busca por

Figura 4 – Ilustração de um *pipeline* para anotação funcional.



Fonte: Adaptada de Lantz *et al.* (2018).

homologia. Em síntese, as saídas das três fontes de informações são agrupadas para elevar o nível de acurácia da caracterização funcional e da atribuição de nomes de produtos gênicos (LANTZ *et al.*, 2018).

Antes da atribuição de nomes de produtos gênicos, é importante caracterizar elementos funcionais adicionais que incluem domínios, motivos, famílias das proteínas, vias metabólicas, localização subcelular da proteína, entre outros. A anotação funcional ainda pode agregar informações específicas que são relevantes para um determinado reino ou filo. Tomados em conjunto, os perfis de função gerais e especializados fornecem uma visão abrangente das características bioquímicas de um genoma, que podem ser correlacionadas com os fenótipos biológicos de uma espécie (HAAS *et al.*, 2011).

A função das proteínas preditas pode ser computacionalmente inferida com base na similaridade entre a sequência de interesse e outras sequências de diferentes repositórios públicos (e.g., BLASTP sobre UniProtKB/Swiss-Prot). Os protocolos de anotação costumam adotar critérios rigorosos para atribuição de nomes de produtos gênicos baseados em similaridade de

sequências.

Por exemplo, para McDonnell, Strasser e Tsang (2018), se o modelo do gene e o seu BLASTP corresponderem a uma identidade $\geq 98\%$ ao longo de todo o seu comprimento, então as duas proteínas são consideradas funcionalmente equivalentes e se a proteína revisada for caracterizada experimentalmente, então pode ser atribuído o mesmo nome de gene e o mesmo nome do produto gênico à proteína que está sendo anotada. Quando a identidade entre as sequências proteicas for $\geq 70\%$ e a cobertura da consulta $\geq 70\%$ é atribuído somente o nome do produto gênico da proteína revisada à proteína que está sendo anotada. Outro critério complementar ainda é adicionado por Haas et al. (2011), indicando que além da identidade e da cobertura $\geq 70\%$, a diferença de comprimento entre a proteína revisada e a proteína que está sendo anotada deve ser $\leq 10\%$.

Identidade e/ou cobertura inferiores aos citados são abordados de diferentes formas nos protocolos de anotação funcional. Para os genes restantes com nomes de produtos não atribuídos, é comum atribuir uma função mais geral com base em seu(s) domínio(s) conservados obtidos a partir da caracterização dos elementos funcionais, principalmente obtidos a partir dos bancos de dados InterPro (MITCHELL *et al.*, 2019) ou Pfam (EL-GEBALI *et al.*, 2019).

Deve-se tomar cuidado ao atribuir resultados baseados em similaridade de sequência, tendo em vista que duas sequências cuja evolução foi independente poderiam ser consideradas homólogas, por compartilharem alguns domínios comuns. Assim, sempre que possível, é aconselhável o uso de sequências ortólogas para fins de anotação, em vez de simplesmente sequências homólogas (LANTZ *et al.*, 2018).

A transferência de anotação funcional baseia-se na conjectura ortologia-função, a qual presume que ortólogos realizam funções biologicamente equivalentes em diferentes organismos (GABALDÓN; KOONIN, 2013). Dessa forma, a utilização de proteínas ortólogas, bem anotadas e revisadas de espécies estreitamente relacionadas, podem servir de embasamento para anotação funcional de um genoma. Essa abordagem de anotação funcional baseada em ortologia é apoiada por diversos bancos de dados de grupos de ortólogos e ferramentas que dão suporte à identificação de ortólogos. Por fim, para a nomenclatura de produtos gênicos, comumente são utilizadas as diretrizes internacionais de nomenclatura adotadas tanto pelo GenBank (SAYERS *et al.*, 2018) quanto pelo UniProtKB/Swiss-Prot (CONSORTIUM, 2019).

Com o crescente número de sequências em repositórios públicos, é possível realizar várias pesquisas e combinar os resultados obtidos para gerar uma anotação consensual. A caracterização funcional dos elementos genômicos é um processo complexo e propenso a erros, os *pipelines* de anotação funcional automatizada acabam por acumular e propagar os erros em bases de dados públicas. Portanto, uma curadoria manual é muitas vezes necessária para avaliar vários tipos de evidências e elevar o grau de confiabilidade da anotação funcional (LANTZ *et al.*, 2018).

No entanto, a análise aprofundada de todo genoma exige esforço e tempo, evidenciando a importância da adoção de critérios rigorosos para atribuição de nomes de produtos gênicos. Assim, em muitos casos a curadoria pode se limitar à uma porção do genoma que seja de interesse. Em última análise, a verificação experimental é a única maneira de ter certeza de que a caracterização funcional dos produtos gênicos está correta.

Além da curadoria manual, outro fator crucial para elevar o grau de confiabilidade da anotação compreende a escolha de ferramentas computacionais e de fontes de dados confiáveis e revisadas. As escolhas influenciam diretamente no esforço e no tempo gasto para essa atividade, sendo um dos fatores determinantes para o sucesso do processo de anotação. Vale ressaltar que as ferramentas computacionais geralmente são muito específicas para determinados tipos de dados de entrada e podem não ser capazes de analisar outros formatos, necessitando conversões ou até inviabilizando seu uso. Diversos softwares são necessários para a anotação e as respectivas instalações devem ocorrer em um sistema operacional baseado em Unix, usando a documentação incluída em cada um deles. Uma vez que o genoma possui uma anotação confiável, os estudos subsequentes compreendem vastas possibilidades de descoberta de conhecimento.

2.3.3 Descoberta de conhecimento

A identificação de padrões em grandes conjuntos de dados estruturados, semiestruturados e não estruturados é frequentemente chamada de descoberta de conhecimento. Nos últimos anos, muitas abordagens de descoberta de conhecimento foram desenvolvidas, incluindo métodos para pré-processar, integrar, analisar e interpretar dados complexos com o objetivo de identificar hipóteses testáveis (SIMSKE, 2019). O processo de descoberta de conhecimento em conjuntos de dados biológicos não seria possível sem a utilização de softwares especializados. Os quais, muitas vezes fazem uso de métodos, técnicas e ferramentas de aprendizado de máquina para realizar inferências biológicas, viabilizando decisões humanas.

O processo de descoberta de conhecimento envolve a transformação e estruturação dos dados biológicos. A alta conectividade entre os dados biológicos facilita a inferência de redes, constituídas por nós e conexões entre eles, compreendendo um modelo de dados flexível que pode capturar grande parte da complexidade e interconectividade dos dados (HUBER *et al.*, 2007). Além disso, as redes são frequentemente apontadas como a camada que conecta os dados genômicos às características fenotípicas (CARTER; HOFREE; IDEKER, 2013).

2.3.3.1 Ortológos e parálogos

Para Gabaldón & Koonin (2013), ortólogos e parálogos são tipos de genes homólogos relacionados por especiação ou duplicação, respectivamente. A conjectura ortologia-função indica que os genes ortólogos mantêm funções idênticas, ou biologicamente equivalentes, em diferentes organismos, além de compartilharem outras propriedades-chave. Por outro lado, a conjectura da ortologia avançada sugere que os parálogos podem sofrer diversificação funcional.

No entanto, os mesmos autores apresentam uma série de falsas implicações decorrentes das conjecturas apresentadas, sugerindo que as mesmas não são verdades absolutas.

Entre as falsas implicações citadas por Gabaldón & Koonin (2013), destacam-se: i) a ortologia não implica uma relação 1:1 entre genes de diferentes organismos. As duplicações gênicas específicas de linhagem geralmente levam a relações de co-ortologia de 1:N e de N:N; ii) ortologia não implica necessariamente que os genes ortólogos sejam as sequências ou estruturas mais semelhantes nos genomas comparados; iii) os genes que são mais similares entre si nos genomas comparados podem não ser ortólogos; iv) a ortologia não implica necessariamente a conservação da função do gene; v) genes com funções equivalentes não são necessariamente ortólogos; vi) apesar de todas as ressalvas, a conjectura generalizada da ortologia presume que, como uma tendência estatística em todos os genomas, os ortólogos são os genes mais semelhantes em diferentes espécies, em termos de sequência, estrutura e função.

Constata-se que as questões apresentadas são altamente relevantes para análises de ortologia e paralogia. Devido à inexatidão de conceitos no âmbito biológico, nota-se que diferentes métodos e softwares para identificação de ortólogos, como ProteinOrtho (LECHNER *et al.*, 2011) ou OrthoMCL (FISCHER *et al.*, 2011), possivelmente podem gerar resultados diferentes. Para Baldauf (2003), o mapeamento dos genes ortólogos entre organismos permite inúmeras aplicações, dentre as quais destacam-se a transferência de anotação funcional e a reconstrução da evolução das espécies, cujas relações ortólogas entre os genes é obviamente indispensável. As filogenias de espécies visam representar o curso dos eventos de especiação e, portanto, presume-se que somente as relações entre os genes ortólogos sirvam a esse propósito.

2.3.3.2 Análises filogenéticas e evolutivas

Análises filogenéticas mais amplas podem utilizar conjuntos grandes de genes ortólogos ou até mesmo proteomas completos. Por exemplo, BUSCO disponibiliza diversos conjuntos de genes de cópia única quase universais, representando conjuntos predefinidos de marcadores confiáveis para diversos reinos, classes e ordens. Esses conjuntos de dados podem ser utilizados para aplicações em análises genômicas comparativas, metagenômica e filogenômica (WATERHOUSE *et al.*, 2018). A partir do mapeamento de genes ortólogos, esses marcadores foram utilizados com sucesso em vários estudos, incluindo os gêneros *Penicillium* e *Aspergillus*, sendo considerados marcadores de grande utilidade para a inferência de relacionamentos evolutivos (STEENWYK *et al.*, 2019).

De forma ainda mais extensiva, é possível fazer uma análise de todo o proteoma de um determinado organismo, frente a outros proteomas disponíveis em bancos de dados. Como é o caso da ferramenta AAI-profiler (MEDLAR; TÖRÖNEN; HOLM, 2018), que realiza uma análise ampla do proteoma por pesquisas de homologia contra proteomas do UniProtKB (CONSORTIUM, 2019). Essa comparação de todas as sequências proteicas é considerada uma maneira rápida de obter uma visão geral das relações taxonômicas e filogenéticas entre espécies.

No entanto, a pressão de seleção costuma atuar de maneira mais direcionada, favorecendo ou desfavorecendo conjuntos de genes específicos em diferentes reinos, classes e ordens. Dessa forma, é importante analisar fatores como o ambiente de crescimento, a fisiologia e o metabolismo dos fungos filamentosos, facilitando assim a identificação de conjuntos de marcadores ideais para compreensão dos relacionamentos evolutivos entre espécies próximas.

A versatilidade biológica do gênero *Penicillium* permite com que esses microrganismos ocorram em todo o mundo, crescendo em ambientes variados, cuja classificação ocorre de acordo com a sua nutrição: saprófitos, parasitas e simbiontes (MUELLER; BILLS, 2004). Essa adaptabilidade às condições ambientais revela elementos estruturais, fisiológicos e metabólicos altamente sujeitos à mudanças, como os elementos que compõem a parede celular e os diferentes tipos de enzimas que o fungo necessita para degradar as macromoléculas disponíveis para sua nutrição.

Muitos dos elementos da parede celular são conservados em diferentes espécies de fungos, enquanto outros componentes são específicos da espécie. No geral, a parede celular é potencialmente a parte da célula que exibe mais diversidade e plasticidade fenotípica (ADAMS, 2004). Além disso, em fungos filamentosos, a parede celular é uma estrutura altamente dinâmica, sujeita a mudanças constantes, por exemplo, durante a germinação de esporos, ramificação das hifas e formação do septo. Proteínas e glucanos que compõem a parede celular, específicos para cada tipo de célula, são gerados em cada estágio do ciclo de vida dos fungos. Durante o ciclo, a parede celular pode ser drasticamente alterada, de acordo com o tipo de célula gerada. A composição da parede celular também é altamente regulada em resposta ao estresse e às condições ambientais, influenciando a ecologia dos fungos (GOW; LATGE; MUNRO, 2017).

A parede celular dos fungos compreende estruturas dinâmicas que desempenham um papel crítico na sobrevivência, crescimento e morfologia. Seus constituintes principais são quitina, quitosana, β-1,3-glucano, β-1,6-glucano, β-1,3-/β-1,4-glucano, α-1,3-glucano, melanina e glicoproteínas. Em geral, a parede celular fúngica é gerada pela reticulação de glucanos, quitina e outras proteínas da parede celular criando uma matriz tridimensional (FREE, 2013).

Os pigmentos de melanina, por sua vez, são formados por polimerização oxidativa de compostos fenólicos. Esses polímeros amorfos de alto peso molecular são amplamente encontrados em bactérias, fungos, plantas e animais. Os grânulos de melanina estão localizados na parede celular, onde possivelmente estão reticulados aos polissacarídeos. A melanina fornece defesa contra tensões ambientais, como luz ultravioleta, agentes oxidantes e radiação ionizante. Além disso, contribui para a capacidade dos fungos sobreviverem em ambientes agressivos (EISENMAN; CASADEVALL, 2012). A melanina fúngica ainda está associada com a virulência em uma série de fungos patogênicos, incluindo *Aspergillus fumigatus* e *Talaromyces marneffei* (LANGFELDER *et al.*, 2003).

Os fungos podem produzir melanina por vias distintas: a eumelanina pelas vias DHN (1,8-di-hidroxinaftaleno) e DOPA (L-3,4-di-hidroxifenilalanina) e as piomelaninas pela via

de degradação da l-tirosina. Acredita-se que o polímero escuro DHN-melanina seja a via biossintética da melanina fúngica mais bem caracterizada. Os homólogos da biossíntese de melanina dessas três vias foram caracterizados em vários fungos filamentosos. No entanto, a caracterização química da melanina pode ser uma tarefa desafiadora, pois o pigmento é altamente heterogêneo, insolúvel em solventes orgânicos, hidrofóbico e resistente à degradação química (PRALEA *et al.*, 2019).

Além da composição da parede celular fúngica, outro conjunto importante de marcadores evolutivos para fungos filamentosos compreende o repertório de enzimas (CAZymes) utilizadas para degradação de macromoléculas vegetais disponíveis no ambiente de crescimento (NYGAARD *et al.*, 2016). O nicho ecológico direciona a produção de enzimas, assim o fungo sob pressão de seleção é obrigado a se especializar para utilizar os tipos de fontes de carbono disponíveis no ambiente para sua nutrição (STAJICH, 2017). Recentemente, foi sugerido que diferentes nichos ecológicos e a disponibilidade dos tipos de fontes de carbono podem motivar a expansão e/ou contração nas famílias de enzimas necessárias para a rápida extração de carboidratos da matéria vegetal (NYGAARD *et al.*, 2016; STAJICH, 2017).

Os genes que codificam os componentes da parede celular, assim como os genes que compõem os sistemas enzimáticos, podem ser muito úteis como marcadores evolutivos. A partir desses marcadores é possível efetuar análises genômicas comparativas e evolutivas com outras espécies de ascomicetos, especialmente espécies de *Penicillium* que ocorrem em diferentes nichos ecológicos. Esse tipo de análise torna-se útil para compreensão da evolução de espécies próximas de *Penicillium*, frente aos diferentes estilos de vida, assim como frente às diferentes fontes de nutrição.

2.3.3.3 Rede regulatória de genes

A compreensão das relações entre os genes e os produtos que eles codificam continua sendo um dos principais desafios da biologia experimental e computacional. A identificação de relações regulatórias entre reguladores da transcrição e seus alvos é essencial para a compreensão de fenômenos biológicos que variam do crescimento e divisão celular à diferenciação e desenvolvimento celular (JACKSON *et al.*, 2020).

Em decorrência da importância das interações regulatórias, surgiu a necessidade de organizar esses dados em forma de uma GRN (do inglês *gene regulatory network*), a qual compreende um grafo direcionado que inclui os nós dos reguladores da expressão gênica, conectados aos nós dos genes-alvo pelas arestas de interação (KARLEBACH; SHAMIR, 2008). Os reguladores da expressão gênica incluem os TFs, que podem atuar como ativadores e repressores da expressão de genes-alvo (TGs), os reguladores de genes também incluem proteínas de ligação a RNA e RNAs reguladores. Esse tipo de rede aborda um desafio-chave na biologia experimental e computacional, ajudando a esclarecer as relações entre os genes e os produtos que eles codificam (JACKSON *et al.*, 2020).

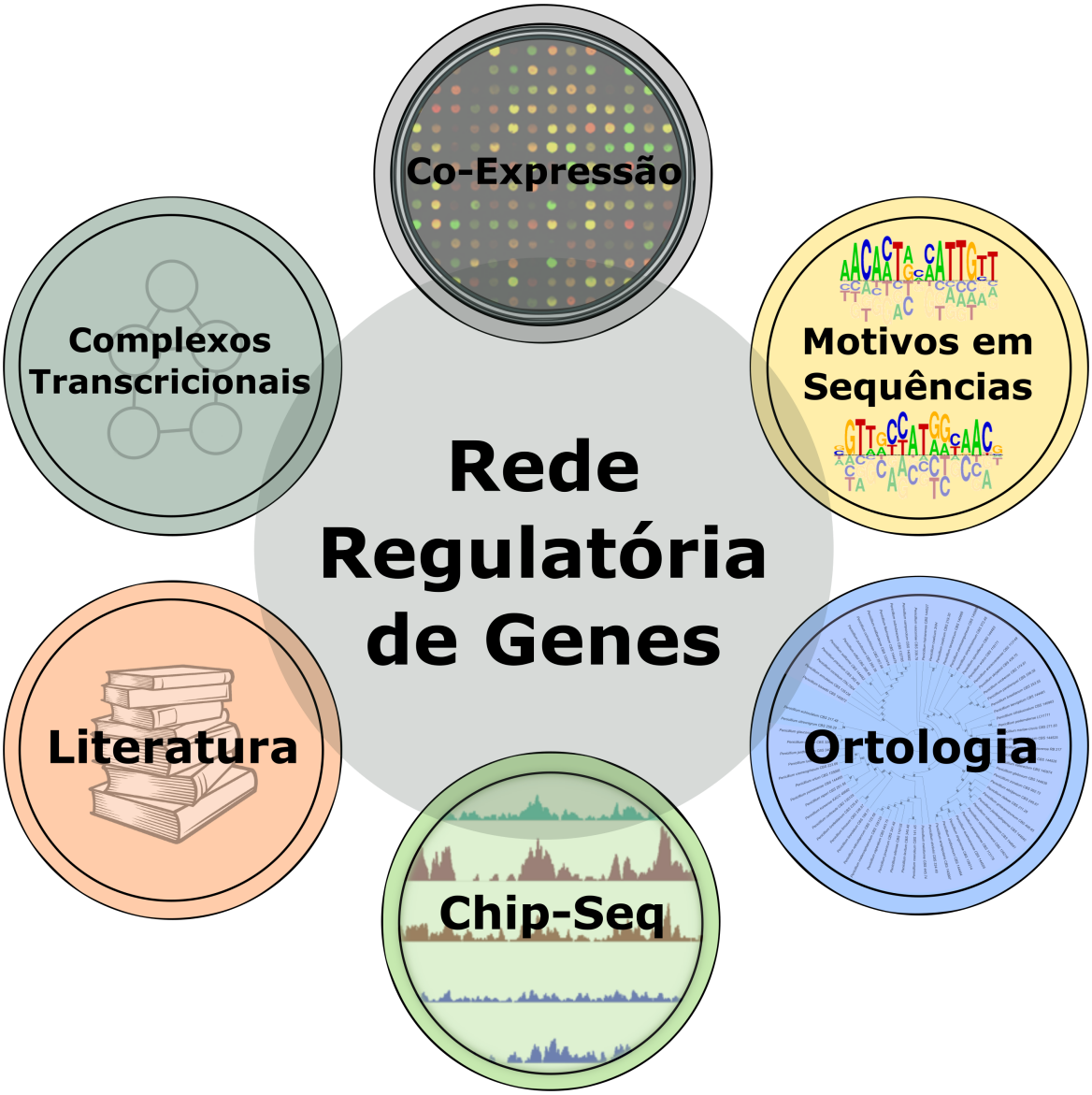
As GRNs combinadas com a descoberta de conhecimento têm um grande potencial para melhorar a interpretação dos dados ômicos, permitindo descobrir como a regulação da transcrição pode controlar processos biológicos e fenótipos (HASSANI-PAK; RAWLINGS, 2017). No escopo dos fungos filamentosos, atualmente apenas *Aspergillus nidulans* FGSC A4 e *Neurospora crassa* OR74A possuem estudos aprofundados para reconstrução de GRNs (HU; QIN; LIU, 2018), enquanto que, para o gênero *Penicillium*, não foram encontradas GRNs globais. Outro microrganismo amplamente estudado que possui uma GRN curada é *Saccharomyces cerevisiae* S288C, para a qual foi realizada a construção da rede global de regulação de genes compreendendo 12.228 interações (JACKSON *et al.*, 2020). A construção dessa GRN foi facilitada pela existência do banco de dados YEASTRACT+, que reune informações de interações para este microrganismo (MONTEIRO *et al.*, 2020).

Os recursos que podem ser utilizados para inferência de uma GRN podem ser classificados em seis classes (Fig. 5), de acordo com as abordagens empregadas e os dados subjacentes: Co-expressão, Motivos em Sequências, Imunoprecipitação de Cromatina (ChIP-Seq), Ortologia, Literatura e Interação Proteína-Proteína, especificamente focados em complexos transcricionais. Quanto mais informações são agregadas, mais precisa a relação TF-TG se torna. Em particular, as GRNs podem se beneficiar do conhecimento baseado em ortologia de espécies intimamente relacionadas; na qual a premissa principal é que uma relação TF-TG comprovada em um organismo possa ser conservada em outro (MERCATELLI *et al.*, 2020). Essa transferência de conhecimento, no entanto, requer métodos confiáveis para definir a ortologia entre genes diferentes para qualquer par TF-TG, além de levar em consideração o posicionamento filogenético das espécies analisadas (FERNANDEZ-VALVERDE; AGUILERA; RAMOS-DÍAZ, 2018).

Conforme citado anteriormente, a conjectura generalizada da ortologia presume que, como uma tendência estatística em genomas, os ortólogos são os genes mais semelhantes em diferentes espécies, em termos de sequência, estrutura e função (GABALDÓN; KOONIN, 2013). A detecção de ortologia é especialmente importante para maximizar o conteúdo e a precisão das informações. Portanto, uma premissa fundamental para construir uma complexa rede TF-TG baseada em ortologia é a presença de genes em diferentes espécies que podem ser atribuídos a um ancestral comum e, também assumindo que a ortologia seja funcional e não apenas baseada em similaridade de sequências (MERCATELLI *et al.*, 2020).

Outro recurso amplamente utilizado para inferência de GRNs compreende motivos conservados em sequências de DNA. Este recurso compreende a identificação de motivos conservados, localizados na região reguladora de genes, reconhecidos por TFs. Estes motivos são conhecidos como TFBSs. Dessa forma, a disponibilidade das sequências genômicas fornece informações valiosas para corroborar as relações regulatórias inferidas por ortologia, elevando assim a acurácia da GRN. Consequentemente, torna-se fundamental a compreensão dos conceitos de transcrição e regulação gênica, propiciando a aquisição de informações essenciais para a construção de uma GRN a partir de biologia computacional.

Figura 5 – Ilustração dos recursos que podem ser utilizados para inferência de uma GRN.



Fonte: Adaptada de Mercatelli *et al.* (2020).

2.3.3.3.1 Transcrição e regulação gênica

As sequências de transcrição tem um papel chave na explicação de diferenças entre as espécies. Dessa maneira, os estudos dos sistemas de transcrição e de sua regulação auxiliam na concepção do conhecimento em relação à funcionalidade dos genes em diferentes espécies, na diferenciação celular em organismos multicelulares, na resposta celular diante das alterações ambientais, entre outros tópicos. Em cada organismo, a expressão de genes assim como a regulação da expressão gênica constituem parte importante ao longo do ciclo de vida. Embora grande parte do processo de regulação pôde ser descoberto nas últimas décadas, ainda existem várias partes desse processo cujas funções são desconhecidas, permanecendo como objetos atuais de pesquisa (ROY; SINGER, 2015).

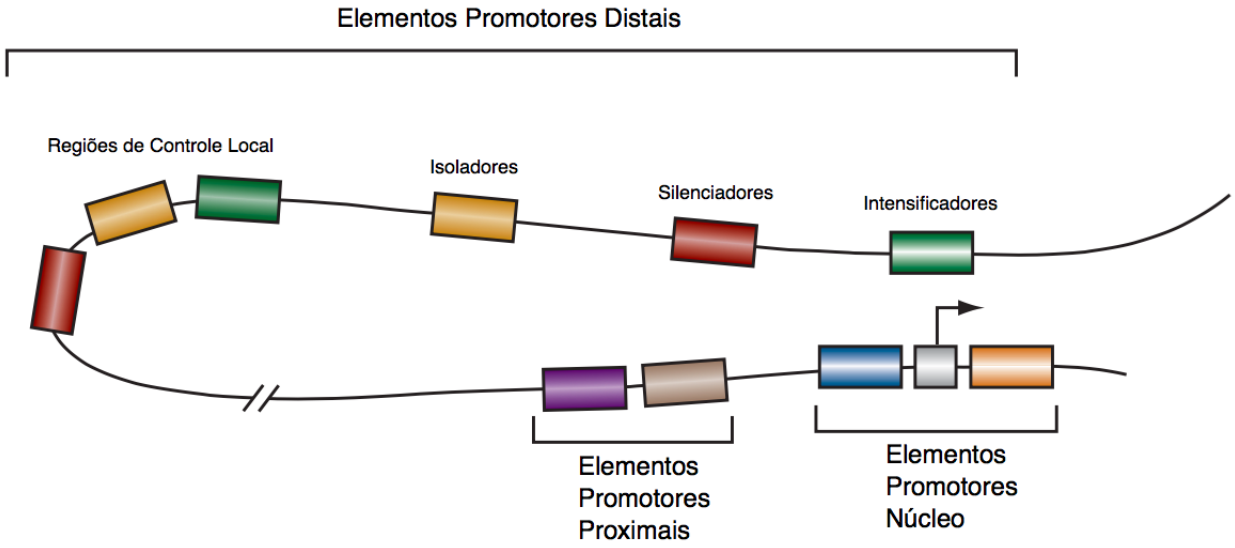
Os promotores referem-se às regiões de controle localizadas nas proximidades (tanto *upstream* quanto *downstream*) dos sítios de início da transcrição (TSSs), que são a base de início da transcrição e são responsáveis por conduzir a ligação da RNA polimerase (SANDELIN *et al.*, 2007; PAL; GUPTA; DAVULURI, 2014). Essas sequências podem ser bastante variáveis, porém, mantêm conservadas regiões responsáveis pela função promotora. No entanto, foi observado que nem todos os promotores estão associados com genes que codificam proteínas e, além disso, nem todos os eventos de transcrição começam no início de um gene. TSSs também foram observados no início de éxons internos e na extremidade 3' de alguns genes (ABEEL; PEER; SAEYS, 2009).

De acordo com Kristiansson et al. (2009), o tamanho da região promotora costuma ser variável para muitas categorias de genes, sendo que os promotores de alguns genes podem ultrapassar os 2000 pb no genoma de *S. cerevisiae*. O comprimento médio dos promotores para este microrganismo é de 455 pb e aproximadamente 5% dos promotores são maiores que 2000 bp. Além disso, foi observado que promotores longos geralmente estão relacionados com a capacidade de resposta a uma variedade de condições de estresse, fenômeno que também foi observado em *Saccharomyces pombe* e *Arabidopsis thaliana*. Dessa forma, promotores relativamente longos permitem a ligação de uma grande variedade de TFs enquanto que, inversamente, outros promotores podem ter sido encurtados para reduzir o tamanho do genoma. Isto posto, os autores sugerem que promotores mais longos de genes regulados por estresse sejam um fenômeno conservado em eucariotos.

A região promotora em si é tipicamente dividida em três partes em organismos eucariotos, conforme a Figura 6: i) o promotor-núcleo, que é a região responsável pela ativação real do sistema de transcrição e que compreende a região entre aproximadamente 40 pb *upstream* e aproximadamente 50 pb *downstream*, incluindo o TSS ; ii) o promotor proximal, que é uma região que contém vários elementos reguladores, variando até algumas centenas de pares de bases *upstream* do TSS; e iii) o promotor distal, que é uma região a partir de aproximadamente 200 pb *upstream* do TSS e contém elementos reguladores adicionais chamados intensificadores e silenciadores que influenciam fortemente os demais elementos promotores (ABEEL *et al.*, 2008). O promotor-núcleo e os elementos proximais, tipicamente, se estendem por menos de 1 kb de

pares de bases.

Figura 6 – Representação esquemática da região promotora de um gene.



Fonte: Maston, Evans e Green (2006).

Três principais mecanismos regulatórios controlam a transcrição de DNA para mRNA em células eucarióticas: cis-reguladores, trans-reguladores e a regulação por interferência de RNA, que é um mecanismo baseado em pequenos fragmentos de RNA de interferência (siRNA). Os mecanismos de regulação gênica são baseados em sequências de DNA adjacentes (cis) ou separadas (trans) dos genes que eles regulam (HALL, 2011).

Entre as sequências reguladoras encontradas nas regiões promotoras proximais e distais, são destacados os elementos cis-reguladores. Os cis-reguladores permitem a ligação dos TFs ao DNA, orquestrando o início da transcrição. Os elementos cis-reguladores são motivos curtos conservados que contém de 5 até 20 nucleotídeos que representam pontos de controle importantes na regulação da expressão do gene (ROMBAUTS *et al.*, 2003).

Os elementos promotores distais (upstream), que podem incluir intensificadores, silenciadores, isoladores, e regiões de controle local, podem estar localizados até 1 Mb pb do promotor proximal. Estes elementos distais podem entrar em contato com o promotor-núcleo ou com o promotor proximal através de um mecanismo de co-regulação, formando um sistema complexo de regulação. A presença de vários elementos reguladores dentro dos promotores confere o controle combinatório da regulação, aumentando exponencialmente o número de padrões regulatórios em potencial para expressão de genes. O maior desafio compreende o entendimento de como diferentes combinações dos mesmos elementos reguladores alteram a expressão gênica (MASTON; EVANS; GREEN, 2006).

Tanto os elementos proximais quanto distais são também chamados de sítios de ligação dos fatores de transcrição (TFBSs). Uma série de estudos examinaram a evolução desses

elementos cis-reguladores comparando com as regiões reguladoras de ortólogos. Foi demonstrado de forma consistente que estes elementos evoluíram a uma velocidade mais lenta do que o DNA não funcional que os rodeia. Elementos cis-reguladores podem ser conservados em espécies mais distantemente relacionadas, mesmo quando as regiões reguladoras ortólogas são divergentes para serem alinhadas de forma precisa (GASCH *et al.*, 2004).

Muita atenção tem sido dada ao investigar a estrutura modular de regiões promotoras que controlam a transcrição de genes eucarióticos. A imprecisão de um sítio de ligação pode ser compensada por uma maior aptidão dos sítios de ligação adjacentes, permitindo assim o posicionamento dos TFs adicionais graças às interações específicas entre as proteínas que compõem o sistema transcricional. Dessa forma, os promotores podem ser descritos como o resultado de uma hierarquia modular, na qual os elementos cis-reguladores individuais constituem o nível mais baixo; em seguida, a expressão dos elementos trans-reguladores e dos siRNAs formam o complexo transcricional que ativa a transcrição e confere a expressão específica de cada gene. A consequência disto é que cada promotor é único e regula especificamente o nível de transcrição do seu gene localizado *downstream*. Essa complexidade têm grandes repercussões sobre a identificação *in silico* de TFBSs (ROMBAUTS *et al.*, 2003).

Em um segundo nível estão os TFs que podem estar localizados em qualquer parte do genoma, inclusive em diferentes cromossomos dos genes que eles regulam, sendo conhecidos como elementos trans-reguladores por esse motivo (HAAS; HIMMELBACH; MASCHER, 2020). Os TFs são proteínas cuja estrutura molecular contêm um ou mais domínios especiais que conferem a capacidade de ligação ao DNA das regiões promotoras de seus genes-alvo, de forma direta ou indireta, possibilitando assim o aumento ou a diminuição da taxa de transcrição desses genes-alvo. É importante salientar que os próprios TFs são regulados por outros TFs ou até mesmo por uma auto-regulação. A taxa de transcrição de um gene específico em uma situação particular é, portanto, determinada pelo equilíbrio entre fatores de regulação positiva e negativa que podem se ligar às suas regiões promotoras e que estão presentes de forma ativa nesta situação particular (LATCHMAN, 2013).

A caracterização dos domínios que ocorrem nas proteínas de um determinado organismo, com base em bancos de dados de domínios como InterPro (MITCHELL *et al.*, 2019) ou Pfam (EL-GEBALI *et al.*, 2019), permite um amplo mapeamento de TFs. A distribuição de TFs em fungos compreende uma vasta quantidade de domínios de ligação ao DNA, incluindo *homeobox*, *forkhead*, *sox*, *helix-loop-helix*, *zinc fingers*, entre outros (TODD *et al.*, 2014).

Em organismos eucariotos, a maioria dos TFs está presente nas células apenas em pequenas quantidades necessárias para iniciar ou silenciar a expressão do gene, e, em muitos casos, eles são induzidos pelas condições para as quais são necessários e são degradados assim que não são mais necessários. Nos fungos, os TFs se ligam aos cis-reguladores das regiões promotoras, geralmente localizados de 1 a 1500 pb *upstream* dos genes-alvos de regulação (HU *et al.*, 2013).

A descoberta *in silico* de novas interações de regulação entre TFs e seus genes-alvo, compreende a caracterização de domínios de ligação ao DNA, localizados nos TFs, e de seus respectivos sítios de ligação ao DNA (TFBSs), localizados nas regiões promotoras dos genes-alvo. Assim, um TFBS pode ser representado sob a forma de uma matriz de frequência de posição (PFM) ou uma matriz de peso de posição (PWM) (HU *et al.*, 2013). Estudos recentes caracterizam a especificidade de TFs em eucariotos, contemplando a representação de diversos TFBSs em forma de matrizes para microrganismos modelo como *S. cerevisiae*, *N. crassa* e *A. nidulans* (LAMBERT *et al.*, 2019).

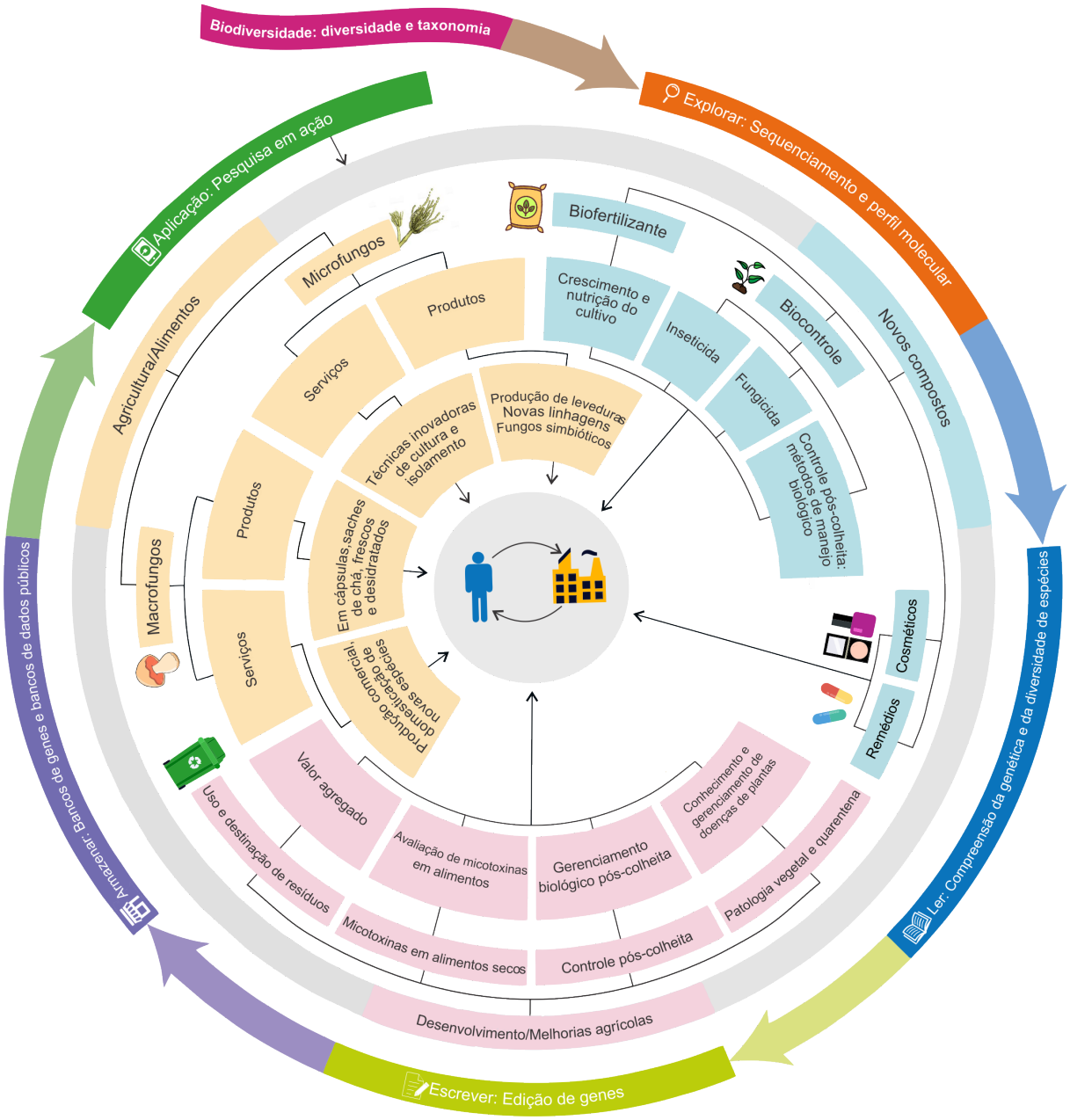
Um dos meios de descoberta de TFBSs se concentra em informações biológicas, como a conservação evolucionária desses motivos curtos. Essa estratégia baseia-se no pressuposto de que as regiões não codificadoras entre as espécies relacionadas provavelmente estão sob força de seleção negativa e, portanto, contêm motivos funcionais conservados. Devido ao rápido aumento de sequências genômicas disponíveis, esse método de busca por TFBSs está se expandindo, principalmente por gerar uma menor taxa de falsos-positivos em comparação a outros métodos (HU *et al.*, 2013). Nota-se que um mapeamento amplo de TFBSs de um organismo compreende informações altamente significativas sobre as interações de regulação TF-TG, contribuindo para acurácia das redes de regulação gênica.

2.4 Aplicações biotecnológicas

Embora os fungos apresentem numerosos usos em potencial e aplicações consolidadas de extrema importância para humanidade, estes notáveis organismos continuam sendo pouco explorados, devido aos recursos escassos para financiamentos de pesquisas relacionadas. Os fungos forneceram ao mundo a penicilina, a lovastatina e outros medicamentos de importância global, sendo largamente empregados em estratégias contra doenças humanas e doenças de plantas. Os fungos também são utilizados para melhoria de cultivos agrícolas e produção de *commodities*, para produção de alimentos e bebidas e para diversas estratégias sustentáveis que incluem: produção de biocombustíveis, biorremediação e degradação de materiais sintéticos. O amplo potencial biotecnológico e industrial desse reino é destacado na Figura 7. O ciclo começa com a pesquisa básica em biodiversidade que, por sua vez, leva ao depósito de culturas em coleções públicas. As culturas são então usadas para pesquisa aplicada que, por sua vez, leva a produtos biotecnológicos (HYDE *et al.*, 2019).

A síntese de misturas enzimáticas relevantes para a indústria, a partir de fungos filamentosos, tem sido uma realidade há mais de 100 anos. No entanto, a maior expansão ocorreu na última década, com o advento da genômica e da pós-genômica. Por exemplo, os fungos *Aspergillus niger* (CAIRNS; NAI; MEYER, 2018) e *Trichoderma reesei* (BISCHOF; RAMONI; SEIBOTH, 2016; GUPTA *et al.*, 2016) têm sido manipulados como biofábricas para produção de diversas moléculas de interesse biotecnológico, principalmente enzimas individuais e sistemas

Figura 7 – Diagrama de uso biotecnológico potencial dos fungos.



Fonte: Hyde *et al.* (2019).

enzimáticos.

O potencial de exploração dos fungos filamentosos como biofábricas para produção de moléculas de interesse biotecnológico está se expandindo bastante, graças aos avanços recentes no conhecimento biológico acerca desses microrganismos e ao desenvolvimento concomitante da bioinformática, das técnicas de cultivo e das ferramentas moleculares. Os fungos filamentosos podem ser vistos como um chassi para produtos que não podem ser produzidos em sistemas bacterianos mais simples, eles são capazes de produzir não apenas proteínas e enzimas em

altas concentrações, mas também produtos farmacêuticos, na maioria das vezes metabólitos secundários expressos por *clusters* de genes, que são benéficos para a saúde humana e animal. Possivelmente, os fungos filamentosos ocupem um papel como principais organismos envolvidos na próxima revolução industrial, compreendendo a mudança de uma economia baseada em uma matriz predominantemente fóssil para uma bioeconomia (CAIRNS; NAI; MEYER, 2018).

Uma aplicação crítica desses microrganismos são os coquetéis enzimáticos responsáveis pela degradação de polissacarídeos vegetais, em que a celulose, a hemicelulose e a pectina podem ser decompostas em oligossacarídeos e monossacarídeos (CAIRNS; NAI; MEYER, 2018). A degradação enzimática de polissacarídeos vegetais de fungos é notável por sua relevância, devido ao seu amplo uso em diversas aplicações industriais, como papel, alimentos, ração animal, produtos químicos e biocombustíveis (HYDE *et al.*, 2019).

Os fungos filamentosos dos gêneros (*Aspergillus*, *Rhizopus*, *Trichoderma*, *Neurospora*, *Penicillium* etc.) são organismos naturalmente especializados na desconstrução da biomassa lignocelulósica e esse recurso representa um enorme potencial para a produção de biocombustíveis a partir de fontes renováveis (DALENA *et al.*, 2019). A exploração desse potencial exige um profundo entendimento da fisiologia celular fúngica e o desenvolvimento de ferramentas de alto desempenho adequadas para as aplicações biotecnológicas pretendidas (GUPTA *et al.*, 2016).

2.4.1 Enzimas degradadoras de biomassa vegetal

A biomassa vegetal é basicamente composta por lignina, celulose e hemicelulose, esta última, composta principalmente pelos seguintes polissacarídeos: xilana, glicurono-xilana, xiloglicano, glicomanana e arabinoxilana com cadeias heterogêneas. O uso de monossacarídeos que constituem esses heteropolímeros necessita uma hidrólise eficiente, que ainda é um grande desafio técnico devido à sua recalcitrância e heterogeneidade (DRUZHININA; KUBICEK, 2017).

A degradação de matéria vegetal em açúcares monoméricos tem grande importância, uma vez que os açúcares fermentáveis podem ser utilizados como matérias-primas em inúmeros processos biotecnológicos, desde a indústria alimentícia até a produção do etanol 2G. Os fungos são os microrganismos que apresentam a maior capacidade degradadora dos principais heteropolímeros que compõem a biomassa vegetal: celulose, hemicelulose e lignina (PÉREZ *et al.*, 2002).

A celulose, principal componente das fibras vegetais, é um homopolissacarídeo composto por unidades de β-glicopiranose (D-glicose) as quais são conectadas por ligações β-(1 -> 4). Tomando a celobiose como a unidade de base, a celulose pode ser considerada um polímero isotático de celobiose. As cadeias de celulose são embaladas em microfibrilas que são estabilizadas por ligações de hidrogênio. Estas fibrilas são ligadas umas às outras por hemiceluloses e cobertas por lignina (BRODEUR *et al.*, 2011).

Os componentes hemicelulósicos da parede celular vegetal tem um peso molecular mais baixo do que a celulose, esses componentes incluem uma grande variedade de heteropolissacarídeos com estruturas lineares e/ou ramificadas e geralmente são classificados de acordo com o principal açúcar presente na sua composição (GLASS *et al.*, 2013). Entre os polímeros constituintes estão as pentoses (xilose, ramnose, arabinose), as hexoses (manose, glicose, galactose) e os açúcares ácidos (ácidos urônicos) e, ao contrário da celulose, a hemicelulose é composta por polímeros facilmente hidrolisáveis. As xilanas são as hemiceluloses mais abundantes e se constituem de heteropolissacarídeos, localizadas entre as moléculas de lignina e o conjunto de fibras de celulose (PÉREZ *et al.*, 2002; BRODEUR *et al.*, 2011).

Já a lignina é um heteropolímero reticulado e amorfo que proporciona rigidez e impermeabilidade à parede da célula vegetal, além de resistência contra ataque microbiano. As microfibrilas de celulose, que estão presentes no centro da hemicelulose e cobertas por lignina, são muitas vezes associadas sob a forma de feixes ou macrofibrilas. Naturalmente, a estrutura destas fibrilas de celulose é cristalina e altamente resistente à ação de enzimas (PÉREZ *et al.*, 2002; BRODEUR *et al.*, 2011).

Uma vez que os substratos são insolúveis, a degradação de biomassa vegetal deve ocorrer fora do micélio fúngico. Dessa forma, os fungos possuem dois tipos de sistemas enzimáticos extracelulares: um sistema hidrolítico que produz hidrolases e é responsável pela degradação da celulose e hemicelulose; e um sistema lignolítico que produz oxidases que despolimerizam a lignina (PÉREZ *et al.*, 2002). Mais recentemente, também foi identificado um grupo de enzimas não hidrolíticas que incluem algumas enzimas com atividades auxiliares. Essas enzimas atuam em um processo sinérgico na clivagem redutiva oxidante da cadeia de celulose (RYTIOJA *et al.*, 2014). As enzimas de atividade auxiliar ajudam a reduzir a dosagem enzimática necessária para a degradação da biomassa e, portanto, tornaram-se enzimas importantes encontradas em formulações comerciais recentes de celulases (HU *et al.*, 2018).

As enzimas que degradam, modificam ou criam ligações glicosídicas são classificadas como CAZymes. Essas enzimas são organizadas em diferentes famílias, de acordo com a sua sequência de aminoácidos e similaridade estrutural. O banco de dados CAZy (http://www.cazy.org) fornece acesso on-line e continuamente atualizado para CAZymes. A utilização deste banco de dados como referência é muito importante para anotação funcional desse conjunto de enzimas em genomas, principalmente para anotação automatizada ou para não-especialistas. Os *pipelines* de anotação funcional automatizada acabam por acumular e propagar os erros em bases de dados públicas. Esses erros ocorrem principalmente devido à variação da modularidade destas enzimas, bem como pelo fato de agrupamentos de enzimas com diferentes especificidades de substrato pertencerem a uma mesma família (LOMBARD *et al.*, 2014).

As seguintes famílias compõem o banco de dados de enzimas CAZy: Glicosil Hidrolases (GHs), responsáveis pela hidrólise e/ou rearranjo de ligações glicosídicas; Glicosil Transferases (GTs), responsáveis pela formação de ligações glicosídicas; Polissacarídeo Liases (PLs), realizam

a clivagem não-hidrolítica de ligações glicosídicas; Carboidrato Esterases (CEs), hidrolisam ésteres de carboidratos; Atividades Auxiliares (AAs), enzimas redox que atuam em conjunto com outras CAZymes; e Módulos de Ligação ao Carboidrato (CBMs) que promovem a adesão da enzima ao carboidrato (LOMBARD *et al.*, 2014).

O Quadro 1 apresenta uma visão geral das enzimas que atuam na degradação da biomassa lignocelulósica. As enzimas são divididas de acordo com os substratos onde atuam, seus números *Enzyme Commission* (EC), abreviaturas e famílias CAZy correspondentes (RYTIOJA *et al.*, 2014).

Quadro 1 – Visão geral das enzimas produzidas por fungos que atuam na degradação da biomassa vegetal.

Substrato	Enzima	EC	Abreviação	Família(s) CAZy
Celulose	Endo-1,4-β-glicanase	3.2.1.4	EGL	GH5,-7,-12,-45
	Celobiohidrolase (final redutor)	3.2.1.176	CBHI	GH7
	Celobiohidrolase (final não redutor)	3.2.1.91	CBHII	GH6
	β-glicosidase	3.2.1.21	BGL	GH1,-3
	Monooxigenase lítica de polissacarídeos	1.1.99.-	LPMO	AA9,-16
	Celobiose Desidrogenase	1.1.99.18	CDH	AA3_1,AA8
Xilana	Endo-1,4-β-xilanase	3.2.1.8	XLN	GH10,-11
	Xilobiohidrolase	3.2.1.-	XBH	
	Xilana 1,4-β-xilosidase	3.2.1.37	BXL	GH3,-43
Galactomanana	Manana Endo-1,4-β-mananase	3.2.1.78	MAN	GH5,-26
	β-manosidase	3.2.1.25	MND	GH2
	α-galactosidase	3.2.1.22	AGL	GH27,-36
	β-galactosidase	3.2.1.23	LAC	GH2,-35
	α-L-arabinofuranosidase	3.2.1.55	ABF	GH51,-54
Xiloglicano	Xiloglicano Endo-1,4-β-glicanase	3.2.1.151	XEG	GH12,-74
	α-L-arabinofuranosidase	3.2.1.55	ABF	GH51,-54
	α-D-xiloside xilohidrolase	3.2.1.177	AXL	GH31
	α-L-fucosidase	3.2.1.51	AFC	GH29,-95
	α-galactosidase	3.2.1.22	AGL	GH27,-36
	β-galactosidase	3.2.1.23	LAC	GH2,-35
Arabinoxilana	α-L-arabinofuranosidase	3.2.1.55	AXH	GH6
	α-glicuronidase	3.2.1.139	AGU	GH67,-115
	α-galactosidase	3.2.1.22	AGL	GH27,-36
	β-galactosidase	3.2.1.23	LAC	GH2,-35
	Acetilxilana esterase	3.1.1.72	AXE	CE1,-5
	Feruloil esterase	3.1.1.73	FAE	CE1

Fonte: Adaptada de Rytioja *et al.* (2014).

2.4.1.1 Sistema celulolítico

A degradação da celulose é mediada por um sistema enzimático celulolítico amplamente utilizado para a produção de biocombustíveis (PANCHAPAKESAN; SHANKAR, 2016). A

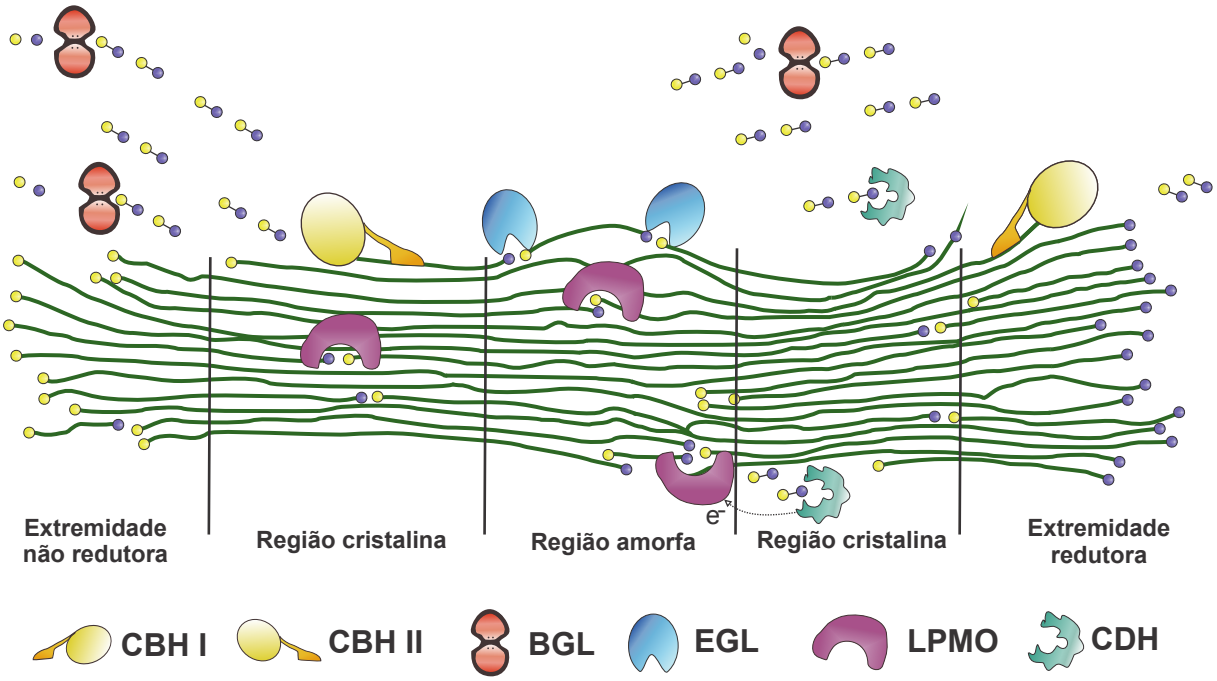
hidrólise da celulose necessita de uma mistura de enzimas celulolíticas, contendo enzimas que atuam sinergicamente (Figura 8): (i) as celobiohidrolases (CBHI) (GH7) atuam nas extremidades redutoras da cadeia de celulose; (ii) as celobiohidrolases (CBHII) (GH6) atuam nas extremidades não redutoras da cadeia de celulose; (iii) as endoglicanases (EGL) (GH5, GH7, GH12 e GH45) atuam nas regiões amorfas da celulose; (iv) as monooxigenases líticas de polissacarídeos (LPMO) (AA9 e AA16) podem atuar tanto em regiões cristalinas quanto amorfas; (v) as celobiose desidrogenases (CDH) (AA3-1 e AA8) atuam na oxidação da celobiose produzindo elétrons que auxiliam a despolimerização da celulose a ser catalisada pelas LPMOs; (vi) os oligossacarídeos são posteriormente hidrolisados em D-glicose por β-glicosidases (BGL) (GH3) (GLASS *et al.*, 2013; RYTIOJA *et al.*, 2014; DRUZHININA; KUBICEK, 2017; FILIATRAULT-CHASTEL *et al.*, 2019).

Além das GHs amplamente estudas, o sistema celulolítico inclui algumas enzimas com atividades auxiliares (AAs). Essas enzimas atuam em um processo sinérgico (CDH - LPMO) na clivagem oxidorredutora da cadeia de celulose (RYTIOJA *et al.*, 2014). As LPMOs são metaloenzimas dependentes de cobre que atuam na celulose empregando mecanismos oxidativos. Essas proteínas oxidativas introduzem cortes na superfície da celulose, fornecendo extremidades extras na cadeia para as GHs agirem, o que melhora a cinética da reação e também aumenta a sinergia enzimática (BERLEMONT, 2017; HU *et al.*, 2018; FILIATRAULT-CHASTEL *et al.*, 2019).

As CDHs, por sua vez, atuam em sinergia com as LPMOs. A CDH é o único flavocitocromo extracelular conhecido que oxida a celobiose e outras celodextrinas em suas lactonas correspondentes. Essa oxidação da celobiose também atua como um doador natural de elétrons, necessários para o processo de catálise da celulose pelas LPMOs. O papel exato das interações eletrostáticas das CDHs na degradação da lignocelulose ainda não está claro, embora haja evidências de sua relevância tanto na maquinaria celulolítica quanto na modificação da lignina (KRACHER; LUDWIG, 2016). O papel das CDHs na modificação da lignina, por sua vez, é apoiado pela capacidade de produzir radicais hidroxila através da reação de Fenton (RYTIOJA *et al.*, 2014). As enzimas auxiliares (CDH/LPMO) ajudam na redução da dosagem enzimática necessária para degradação da biomassa e, portanto, tornaram-se enzimas importantes encontradas em formulações comerciais recentes de celulases (HU *et al.*, 2018).

As enzimas celulolíticas são produzidas por diversos microrganismos, tais como bactérias, fungos basidiomicetos e fungos ascomicetos. Devido à sua excepcional capacidade de expressar e secretar proteínas, os fungos tornaram-se indispensáveis para a produção dessas enzimas. Enzimas nativas ou recombinantes são produzidas principalmente por *A. niger*, *T. reesei* e *Penicillium oxalicum* (PUNT *et al.*, 2002). Além da produtividade elevada de enzimas, os fungos são capazes de secretar as enzimas produzidas de forma eficiente devido ao seu sistema robusto de secreção, envolvendo retículo endoplasmático e complexo golgiense, enquanto que isso não é possível na maioria das bactérias (JOUZANI; TAHERZADEH, 2015).

Figura 8 – Ilustração do mecanismo de ação de enzimas que atuam na degradação da celulose.



Fonte: Adaptada de Wang *et al.* (2012).

Recentemente foi demonstrado que espécies do gênero *Penicillium* podem produzir misturas enzimáticas com níveis de atividade comparáveis aos coquetéis comerciais, geralmente produzindo misturas de enzimas degradadoras da biomassa vegetal mais equilibradas que outros gêneros, especialmente pela atividade de β-glicosidases (VAISHNAV *et al.*, 2018). Esses resultados podem ser explicados pela fonte de carbono preferencial, resultante da evolução de cada espécie e de sua respectiva capacidade de degradar substratos específicos. As preferências degradadoras de um fungo fornecem uma nova perspectiva em relação à composição do sistema de enzimas celulolíticas. A possibilidade de melhorar a produção de enzimas celulolíticas deve levar esse fato em consideração, direcionando a escolha de microrganismos naturalmente especializados em cada substrato de interesse biotecnológico. Dessa forma, o conhecimento das vias regulatórias pode permitir que a construção de novas linhagens amplifique a produção de celulases em espécies de *Penicillium* para competir com os produtores de enzimas comerciais *T. reesei* e *A. niger*.

2.4.1.2 Sistema hemicelulolítico

A hemicelulose é o segundo componente mais abundante na parede vegetal. Também é mais complexa que a celulose e sua degradação completa exige a degradação de todas as ramificações, no entanto as ligações entre os monômeros da hemicelulose são facilmente hidrolisáveis pelas enzimas. O que difere são os tipos de enzimas necessárias, necessitando de pelo menos nove diferentes enzimas para uma hidrólise completa. A hidrólise da hemicelulose ocorre por meio da ação sinérgica de endo-enzimas responsáveis pela clivagem interna da

cadeia principal, além de exo-enzimas que liberam açúcares monoméricos e outras enzimas que hidrolisam as cadeias laterais de polímeros ou oligossacarídeos, liberando assim diversos mono ou dissacarídeos dependendo do tipo de hemicelulose clivada (RYTIOJA *et al.*, 2014).

As hemicelulases são frequentemente classificadas de acordo com sua ação sobre substratos distintos, assim, destacam-se as xilanases que formam o sistema enzimático hidrolítico que compreende uma variedade de enzimas responsáveis pela degradação completa da xilana, o principal componente da hemicelulose (PÉREZ *et al.*, 2002). O mecanismo clássico de degradação da xilana por fungos inclui endo-1,4-βxilanases (GH10, GH11) que hidrolisam internamente a cadeia principal da xilana entre as ligações 1,4-β, β-xilosidases (GH43 e GH3) que liberam xilose da xilobiose e pequenas cadeias de xilo-oligossacarídeos, α-L-arabinofuranosidases (GH3, GH10, GH43, GH51, GH54, e GH62), acetilxilana esterases (CE1–CE7), feruloil esterases (CE1), dentre outras (GLASS *et al.*, 2013).

2.4.1.3 Indução da expressão do sistema celulolítico por celodextrinas

A compreensão das diferentes estratégias empregadas pelos fungos filamentosos para degradar a biomassa lignocelulósica é um fator fundamental para o aprimoramento da produção enzimática. Assim, a capacidade dos fungos de crescer, transportar e fermentar diferentes tipos de açúcar continua sendo um grande desafio para a produção de biocombustíveis a partir da biomassa. Desse modo, as análises de genomas, secretomas e transcriptomas de fungos filamentosos são destacadas por permitirem a compreensão da fisiologia e do metabolismo desses microrganismos (HYDE *et al.*, 2019).

A produção de enzimas (celulases, hemicelulases, ligninases e pectinases) degradadoras da parede celular é regulada principalmente no nível transcricional dos fungos filamentosos. A expressão gênica dessas enzimas é regulada por vários fatores ambientais e celulares, alguns dos quais são comuns enquanto outros são específicos de espécie ou classe enzimática. Esses genes são induzíveis na presença da fonte de carbono ou moléculas derivadas da fonte de carbono, enquanto a repressão ocorre em condições de crescimento em que a produção dessas enzimas não é necessária, como na presença de glicose (ARO; PAKULA; PENTTILÄ, 2005).

A celobiose é o principal produto final gerado pela degradação da celulose pelas celulases. Foi demonstrado que a produção de celulases é induzida pela celobiose em muitas espécies de fungos (ARO; PAKULA; PENTTILÄ, 2005), e resultados de pesquisas recentes suportam que o acúmulo de celodextrinas intracelulares (principalmente celobiose) pode aumentar a secreção de celulases por uma via de sinalização em cascata (YAO *et al.*, 2016; CAI *et al.*, 2015).

Em *T. reesei*, as β-glicosidases intracelulares (iBGLs) podem atuar na celobiose produzindo produtos de transglicosilação, que induzem a expressão de genes de celulases (SHIDA *et al.*, 2015). A transglicosilação é um mecanismo para formação de ligações glicosídicas durante a síntese de polissacarídeos. Por outro lado, um experimento de indução em *P. oxalicum* mostrou que o acúmulo de celobiose induz a expressão de enzimas lignocelulolíticas, porém a indução

não está relacionada à formação de produtos de transglicosilação a partir da celobiose (CHEN *et al.*, 2013).

Além disso, os fungos filamentosos são capazes de transportar dissacarídeos como a celobiose para o interior da célula através de transportadores específicos de celobiose. As celodextrinas podem atuar como transdutores de sinal de duas maneiras: i) as celodextrinas são transportadas para as células ativando sensores intracelulares e ii) as celodextrinas extracelulares ativam sensores da membrana plasmática, são proteínas similares a transportadores ou proteínas G acopladas a receptores de membrana (REIS *et al.*, 2016).

Resultados experimentais indicam a existência do primeiro mecanismo citado em *T. reesei* (ZHANG *et al.*, 2013) e *P. oxalicum* (LI *et al.*, 2013). O acúmulo de celobiose extracelular permite verificar a existência de sensores de celobiose na membrana plasmática. Em *P. oxalicum*, essa hipótese foi refutada pela exclusão da principal β-glicosidase *bgl1* extracelular, indicando que não há sensores correspondentes ao segundo mecanismo (CHEN *et al.*, 2013). Ambos os mecanismos foram identificados em *A. nidulans* (REIS *et al.*, 2016) e *N. crassa* (CAI *et al.*, 2015). No entanto, poucos transportadores de celodextrinas foram funcionalmente caracterizados em fungos filamentosos.

Considerando isso, a indução do sistema celulolítico pode ser proporcionada por perturbações ou alterações na expressão de iBGLs e transportadores de celodextrinas, resultando no acúmulo intracelular de celobiose. Dessa forma, o acúmulo de celobiose coordena a indução do sistema celulolítico por uma via de sinalização em cascata. Esse pressuposto fornece oportunidades para explorar mecanismos de expressão do sistema celulolítico, tanto em fungos comerciais estabelecidos como também em novas espécies. Esse conhecimento contribui para a concepção de novas linhagens hiperprodutoras de enzimas celulolíticas, colaborando para viabilizar a produção de biocombustíveis a partir de biomassa vegetal em larga escala.

2.4.2 Etanol 2G

Nos últimos anos, amplos esforços de pesquisa industrial e acadêmica foram feitos com intuito de reduzir custos e permitir a produção em larga escala de biocombustíveis como recursos renováveis. Considerando que a substituição de combustíveis fósseis por biocombustíveis tem o potencial de gerar numerosos benefícios (DALENA *et al.*, 2019). Dessa forma, torna-se necessário compreender de forma mais aprofundada o papel do etanol 2G diante da demanda energética global e nacional.

2.4.2.1 Consumo de energia

O consumo de energia tem aumentado progressivamente em âmbito global, tendo os combustíveis fósseis como principal recurso para atender às crescentes demandas. Além das fontes fósseis não serem renováveis, sabe-se que os combustíveis fósseis causam uma infinidade

de impactos negativos difíceis de serem mensurados.

Nos últimos anos, uma atenção considerável tem sido concentrada na redução dos custos de biocombustíveis, nas emissões de gases do efeito estufa (GEE), nas necessidades de recursos terrestres e hídricos, e na melhoria da compatibilidade entre sistemas de distribuição de combustível e mecanismos de veículos. Existem fortes evidências mostrando a expansão das mudanças climáticas em nível global. As causas estão ligadas às atividades antropogênicas para suprir a necessidade voraz por energia, cuja produção e consumo criaram efeitos perigosos iminentes. Os efeitos incluem o aumento de temperatura do planeta, precipitação irregular, derretimento de geleiras, elevação do nível do mar e clima extremo (ABDULLAH *et al.*, 2019).

O crescente consumo de combustíveis fósseis provoca impactos negativos através de processos industriais não-combustivos, poluição e degradação ambiental resultantes da extração, assim como poluição direta do ar devido à combustão. A combustão, por sua vez, tem aumentando a cada dia a liberação na atmosfera de GEE (SINGHANIA *et al.*, 2014; ANIL *et al.*, 2016). As informações mais atuais sobre a matriz energética mundial correspondem ao ano de 2018. Os números indicam que aproximadamente 85% da matriz energética advém de fontes de carbono fóssil, sendo 33,63% de petróleo, 27,21% de carvão mineral e 23,87% de gás natural. Enquanto que as fontes renováveis correspondem apenas a 4,05% da matriz energética global (BPSTATS, 2019).

O relatório da Agência Internacional de Energia Renovável (IRENA) (ANIL *et al.*, 2016) indica que a transição energética depende das energias renováveis modernas, as quais incluem vários tipos de bioenergia e biocombustíveis sustentáveis e de baixo custo. Em um cenário prospectivo, uma possível duplicação da quota de energias renováveis no âmbito energético global poderia resultar em economias significativas de combustíveis fósseis. As estimativas indicam que os suprimentos de carvão, petróleo e gás natural podem ser reduzidos em 22%, 11% e 11%, respectivamente, enquanto a oferta de energia renovável primária pode aumentar em até 46% nos próximos 10 anos. Como benefícios positivos, as emissões que resultam em sérios efeitos adversos sobre a saúde humana poderiam ser reduzidas em 33% e poderiam salvar até 4 milhões de vidas por ano até 2030.

A preocupação mundial com as consequências do avanço do aquecimento global e da elevação dos preços do petróleo resultaram no aumento da relevância atribuída aos biocombustíveis. Geralmente, os tipos de biomassa utilizados como matérias-primas dos biocombustíveis são gramíneas e plantas oleaginosas que possuem açúcares, amido ou substâncias em formas de óleos e gorduras que podem ser extraídas a partir de determinados processos. Entre os vegetais mais comumente empregados estão a cana-de-açúcar, o milho, a mamona, a palma, o girassol, o babaçu e a soja (PANAHI *et al.*, 2019).

A produção e o consumo de biocombustíveis apresentam diversas vantagens em relação aos combustíveis fósseis: menor índice de poluição durante a combustão e processamento, emitindo menos GEE; a matéria-prima pode ser cultivada ou reaproveitada, portanto, trata-se de

uma fonte renovável; geram empregos em sua cadeia produtiva; e ajudam a suprir a demanda global crescente. As energias renováveis podem suprir até dois terços da demanda total de energia global e contribuir com a maior parte das reduções de emissões de GEE que são necessárias entre 2019 e 2050 para limitar abaixo de 2 °C o aumento médio da temperatura do planeta (GIELEN *et al.*, 2019).

O etanol de primeira geração (1G) é conhecido por diversos problemas: (i) impacto negativo nos preços de alimentos e competição pela área de cultivo; (ii) ameaça à biodiversidade; (iii) balanço de carbono insatisfatório; e (iv) tecnologias de conversão relativamente ineficientes. O etanol 2G, por sua vez, utiliza resíduos de biomassa vegetal, uma matéria-prima de baixo custo e amplamente disponível na biosfera. Este processo elimina a competição entre alimentos e combustível e oferece um balanço de carbono notável (DUTTA; DAVEREY; LIN, 2014).

O etanol 2G pode ser produzido através da hidrólise enzimática de diferentes tipos de resíduos lignocelulósicos (SHEN *et al.*, 2012), os quais podem ser obtidos do bagaço de cana-de-açúcar (MAEDA *et al.*, 2011; RAMOS *et al.*, 2015), resíduos da produção do milho, madeira (SOUDHAM *et al.*, 2015), palha de arroz (RAN *et al.*, 2012), capim-elefante (CARDONA *et al.*, 2014; MENEGOL *et al.*, 2016), etc. Essas biomassas lignocelulósicas, devido a grande disponibilidade e baixo custo, tornam-se uma contribuição relevante para o sucesso da tecnologia do etanol 2G, com consequente aumento da produção do combustível, sem aumentar a área cultivada (JOUZANI; TAHERZADEH, 2015). Dessa forma, os resíduos vegetais agrícolas e industriais podem ser convertidos via rota bioquímica, a partir de enzimas produzidas por microrganismos dedicados que degradam a celulose para obter os açúcares contidos na biomassa.

O etanol celulósico têm um balanço de carbono excelente e a redução das emissões de CO_2 podem variar entre 60% e 90%, quando comparado aos combustíveis fósseis. Além disso, as emissões de outros poluidores importantes (NOx, SOx) é reduzida. Por outro lado, são necessários processos caros para hidrolizar a matéria-prima lignocelulósica, exigindo tecnologias de ponta para permitir a competitividade do processo. Tornando-se fundamental a redução dos custos da hidrólise enzimática da biomassa (DUTTA; DAVEREY; LIN, 2014).

Diferentes configurações de processos podem ser definidas para a produção de etanol 2G, entretanto, o processo simplificado de produção de etanol utilizando biomassa envolve 3 etapas: (i) o pré-tratamento dos materiais lignocelulósicos presentes na biomassa, (ii) a hidrólise da celulose através de tratamentos enzimáticos para a obtenção de glicose e (iii) a fermentação que produz o etanol (LYND *et al.*, 2002; JOUZANI; TAHERZADEH, 2015).

O desenvolvimento de tecnologias para produção de etanol 2G, obtido a partir de biomassa lignocelulósica, tem ganhado força nos últimos anos. Embora alguns empecilhos estejam dificultando a ampliação da produção em larga escala, que está mais lenta do que o esperado (GE; LI, 2018). Como consequência do desenvolvimento das tecnologias de produção, é possível obter um aumento na produção deste combustível, dando um passo importante para tentar alcançar a independência em relação ao petróleo (BAYRAKCI; KOÇAR, 2014).

O etanol é um combustível renovável com baixa toxicidade, alta densidade energética e fácil produção através da fermentação de diferentes fontes de açúcares. Essas características tornam o etanol um combustível atraente para dispositivos de conversão de energia, especialmente para células combustíveis. Essa linha de pesquisa promissora compreende a adaptação de mecanismos veiculares para utilização de etanol em veículos elétricos (JABLONSKI; LEWERA, 2015).

Recentemente, pesquisadores produziram um catalisador altamente eficiente para extrair energia elétrica do etanol, o catalisador orienta a eletro-oxidação do etanol por um caminho químico ideal que libera todo o potencial de energia armazenada do combustível líquido. Este catalisador representa um evento histórico que viabiliza o uso de células a combustível de etanol como uma fonte promissora de energia elétrica e de alta densidade energética. As células de combustível de etanol são leves em comparação com as baterias, fornecendo energia suficiente para operar *drones* a partir de etanol. Este grande salto tecnológico deve permitir, em um futuro próximo, que as células a combustível gerem energia suficiente para fazer rodar um carro elétrico movido a etanol (LIANG *et al.*, 2019).

Veículos elétricos, por sua vez, podem contribuir para a redução da poluição atmosférica local. No entanto, a sua contribuição para a redução das emissões de GEE depende totalmente da forma como é produzida a eletricidade usada para carga. Embora este seja um problema menor em países com disponibilidade de fontes renováveis como a Noruega, Áustria ou Suécia, na maioria dos outros países que utilizam carvão para geração de eletricidade (e.g., China, Turquia, Grécia), os carros elétricos não apresentam vantagens. Atualmente, em muitos países, as emissões de CO_2 por kWh de eletricidade gerada é muito alta, levando ao efeito de que virtualmente não ocorra redução nas emissões de GEE por veículos elétricos (AJANOVIC; HAAS, 2018).

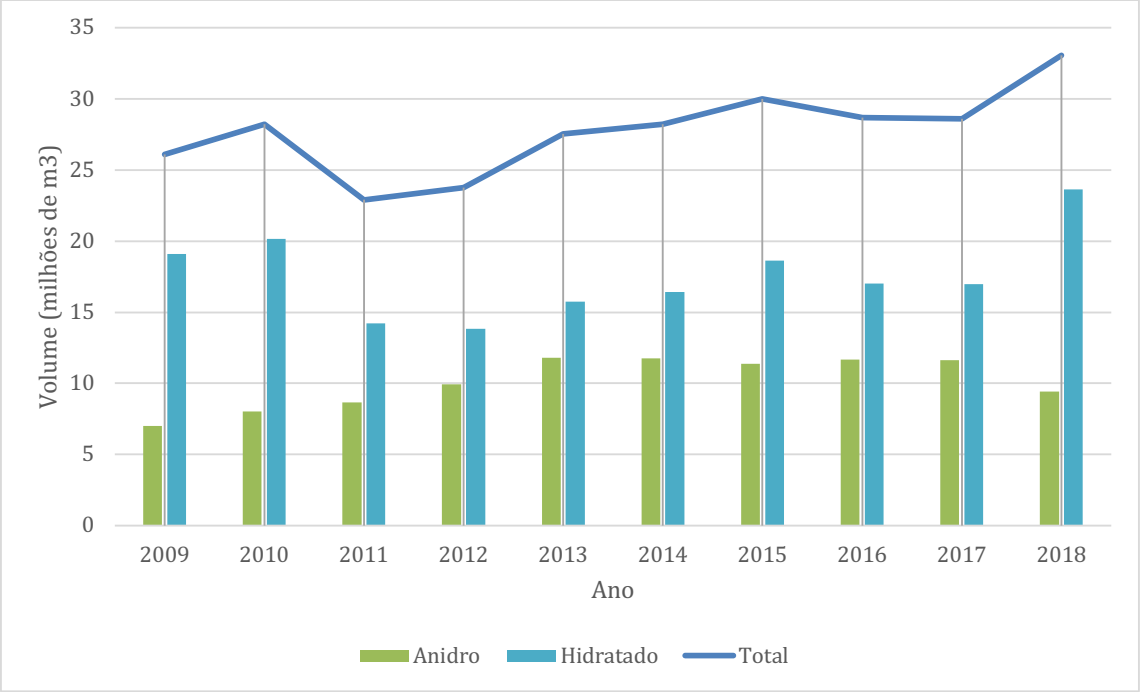
A seguir são apresentados os panoramas nacional e global referentes ao histórico e prospecções da produção e consumo de etanol convencional e etanol 2G.

2.4.2.1.1 Panorama nacional

O gráfico de evolução da produção de etanol no Brasil entre 2009 e 2018, disponibilizado pela Agência Nacional do Petróleo (ANP), demonstra um aumento na produção nos últimos anos, totalizando 33,06 milhões de m^3 em 2018, quando somadas a produção do etanol anidro combustível e a produção de etanol hidratado combustível. A região sudeste do Brasil lidera a produção nacional, seguida pela região centro-oeste (ANP, 2019).

No Balanço Energético Nacional do ano base 2018 (EPE, 2019b), fornecido pelo Ministério de Minas e Energia (MME), a porcentagem de oferta interna de energia renovável produzida foi de 45,3%, a biomassa da cana-de açúcar compreende 17,4% em relação à oferta interna de todas as fontes energéticas. O pico de oferta interna dessa fonte energética ocorreu em 2009, totalizando 18,1% do total, no entanto esse percentual decaiu bastante em seguida e voltou

Figura 9 – Gráfico de evolução da produção de etanol entre 2009 e 2018 no Brasil.



Fonte: ANP (2019).

a crescer nos últimos anos. O relatório também destaca as energias não renováveis, as quais apresentam queda de 5,5% no ano de 2018, em relação à 2017. Essa queda foi impulsionada pelo petróleo e seus derivados cuja oferta foi reduzida em 6,5%, totalizando 34,4%, em relação à oferta interna de todas as fontes energéticas em 2018.

Já o consumo final de etanol compreendeu 6,4% em relação ao consumo de todas as fontes em 2018. Enquanto que o consumo final de gasolina e óleo diesel somados totalizaram 25,1% no mesmo ano. Em relação ao ano de 2017, o consumo de etanol aumentou 0,7% e o consumo de gasolina e óleo diesel somados aumentaram 2,1% (EPE, 2019b). Ainda em 2018, as emissões evitadas pelo uso de etanol (anidro e hidratado) e biodiesel, em comparação aos equivalentes fósseis (gasolina e óleo diesel), foi de 63,7 $MtCO_2$ (EPE, 2019a).

A análise de conjectura dos biocombustíveis, ano base 2018, informa que a implementação mundial da produção comercial do etanol 2G segue em ritmo lento. As três plantas com projetos de segunda geração localizadas em território nacional correspondem à duas usinas de escala comercial em operação (GranBio e Raízen) e uma usina piloto do Centro de Tecnologia Canavieira (CTC). As três plantas enfrentaram desafios técnicos e realizaram ajustes em seus processos, permanecendo em funcionamento abaixo da capacidade nominal. A GranBio, por exemplo, operou durante alguns meses no começo de 2019 e deve retomar as operações no início da safra da região Nordeste. Para o ano de 2020 no Brasil, espera-se uma produção entre 40 e 45

milhões de litros de etanol 2G (EPE, 2019a).

Pelo quinto ano consecutivo (2015-2019), o Brasil foi o principal destino do etanol produzido nos Estados Unidos, ocupando 24% das exportações do etanol norte americano (KOEHLER *et al.*, 2020). Analisando o histórico da produção e de consumo de etanol no Brasil e a necessidade de importação de etanol dos Estados Unidos, conclui-se que a produção de etanol no Brasil está muito aquém das necessidades. As prospecções para transição energética global sugerem um aumento significativo da oferta e consumo de combustíveis renováveis enquanto a oferta e o consumo de combustíveis fósseis deveriam se manter estáveis ou até mesmo serem reduzidos. No entanto, observa-se que o consumo de combustíveis fósseis no Brasil segue em plena expansão, na contramão das metas globais de redução.

2.4.2.1.2 Panorama global

De acordo com a Agência Internacional de Energia (IEA) (2019), os biocombustíveis têm feito algumas incursões importantes na área de transporte rodoviário, mas, em 2018, ainda se encontravam apenas com 3,7% da demanda de combustíveis nessa área. A produção global de biocombustíveis em 2018 aumentou 10 Bi de litros, atingindo seu recorde de 154 Bi de litros em nível global e a concentração da produção e do consumo ocorrem nos Estados Unidos. A perspectiva é que ocorra um aumento de 25% até 2024 no Brasil, Estados Unidos e especialmente na China, podendo atingir 190 Bi de litros. Observa-se também a expansão para novos mercados promissores, englobando países da América e da Ásia, como a Tailândia.

Considerando a perspectiva até 2024, a China possui o maior aumento na produção de biocombustíveis quando comparada a outros países. Isso ocorre principalmente porque a produção de etanol está prevista para mais do que o triplo atual, pois são tomadas medidas concretas para aumentar o consumo de etanol de cerca de 2% para a meta de 10% da demanda nacional de gasolina. Os programas de mistura de etanol estão se expandindo para mais províncias e a capacidade de produção deve aumentar quase 50% em 2021 (IEA, 2019).

Enquanto no Brasil a matéria-prima mais utilizada é a cana-de-açúcar, os Estados Unidos produzem o etanol a partir do milho. No ano de 2019, o Brasil respondeu por 30% da produção global, enquanto os Estados Unidos assumiram 54%. Vinte usinas fecharam as portas permanente ou temporariamente em 2019 e a produção de etanol nos Estados Unidos caiu para 15,8 bilhões de galões, uma redução de 300 milhões de galões em relação à 2018. Ainda em 2019, aproximadamente 6% da produção de etanol dos Estados Unidos ocorreu a partir de fontes não convencionais, principalmente biomassa celulósica (KOEHLER *et al.*, 2020).

De acordo com o Relatório Especial em nível Mundial de Energia e Mudanças Climáticas da IEA (2015), em um cenário de prospecção, o uso de eletricidade, através de veículos elétricos, e de biocombustíveis avançados, como o etanol 2G, são as alternativas de combustível que proporcionam a redução das emissões de GEE mais profundas, exigidas no setor de transportes

terrestres. A prospecção indica que essas duas fontes juntas podem reduzir o consumo de petróleo em cerca de 13,8 milhões de barris por dia em 2040 e as emissões de CO_2 em cerca de 11,5 Gt até 2040.

A mesma prospecção (IEA, 2015) ainda indica que cerca de 30% do aumento na produção de biocombustíveis pode ser direcionada para o setor da aviação, que tem poucas alternativas viáveis de combustíveis por causa das limitações de armazenamento que excluem o hidrogênio e a eletricidade. Neste cenário, os biocombustíveis devem ocupar uma quota de 17% da demanda mundial em todos os setores de transporte até 2040. O relatório ainda enfatiza que os preços baixos dos derivados de petróleo são o maior desafio para as fontes de energia renováveis. Essa afirmação é corroborada ao analisarmos o panorama brasileiro referente aos preços do etanol no país (EPE, 2019b).

Conforme o relatório de custos da IRENA (2013), os três principais desafios associados à produção de biocombustíveis avançados compreendem desafios técnicos no processo, redução dos custos de produção e a garantia da sustentabilidade. A maioria das tecnologias de produção de biocombustíveis avançados têm custos totais de produção significativamente superiores ao preço sem impostos de produtos petrolíferos. A cotação do petróleo convencional será de importância fundamental para as perspectivas competitivas de longo prazo para os biocombustíveis. A segunda questão compreende a sustentabilidade dos biocombustíveis convencionais, que tem sido colocada em questão por muitos anos, sendo o principal fator de obstrução da expansão do consumo de etanol na União Europeia. Portanto, é necessário garantir que os biocombustíveis avançados não levantem as mesmas preocupações, como a expansão agrícola, desmatamento e a competição com a área de produção de alimentos (PANAHI *et al.*, 2019).

2.4.2.2 Processo de produção de etanol 2G

O etanol 2G obteve um interesse considerável devido ao seu potencial para mitigar as mudanças climáticas globais e melhorar a segurança energética fazendo uso de resíduos vegetais. A biomassa vegetal é o recurso renovável mais abundante da biosfera (PAULY; KEEGSTRA, 2008; JOUZANI; TAHERZADEH, 2015). No entanto, a demanda global por produtos baseados em biomassa deverá crescer nas próximas décadas e estima-se que a demanda por biomassa pode dobrar até 2050, gerando uma demanda possivelmente maior que a produção (MAUSER *et al.*, 2015).

Geralmente, a biomassa vegetal lignocelulósica é obtida a partir de quatro fontes principais: (i) resíduos agrícolas (palha de milho, palha de arroz, palha de cana-de-açúcar, bagaço de cana de açúcar, etc.); (ii) resíduos florestais (madeiras, ramos, folhas, etc.); (iii) culturas energéticas (capim-elefante, *switchgrass*, álamo amarelo, etc.); e (iv) resíduos de celulose (resíduos sólidos urbanos e resíduos de alimentos) (PARISUTHAM; KIM; LEE, 2014).

A biomassa tipicamente contém 50 a 80% de polissacarídeos representando um desafio para a produção de etanol 2G (JOUZANI; TAHERZADEH, 2015). Esses polissacarídeos que

compõem as células vegetais variam em quantidade para cada tipo de biomassa, compreendendo 40-60% de celulose, 20-40% de hemicelulose e 15-25% de lignina (PÉREZ *et al.*, 2002; BALAT; BALAT; ÖZ, 2008).

Devido à complexidade estrutural e à resistência aos processos de transformação, a produção rentável e eficiente de combustíveis líquidos a partir de biomassa lignocelulósica continua sendo um desafio (XIA *et al.*, 2016). Conforme citado anteriormente, esta estrutura robusta e complexa requer um processo que, de modo simplificado, inclui três etapas para converter a matéria vegetal em bioetanol: o pré-tratamento, a hidrólise enzimática e a fermentação. Esse processo aumenta o custo de produção deste biocombustível significativamente, tendo a hidrólise enzimática das lignoceluloses como a etapa de maior custo (LYND *et al.*, 2002; HO *et al.*, 2012; REIS *et al.*, 2014; KUMAGAI *et al.*, 2014).

Uma avaliação de custo de coquetéis celulolíticos para produção de etanol celulósico em escala industrial sugeriu que o custo das enzimas leva o preço do etanol celulósico abaixo do ponto mínimo de lucro, quando são utilizados coquetéis enzimáticos comprados no mercado de enzimas industriais. Esse fato demanda a produção inovadora de coquetéis celulolíticos para produzir um suprimento de enzimas economicamente viável para a produção de etanol celulósico (LIU; ZHANG; BAO, 2016).

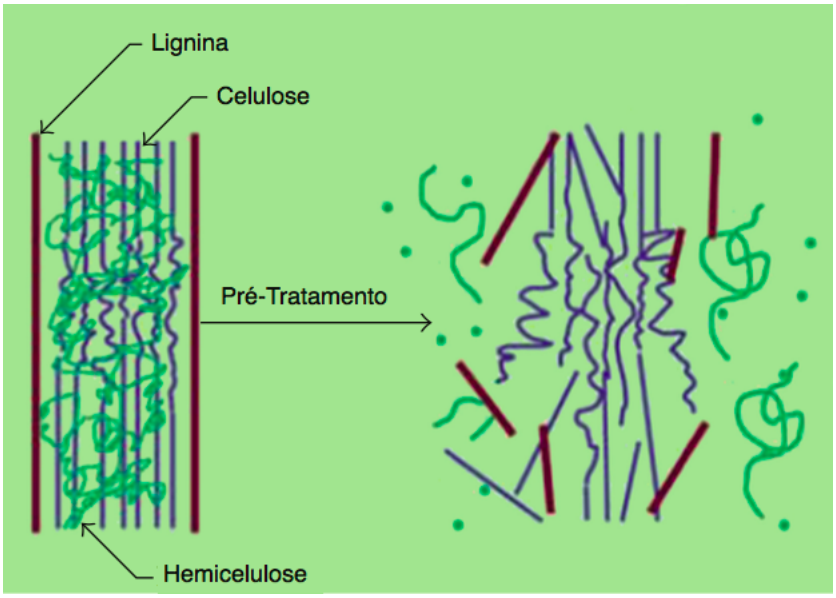
A etapa de pré-processamento compreende 18-20% do custo do processamento da biomassa. Essa etapa consiste no emprego de um método que amenize a interação entre os principais componentes da biomassa e os torne propensos à conversão em combustíveis, conforme pode-se observar na Figura 10. Os métodos de pré-tratamento podem ser físicos, químicos, biológicos ou combinados. Embora cada método apresente vantagens, não existe um método ideal que contemple todos os tipos de biomassa, pois cada método de pré-tratamento apresenta limitações em sua aplicação (JOUZANI; TAHERZADEH, 2015).

A segunda etapa do processo, conhecida como hidrólise enzimática, consiste na atuação cooperativa e sinérgica das enzimas lignocelulolíticas na degradação da celulose e da hemicelulose em glicose e outros açúcares. Por fim, a última etapa do processo compreende a fermentação dos açúcares resultantes da hidrólise enzimática, sendo efetuada por algum microrganismo produtor de etanol (LYND *et al.*, 2002).

A hidrólise enzimática demanda elevadas quantidades de enzimas, podendo tornar o processo de produção de etanol 2G desvantajoso economicamente. O alto custo de produção desse sistema enzimático representa cerca de 40% do custo total da produção do etanol a partir da biomassa. Assim, destaca-se a importância do estudo dos microrganismos produtores de sistemas enzimáticos, visto que são essenciais para viabilizar economicamente a produção do etanol 2G em larga escala (LYND *et al.*, 2002).

Nas três plantas brasileiras de produção de etanol 2G, desafios técnicos e a necessidade ajustes nos processos tem dificultado a produção comercial deste biocombustível (EPE, 2019a).

Figura 10 – Ilustração da etapa de pré-tratamento de biomassa vegetal.



Fonte: Brodeur *et al.* (2011).

Dessa forma, percebe-se a necessidade de desenvolvimento de mais pesquisas aplicadas nessa área, com intuito de fomentar a indústria nacional e alavancar a produção do etanol 2G.

A ampla adoção deste biocombustível celulósico está ocorrendo em um ritmo mais lento do que o esperado. Um grande desafio é que a produção do etanol 2G ainda consome muita energia. Em um estudo recente, as relações entre energia e parâmetros de produção foram estudadas e, a partir da condução de um estudo de caso, foram identificados os principais fatores energéticos. Dessa forma, foi possível realizar a otimização do processo que resultou em uma redução de 21,09% no consumo total de energia (GE; LI, 2018).

Outra estratégia, recentemente proposta para produção de etanol, utiliza um consórcio microbiano contendo microrganismos com diferentes capacidades celulolíticas e de fermentação, como ocorre normalmente na natureza. Esses consórcios partem do princípio que a sinergia entre diferentes microrganismos pode aumentar a utilização do substrato, incrementando a eficiência do processo de produção do etanol 2G (BRETHAUER; STUDER, 2014; JOUZANI; TAHERZADEH, 2015).

Outro eixo de pesquisa que pretende reduzir os custos do processo de produção enfatiza o aperfeiçoamento dos sistemas enzimáticos. As enzimas são obtidas a partir de ambientes muito diversos e a composição dos sistemas varia de lugar para lugar e ao longo do tempo. Por isso, ao invés de utilizar coquetéis celulolíticos comerciais como ocorre nas plantas nacionais, uma planta de etanol celulósico necessita desenvolver os processos de pré-tratamento, hidrólise e fermentação de forma eficiente e integrada, utilizando fungos eficientes capazes de lidar com a conversão eficaz da biomassa (PARISUTHAM; KIM; LEE, 2014; AKINOSHO *et al.*, 2014).

3 MATERIAL E MÉTODOS

Neste capítulo, são descritos os materiais e a metodologia utilizados para atingir os objetivos traçados nesta tese. Esta pesquisa é classificada como aplicada e exploratória, de acordo com a sua natureza e objetivos, respectivamente. Conforme descrito anteriormente, a motivação desta tese surgiu da necessidade de empregar os dados genômicos do sequenciamento realizado em 2013 para obtenção de conhecimento acerca do fungo estudado, possibilitando a engenharia de linhagens comerciais hiperprodutoras de enzimas celulolíticas. Assim, torna-se imprescindível a contextualização das fases anteriores do Projeto Genoma. Com intuito de facilitar a compreensão da metodologia, a Figura 11 apresenta o fluxograma metodológico das fases e etapas do Projeto Genoma das linhagens 2HH e S1M29 identificados como *P. echinulatum*. Os métodos detalhados utilizados nesta tese estão descritos somente nos três manuscritos que contemplam os resultados desta tese, portanto, não são contemplados neste capítulo. As seções subsequentes compreendem as fases do fluxograma metodológico, assim como a descrição dos organismos e linhagens utilizados nesta tese.

3.1 Fases e etapas do Projeto Genoma

O Projeto Genoma foi concebido em 2013 a partir de uma parceria entre a Universidade de Caxias do Sul, o Laboratório Nacional de Ciência e Tecnologia do Bioetanol (CTBE) e a Universidade de Oklahoma. Essa primeira fase do projeto englobou o sequenciamento completo dos genomas da linhagem selvagem 2HH e do mutante S1M29. O sequenciamento foi realizado pela empresa californiana Ambry Genetics. Posteriormente os genomas foram montados na Universidade de Oklahoma, onde também foi realizada a predição de genes. Entre os anos de 2014 e 2016, foi conduzido um estudo de secretômica das duas linhagens. Este estudo foi publicado em 2016 (SCHNEIDER *et al.*, 2016) e contempla as características gerais dos genomas e o cruzamento dos dados de secretômica com as predições de genes realizadas em Oklahoma, esta etapa ocorreu no CTBE em Campinas. É importante ressaltar que até esse momento do projeto, todas os estudos genômicos ocorreram fora da UCS, observa-se também que os genomas completos não foram depositados em bases de dados públicas. Ainda no ano de 2016, as montagens e respectivas predições de genes foram disponibilizadas para o grupo de pesquisa da UCS. No entanto, foi observado que o tamanho da montagem do genoma do mutante S1M29 disponibilizada pela Universidade de Oklahoma estava diferente das características publicadas no artigo, sendo que o arquivo FASTA em posse do CTBE também estava incompleto.

A fase 2 do projeto iniciou-se com a decisão de realizar novas montagens e predições de genes no CTBE em Campinas. Considerando que os arquivos originais dos sequenciamentos, gerados pelo sequenciador Illumina, também estavam em posse do CTBE. Apesar disso, essa nova

fase do projeto não gerou os resultados esperados pelo grupo de pesquisa da UCS, uma vez que a metodologia para realização das novas montagens e predições de genes realizadas em Campinas estava incompleta e não reprodutível. O conjunto de arquivos da anotação funcional continha resultados duvidosos e poderiam comprometer publicações futuras. Essas inconsistências foram detectadas em março de 2017 no início das atividades deste doutorado. É importante ressaltar que o projeto de tese apresentado ao PPGBIO compreendia a análise dos genomas da linhagem 2HH e do mutante S1M29 para identificação de sequências regulatórias e caracterização de suas funções e, portanto, dependia de genomas bem montados e anotados como premissa básica para execução.

Isto posto, foram solicitados ao CTBE os arquivos originais dos sequenciamentos gerados pelo Illumina. Em maio de 2017 esses arquivos foram recebidos em um HD externo enviado pelos Correios, dando início a fase 3 do Projeto Genoma e resultando em um replanejamento completo desta tese que passou a focar na obtenção de montagens e anotações confiáveis para os genomas em questão. Como não existia expertise para realização destas atividades na UCS, foi necessária uma etapa de revisão bibliográfica sobre o assunto para posterior montagem de um ambiente computacional baseado em Unix que suportasse as montagens de genomas fúngicos.

Para tanto, foi adquirida uma *Workstation* HP Z600 com dois processadores Intel Xeon (2,67 GHz) de 6 cores cada, incluindo também 64GB de memória DDR3 e um total de 2,3TB de capacidade de armazenamento em disco rígido. Subsequentemente, foi realizada a configuração do ambiente computacional, a partir da instalação de um vasto repertório de softwares e bibliotecas necessários para montagem e anotação genômica. Nota-se que a aquisição desta máquina foi crucial para realização desta tese, considerando o grande volume de dados analisados e principalmente a falta de expertise na área.

Conforme sugerido por (EKBLOM; WOLF, 2014), foram testadas diversas metodologias e ferramentas para montagem dos genomas, permitindo a identificação da melhor alternativa para os dados disponíveis. É importante levar em consideração que o sequenciamento Illumina realizado em 2013 contemplou somente leituras de tamanho curto (100pb), fato que dificulta a montagem contígua do genoma e inviabiliza o uso de ferramentas computacionais mais robustas que exigem a disponibilidade de leituras de tamanho médio e/ou longo (mate-pair ou PACBIO). Após a obtenção de resultados adequados na avaliação de qualidade das montagens, a execução das fases de predição de genes e anotação funcional se mostraram ainda mais complexas, uma vez que os softwares automáticos são propensos a erros e a curadoria dos modelos de genes e da anotação funcional tornam-se essenciais.

A inspeção dos modelos de genes e a caracterização funcional das proteínas preditas gerou mais uma demanda inesperada no projeto. Posto que, naquele momento do projeto, os softwares gratuitos para realização destas atividades não atendiam às necessidades. Consequentemente, essa demanda resultou em um projeto de mestrado paralelo para concepção de uma ferramenta web colaborativa, capaz de suprir as necessidades. Essa ferramenta, denominada

Seq2Annot, foi concebida com intuito de permitir a atualização de modelos de genes inconsistentes e, principalmente, efetuar a anotação funcional de forma automática a partir de consultas a bancos de dados e serviços web de anotação funcional. A ferramenta permite o trabalho colaborativo de uma equipe, provendo o acesso facilitado a conjuntos de genes específicos e suas respectivas anotações. Nota-se que essa etapa consome inúmeras horas, dada a quantidade de modelos de genes em um fungo filamentoso. A curadoria manual oferecida pela ferramenta foi a etapa mais custosa deste projeto e também uma das mais importantes para permitir o depósito dos genomas nas bases de dados públicas DDBJ/ENA/GenBank.

As versões dos genomas das linhagens selvagem 2HH e do mutante S1M29 foram depositadas no GenBank, sob os números de acesso WIWU00000000 e WIWV00000000, respectivamente. Destaca-se o fato de que os dois genomas foram depositados sem a identificação da espécie, ou seja, foram depositados como *Penicillium* sp. Essa decisão foi tomada em virtude dos resultados obtidos pela caracterização molecular que indica que a linhagem 2HH não corresponde à espécie *Penicillium echinulatum*, demandando um estudo taxônomico para reposicionamento e caracterização da espécie e posterior identificação no GenBank.

O depósito dos dois genomas deu início a etapa de descoberta de conhecimento de maneira exploratória. Os resultados desta tese estão organizados em forma de três manuscritos de artigos científicos para revistas da área. Os métodos detalhados para atingir os objetivos desta tese estão descritos em cada um dos manuscritos e todo o material suplementar gerado nesta tese está disponível no projeto FunRegulation no Git http://www.github.com/alexandrelenz/funregulation.git.

Este doutoramento ainda englobou um intercâmbio científico realizado na Universidad Nacional Autónoma de México, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Unidad Académica Yucatán, Mérida, México. Esse intercâmbio resultou na publicação de um artigo na revista *Frontiers in Microbiology* sob DOI: 10.3389/fmicb.2020.588263, abrangendo não somente a linhagem 2HH de *P. echinulatum*, mas também a linhagem 114-2 de *P. oxalicum*, devido à sua proximidade filogenética. Esse artigo contempla: i) a caracterização funcional de TFomas; ii) a identificação de genes ortólogos compartilhados entre espécies relacionadas; iii) a construção de GRNs; iv) a descoberta de conhecimento a partir da análise das GRNs; v) a sugestão de genes-alvo para engenharia de linhagens hiperprodutoras de complexos enzimáticos. É importante ressaltar que as GRNs deste estudo compreendem as primeiras GRNs inferidas para o gênero *Penicillium*.

O manuscrito seguinte foi submetido à revista *Fungal Genetics & Biology* e está em fase de revisão, contemplando a principal aplicação biotecnológica deste fungo, a obtenção de complexos enzimáticos para produção de etanol 2G. Esse manuscrito inclui: i) a caracterização funcional do CAZyoma e do transportoma de açúcares do fungo; ii) as análises filogenéticas de iBGLs e de transportadores de açúcares; iii) sugestões evolutivas relacionadas às enzimas CAZy necessárias para degradação dos polímeros disponíveis no meio de crescimento; iv) a sugestão de genes-alvo para engenharia de linhagens hiperprodutoras de

complexos enzimáticos. Figuras e material suplementar deste manuscrito estão disponíveis no Git (https://github.com/alexandrelenz/funregulation/tree/master/Manuscripts/Article2)

Por fim, o último manuscrito, formatado de acordo com as normas para publicação na revista *International Journal of Systematic and Evolutionary Microbiology*, engloba: i) o depósito dos dois genomas nas bases de dados públicas; ii) a comparação entre os dois genomas identificando mutações que levaram à hiperprodução de celulases e ao albinismo do mutante S1M29; iii) a identificação dos genes envolvidos na composição da parede celular e na produção de melanina; iv) sugestões evolutivas relacionadas à composição da parede celular durante uma possível simbiose mutualística a longo prazo; e v) a identificação molecular da espécie.

Ressalta-se que esse manuscrito depende da caracterização morfológica e do perfil de extrólitos da nova espécie, não contempladas por esta tese. Essa investigação está sendo realizada em parceria com o PhD. Jos Houbraken, Westerdijk Fungal Biodiversity Institute, Utrecht, Holanda, referência na identificação e caracterização de espécies do gênero *Penicillium*. A pandemia global afetou diretamente a caracterização da espécie, que aguarda a retomada das atividades do pesquisador holandês. É fundamental destacar que a classificação *Penicillium ucsensis* **sp. nov.** será válida somente após a caracterização da nova espécie e depósito no MycoBank. Nesse sentido, esse manuscrito deve ser avaliado como um projeto em andamento. Figuras e material suplementar deste manuscrito estão disponíveis no Git (https://github.com/alexandrelenz/funregulation/tree/master/Manuscripts/Article1)

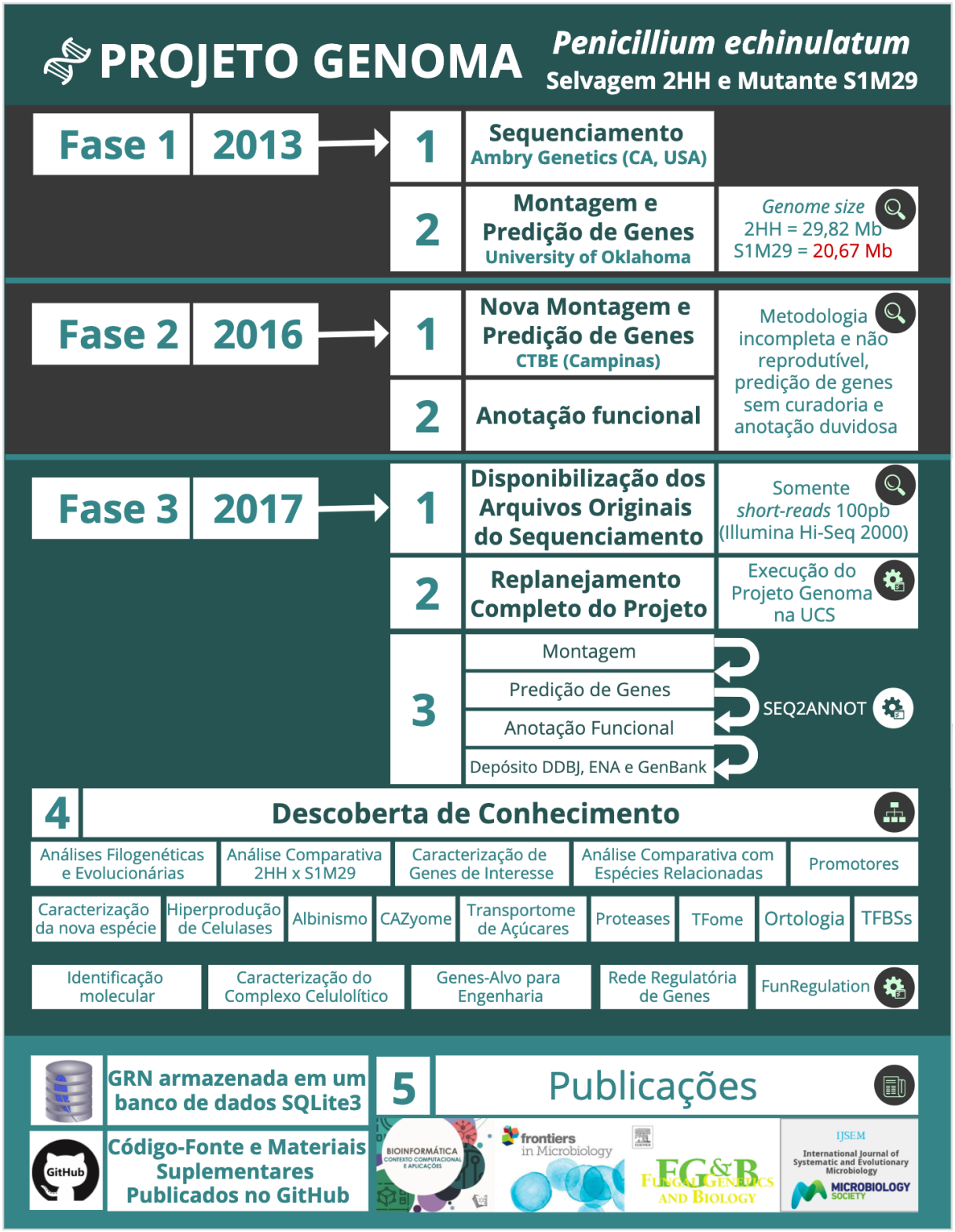
3.2 Organismos e linhagens

Para os experimentos laboratoriais, foram utilizados o isolado selvagem 2HH e o mutante S1M29 do gênero *Penicillium*. As linhagens pertencem à coleção de microrganismos do Laboratório de Enzimas e Biomassas da Universidade de Caxias do Sul. Para fins de identificação e caracterização da nova espécie, a linhagem 2HH foi enviada para Holanda, seguindo as devidas regras nacionais (Conselho de Gestão do Patrimônio Genético) e internacionais para envio de remessas biológicas para o exterior.

Para as análises genômicas foram utilizados linhagens selvagens de fungos modelo, fungos filogeneticamente relacionados e fungos usados comercialmente para a produção de sistemas enzimáticos celulolíticos. Os genomas e proteomas (preditos a partir dos genomas) foram obtidos do UniprotKB e/ou GenBank, os respectivos números de acesso encontram-se entre parênteses: a) *Penicillium* sp. 2HH (GCA014839855.1); b) *Penicillium* sp. S1M29 (GCA014839625.1); c) *P. oxalicum* 114-2 (GCA000346795.1 e UP000019376); d) *Penicillium chrysogenum* P2niaD18 (GCA000710275.1 e UP000076449); e) *Penicillium digitatum* Pd1 (GCF000315645.1 e UP000009886); f) *Penicillium brasilianum* MG11 (GCA001048715.1 e UP000042958); g) *Penicillium subrubescens* CBS 132785 (GCA001908125.1 e UP000186955); h) *A. niger* CBS 513.88 (GCF000002855.3 e UP000006706); i) *A. nidulans* FGSC A4 (GCF000149205.2

e UP000000560); j) *Aspergillus oryzae* RIB40 (GCA000184455.3 e UP000006564); k) *Aspergillus fumigatus* Af293 (GCA000002655.1 e UP000002530); l) *T. reesei* Qm6a (GCA000167675.2 e UP000008984); m) *N. crassa* OR74A (GCF000182925.2 e UP000001805); e n) *S. cerevisiae* S288c (GCF000146045.2).

Figura 11 – Fluxograma metodológico



Fonte: Elaborada pelo autor.

4 RESULTADOS E DISCUSSÃO

Este capítulo apresenta os resultados obtidos a partir da metodologia descrita no capítulo anterior de forma a cumprir os objetivos traçados nesta tese. Os resultados estão organizados em três manuscritos de artigos científicos, formatados de acordo com as normas das revistas *Frontiers in Microbiology*, *Fungal Genetics & Biology* e *International Journal of Systematic and Evolutionary Microbiology*.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 27 October 2020
doi: 10.3389/fmicb.2020.588263

Gene Regulatory Networks of *Penicillium echinulatum* 2HH and *Penicillium oxalicum* 114-2 Inferred by a Computational Biology Approach

Alexandre Rafael Lenz[1,2,3]*, *Edgardo Galán-Vásquez*[4], *Eduardo Balbinot*[2], *Fernanda Pessi de Abreu*[2], *Nikael Souza de Oliveira*[2,5], *Letícia Osório da Rosa*[5], *Scheila de Avila e Silva*[2], *Marli Camassola*[5], *Aldo José Pinheiro Dillon*[5] and *Ernesto Perez-Rueda*[1,6]*

[1] Unidad Académica Yucatán, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de Mexico, Mérida, Mexico, [2] Laboratório de Bioinformática e Biologia Computacional, Instituto de Biotecnologia, Universidade de Caxias do Sul, Caxias do Sul, Brazil, [3] Departamento de Ciências Exatas e da Terra, Universidade do Estado da Bahia, Salvador, Brazil, [4] Departamento de Ingeniería de Sistemas Computacionales y Automatización, Instituto de Investigaciones en Matemàticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de Mexico, Ciudad Universitaria, Mexico, [5] Laboratório de Enzimas e Biomassas, Instituto de Biotecnologia, Universidade de Caxias do Sul, Caxias do Sul, Brazil, [6] Facultad de Ciencias, Centro de Genómica y Bioinformática, Universidad Mayor, Santiago, Chile

Penicillium echinulatum 2HH and *Penicillium oxalicum* 114-2 are well-known cellulase fungal producers. However, few studies addressing global mechanisms for gene regulation of these two important organisms are available so far. A recent finding that the 2HH wild-type is closely related to *P. oxalicum* leads to a combined study of these two species. Firstly, we provide a global gene regulatory network for *P. echinulatum* 2HH and *P. oxalicum* 114-2, based on TF-TG orthology relationships, considering three related species with well-known regulatory interactions combined with TFBSs prediction. The network was then analyzed in terms of topology, identifying TFs as hubs, and modules. Based on this approach, we explore numerous identified modules, such as the expression of cellulolytic and xylanolytic systems, where XlnR plays a key role in positive regulation of the xylanolytic system. It also regulates positively the cellulolytic system by acting indirectly through the cellodextrin induction system. This remarkable finding suggests that the XlnR-dependent cellulolytic and xylanolytic regulatory systems are probably conserved in both *P. echinulatum* and *P. oxalicum*. Finally, we explore the functional congruency on the genes clustered in terms of communities, where the genes related to cellular nitrogen, compound metabolic process and macromolecule metabolic process were the most abundant. Therefore, our approach allows us to confer a degree of accuracy regarding the existence of each inferred interaction.

Keywords: *Penicillium*, regulatory network, orthologous, gene regulation, genomics, fungi

OPEN ACCESS

Edited by:
George Tsiamis,
University of Patras, Greece

Reviewed by:
Yinbo Qu,
Shandong University, China
Shuai Zhao,
Guangxi University, China
Yuqi Qin,
Shandong University, China

***Correspondence:**
Alexandre Rafael Lenz
arlenz@ucs.br
Ernesto Perez-Rueda
ernesto.perez@iimas.unam.mx

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *28 July 2020*
Accepted: *23 September 2020*
Published: *27 October 2020*

Citation:
Lenz AR, Galàn-Vàsquez E, Balbinot E, Abreu FP, Souza de Oliveira N, Rosa LO, Avila e Silva S, Camassola M, Dillon AJP and Perez-Rueda E (2020) Gene Regulatory Networks of Penicillium echinulatum 2HH and Penicillium oxalicum 114-2 Inferred by a Computational Biology Approach. Front. Microbiol. 11:588263. doi: 10.3389/fmicb.2020.588263

1. INTRODUCTION

The connectivity between biological data facilitates the inference of networks. These graphs are made up of nodes and connections between them, comprising a flexible model that can capture the complexity and interconnectivity of biological information (Huber et al., 2007). The elucidation of regulatory relationships between transcription regulators and their target genes is essential to understand various biological processes. These processes range from cell growth and division, cell differentiation in multicellular organisms and cell response to environmental changes. In addition, networks are often identified as the layer that connects genomic data to phenotypic characteristics (Carter et al., 2013).

A gene regulatory network (GRN) is a directed graph in which gene regulators are connected to target genes (TGs) by interaction edges (Karlebach and Shamir, 2008). In addition to transcription factors (TFs) that can act as both activators and repressors, gene regulators also include RNA-binding proteins and regulatory RNAs. This type of network addresses a key challenge in experimental and computational biology, helping to clarify the relationships between genes and the products they encode (Jackson et al., 2020). GRNs combined with knowledge mining has a major potential to improve omics data interpretation, allowing the discovery of how transcription regulation may control biological processes, phenotypes, and diseases (Hassani-Pak and Rawlings, 2017).

To date, GRNs have been reconstructed only for a few model organisms (Gerstein et al., 2012; Chen et al., 2018; Hu et al., 2018; Jackson et al., 2020), since their reconstruction depends largely on experimental approaches. On the other hand, GRN inferences become a viable alternative (Filho et al., 2019; Staunton et al., 2019) supported by several information resources and bioinformatic tools (Glenwinkel et al., 2014; Penfold et al., 2015; Lam et al., 2016; Koch et al., 2017; Kulkarni et al., 2017).

GRN inference resources may be categorized into six classes, according to the approach employed and the underlying data used: Coexpression, Sequence Motifs, Chromatin Immunoprecipitation (ChIP), Orthology, Literature, and Protein-Protein Interaction (PPI) specifically focused on transcriptional complexes. The more information is aggregated, the more accurate a TF-TG relationship becomes. In particular, GRNs can benefit from orthology-based knowledge from closely related species; where the key concept is that a TF-TG relationship proven in one organism can be conserved in another one (Mercatelli et al., 2020). However, this knowledge transfer requires reliable methods to define orthology between different genes for any TF-TG pair, as well as taking into account the phylogenetic positioning of the species analyzed (Fernandez-Valverde et al., 2018).

Orthologous genes are the most similar genetic elements in different species, in terms of sequence, structure, and function (Gabaldón and Koonin, 2013). Detecting orthology is especially important to maximize information content and accuracy. Therefore, the premise of constructing a TF-TG network is based on the presence of genes that can be traced to a common ancestor between different species, also taking into account functionality besides nucleotide sequence (Mercatelli et al., 2020).

Another widespread GRN-inference resource comprehends sequence motifs. This resource comprises the identification of conserved DNA sequence motifs recognized by TFs in the regulatory region of genes. These motifs, known as transcription factor binding sites (TFBSs), can be represented in the form of a position frequency matrix (PFM) or a position weight matrix (PWM) (Hu et al., 2013), which may be useful to increase the accuracy of TF-TG relationships (Mercatelli et al., 2020). Recent studies, characterizing the specificity of TFs in eukaryotes, include the representation of several TFBSs in the form of matrices for model microorganisms such as *Aspergillus nidulans*, *Neurospora crassa*, and *Saccharomyces cerevisiae* (Lambert et al., 2019).

The identification of TFBSs is based on the assumption that some non-coding regions among related species are likely to be under negative selection and therefore contain conserved functional motifs (Hu et al., 2013). Cis-regulatory elements can be conserved in more distantly related species, even when the orthologous regulatory regions are divergent to be precisely aligned. However, despite the conservation of these motifs, the regulation role may not be conserved, suggesting possible different functions (Gasch et al., 2004).

So far, in the fungi scope, *S. cerevisiae* S288C, *N. crassa* OR74A, and *A. nidulans* FGSC A4 have in-depth studies for GRN reconstruction (Hu et al., 2018; Jackson et al., 2020), whereas, for the genus *Penicillium*, no global GRNs have been described. The reconstruction of GRN in *S. cerevisiae* was, in particular, facilitated by the YEASTRACT+ database, which gathers interaction information for this organism, comprising 12,228 interactions (Jackson et al., 2020; Monteiro et al., 2020). In this regard, curated data of *A. nidulans* and *N. crassa* regulatory interactions may be useful as a catalog for gene regulation studies in filamentous fungi, since 33 conserved regulatory interactions, supported by classical experiments were identified in both species (Hu et al., 2018).

The 2HH wild-type was previously classified by morphology as *Penicillium echinulatum* in the 80's. Long-term 2HH strain improvement studies use this classification (Camassola et al., 2004; Dillon et al., 2006, 2011; Camassola and Dillon, 2007a,b, 2009, 2010, 2012; Rubini et al., 2010; Ribeiro et al., 2012; Dos Reis et al., 2013; Novello et al., 2014; Schneider et al., 2014, 2016, 2018, 2020). However, whole genome sequences of 2HH strain, deposited recently at GenBank and released in this work, provided evidence that 2HH strain is closely related to *P. oxalicum*, suggesting a taxonomic study for the repositioning and characterization of this strain. The close relationship between these two filamentous fungi leads to a combined study, both *P. echinulatum* 2HH and *P. oxalicum* 114-2 (Liu et al., 2013d) are well-known cellulase producers studied extensively in Brazil and China, respectively. Despite advances to understand the mechanisms responsible for regulating the expression of cellulases in *P. oxalicum* (Liu et al., 2013a,b,c,d; Li et al., 2015; Yao et al., 2015), studies addressing the global mechanisms of gene regulation of these two important organisms of biotechnological

interest are scarce. It also highlights the relevance of *Penicillium* combined studies due to its potential superiority over existing cellulase producers (Vaishnav et al., 2018).

In the present work, we propose the inference of GRNs for *P. echinulatum* 2HH and *P. oxalicum* 114-2, based on TF-TG orthology relationships of three related species with well-known regulatory interactions, combined with TFBSs prediction. First, GRNs of related species (*A. nidulans*, *N. crassa*, and *S. cerevisiae*) allow the mapping of orthologous interactions. Further, the TFBSs prediction provides accuracy to TF-TG relationships. The reconstructed GRNs were posteriorly analyzed in terms of topology, identifying TFs as hubs, and modules were inferred by using a community approach algorithm. Therefore, our approach allows us to confer a degree of accuracy regarding the existence of each inferred interaction.

2. MATERIALS AND METHODS

The schematic workflow (**Figure 1**) describes procedure steps of the network inference. Details on each step herewith the input and output data are described below.

2.1. Fungal Genomes Analyzed

The information of five fungal genomes and proteomes used in this study were downloaded from the NCBI server for (a) *P. echinulatum* 2HH (GCA_014839855.1 UCS_PECH_1.0); (b) *P. oxalicum* 114-2 (GCA_000346795.1 pdev1.0); (c) *A. nidulans* FGSC A4 (GCF_000149205.2 ASM14920v2); (d) *N. crassa* OR74A (GCF_000182925.2 NC12); and (e) *S. cerevisiae* S288c (GCF_000146045.2 R64).

2.2. Identification of Orthologous Proteins

To identify orthologous proteins between each *Penicillium* proteome and the proteomes of *A. nidulans*, *N. crassa*, and *S. cerevisiae*, we used the program ProteinOrtho (V6.0.15) (Lechner et al., 2011). ProteinOrtho analyses were performed with default parameters, except by the report of singleton genes without any hit. Further, OrthoVenn2 (Xu et al., 2019) was used to identify orthologous clusters in the five proteomes and to perform GO enrichment for each cluster.

2.3. Identification of Transcription Factors

To assess TFs diversity, protein sequences of whole proteomes were used to search TF domains using InterProScan (v5.25-64.0) (Jones et al., 2014) and hmmscan (v3.1b2) (Potter et al., 2018). InterProScan was used to map Interpro families and domains, while hmmscan was used to identify domains over the PFAM database (v31.0-2017-02) (El-Gebali et al., 2019) using default parameters. Afterwards, PFAM and InterPro predictions of each species were compiled making use of the 91 DNA-binding domains described in the catalog of the main eukaryotic transcription factor families (Weirauch and Hughes, 2011), also used by CIS-BP database (Weirauch et al., 2014). TF distribution is available in **Supplementary Table 1**. Finally, a heatmap was generated including 37 of the 91 domains, which were found in the analyzed fungal proteomes (**Supplementary Material Script "funregulation_tf_heatmap.py"**).

2.4. Collection of Regulatory Interactions From Related Species

Regulatory interactions from *A. nidulans*, *N. crassa* (Hu et al., 2018), and *S. cerevisiae* available in YEASTRACT+ (Monteiro et al., 2020) were collected and organized in tab-delimited files. All interactions are available in **Supplementary Table 2**. A cross-validation was performed to check locus tag and gene name for each regulatory interaction, crossing information from the reference genomes and regulatory interactions. This step was essential to guarantee input data accuracy, especially for *S. cerevisiae* regulatory interactions, as it had already been described in other fungal models (Hu et al., 2018). It is important to note that some locus tags and/or gene names were corrected and others were not found in the *S. cerevisiae* reference genomes (SGD and Genbank), and therefore have been discarded from this study.

2.5. Upstream Sequences

Annotation in gff3 format and whole genome sequences of *P. echinulatum* 2HH and of *P. oxalicum* 114-2 were used to extract the DNA sequences comprising 1000bp upstream of each gene (**Supplementary Material Script "funregulation_promoter_extract.py"**).

2.6. Weight Matrices Used to Identify TFBS

PWMs from *A. nidulans*, *N. crassa*, and *S. cerevisiae* were obtained from CIS-BP Database (Weirauch et al., 2014). A cross-validation was also performed to check locus tag and gene name for each transcription factor, crossing information from the reference genomes and CIS-BP. Some locus tags and/or gene names were corrected, especially for *A. nidulans*.

2.7. SQLite3 Database

The large volume of data and its interconnectivity restricts the access to specific records in text files and makes it impossible to load complete files on machines with low memory capacity. Consequently, we chose a SQLite database by modeling six tables: "gene," "ortho," "pwm," "regulation," "tfbs_prediction," and "network_node." Input data obtained in the previous steps were inserted in the tables: "gene," "ortho," "pwm," and "regulation" (see **Supplementary Material**).

2.8. Inference of Global Regulatory Networks (GRN)

In order to reconstruct the GRN of *P. echinulatum* 2HH and *P. oxalicum* 114-2, a number of steps were considered, as described below (**Figure 1**).

2.8.1. Identification of TF-TG Relationships by Orthology

The identification of potential TF-TG interactions in *P. echinulatum* 2HH and in *P. oxalicum* 114-2 was performed (**Supplementary Material Script "funregulation_network_inference.py"**). This step considered the regulatory interactions from *A. nidulans*, *N. crassa*, and *S. cerevisiae* in addition to the orthology relationships previously mapped. For each known TF-TG interaction of these three

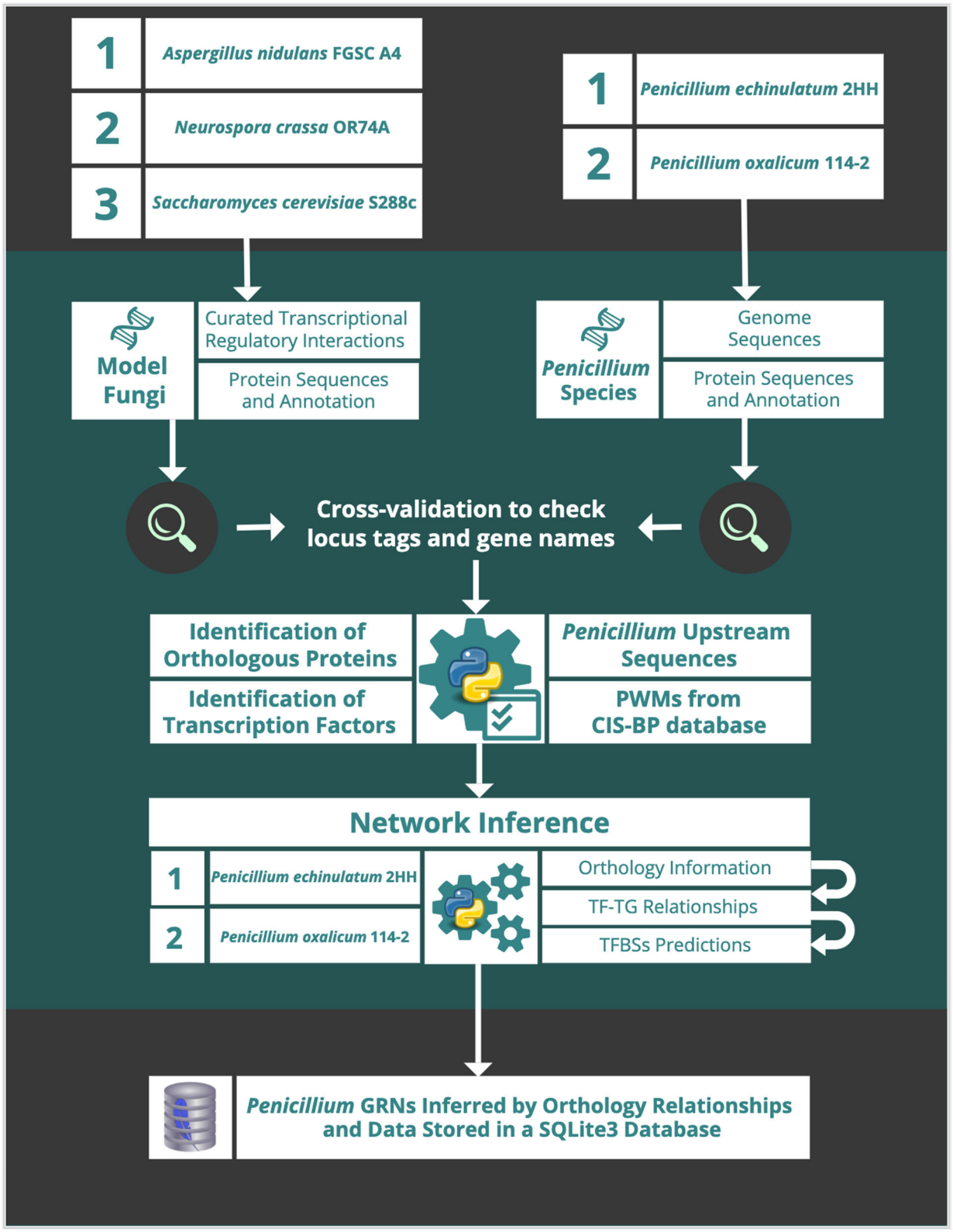


FIGURE 1 | Schematic workflow of the network inference procedure steps.

species, only when orthologous were found for both TF and TG in the *Penicillium* species, a new TF-TG interaction was created for the respective *Penicillium* species and these data were inserted in the "regulation" table. Foreign-keys cross reference from which species derived each new TF-TG interaction of the *Penicillium* species. This information allows us to analyze the conservation of TF-TG interactions between the species analyzed. Besides, it also allows us to assign different weights for the new TF-TG interactions, according to the species phylogenetic distances of which the TFs-TGs interactions were inferred.

2.8.2. TFBSs Prediction

For each TF-TG interaction, TFBS prediction was carried out. RSAT matrix-scan (Turatsinze et al., 2008) was used to predict the TFBSs using all the respective PWMs from related species, obtained from CIS-BP Database (Weirauch et al., 2014). RSAT matrix-scan analyses were performed with "cis-bp" as matrix format. Other default parameters were maintained, including an *e*-value <1e-4 as upper threshold *P*-value. RSAT results for each TF-TG interaction were stored in single text files and also in the "tfbs_prediction" table.

2.8.3. Network Nodes

For *P. echinulatum* 2HH and in *P. oxalicum* 114-2, unique TF-TG interactions were inserted in the "network_node" table, checking out from which model organisms the relationship originated and counting how many TFBS were predicted for it (**Supplementary Material Script "funregulation_network_inference.py"**). All data from the "network_node" and "tfbs_prediction" tables were exported to the tab-delimited output files.

2.8.4. Supplementary Material

Finally, the GRN inference was performed by Python scripts (Python Software Foundation, 2020) (v3.8.2) and data were stored in a SQLite database (SQLite Consortium, 2020) (v.3.31.1). All scripts and data are available in the FunRegulation project at Git http://www.github.com/alexandrelenz/funregulation.git.

2.9. Topological Analysis of the Networks

In order to topologically characterize the GRN, numerous metrics, such as node degree, clustering coefficient, centrality, hubs, and communities were determined (Junker and Schreiber, 2011). Degree of a node (K) is defined as the number of interactions that it has with other nodes. In directed networks, input (Kin) and output degree ($Kout$) are defined as the number of arrows that enter and leave from a node, respectively, which corresponds to the number of TFs that affect a certain TG, and the number of TGs that a TF regulates (Barabási and Oltvai, 2004).

Centrality (C) is a function which assigns every $v \in V$ of a given graph G, where the value $C(v) \in \Re$. Thus, to get a ranking of the node for a given G we choose the convention that a node u is more important than a node v if $C(u) > C(v)$. Lastly, we computed assorted centrality metrics, including degree, closeness, betweenness, and eigenvector centrality (Junker and Schreiber, 2011).

In a network, connectivity refers to the connections between each pair of nodes, and these connections can be via a direct or indirect link. Therefore, the connected component was defined as a set of nodes that are linked to each other by paths and give us information about how much the elements are connected in a network and their module structure (Junker and Schreiber, 2011).

To identify communities, we used the algorithm proposed by Blondel et al. (2008), that assigns a different community to each node of the network. When a node is moved to one of its neighbors' community, it achieves the highest positive contribution to modularity. This step is repeated for all nodes until no further improvement can be reached. Then, each community is considered as a single node on its own, and a subsequent move is repeated until there is only a single node left or when the modularity cannot be increased in a single step.

3. RESULTS AND DISCUSSION

3.1. Identification of Common Proteins in All the Genomes

One of the premises for knowledge transfer of regulatory relationships is a close phylogenetic positioning of the selected species, resulting in a high rate of shared orthologous genes. In this study, we use regulatory relationships of three model fungi (*A. nidulans*, *N. crassa*, and *S. cerevisiae*) for regulatory transfer to *P. echinulatum* and *P. oxalicum*. In order to analyze shared orthologous proteins, the complete proteomes were displayed into OrthoVenn2 (**Figure 2**). The Veen diagram indicates that 2,731 clusters of protein orthologous were identified as common to all proteomes, corresponding to 36.06 and 27.93% of *P. echinulatum* and *P. oxalicum* proteomes, respectively. The diagram also displays 5 clusters including 11 proteins exclusively associated to *P. echinulatum* and 30 clusters containing 70 proteins exclusively associated to *P. oxalicum*, suggesting that those proteins are species-specific.

A functional analysis of the 14,445 proteins placed in the 2,731 clusters identified as common to all proteomes, showed that DNA-dependent transcription (GO:0006351) (*p*-val: 6.13e-21) and rRNA processing (GO:0006364) (*p*-val: 1.37e-14) are the most represented GO terms. This functional analysis denotes that the core proteins shared between the five fungal species includes proteins related to cellular synthesis of RNA on a template of DNA (transcription regulator activity) and conversion of rRNA transcripts into mature rRNA molecules (rRNA maturation). These results suggest that probably the vast majority of transcription factors are conserved in the five fungal species, supporting the transference of regulatory knowledge from the model fungi to the *Penicillium* species.

3.2. The Repertoire of TFs Comprises 37 Families

The TFs (TFome) repertoire of a species comprises a set of essential proteins responsible for the regulation of gene expression in a cell. Around 80 families of TFs have been described in fungi and the proportion of transcription factors

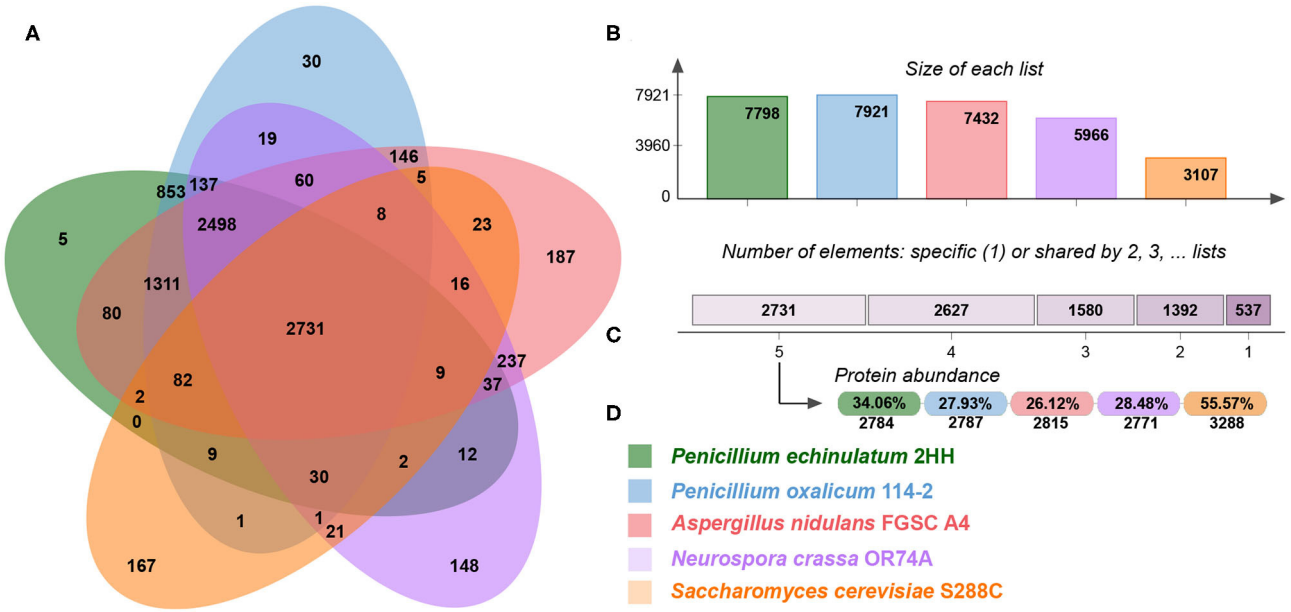


FIGURE 2 | (A) Venn diagram comparing orthologous clusters of whole-proteomes. **(B)** The barplot shows the number of orthologous clusters by organism. **(C)** The plot indicates the number of clusters that are organism-specific or shared by 2, 3, 4, or 5 organisms. **(D)** For the 2,731 clusters shared by 5 organisms, the protein abundance in percentage and absolute numbers are shown for each organism.

in genomes increases as a function of genome size, where larger genomes have more TFs. However, the increment is largely limited to three main families, Zn2Cys6 clusters, C2H2-like Zn fingers, and homeodomain-like (Shelest, 2017).

We compiled InterPro and PFAM predictions in each fungal species, taking as reference the catalog of 91 DNA-binding domains of the main eukaryotic transcription factor families (Weirauch and Hughes, 2011), also used by CIS-BP database (Weirauch et al., 2014). We found 37 TF families of the catalog in the analyzed fungal proteomes, depicting the diversity of TF families shown in **Figure 3**. These families include the most important TFs of the five species, previously described in the literature. The identified TFome of *P. echinulatum* and *P. oxalicum* include 478 and 463 TFs, respectively.

From this analysis, we found that the largest family of TFs in *P. oxalicum* and *P. uscensis* corresponds to the fungal-specific Zn2Cys6 binuclear cluster. Zn2Cys6 binuclear clusters have been identified in all fungal species analyzed so far. For example, Zn2Cys6 zinc finger proteins, such as AmyR (PDE_03964), XlnR (PDE_07674) (Li et al., 2015), and AraR (PDE_04461) (Gao et al., 2019) act in the regulation of carbon metabolism in *P. oxalicum* 114-2. Based on this knowledge, this domain could be considered ubiquitous to fungi. Based on its abundance and distribution, it has been proposed that Zn2Cys6 clusters expand much faster in ascomycetes when correlated to proteome size growth, suggesting a particular role in the evolutionary history of this phylum (Shelest, 2017). In this regard, Zn2Cys6 clusters have been found in proteins regulating a wide range of processes, including carbon and nitrogen metabolism, amino acid and vitamin synthesis, stress response, pleiotropic drug resistance, meiosis, and morphogenesis, to name but a few (MacPherson et al., 2006).

The second more abundant family identified in both *Penicillium* corresponds to the highly conserved family of C2H2 zinc fingers. The functional roles associated with members of this family are extraordinarily diverse and include DNA recognition, transcription, mRNA trafficking, cytoskeleton organization, epithelial development, chromatin remodeling, and zinc sensing, amongst others (Laity et al., 2001). In *P. oxalicum* 114-2, a C2H2 transcription factor FlbC (PDE_08372) regulates fungal asexual development and acts as an essential activator of genes encoding cellulases, hemicellulases, and other proteins with functions in lignocellulose degradation (Yao et al., 2016). Another C2H2 transcription factor BrlA (PDE_00087) has not only a key role in regulating conidiation, but it also regulates secondary metabolism extensively as well as the expression of cellulase genes (Qin et al., 2013).

At the same time, two other relevant TF families in fungi are the basic leucine zipper (bZIP) and the zinc finger GATA-type. In fungi, bZIP comprehends important regulation mechanisms, responding to oxidative stress, DNA-damage, and amino acid starvation (Tian et al., 2011). The bZIP transcription factor ClrC (PDE_09023) in *P. oxalicum* 114-2, positively regulates multiple stress responses, conidiation and the transcription levels of major cellulase genes, as well as two cellulase transcriptional activator genes, ClrB and XlnR (Lei et al., 2016). Still in *P. oxalicum* 114-2, another bZIP transcription factor CpcA (PDE_08488), is a conserved transcriptional activator for the cross-pathway control of amino acid biosynthetic genes, supporting normal growth and extracellular enzyme production under amino acid

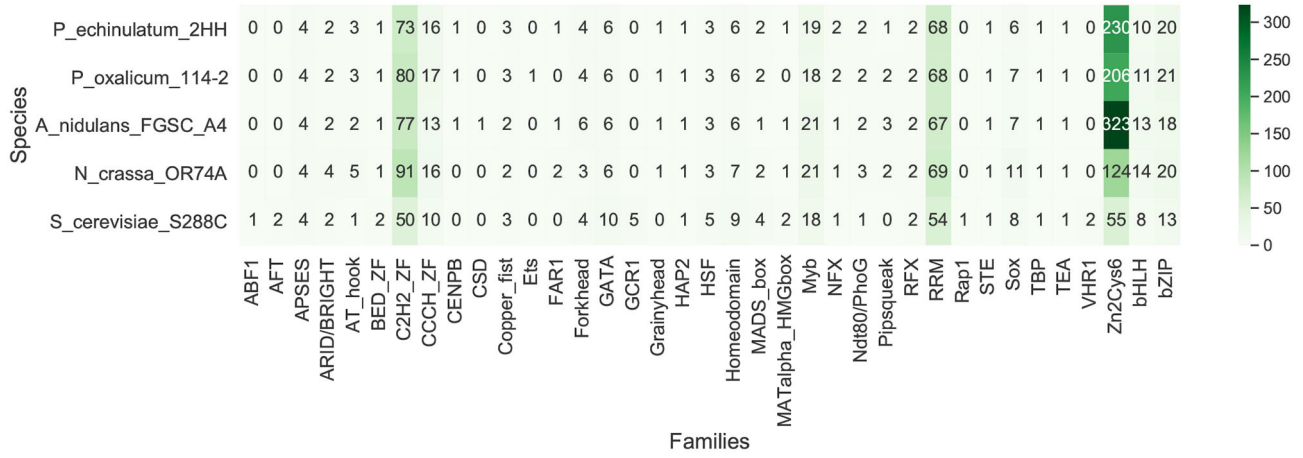


FIGURE 3 | Abundance and distribution of transcription factor families in fungal genomes. TFs of *P. echinulatum* 2HH containing multiple domains: Zn2Cys6+C2H2_ZF (5); Homeodomain+C2H2_ZF (1); BED_ZF+C2H2_ZF (1); STE+C2H2_ZF (1); CCCH_ZF+NFX (1); RRM+CCCH_ZF (3); RRM+Zn2Cys6 (1). TFs of *P. oxalicum* 114-2 containing multiple domains: Zn2Cys6+C2H2_ZF (7); Homeodomain+C2H2_ZF (1); BED_ZF+C2H2_ZF (1); STE+C2H2_ZF (1); CCCH_ZF+NFX (1); RRM+CCCH_ZF (3).

non-starvation condition (Pan et al., 2020). GATA-type fungal TFs regulate nitrogen metabolism, light induction, siderophore biosynthesis, and mating-type switching, playing global roles in growth and development (Scazzocchio, 2000). Some TFs of this family are widely studied, for example, the light-responsive WC-1 and WC-2 in *N. crassa* (Grimaldi et al., 2006), or the nitrogen regulators AreA and AreB in *A. nidulans* (Macios et al., 2012). In *P. oxalicum* HP7-1, the GATA-type transcription factor NsdD, ortholog to PDE_02029 in *P. oxalicum* 114-2, regulates the expression of major genes involved in starch, cellulose, and hemicellulose degradation, conidiation, and pigment biosynthesis (He et al., 2018). In *A. nidulans*, NsdD is an activator of sexual development and key repressor of conidiation (Han et al., 2001).

A considerable number of TFs identified in low proportions were identified in both fungal genomes, as similar to the model organisms. As examples we can name RRM, Myb, bHLH, Sox, homeobox, forkhead, APSES, HSF, AT hook, copper fist, and CAAT-binding, amongst others. Although these proteins were identified in low numbers of copies, they play important functional roles, such as fungal adaptation to host and environment described in the group of forkhead proteins (Wang et al., 2015) and fungal differentiation and secondary metabolism (Son et al., 2020). In addition, diverse families have been proposed as specific to fungi, such as APSES and copper fist, because they were found exclusively in fungal genomes (Shelest, 2017).

In summary, the TF families identified in both *Penicillium* species are very similar to the fungal genomes used as reference; where three families (Zn2Cys6, C2H2 ZF, and RRM) include around 75% of the total of the DNA-binding domains identified in *P. echinulatum* and *P. oxalicum* genomes. When the role of these core families was explored in the fungi used as reference, central processes were found, such as carbon and nitrogen metabolism, amino acid and vitamin synthesis, growth and development (Scazzocchio, 2000), and fungal differentiation and secondary metabolism, among others.

3.3. General Properties of the Regulatory Network

The GRN in the *Penicillium* species was inferred considering orthology information and curated regulatory interactions from the model organisms *A. nidulans*, *N. crassa*, and *S. cerevisiae*. A regulatory interaction was defined as a TF-TG relation, where TF is the regulator gene and TG is the target gene. For each regulatory interaction of the model fungi, when orthologous were found for both TF and TG in the *Penicillium* species, a new TF-TG relation for the respective *Penicillium* species was created. Besides the TF-TG orthology, we also performed TFBSs predictions for each TF-TG identified. Based on these data, we constructed the global regulatory networks of *P. echinulatum* and *P. oxalicum*.

Therefore, the GRN of *P. echinulatum* shown in **Figure 4A** contains 5,862 nodes and 21,184 regulatory interactions. Based on the TF-TG orthology, 96 TFs and 5,853 TGs were identified in the GRN, and that covers 71.7% of the *P. echinulatum* proteome. From these 96 TFs, 87 are also TGs; i.e., they could be self-regulated or regulated by another TF. The vast majority of regulatory interactions inferred for *P. echinulatum* came from orthology related to only one model fungus. This group totalized 21,067 interactions, of which 5,962 resulted from *A. nidulans*, 10,723 derived from *N. crassa*, and 4,382 originated from *S. cerevisiae*. Another group of 115 regulatory interactions inferred for *P. echinulatum* came from orthology found in two model fungi, of which 90 interactions came from curated regulatory interactions found in both *A. nidulans* and *N. crassa*; other three interactions were derived from *A. nidulans* and *S. cerevisiae*; and 22 interactions came from *N. crassa* and *S. cerevisiae*. Finally,

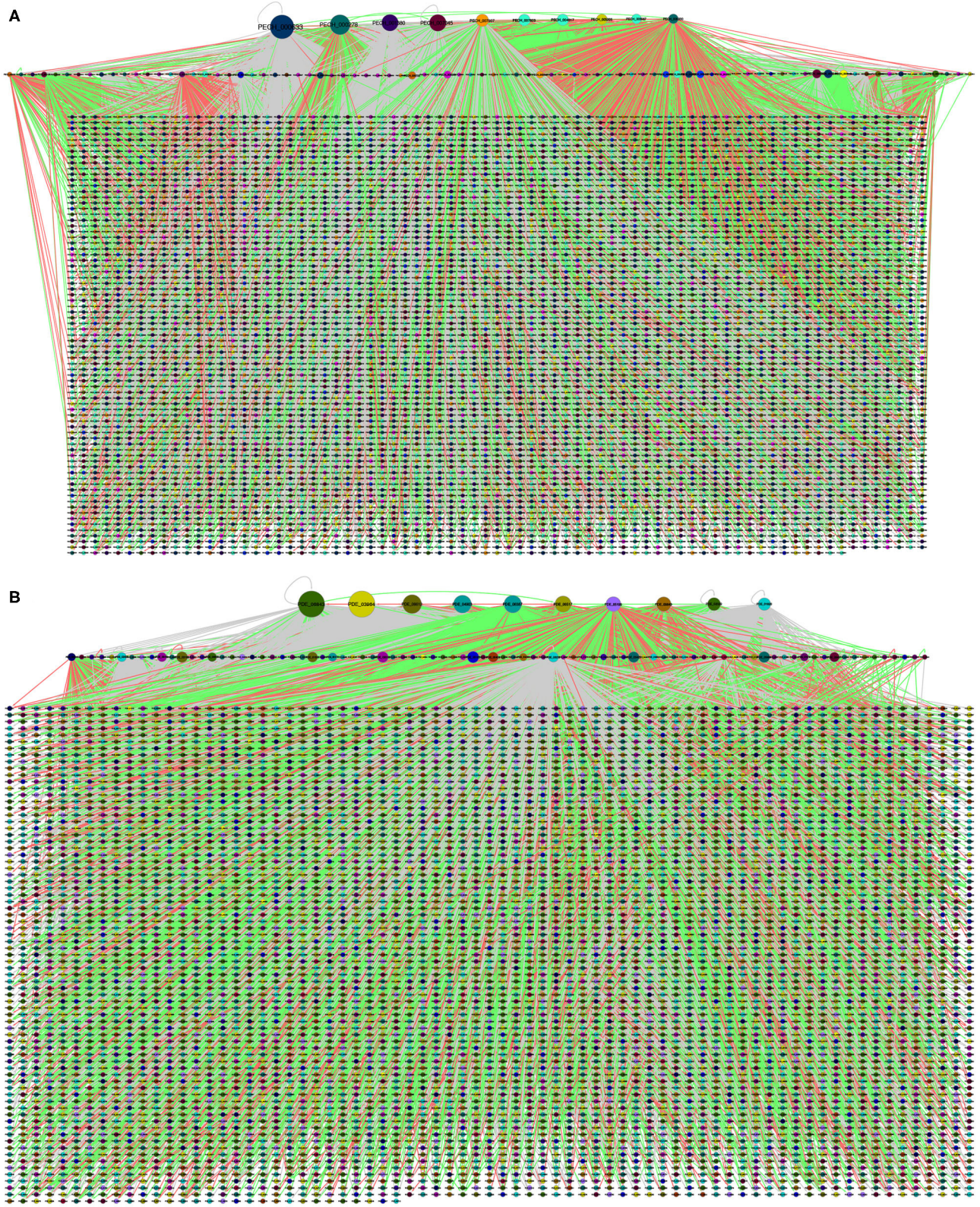


FIGURE 4 | Global regulatory networks of **(A)** *P. echinulatum* 2HH and **(B)** *P. oxalicum* 114-2. Each community is represented in a different color. The node size is proportional to the output degree. In the first level, we show the hub nodes, in the second level the remained TFs and finally, the target nodes.

TABLE 1 | General properties of the regulatory networks.

	***P. echinulatum* 2HH**	***P. oxalicum* 114-2**
Total of nodes	5,862	5,528
Interactions	21,184	16,775
Auto-regulations	23	19
Positive edges	8,078	6,901
Negative edges	4,697	3,952
Unknown edges	8,409	5,922
Average degree	7.2	6.0
Connected components	1	1
Giant component	5,862	5,528
Maximum out degree	2,502 (PECH_000633)	1619 (PDE_06843)
Maximum in degree	24 (PECH_007435)	24 (PDE_00087)
Communities	20	19

two interactions were inferred by orthology of curated regulatory interactions that are conserved in the three model fungi.

A similar behavior was observed in the inference of the *P. oxalicum* GRN shown in **Figure 4B** that contains 5,528 nodes and 16,775 regulatory interactions, of which 99 are TFs and 5,516 are TGs. From these 99 TFs, 86 are also TGs. This inferred GRN covers 55.4% of the *P. oxalicum* proteome. The vast majority of regulatory interactions also came from orthology related to only one model fungus. This group totalize 16,685 interactions, of which 6,099 resulted from *A. nidulans*, 8,677 derived from *N. crassa*, and 1,909 originated from *S. cerevisiae*. The second group of 90 regulatory interactions came from orthology found in two model fungi, of which 87 interactions came from *A. nidulans* and *N. crassa*; two interactions derived from *A. nidulans* and *S. cerevisiae*; and one interaction came from *N. crassa* and *S. cerevisiae*. Finally, no interactions were inferred for *P. oxalicum* by orthology of curated regulatory interactions that are conserved in the three model fungi.

3.4. Topological Properties of the Regulatory Network

In order to characterize the structure of GRN of *Penicillium* species, the general structure of both networks was analyzed in **Table 1**. We identified that the networks are structured into a single giant component in which there is a path between each pair of nodes.

Taking these data into account, we identified that 1,404 nodes in *P. echinulatum* and 1,776 nodes in *P. oxalicum* are regulated by only one TF, i.e., they have an input degree of 1; while the BrlA (PDE_00087) and its orthologous PECH_007435 are regulated by 24 TFs in *P. oxalicum* and *P. echinulatum*, respectively, making them the most regulated genes. BrlA (PDE_00087), identified as a member of the C2H2 transcription factor family, plays a key role in regulating conidiation, affecting also the regulation of secondary metabolism and the expression of cellulase genes (Qin et al., 2013). Concerning the output degree, the most connected nodes are influencing 2,502 and 1,619 nodes, representing 42 and 29% of total nodes in the GRNs of *P. echinulatum* and *P. oxalicum*, respectively.

We also found that the highest clustering coefficient is 1, meaning that nodes whose neighbors are connected between them are forming complete graphs. We identified this property for 78 nodes in *P. echinulatum* and 36 in *P. oxalicum*, suggesting the existence of substructures, such as triangles or more complex motifs. On the other hand, 2,143 nodes in *P. echinulatum* and 2,902 in *P. oxalicum* have a clustering coefficient equal to 0; whereas the average clustering coefficient for the network was 0.20 for *P. echinulatum* and 0.13 for *P. oxalicum*. This result indicates that neighbors have, on average, $\frac{1}{5}$ of connections for *P. echinulatum* and $< \frac{1}{5}$ for *P. oxalicum*. In this regard, when the clustering coefficient is large, a small world network structure can be described, which is not the case analyzed here.

3.5. Identification of Communities on the GRN

In order to identify the most connected elements, we analyzed the network in terms of communities. A community was defined as a subset of nodes densely connected in comparison with the rest of the network, and its identification may help to discover relations not previously identified (Radicchi et al., 2004). Based on this approach, we identified 20 communities in the GRN of *P. echinulatum*, where the largest one contains 1,170 nodes and the smallest one contains 38 nodes; while for the GRN of *P. oxalicum*, 19 communities were identified, where the largest one contains 706 nodes and the smallest one contains 23 nodes.

To determine the most abundant function, each community was analyzed with the Gene Ontology (GO) terms enrichment (**Figure 5**). Based on this approach, we identified that community 4 for *P. oxalicum* and 19 for *P. echinulatum* are enriched of genes related to cellular process, metabolic process and localization. On the other hand, communities 5, 11, and 16 in *P. oxalicum* and the community 19 in *P. echinulatum* are the most diverse, where the most abundant functions are catalytic activity, binding, and transporter activity (**Figure 5**).

Therefore, we found that communities 19 of *P. echinulatum* and 16 of *P. oxalicum* share a high proportion of orthologous proteins, and similarities at molecular function level. In contrast, community 4 of *P. oxalicum* contains orthologs from various communities of *P. echinulatum*, such as 19, 17, and 4. In addition, communities 3 and 11 in *P. echinulatum* and community 2 in *P. oxalicum* contain genes involved in the xylanolytic and cellulolytic transcriptional activator systems, among others. In summary, we not only identified that both *Penicillium* species share similar communities at sequence and functional levels, but also, particular communities in each species, which shows their diversity.

3.6. Mining the Regulatory Network

In order to identify the most connected nodes associated with the network, hubs were identified. Hence, a hub was defined as a TF with connections to many other nodes, i.e., a large output degree.

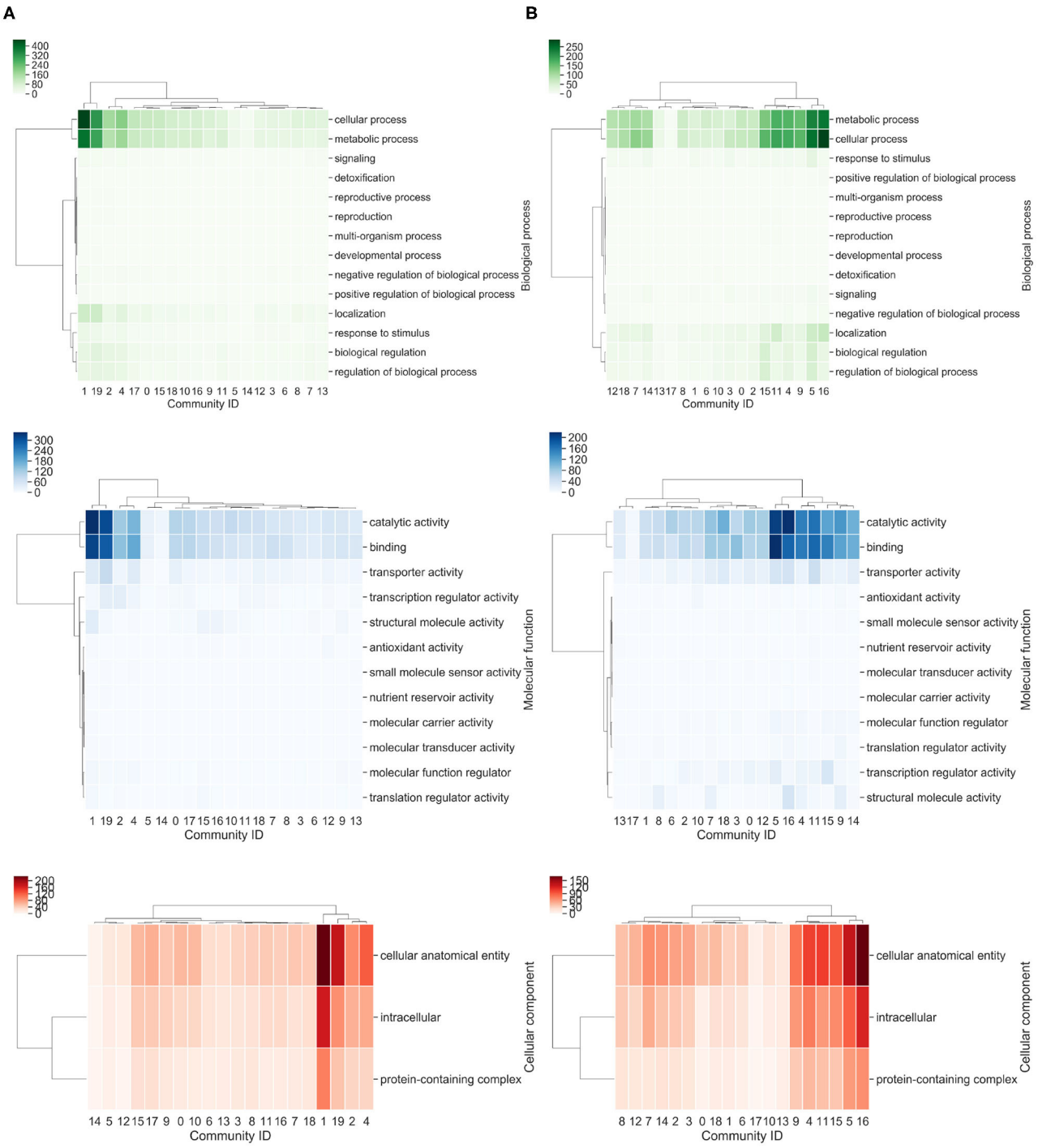


FIGURE 5 | Heatmap of Gene Ontology terms associated with communities. **(A)** *P. echinulatum* 2HH and **(B)** *P. oxalicum* 114-2. The richest GO terms for each community were identified. A hierarchical clustered based on Euclidean distance measure and Ward's method for linkage analysis was achieved. Each row represents the GO term and each column represents community ID.

Therefore, we showed the ten top hubs in *P. echinulatum* and *P. oxalicum* (**Table 2**).

In the GRN of *P. echinulatum*, the most connected node is PECH_000633 that consists of 2,502 inferred interactions. This bZIP transcription factor is orthologous to CpcA of *A. nidulans*, CPC-1 of *N. crassa*, and Gcn4p of *S. cerevisiae*. In *S. cerevisiae*, Gcn4p stimulates the transcription of 12 different pathways related genes, and also genes encoding various aminoacyl-tRNA

TABLE 2 | Top 10 of hub nodes.

P. echinulatum 2HH		*P. oxalicum* 114-2			
Transcription factor	**Out degree**	**Transcription factor**	**Out degree**	**TF family**	**Regulated process**
PECH_000633 (CpcA*)	2502	PDE_08488 (CpcA*)	#	bZIP	Protein synthesis.
PECH_000278 (COL-26**)	2011	PDE_04455 (COL-26**)	#	Zn2Cys6	Starch utilization.
PECH_007045 (FF-7**)	1611	PDE_06843 (FF-7**)	1619	Zn2Cys6	Sexual development.
PECH_001380 (AmyR*)	1595	PDE_03964 (AmyR*)	1612	Zn2Cys6	Starch utilization.
PECH_007907 (LreB*)	1053	PDE_08612 (LreB*)	1046	GATA	Blue-light responsive differentiation.
PECH_007823 (AtfA*)	957	PDE_04903 (AtfA*)	972	bZIP	Oxidative and osmotic stress-responsive genes.
PECH_004317 (FlbB*)	932	PDE_06387 (FlbB*)	950	bZIP_YAP	Conidiophore development.
PECH_005206 (RES-1**)	844	PDE_06517 (RES-1**)	857	C2H2	Endoplasmic reticulum stress response.
PECH_006987 (MetZ*)	723	PDE_05199 (MetZ*)	745	bZIP	Sulfur metabolism.
PECH_000930 (CCG-8**)	715	PDE_09849 (CCG-8**)	723	TF_Opi1	Biological processes (Clock-controlled gene).
		PDE_02029 (NsdD*)	609	GATA	Conidiation and cell wall stress resistance.
		PDE_01826 (GAL4***)	574	Zn2Cys6	Galactose-inducible genes.

*TFs are ordered according to the out degree value. Proteins from columns 1 and 3 are orthologous. Gene names are shown in brackets according to the orthologous proteins of *A. nidulans. **N. crassa. ***S. cerevisiae. #Incorrect annotation, orthology not identified by our approach. See alignment in* ***Supplementary Data Sheet 1****.*

synthetases and pathway-specific activators. This cross-pathway regulatory network of amino acid biosynthesis is known as GAAC in yeast and CPC in *Neurospora* and *Aspergillus* (Hoffmann et al., 2001; Hinnebusch, 2005). Recently, Gcn4p was described as a central regulator of protein synthesis, holding a key role in stress response and longevity. The reduction of the protein synthesis capacity by this regulator extends yeast lifespan (Mittal et al., 2017). The control of amino acid starvation by CpcA in *A. nidulans* also regulates sexual development, revealing a connection between metabolism and sexual development in filamentous fungi (Hoffmann et al., 2000). Therefore, PECH_000633 could be considered as a central regulatory protein associated with fundamental physiological processes, similar to Gcn4p, and CPC transcription factors. The conservation of this important regulatory system suggests subsequent studies regarding the fungal lifespan, given that longer lifetime is highly important to improve the production of cellulolytic enzymes by *Penicillium* spp.

The second most connected node of the *P. echinulatum* GRN is PECH_000278 with 2,011 regulatory interactions. This Zn2Cys6 binuclear cluster transcription factor is orthologous to COL-26 of *N. crassa*. In *N. crassa*, COL-26 is necessary for the expression of amylolytic genes and is required for the utilization of maltose and starch. This TF also acts as a regulator of glucose metabolism and its loss causes resistance to carbon catabolite repression, affecting integration of carbon and nitrogen metabolisms (Xiong et al., 2017). Several regulatory interactions of biotechnological interest were inferred for PECH_000278, covering the major cellulolytic enzymes of *P. echinulatum*: intracellular β-glucosidase (PECH_005648), cellobiohydrolase (PECH_007386), endoglucanase EGL1 (PECH_009029), β-glucosidases (PECH_005824 and PECH_002471), and xylanases (PECH_006995 and PECH_007282). In addition, regulatory interactions related to the major amylolytic genes were inferred for PECH_000278, including α-amylases (PECH_008724 and PECH_000987), α-glucosidases (PECH_001379 [GH31] and PECH_005310 [GH15]), and a lytic starch monooxygenase (PECH_007113). These inferred regulatory interactions suggest that PECH_000278 is a promising target for industrial strains improvement, due to its role in sugar metabolism regulation.

In *P. oxalicum*, our approach could not identify the orthologous of these two most connected TFs, highlighted in the GRN of *P. echinulatum*. The blastp querying PDE_08488 against UniprotKB showed the following results: AN3675-CpcA of *A. nidulans* (*e*-value: 3.4e-36; ident.: 46.0%; coverage length: 100%, NCU04050-CPC-1 of *N. crassa* (*e*-value: 1.3e-16; ident.: 31.7%; coverage length: 100%) and YEL009C-GCN4 of *S. cerevisiae* (*e*-value: 2e-10; ident.: 29.9%; coverage length: 100%). The results showed that the length of PDE_08488 is quite larger when compared to the reviewed CpcA, CPC-1, and GCN4. Our results are in agreement with the previously identified incorrect annotation of CpcA in *P. oxalicum* (Pan et al., 2020). In a blastp using PDE_04455 as query against UniprotKB, the best hit is PENSUB_363 (*P. subrubescens*) (*e*-value: 0.0; ident.: 58.8%; coverage length: 88.1%). In PDE_04455, only the domain IPR007219 was identified, different from their homologs in closely related *Penicillium* species. The blastp result also suggests that its transcription start site is probably misannotated and possibly there is a Zn2Cys6 binuclear cluster upstream of PDE_04455, as occurs in *P. subrubescens* and *P. echinulatum* (PECH_000278). These results suggest that both PDE_08488 and PDE_04455 are misannotated in *P. oxalicum*.

As shown in **Table 2**, the third most connected node in the GRN of *P. echinulatum* is orthologous to the first most connected node in the GRN of *P. oxalicum*. The subsequent most connected nodes in the two GRNs follow this orthological relationship, up to the tenth most connected node in the GRN of *P. echinulatum* that corresponds to the eighth most connected node as its ortholog in the GRN of *P. oxalicum*.

The most connected node (PDE_06843) in the GRN of *P. oxalicum*, is a Zn2Cys6 binuclear cluster transcription factor, ortholog to the third most connected node (PECH_007045)

in the GRN of *P. echinulatum* and also ortholog to FF-7 in *N. crassa*. In *N. crassa*, the female fertility-7 (FF-7) is required for initiation of sexual development, and Δ*ff-7* mutant does not produce protoperithecia and perithecia as well as ascospores (Carrillo et al., 2017). In *P. oxalicum* HP7-1, it was demonstrated that POX09752, ortholog of PDE_06843 in *P. oxalicum* 114-2, is involved in the regulation of raw-starch-digesting enzymes (Zhang et al., 2019). This is consistent with the inferred regulatory interactions where PDE_06843 regulates one α-amylase encoding gene PDE_04683 (GH13-5) in *P. oxalicum*; and PECH_007045 regulates two α-amylase encoding genes PECH_000987 (GH13-5) and PECH_000986 (GH13-1) in *P. echinulatum*.

The next hub node refers to AmyR orthologs, also belonging to the Zn2Cys6 binuclear cluster family. Its orthologs are PECH_001380 and PDE_03964 in *P. echinulatum* and *P. oxalicum*, respectively. AmyR has been described as a positive regulator of amylase encoding genes, involved in starch utilization in *Aspergillus* species (Benocci et al., 2017). Its orthologs were found in several *Ascomycetes*, and its function has been substantially studied in *P. oxalicum*, in which the Δ*amyR* mutant resulted in a substantial increase of cellulase activity (Li et al., 2015).

In this regard, the Δ*amyR* mutant of *P. oxalicum* was reported as deficient for transcribing the major raw starch-digesting glucoamylase gene *gluA* (PDE_09417) when grown on cellulose. The lack of AmyR also affects a wide range of CAZymes involved in the starch metabolism (Li et al., 2015). In the GRN of *P. oxalicum*, AmyR exhibited 1,612 inferred regulatory interactions, including some of the major downregulated CAZymes involved in the starch metabolism of the Δ*amyR* mutant: PDE_01201 α-amylase Amy13A (GH13-1), PDE_09417 glucoamylase GluA (GH15), PDE_03966 α-glucosidase (GH31), and PDE_01354 lytic starch monooxygenase (AA13). In addition, our inferred regulatory interactions also include some of the major cellulase encoding genes, upregulated in the Δ*amyR* mutant: PDE_07124 cellobiohydrolase (GH6), PDE_09226 endoglucanase (GH5-5), and PDE_04251 β-glucosidase (GH3).

In *P. echinulatum* GRN, 1,595 regulatory interactions were predicted for AmyR, including all ortholog genes related to starch and cellulose metabolisms described for *P. oxalicum*. As discussed above, our results are in agreement with the reported regulatory role of AmyR in *Aspergillus* spp. and *P. oxalicum*, affecting positively the expression of starch-related enzymes and negatively the expression of cellulose-related enzymes.

Although various TFs were identified as hubs in the GRNs, additional transcription factors not identified in the top ten of highly connected nodes according to the metric of output degree are relevant to the regulation of the cellulolytic system, such as XlnR, ClrB, and CreA. ClrB is a Zn2Cys6 zinc finger transcription factor, known for its role in the positive regulation of cellulolytic enzymes in filamentous fungi. In *P. oxalicum* 114-2, we identified that ClrB (PDE_05999) could regulate 472 genes; and experimental evidence showed its positive regulating role of major cellulolytic enzymes (Li et al., 2015). Inferred regulatory interactions found in our GRN of *P. oxalicum* are in agree with previous experimental evidence, including cellobiohydrolases (PDE_07124, PDE_07945, and PDE_05445), endoglucanases (PDE_09226, PDE_03711, PDE_05193, PDE_07929, and PDE_02886), LPMOs (PDE_06768 and PDE_01261), and BGL2 (PDE_00579). Similarly, in *P. echinulatum*, ClrB (PECH_005720) could regulate 479 genes, including cellobiohydrolases (PECH_006365, PECH_008028, and PECH_007386), endoglucanases (PECH_009029, PECH_007371, and PECH_005815), LPMOs (PECH_008064 and PECH_007161), and BGL2 (PECH_005648).

In contrast, CreA plays a key role in the carbon catabolite repression regulatory system which prevents wasting energy on the production of extracellular enzymes, as well as metabolic routes that are not needed. Therefore, this C2H2 transcription factor represses the expression of cellulolytic genes. In *P. oxalicum* 114-2, CreA (PDE_03168) represses the expression of major cellulases (Li et al., 2015). These previous experimental evidences are in accordance with the regulatory interactions found in our GRN of *P. oxalicum* 114-2 (464 regulated genes), including cellobiohydrolases (PDE_07124 and PDE_07945), endoglucanases (PDE_09226, PDE_03711, PDE_05193, and PDE_07929), LPMOs (PDE_06768 and PDE_01261), and BGL2 (PDE_00579). Similarly, in the GRN of *P. echinulatum* 2HH we found that CreA (PECH_004563) regulates 457 genes; where interactions covering cellobiohydrolases (PECH_006365 and PECH_007386), endoglucanases (PECH_009029 and PECH_007371), LPMOs (PECH_008064 and PECH_007161), and BGL2 (PECH_005648), were identified.

The xylanolytic and cellulolytic transcriptional activator XlnR of *P. echinulatum* contains a repetitive sequence in the coding region, resulting in fragments placed in two different scaffolds of the WGS. However, the complete protein was used in this study, once this gene was sequenced and deposited at GenBank (accession number: MT676450 and locus tag: PECH_002137). This TF comprehends 250 inferred regulatory interactions in *P. echinulatum* while its ortholog PDE_07674 in *P. oxalicum* include 251 interactions. As expected, quite a few regulatory interactions involving xylanolytic enzymes were inferred for both species, including the major xylosidases of the CAZy families GH3 (2 genes) and GH43 (2 genes), the main xylanases of families GH10 (1 gene) and GH11 (2 genes) and the major xylose transporters (2 genes). The major XlnR-dependent proteins involved in the xylanolytic and cellulolytic systems are detailed in **Supplementary Table 3** and presented in **Figure 6**.

In order to provide accuracy to the TF-TG relationships inferred for XlnR, predictions of TFBSs were performed in the upstream sequences of each TG. Our approach used the PWM matrix of XLR-1 (M02621_2.00), the XlnR ortholog found in *N. crassa*. For the major β-xylosidase (GH3), despite a significant difference in the composition of nucleotides, two conserved motifs were predicted in both species: XlnR-PECH_007407 predicted (CGGCTAATA) and (CGGTTACGT); XlnR-PDE_00049 predicted (TGTATATAT) and (TATATATAC). For both fungi, no motifs were predicted for the TF-TG relationships of the second β-xylosidase (GH3): XlnR-PECH_003595 and XlnR-PDE_08037. For the xylanase of the GH10 family, the same motif (CGGCTAAAA) was identified for both fungi relationships: XlnR-PECH_004799 and

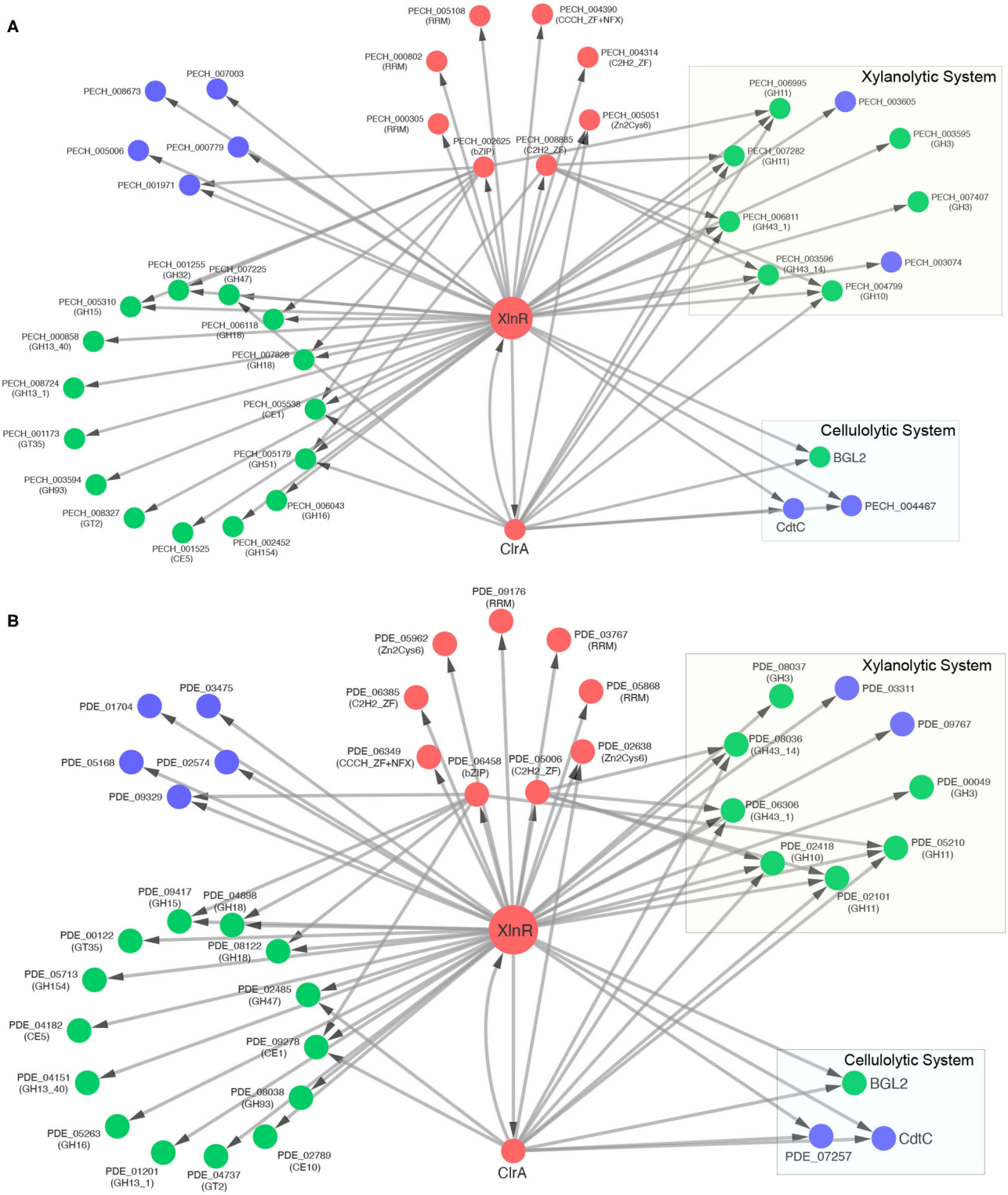


FIGURE 6 | Regulatory networks of XlnR-TGs containing only CAZymes, TFs, and sugar transporters. **(A)** *P. echinulatum* 2HH and **(B)** *P. oxalicum* 114-2. CAZymes are colored in green, TFs in red, and sugar transporters in blue. Rectangles highlight the major CAZymes and sugar transporters involved in xylanolytic and cellulolytic systems. Arrows point the regulatory direction. The regulatory relationship between XlnR and ClrA is bidirectional.

XlnR-PDE_02418. The DNA-binding site associated to XlnR in *Penicillium* was predicted by using the PWM matrix of XLR-1 (M02621 2.00), This weight matrix identified diverse probable sites recognized by this TF, such as the sequence CGGCTAATA and CGGTTACGT of XlnR-PECH_007407 and TGTATATAT and TATATATA for XlnR-PDE_00049. One of these motifs is similar to the identified in the *gsn* gene 5'-flanking region of *N. crassa* (GGCTGA) (Gonçalves et al., 2011). However, we must remember that our approach considers a PWM matrix that is a model for the binding specificity of a TF and can be used to scan a sequence for the presence of DNA sites that are significantly more similar to the PWM than to the background (Stormo, 2013). The complete prediction dataset is provided in **Supplementary Table 3**.

For the first xylanase of the GH11 family, the motif (CGGGTAAAT) was predicted for XlnR-PECH_006995 in *P. echinulatum*, while for XlnR-PDE_05210 in *P. oxalicum* no motifs were predicted. Aligning these promoter regions, we observed that the last "T" of the predicted motif in *P. echinulatum* is an "A" in *P. oxalicum*. In contrast, for the second xylanase of the GH11 family, there was found a motif (CGGATAAAT) only for XlnR-PDE_02101 in *P. oxalicum*, while for XlnR-PECH_007282 it was not predicted. For the β-xylosidase (GH43-14) a XlnR binding-site (CGGTTAACG) was predicted for PECH_003596 in *P. echinulatum*, while in *P. oxalicum* this binding-site was not found for PDE_08036. In contrast, for the β-xylosidase (GH43-1) XlnR binding-sites were predicted for both species. In *P. echinulatum*, (CGGTTAATT) was predicted for PECH_006811 and, in *P. oxalicum*, the motif (CGGCTAAAC) was predicted for PDE_06306. For the xylose transporters XlnR binding motifs were not found in both species.

Regulatory interactions for the major cellulolytic enzymes were not inferred for XlnR. Nevertheless, XlnR includes inferred regulatory interactions for some regulators of cellulase expression previously described. In *P. oxalicum*, ClrA (PDE_04046) is a Zn2Cys6 positive regulator of cellulase transcription (Liu et al., 2013c), orthologous to PECH_001863, CLR-1 and ClrA in *P. echinulatum*, *N. crassa*, and *Aspergillus* spp., respectively. Our approach found a XlnR binding-site (ACTTATACT) in the upstream region of *clrA* in *P. oxalicum*, while no motifs were predicted for *clrA* in *P. echinulatum*.

Cellobiose is the primary end product generated from cellulose degradation by cellulolytic enzymes. It has been shown that cellulase production is induced by cellobiose and other cellodextrins in many species of fungi (Aro et al., 2005). We inferred a regulatory interaction of XlnR to the major intracellular β-glucosidase BGL2 (PECH_005648/PDE_00579). However, no XlnR binding-sites were predicted for *bgl2* in both *P. oxalicum* and *P. echinulatum*. In *P. oxalicum*, BGL2 has been previously identified as a negative regulator of cellulase expression, considering that the Δ*bgl2* mutant raised remarkably the secretion of cellulolytic enzymes (Chen et al., 2013).

Besides BGL2, cellulase expression is also affected by cellodextrin transporters. Two major XlnR-dependent cellodextrin transporters were inferred for both fungi. The first cellodextrin transporter CdtC (PDE_00607) increased cellulase production in *P. oxalicum* mutant when overexpressed, denoting that CdtC played a positive regulatory role (Li et al., 2013). CdtC is ortholog to PECH_005610 in *P. echinulatum*, LacpB-CltB in *A. nidulans* (Dos Reis et al., 2016; Fekete et al., 2016) and CDT-1 in *N. crassa* (Znameroski et al., 2012). A XlnR binding-site (GGTATATAA) was predicted for *cdtC* in *P. echinulatum*, while in *P. oxalicum* this binding-site was not found. However, when both promoter regions (PECH_005610 and PDE_00607) were aligned, the same motif was found in *P. oxalicum*. We suggest that this RSAT misprediction could be a false negative. Further, the orthologs of *N. crassa* cellodextrin transporter CDT-2 (Znameroski et al., 2012) were also inferred as XlnR-dependent in the GRNs of *P. echinulatum* (PECH_004467) and *P. oxalicum* (PDE_07257). Lastly, for CDT-2 orthologs, no XlnR binding-sites were predicted.

Our results suggest that XlnR may play a background regulatory role in the expression of the cellulolytic system, acting through the cellodextrin induction system. On the other hand, XlnR plays a key role in positive regulation of the xylanolytic system. This remarkable finding of our study suggests that the XlnR-dependent cellulolytic and xylanolytic regulatory systems are probably conserved in both *P. oxalicum* and *P. echinulatum*. Our results are in consonance with *P. oxalicum* experimental evidence previously reported and discussed below.

The Δ*xlnR* mutant of *P. oxalicum* revealed low expressions for cellobiohydrolase *cbh1* (PDE_07945) and endoglucanase *eg2* (PDE_09226) transcripts; and no expression for xylanase *xyn1* (PDE_08094) when compared with the wild-type in cellulose growth medium. Besides other experimental evidence, it was suggested that XlnR is a general TF that facilitates the induction of cellulase expression under cellulose growth conditions. XlnR might participate in a transcriptional cascade that regulates the expression of genes coding for cellulolytic enzymes in *P. oxalicum* (Li et al., 2015).

The overexpression of *xlnR* was expected to upregulate the expression of the major extracellular β-xylosidase xyl3A (PDE_00049) in *P. oxalicum*. This was confirmed by OE*xlnR*Δ*laeA* mutant, which induced a remarkable 28.5 fold increase in the expression of *xyl3A* in relation to the wild-type. This result also showed that the regulation of *xyl3A* is not LaeA-dependent as the regulatory system of most of the cellulases or hemicellulases (Li et al., 2016).

In summary, the GRNs of *P. echinulatum* 2HH and *P. oxalicum* 114-2 showed similar components, such as those TFs identified as hubs according to the metric of output degree. In **Table 2**, the hub nodes were ordered according to TF orthology to highlight the high similarity between the two GRNs. Similar to the orthology identified among the hub nodes of both GRNs, most TGs regulated by each TF also preserve the orthology relationship in both GRNs. For example, 1526 TGs regulated by AmyR in *P. echinulatum* have their respective orthologous regulated by AmyR in *P. oxalicum*, while 69 are unique to the GRN of *P. echinulatum* and 86 are unique to the GRN of *P. oxalicum*. In this regard, both GRNs are highly similar, considering that both share

mainly ortholog TFs and regulatory relationships for mainly ortholog TGs.

4. CONCLUSIONS

Our reconstructed network is a valuable resource of regulatory interactions occurring within *Penicillium* spp., and it may integrate with global expression data available for these fungal organisms in order to improve global interaction data models. In this regard, we found a group of proteins shared among the two *Penicillium* and three model fungi, involved in transcription and rRNA processing, i.e., genetic information flow. In addition, we demonstrate, through our analysis, the existence of large protein sets devoted to regulate gene expression in these fungal systems, where three families (Zn2Cys6, C2H2 ZF, and RRM) comprehend around 75% of the total of the DNA-binding domains identified in both fungal genomes. These genes are involved in regulation of central processes, carbon and nitrogen metabolism, amino acid and vitamin synthesis, growth and development, among others. Concerning the GRN, we found similar topological properties identified in other biological networks, highlighting the existence of at least ten global regulators. To name but a few of them in *P. echinulatum*, PECH_000633 could be considered fundamental in control of amino acid starvation, longevity and sexual development; PECH_000278 could be highly involved in carbon and nitrogen metabolisms; and PECH_007045 could be involved in the regulation of raw-starch-digesting enzymes. Finally, we explore diverse identified modules, such as the expression of cellulolytic and xylanolytic systems, where XlnR plays a key role in positive regulation of the xylanolytic system. It also regulates positively the cellulolytic system by acting indirectly through the cellodextrin induction system. This remarkable finding of our study suggests that the XlnR-dependent cellulolytic and xylanolytic regulatory systems are probably conserved in both *P. oxalicum* and *P. echinulatum*. Our results are in consonance with *P. oxalicum* experimental evidence previously reported. Finally, we explored the functional congruency on the genes clustered in terms of communities where the genes related to cellular nitrogen, compound metabolic process and macromolecule metabolic process were the most abundant.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/**Supplementary Material** or available in the FunRegulation project at Git http://www.github.com/alexandrelenz/funregulation.git.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

We are grateful to the Coordination for the Improvement of Higher Education Personnel (CAPES) for the PhD scholarship (88887.158496/2017-00 to ARL). This research was supported by grants from Dirección General de Asuntos del Personal Académico-Universidad Nacional Autónoma de Mexico (UNAM) (IN-209620), CAPES (3255/2013), and the National Council for Scientific and Technological Development (CNPq) (472153/2013-7). MC and AJPD are CNPq Research Fellowship. We are grateful to Bahia State University (UNEB) for the leave of absence (3.145/2016 to ARL) and financial support.

ACKNOWLEDGMENTS

The authors would like to thank CAPES, CNPq, UNEB, UNAM, and UCS.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.588263/full#supplementary-material

Supplementary Table 1 | Distribution of transcription factor domains.

Supplementary Table 2 | Curated transcriptional regulatory interactions in model fungi.

Supplementary Table 3 | Major XlnR-dependent proteins involved in the xylanolytic and cellulolytic systems.

Supplementary Data Sheet 1 | Alignment of incorrect annotated proteins of *Penicillium oxalicum* 114-2.

REFERENCES

Aro, N., Pakula, T., and Penttilä, M. (2005). Transcriptional regulation of plant cell wall degradation by filamentous fungi. *FEMS Microbiol. Rev.* 29, 719–739. doi: 10.1016/j.femsre.2004.11.006

Barabási, A.-L., and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113. doi: 10.1038/nrg1272

Benocci, T., Aguilar-Pontes, M. V., Zhou, M., Seiboth, B., and de Vries, R. P. (2017). Regulators of plant biomass degradation in ascomycetous fungi. *Biotechnol. Biofuels* 10:152. doi: 10.1186/s13068-017-0841-x

Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.* 2008:P10008. doi: 10.1088/1742-5468/2008/10/P10008

Camassola, M., De Bittencourt, L. R., Shenem, N. T., Andreaus, J., and Dillon, A. J. P. (2004). Characterization of the cellulase complex of *Penicillium echinulatum*. *Biocatal. Biotransform.* 22, 391–396. doi: 10.1080/10242420400024532

Camassola, M., and Dillon, A. J. (2010). Cellulases and xylanases production by *Penicillium echinulatum* grown on sugar cane bagasse in solid-state fermentation. *Appl. Biochem. Biotechnol.* 162, 1889–1900. doi: 10.1007/s12010-010-8967-3

Camassola, M. and Dillon, A. J. P. (2007a). Effect of methylxanthines on production of cellulases by *Penicillium echinulatum*. *J. Appl. Microbiol.* 102, 478–485. doi: 10.1111/j.1365-2672.2006.03098.x

Camassola, M., and Dillon, A. J. P. (2007b). Production of cellulases and hemicellulases by *Penicillium echinulatum* grown on pretreated sugar cane bagasse and wheat bran in solid-state fermentation. *J. Appl. Microbiol.* 103, 2196–2204. doi: 10.1111/j.1365-2672.2007.03458.x

Camassola, M., and Dillon, A. J. P. (2009). Biological pretreatment of sugar cane bagasse for the production of cellulases and xylanases by *Penicillium echinulatum*. *Indus. Crops Prod.* 29, 642–647. doi: 10.1016/j.indcrop.2008.09.008

Camassola, M., and Dillon, A. J. P. (2012). Steam-exploded sugar cane bagasse for on-site production of cellulases and xylanases by *Penicillium echinulatum*. *Energy Fuels* 26, 5316–5320. doi: 10.1021/ef3009162

Carrillo, A. J., Schacht, P., Cabrera, I. E., Blahut, J., Prudhomme, L., Dietrich, S., et al. (2017). Functional profiling of transcription factor genes in *Neurospora crassa*. *G3* 7, 2945–2956. doi: 10.1534/g3.117.043331

Carter, H., Hofree, M., and Ideker, T. (2013). Genotype to phenotype via network analysis. *Curr. Opin. Genet. Dev.* 23, 611–621. doi: 10.1016/j.gde.2013.10.003

Chen, D., Yan, W., Fu, L.-Y., and Kaufmann, K. (2018). Architecture of gene regulatory networks controlling flower development in *Arabidopsis thaliana*. *Nat. Commun.* 9:4534. doi: 10.1038/s41467-018-06772-3

Chen, M., Qin, Y., Cao, Q., Liu, G., Li, J., Li, Z., et al. (2013). Promotion of extracellular lignocellulolytic enzymes production by restraining the intracellular β-glucosidase in Penicillium decumbens. *Bioresour. Technol.* 137, 33–40. doi: 10.1016/j.biortech.2013.03.099

Dillon, A. J., Bettio, M., Pozzan, F. G., Andrighetti, T., and Camassola, M. (2011). A new *Penicillium echinulatum* strain with faster cellulase secretion obtained using hydrogen peroxide mutagenesis and screening with 2-deoxyglucose. *J. Appl. Microbiol.* 111, 48–53. doi: 10.1111/j.1365-2672.2011.05026.x

Dillon, A. J., Zorgi, C., Camassola, M., and Henriques, J. A. P. (2006). Use of 2-deoxyglucose in liquid media for the selection of mutant strains of *Penicillium echinulatum* producing increased cellulase and β-glucosidase activities. *Appl. Microbiol. Biotechnol.* 70, 740–746. doi: 10.1007/s00253-005-0122-7

Dos Reis, L., Fontana, R. C., da Silva Delabona, P., da Silva Lima, D. J., Camassola, M., da Cruz Pradella, J. G., et al. (2013). Increased production of cellulases and xylanases by *Penicillium echinulatum* S1M29 in batch and fed-batch culture. *Bioresour. Technol.* 146, 597–603. doi: 10.1016/j.biortech.2013.07.124

Dos Reis, T. F., De Lima, P. B. A., Parachin, N. S., Mingossi, F. B., De Castro Oliveira, J. V., Ries, L. N. A., et al. (2016). Identification and characterization of putative xylose and cellobiose transporters in *Aspergillus nidulans*. *Biotechnol. Biofuels* 9:204. doi: 10.1186/s13068-016-0611-1

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., et al. (2019). The Pfam protein families database in 2019. *Nucleic Acids Res.* 47, D427–D432. doi: 10.1093/nar/gky995

Fekete, E., Orosz, A., Kulcsár, L., Kavalecz, N., Flipphi, M., and Karaffa, L. (2016). Characterization of a second physiologically relevant lactose permease gene (lacpB) in *Alespergillus nidulans*. *Microbiology* 162, 837–847. doi: 10.1099/mic.0.000267

Fernandez-Valverde, S. L., Aguilera, F., and Ramos-Díaz, R. A. (2018). Inference of developmental gene regulatory networks beyond classical model systems: new approaches in the post-genomic era. *Integr. Compar. Biol.* 58, 640–653. doi: 10.1093/icb/icy061

Filho, F. M., do Nascimento, A. P. B., dos Santos, M. T., Carvalho-Assef, A. P. D., and da Silva, F. A. B. (2019). Gene regulatory network inference and analysis of multidrug-resistant *Pseudomonas aeruginosa*. *bioRxiv* 610493. doi: 10.1101/610493

Gabaldón, T., and Koonin, E. V. (2013). Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.* 14, 360–366. doi: 10.1038/nrg3456

Gao, L., Li, S., Xu, Y., Xia, C., Xu, J., Liu, J., et al. (2019). Mutation of a conserved alanine residue in transcription factor AraR leads to hyperproduction of α-l-arabinofuranosidases in *Penicillium oxalicum*. *Biotechnol. J.* 14:1800643. doi: 10.1002/biot.201800643

Gasch, A. P., Moses, A. M., Chiang, D. Y., Fraser, H. B., Berardini, M., and Eisen, M. B. (2004). Conservation and evolution of cis-regulatory systems in ascomycete fungi. *PLoS Biol.* 2:e398. doi: 10.1371/journal.pbio.0020398

Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., et al. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature* 489, 91–100. doi: 10.1038/nature11245

Glenwinkel, L., Wu, D., Minevich, G., and Hobert, O. (2014). TargetOrtho: a phylogenetic footprinting tool to identify transcription factor targets. *Genetics* 197, 61 LP–76. doi: 10.1534/genetics.113.160721

Gonçalves, R. D., Cupertino, F. B., Freitas, F. Z., Luchessi, A. D., and Bertolini, M. C. (2011). A genome-wide screen for *Neurospora crassa* transcription factors regulating glycogen metabolism. *Mol. Cell. Proteomics* 10:M111.007963. doi: 10.1074/mcp.M111.007963

Grimaldi, B., Coiro, P., Filetici, P., Berge, E., Dobosy, J. R., Freitag, M., et al. (2006). The *Neurospora crassa* White Collar-1 dependent blue light response requires acetylation of histone H3 lysine 14 by NGF-1. *Mol. Biol. Cell* 17, 4576–4583. doi: 10.1091/mbc.e06-03-0232

Han, K.-H., Han, K.-Y., Yu, J.-H., Chae, K.-S., Jahng, K.-Y., and Han, D.-M. (2001). The nsdD gene encodes a putative GATA-type transcription factor necessary for sexual development of *Aspergillus nidulans*. *Mol. Microbiol.* 41, 299–309. doi: 10.1046/j.1365-2958.2001.02472.x

Hassani-Pak, K., and Rawlings, C. (2017). Knowledge discovery in biological databases for revealing candidate genes linked to complex phenotypes. *J. Integr. Bioinformatics* 14:20160002. doi: 10.1515/jib-2016-0002

He, Q.-P., Zhao, S., Wang, J.-X., Li, C.-X., Yan, Y.-S., Wang, L., et al. (2018). Transcription factor NsdD regulates the expression of genes involved in plant biomass-degrading enzymes, conidiation, and pigment biosynthesis in *Penicillium oxalicum*. *Appl. Environ. Microbiol.* 84:e01039-18. doi: 10.1128/AEM.01039-18

Hinnebusch, A. G. (2005). Translational regulation of Gcn4 and the general amino acid control of yeast. *Annu. Rev. Microbiol.* 59, 407–450. doi: 10.1146/annurev.micro.59.031805.133833

Hoffmann, B., Valerius, O., Andermann, M., and Braus, G. H. (2001). Transcriptional autoregulation and inhibition of mRNA translation of amino acid regulator gene cpcA of filamentous fungus *Aspergillus nidulans*. *Mol. Biol. Cell* 12, 2846–2857. doi: 10.1091/mbc.12.9.2846

Hoffmann, B., Wanke, C., LaPaglia, S. K., and Braus, G. H. (2000). c-Jun and RACK1 homologues regulate a control point for sexual development in *Aspergillus nidulans*. *Mol. Microbiol.* 37, 28–41. doi: 10.1046/j.1365-2958.2000.01954.x

Hu, J., Chen, C., Huang, K., and Mitchell, T. K. (2013). A distribution pattern assisted method of transcription factor binding site discovery for both yeast and filamentous fungi. *Adv. Biosci. Biotechnol.* 4, 509–517. doi: 10.4236/abb.2013.44067

Hu, Y., Qin, Y., and Liu, G. (2018). Collection and curation of transcriptional regulatory interactions in *Aspergillus nidulans* and *Neurospora crassa* reveal structural and evolutionary features of the regulatory networks. *Front. Microbiol.* 9:27. doi: 10.3389/fmicb.2018.02713

Huber, W., Carey, V. J., Long, L., Falcon, S., and Gentleman, R. (2007). Graphs in molecular biology. *BMC Bioinformatics* 8(Suppl. 6):S8. doi: 10.1186/1471-2105-8-S6-S8

Jackson, C. A., Castro, D. M., Saldi, G. A., Bonneau, R., and Gresham, D. (2020). Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments. *eLife* 9:e51254. doi: 10.7554/eLife.51254

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Junker, B. H., and Schreiber, F., (eds.). (2011). *Analysis of Biological Networks*. Hoboken, NJ: Wiley Online Books. John Wiley & Sons, Inc.

Karlebach, G., and Shamir, R. (2008). Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell Biol.* 9, 770–780. doi: 10.1038/nrm2503

Koch, C., Konieczka, J., Delorey, T., Lyons, A., Socha, A., Davis, K., et al. (2017). Inference and evolutionary analysis of genome-scale regulatory networks in large phylogenies. *Cell Syst.* 4, 543–558.e8. doi: 10.1016/j.cels.2017.04.010

Kulkarni, S. R., Vaneechoutte, D., Van de Velde, J., and Vandepoele, K. (2017). TF2Network: predicting transcription factor regulators and gene regulatory networks in Arabidopsis using publicly available binding site information. *Nucleic Acids Res.* 46:e31. doi: 10.1101/173559

Laity, J. H., Lee, B. M., and Wright, P. E. (2001). Zinc finger proteins: new insights into structural and functional diversity. *Curr. Opin. Struct. Biol.* 11, 39–46. doi: 10.1016/S0959-440X(00)00167-6

Lam, K. Y., Westrick, Z. M., Müller, C. L., Christiaen, L., and Bonneau, R. (2016). Fused regression for multi-source gene regulatory network inference. *PLoS Comput. Biol.* 12:e1005157. doi: 10.1371/journal.pcbi.1005157

Lambert, S. A., Yang, A. W. H., Sasse, A., Cowley, G., Albu, M., Caddick, M. X., et al. (2019). Similarity regression predicts evolution of transcription factor sequence specificity. *Nat. Genet*. 51, 981–989. doi: 10.1038/s41588-019-0411-1

Lechner, M., Findeiß, S., Steiner, L., Marz, M., Stadler, P. F., and Prohaska, S. J. (2011). Proteinortho: detection of (Co-)orthologs in large-scale analysis. *BMC Bioinformatics* 12:124. doi: 10.1186/1471-2105-12-124

Lei, Y., Liu, G., Yao, G., Li, Z., Qin, Y., and Qu, Y. (2016). A novel bZIP transcription factor ClrC positively regulates multiple stress responses, conidiation and cellulase expression in *Penicillium oxalicum*. *Res. Microbiol*. 167, 424–435. doi: 10.1016/j.resmic.2016.03.001

Li, J., Liu, G., Chen, M., Li, Z., Qin, Y., and Qu, Y. (2013). Cellodextrin transporters play important roles in cellulase induction in the cellulolytic fungus *Penicillium oxalicum*. *Appl. Microbiol. Biotechnol*. 97, 10479–10488. doi: 10.1007/s00253-013-5301-3

Li, Y., Zheng, X., Zhang, X., Bao, L., Zhu, Y., Qu, Y., et al. (2016). The different roles of *Penicillium oxalicum* LaeA in the production of extracellular cellulase and β-xylosidase. *Front. Microbiol*. 7:2091. doi: 10.3389/fmicb.2016.02091

Li, Z., Yao, G., Wu, R., Gao, L., Kan, Q., Liu, M., et al. (2015). Synergistic and dose-controlled regulation of cellulase gene expression in *Penicillium oxalicum*. *PLoS Genet*. 11:e1005509. doi: 10.1371/journal.pgen.1005509

Liu, G., Qin, Y., Hu, Y., Gao, M., Peng, S., and Qu, Y. (2013a). An endo-1,4-β-glucanase PdCel5C from cellulolytic fungus *Penicillium decumbens* with distinctive domain composition and hydrolysis product profile. *Enzyme Microb. Technol*. 52, 190–195. doi: 10.1016/j.enzmictec.2012.12.009

Liu, G., Qin, Y., Li, Z., and Qu, Y. (2013b). Improving lignocellulolytic enzyme production with Penicillium: from strain screening to systems biology. *Biofuels* 4, 523–534. doi: 10.4155/bfs.13.38

Liu, G., Zhang, L., Qin, Y., Zou, G., Li, Z., Yan, X., et al. (2013c). Long-term strain improvements accumulate mutations in regulatory elements responsible for hyper-production of cellulolytic enzymes. *Sci. Rep*. 3:1569. doi: 10.1038/srep01569

Liu, G., Zhang, L., Wei, X., Zou, G., Qin, Y., Ma, L., et al. (2013d). Genomic and secretomic analyses reveal unique features of the lignocellulolytic enzyme system of *Penicillium decumbens*. *PLoS ONE* 8:e55185. doi: 10.1371/journal.pone.0055185

Macios, M., Caddick, M. X., Weglenski, P., Scazzocchio, C., and Dzikowska, A. (2012). The GATA factors AREA and AREB together with the co-repressor NMRA, negatively regulate arginine catabolism in *Aspergillus nidulans* in response to nitrogen and carbon source. *Fungal Genet. Biol*. 49, 189–198. doi: 10.1016/j.fgb.2012.01.004

MacPherson, S., Larochelle, M., and Turcotte, B. (2006). A fungal family of transcriptional regulators: the zinc cluster proteins. *Microbiol. Mol. Biol. Rev*. 70, 583–604. doi: 10.1128/MMBR.00015-06

Mercatelli, D., Scalambra, L., Triboli, L., Ray, F., and Giorgi, F. M. (2020). Gene regulatory network inference resources: a practical overview. *Biochim. Biophys. Acta* 1863:194430. doi: 10.1016/j.bbagrm.2019.194430

Mittal, N., Guimaraes, J. C., Gross, T., Schmidt, A., Vina-Vilaseca, A., Nedialkova, D. D., et al. (2017). The Gcn4 transcription factor reduces protein synthesis capacity and extends yeast lifespan. *Nat. Commun*. 8:457. doi: 10.1038/s41467-017-00539-y

Monteiro, P. T., Oliveira, J., Pais, P., Antunes, M., Palma, M., Cavalheiro, M., et al. (2020). YEASTRACT+: a portal for cross-species comparative genomics of transcription regulation in yeasts. *Nucleic Acids Res*. 48, D642–D649. doi: 10.1093/nar/gkz859

Novello, M., Vilasboa, J., Schneider, W. D. H., Reis, L. D., Fontana, R. C., and Camassola, M. (2014). Enzymes for second generation ethanol: exploring new strategies for the use of xylose. *RSC Adv*. 4, 21361–21368. doi: 10.1039/c4ra00909f

Pan, Y., Gao, L., Zhang, X., Qin, Y., Liu, G., and Qu, Y. (2020). The role of cross-pathway control regulator CpcA in the growth and extracellular enzyme production of *Penicillium oxalicum*. *Curr. Microbiol*. 77, 49–54. doi: 10.1007/s00284-019-01803-8

Penfold, C. A., Millar, J. B. A., and Wild, D. L. (2015). Inferring orthologous gene regulatory networks using interspecies data fusion. *Bioinformatics* 31, i97–i105. doi: 10.1093/bioinformatics/btv267

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., and Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Res*. 46, W200–W204. doi: 10.1093/nar/gky448

Python Software Foundation (2020). *Python*.

Qin, Y., Bao, L., Gao, M., Chen, M., Lei, Y., Liu, G., et al. (2013). Penicillium decumbens BrlA extensively regulates secondary metabolism and functionally associates with the expression of cellulase genes. *Appl. Microbiol. Biotechnol*. 97, 10453–10467. doi: 10.1007/s00253-013-5273-3

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D. (2004). Defining and identifying communities in networks. *Proc. Natl. Acad. Sci. U.S.A*. 101, 2658–2663. doi: 10.1073/pnas.0400054101

Ribeiro, D. A., Cota, J., Alvarez, T. M., Brüchli, F., Bragato, J., Pereira, B. M., et al. (2012). The *Penicillium echinulatum* secretome on sugar cane bagasse. *PLoS ONE* 7:e50571. doi: 10.1371/journal.pone.0050571

Rubini, M. R., Dillon, A. J., Kyaw, C. M., Faria, F. P., Poças-Fonseca, M. J., and Silva-Pereira, I. (2010). Cloning, characterization and heterologous expression of the first *Penicillium echinulatum* cellulase gene. *J. Appl. Microbiol*. 108, 1187–1198. doi: 10.1111/j.1365-2672.2009.04528.x

Scazzocchio, C. (2000). The fungal GATA factors. *Curr. Opin. Microbiol*. 3, 126–131. doi: 10.1016/S1369-5274(00)00063-1

Schneider, W. D. H., Dos Reis, L., Camassola, M., and Dillon, A. J. P. (2014). Morphogenesis and production of enzymes by *penicillium echinulatum* in response to different carbon sources. *BioMed Res. Int*. 2014:10. doi: 10.1155/2014/254863

Schneider, W. D. H., Fontana, R. C., Baudel, H. M., de Siqueira, F. G., Rencoret, J., Gutiérrez, A., et al. (2020). Lignin degradation and detoxification of eucalyptus wastes by on-site manufacturing fungal enzymes to enhance second-generation ethanol yield. *Appl. Energy* 262:114493. doi: 10.1016/j.apenergy.2020.114493

Schneider, W. D. H., Gonçalves, T. A., Uchima, C. A., Couger, M. B., Prade, R., Squina, F. M., et al. (2016). *Penicillium echinulatum* secretome analysis reveals the fungi potential for degradation of lignocellulosic biomass. *Biotechnol. Biofuels* 9:66. doi: 10.1186/s13068-016-0476-3

Schneider, W. D. H., Gonçalves, T. A., Uchima, C. A., dos Reis, L., Fontana, R. C., Squina, F. M., et al. (2018). Comparison of the production of enzymes to cell wall hydrolysis using different carbon sources by *Penicillium echinulatum* strains and its hydrolysis potential for lignocelullosic biomass. *Process Biochem*. 66, 162–170. doi: 10.1016/j.procbio.2017.11.004

Shelest, E. (2017). Transcription factors in fungi: TFome dynamics, three major families, and dual-specificity TFs. *Front. Genet*. 8:53. doi: 10.3389/fgene.2017.00053

Son, S.-H., Son, Y.-E., Cho, H.-J., Chen, W., Lee, M.-K., Kim, L.-H., et al. (2020). Homeobox proteins are essential for fungal differentiation and secondary metabolism in *Aspergillus nidulans*. *Sci. Rep*. 10:6094. doi: 10.1038/s41598-020-63300-4

SQLite Consortium (2020). *SQLite*.

Staunton, P. M., Miranda-CasoLuengo, A. A., Loftus, B. J., and Gormley, I. C. (2019). BINDER: computationally inferring a gene regulatory network for *Mycobacterium abscessus*. *BMC Bioinformatics* 20:466. doi: 10.1186/s12859-019-3042-8

Stormo, G. D. (2013). Modeling the specificity of protein-DNA interactions. *Quant. Biol*. 1, 115–130. doi: 10.1007/s40484-013-0012-4

Tian, C., Li, J., and Glass, N. L. (2011). Exploring the bZIP transcription factor regulatory network in *Neurospora crassa*. *Microbiology* 157(Pt 3), 747–759. doi: 10.1099/mic.0.045468-0

Turatsinze, J.-V., Thomas-Chollier, M., Defrance, M., and van Helden, J. (2008). Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nat. Protoc*. 3, 1578–1588. doi: 10.1038/nprot.2008.97

Vaishnav, N., Singh, A., Adsul, M., Dixit, P., Sandhu, S. K., Mathur, A., et al. (2018). Penicillium: the next emerging champion for cellulase production. *Bioresour. Technol. Rep*. 2, 131–140. doi: 10.1016/j.biteb.2018.04.003

Wang, J.-J., Qiu, L., Cai, Q., Ying, S.-H., and Feng, M.-G. (2015). Transcriptional control of fungal cell cycle and cellular events by Fkh2, a forkhead transcription factor in an insect pathogen. *Sci. Rep*. 5:10108. doi: 10.1038/srep10108

Weirauch, M. T., and Hughes, T. R. (2011). "A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution," in *A Handbook of Transcription Factors*, ed T. R. Hughes (Dordrecht: Springer Netherlands), 25–73. doi: 10.1007/978-90-481-9069-0_3

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158, 1431–1443. doi: 10.1016/j.cell.2014.08.009

Xiong, Y., Wu, V. W., Lubbe, A., Qin, L., Deng, S., Kennedy, M., et al. (2017). A fungal transcription factor essential for starch degradation affects integration of carbon and nitrogen metabolism. *PLoS Genet*. 13:e1006737. doi: 10.1371/journal.pgen.1006737

Xu, L., Dong, Z., Fang, L., Luo, Y., Wei, Z., Guo, H., et al. (2019). OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res*. 47, W52–W58. doi: 10.1093/nar/gkz333

Yao, G., Li, Z., Gao, L., Wu, R., Kan, Q., Liu, G., et al. (2015). Redesigning the regulatory pathway to enhance cellulase production in *Penicillium oxalicum*. *Biotechnol. Biofuels* 8:71. doi: 10.1186/s13068-015-0253-8

Yao, G., Li, Z., Wu, R., Qin, Y., Liu, G., and Qu, Y. (2016). *Penicillium oxalicum* PoFlbC regulates fungal asexual development and is important for cellulase gene expression. *Fungal Genet. Biol*. 86, 91–102. doi: 10.1016/j.fgb.2015.12.012

Zhang, M.-Y., Zhao, S., Ning, Y.-N., Fu, L.-H., Li, C.-X., Wang, Q., et al. (2019). Identification of an essential regulator controlling the production of raw-starch-digesting glucoamylase in *Penicillium oxalicum*. *Biotechnol. Biofuels* 12:7. doi: 10.1186/s13068-018-1345-z

Znameroski, E. A., Coradetti, S. T., Roche, C. M., Tsai, J. C., Iavarone, A. T., Cate, J. H., et al. (2012). Induction of lignocellulose-degrading enzymes in *Neurospora crassa* by cellodextrins. *Proc. Natl. Acad. Sci. U.S.A*. 109, 6012–6017. doi: 10.1073/pnas.1118440109

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

CAZyome and Sugar Transportome Unveil Singular Aspects of the Lignocellulolytic Enzyme System of *Penicillium echinulatum* 2HH

Alexandre Rafael Lenz[a,c,d,*], Eduardo Balbinot[a], Nikael Souza de Oliveira[a,b], Fernanda Pessi de Abreu[a], Scheila de Avila e Silva[a], Marli Camassola[b], Ernesto Perez-Rueda[d], Aldo José Pinheiro Dillon[b]

[a] *Laboratório de Bioinformática e Biologia Computacional, Instituto de Biotecnologia, Universidade de Caxias do Sul (UCS). Rua Francisco Getúlio Vargas, 1130, 95070-560, Caxias Do Sul, RS, Brasil.*
[b] *Laboratório de Enzimas e Biomassas, Instituto de Biotecnologia, Universidade de Caxias do Sul (UCS). Rua Francisco Getúlio Vargas, 1130, 95070-560, Caxias Do Sul, RS, Brasil.*
[c] *Departmento de Ciências Exatas e da Terra, Universidade do Estado da Bahia (UNEB). Rua Silveira Martins, 2555, 41150-000, Salvador, BA, Brasil.*
[d] *Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas. Unidad Académica Yucatán. Universidad Nacional Autónoma de México (UNAM), C.P. 97302, Mérida, Yucatán, México.*

Abstract

Penicillium echinulatum 2HH is a widely studied ascomycete, known by its efficient cellulolytic cocktails. Understanding both lignocellulolytic and sugar uptake systems is essential to obtain industrial strains with adequate efficiency for bioethanol production. We report a comprehensive *in silico* characterization of CAZymes and sugar transporters of *P. echinulatum* 2HH. The CAZyome reveals an outstanding repertoire of enzymes involved in plant biomass

*Corresponding author
Email addresses: alenz@uneb.br/arlenz@ucs.br (Alexandre Rafael Lenz), ebalbinot@ucs.br (Eduardo Balbinot), nsoliveira4@ucs.br (Nikael Souza de Oliveira), fpabreu1@ucs.br (Fernanda Pessi de Abreu), sasilva6@ucs.br (Scheila de Avila e Silva), mcamasso@ucs.br (Marli Camassola), ernesto.perez@iimas.unam.mx (Ernesto Perez-Rueda), ajpdillo@ucs.br (Aldo José Pinheiro Dillon)
URL: https://orcid.org/0000-0001-6699-2899 (Alexandre Rafael Lenz), https://orcid.org/0000-0002-1322-3153 (Eduardo Balbinot), https://orcid.org/0000-0003-4713-4527 (Nikael Souza de Oliveira), https://orcid.org/0000-0002-5006-4833 (Fernanda Pessi de Abreu), https://orcid.org/0000-0002-3472-3907 (Scheila de Avila e Silva), https://orcid.org/0000-0001-7410-4337 (Marli Camassola), https://orcid.org/0000-0002-6879-0673 (Ernesto Perez-Rueda), https://orcid.org/0000-0002-1969-1740 (Aldo José Pinheiro Dillon)
[1]# The authors contributed equally to this work.
[2]Running title: *Lenz et al / Cellulolytic Enzyme System of Penicillium echinulatum*

degradation. Among them, we highlight the cellulolytic enzyme system, whose genes are predominantly orthologous to *Penicillium oxalicum* 114-2, demonstrating the high similarity of these phylogenetically related enzyme producers. We also report a LPMO-type enzyme of the AA16 family described for the first time in these fungi. In addition, we found peculiarities in the gene composition of enzymes required to degrade wood, reinforcing the hypothesis of environment-specific adaptations in *P. echinulatum* 2HH during a potential long-term mutualistic symbiosis with coleoptera larvae. Our phylogenetic analysis of the sugar transportome suggests that *P. echinulatum* 2HH diversity and specificity of STs include eight major families with specificity to different groups of sugars. Finally, our phylogenetic analyses enabled the identification of several iBGLs and STs potentially involved in the accumulation of intracellular cellodextrins. Overall, both CAZyome and sugar transportome of *P. echinulatum* 2HH revealed new insights into the mechanisms underlying a flexible and highly functional metabolism to degrade plant biomass. Furthermore, the first phylogenetic classification of STs and iBGLs shed new light into the role of these genes regarding the preferred carbon source during fungal growth. Along these lines, these iBGLs and STs comprise novel gene targets to understand the regulatory mechanisms underlying cellulolytic enzymes and to design hypersecreting strains with adequate efficiency for industry.

Keywords: lignocellulolytic, CAZy, sugar transporter, cellobiose, cellulase

1. Introduction

Second-generation bioethanol produced by using lignocellulosic feedstock has proved to be a cutting-edge technology in optimizing the production of biofuels. Renewable sources used for generating bioethanol exhibit a particularly rich lignocellulose content. The high cost of bioethanol production from second-generation feedstocks results from the pretreatment and enzymatic hydrolysis processes, required to convert this lignocellulosic biomass into monomeric sugars more rapidly and with greater yields Dalena et al. (2019). Some *Penicillium*

species have been highlighted due to its superiority over existing enzyme producers, especially due to the production of balanced cellulolytic cocktails rich in β-glucosidases. Consequently, these enzymatic mixtures result in improved enzymatic hydrolysis yields Vaishnav et al. (2018).

P. echinulatum is widely studied for biofuels production from lignocellulosic biomass, mainly agricultural residues, such as sugar-cane bagasse and elephant grass Ribeiro et al. (2012); Camassola and Dillon (2012); Scholl et al. (2015); Menegol et al. (2016). *P. echinulatum* has been studied for about 40 years, since the isolation of the 2HH wild-type. Long-term strain improvements resulted in the S1M29 mutant that yields an expressive increase in cellulase titers and provides a better lignocellulosic biomass hydrolysis Schneider et al. (2016, 2018).

Filamentous fungi genome, secretome and transcriptome analyses are quite relevant for enzymatic cocktails design, particularly for biofuel production. The design of hypersecreting strains is essential to achieve adequate efficiency of enzyme mixtures, considering the established potential of *P. echinulatum* 2HH for the production of cellulolytic enzymes. In summary, it is necessary to understand the different strategies employed by *P. echinulatum* 2HH to degrade lignocellulosic biomass, enabling the enhancement of enzyme production. The ability of fungi to grow, to transport, and to ferment different types of sugars remains a major challenge for biofuels production from lignocellulosic biomass Hyde et al. (2019). The plant lignocellulosic biomass is primarily made up of the glucohomopolysaccharide cellulose (20-50%, w/w), hemicellulose (15-35%, w/w) and lignin. The polysaccharides that constitute the hemicellulose include xylan, glucuronoxylan, xyloglucan, glucomannan and arabinoxylan backbones with heterogeneous side-chains. The use of monosaccharides that constitute plant biomass polymers implies their efficient hydrolysis, which is still a major technical challenge because of its recalcitrance and heterogeneity Druzhinina and Kubicek (2017).

P. echinulatum 2HH was first isolated from the digestive-tract of coleoptera larvea, commonly known as furniture beetles. As the name suggests, it is known to feed on wood and has the potential to reduce wooden objects to fine dust.

Moreover, *Anobium punctatum* larvae normally live in coarse wood debris, which are weakened and predigested fallen tree trunks, allowing the larva to move through the cracks Wheeler and Crowson (1982). Assuming that lignin, cellulose and hemicellulose work together to provide a structural function in plants and lignin is responsible for stiffness and rigidity Glass et al. (2013), it can be stated that the larvae diet is basically composed by cellulose and hemicellulose, and may contain some lignin residues. In this sense, the first enzymatic production experiments with 2HH strain raised evolutionary speculations, suggesting a possible long-term mutualistic symbiosis between the 2HH strain and *A. punctatum* larvae.

Cellulose degradation is mediated by the cellulolytic enzyme system, widely used for biofuel production Panchapakesan and Shankar (2016). Besides that, different biopolymers require specific enzymatic systems for their degradation, such as starch-degrading enzymes responsible for starch degradation and xylan-degrading enzymes responsible for hemicellulose major component depolymerization. Enzymes that degrade, modify or create glycosidic bonds are known as CAZymes and are classified by the Carbohydrate-Active Enzyme (CAZy) database. These enzymes are organized into different families, according to their amino acid sequence and structural similarity: i) Glycoside Hydrolases (GHs) are responsible for hydrolysis and / or rearrangement of glycosidic bonds; ii) Glycosyl Transferases (GTs) are responsible for the formation of glycosidic bonds; iii) Polysaccharide Lyases (PLs) perform non-hydrolytic cleavage of glycosidic bonds; iv) Carbohydrate Esterases (CEs) hydrolyze carbohydrate esters; v) Auxiliary Activities (AAs) are redox enzymes that act synergically with other CAZymes; and vi) Carbohydrate Binding Modules (CBMs) promote the adhesion of the enzyme to the carbohydrate Lombard et al. (2014).

The cellulolytic system comprises a variety of enzymes that act synergistically: (i) cellobiohydrolases (CBHI) (GH7) cleave at the reducing ends of the cellulose chain; (ii) cellobiohydrolases (CBHII) (GH6) cleave at the non-reducing ends of the cellulose chain; (iii) endoglucanases (EGL) (GH5, GH7, GH12 and GH45) cleave in amorphous cellulose regions; (iv) lytic polysaccha-

ride monooxygenases (LPMO) (AA9 and AA16) can act on both crystalline and less-crystalline regions; (v) cellobiose dehydrogenases (CDH) (AA3-1 and AA8) act on cellobiose oxidation, producing electrons that help the depolymerization of cellulose to be catalyzed by LPMOs; and (vi) the oligosaccharides are further hydrolysed to D-glucose by β-glucosidases (BGL) (GH3) Glass et al. (2013); Rytioja et al. (2014); Panchapakesan and Shankar (2016); Druzhinina and Kubicek (2017); Filiatrault-Chastel et al. (2019). In addition to the widely studied GHs, the cellulolytic complex includes some enzymes with auxiliary activities (AAs). These enzymes act in a synergistic process (CDH-LPMO) in the oxidizing reductive cleavage of the cellulose chain Rytioja et al. (2014). Auxiliary activity enzymes help to reduce the enzyme dosage required for biomass degradation and, therefore, have become important enzymes found in recent commercial cellulases formulations Hu et al. (2018).

The production of plant cell wall degrading enzymes, cellulases, hemicellulases, ligninases and pectinases, is regulated mainly at the transcriptional level in filamentous fungi. Gene expression of these enzymes is regulated by various environmental and cellular factors, some of which are common while others are species-specific or enzyme class-specific. These genes are inducible in presence of the carbon source or molecules derived from the carbon source, whereas repression occurs under growth conditions where the production of these enzymes is not necessary, such as in presence of glucose. Along these lines, it has been shown that cellulolytic enzyme expression is induced by cellobiose in many species of fungi, regarding cellobiose is the primary end product generated from cellulose degradation by cellulases Aro et al. (2005). Recent research results support that accumulation of intracellular cellodextrins (mainly cellobiose) may raise cellulases secretion by a cascade signaling pathway in *P. oxalicum* Yao et al. (2016) 114-2 and *Neurospora crassa* OR74A Cai et al. (2015).

Filamentous fungi are able to transport a wide diversity of sugars by transmembrane proteins. The vast majority of Sugar Transporters (STs) characterized so far, belong to the subfamily (PF00083) of the major facilitator superfamily (MFS). Members in this subfamily include various STs, which are

responsible for the binding and transport of various carbohydrates, organic alcohols, and acids Gonçalves et al. (2015); Peng et al. (2018). Among these sugars, filamentous fungi are able to transport disaccharides such as cellobiose into the cell through specific transporters. Cellobiose and other cellodextrins can act as signal transducers in two ways: i) cellodextrins are transported into cells activating intracellular sensors, and ii) extracellular cellodextrins activate plasma membrane sensors, such as transporter-like proteins or protein-coupled G receptors Dos Reis et al. (2016).

Recently, the whole genome sequences (WGS) of both 2HH wild-type and S1M29 mutant of *P. echinulatum* were deposited at GenBank, allowing studies to discover novel features of the lignocellulolytic enzyme system. These WGS provided evidence that 2HH wild-type strain is closely related to *P. oxalicum*, leading to a taxonomic revision study of this fungus. In this study, we explore the genomic content of this fungus to discover novel features of the lignocellulolytic expression mechanisms, including the characterization of CAZymes and STs involved in the accumulation of intracellular cellodextrins. We also link experimental results of cellulase enhancement secretion from established cellulase producers to *P. echinulatum* 2HH, by analyzing amino acid sequences similarities and their likely roles.

Here, we analyzed the genes that constitute the cellulolytic enzyme system of *P. echinulatum* 2HH using a comparative genomic approach. Besides, *P. echinulatum* 2HH phylogenetic adjacency to *P. oxalicum* 114-2 enabled genomic comparisons to identify singularities in the lignocellulolytic enzyme system of these two well-known enzyme producers. Characterization of CAZymes and STs also provided new evidences to elucidate the adaptation of *P. echinulatum* 2HH to the coleoptera larvae diet during a potential long-term mutualistic symbiosis. Additionally, our *in silico* approach allowed the discovery of new gene targets and suggests a path to engineer *P. echinulatum* 2HH for industrial use. Finally, our results point to a broad number of genes involved on cellulolytic expression mechanisms, revealing new targets to design hypersecreting strains.

2. Results and Discussion

P. echinulatum CAZyome

There are no significant differences between the vast majority of putative proteins of the 2HH wild-type and the S1M29 mutant. The results and subsequent discussions used the 2HH strain as reference. Among all annotated CAZymes in the putative proteome of *P. echinulatum*, we highlight the protein encoding-genes for the cellulolytic enzyme system in **Table 1**. The Supplementary file S01.1 include all annotated CAZymes of *P. echinulatum*, including also the identified differences between 2HH and S1M29 strains.

Cellulase mixtures of *Penicillium* spp. are known to be rich in β-glucosidase Vaishnav et al. (2018), they are found in GH1 and GH3 families in *P. echinulatum* 2HH. Those proteins belonging to GH1 are probably intracellular enzymes, whereas five of nine GH3 β-glucosidases contain a signal peptide and are probably secreted into the medium. Furthermore, all cellobiohydrolases also contain a signal peptide, one of GH6 family and two of GH7 family. In addition, endo-1,4-β-D-glucanases are found in GH5-4, GH5-5, GH5-22, GH7, GH12 and GH45 families, where GH5-22 family is probably an intracellular enzyme, whereas all others contain a signal peptide. Considering AA enzymes, four LPMOs of AA9 family, the LPMO of AA16 family and the CDH enzyme of AA3-1 family contain a signal peptide, with the exception of the CDH enzyme of AA8 family. These AAs act synergically with the GHs, playing a crucial role in cellulose degradation system. The LPMO-type enzyme of AA16 family acts on cellulose with oxidative cleavage at the C1 position of the glucose unit Filiatrault-Chastel et al. (2019). In this study we identified this enzyme for the first time in *P. echinulatum* 2HH and *P. oxalicum* 114-2. This AA16 coding gene was also found in the other analyzed species of *Penicillium* and *Aspergillus*, being absent only in *Trichoderma reesei* QM6a and *N. crassa* OR74A.

Among the putative cellulases, only the most secreted cellulase named EGL1 has been cloned, characterized and heterologous expressed Rubini et al. (2010). This characterization showed that EGL1 optimal temperature is 60 °C and the

optimal activity occurs over a broad pH range (5–9). Furthermore, the EGL1 secreted by a *Pichia pastoris* recombinant also showed high thermostability (84% of residual activity after 1h of pre-incubation at 70 °C) and calcium exerted a strong stimulatory effect over EGL1 activity Rubini et al. (2010).

Another putative EGL (PECH_006176/PECM_002589) was predicted to contain a near C-terminal CBM_X2 domain (Pfam ID: PF03442, InterPro ID: IPR005102) in addition to a CBM1 (Pfam ID: PF00734, InterPro ID: IPR000254) and a GH5-4 typical catalytic domain (Pfam ID: PF00150, InterPro ID: IPR001547). Homology search suggested that proteins containing cellulase catalytic domain followed by CBM_X2 were present in many fungal species like PdCel5C (PDE_09969) in *P. oxalicum* 114-2 Liu et al. (2013a), Cel5C (PMG11_08470) in *P. brasilianum* MG11, Endoglucanase B (EN45_076530) in *P. chrysogenum* P2niaD18 and (PENSUB_1985) in *P. subrubescens*. It is important to highlight that GH5-4 catalytic domain is not found in cellulolytic complexes of *T. reesei* QM6a and *N. crassa* OR74A.

Several CAZymes contains accessory non-catalytic domains (*e.g.*: CBMs). InterPro, PROSITE, Pfam and dbCAN2 were used to refine CBM predictions and the results are presented in Table S01.2. Specifically, we found 58 proteins containing at least one CBM domain, including 24 proteins with a CBM1 domain targeting cellulose. The main CBM1-containing cellulases are featured in Table 1. In addition to the cellulolytic enzymes of GH families 5, 6, 7, 45 and AA9, CBM1 was also observed in association with different types of catalytic domains from CE2 family and GH families 10, 11, 26, 30, 43 and 62. Furthermore, CBM1 was observed as associated with CBM63 domain, also targeting cellulose in a swollenin encoding-gene. We also found an expansin encoding-gene, containing a CBM63 domain. Many expansin-like proteins have been reported and demonstrated to bind and act on cellulosic networks. Some of them have shown to act synergistically with cellulases and xylanases Georgelis et al. (2014).

P. echinulatum 2HH exhibit a broad profile of plant biomass degrading enzymes (detailed in Supplementary file S01.1), including a variety of enzymes required for xylan degradation: endo-β-1,4 xylanases (GH10, GH11), α-L-arabinofuranosidases

(GH43, GH51, GH54, and GH62), β-xylosidases (GH43 and GH3), acetylxylan esterases (CE1, CE5), α-glucuronidase (GH67), and ferulic acid esterases (CE1). It also contains enzymes required for xyloglucan degradation, including α-xylosidases (GH31), β-galactosidases (GH2 and GH35), α-L-arabinofuranosidases (GH43, GH51, GH54, and GH62) and α-fucosidases (GH29 and GH95). Yet, it includes enzymes required for galactomannan degradation: endo-β-1,4-mannases (GH5-7) and α-galactosidases (GH27, GH36). A wide range of enzymes necessary for efficient degradation of pectin was also observed, including polygalacturonases and rhamnogalacturonases of GH28 family, pectin and pectate lyases (PL1), and rhamnogalacturonan lyases (PL4), as well as pectin esterases (CE8) and rhamnogalacturonan acetylesterases (CE12). Enzymes that function on RG-I side chain substitutions include arabinanases (GH43, GH93), arabinosidases / α-arabinofuranosidases (GH43, GH51, GH54, and GH62), galactanases (GH16, GH30 and GH53), α- and β-galactosidases (GH2, GH27, GH35, GH36, and GH43), β-glucuronidases (GH67, and GH79), and feruloyl esterases (CE1).

In summary, the CAZyome characterization confirms the natural potential of *P. echinulatum* 2HH for the production of cellulolytic mixtures. Considering that *P. echinulatum* 2HH is known to produce a vast range of CAZymes, primarily cellulases and xylanases, previously described by secretome analysis Schneider et al. (2016). These results provide an important step in the molecular understanding of this microorganism, allowing strain improvements using advanced techniques and further elevating the importance of the genus *Penicillium* in biotechnology for biofuels.

CAZymes as evolutionary markers

Whole genome sequences of 2HH wild-type strain, deposited recently at GenBank, provided evidence that 2HH is closely related to *P. oxalicum*, leading to a taxonomic revision study of this fungus. The phylogenetic proximity between *P. echinulatum* 2HH and *P. oxalicum* 114-2 and the orthology of genes belonging to the cellulolytic system (detailed in Supplementary file S01.1) denote that the cellulolytic system of *P. echinulatum* 2HH and *P. oxalicum* 114-2 are highly

similar. We found the respective orthologous in *P. oxalicum* 114-2 for almost all cellulolytic genes of *P. echinulatum* 2HH.

The isolation method of 2HH strain hypothesizes a potential natural adaptation for the secretion of cellulolytic enzymes, as a possible adaptation to *A. punctatum* larvae diet as the only growth condition available for the fungus. To date, this evolutionary hypothesis was supported only by the mixture of cellulases secreted by the 2HH strain, which provides an effective enzymatic formulation for complete saccharification of plant residues rich in cellulose and hemicellulose Schneider et al. (2016). This hypothesis encourages the search for new insights into how *P. echinulatum* 2HH uptakes carbon sources available in the environment.

In order to understand evolutionary relationships, we analyzed the differences in the CAZyome composition of *P. echinulatum* 2HH and *P. oxalicum* 114-2. First, we analyzed CAZymes that showed low transcription level in *P. oxalicum* 114-2 Liu et al. (2013b) and their respective orthology in *P. echinulatum* 2HH. In *P. oxalicum* 114-2, PDE_05193 is an endo-1,4-β-D-glucanase with a signal peptide and it is orthologous of PECH_006981 in *P. echinulatum* 2HH. The secretion profile of *P. oxalicum* 114-2 does not show secreted protein ratio for PDE_05193 Liu et al. (2013b). In addition, the transcription levels for PDE_05193 in CW medium is very low Liu et al. (2013b). When we aligned the nucleotides of *P. oxalicum* 114-2 and *P. echinulatum* 2HH, it was possible to observe gap occurrence above 10%, suggesting that this gene may have been disabled by mutations.

In the same way, extracellular β-glucosidase BGL3 (PDE_01277) of *P. oxalicum* 114-2 did not show activities on both pNPG and salicin in vitro and its role is not yet known Yao et al. (2016). In *P. echinulatum* 2HH we have not found the ortholog of this β-glucosidase of GH1 family. To confirm the absence of this ortholog, we performed a syntenic comparison of the genome region that was supposed to contain the orthologous of PDE_01277 in *P. echinulatum* 2HH. We observed that this part of the sequence was missing, comprising the location of both PDE_01277 and PDE_01278 orthologs. Although, nearby synteny

is preserved, both before (PDE_01276) and after (PDE_01278), exhibiting their respective orthologs (PECH_007174 and PECH_007175) in *P. echinulatum* 2HH.

In contrast, by searching for orthologs in *P. brasilianum* MG11, *P. subrubescens* CBS 132785 and *P. chrysogenum* P2niaD18, we found the respective orthologs of both endo-1,4-β-D-glucanase and β-glucosidase of *P. oxalicum* 114-2. These orthologs in related *Penicillium* spp. suggest that the absence of both genes are particular evolutionary characteristics of *P. echinulatum* 2HH. It is remarkable that these two genes are also the only differences in the proteins containing signal peptide, when we compare the cellulolytic complexes of *P. echinulatum* 2HH and *P. oxalicum* 114-2. An adaptive characteristic, such as the ability to survive within a specific host, may culminate in "conditionally dispensable" genes, reflecting their importance in some, but not all, growth conditions. Gene loss during the evolution can be an adaptive evolutionary force that is especially effective when organisms are faced with abrupt environmental challenges Albalat and Cañestro (2016).

Additionally, orthology analysis revealed other important CAZymes of *P. oxalicum* 114-2 for which the respective orthologs were not found in *P. echinulatum* 2HH, including: (PDE_02801 - GH2) α-glucuronidase, (PDE_02413 - GH3) xylan 1,4-β-xylosidase, (PDE_01949 - GH79) α-glucuronidase, (PDE_02902 - GH88) d-4,5 unsaturated β-glucuronyl hydrolase, (PDE_02128 - AA1) multicopper oxidase, (PDE_01302 - CE2) acetyl xylan esterase and (PDE_03849 - CE16) acetylesterase. Furthermore, it is also noticeable that *P. oxalicum* 114-2 putative proteome does not include enzymes of GH29 family, while *P. echinulatum* 2HH contains one α-fucosidase. The putative proteome of *P. echinulatum* 2HH includes also an additional intracellular β-glucosidase (iBGL)-encoding gene (PECH_004285) when compared to *P. oxalicum* 114-2. This additional β-glucosidase is orthologous to Cel3C in *T. reesei* Qm6a and to PMG11_03092 in *P. brasilianum* MG11. Moreover, *P. echinulatum* 2HH showed four additional α-L-rhamnosidase coding genes of GH78 family, when compared to *P. oxalicum* 114-2. A plausible reason for the existence of these additional genes could be the abundance of undigested plant compounds in the larvae gut, considering that

α-l-rhamnose and fucose are found in plants as components of polysaccharides, such as pectins.

The coleoptera larvae diet, mainly composed of cellulose, hemicellulose and residues of lignin from decayed wood Wheeler and Crowson (1982), led us to hypothesize the presence of environment-specific adaptations in *P. echinulatum* 2HH to degrade these biopolymers. Along these lines, we investigate a specific set of enzymes required to degrade wood, comparing *P. echinulatum* 2HH and *P. oxalicum* 114-2. We included 18 families of peroxidases and CAZymes in our analyses, as it was previously suggested Nagy et al. (2016). **Table 2** shows the number of encoding genes of each enzyme family, which were organized into three major groups: oxidoreductases related to the degradation of lignin or lignin-like compounds, CAZys active on polysaccharide main chains; and other CAZys related to wood decay.

Fewer oxidoreductases encoding-genes of AA1 family and HTP-type could be explained by the larvae diet, which probably does not comprise unaffected lignin, but contains lignin-related compounds. Surprisingly, an encoding-gene of dye-decolorizing peroxidase (DyP) was found in *P. echinulatum* 2HH. This family of heme-containing peroxidases is active on lignin-related compounds and contains important properties for lignocellulose biorefineries Brissos et al. (2017). Apparently, this enzyme is not a particular adaptation of *P. echinulatum* 2HH, considering that the orthologs of this DyP peroxidase were found in *P. brasilianum* MG11, *P. subrubescens* CBS 132785 and *P. chrysogenum* P2niaD18, with the exception of *P. oxalicum* 114-2.

In summary, the major differences of *P. echinulatum* 2HH in relation to *P. oxalicum* 114-2 include: i) fewer encoding-genes for oxidoreductases of AA1 family and HTP-type, contrasting with an additional encoding-gene for a DyP peroxidase, which all related to lignin compounds degradation; ii) one less endo-1,4-β-glucanase of GH5 family and one less extracellular β-glucosidase of GH1 family, where both enzymes showed low transcription levels in *P. oxalicum* 114-2; iii) four additional α-L-rhamnosidases of GH78 family, comprising a remarkable range of enzymes related to pectin degradation; iv) one less β-xylosidase

and one additional iBGL of GH3 family, which is ortholog of Cel3C in *T. reesei* Qm6a; and v) one less acetyl-xylan esterase of CE16 family, one additional endomannanase of GH5-7 family and one additional α-fucosidase of GH29 family. Our results reinforce the hypothesis that a potentially host–symbiont association may lead to environment-specific adaptations in the symbiont (fungus), particularly due to the host (insect larvae) diet, although it is still a hypothesis. *P. echinulatum* 2HH and *P. oxalicum* 114-2 may have a very close common ancestor and this is a remarkable finding that might grant a status to *P. echinulatum* 2HH on the global market of enzymes producers, particularly owing to its potential natural evolution for cellulase production.

Comparative analysis of Plant Biomass Degrading CAZymes (PBDC) in related fungi

We performed a comparative analysis of the number of CAZy coding proteins, which are related to the degradation of plant biomass, and which also contain signal peptide. First, we identified putative CAZymes, then each one was classified according to its connection to plant biomass degradation, as it was previously suggested Peng et al. (2017). Finally, we crossed both PBDC and signal peptide predictions. Our results provide a comprehensive comparative analysis of plant biomass degradation profile between twelve filamentous fungi species. The stacked barplot (**Figure 1**) shows the distribution of potentially extracellular CAZymes (number of proteins) involved in degradation of plant biopolymers. Complementary information about PBDC predictions are available in Supplementary file S02. Comparative studies like this one, contribute with the identification of nature and peculiarities for each species and how each one can be used for commercial enzymatic production.

As can be observed in the stacked barplot (Figure 1), when we totalize the number of proteins in all CAZy classes, the total number of PBDC in analyzed *Ascomycetes* and also between the *Penicillium* species varies greatly. Our results revealed an expressive higher number of PBDC for *P. subrubescens* CBS 132785, totalizing 181 proteins. *Aspergillus oryzae* RIB40 also displays a wide range of

PBDC, being the two fungi to outpace the 180 proteins, while *P. digitatum* Pd1 includes only 71 proteins. In contrast, *P. echinulatum* 2HH and *P. oxalicum* 114-2 include an average number of PBDC, comprising 125 and 121 proteins, respectively. Although the number of PBDC is not a key factor for efficient breakdown of plant biomass, our comparison shows the potential for enzyme secretion in relevant filamentous fungi.

A previous study observed CAZy families that were present in *P. chrysogenum* P2niaD18 and *Aspergillus niger* CBS 513.88 but not in *T. reesei* Qm6a Daly et al. (2017). These enzymes comprise CE8 and CE12 families whose proteins encode for pectin methylesterase and rhamnose acetylesterase activities, respectively. Also, PL1-7 and PL4-1 families that encode for pectate lyase and rhamnogalacturonan lyase activities respectively, and endo-α-1,5-L-arabinosidase from GH43-6 family. In our study we observed that the same families were present in almost all analyzed species but not in *T. reesei* Qm6a, indicating that *T. reesei* Qm6a does not include enzymes with these activities. Except for *N. crassa* OR74A which also does not include GH43-6 enzymes. In summary, *T. reesei* Qm6a provides a narrow range of enzymes for pectin degradation when compared to the other fungi.

A larger difference was observed in *P. echinulatum* 2HH when compared to other *Penicillium* species. Most of this difference can be attributed to GH11, GH43 and AA9, which account for 19 PBDC in *P. echinulatum* 2HH, 16 in *P. oxalicum* 114-2, 12 in *P. chrysogenum* P2niaD18, 16 in *P. brasilianum* MG11, 29 in *P. subrubescens* CBS 132785 and only 2 in *P. digitatum* Pd1. Enzymes of these families are mostly related to the degradation of hemicellulose and cellulose, including xylanases, xylosidades, arabinofuranosidases and LPMOs. As already mentioned, *P. echinulatum* 2HH and *P. oxalicum* 114-2 showed a quite similar profile of putative cellulolytic enzymes, leading to their comparison with other filamentous fungi to discover peculiarities of cellulolytic enzyme profiles. In this context, we highlight the ability of these two fungi to produce effective cellulolytic cocktails. We made interesting discoveries by reducing the analysis scope to putative proteins of the cellulolytic system that carry signal peptide

(**Figure 2**).

In *T. reesei* Qm6a, despite the regular number of β-glucosidase putative proteins, it is known that the low β-glucosidase activity of the cellulolytic complex leads to inefficiency in biomass degradation, requiring genetic engineering for secretion enhancement of this enzyme Li et al. (2016). Additionally, the conservation of LPMO-type enzyme (AA16) in *Aspergilli* and *Penicillia* analyzed is notable, while this enzyme family was not found in *Neurospora* and *Trichoderma* genomes. Another interesting finding comprises putative cellulases of GH45 family. Here, we demonstrated that only four of the twelve filamentous fungi investigated possess GH45 proteins containing signal peptide. Lastly, GH5-4 typical catalytic domain was not found in the cellulolytic complex of *T. reesei* Qm6a, *N. crassa* OR74A and *A. niger* CBS 513.88.

In summary, our results showed that both *P. echinulatum* 2HH and *P. oxalicum* 114-2 enclose a quite similar profile of PBDC. Cellulolytic activities of AA16, GH5-4 and GH45 deserve attention when it comes to understanding their roles in the cellulose degradation system of *P. echinulatum* 2HH and *P. oxalicum* 114-2. Considering that these enzymes comprehend the main differences in the cellulolytic complexes, when both are compared to commercial producers. The peculiarities found in our study may contribute to highlight *P. echinulatum* 2HH and *P. oxalicum* 114-2 cellulolytic complexes, contributing to the commercial ascendance of these two fungi. In addition to the high level of extracellular β-glucosidase activity, our results help to support the commercial application of *P. echinulatum* 2HH for cellulolytic enzymes production.

Cellulolytic system expression induced by cellodextrins

Cellodextrins are saccharide polymers of varying length (two or more glucose monomers) resulting from cellulolysis, the breakdown of cellulose. The primary end product generated from cellulose degradation is disaccharide cellobiose. Cellodextrin classification occurs by its degree of polymerization (DP) including different saccharide polymers, such as cellobiose (DP2), cellotriose (DP3), cellotetraose (DP4) and so forth. Previous research results support the existence

of a cascade signaling pathway conserved in filamentous fungi Aro et al. (2005). This pathway acts when cellobiose or other cellodextrins accumulates into the cell, raising the secretion of cellulases. In this context, intracellular accumulation of cellodextrins can occur in two ways: i) by low activity of iBGLs, which reduces the hydrolysis of cellodextrins to D-glucose Yao et al. (2016); Shida et al. (2015); and ii) by the high expression of STs, which are able to transport cellodextrins from the medium into the cell Cai et al. (2015); Li et al. (2013). In order to discover the features behind this cellodextrin induction system in *P. echinulatum* 2HH, we performed several genomic analyses including iBGLs and STs, as presented below.

Intracellular β-glucosidases (iBGLs)

Phylogenetic analyses were performed using two datasets containing 19 and 32 iBGLs sequences, respectively for GH1 and GH3 families. The GH1 dataset includes 3 putative iBGLs of *P. echinulatum* 2HH, 13 putative iBGLs from related fungi, as well as, 3 BGLs of *Arabidopsis thaliana* used as outgroup. The GH3 dataset includes 4 putative iBGLs of *P. echinulatum* 2HH, 26 putative iBGL sequences of the related fungi, as well as, 2 BGLs of *A. thaliana* used as outgroup. Detailed information of iBGLs is available in the Supplementary file S03. **Figure 3** shows the phylogenetic classification of iBGLs, comprising enzymes of GH1 (a) and GH3 (b) families. The roles of proteins highlighted in bold in both trees were verified and may provide evidence of likely conserved roles in filamentous fungi.

In *T. reesei* Qm6a, CEL1a and CEL1b iBGLs may not participate directly into cellobiose hydrolysis, however they may contribute to the accumulation of cellobiose as signal inducers. CEL1a plays an important role in cellulase induction in *T. reesei* PC-3-7, since the *cel1a* single-nucleotide mutation in strain PC-3-7 resulted in high cellulase expression on cellobiose Zhou et al. (2012); Xu et al. (2014); Shida et al. (2015). In *T. reesei* QM6a, CEL1a and CEL1b were also functionally equivalent in mediating the induction of cellulase genes by lactose and the simultaneous absence of these iBGLs abolished *cbh1* gene

expression. Still in *T. reesei* QM6a, CEL1a protein and its glycoside hydrolytic activity were indispensable for cellulase induction by lactose. Intracellular BGL-mediated lactose induction is further conveyed to XYR1 to ensure the efficiently induced expression of cellulase genes Xu et al. (2014).

Moreover, several studies were carried out involving iBGLs of GH3 family in *T. reesei* QM6a. Deletion of *bgl3i* gene significantly increased cellulase activities, it had no influence on fungal growth though. Deletion of *bgl3i* also enhanced transcription levels of CEL1a, CEL1b and XYR1 regulator, which are all crucial for lactose induction in *T. reesei* QM6a Zou et al. (2018). Still in *T. reesei* QM6a, Δ*cel3c* mutant had no significant influence on the expression of secreted proteins Qin et al. (2018), while dysfunction of *cel3d* resulted in higher secretion of cellulases Li et al. (2016).

In *N. crassa* OR74A, individual BGL deletion strains (Δ*gh1-1*, Δ*gh3-3*, or Δ*gh3-4*) did not showed a significant induction of major cellulase genes, whereas a Δ*gh1-1*Δ*gh3-3* double deletion mutant showed some cellulase gene induction. However, a strain carrying deletions for all three BGL genes (Δ*gh1-1*, Δ*gh3-3*, and Δ*gh3-4*) resulted in a strain that produces higher concentrations of secreted active cellulases on cellobiose Znameroski et al. (2012). Still in *N. crassa* OR74A, double BGL deletion strain Δ*gh3-2*Δ*gh3-5* had similar intracellular activity as the wild-type. These two BGL genes do not contribute to BGL activity production, while GH3-6 is the main responsible for intracellular BGL activity Wu et al. (2013).

In *P. oxalicum* 114-2, BGL2 (PDE_00579) is the major iBGL and was found to be dependent on ClrB at the transcription level. The deletion of *bgl2* facilitates the synergistic expression of cellulase genes. Lack of this iBGL facilitates the accumulation of intracellular cellodextrins, which can trigger signaling cascades that include expression of cellulase genes Li et al. (2015); Yao et al. (2016).

In *P. echinulatum*, protein sequence alignments of BGL2 orthologs between S1M29 mutant (PECM_002864) and 2HH wild-type (PECH_005648), revealed a single amino acid substitution (D194P), which occurred in the major iBGL of GH1 family. Although several mutations have been identified, this mutation is

probably the major source for cellulase hyperproduction by the S1M29 mutant. Despite amino acid substitution not affecting BGL2 catalytic domain in *P. echinulatum* S1M29, a single amino acid substitution could negatively affect the enzyme or reduce its activity, as occurred in BGL2 ortholog of *T. reesei* QM6a Shida et al. (2015).

In summary, the influence of iBGLs on the induction of cellulolytic enzyme systems in filamentous fungi is undeniable. However, related fungi results report the complexity and specificities of each species. All iBGLs found in *P. echinulatum* 2HH are highlighted with blue stars in the phylogenetic trees, which allows opportunities to figure out molecular mechanisms underlying the regulation of cellulolytic enzymes secretion. Therefore, we suggest these highlighted genes as potential engineering targets, aiming to improve the expression of cellulolytic enzymes in *P. echinulatum* 2HH.

Sugar transporters (STs)

Sugar transportome of *P. echinulatum* 2HH includes 64 putative ST encoding genes, found by searching the conserved ST domain (PF00083) on putative proteome. We also found considerable diversity in the numbers of PF00083-containing proteins in the fungi investigated, with more than three-fold differences between related species. Our results are consistent with the *Aspergillus* phylogenetic study of STs de Vries et al. (2017). The largest and smallest numbers of PF00083-containing proteins correspond to *P. subrubescens* CBS 132785 (116) and *N. crassa* OR74A (35), respectively. Even species from the same *Penicillium* section do not have similar numbers of loci, such as *Lanata-Divaricata* where *P. oxalicum* 114-2 includes 59 genes and *P. subrubescens* CBS 132785 includes 116 genes.

Furthermore, a phylogenetic analysis was performed using a dataset containing 200 ST sequences, including 64 putative STs of *P. echinulatum* 2HH, 85 putative ST sequences of *A. niger* CBS 513.88 Peng et al. (2018), 44 ST protein sequences of related fungi reported in literature, as well as, 7 STs of *A. thaliana* used as outgroup. Detailed information of STs is available in the Supplemen-

tary file S03. **Figure 4** shows the outcome tree representing the phylogenetic classification of STs, where nine different clades, supported by bootstrap values above 70%, are clearly distinguished. Our results are accordant with previous phylogenetic studies of STs Peng et al. (2018); de Vries et al. (2017). Putative sugar specificity to each clade were suggested based on the previously reported properties of the STs included in the phylogenetic tree.

With the exception of unknown STs highlighted in red, we numbered the clades coursing the tree counterclockwise. The first clade contains 11 STs of *P. echinulatum* 2HH and 9 known cellodextrin/lactose transporters from related fungi. This clade represents the most important group of STs for understanding the cellodextrin induction system in *P. echinulatum* 2HH. Previous studies and established functions of these STs in related fungi, particularly in *P. oxalicum* 114-2, help to clarify and provide insights on the influence of cellodextrin transporters in the cellulolytic induction system. In *P. oxalicum* 114-2, overlapping activity of isoproteins was observed between cellodextrin transporters (*cdtC*, *cdtD* or *cdtG*). Deletion of a single gene resulted in no observable effect on cellulase expression. Nonetheless, simultaneous deletion of *cdtC* and *cdtD* resulted in remarkable decrease in cellobiose consumption and low growth on cellulose, resulting also in lower extracellular activity of cellulases. Besides, overexpression of cellodextrin transporter genes (*cdtC*, *cdtD* or *cdtG*) improved cellulase production in *P. oxalicum* mutants, with the highest fold changes in *cdtG* overexpressed mutant Li et al. (2013). Orthologous of these three cellodextrin transporters (*cdtC*, *cdtD* or *cdtG*) were found in *P. echinulatum* 2HH: PECH_005610, PECH_006659 and PECH_005330, respectively.

In *Aspergillus nidulans* FGSC A4, CltA is a cellobiose-specific transporter, while CltB/LacpB is able to transport cellobiose and lactose. However, this protein is a cellulose signaling sensor rather than a cellobiose transporter Dos Reis et al. (2016). Still in *A. nidulans* FGSC A4, deletion of *cltB/lacpB* resulted in reduced growth and extracellular cellulase activity, indicating that cellulose and lactose catabolic systems operate with common components. Yet, deletion of *cltA* showed no significant effect on cellulase expression in the presence of

cellobiose Fekete et al. (2016). In *P. echinulatum* 2HH, *cltA* was also identified as orthologous to PECH_006659, while *cltB/lacpB* is also an ortholog of PECH_005610. *P. echinulatum* 2HH appears to lack orthologs of *lacpA* (high-affinity lactose permease) of *A. nidulans* FGSC A4, reinforcing the hypothesis that lactose is not among the preferred carbon sources of 2HH strain. Previous experiments have shown reduced extracellular cellulase activity in lactose medium by 9A02S1 mutant, obtained from the 2HH strain Sehnem et al. (2006). In *T. reesei* QM6a, Crt1 plays a crucial role in lactose induction of cellulase genes, either as a lactose transporter or a cellulose sensor Ivanova et al. (2013). The absence of *crt1* abolished cellulase gene expression, being essential in cellulase gene induction independent of intracellular sugar delivery Zhang et al. (2013). In *P. echinulatum* 2HH, *crt1* was also identified as orthologous to PECH_005610.

In *N. crassa* OR74A, CDT1 and CDT2 present dual function, acting as cellodextrin transporters and also holding a key role as cellulose signaling sensors, involved in the induction of cellulases. Still in *N. crassa* OR74A, CLP1 is a putative cellodextrin transporter-like protein that is a critical component of cellulase induction pathway. Although CLP1 protein cannot transport cellodextrin, this signaling sensor may suppress cellulase induction. The co-disruption of *cdt2* and *clp1* enhanced 6.9-fold the cellulase production with cellobiose induction in the strain $\Delta 3\beta$G Cai et al. (2015). In *P. echinulatum* 2HH, *clp1* was identified as orthologous to PECH_007291, while *cdt1* is orthologous to PECH_005610 and *cdt2* is orthologous to PECH_004467. In addition to the five orthologs listed so far, six more putative cellodextrin transporters and/or sensors were mapped in *P. echinulatum* 2HH: PECH_007978, PECH_001261, PECH_002010, PECH_008603, PECH_005239, PECH_003597. These transporters lack reviewed orthologs in related fungi, demanding experimental studies to clarify their roles in *P. echinulatum* 2HH cellulase induction.

Following the tree counterclockwise, Clade 2 carries mainly pentose and glycerol transporters, containing ten STs of *P. echinulatum* 2HH and twenty-one STs of *A. niger* CBS 513.88, as well as two pentose transporters XAT1 and AN25 of

N. crassa OR74A, the first with specificity for D-xylose and L-arabinose and the second is a D-xylose-specific transporter Li et al. (2014). Besides that, it carries glycerol transporter (MfsA) of *Aspergillus fumigatus* Af293 Morton et al. (2011). Transporters expressed by filamentous fungi can often transport more than one type of sugar. For example, *A. nidulans* FGSC A4 transporter XtrD is able to transport xylose, glucose and several other monosaccharides, whereas STP1 is involved in glucose and cellobiose uptake in *T. reesei* QM6a. Disruption of *stp1* in *T. reesei* QM6a comprised major cellulase and hemicellulase genes induction on cellobiose but not on sophorose Zhang et al. (2013). In *P. echinulatum* 2HH, *stp1* was identified as orthologous to PECH_001072.

Clade 3 contains diverse pentose and hexose transporters including three STs of *P. echinulatum* 2HH and L-arabinose LAT1 of *N. crassa* OR74A Benz et al. (2014b), D-xylose XltC of *A. niger* CBS 513.88 Sloothaak et al. (2016), hexose hxtA of *A. nidulans* FGSC A4 Wei et al. (2004), D-glucose MstF of *A. niger* CBS 513.88 Jørgensen et al. (2007), and D-glucose HXT1 of *T. reesei* QM6a Zhang et al. (2015). Clade 4 enclose D-xylose transporters of *A. niger* CBS 513.88 and *T. reesei* QM6a Sloothaak et al. (2016), in addition to four STs of *P. echinulatum* 2HH. Following, Clade 5 contains quinic acid transporters Whittington et al. (1987); Tang et al. (2011) and D-galacturonic acid transporters GatA of *A. niger* CBS 513.88 Sloothaak et al. (2014) and GalA of *N. crassa* OR74A Benz et al. (2014a), enclosing also eight STs of *P. echinulatum* 2HH.

Clade 6 includes seven STs of *P. echinulatum* 2HH, one pentose transporter XYT-1 Li et al. (2014), one glucose sensor RCO-3 Madi et al. (1997). It also comprises D-glucose transporters including Hgt-1/-2 and Glt1 Wang et al. (2017) of *N. crassa* OR74A and some other *Aspergillus* species [mstA, mstC, mstD, mstE, mstG and mstH of *A. nidulans* FGSC A4 and *A. niger* CBS 513.88] Peng et al. (2018). Clade 7 comprises nine STs of *P. echinulatum* 2HH, ten STs of *A. niger* CBS 513.88 Peng et al. (2018) and one known maltose permease of *A. oryzae* RIB40 Hasegawa et al. (2010). Lastly, Clade 8 includes nine inositol and fructose transporters of *A. niger* CBS 513.88 Peng et al. (2018) and five STs of *P. echinulatum* 2HH.

In summary, our phylogenetic analysis, including sugar transportome of *P. echinulatum* 2HH, follows the same classification observed in *A. niger* CBS 513.88 Peng et al. (2018). Eight major families of STs with specificity to different groups of sugar molecules were identified, involving hexoses, pentoses, di-/oligosaccharides, and galacturonic/quinic acid. Furthermore, we also identified 11 STs of *P. echinulatum* 2HH and 9 known cellodextrin and lactose transporters from related fungi, which are grouped in a specific clade in our phylogenetic analysis. Of these 11 STs of *P. echinulatum* 2HH, 5 STs correspond to orthologous reported in literature of the related fungi, which are proven to affect the induction of cellulolytic enzymes. These results suggest that *P. echinulatum* 2HH diversity and specificity of STs are consistent to other cellulase producers. The putative STs provide new insights into metabolism and nutritional behavior of *P. echinulatum* 2HH. Finally, the genes highlighted with blue stars in the tree comprise the basement to comprehend the role of cellobiose induction on the cellulolytic expression mechanisms of *P. echinulatum* 2HH. These gene targets can be applied to different industrial processes and represent an important tool to engineer *P. echinulatum* 2HH for the biofuel industry.

In the genomic analyses we found out various novel features of the lignocellulolytic enzyme system of *P. echinulatum* 2HH. The CAZyome characterization exhibits the outstanding repertoire of enzymes involved in the degradation of lignocellulolytic biomass offered by *P. echinulatum* 2HH. In fact, the genes that constitute the cellulolytic enzyme system of 2HH strain are predominantly orthologs to the cellulolytic enzyme system of *P. oxalicum* 114-2, revealing the phylogenetic proximity of these filamentous fungi. Cellulolytic activities of AA16, GH5-4 and GH45 deserve attention when it comes to understanding their roles in the cellulose degradation system of *P. echinulatum* 2HH and *P. oxalicum* 114-2, considering that these encoding genes comprehend the main differences in the cellulolytic complexes, when both are compared to the commercial producers. Both cellulolytic systems include a LPMO-type enzyme of the AA16 family, which acts on cellulose with oxidative cleavage at the C1

position of the glucose unit, described for the first time in these fungi.

Besides the similarities, we also highlight the singularities in the lignocellulolytic enzyme system of *P. echinulatum* 2HH. Considering CAZymes as evolutionary markers, we compared *P. echinulatum* 2HH to *P. oxalicum* 114-2, reinforcing the previous reported hypothesis of environment-specific adaptations in *P. echinulatum* 2HH during a potential long-term mutualistic symbiosis with *A. punctatum* larvae. We suggest that adaptations to the symbiotic environment associated to the larvae restricted diet could potentially explain some differences in the gene composition of enzymes required to degrade wood. Major differences include: i) fewer encoding-genes for oxidoreductases related to degradation of lignin in *P. echinulatum* 2HH; and ii) functional genes with low transcription levels corresponding to cellulolytic enzymes in *P. oxalicum* 114-2, whose orthologs were absent or identified as pseudogene in *P. echinulatum* 2HH. However, our results are not conclusive and the potential long-term mutualistic symbiosis persists as a hypothesis.

In addition to the CAZyome, we also characterized the sugar transportome of *P. echinulatum* 2HH. Our phylogenetic analysis suggests that *P. echinulatum* 2HH diversity and specificity of STs are consistent to other enzyme producers, including eight major families of STs with specificity to different groups of sugars. Phylogenetic classification of STs helps to clarify the role of STs regarding the preferred carbon source of *P. echinulatum* 2HH. Furthermore, the phylogenetic analyses of iBGLs and STs enabled the identification of several iBGLs and STs involved in the accumulation of intracellular cellodextrins, bringing about a few candidate target genes for rational engineering of industrial strains. Considering the intracellular cyclodextrin accumulation mechanism that plays a key role inducting the expression of cellulolytic enzymes in filamentous fungi by a signalling cascade pathway. Overall, a significant number of putative iBGLs and STs of *P. echinulatum* 2HH correspond to orthologs reported in the literature of related fungi, which are proven to affect the induction of cellulolytic enzymes. These iBGLs and STs comprise valuable gene targets to understand the mechanisms underlying the regulation of cellulolytic enzymes and to design

hypersecreting strains of *P. echinulatum* 2HH.

3. Conclusions

Our results provide the first *in silico* characterization of CAZymes and STs of *P. echinulatum* 2HH, revealing surprising features related to PBDC, particularly related to the cellulolytic enzyme system and its regulatory components. Knowledge of CAZys and STs provide valuable instruments for strain improvements for ethanol production, regarding adequate enzymatic balance for the attributes of second-generation feedstocks, such as crop residue of corn or sugarcane bagasse. Our results also provide new evidences on the close phylogenetic relationship between *P. echinulatum* 2HH and *P. oxalicum* 114-2, well-known cellulase producers studied extensively in Brazil and China, respectively. Finally, our results confer important steps into building molecular understanding of *P. echinulatum* 2HH, highlighting the distinguished cellulolytic system and promoting this species as an important biotechnological ally for lignocellulosic biofuel production.

Materials and Methods

Fungal putative proteomes analyzed

Putative proteomes of *P. echinulatum* 2HH and S1M29 are available at National Center for Biotechnology Information (NCBI) under the accession numbers WIWU00000000 and WIWV00000000, respectively. We were used wild-type strains of model fungi, phylogenetically related fungi and fungi used commercially for the production of cellulolytic enzyme systems for comparison with *P. echinulatum*. The information of the other eleven fungal putative proteomes used in this study were downloaded from the UniprotKB server for a) *P. oxalicum* 114-2 (UP000019376) Liu et al. (2013c); b) *P. chrysogenum* P2niaD18 (UP000076449); c) *P. digitatum* Pd1 (UP000009886); d) *P. brasilianum* MG11 (UP000042958); e) *P. subrubescens* CBS 132785 (UP000186955); f) *A. niger* CBS 513.88 (UP000006706); g) *A. nidulans* FGSC A4 (UP000000560); h) *A.*

oryzae RIB40 (UP000006564); i) *A. fumigatus* Af293 (UP000002530); j) *T. reesei* Qm6a (UP000008984); and k) *N. crassa* OR74A (UP000001805).

CAZyome annotation

Recently, the systematic analysis of bacterial Talamantes et al. (2016) and fungal Berlemont (2017) genomes highlighted the distribution and the variability of enzymes involved in polysaccharide degradation. These approaches provide a framework to investigate CAZymes diversity, and to identify new enzymes with potential for the biopolymer degradation industry.

In this study we used CAZy database (07/31/2019), which comprises full-length protein sequences included in CAZyDB, downloaded on 30/08/2019 from http://bcb.unl.edu/dbCAN2/download/Databases/. In addition, we used dbCAN HMMdb release 8.0 (08/08/2019), comprehending 641 HMMs of CAZymes (421 family HMMs + 3 cellulosome HMMs + 217 subfamily HMMs), data based on CAZyDB released on 07/26/2019. Two approaches were combined in order to improve the CAZyome annotation accuracy: (i) running BLASTp (v2.7.1+) Camacho et al. (2009) against the protein sequences included in the CAZyDB (07/31/2019). All putative protein sequences of *P. echinulatum* were first compared to the full-length protein sequences of CAZyDB by running BLASTP. Query sequences that produced an e-value <10-6 and aligned over at least 95% with a protein in CAZyDB with >50% identity were assigned to the same family as the subject sequence; (ii) running dbCAN2 Zhang et al. (2018) server against the HMMs included in the HMMdb release 8.0. All putative protein sequences of *P. echinulatum* were also subjected to dbCAN2 searches against HMMdb using specific HMMs for each CAZy module family. dbCAN2 combine three tools outputs from HMMER cut-off (e-value<1e-15, coverage>0.35), DIAMOND cut-off (e-Value<1e-102) and Hotpep cut-off (frequency>6.0, hits>2.6).

In addition to dbCAN2 predictions, InterPro Mitchell et al. (2019), PROSITE Sigrist et al. (2013) and Pfam El-Gebali et al. (2019) were also used to improve accuracy of carbohydrate binding modules assignment. To perform a reliable CAZyome annotation is not a simple task, especially when dealing with

a novel species. For dbCAN authors Zhang et al. (2018), the reliability of their tool depends on CAZy predictions by more than one tool. Our experience in this annotation suggests that the two approaches used in this study are complementary and relevant to assign CAZy families, although they are not yet sufficient to assign putative protein products. Along these lines, all putative CAZymes sequences were also subjected to manual curation which involved BLASTp searches against UniProtKB/Swiss-Prot/TrEMBL Consortium (2019) and orthologous inspection from related fungi to assign function to each CAZyme. Orthologous groups were found by searches using ProteinOrtho (V5.16b) Lechner et al. (2011).

Comparison of Plant Biomass Degrading CAZymes (PBDC)

In order to explore the ability of the twelve investigated filamentous fungi (*P. echinulatum* 2HH and the eleven related fungi) to degrade plant biomass, all encoded protein sequences (Supplementary file S02) were first subjected to dbCAN2 Zhang et al. (2018) server against HMMdb release 8.0 using specific HMM models for each CAZy module family using default cut-off values (previous detailed in *P. echinulatum* CAZyome annotation). Next, CAZy families involved in degradation of plant biomass, previously described in Peng et al. (2017), were used to classify all putative CAZymes predicted by dbCAN2. Then, all putative CAZymes sequences were subjected to SignalP Server (v5.0) Almagro Armenteros et al. (2019) to predict the presence and location of signal peptide cleavage sites. Finally, ggplot2 Wickham (2016) was used to plot all charts.

CAZymes as evolutionary markers

All putative proteomes detailed in the section *Fungal putative proteomes analyzed* were used to find orthologous groups in whole genome-wide searches using ProteinOrtho (V5.16b) Lechner et al. (2011).

Phylogenetic analysis of iBGLs

The putative proteomes of *P. echinulatum* 2HH and the other eleven related fungi were investigated to find iBGL protein sequences. Those belonging to GH1

and GH3 families were added to separate datasets. BGLs from *A. thaliana* Xu et al. (2004); Henrissat et al. (2001) were used as outgroups in the phylogenetic analyses. Detailed information about these datasets are available in the Supplementary file S03.

Each collected iBGL dataset was aligned using the protein alignment tool M-Coffee Wallace et al. (2006), with default parameters. The CIPRES Science Gateway (v3.3) Miller et al. (2010) was used to perform RAxML-HPC2 (v8.2.8) Stamatakis (2014). The workflow analysis was used for bootstrap support (BS), setting PROTGAMMAWAG, executing Maximum Likelihood search and thereafter a thorough bootstrap with 1000 iterations, for each dataset (GH1 and GH3). The resulting trees were visualized and configured using iTOL Letunic and Bork (2019).

Identification of STs

To assess the diversity of STs, the ST domain (PF00083) profile extracted from PFAM database El-Gebali et al. (2019) was used to search against the putative proteomes of *P. echinulatum* and the other eleven related fungi with hmmsearch (v3.1b2) Johnson et al. (2010), choosing hmmsearch score ≥ 238 as cutoff Peng et al. (2018).

Phylogenetic analysis of STs

A dataset was created including 64 putative STs of *P. echinulatum* 2HH, 85 putative ST sequences of *A. niger* CBS 513.88 Peng et al. (2018), as well as, 44 fungal ST protein sequences of related fungi (Only protein sequences reviewed or referenced in previous studies were included for better tree visualization). In addition, 7 STs from *A. thaliana* Büttner (2010) were used as outgroup in the phylogenetic analysis. Detailed information about this dataset containing 200 ST protein sequences is available in the Supplementary file S03.

The collected ST sequences were aligned using TM-Coffee Floden et al. (2016), a transmembrane protein alignment tool. The parameter *sequence type* was set as *transmembrane* and the parameter *homology extension* was set as

UniRef100. The CIPRES Science Gateway (v3.3) Miller et al. (2010) was used to perform RAxML-HPC2 (v8.2.8) Stamatakis (2014). The workflow analysis for ST dataset was performed to obtain bootstrap support (BS), setting PROTGAMMAWAG, executing Maximum Likelihood search and thereafter a thorough bootstrap with 500 iterations. The resulting tree was visualized and configured using iTOL Letunic and Bork (2019).

Author's contributions

The authors contributed equally to this work. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

We are grateful to the Coordination for the Improvement of Higher Education Personnel (CAPES) for the PhD scholarship (88887.158496/2017-00 to Lenz, A.R.). This research was supported by grants from CAPES (3255/2013), the National Council for Scientific and Technological Development (CNPq) (472153/2013-7) and Dirección General de Asuntos del Personal Académico-Universidad Nacional Autónoma de México (UNAM) (IN-209620). Camassola, M. and Dillon, A.J.P. are CNPq Research Fellowship.

Acknowledgements

We are thankful to CAPES, Bahia State University (UNEB) and University of Caxias do Sul (UCS), especially to UNEB for the leave of absence (3.145/2016 to Lenz, A.R.).

Availability of data and material

All data generated or analyzed during this study are included in supplementary files or available in public databases. Genomic data of *P. echinulatum* are available in NCBI database. The 2HH wild-type data were deposited under the accession numbers PRJNA520890 (BioProject); SRX6631912, SRX6631913 and SRX6631914 (SRA); and WIWU00000000 (WGS). The S1M29 mutant data were deposited under the accession numbers PRJNA521489 (BioProject); SRX6631956, SRX6631957 and SRX6631958 (SRA); and WIWV00000000 (WGS).

List of abbreviations

AA: Auxiliary Activity CBM: Carbohydrate Binding Module CE: Carbohydrate Esterase CAZy: Carbohydrate- Active Enzyme CDH: Cellobiose dehydrogenase EGL: Endoglucanase GH: Glycoside Hydrolase GT: Glycosyl Transferase BGL: β-glucosidase iBGL: intracellular β-glucosidase LPMO: Lytic polysaccharide monooxygenase MFS: Major facilitator superfamily PBDC: Plant Biomass Degrading CAZymes PL: Polysaccharide Lyase SRA: Sequence Read Archives ST: Sugar Transporter WGS: Whole genome sequences

References

Albalat R, Cañestro C. Evolution by gene loss. Nature Reviews Genetics 2016;17(7):379–91. URL: https://doi.org/10.1038/nrg.2016.39. doi:10.1038/nrg.2016.39.

Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nature Biotechnology 2019;37(4):420–3. URL: https://doi.org/10.1038/s41587-019-0036-z. doi:10.1038/s41587-019-0036-z.

Aro N, Pakula T, Penttilä M. Transcriptional regulation of plant cell wall degradation by filamentous fungi. FEMS Microbiology Reviews 2005;29(4):719–39.

URL: https://doi.org/10.1016/j.femsre.2004.11.006. doi:10.1016/j.femsre.2004.11.006.

Benz JP, Protzko RJ, Andrich JM, Bauer S, Dueber JE, Somerville CR. Identification and characterization of a galacturonic acid transporter from Neurospora crassa and its application for Saccharomyces cerevisiae fermentation processes. Biotechnology for Biofuels 2014a;7(1):20. URL: https://doi.org/10.1186/1754-6834-7-20. doi:10.1186/1754-6834-7-20.

Benz PJ, Chau BH, Zheng D, Bauer S, Glass NL, Somerville CR. A comparative systems analysis of polysaccharide-elicited responses in Neurospora crassa reveals carbon source-specific cellular adaptations. Molecular Microbiology 2014b;91(2):275–99. URL: https://doi.org/10.1111/mmi.12459. doi:10.1111/mmi.12459.

Berlemont R. Distribution and diversity of enzymes for polysaccharide degradation in fungi. Scientific Reports 2017;7(1):222. URL: https://doi.org/10.1038/s41598-017-00258-w. doi:10.1038/s41598-017-00258-w.

Brissos V, Tavares D, Sousa AC, Robalo MP, Martins LO. Engineering a Bacterial DyP-Type Peroxidase for Enhanced Oxidation of Lignin-Related Phenolics at Alkaline pH. ACS Catalysis 2017;7(5):3454–65. URL: https://doi.org/10.1021/acscatal.6b03331. doi:10.1021/acscatal.6b03331.

Büttner M. The Arabidopsis sugar transporter (AtSTP) family: An update. Plant Biology 2010;12(SUPPL. 1):35–41. URL: https://doi.org/10.1111/j.1438-8677.2010.00383.x. doi:10.1111/j.1438-8677.2010.00383.x.

Cai P, Wang B, Ji J, Jiang Y, Wan L, Tian C, Ma Y. The putative cellodextrin transporter-like protein CLP1 is involved in cellulase induction in Neurospora crassa. Journal of Biological Chemistry 2015;290(2):788–96. URL: https://doi.org/10.1074/jbc.M114.609875. doi:10.1074/jbc.M114.609875.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: Architecture and applications. BMC Bioinformat-

ics 2009;10(1):421. URL: https://doi.org/10.1186/1471-2105-10-421. doi:10.1186/1471-2105-10-421.

Camassola M, Dillon AJP. Steam-Exploded Sugar Cane Bagasse for On-Site Production of Cellulases and Xylanases by Penicillium echinulatum. Energy & Fuels 2012;26(8):5316–20. URL: https://doi.org/10.1021/ef3009162. doi:10.1021/ef3009162.

Consortium TU. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research 2019;47(D1):D506–15. URL: https://doi.org/10.1093/nar/gky1049. doi:10.1093/nar/gky1049.

Dalena F, Senatore A, Iulianelli A, Di Paola L, Basile M, Basile A. Chapter 2 - ethanol from biomass: Future and perspectives. In: Basile A, Iulianelli A, Dalena F, Veziroğlu TN, editors. Ethanol. Amsterdam, The Netherlands: Elsevier; 2019. p. 25–59. URL: https://doi.org/10.1016/b978-0-12-811458-2.00002-x. doi:10.1016/b978-0-12-811458-2.00002-x.

Daly P, van Munster JM, Kokolski M, Sang F, Blythe MJ, Malla S, Velasco de Castro Oliveira J, Goldman GH, Archer DB. Transcriptomic responses of mixed cultures of ascomycete fungi to lignocellulose using dual RNA-seq reveal inter-species antagonism and limited beneficial effects on CAZyme expression. Fungal Genetics and Biology 2017;102:4–21. URL: https://doi.org/10.1016/j.fgb.2016.04.005. doi:10.1016/j.fgb.2016.04.005.

Dos Reis TF, De Lima PBA, Parachin NS, Mingossi FB, De Castro Oliveira JV, Ries LNA, Goldman GH. Identification and characterization of putative xylose and cellobiose transporters in Aspergillus nidulans. Biotechnology for Biofuels 2016;9(1):204. URL: https://doi.org/10.1186/s13068-016-0611-1. doi:10.1186/s13068-016-0611-1.

Druzhinina IS, Kubicek CP. Genetic engineering of Trichoderma reesei cellulases and their production. Microbial Biotechnology 2017;10(6):1485–

99. URL: https://doi.org/10.1111/1751-7915.12726. doi:10.1111/1751-7915.12726.

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer EL, Hirsh L, Paladin L, Piovesan D, Tosatto SC, Finn RD. The Pfam protein families database in 2019. Nucleic Acids Research 2019;47(D1):D427–32. URL: https://doi.org/10.1093/nar/gky995. doi:10.1093/nar/gky995.

Fekete E, Orosz A, Kulcsár L, Kavalecz N, Flipphi M, Karaffa L. Characterization of a second physiologically relevant lactose permease gene (lacpB) in aspergillus nidulans. Microbiology (United Kingdom) 2016;162(5):837–47. URL: https://doi.org/10.1099/mic.0.000267. doi:10.1099/mic.0.000267.

Filiatrault-Chastel C, Navarro D, Haon M, Grisel S, Herpoël-Gimbert I, Chevret D, Fanuel M, Henrissat B, Heiss-Blanquet S, Margeot A, Berrin JG. AA16, a new lytic polysaccharide monooxygenase family identified in fungal secretomes. Biotechnology for Biofuels 2019;12(1):55. URL: https://doi.org/10.1186/s13068-019-1394-y. doi:10.1186/s13068-019-1394-y.

Floden EW, Tommaso PD, Chatzou M, Magis C, Notredame C, Chang JM. PSI/TM-Coffee: a web server for fast and accurate multiple sequence alignments of regular and transmembrane proteins using homology extension on reduced databases. Nucleic acids research 2016;44(W1):W339–43. URL: https://doi.org/10.1093/nar/gkw300. doi:10.1093/nar/gkw300.

Georgelis N, Nikolaidis N, Cosgrove DJ. Biochemical analysis of expansin-like proteins from microbes. Carbohydrate Polymers 2014;100:17–23. URL: https://doi.org/10.1016/j.carbpol.2013.04.094. doi:10.1016/j.carbpol.2013.04.094.

Glass NL, Schmoll M, Cate JH, Coradetti S. Plant Cell Wall Deconstruction by Ascomycete Fungi. Annual Review of Microbiology 2013;67(1):477–98. URL: https://doi.org/10.1146/annurev-micro-092611-150044. doi:10.1146/annurev-micro-092611-150044.

Gonçalves C, Coelho MA, Salema-Oom M, Gonçalves P. Stepwise Functional Evolution in a Fungal Sugar Transporter Family. Molecular Biology and Evolution 2015;33(2):352–66. URL: https://doi.org/10.1093/molbev/msv220. doi:10.1093/molbev/msv220.

Hasegawa S, Takizawa M, Suyama H, Shintani T, Gomi K. Characterization and expression analysis of a maltose-utilizing (MAL) cluster in Aspergillus oryzae. Fungal Genetics and Biology 2010;47(1):1–9. URL: https://doi.org/10.1016/J.FGB.2009.10.005. doi:10.1016/j.fgb.2009.10.005.

Henrissat B, Coutinho PM, Davies GJ. A census of carbohydrate-active enzymes in the genome of Arabidopsis thaliana. Plant Molecular Biology 2001;47(1-2):55–72. URL: https://doi.org/10.1023/A:1010667012056. doi:10.1023/A:1010667012056.

Hu J, Tian D, Renneckar S, Saddler JN. Enzyme mediated nanofibrillation of cellulose by the synergistic actions of an endoglucanase, lytic polysaccharide monooxygenase (LPMO) and xylanase. Scientific Reports 2018;8(1):3195. URL: https://doi.org/10.1038/s41598-018-21016-6. doi:10.1038/s41598-018-21016-6.

Hyde KD, Xu J, Rapior S, Jeewon R, Lumyong S, Niego AGT, Abeywickrama PD, Aluthmuhandiram JV, Brahamanage RS, Brooks S, Chaiyasen A, Chethana KW, Chomnunti P, Chepkirui C, Chuankid B, de Silva NI, Doilom M, Faulds C, Gentekaki E, Gopalan V, Kakumyan P, Harishchandra D, Hemachandran H, Hongsanan S, Karunarathna A, Karunarathna SC, Khan S, Kumla J, Jayawardena RS, Liu JK, Liu N, Luangharn T, Macabeo APG, Marasinghe DS, Meeks D, Mortimer PE, Mueller P, Nadir S, Nataraja KN, Nontachaiyapoom S, O'Brien M, Penkhrue W, Phukhamsakda C, Ramanan US, Rathnayaka AR, Sadaba RB, Sandargo B, Samarakoon BC, Tennakoon DS, Siva R, Sriprom W, Suryanarayanan TS, Sujarit K, Suwannarach N, Suwunwong T, Thongbai B, Thongklang N, Wei D, Wijesinghe SN, Winiski J, Yan J, Yasanthika E, Stadler M. The amazing po-

tential of fungi: 50 ways we can exploit fungi industrially. Fungal Diversity 2019;97(1):1–136. URL: https://doi.org/10.1007/s13225-019-00430-9. doi:10.1007/s13225-019-00430-9.

Ivanova C, Bååth JA, Seiboth B, Kubicek CP. Systems Analysis of Lactose Metabolism in Trichoderma reesei Identifies a Lactose Permease That Is Essential for Cellulase Induction. PLoS ONE 2013;8(5):e62631. URL: https://doi.org/10.1371/journal.pone.0062631. doi:10.1371/journal.pone.0062631.

Johnson LS, Eddy SR, Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics 2010;11(1):431. URL: https://doi.org/10.1186/1471-2105-11-431. doi:10.1186/1471-2105-11-431.

Jørgensen TR, VanKuyk PA, Poulsen BR, Ruijter GJ, Visser J, Iversen JJ. Glucose uptake and growth of glucose-limited chemostat cultures of Aspergillus niger and a disruptant lacking MstA, a high-affinity glucose transporter. Microbiology 2007;153(6):1963–73. URL: https://doi.org/10.1099/mic.0.2006/005090-0. doi:10.1099/mic.0.2006/005090-0.

Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho: Detection of (Co-)orthologs in large-scale analysis. BMC Bioinformatics 2011;12:124. URL: https://doi.org/10.1186/1471-2105-12-124. doi:10.1186/1471-2105-12-124.

Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic acids research 2019;47(W1):W256–9. URL: https://doi.org/10.1093/nar/gkz239. doi:10.1093/nar/gkz239.

Li C, Lin F, Li Y, Wei W, Wang H, Qin L, Zhou Z, Li B, Wu F, Chen Z. A β-glucosidase hyper-production Trichoderma reesei mutant reveals a potential role of cel3D in cellulase production. Microbial Cell Factories 2016;15(1):151. URL: https://doi.org/10.1186/s12934-016-0550-3. doi:10.1186/s12934-016-0550-3.

Li J, Lin L, Li H, Tian C, Ma Y. Transcriptional comparison of the filamentous fungus Neurospora crassa growing on three major monosaccharides D-glucose, D-xylose and L-arabinose. Biotechnology for Biofuels 2014;7(1):31. URL: https://doi.org/10.1186/1754-6834-7-31. doi:10.1186/1754-6834-7-31.

Li J, Liu G, Chen M, Li Z, Qin Y, Qu Y. Cellodextrin transporters play important roles in cellulase induction in the cellulolytic fungus Penicillium oxalicum. Applied Microbiology and Biotechnology 2013;97(24):10479–88. URL: https://doi.org/10.1007/s00253-013-5301-3. doi:10.1007/s00253-013-5301-3.

Li Z, Yao G, Wu R, Gao L, Kan Q, Liu M, Yang P, Liu G, Qin Y, Song X, Zhong Y, Fang X, Qu Y. Synergistic and Dose-Controlled Regulation of Cellulase Gene Expression in Penicillium oxalicum. PLoS Genetics 2015;11(9):e1005509–. URL: https://doi.org/10.1371/journal.pgen.1005509. doi:10.1371/journal.pgen.1005509.

Liu G, Qin Y, Hu Y, Gao M, Peng S, Qu Y. An endo-1,4-β-glucanase PdCel5C from cellulolytic fungus Penicillium decumbens with distinctive domain composition and hydrolysis product profile. Enzyme and Microbial Technology 2013a;52(3):190–5. URL: https://doi.org/10.1016/j.enzmictec.2012.12.009. doi:10.1016/j.enzmictec.2012.12.009.

Liu G, Zhang L, Qin Y, Zou G, Li Z, Yan X, Wei X, Chen M, Chen L, Zheng K, Zhang J, Ma L, Li J, Liu R, Xu H, Bao X, Fang X, Wang L, Zhong Y, Liu W, Zheng H, Wang S, Wang C, Xun L, Zhao GP, Wang T, Zhou Z, Qu Y. Long-term strain improvements accumulate mutations in regulatory elements responsible for hyper-production of cellulolytic enzymes. Scientific Reports 2013b;3:1569–. URL: https://doi.org/10.1038/srep01569. doi:10.1038/srep01569.

Liu G, Zhang L, Wei X, Zou G, Qin Y, Ma L, Li J, Zheng H, Wang S, Wang C, Xun L, Zhao GP, Zhou Z, Qu Y. Genomic and Secretomic Analyses Reveal

Unique Features of the Lignocellulolytic Enzyme System of Penicillium decumbens. PLoS ONE 2013c;8(2):e55185. URL: https://doi.org/10.1371/journal.pone.0055185. doi:10.1371/journal.pone.0055185.

Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Research 2014;42(D1):D490–. URL: https://doi.org/10.1093/nar/gkt1178. doi:10.1093/nar/gkt1178.

Madi L, McBride SA, Bailey LA, Ebbole DJ. rco-3, a gene involved in glucose transport and conidiation in Neurospora crassa. Genetics 1997;146(2):499–508. URL: https://www.ncbi.nlm.nih.gov/pubmed/9178001.

Menegol D, Fontana RC, Dillon AJP, Camassola M. Second-generation ethanol production from elephant grass at high total solids. Bioresource Technology 2016;211:280–90. URL: https://doi.org/10.1016/j.biortech.2016.03.098. doi:10.1016/j.biortech.2016.03.098.

Miller MA, Pfeiffer W, Schwartz T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. 2010 Gateway Computing Environments Workshop, GCE 2010 2010;:1–8URL: https://doi.org/10.1109/GCE.2010.5676129. doi:10.1109/GCE.2010.5676129.

Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, Brown SD, Chang HY, El-Gebali S, Fraser MI, Gough J, Haft DR, Huang H, Letunic I, Lopez R, Luciani A, Madeira F, Marchler-Bauer A, Mi H, Natale DA, Necci M, Nuka G, Orengo C, Pandurangan AP, Paysan-Lafosse T, Pesseat S, Potter SC, Qureshi MA, Rawlings ND, Redaschi N, Richardson LJ, Rivoire C, Salazar GA, Sangrador-Vegas A, Sigrist CJ, Sillitoe I, Sutton GG, Thanki N, Thomas PD, Tosatto SC, Yong SY, Finn RD. InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. Nucleic Acids Research 2019;47(D1):D351–60. URL: https://doi.org/10.1093/nar/gky1100. doi:10.1093/nar/gky1100.

Morton CO, Varga JJ, Hornbach A, Mezger M, Sennefelder H, Kneitz S, Kurzai O, Krappmann S, Einsele H, Nierman WC, Rogers TR, Loeffler J. The temporal dynamics of differential gene expression in Aspergillus fumigatus interacting with human immature dendritic cells in vitro. PLoS ONE 2011;6(1):e16016–. URL: https://doi.org/10.1371/journal.pone.0016016. doi:10.1371/journal.pone.0016016.

Nagy LG, Riley R, Tritt A, Adam C, Daum C, Floudas D, Sun H, Yadav JS, Pangilinan J, Larsson KH, Matsuura K, Barry K, Labutti K, Kuo R, Ohm RA, Bhattacharya SS, Shirouzu T, Yoshinaga Y, Martin FM, Grigoriev IV, Hibbett DS. Comparative genomics of early-diverging mushroom-forming fungi provides insights into the origins of lignocellulose decay capabilities. Molecular Biology and Evolution 2016;33(4):959–70. URL: https://doi.org/10.1093/molbev/msv337. doi:10.1093/molbev/msv337.

Panchapakesan A, Shankar N. Fungal Cellulases: An Overview. New and Future Developments in Microbial Biotechnology and Bioengineering: Microbial Cellulase System Properties and Applications 2016;:9–18URL: https://doi.org/10.1016/B978-0-444-63507-5.00002-2. doi:10.1016/B978-0-444-63507-5.00002-2.

Peng M, Aguilar-Pontes MV, de Vries RP, Mäkelä MR. In silico analysis of putative sugar transporter genes in Aspergillus niger using phylogeny and comparative transcriptomics. Frontiers in Microbiology 2018;9(MAY):1045. URL: https://doi.org/10.3389/fmicb.2018.01045. doi:10.3389/fmicb.2018.01045.

Peng M, Dilokpimol A, Mäkelä MR, Hildén K, Bervoets S, Riley R, Grigoriev IV, Hainaut M, Henrissat B, de Vries RP, Granchi Z. The draft genome sequence of the ascomycete fungus Penicillium subrubescens reveals a highly enriched content of plant biomass related CAZymes compared to related fungi. Journal of Biotechnology 2017;246:1–3. URL: https://doi.org/10.1016/j.jbiotec.2017.02.012. doi:10.1016/j.jbiotec.2017.02.012.

Qin L, Jiang X, Dong Z, Huang J, Chen X. Identification of two integration sites in favor of transgene expression in Trichoderma reesei. Biotechnology for Biofuels 2018;11(1):142. URL: https://doi.org/10.1186/s13068-018-1139-3. doi:10.1186/s13068-018-1139-3.

Ribeiro DA, Cota J, Alvarez TM, Brüchli F, Bragato J, Pereira BM, Pauletti BA, Jackson G, Pimenta MT, Murakami MT, Camassola M, Ruller R, Dillon AJ, Pradella JG, Paes Leme AF, Squina FM. The Penicillium echinulatum Secretome on Sugar Cane Bagasse. PLoS ONE 2012;7(12):e50571–. URL: https://doi.org/10.1371/journal.pone.0050571. doi:10.1371/journal.pone.0050571.

Rubini MR, Dillon AJ, Kyaw CM, Faria FP, Poças-Fonseca MJ, Silva-Pereira I. Cloning, characterization and heterologous expression of the first Penicillium echinulatum cellulase gene. Journal of Applied Microbiology 2010;108(4):1187–98. URL: https://doi.org/10.1111/j.1365-2672.2009.04528.x. doi:10.1111/j.1365-2672.2009.04528.x.

Rytioja J, Hildén K, Yuzon J, Hatakka A, de Vries RP, Mäkelä MR. Plant-Polysaccharide-Degrading Enzymes from Basidiomycetes. Microbiology and Molecular Biology Reviews 2014;78(4):614–49. URL: https://doi.org/10.1128/mmbr.00035-14. doi:10.1128/mmbr.00035-14.

Schneider WDH, Gonçalves TA, Uchima CA, Couger MB, Prade R, Squina FM, Dillon AJP, Camassola M. Penicillium echinulatum secretome analysis reveals the fungi potential for degradation of lignocellulosic biomass. Biotechnology for Biofuels 2016;9(1):66. URL: https://doi.org/10.1186/s13068-016-0476-3. doi:10.1186/s13068-016-0476-3.

Schneider WDH, Gonçalves TA, Uchima CA, dos Reis L, Fontana RC, Squina FM, Dillon AJP, Camassola M. Comparison of the production of enzymes to cell wall hydrolysis using different carbon sources by Penicillium echinulatum strains and its hydrolysis potential for lignocelullosic biomass. Process Bio-

chemistry 2018;66:162–70. URL: https://doi.org/10.1016/j.procbio.2017.11.004. doi:10.1016/j.procbio.2017.11.004.

Scholl AL, Menegol D, Pitarelo AP, Fontana RC, Filho AZ, Ramos LP, Dillon AJP, Camassola M. Ethanol production from sugars obtained during enzymatic hydrolysis of elephant grass (Pennisetum purpureum, Schum.) pretreated by steam explosion. Bioresource Technology 2015;192:228–37. URL: https://doi.org/10.1016/j.biortech.2015.05.065. doi:10.1016/j.biortech.2015.05.065.

Sehnem NT, De Bittencourt LR, Camassola M, Dillon AJ. Cellulase production by Penicillium echinulatum on lactose. Applied Microbiology and Biotechnology 2006;72(1):163–7. URL: https://doi.org/10.1007/s00253-005-0251-z. doi:10.1007/s00253-005-0251-z.

Shida Y, Yamaguchi K, Nitta M, Nakamura A, Takahashi M, Kidokoro SI, Mori K, Tashiro K, Kuhara S, Matsuzawa T, Yaoi K, Sakamoto Y, Tanaka N, Morikawa Y, Ogasawara W. The impact of a single-nucleotide mutation of bgl2 on cellulase induction in a Trichoderma reesei mutant. Biotechnology for Biofuels 2015;8(1):230. URL: https://doi.org/10.1186/s13068-015-0420-y. doi:10.1186/s13068-015-0420-y.

Sigrist CJ, De Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I. New and continuing developments at PROSITE. Nucleic Acids Research 2013;41(D1):D344–7. URL: https://doi.org/10.1093/nar/gks1067. doi:10.1093/nar/gks1067.

Sloothaak J, Schilders M, Schaap PJ, de Graaff LH. Overexpression of the Aspergillus niger GatA transporter leads to preferential use of D-galacturonic acid over D-xylose. AMB Express 2014;4(1):1–9. URL: https://doi.org/10.1186/s13568-014-0066-3. doi:10.1186/s13568-014-0066-3.

Sloothaak J, Tamayo-Ramos JA, Odoni DI, Laothanachareon T, Derntl C, Mach-Aigner AR, Martins Dos Santos VA, Schaap PJ. Identification and

functional characterization of novel xylose transporters from the cell factories Aspergillus Niger and Trichoderma reesei. Biotechnology for Biofuels 2016;9(1):148. URL: https://doi.org/10.1186/s13068-016-0564-4. doi:10.1186/s13068-016-0564-4.

Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;30(9):1312–3. URL: https://doi.org/10.1093/bioinformatics/btu033. doi:10.1093/bioinformatics/btu033.

Talamantes D, Biabini N, Dang H, Abdoun K, Berlemont R. Natural diversity of cellulases, xylanases, and chitinases in bacteria. Biotechnology for Biofuels 2016;9(1):133. URL: https://doi.org/10.1186/s13068-016-0538-6. doi:10.1186/s13068-016-0538-6.

Tang X, Dong W, Griffith J, Nilsen R, Matthes A, Cheng KB, Reeves J, Schuttler HB, Case ME, Arnold J, Logan DA. Systems biology of the qa gene cluster in Neurospora crassa. PLoS ONE 2011;6(6):e20671. URL: https://doi.org/10.1371/journal.pone.0020671. doi:10.1371/journal.pone.0020671.

Vaishnav N, Singh A, Adsul M, Dixit P, Sandhu SK, Mathur A, Puri SK, Singhania RR. Penicillium: The next emerging champion for cellulase production. Bioresource Technology Reports 2018;2:131–40. URL: https://doi.org/10.101et6/j.biteb.2018.04.003. doi:10.1016/j.biteb.2018.04.003.

de Vries RP, Riley R, Wiebenga A, Aguilar-Osorio G, Amillis S, Uchima CA, Anderluh G, Asadollahi M, Askin M, Barry K, Battaglia E, Bayram Ö, Benocci T, Braus-Stromeyer SA, Caldana C, Cánovas D, Cerqueira GC, Chen F, Chen W, Choi C, Clum A, dos Santos RAC, de Lima Damásio AR, Diallinas G, Emri T, Fekete E, Flipphi M, Freyberg S, Gallo A, Gournas C, Habgood R, Hainaut M, Harispe ML, Henrissat B, Hildén KS, Hope R, Hossain A, Karabika E, Karaffa L, Karányi Z, Kraševec N, Kuo A, Kusch H, LaButti K, Lagendijk EL, Lapidus A, Levasseur A, Lindquist E, Lipzen A,

Logrieco AF, MacCabe A, Mäkelä MR, Malavazi I, Melin P, Meyer V, Mielnichuk N, Miskei M, Molnár ÁP, Mulé G, Ngan CY, Orejas M, Orosz E, Ouedraogo JP, Overkamp KM, Park HS, Perrone G, Piumi F, Punt PJ, Ram AF, Ramón A, Rauscher S, Record E, Riaño-Pachón DM, Robert V, Röhrig J, Ruller R, Salamov A, Salih NS, Samson RA, Sándor E, Sanguinetti M, Schütze T, Sepčić K, Shelest E, Sherlock G, Sophianopoulou V, Squina FM, Sun H, Susca A, Todd RB, Tsang A, Unkles SE, van de Wiele N, van Rossen-Uffink D, de Castro Oliveira JV, Vesth TC, Visser J, Yu JH, Zhou M, Andersen MR, Archer DB, Baker SE, Benoit I, Brakhage AA, Braus GH, Fischer R, Frisvad JC, Goldman GH, Houbraken J, Oakley B, Pócsi I, Scazzocchio C, Seiboth B, VanKuyk PA, Wortman J, Dyer PS, Grigoriev IV. Comparative genomics reveals high biological diversity and specific adaptations in the industrially and medically important fungal genus Aspergillus. Genome Biology 2017;18(1):28. URL: https://doi.org/10.1186/s13059-017-1151-0. doi:10.1186/s13059-017-1151-0.

Wallace IM, O'Sullivan O, Higgins DG, Notredame C. M-Coffee: Combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Research 2006;34(6):1692–9. URL: https://doi.org/10.1093/nar/gkl091. doi:10.1093/nar/gkl091.

Wang B, Li J, Gao J, Cai P, Han X, Tian C. Identification and characterization of the glucose dual-affinity transport system in Neurospora crassa: Pleiotropic roles in nutrient transport, signaling, and carbon catabolite repression. Biotechnology for Biofuels 2017;10(1):17. URL: https://doi.org/10.1186/s13068-017-0705-4. doi:10.1186/s13068-017-0705-4.

Wei H, Vienken K, Weber R, Bunting S, Requena N, Fischer R. A putative high affinity hexose transporter, hxtA, of Aspergillus nidulans is induced in vegetative hyphae upon starvation and in ascogenous hyphae during cleistothecium formation. Fungal Genetics and Biology 2004;41(2):148–56. URL: https://doi.org/10.1016/j.fgb.2003.10.006. doi:10.1016/j.fgb.2003.10.006.

Wheeler Q, Crowson RA. The Biology of the Coleoptera. Systematic Zoology 1982;31(3):342. URL: https://doi.org/10.2307/2413243. doi:10.2307/2413243.

Whittington HA, Grant S, Roberts CF, Lamb H, Hawkins AR. Identification and isolation of a putative permease gene in the quinic acid utilization (QUT) gene cluster of Aspergillus nidulans. Current Genetics 1987;12(2):135–9. URL: https://doi.org/10.1007/BF00434668. doi:10.1007/BF00434668.

Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2nd ed.; volume 35. Cham, Switzerland: Springer, 2016. doi:10.1007/978-3-319-24277-4.

Wu W, Kasuga T, Xiong X, Ma D, Fan Z. Location and contribution of individual β-glucosidase from Neurospora crassa to total β-glucosidase activity. Archives of Microbiology 2013;195(12):823–9. URL: https://doi.org/10.1007/s00203-013-0931-5. doi:10.1007/s00203-013-0931-5.

Xu J, Zhao G, Kou Y, Zhang W, Zhou Q, Chen G, Liu W. Intracellular β-glucosidases CEL1a and CEL1b are essential for cellulase induction on lactose in Trichoderma reesei. Eukaryotic Cell 2014;13(8):1001–13. URL: https://doi.org/10.1128/EC.00100-14. doi:10.1128/EC.00100-14.

Xu Z, Escamilla-Treviño LL, Zeng L, Lalgondar M, Bevan DR, Winkel BS, Mohamed A, Cheng CL, Shih MC, Poulton JE, Esen A. Functional genomic analysis of Arabidopsis thaliana glycoside hydrolase family 1. Plant Molecular Biology 2004;55(3):343–67. URL: https://doi.org/10.1007/s11103-004-0790-1. doi:10.1007/s11103-004-0790-1.

Yao G, Wu R, Kan Q, Gao L, Liu M, Yang P, Du J, Li Z, Qu Y. Production of a high-efficiency cellulase complex via β-glucosidase engineering in Penicillium oxalicum. Biotechnology for Biofuels 2016;9(1):78. URL: https://doi.org/10.1186/s13068-016-0491-4. doi:10.1186/s13068-016-0491-4.

Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, Busk PK, Xu Y, Yin Y. DbCAN2: A meta server for automated carbohydrate-active enzyme

annotation. Nucleic Acids Research 2018;46(W1):W95–. URL: https://doi.org/10.1093/nar/gky418. doi:10.1093/nar/gky418.

Zhang W, Cao Y, Gong J, Bao X, Chen G, Liu W. Identification of residues important for substrate uptake in a glucose transporter from the filamentous fungus Trichoderma reesei. Scientific Reports 2015;5(1):13829. URL: https://doi.org/10.1038/srep13829. doi:10.1038/srep13829.

Zhang W, Kou Y, Xu J, Cao Y, Zhao G, Shao J, Wang H, Wang Z, Bao X, Chen G, Liu W. Two major facilitator superfamily sugar transporters from Trichoderma reesei and their roles in induction of cellulase biosynthesis. Journal of Biological Chemistry 2013;288(46):32861–72. URL: https://doi.org/10.1074/jbc.M113.505826. doi:10.1074/jbc.M113.505826.

Zhou Q, Xu J, Kou Y, Lv X, Zhang X, Zhao G, Zhang W, Chen G, Liu W. Differential involvement of β-glucosidases from Hypocrea jecorina in rapid induction of cellulase genes by cellulose and cellobiose. Eukaryotic Cell 2012;11(11):1371–81. URL: https://doi.org/10.1128/EC.00170-12. doi:10.1128/EC.00170-12.

Znameroski EA, Coradetti ST, Roche CM, Tsai JC, Iavarone AT, Cate JH, Glass NL. Induction of lignocellulose-degrading enzymes in Neurospora crassa by cellodextrins. Proceedings of the National Academy of Sciences of the United States of America 2012;109(16):6012–7. URL: https://doi.org/10.1073/pnas.1118440109. doi:10.1073/pnas.1118440109.

Zou G, Jiang Y, Liu R, Zhu Z, Zhou Z. The putative β-glucosidase BGL3I regulates cellulase induction in Trichoderma reesei. Biotechnology for Biofuels 2018;11(1):314. URL: https://doi.org/10.1186/s13068-018-1314-6. doi:10.1186/s13068-018-1314-6.

Figures

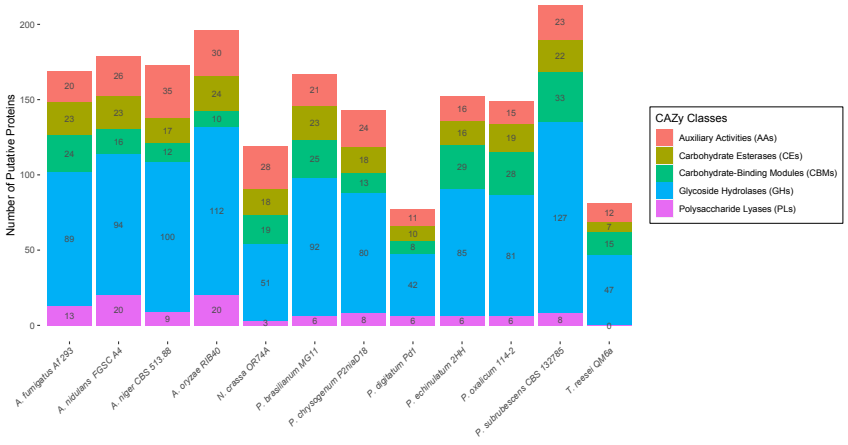


Figure 1: Stacked barplot comparing the number of putative proteins classified as PBDC in twelve filamentous fungi. Only proteins containing signal peptide were counted. Distribution of PBDC sums were grouped accordingly to CAZy classes.

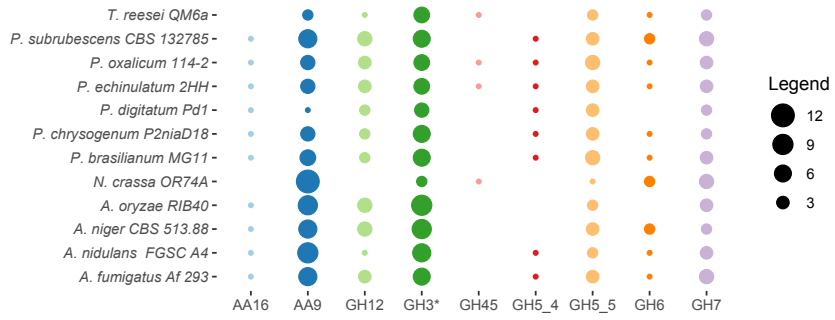


Figure 2: Bubble chart comparing the number of putative CAZymes involved in cellulose degradation. *GH3 count only for β-glucosidase; Only proteins containing signal peptide were counted.

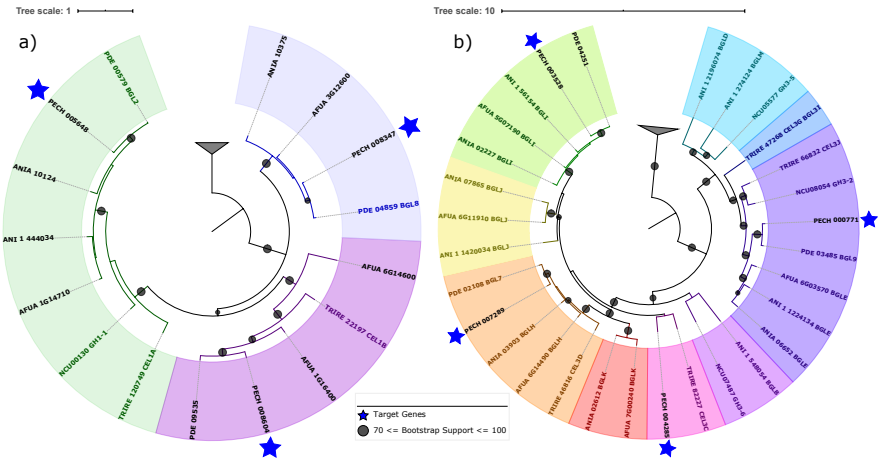


Figure 3: Phylogenetic classification of iBGLs: **a)** GH1 and **b)** GH3 families. The gene names of characterized iBGLs of related fungi are highlighted in bold and in the same colour font of the respective clade; Blue stars are highlighting target genes; Branches with bootstrap values ≥70% were indicated with circles; The gene prefix correspond to the abbreviation of fungal species name (PECH = *P. echinulatum*, PDE = *P. oxalicum*, ANIA = *A. nidulans*, ANI = *A. niger*, AFUA = *A. fumigatus*, NCU = *N. crassa*, TRIRE = *T. reesei*).

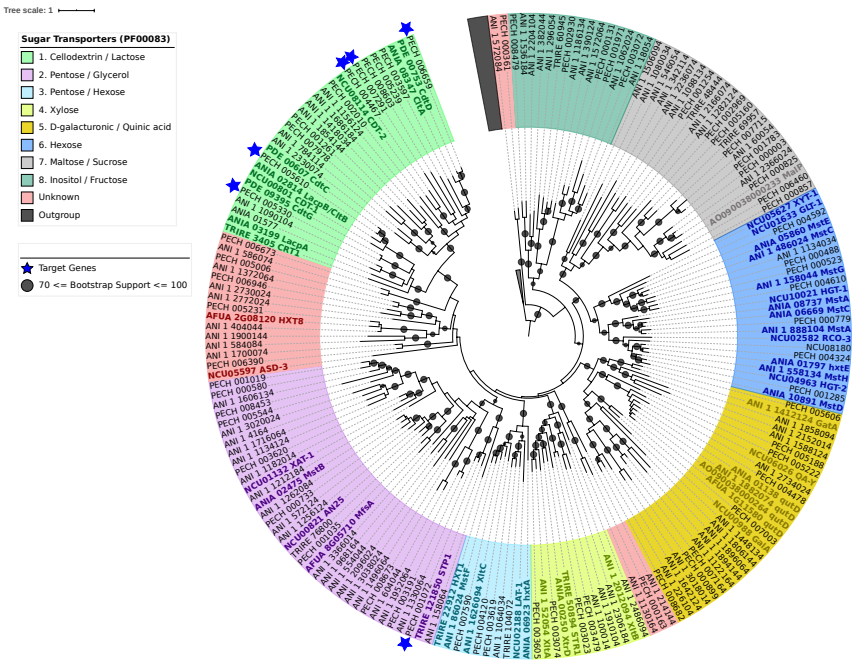


Figure 4: Phylogenetic classification of STs. The gene names of characterized STs of related fungi are highlighted in bold and in the same colour font of the respective clade; Blue stars are highlighting target genes; Branches with bootstrap values $\geq$70% were indicated with circles; The gene prefix correspond to the abbreviation of fungal species name (PECH = *P. echinulatum*, PDE = *P. oxalicum*, ANIA = *A. nidulans*, ANI = *A. niger*, AO = *A. oryzae*, AFUA = *A. fumigatus*, NCU = *N. crassa*, TRIRE = *T. reesei*).

Tables

Table 1: *P. echinulatum* cellulolytic complex encoding genes

2HH ID	S1M29 ID	EC number	CAZy family	Full name	Signal Peptide
PECH_006365	PECM_006949	3.2.1.91	GH6	1,4-β-D-glucan cellobiohydrolase[3]	Y (18-19)
PECH_007386	PECM_003867	3.2.1.91	GH7	1,4-β-D-glucan cellobiohydrolase[3]	Y (27-28)
PECH_008028	PECM_007794	3.2.1.91	GH7	1,4-β-D-glucan cellobiohydrolase	Y (17-18)
PECH_006176	PECM_002589	3.2.1.4	GH5_4	Endo-1,4-β-D-glucanase[3]	Y (18-19)
PECH_009029	PECM_008781	3.2.1.4	GH5_5	Endo-1,4-β-D-glucanase EGL1[1,3]	Y (16-17)
PECH_002030	PECM_004329	3.2.1.4	GH5_5	Endo-1,4-β-D-glucanase[3]	Y (18-19)
PECH_003801	PECM_006072	3.2.1.4	GH5_5	Endo-1,4-β-D-glucanase[3]	Y (21-22)
PECH_001606	PECM_001417	3.2.1.4	GH5_22	Endo-1,4-β-D-glucanase	N
PECH_007371	PECM_003852	3.2.1.4	GH7	Endo-1,4-β-D-glucanase[3]	Y (21-22)
PECH_003013	PECM_002481	3.2.1.4	GH12	Endo-1,4-β-D-glucanase	Y (16-17)
PECH_007370	PECM_003851	3.2.1.4	GH45	Endo-1,4-β-D-glucanase[3]	Y (18-19)
PECH_006981	PECM_005047	-	-	Endo-1,4-β-D-glucanase[2]	N
PECH_004782	PECM_007582	3.2.1.21	GH3	β-glucosidase	Y (19-20)
PECH_002471	PECM_006691	3.2.1.21	GH3	β-glucosidase	Y (21-22)
PECH_005824	PECM_000560	3.2.1.21	GH3	β-glucosidase	Y (22-23)
PECH_003879	PECM_005646	3.2.1.21	GH3	β-glucosidase	Y (20-21)
PECH_007973	PECM_006421	3.2.1.21	GH3	β-glucosidase	Y (19-20)
PECH_004285	PECM_001151	3.2.1.21	GH3	β-glucosidase	N
PECH_000771	PECM_004769	3.2.1.21	GH3	β-glucosidase	N
PECH_007289	PECM_007134	3.2.1.21	GH3	β-glucosidase	N
PECH_003528	PECM_007417	3.2.1.21	GH3	β-glucosidase	N
PECH_002378	PECM_003008	3.2.1.4	AA9	Lytic cellulose monooxygenase	Y (22-23)
PECH_007161	PECM_000770	3.2.1.4	AA9	Lytic cellulose monooxygenase[3]	Y (18-19)
PECH_001644	PECM_001602	3.2.1.4	AA9	Lytic cellulose monooxygenase	Y (21-22)
PECH_008064	PECM_003383	3.2.1.4	AA9	Lytic cellulose monooxygenase[3]	Y (19-20)
PECH_004020	PECM_004700	1.14.99.54	AA16	Lytic cellulose monooxygenase (C1-hydroxylating)	Y (20-21)
PECH_000306	PECM_008697	1.1.99.18	AA3_1	Cellobiose dehydrogenase	Y (22-23)
PECH_005234	PECM_006332	1.1.99.18	AA8	Cellobiose dehydrogenase (cytochrome)	N

[1] Characterized enzyme

[2] Pseudogene

[3] CBM1-containing

Table 2: Distribution of enzymes required to degrade wood in *P. echinulatum* 2HH and *P. oxalicum* 114-2

Enzyme Family	Activity Related to Wood Degradation	*P. oxalicum* 114-2	*P. echinulatum* 2HH
Oxidoreductases		12	11
AA1	Lignin degradation	6	5
AA2	Lignin degradation	3	3
DyP (PF04261)	Acting on lignin or lignin derivatives	0	1
HTP (PF01328)	Possible action on lignin derivatives	3	2
CAZys active on polysaccharide main chains		59	60
GH5_5	Endoglucanase	3	3
GH5_7	Endomannanase	1	2
GH6	Cellobiohydrolase	1	1
GH7	Cellobiohydrolase	3	3
AA9	Cellulose cleaving oxidoreductase	4	4
AA16	Cellulose cleaving oxidoreductase	1	1
GH10	Endoxylanase	3	3
GH12	Endoglucanase	3	3
GH28	Pectinase activity	12	12
GH45	Endoglucanase	1	1
GH74	Xyloglucanase	0	0
GH3	β-glucosidase/β-xylosidase	14	14
GH43	α-Arabinofuranosidase/β-xylosidase	14	14
Other CAZys related to wood decay		7	6
CE1	Esterases (acetyl-xylan, ferruloyl, cinnamoyl)	4	4
CE16	Esterases (acetyl-xylan, ferruloyl, cinnamoyl)	3	2

Supplementary files

Supplementary file S01 — CAZyome Annotation of P. echinulatum.

Table page S01.1 refers to CAZy families assignments; Table page S01.2 refers to CBM assignments.

Supplementary file S02 — PBD enzymes assignments in P. echinulatum 2HH and eleven related fungi.

Table page S02.1 refers to putative proteomes used in the PBD enzymes analyses; Table page S02.2 refers to PBD enzymes summarized in the twelve fungi; Table page S02.3 refers to cellulolytic enzymes summary in the twelve fungi; Table pages 0 to 11 refers to PBD enzymes assignments in the twelve fungi.

Supplementary file S03 — Identification and phylogenetic analyses of iBGLs and STs in P. echinulatum 2HH and eleven related fungi.

Table page S03.1 refers to putative proteomes used in the iBGLs and STs analyses; Table page S03.2 refers to the summary of iBGLs of GH1 and GH3 families and sugar transporters (PF00083); Table page S03.3 refers to iBGL ortholog groups, highlighting reviewed proteins; Table page S03.4 refers to GH1 protein sequences used in the phylogenetic analysis; Table page S03.5 refers to GH3 protein sequences used in the phylogenetic analysis; Table page S03.5 refers to GH3 protein sequences used in the phylogenetic analysis; Table page S03.6 refers to protein ids of the sugar transporters (PF00083) identified in the twelve fungi; Table page S03.7 refers to sugar transporter protein sequences used in the phylogenetic analysis.

INTERNATIONAL
JOURNAL OF SYSTEMATIC
AND EVOLUTIONARY
MICROBIOLOGY

Taxonomy, comparative genomics and evolutionary insights of *Penicillium ucsensis*: a novel species of the *Oxalica* Series

***Alexandre Rafael Lenz*[1,3,*], *Jos Houbraken*[4], *Eduardo Balbinot*[1], *Fernanda Pessi de Abreu*[1], *Nikael Souza de Oliveira*[1], *Roselei Claudete Fontana*[2], *Scheila de Avila e Silva*[1], *Marli Camassola*[2] *and Aldo José Pinheiro Dillon*[2]**

[1]*Bioinformatics and Computational Biology Laboratory, Institute of Biotechnology, University of Caxias do Sul, Francisco Getúlio Vargas Street 1130, 95070-560 Caxias do Sul, RS, Brazil.*
[2]*Enzymes and Biomass Laboratory, Institute of Biotechnology, University of Caxias do Sul, Francisco Getúlio Vargas Street 1130, 95070-560 Caxias do Sul, RS, Brazil.*
[3]*Bahia State University, Silveira Martins Street 2555, 41150-000 Salvador, BA, Brazil*
[4]*Westerdijk Fungal Biodiversity Institute, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands*

***Correspondence:** Alexandre R. Lenz, arlenz@ucs.br/alenz@uneb.br

ABSTRACT

The former *Penicillium echinulatum* 2HH was isolated from the digestive tract of *Anobium punctatum* larvae in 1979. The wild-type originated the S1M29 mutant by a long-term mutagenesis process, resulting in a cellulase hypersecretory strain. Both genomes were sequenced and analyzed in order to identify molecular features related to cellulase hyperproduction and albinism phenotype of the mutant. This investigation led to hypothesize the requirement of a taxonomy reclassification. Hence, the genome sequence data were used, not only to figure out about the cellulase hyperproduction, but also to reposition the species classification, including also insights on the evolutionary relationships of cell wall-associated proteins. The phylogenetic results lead us to the description of the 2HH wild-type as a novel *Penicillium* species placed in the *Oxalica* Series, for which the name *Penicillium ucsensis* sp. nov. is proposed. The genomic comparison of S1M29 and 2HH strains highlighted single amino-acid substitutions in two major proteins (BGL2 and FlbA) that can be associated to the hyperproduction of cellulases. The study of the DHN-melanin pathway shows that the S1M29 albinism phenotype resulted from a single amino-acid substitution in the enzyme ALB1, precursor of the DHN-melanin biosynthesis. The composition of cell wall-associated proteins of *P. ucsensis* shows 5 less chitinases; considerable differences in the number of proteins related to β-D-glucan metabolism; and two potential pseudogenes; when compared to *Penicillium oxalicum* 114-2, its closest free-living relative available. We suggest that these differences could be potentially explained by specific-environment interactions resulting from a possible long-term mutualistic symbiosis between the fungus and their host.

Keywords

Filamentous fungi; *Penicillium ucsensis*; *Oxalica* series; revised taxonomy; lignocellulolytic enzymes; cellulase; albinism.

Author Notes

Five supplementary tables and four supplementary figures are available with this article manuscript.

The authors contributed equally to this work.

Abbreviations

SRA, Sequence Read Archives
WGS, Whole Genome Shotgun
UCS, Universidade de Caxias do Sul
DHN, 1,8-dihydroxynaphthalene
DOPA, L-3,4-dihydroxyphenylalanine
DDBJ, DNA Data Bank of Japan
ENA, European Nucleotide Archive
NCBI, National Center for Biotechnology Information
NR, non-redundant database
ITS, fragments containing the internal transcribed spacers (ITS1 and ITS2), the 5.8S subunit, and the D1/D2 region of the 28S subunit
BenA, β-tubulin
CaM, calmodulin
RPB2, subunit of RNA polymerase II
ITS-RefSeq, ITS Targeted Loci project that provides a curated set of records with public collection data and correct taxonomic names
tRNA, transfer RNA
rRNA, ribosomal RNA
AICc, Corrected version of Akaike Information Criterion
BIC, Bayesian Information Criterion
ML, Maximum Likelihood
pp, posterior probabilities
bs, bootstrap support
QC, quality control
GO, gene ontology
KO, KEGG Orthology
EC, Enzyme Commission
COG, Cluster of Orthologous Groups
CAZyme, Carbohydrate-Active Enzyme
GH, Glycoside Hydrolase
GT, Glycosyl Transferase
SNP, single-nucleotide polymorphism
MEA, Malt Extract agar
MEAbl, Blakeslee's Malt extract agar
CYA, Czapek Yeast Autolysate agar
IHMM, In-house Maintenance medium
DG18, Dichloran Glycerol agar
YES, Yeast Extract Sucrose medium

Repositories

P. ucsensis 2HH/CBS 146492 wild-type: BioProject PRJNA520890 and Whole Genome Shotgun (WGS) project deposited at DDBJ/ENA/GenBank under the accession WIWU00000000. The version described in this paper corresponds to WIWU01000000. The 2HH/CBS 146492 strain was deposited at the Laboratory of Enzymes and Biomass - Collection of Fungal Biotechnology Cultures at the University of Caxias do Sul (UCS) and also deposited at the CBS collection (www.cbs.knaw.nl).

P. ucsensis S1M29 mutant: BioProject PRJNA521489 and WGS project deposited at DDBJ/ENA/GenBank under the accession WIWV00000000. The version described in this paper corresponds to WIWV01000000. The S1M29 mutant was deposited at the Laboratory of Enzymes and Biomass - Collection of Fungal Biotechnology Cultures at the UCS.

INTRODUCTION

Enzymes are notable for their various industrial applications, such as paper, food, animal feed, chemicals and biofuels. In this context, the degradation of plant polysaccharides by fungal enzymes is widely employed in large scale. In fact, the screening for enzyme-producing microorganisms, especially cellulose-degrading fungi, has been the subject of research for many decades [1].

In some instances, symbiotic relationships are established between cellulose-degrading microorganisms and organisms that consumes plant biomass but lacks the machinery to process it. One example of this mutualistic exchange are symbiont fungi found in the digestive tract of wood beetle larvae, such as *Anobium punctatum*. Moreover, the breakdown of cellulose in these beetles is achieved through a collective effort between various insect digestive, bacterial and fungal enzymes [2].

In 1979, a fungus of the genus *Penicillium* was isolated from the digestive tract of *A. punctatum* larvae, found on a wood wall mural in the Rectory building of the University of Caxias do Sul, Rio Grande do Sul-Brazil. This filamentous fungus was found as a symbiont living in the gut of the beetle and the strain was named 2HH. The isolation of this fungus was published in the International Symposium on Genetic Engineering, that occurred in São Paulo-Brazil [3].

The 2HH wild-type was previously classified by morphology as *P. echinulatum* in the 90's. Long-term 2HH strain improvement studies published before this work use this classification [4-19]. At that time, misleading classifications occurred more often, since a morphological concept was used for classification and identification. The reliability of molecular markers facilitates the taxonomy and classification, however, the identification of novel species of the genus *Penicillium* still proves problematic [20].

Long-term 2HH strain improvements resulted in the S1M29 mutant, obtained from the 9A02S1 mutant through the employment of hydrogen peroxide mutagenesis and a selection of mutants in a medium supplemented with 2-deoxyglucose [11]. The S1M29 mutant provides a better biomass hydrolysis and is the best mutant to date, due to a significant increase in enzyme titers as a result of several accumulated mutations [17,18]. Moreover, the lack of melanin production in the S1M29 mutant comprises another important phenotypic difference between the 2HH wild-type and the mutant.

Whole genomes of the 2HH wild-type and the S1M29 mutant were sequenced in 2013 and analyzed in this study in order to identify molecular features related to cellulase hyperproduction. Along these lines, the genomic data were used to track the origins of the albinism and the hyperproduction of cellulases between 2HH and S1M29. In addition, this investigation led us to hypothesize the requirement of a taxonomy reclassification and genomic sequences were used for molecular identification and species repositioning.

The enzymes secreted by the 2HH wild-type provide an effective enzyme formulation for complete saccharification of plant residues, being rich in cellulases and hemicellulases [17]. In this sense, we hypothesized the existence of a stable symbiosis association between the 2HH strain and *A. punctatum* larvae, which could require adaptations to survive in the available growth conditions. Consequently, the larvae diet and the symbiotic environment could result in genomic adaptations in relation to free-living relatives. Additionally, this potential stable association provides opportunities to explore genome content in terms of comparisons with related fungi.

Microorganisms living within an insect have an advantage over those that are free-living, because within the gut the fungi are bathed by a regular supply of nutrients. Another benefit might be direct dispersal. Free-living fungi usually deplete their substrate, and dispersal to a new substrate occurs by wind, water, or animals. In the case of organisms that live within insects, however, dispersal shifts become highly dependent on the insects. This mode of hijacking transportation can result in symbiotic lineages, which remain within the host with vertical transmission [21].

On the whole, the cell wall is potentially the part of the cell that exhibits the most phenotypic diversity and plasticity. Many cell wall elements are conserved in different fungal species, while other components are species-specific. Furthermore, the cell wall in filamentous fungi is a highly dynamic structure subject to constant change. For instance, distinct cell wall proteins and glucans are generated at each stage of the fungal life cycle; for example, during spore germination, hyphal branching and septum formation. Throughout this time, the cell wall can be drastically altered according to the type of cell that proliferates. Cell wall composition is also highly regulated in response to stress and environmental conditions, influencing fungal ecology [22,23]. The overwhelming diversity in cell wall structure combined with variant and recurrent features comprise an attractive set of elements to explore genome evolution.

The fungal cell walls are dynamic structures that play a critical role in fungal survival, growth, and morphology. Its major constituents are chitin, chitosan, β-1,3-glucan, β-1,6-glucan, mixed β-1,3-/β-1,4-glucan, α-1,3-glucan, melanin, and glycoproteins. In general, the fungal cell wall is generated by the cross-linking of glucans, chitin, melanin and other cell wall proteins to create a three-dimensional matrix [24].

Melanin pigments are formed by oxidative polymerization of phenolic compounds. These high molecular weight amorphous polymers are widely found in bacteria, fungi, plants and animals. Many fungi synthesize melanins, and several types of melanin are known to exist in the fungal kingdom [25]. Melanin provides defense against environmental stresses such as ultraviolet light, oxidizing agents and ionizing radiation. In addition, it contributes to the ability of fungi to survive in harsh environments. Also, fungal melanin contributes to virulence in an array of human pathogen fungi, including *Cryptococcus neoformans*, *Paracoccidioides brasiliensis*, *Aspergillus fumigatus* and *Talaromyces marneffei* [25,26].

Fungi may produce melanin via distinct pathways: the eumelanin via the 1,8-dihydroxynaphthalene (DHN) and L-3,4-dihydroxyphenylalanine (DOPA) pathways, and the pyomelanin via L-tyrosine degradation pathway. DHN-melanin, responsible by dark polymers production, is probably the best characterized fungal melanin biosynthetic pathway. Melanin biosynthesis homologues of these three pathways have been characterized in several filamentous fungi. However, chemical characterization of melanin can be a challenging task as the pigment is highly heterogeneous, insoluble in organic solvents, hydrophobic, and resistant to chemical degradation [27].

In this study, we present the first draft genome assembly of this fungus. *P. ucsensis* was nominated in recognition to the UCS, where the 2HH wild-type was isolated and also where it has been studied for the last 40 years. The draft genomes of the 2HH wild-type and the S1M29 mutant were assembled, annotated and published at DDBJ/ENA/GenBank. The massive sequencing data allowed to explore the differences between the wild-type and the mutant, and particularly, allowed the molecular identification of the wild-type, leading to species revision of this fungus. This study also revealed surprising insights into genome evolution of this species, considering that genome information from non-model organisms is highly important, as they represent specific phenotypes that aid in disengaging the common parts of gene sets from those that have evolved as adaptations to specific ecosystems like symbiotic evolution.

METHODS

2.1 Taxonomy and species revision

The standard polyphasic working method for *Penicillium* species descriptions and identifications was adopted. This method is composed by macro and micromorphological observation, extrolite analysis and molecular phylogenetic analysis [20].

2.1.1 Cultures and morphological observation

Notes: Future work

2.1.2 Extrolite analysis

Notes: Future work

2.1.3 Molecular phylogenetic analysis

For molecular identification, sets of sequences were generated for the four *Penicillium* standard molecular markers (ITS, BenA, CaM and RPB2). ITS dataset comprises ribosomal DNA (rDNA) fragments containing the Internal Transcribed Spacers (ITS1 and ITS2), the 5.8S subunit, and the D1/D2 region of the 28S subunit. The BenA dataset comprises partial sequences of the β-tubulin gene (*benA*). The CaM dataset comprises partial sequences of the calmodulin gene (*caM*), and the RPB2 dataset is composed by partial sequences of the subunit of RNA polymerase II gene (*rpb2*).

Complete sequences of each marker were identified and annotated during the whole genome annotation of the 2HH wild-type (Section 2.2.6.1). The set of sequences for each marker was obtained by combining the newly obtained sequences from 2HH wild-type with reference sequences of *Penicillium* genus (preferably ex-type) [20, 28-31]. Publicly available sequences at NCBI (NR/WGS/ITS-RefSeq) were used in this study. Strains and respective sequence accession numbers are reported in Supplementary Table S01.

Five DNA sequence files were generated, one file for each of the four loci (ITS, BenA, CaM and RPB2) and a fifth file concatenating the sequences of the four loci. In addition, the concatenated data set was obtained by FaBox (v1.5) [32], missing sequences from CaM and RPB2 loci were replaced by N's to indicate missing data. Four data partitions were defined for the concatenated dataset, describing ITS, BenA, CaM and RPB2. Alignment of each sequence file was performed using the MAFFT online service (v7.452) [33], then the manual adjustment was made to optimize the homology using AliView (v1.26) [34].

Successively, JModelTest2 (v2.1.6) [35] was used to find the preferred model of evolution for each dataset, models were selected according to the Corrected version of Akaike Information Criterion (AICc) for both tools. When taxon number and heterogeneity are small, AICc likely to perform well [36]. The following tools were used to infer the phylogenetic trees: a) MrBayes (v3.2.6) [37] for posterior probabilities (pp); and b) RAxML-HPC2 (v8.2.8) [38] for bootstrap support (bs).

The CIPRES Science Gateway (v3.3) [39] was used to perform the analyzes: a) MrBayes analysis, setting GTR (nst=6) + GAMMA, 10^7 generations, sampling every 1,000 generations with a burnin fraction of 0.25; and b) RAxML-HPC2 Workflow analysis, setting GTR + GAMMA, executing Maximum Likelihood (ML) search and thereafter a thorough bootstrap with 1,000 iterations.

Trees were visualized in FigTree (v1.4.4) [40] and edited in Inkscape (v0.92.2) [41]. Posterior probabilities (pp) values and bootstrap support (bs) values are labelled across the top of a branch. Values less than 0.95 pp and 80% bootstrap support are not shown. Branches with full support in Bayesian and RAxML analyses are thickened. Values below 0.95 pp and 80% are not shown and indicated with a hyphen. *P. echinulatum* (sect. *Fasciculata*) was chosen as an outgroup.

2.2 Genome assembly, annotation and general features

2.2.1 Previous Shotgun Genome Sequencing

High-molecular-weight genomic DNA was extracted from 2HH wild-type and S1M29 mutant using a protocol for DNA isolation [42]. The extracted DNA was used to generate libraries for Illumina Sequence by Synthesis (Illumina-SBS) using an unmodified Illumina TruSeq DNA protocol [43]. Both genomes were sequenced using Illumina HiSeq 2000 platform conducted by the commercial provider Ambry Genetics (Aliso Viejo, CA, USA) in 2013. Libraries of 100-bp paired-end (PE) were generated for each strain sequencing. 29,316,764,800 bp represents a raw coverage of 961× for the S1M29 mutant, and 25,514,587,400 bp represents a raw coverage of 836× for the 2HH wild-type. The high coverage was required to precisely identify the single-nucleotide polymorphisms (SNPs) that occurred during the S1M29 mutagenesis process.

2.2.2 Reads assessment and quality trimming

Raw data quality control (QC) assessment using FastQC (v0.11.5) [44] was performed to check the overall sequence quality, the GC percentage distribution and the presence/absence of overrepresented sequences. Using Trim Galore (v0.4.4) [45], small fragments (length, 30 bp) were abandoned, following adapter clipping and quality trimming. Low quality bases from the 3' and 5' ends were removed before being cut and quality trimmed (sequencing quality values, Q28).

2.2.3 Assemblies

The high-quality reads were assessed using KmerGenie (v1.7044) [46], to predict the best k-mer and genome size. The high-quality reads were assembled with SPAdes (v3.11.0) [47]. SPAdes is a multi-k-mer assembler, we used odd values ($21 \leq k \leq 83$). Further, scaffolds shorter than 500 bp were removed. Quast (v5.0.2) [48] with default parameters was used to statistically evaluate the assemblies. BUSCO (v3.0.2) [49] assembly mode was used to assess the assemblies, providing quantitative measures based on evolutionarily expectations of gene content from near-universal single-copy orthologous selected from OrthoDB (v9) [50]. Five conserved orthologous datasets were used for evaluation: *Eukaryota* (303 genes), *Fungi* (290 genes), *Dikarya* (1312 genes), *Ascomycota* (1315 genes) and *Eurotiomycetes* (4046 genes).

2.2.4 Previous RNA-seq for gene model prediction and annotation

Non-stranded RNA-seq libraries were constructed using the Illumina TruSeq protocol [51] with poly-A selection from the S1M29 mutant to help in gene model prediction. RNA sequencing was performed using the Illumina Hiseq 2000 platform in 2014. Reads assessment and quality trimming were performed in the same way as genome sequences. Paired-end reads were aligned over both genomes using hisat2 (v2.0.5) [52], and assembled into transcripts using Trinity (v2.5.1) [53]. In addition, Trinity de novo transcriptome assembly was performed. All transcripts were used to train gene finders for gene model prediction.

2.2.5 Gene calling

Repeat sequences of both assemblies were masked using RepeatMasker (v4.0.7 2) [54], RepeatModeler (v1.0.8 1) [54] and RepBase library (2017-01-27) [55]. Repeat masked assemblies were used for coding gene prediction, performed independently with a set of gene finders. The first group uses an *ab-initio* approach to predict genes directly from nucleotide sequences, including Augustus (v3.2.2) [56] and GeneMark-ES (v4.33) [57]. Augustus parameters were trained on gene models of the S1M29 mutant with the transcriptome data as hints. Pre-trained parameters from *Aspergillus* (*A. fumigatus*, *A. nidulans*, *A. oryzae* and *A. terreus*) were also used in Augustus predictions.

The second group uses a similarity-based approach to identify gene structure using a sequence alignment between genomic sequence and transcript or protein databases. BLAT (v36) [58] and GMAP-GSNAP (v2017-06-20) [59] were used to align the RNA-Seq transcripts of the S1M29 mutant, while Exonerate (v2.2.0) [60] was used to align UniProtKB/Swiss-Prot [61] and the proteome of *P. oxalicum* 114-2 [62]. Additionally, tRNAscan-SE (v2.0.3) [63] was used to predict transfer RNAs (tRNAs). EVidence Modeler (v0.1.3) [64] was used to take all gene prediction inputs, outputting consensus gene models and generating the predicted gene

set for each strain. *Penicillium* spp. was used to revise and complement the predicted genes by homology searches using Exonerate [65], GeneWise [66] and SoftBerry Fgenesh+ [67]. The annotation completeness was assessed by running BUSCO (v3.0.2) [49] in protein mode, using the same datasets as the assembly assessment.

2.2.6 Functional annotation

Predicted proteins were functionally annotated using the standard protocol [68], implemented by the tool Seq2Annot. Seq2Annot is an in-house web tool that implements a workflow for genomic annotation and manual curation. Functional annotation is a process that involves several biological databases, each one with its peculiarities and specific output formats. Structural and functional information about genes can be edited, supporting the manual curation process. After completing the annotation, the system allows the generation of the files required for WGS deposit at GenBank.

All predicted gene models were functionally annotated using: SignalP Server (v5.0) [69] to predict the presence and location of signal peptide cleavage sites; TMHMM Server (v2.0) [70] was used to identify transmembrane helices and membrane-bounds; InterProScan (v5.25-64.0) [71] was used to map Interpro families, domains and gene ontology (GO) terms; hmmscan (v3.1b2) [72] was used to identify PFAM domains over PFAM database (v31.0 - 2017-02) [73] using gathering cutoffs; KO (KEGG Orthology [74]) assignments and automatically generated KEGG pathways were assigned to predicted proteins using KAAS [75]; KEGG hints were also used to assign Enzyme Comission (EC) numbers; EggNOG-Mapper (v2) [76] was used to map general functional categories from Clusters of Orthologous Groups (COGs) and orthologous precomputed eggNOG clusters from the eggNOG (v5.0) [77] (ascNOG) database; secondary metabolite biosynthetic gene clusters were predicted by antiSMASH fungal (v3.0) [78], and mibig database (v1.3) [79].

Moreover, product assignment of molecular markers, putative peptidases, Carbohydrate-Active Enzymes (CAZymes), sugar transporters and transcription factors were analyzed manually, using orthologous from related fungi, BLASTp (v2.7.1+) [80] searches over UniProtKB/Swiss-Prot/TrEMBL [61], and specific databases, as described below.

2.2.6.1 Annotation of standard molecular markers

The WGS allowed the annotation of the complete sequences of the standard molecular markers of *Penicillium* genus. Five rDNA sequences were identified: the three subunits 18S, 5.8S and 28S; and the sequences of the internal transcribed spacers (ITS1 and ITS2). In addition, secondary identification markers were also annotated: β-tubulin BenA, calmodulin calcium-binding protein CaM and RNA Polymerase II subunit RPB2. These sequences are included in the WGS deposited at GenBank.

2.2.6.2 Annotation of CAZymes

Two approaches were combined to improve the CAZyome annotation accuracy: (i) dbCAN2 [81] server over HMMdb release 8.0, and (ii) BLASTp (v2.7.1+) [80] over CAZyDB (07/31/2019). The putative encoded protein sequences were compared to the full-length sequences of the CAZyDB. Query sequences that produced an e-value <10-6 and aligned over at least 95% with a protein in the CAZyDB with >50% identity, were assigned to the same

family as the subject sequence. All putative encoded protein sequences were also subjected to dbCAN2 searches over HMMdb using specific models for each CAZy module family, requiring both methods to yield the same family assignment.

2.2.6.3 Annotation of Peptidases

To assess the diversity of peptidases, protein sequences of whole proteomes were used in BLASTp (v2.7.1+) [80] searches (e-value cut-off = 1e-05) over the file 'merops scan.lib', which is a non-redundant library of 4,968 protein included in MEROPS database [82] (Release 12.0): aspartic (A), cysteine (C), serine (S), metallo (M) and threonine (T) peptidases, class with unknown activities (U) and peptidase inhibitors (I). Putative peptidases were classified according to their best hits in a BLASTp. In addition, for MEROPS assignment of *P. ucsensis*, all hints were also used as input to BLASTp searches over 'Peptidase Full-length Sequences' from MEROPS database and over UniProtKB/Swiss-Prot [61] to assign protein products.

2.2.6.4 Annotation of Sugar Transporters

To assess the diversity of sugar transporters, the domain PF00083 (08/05/2018) profile extracted from the PFAM database [73] was used to search over *P. ucsensis* proteomes with hmmsearch (v3.1b2) [72], choosing hmmsearch score ≥ 238 as cutoff [83]. This class of sugar transporters was chosen for its biotechnological relevance, related the production of cellulolytic enzymes.

2.2.6.5 Annotation of Transcription Factors

To assess the diversity of transcription factors, Interpro and PFAM domains of whole proteomes were used to search TF domains using those protein domains described in the library of TFs used by CIS-BP Database [84].

2.3 Comparative and evolutionary genomic analyzes

2.3.1 Tracking mutations in the S1M29 mutant

PogressiveMauve [85] (v2015-02-13) build 0 was used to align both *P. ucsensis* draft genomes. In order to identify amino-acid substitutions, protein alignments were obtained using MAFFT online service (v7.452) [33].

2.3.1.1 Origin tracking of the cellulase hyperproduction

Protein sequences of both proteomes (2HH and S1M29) were aligned in order to find amino-acid substitutions generated by the mutations. Proteins that contained amino-acid substitutions in the S1M29 mutant were classified in four levels of probable impact on cellulase hyperproduction, this classification was supported by experts in fungal metabolism. The main proteins affected by the SNPs, classified at levels 3 and 2, exhibit potential involvement in the expression of cellulase coding genes. These levels include transporter proteins, transcription factors and other proteins strongly related to gene regulation. In contrast, levels 1 and 0 comprise a diverse series of proteins with low probability of impact in the expression of cellulolytic enzymes.

2.3.1.2 Origin tracking of albinism

Previously released melanin-associated proteins from *A. fumigatus* Af293, *Aspergillus niger* CBS 513.88, *Penicillium chrysogenum* P2niaD18 [86] and *T. marneffei* ATCC 18224 [87] were used to identify orthologous melanin-related proteins in *P. ucsensis* and *P. oxalicum* 114-2. Proteomes obtained from UniProtKB [61] were used to find orthologous groups in whole genome-wide searches using ProteinOrtho (V5.16b) [88].

Macromorphology and other phenotypic characters were observed in four culture media: (i) Malt Extract agar (MEA); (ii) Blakeslee's Malt extract agar (MEAbl); (iii) Czapek Yeast Autolysate agar (CYA); and (iv) In-house Maintenance Medium (IHMM). The first three culture media are described in [20], and the last culture medium is described in [5].

Cultures were inoculated using a dense conidium suspension of both strains 2HH and S1M29 on the four media. After 24h, the inocula was transferred using a three-point method to the same medium in 9 cm glass Petri dishes and incubated in the dark at 28 °C. The cultures were examined after 7 and 14 days of growth and photos of the dishes were taken and edited in GIMP (v2.10.14) [89].

2.3.2 Evolutionary relationships

2.3.2.1 General evolutionary relationships

AAI-profiler [90] was used to perform a proteome-wide analysis by homology searches over UniProtKB proteomes and to infer the histogram of amino-acid sequence identity between *P. ucsensis* 2HH and *P. oxalicum* 114-2 [62].

Besides, the evolutionary relationships were inferred by a BUSCO approach. Firstly, 14,768 orthologous groups were found in whole genome-wide searches using ProteinOrtho (V5.16b) [88], secondly 2,383 (58.9%) highly conserved single-copy BUSCO proteins (Supplementary Table S02) from *Eurotiomycetes* dataset were selected and concatenated to infer the phylogenetic relationships (*Trichoderma reesei* and *Neurospora crassa* do not belong to the *Eurotiomycetes* class, reducing the number of selected conserved groups). Multiple alignments for this data matrix were obtained using MAFFT online service (v7.452) [33].

ModelTest-NG (v.0.1.5) [91] was used to find the preferred model of evolution for the concatenated dataset, the best-fit model (JTT+I+G4+F) was selected according to the Bayesian Information Criterion (BIC). RAxML-HPC2 (v8.2.8) [38] was used for bootstrap support (bs) using the model of evolution defined by ModelTest-NG on the concatenated dataset. The CIPRES Science Gateway (v3.3) [39] was used to perform RAxML-HPC2 Workflow analysis, setting JTT+GAMMA+I+F, executing ML search and thereafter a thorough bootstrap with 100 iterations for the concatenated dataset. The tree was generated and configured using iTOL [92].

2.3.2.2 Evolutionary relationships of cell wall-associated proteins

Previously released cell wall-associated proteins from *A. fumigatus* Af293 [24,86,93] and *N. crassa* OR74A [94] were used to find orthologous cell wall-related proteins in *P. ucsensis* and *P. oxalicum* 114-2. Proteomes obtained from UniProtKB [61] were used to find orthologous groups in whole genome-wide searches using ProteinOrtho (V5.16b) [88].

RESULTS AND DISCUSSION

3.1 Taxonomy and species revision

After genome assemblies, preliminary analyzes raised the suspicion of mistaken identification of the species, designated as *P. echinulatum* (Section *Fasciculata*), based on the morphological characterization carried out in the 1990's. As a result of this suspicion, the taxonomic revision of the isolate and the characterization of the novel species were carried out. Currently, the taxonomic structure of the genus is well defined, based on a polyphasic approach that includes morphology, extrolite profiles and molecular markers.

Penicillium ucsensis Lenz et al., ***sp. nov.*** MycoBank MBxxxxx. Fig. 1

Etymology: Named from the term UCS, which refers to the University of Caxias do Sul, considering that the fungus was isolated from the digestive tract of coleoptera larvae that were collected in the Rectory building of this University in 1979 and also where it has been studied for the last 40 years.

Type: Brazil, Coleoptera intestinal tract, 1979, collected by Carrau, J. L. & Ribeiro, R. T. (holotype CBS H-24331, culture ex-type: CBS 146492 = DTO 426-B1 = 2HH).

In: *Penicillium* subgenus *Aspergilloides*, section *Lanata-Divaricata*, series *Oxalica*.

ITS barcode: Five sequences of rDNA were annotated in the WGS: 18S ribosomal RNA (PECH_004920), Internal transcribed spacer 1 (PECH_004924), 5.8S ribosomal RNA (PECH_004925), Internal transcribed spacer 2 (PECH_004926) and 28S ribosomal RNA (PECH_004965). In addition, the secondary identification markers were annotated: BenA (PECH_001952), CaM (PECH_007232) and RPB2 (PECH_007722). These sequences are included in the WGS deposited at GenBank.

Description: Colony diam, 7 d, in mm: CYA 57–60; CYA15°C 19–20; CYA30°C 58–61; CYA37°C no growth; DG18 26–30; MEA 45–50; YES 55–59; creatine agar 18–22, poor growth, acid and base production absent. CYA, 25°C: Colonies radially sulcate, slightly raised in the centre; margin entire; mycelium white; sporulation absent; soluble pigments absent; exudates absent; conidial color not determined; reverse pale yellow-brown. YES, 25°C: Colonies radially sulcate, slightly elevated; margin entire, slightly feathery; mycelium pale brown; texture velvety; sporulation absent; soluble pigments absent; exudates absent; reverse pale brown. MEA, 25°C: Colonies slightly radially sulcate in centre, moderately high; margin entire; mycelium white; colony texture floccose; sporulation absent at the edge, poor in centre and good in a ring between centre and edge; soluble pigments absent; exudates absent; conidial colour en masse dull green; reverse unchanged in centre, light brown at edge. DG18, 25°C: Colonies plane, raised at centre; margin entire; mycelium white; texture velvety; sporulation absent; soluble pigments absent; exudates absent; reverse brownish white in centre, uncolored at edge. Sclerotia absent. Conidiophores xx–xx long, xx–xx µm in width, smooth, thin, predominantly monoverticillate, rarely biverticillate, non-vesiculate; phialides cylindrical, xx–xx × xx–xx µm; conidia broadly ellipsoidal, rough walled, xx–xx × xx–xx µm.

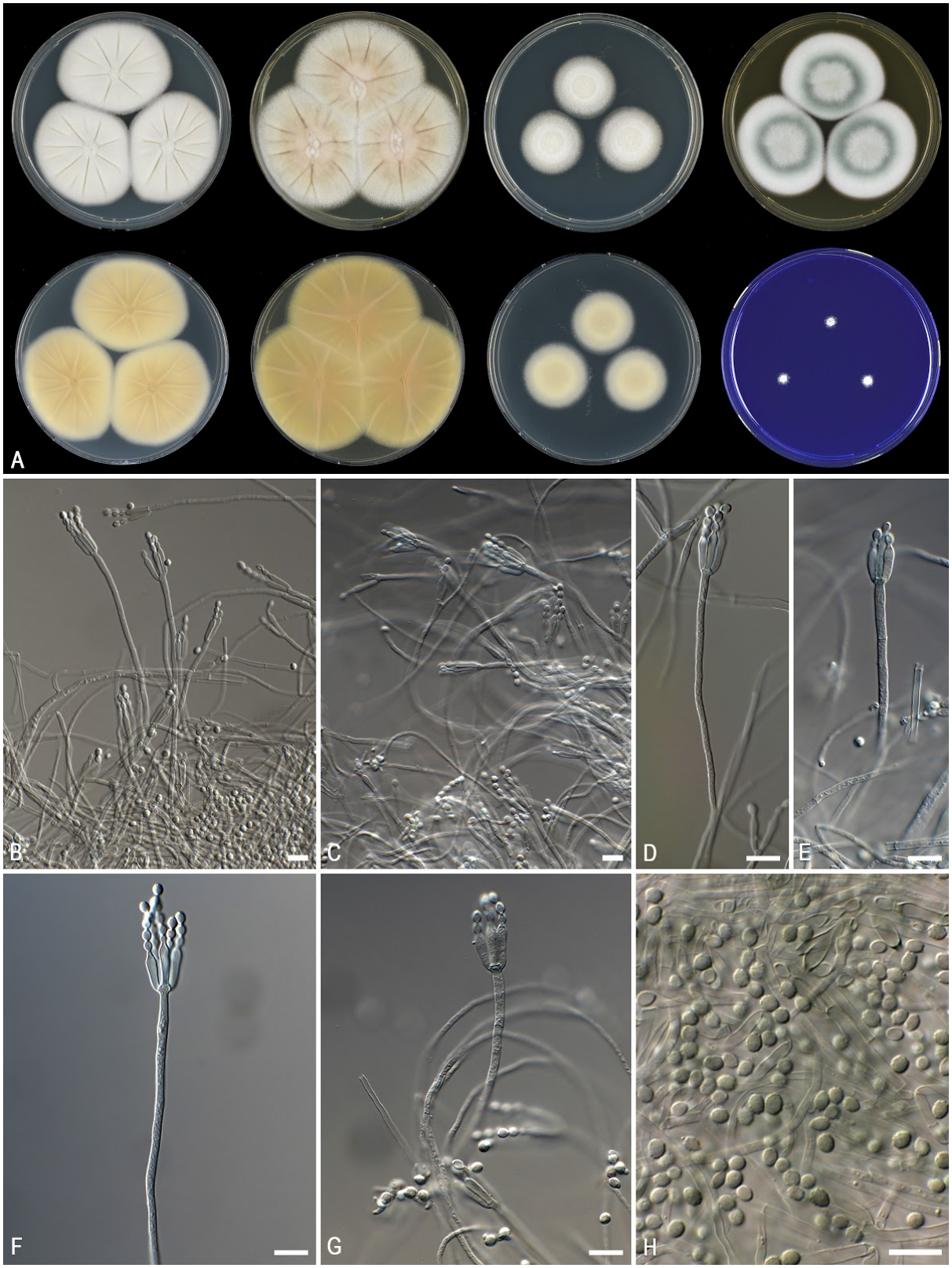


Figure 1. Photographs of macromorphological and micromorphological observation of *P. ucsensis* 2HH. A. Colonies: top row left to right, obverse CYA, YES, DG18 and MEA; bottom row left to right, reverse CYA, reverse YES, reverse DG18 and CREA. B–G. Conidiophores. H. Conidia. Scale bars: B–H = 10 µm.

Notes: Future work, unfinished descriptions

3.1.1 Cultures and morphological observation

Notes: Future work

3.1.2 Extrolite analysis

Notes: Future work

3.1.3 Molecular phylogenetic analysis

For molecular identification of the 2HH wild-type, the rDNA region, BenA, CaM and RPB2 were aligned with reference *Penicillium* sequences obtained from GenBank (detailed in Supplementary Table S01). Phylogenetic analyses were performed including individual and concatenated datasets. The results of each analysis are presented in the phylogenetic trees. All phylogenetic trees were configured in the same way: (i) thickened branches correspond to 1.00 posterior probability (pp) and 100% bootstrap support (bs); (ii) branch support in nodes higher than 0.95 pp and/or 80% bs are shown; (iii) hyphen = support lower than 0.95 pp and/or 80% bs; (iv) *P. echinulatum* (sect. *Fasciculata*) was chosen as outgroup; (v) ^{T} = ex-type; (vi) Species analyzed in this study is indicated in bold; (vii) Sequence accession numbers are available in Supplementary Table S01.

The ITS phylogenetic analysis (Supplementary Fig. 1) showed, with high statistical support, that 2HH wild-type belongs to the Section *Lanata-Divaricata*, being a phylogenetic sister species of *P. oxalicum*, located in the clade of newly characterized species *Penicillium diatomitis* and *Penicillium soosanum* [31]. In addition, *Penicillium* sp. SFC101850 [95] also appears in the same clade. These related species were identified as Series *Oxalica*.

ITS is not variable enough for distinguishing all closely related *Penicillium* species. Because of the limitations associated with ITS as a species marker in *Penicillium*, a secondary identification marker is often needed for identifying isolates to species level [20]. The phylogenetic analysis of BenA as a secondary identification marker (Supplementary Fig. 2) revealed the presence of two well-supported clades that delineate the Series *Oxalica*: the clade of *P. oxalicum* and the clade of *P. diatomitis*, *P. soosanum*, 2HH isolate and SFC101850 isolate.

The phylogenetic analysis of CaM indicates the same mentioned clades of BenA analysis (Supplementary Fig. 3). However, the thickened branches denote that CaM analysis achieved a better statistical support than BenA. It is important to note the lack of sequences of this marker for many strains important for this study. In the phylogenetic analysis of RPB2 (Supplementary Fig. 4), the clade of *P. oxalicum* is more distant in relation to the 2HH isolate. Secondary markers are essential to distinguish *P. oxalicum* from the clade of *P. diatomitis*, *P. soosanum* and 2HH isolate. Still, single markers were not sufficient to obtain high support and to differentiate 2HH from *P. diatomitis* and *P. soosanum*, requiring a combined phylogenetic analysis with more than one molecular marker.

Phylogenetic analysis with all concatenated markers includes the description of specific partitions for each marker, using the same evolutionary model configurations of each single

marker, already mentioned in the individual trees. The combined phylogenetic analysis (Fig. 2) revealed two main clades. The first clade shapes a monophyletic group representing the Series *Oxalica*, which is grouped separately from any of the other related species of Section *Lanata-Divaricata*. These results are corroborated by previous studies that characterize the Series *Oxalica* [31].

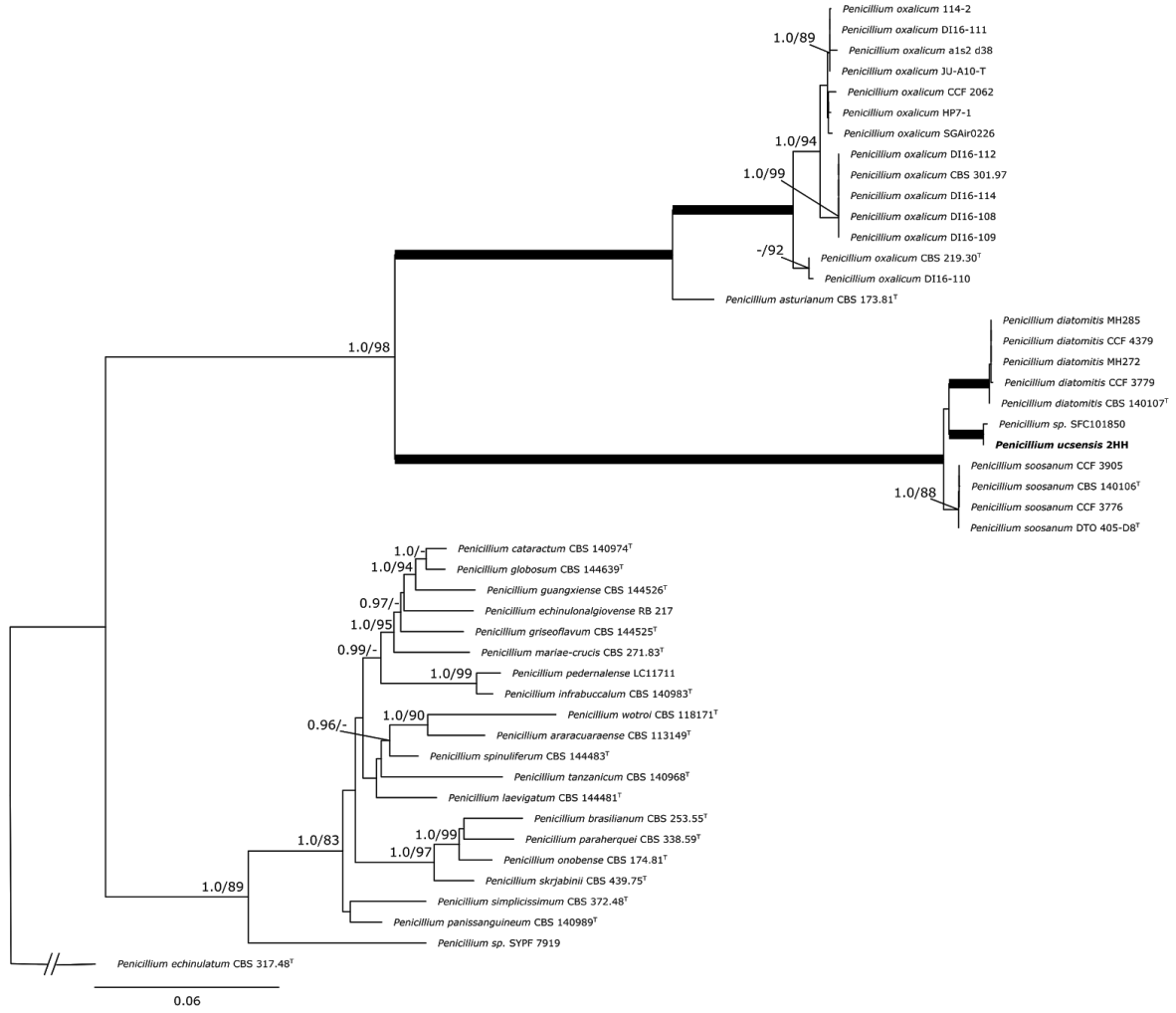


Figure 2. Phylogenetic tree for ITS, BenA, CaM & RPB2 concatenated

For the Series *Oxalica*, there are two fully supported clades. The first clade comprises isolates designated as *P. oxalicum*, including the ex-type strain of *P. oxalicum* CBS 219.30. While in the second clade a triple division is observed, where the clade of *P. diatomitis* appears fully supported and the clade of *P. soosanum* appears with support equal to 1.0pp/88bs. Finally, the combination of all reference markers revealed a fully supported clade, differentiated from *P. diatomitis* and *P. soosanum*, containing the 2HH and SFC101850 isolates.

Penicillium sp. SFC101850 [95], a strain isolated from *Arctoscopus japonicus* (sailfin sandfish) egg masses in Korea, was the only isolate belonging to the same clade as the 2HH isolate. It is important to note that CaM and RPB2 are not available for this isolate. The fully supported clade where the 2HH and SFC101850 isolates are placed, supports a novel species of the Series *Oxalica*.

3.2 Genome assembly, annotation and general features

Whole genomes of *P. ucsensis* (2HH and S1M29) were sequenced using Illumina HiSeq 2000 platform in 2013, conducted by the commercial provider Ambry Genetics. These sequencing data were used in previous studies [17], however, WGS have not been published yet. In this study we publish the Sequence Read Archives (SRA) and WGS of both strains to the scientific community.

The genome assembly of the 2HH wild-type comprises 697 scaffolds, totalizing 30.43 Mb and scaffold N50 equals 151.4 kb. While the genome assembly of the S1M29 mutant comprises 673 scaffolds, totalizing 30.41 Mb, and scaffold N50 is 185.175 kb. Furthermore, the GC contents of both assemblies are 50.3% and unclosed gap regions represented less than 2% for each strain assembly as assessed by Quast (Table 1).

Table 1. Genome features of *P. ucsensis* draft genomes

Features (# means the number)	2HH	S1M29
Raw reads (bp)	255,145,874	293,167,648
Paired-end read length (bp)	100	100
Filtered reads (bp)	239,270,048	273,856,380
Genome assembly size (Mb)	30.43	30.41
# of scaffolds	697	673
GC content (%)	50.3	50.3
# of N's	533	621
Scaffold N50	151,400	185,175
Scaffold N75	86,997	98,992
Scaffold L50	62	50
Scaffold L75	126	109
Average Scaffold (bp)	43,655	45,180
Largest Scaffold (bp)	511,359	802,811
# of gene models	8,366	8,375
# of tRNAs	188	197
# of putative proteins	8,173	8,173
Average of exons per gene	3.2	3.2
Average of introns per gene	2.7	2.7
Smallest protein length (aa)	51	51
Average protein length (aa)	519.2	519.2
Largest protein length (aa)	7,243	7,243
# of Complete and single-copy BUSCOs of *Eurotiomycetes* class	3,962	3,963
# of Complete and duplicated BUSCOs of *Eurotiomycetes* class	10	10
# of Fragmented BUSCOs of *Eurotiomycetes* class	40	39
# of Missing BUSCOs of *Eurotiomycetes* class	34	34
BUSCO Completeness (%) of *Eurotiomycetes* class	98.1	98.1
# of PFAM annotations	9,505	9,503
# of InterPro annotations	22,602	22,591
# of GO terms annotations	15,103	15,099
# of Signal peptide annotations	652	651
# of Transmembrane annotations	1,666	1,668
# of BUSCO annotated (All datasets)	7,241	7,245
# of EGGNOG annotated	7,982	7,982
# of COG annotated	7,528	7,527
# of KEGG annotated	3,510	3,512
# of MEROPS annotated	276	276
# of CAZymes annotated	388	389
# of Secondary metabolite clusters	72	71

We evaluate both assemblies using five BUSCO conserved orthologous datasets. The completeness average of the assemblies is higher than 98%, where 1.2% of the core orthologous are missing. The proportion of fragmented orthologous comprise less than 0.6%.

Next, we performed gene calling independently with a set of gene finders, resulting in 8,366 gene models predicted for 2HH, while 8,375 gene models were predicted for S1M29 (Table 1). The annotation completeness was assessed by running BUSCO in protein mode, using the same datasets of the assembly assessment. The average of complete proteins is 99.4%, where just 36 of 7266 orthologous are missing in each strain. The proportion of fragmented proteins comprise less than 0.4%.

Previous proteome completeness of 28 fungal species showed that the average of complete proteins is higher than 95% in most species. High numbers of fragmented proteins (more than 150 in some species) were observed regardless of the genus, technology, assembler or automatic gene calling methodology [96]. Our draft assembly evaluation showed results above this average, allowing a plentiful gene calling and comparison with other fungi. However, misleading in gene prediction may be present given that the gene inspection focused only on interest groups encompassing: CAZymes, transcription factors, sugar transporters and peptidases.

The 8,173 predicted proteins for each strain were functionally annotated using standard protocols [68] and supported by the in-house web tool Seq2Annot. The summary results are presented in Table 1.

3.3 Comparative and evolutionary genomic analyzes

In recent years, *P. ucsensis* (still described as *P. echinulatum*) has been the subject of several studies employing random mutagenesis. The improvements in the strains accumulate several mutations, resulting in higher enzyme yield and leading to a significant increase in the volume of secreted cellulases compared to the parental strain 2HH. In 2011, after several rounds of mutagenesis, the mutant S1M29 was obtained. This mutant has the best enzymatic yields for cellulases of this species [11,14,17].

Despite the efforts and advances achieved with classical mutagenesis, molecular knowledge of the mutations accumulated in the mutant S1M29 are important to design new strains. In this sense, ultra-deep WGS data of both strains allowed the accurate SNP calling, discovering variations between the mutant S1M29 and the parental strain 2HH.

3.3.1 Tracking mutations in the S1M29 mutant

Both draft genomes were aligned using progressiveMauve resulting in 4.3% missed and/or extra bases between the drafts. These differences may be explained by the haziness of draft assemblies, particularly in this case due to lack of long-reads. The progressiveMauve also identified 1,337 SNPs, the most common type of genetic variation. We identified 8,067 identical proteins in both strains, representing 98.7% of putative proteins. In addition, 52 proteins were affected by SNPs and contain amino-acid substitutions, representing 0.64% of the proteome. Alignment results and amino-acid substitutions are available in Supplementary Table S03.

The origin tracking of the two main characteristics that differentiate the S1M29 mutant from the parental 2HH are presented below: i) the hyperproduction of cellulases and ii) the absence of green pigmentation in solid culture medium.

3.3.1.1 Origin tracking of the cellulase hyperproduction

There are no significant differences in the size of the genome and in the composition of proteins, the majority of the sequences of the CAZymes are also identical in the S1M29 mutant and in the 2HH wild-type. Also, we verified that the well-known transcription factors of *Penicillium* species, involved in cellulases and hemicellulases transcription regulation, have the same amino-acid sequences in both strains. Consequently, the variation on enzymatic production by both strains likely could be explained at the transcription level and not due to possible large-scale genomic rearrangements.

In order to classify the degree of the impact of the mutations in the improvement of cellulase expression of the mutant strain, we categorized proteins affected by SNPs into four levels of likely impact. The main proteins affected by the SNPs, classified at levels 3 and 2 of potential involvement on cellulase expression increment, are summarized in Table 2. In that, it is possible to verify the amino-acid substitution represented by the amino-acid in the parental 2HH, its location in the protein sequence and the respective mutation identified in the S1M29 mutant and the annotation associated.

Table 2. Major proteins potentially involved in cellulase hyperproduction by the S1M29 mutant

Protein Id 2HH / S1M29	Mutation	Impact	Annotation
PECH_005648 / PECM_002864	D194P	3	β-glucosidase BGL2
PECH_004634 / PECM_006143	S259P	3	Developmental regulator FlbA
PECH_007011 / PECM_006829	Y370D & I381E	2	Zn2Cys6 binuclear cluster domain
PECH_006810 / PECM_005158	P277S	2	MFS-type transporter
PECH_002170 / PECM_001690	L203F	2	Aromatic amino acid and leucine permease
PECH_008009 / PECM_007813	S371F	2	Pre-mRNA-processing factor
PECH_006814 / PECM_005162	A712D	2	Pre-mRNA-splicing helicase
PECH_008673 / PECM_006103	Promoter region	2	MFS-type Sugar/inositol transporter

Besides the mutations in BGL2 and FlbA, classified with a potential 3 of positive impact on the expression of cellulases, six other proteins were listed in Table 2, classified as level 2. This group comprises transporters and transcription factors that are commonly involved in numerous regulatory pathways, affecting directly or indirectly the cellular metabolism. In addition, pre-mRNA-splicing and pre-mRNA-processing factors may also play important roles in regulatory systems. While the level 1 encompasses proteins related to RNA polymerase, RNA interference, gene silencing, and less influential transcription factors and transporters. Finally, the level 0 covers proteins which no involvement on the cellulase expression system were found. Mutations classified as level 0 and 1 are shown in the Supplementary Table S03.

On the whole, there were no significant changes in encoding CAZymes regions between the parental and the mutant. We found just four single amino-acid substitutions in CAZymes. The first three enzymes of glycoside hydrolase (GH) and glycosyl transferase (GT) families probably do not impact the cellulase expression: (i) Mannan endo-1,6-α-mannosidase (A5V); (ii) Glycoside Hydrolase Family 78 protein (A124T); and (iii) Glycosyl Transferase Family 90 protein (A287V).

The fourth CAZyme containing amino-acid substitution corresponds to a key finding of this study. A single amino-acid substitution occurs at position 194 in BGL2, changing an aspartic acid in the parental 2HH (PECH_005648) for a proline in the mutant S1M29 (PECM_002864). Considering this SNP of the major intracellular β-glucosidase BGL2,

605 belonging to the GH1 family, it was reported that BGL2 orthologous play an important role in
606 induction of cellulases in *T. reesei* and in *P. oxalicum* 114-2.

607 In a *T. reesei* mutant, it was observed that a single-nucleotide mutation of *bgl2* explains the
608 enhanced cellulase expression, when the strains were cultivated on cellulose and cellobiose
609 [97]. Besides, in *P. oxalicum*, the major intracellular β-glucosidase BGL2 (PDE 00579) was
610 found to be dependent on ClrB at the transcription level and *bgl2* deletion facilitates the
611 synergistic expression of cellulase genes. Lack of this β-glucosidase facilitates the
612 accumulation of intracellular cellodextrins, which can trigger signaling cascades that include
613 expression of cellulase genes [98,99].

614 Evolutionary analysis of BGL2 was performed using PANTHER-PSEP [100]. PSEP
615 (position-specific evolutionary preservation) measures the length of time (in millions of years)
616 a position in a current protein has been preserved by tracing back to its reconstructed direct
617 ancestors. The PSEP predicted for the amino-acid substitution of BGL2 was 674 million years.
618 According to the tool output, time longer than 450 million years represents "probably
619 damaging", corresponding to a false positive rate of 0.2.

620 As well as the mutation in *T. reesei*, in the S1M29 mutant the single amino-acid substitution
621 is not in the catalytic domain of the enzyme BGL2. Surprisingly, maybe a single amino-acid
622 substitution may be the major factor influencing the increase of cellulase expression in the
623 S1M29 mutant. Our analyzes led us to suggest that the SNP in *bgl2* may have reduced the
624 activity of this enzyme, affording the hyperproduction of cellulases by the mutant S1M29. The
625 effect of this amino-acid substitution on cellulase production could be also analyzed by a
626 combination of gene complementation and disruption techniques. Despite these experiments
627 are outside of the scope of this work.

628 Another key mutation is S259P identified in FlbA, which is a fungal Zn(II)2-Cys6
629 transcription regulator of conidiospore. The conidiation is essential for industrial applications
630 since the conidia are used as starters in the first step of fermentation. In this way, the regulatory
631 pathway that controls conidiophore development and spore maturation is well-described.
632 Various upstream developmental activators known as FLBs have been identified in filamentous
633 fungi, reported to be involved in the regulation of conidiospore development [101].

634 In *A. niger*, proteomics analysis indicates an increase in the number of total proteins and
635 CAZymes during growth of the Δ*flbA* strain, suggesting that *flbA* gene is an interesting target
636 for improvement of industrial strains. Deletion of *flbA* in *A. nidulans* results in strongly reduced
637 sporulation and excessive growth of aerial or submerged hyphae followed by autolytic collapse
638 of the mycelium [101].

639 Experimental results from *P. oxalicum* [98,99] and *T. reesei* [97] suggest that BGL2 likely
640 is the major mutation involved in cellulase hyperproduction of the S1M29 mutant. However,
641 experimental evidence of *A. niger* [101] suggests that the mutation in FlbA may also have an
642 impact on the production of cellulases. Our results suggest that these mutations probably
643 influence directly the regulatory pathways, resulting in overexpression of the main cellulases.
644 However, other identified mutations may also have had a less positive influence on cellulase
645 expression. Finally, we suggest that the group of mutations was responsible for the increased
646 expression of cellulases in the mutant and not just one isolated mutation.

647 ### *3.3.1.2 Origin tracking of albinism*

The lack of pigmentation is one of the striking phenotypic characteristics of the S1M29 mutant, motivating comparative analysis between the *P. ucsensis* proteome and the proteins associated with the three melanin biosynthesis pathways, characterized in the following filamentous fungi: *A. fumigatus* Af293 [86], *A. niger* CBS 513.88 [86], *T. marneffei* ATCC 18224 [87], *P. chrysogenum* P2niaD18 [87].

Our comparative analyses revealed that *P. ucsensis* possesses all the orthologous for the production of eumelanin by the DHN pathway, suggesting that the DHN pathway is the major route for melanin production in *P. ucsensis*. In addition, we observed that L-tyrosine degradation is the most conserved pathway across the analyzed fungi. In *A. niger* CBS 513.88, DOPA-melanin pathway is responsible for melanin production and this route is composed by tyrosinase coding genes and a vast number of laccase coding genes [102]. In contrast to *A. niger*, we did not find homologues for these laccases in *P. ucsensis* (Supplementary Table S04).

DHN-melanin genes are frequently encoded in biosynthetic gene clusters [25]. As presented in Fig. 3, we did not verify this organization in *A. niger* CBS 513.88, *P. chrysogenum* P2niaD18, *P. oxalicum* 114-2 and *P. ucsensis*. The DHN-melanin encoding genes were probably acquired by the common ancestor of *Penicillium* spp. and *Aspergillus* spp., and subsequent divergence and gene re-arrangement resulted in different gene orders and orientations of the individual genes in the different fungi [87].

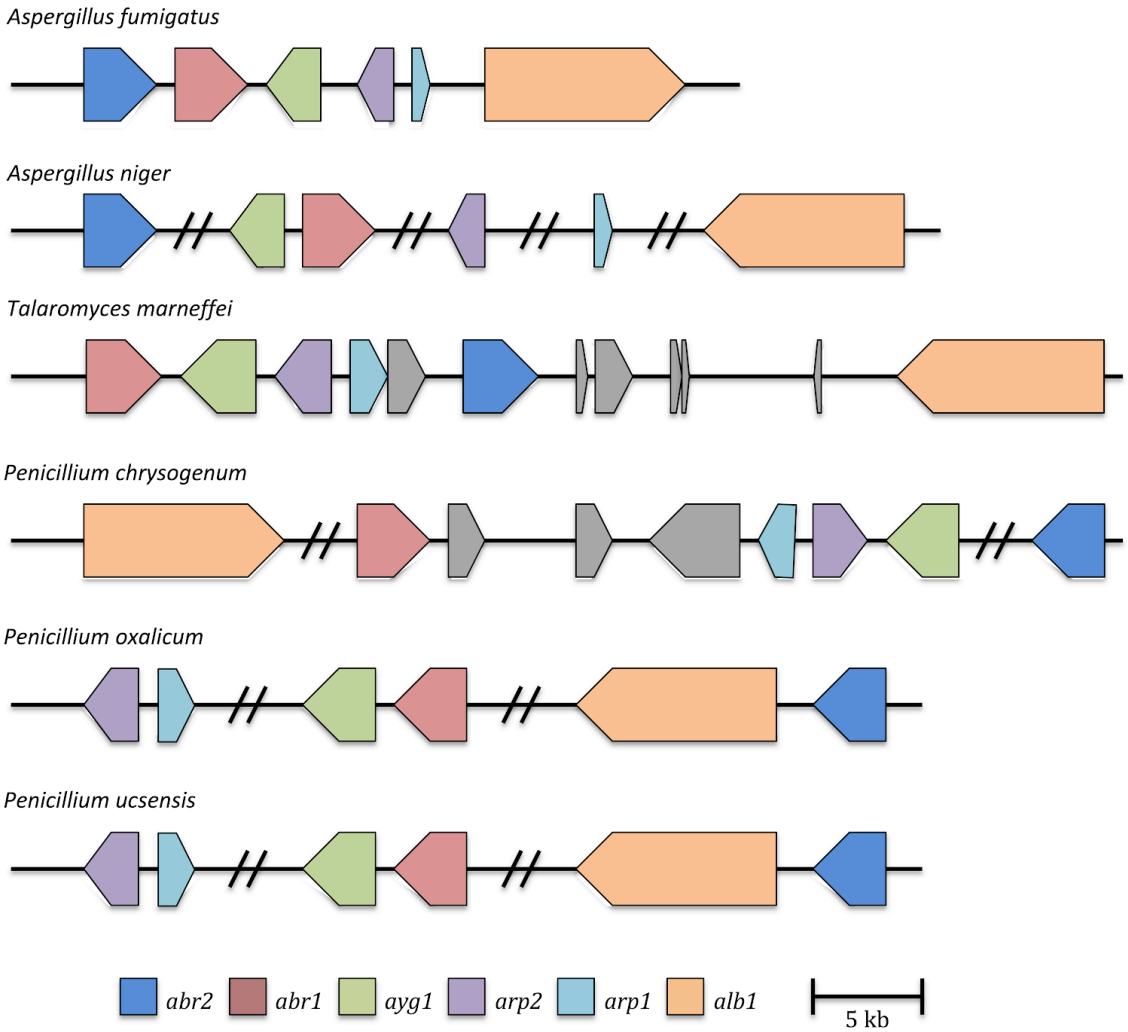


Figure 3. Map of the DHN melanin-biosynthesis gene organization in *P. ucsensis* and closely related fungi, based on [87]. *abr2*, conidial pigment biosynthesis oxidase; *abr1*, conidial pigment biosynthesis oxidase; *ayg1*, conidial pigment biosynthesis protein; *arp2*, conidial pigment biosynthesis 1,3,6,8-tetrahydroxynaphthalene reductase; *arp1*, conidial pigment biosynthesis scytalone dehydratase; *alb1*, conidial pigment polyketide synthase.

In *A. fumigatus*, a cluster of six genes includes the *alb1* gene encoding a conidial pigment polyketide synthase, an initial precursor of DHN-melanin, involved in the production of the heptaketide naphtopyrone YWA1. Also in *A. fumigatus*, *alb1* disruption resulted in the albino conidial phenotype [103]. Analyzing the DHN-melanin pathway of *P. ucsensis* strains, we found the same amino-acid sequences for all proteins, except a single amino-acid substitution at amino-acid 526 of ALB1, orthologue of *A. fumigatus* Af293. ALB1 exchanged a glycine in the wild-type strain 2HH (PECH_008565) for aspartic acid in the mutant S1M29 (PECM_000136).

In order to observe the effect of this mutation, fungal morphology was monitored by macromorphological observations of both strains after 7 and 14 days of growth on four different growth media. In Fig. 4, we observe visible changes in mycelium pigmentation in all growth media.

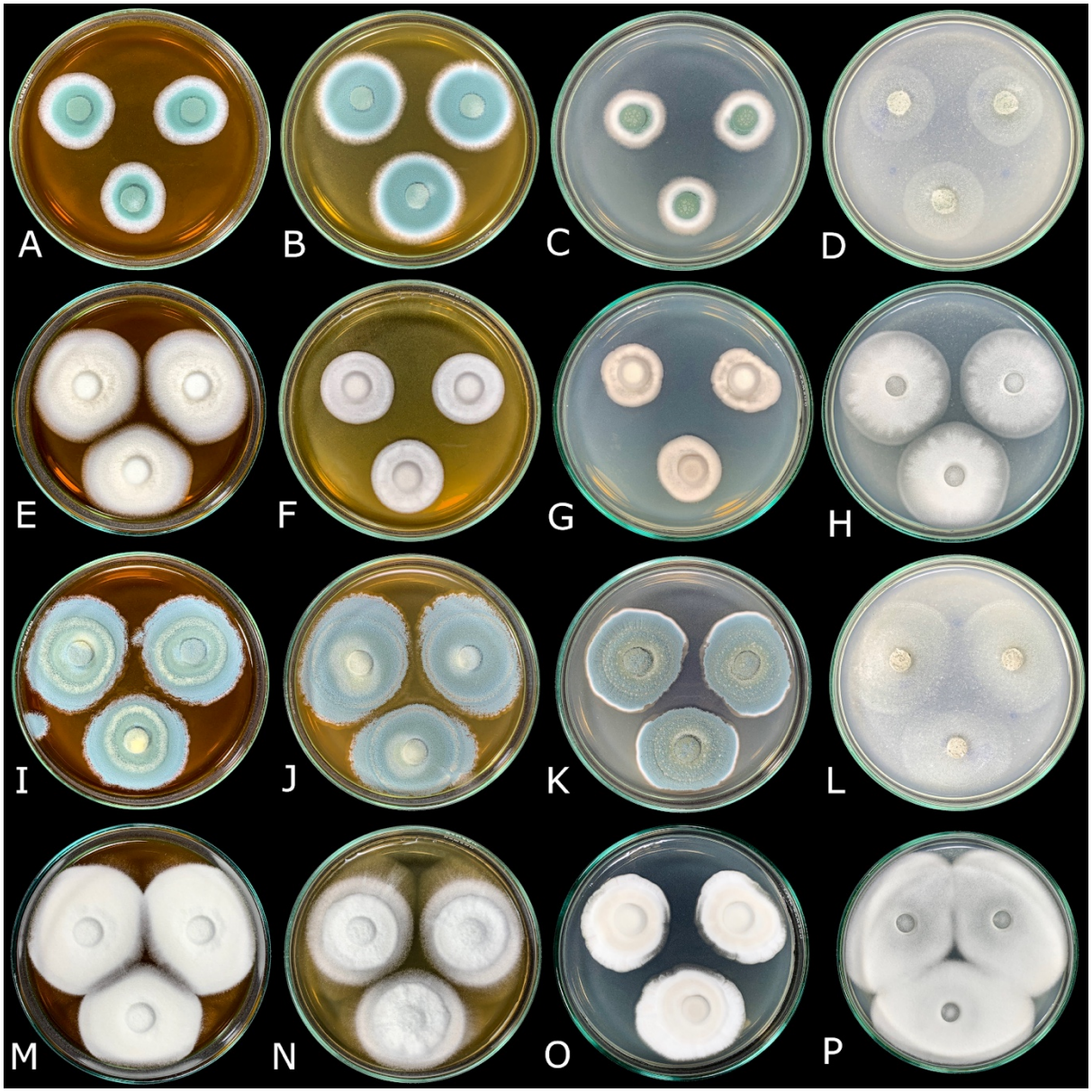


Figure 4. Macromorphological observation photographs of the impact of the *alb1* mutation in *Penicillium ucsensis* A) 2HH colonies on MEA after 7 days at 28°C. B) 2HH colonies on MEAbl after 7 days at 28°C. C) 2HH colonies on CYA after 7 days at 28°C. D) 2HH colonies on IHMM after 7 days at 28°C. E) S1M29 colonies on MEA after 7 days at 28°C. F) S1M29 colonies on MEAbl after 7 days at 28°C. G) S1M29 colonies on CYA after 7 days at 28°C. H) S1M29 colonies on IHMM after 7 days at 28°C. I) 2HH colonies on MEA after 14 days at 28°C. J) 2HH colonies on MEAbl after 14 days at 28°C. K) 2HH colonies on CYA after 14 days at 28°C. L) 2HH colonies on IHMM after 14 days at 28°C. M) S1M29 colonies on MEA after 14 days at 28°C. N) S1M29 colonies on MEAbl after 14 days at 28°C. O) S1M29 colonies on CYA after 14 days at 28°C. P) S1M29 colonies on IHMM after 14 days at 28°C.

While the wild-type grow melanized revealing the characteristic green pigmentation of the conidial color in *P. ucsensis*, the mutant conidia grow nonmelanized, exhibiting the albino phenotype, similar to *alb1* gene disruption observed in *A. fumigatus*. These results confirm that DHN is the pathway responsible for melanin production in *P. ucsensis*.

3.3.2 Evolutionary relationships

The taxonomic revision of the 2HH wild-type raises questions about the evolutionary relationships of this isolate with other filamentous fungi. The main evolutionary factor to be observed is the ecological niche of *P. ucsensis*, isolated as a symbiont of *A. punctatum* larvae.

3.3.2.1 General evolutionary relationships

For a general analysis of evolutionary relationships, we use a methodology based on highly conserved markers available for the class *Eurotiomycetes*. In addition to the *P. ucsensis* 2HH proteome, eleven other species of related filamentous fungi were used. All data used in this analysis can be found in the Supplementary Table S02. Phylogenetic relationships of the analyzed fungi are shown in Fig. 5. When *Penicillium brasilianum* MG11 is compared with *Penicillium subrubescens* CBS 132785, a smaller difference in length of branches between them is noticeable, when compared to the difference in length between branches of *P. oxalicum* 114-2 and *P. ucsensis* 2HH. Furthermore, AAI-profiler assessment shows that the average amino-acid sequence identity between *P. ucsensis* 2HH and *P. oxalicum* 114-2 is 85%. The histogram of amino-acid sequence identity is presented in Fig. 6. The literature suggests >95% identity at the species boundary [90]. The general evolutionary relationships presented corroborate the taxonomic revision presented above, providing even more evidence that *P. ucsensis* 2HH is a novel species placed at the Series *Oxalica*, differentiated from *P. oxalicum*.

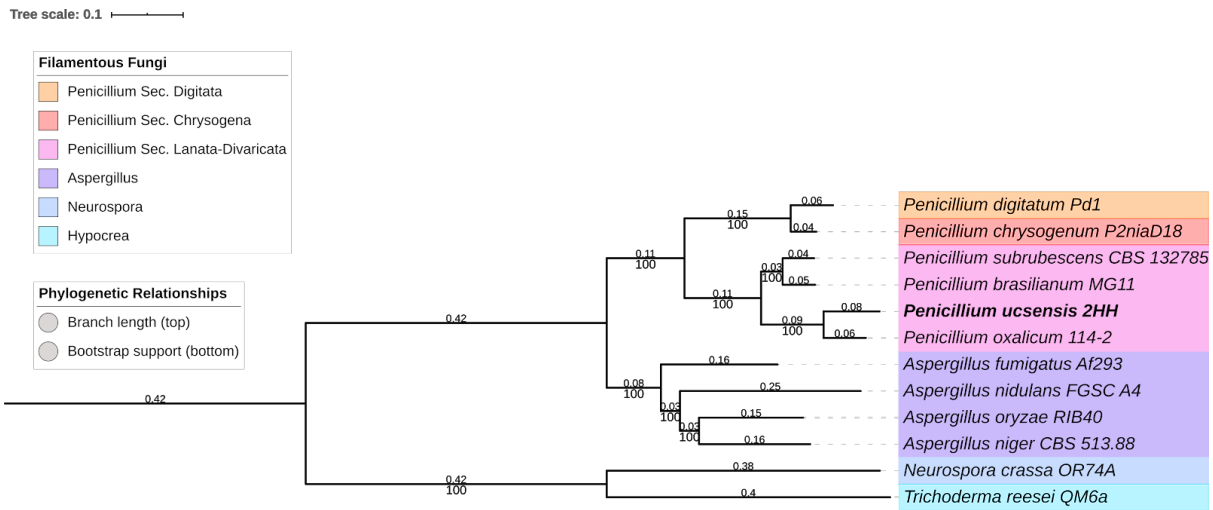


Figure 5. Evolutionary relationships of *P. ucsensis* 2HH; Bootstrap support (bs) values are labelled across the bottom of a branch, while branch length across the top.

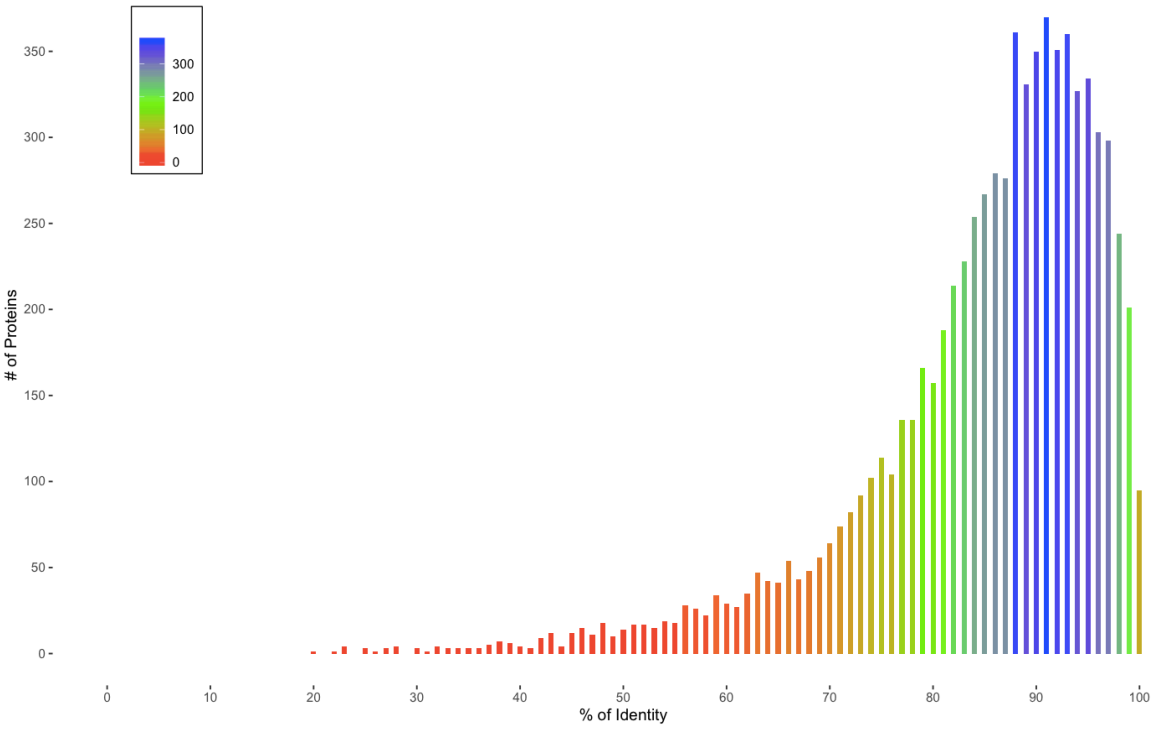


Figure 6. Histogram of amino-acid sequence identity between *P. ucsensis* 2HH and *P. oxalicum* 114-2.

3.3.2.2 Evolutionary relationships of cell wall-associated proteins

Comparative analyses between previously released cell wall-associated proteins from the well-characterized fungi *A. fumigatus* Af293 [24,86,93] and *N. crassa* OR74A [94] and our dataset revealed that *P. ucsensis* possess several orthologous of known fungal cell wall-associated proteins. The orthologous groups (Supplementary Table S05) also show that there is a high variety of cell wall-related proteins between the analyzed genera, pointing that the group of proteins related to assembly and changes in fungal cell wall are a valuable option to evaluate conserved and species-specific elements.

Overall, cell wall-associated proteins remain conserved in the analyzed fungi, especially those involved in biosynthesis of the main cell wall structural components: chitin, β-1,3-glucan, lichenin and α-1,3-glucan. We also observed that the major regulatory proteins required for cell wall assembly and modifications are conserved. On the other hand, some components are genus-specific or species-specific, particularly those involved in changing the cell wall during the fungus life cycle. Lastly, some cell wall components are seemingly genus-specific, such some glycoproteins and proteins involved in biogenesis of the cell wall as a three-dimensional matrix.

When we compared the proteins of *P. ucsensis* with *P. oxalicum* 114-2, the closest and available free-living relative, the major differences in cell wall formation are components involved in changing the cell wall throughout the fungus life cycle. We observed differences in the number of coding genes for chitinase (GH18), β-acetylhexosaminidase (GH84), endo-1,3(4)-β-glucanase (GH16), glucan endo-1,3-β-D-glucosidase (GH16), glucan endo-1,6-β-glucosidase (GH5) and 1,3-β-glucosidase (GH55). These enzymes are required during the biogenesis of the cell wall, spore germination, hyphal branching and septum formation in

filamentous fungi. Chitinases and β-hexosaminidases lead to re-modelling of the cell wall during growth and morphogenesis, while the other enzymes participate in the metabolism of β-D-glucan, the main structural component of the cell wall [22].

Evolutionary modifications in *P. ucsensis* include a significant reduction (5 genes) in the number of chitinases from the GH18 family. Differences in the number of proteins related to β-D-glucan metabolism are less expressive, but not less important. These changes in cell wall biosynthesis likely indicate modifications on growth and morphogenesis when compared to free-living relatives. This is consistent with symbiotic co-adaptations between *Leucoagaricus gongylophorus* and leaf-cutting ants that led to increased chitin and fungal cell wall thickness [104]. In addition, we also observed an endo-1,6-β-glucosidase (GH5) and a β-acetylhexosaminidase (GH84) which apparently are pseudogenes, predicted by gene finders, but not conserved when aligned with other fungi.

The cell wall represents a major organelle of filamentous fungi, dynamically responsive to environmental changes. Our results lead us to suggest that the differences in the composition of cell wall-associated proteins could be explained by specific-environment interactions resulting from a possible long-term mutualistic symbiosis between *P. ucsensis* and *A. punctatum*. Even more suggestively, the significant differences in the composition of proteins related to modifications on growth and morphogenesis could be explained by vertical transmission that makes the dispersion of the fungus highly dependent on the insect. In this sense, the reductive evolution refers to dispensability and loss of genetic material usually observed during the evolution of symbiotic species [105].

CONCLUSIONS

The results of our study shed considerable new light on the studies of *P. ucsensis*. First, the description of the novel species and its repositioning in the Section *Lanata-Divaricata* are essential to carry out comparative studies with other fungal species. The repositioning of *P. ucsensis* in the Series *Oxalica* clarifies the high levels of cellulase secretion, common in fungi of this series, like *P. oxalicum* widely used for commercial production of cellulolytic complexes in China.

Second, both draft genomes of *P. ucsensis* 2HH/S1M29 have been deposited at GenBank affording the molecular understanding of this microorganism and strain improvements using advanced techniques. Ongoing studies aim to identify the cellulolytic complex encoding genes and their main expression regulators, comprising transcription factors and sugar transporters. Besides, the global gene regulatory network [106] was inferred using the WGS data generated in this study.

Third, the genomic comparison of the mutant and the wild-type strains highlighted a wide set of mutations, of which only a few have been analyzed in more detail. BGL2 and FlbA likely are the major mutations involved in cellulase hyperproduction of the S1M29 mutant. Moreover, the single amino-acid substitution in ALB1, precursor enzyme of the DHN-melanin biosynthesis is the responsible for the complete loss of the conidiospores color, evidencing that DHN-melanin biosynthesis pathway is the major responsible for melanin production in *P. ucsensis*.

Four, the composition of cell wall-associated proteins of *P. ucsensis* shows considerable differences in the number of proteins, when compared to *P. oxalicum* 114-2, its closest free-

living relative available. The major differences comprise less chitinases, proteins related to β-D-glucan metabolism and two potential pseudogenes. We suggest that these differences could be potentially explained by specific-environment interactions resulting from a possible long-term mutualistic symbiosis between *P. ucsensis* and *A. punctatum*. With these ecological and evolutionary speculations, one factor is of particular biotechnological interest: a potential natural adaptation for cellulolytic enzymes production related to the larvae diet. Obviously, these speculations need experimental evidence to be considered.

Finally, our study provides an important step in building toward understanding the molecular machinery, the cellulolytic system, the melanin production and the evolutionary sphere of this notable fungus.

AUTHOR STATEMENTS

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgements

The authors thank the Coordination of Improvement of Higher Education Personnel (CAPES), National Council for Scientific and Technological Development (CNPq), the Bahia State University (UNEB) and the University of Caxias do Sul (UCS).

Funding information

We are grateful to the Coordination for the Improvement of Higher Education Personnel (CAPES) for the PhD scholarship (88887.158496/2017-00 to ARL). This research was supported by grants from CAPES (3255/2013) and the National Council for Scientific and Technological Development (CNPq) (472153/2013-7). MC and AJPD are CNPq Research Fellowship. We are grateful to Bahia State University (UNEB) for the leave of absence (3.145/2016 to ARL) and financial support.

REFERENCES

1. **Hyde KD, Xu J, Rapior S, Jeewon R, Lumyong S, Niego AGT, et al.** The amazing potential of fungi: 50 ways we can exploit fungi industrially. *Fungal Divers*. 2019;97(1):1–136. Available from: https://doi.org/10.1007/s13225-019-00430-9
2. **Parkin EA.** The Digestive Enzymes of Some Wood-Boring Beetle Larvae. *J Exp Biol.* 1940 Nov;17(4):364–77. Available from: http://jeb.biologists.org/content/17/4/364.abstract
3. **Carrau JL, Dillon AJP, Ribeiro RTS, Leygue-Alba NMR, Azevedo JL.** *Produção de enzimas celulolíticas por microrganismos.* In: Simpósio Internacional de Engenharia Genética. Piracicaba, SP; 1981. p. 39.
4. **Camassola M, De Bittencourt LR, Shenem NT, Andreaus J, Dillon AJP.** Characterization of the Cellulase Complex of *Penicillium echinulatum*. Biocatal Biotransformation. 2004;22(5–6):391–6. Available from: https://doi.org/10.1080/10242420400024532
5. **Dillon AJP, Zorgi C, Camassola M, Henriques JAP.** Use of 2-deoxyglucose in liquid media for the selection of mutant strains of *Penicillium echinulatum* producing

increased cellulase and β-glucosidase activities. *Appl Microbiol Biotechnol*. 2006;70(6):740–6. Available from: http://dx.doi.org/10.1007/s00253-005-0122-7

6. **Camassola M, Dillon AJP.** Production of cellulases and hemicellulases by *Penicillium echinulatum* grown on pretreated sugar cane bagasse and wheat bran in solid-state fermentation. *J Appl Microbiol*. 2007;103(6):2196–204. Available from: https://doi.org/10.1111/j.1365-2672.2007.03458.x
7. **Camassola M, Dillon AJP.** Effect of methylxanthines on production of cellulases by *Penicillium echinulatum*. *J Appl Microbiol*. 2007;102(2):478–85. Available from: https://doi.org/10.1111/j.1365-2672.2006.03098.x
8. **Camassola M, Dillon AJP.** Biological pretreatment of sugar cane bagasse for the production of cellulases and xylanases by *Penicillium echinulatum*. *Ind Crops Prod*. 2009;29(2):642–7. Available from: https://doi.org/10.1016/j.indcrop.2008.09.008
9. **Rubini MR, Dillon AJP, Kyaw CM, Faria FP, Poças-Fonseca MJ, Silva-Pereira I.** Cloning, characterization and heterologous expression of the first *Penicillium echinulatum* cellulase gene. *J Appl Microbiol*. 2010;108(4):1187–98. Available from: https://doi.org/10.1111/j.1365-2672.2009.04528.x
10. **Camassola M, Dillon AJP.** Cellulases and xylanases production by *Penicillium echinulatum* grown on sugar cane bagasse in solid-state fermentation. *Appl Biochem Biotechnol*. 2010;162(7):1889–900. Available from: http://dx.doi.org/10.1007/s12010-010-8967-3
11. **Dillon AJP, Bettio M, Pozzan FG, Andrighetti T, Camassola M.** A new *Penicillium echinulatum* strain with faster cellulase secretion obtained using hydrogen peroxide mutagenesis and screening with 2-deoxyglucose. *J Appl Microbiol*. 2011 Jul 1;111(1):48–53. Available from: https://doi.org/10.1111/j.1365-2672.2011.05026.x
12. **Camassola M, Dillon AJP.** Steam-Exploded Sugar Cane Bagasse for On-Site Production of Cellulases and Xylanases by *Penicillium echinulatum*. *Energy & Fuels*. 2012 Aug 16;26(8):5316–20. Available from: https://doi.org/10.1021/ef3009162
13. **Ribeiro DA, Cota J, Alvarez TM, Brüchli F, Bragato J, Pereira BMP, et al.** The *Penicillium echinulatum* Secretome on Sugar Cane Bagasse. *PLoS One*. 2012;7(12):e50571–e50571. Available from: https://doi.org/10.1371/journal.pone.0050571
14. **dos Reis L, Fontana RC, da Silva Delabona P, da Silva Lima DJ, Camassola M, da Cruz Pradella JG, et al.** Increased production of cellulases and xylanases by *Penicillium echinulatum* S1M29 in batch and fed-batch culture. *Bioresour Technol*. 2013;146:597–603. Available from: https://doi.org/10.1016/j.biortech.2013.07.124
15. **Novello M, Vilasboa J, Schneider WDH, Reis L Dos, Fontana RC, Camassola M.** Enzymes for second generation ethanol: Exploring new strategies for the use of xylose. *RSC Adv*. 2014;4(41):21361–8. Available from: https://doi.org/10.1039/c4ra00909f
16. **Schneider WDH, Dos Reis L, Camassola M, Dillon AJP.** Morphogenesis and production of enzymes by *Penicillium echinulatum* in response to different carbon sources. *Biomed Res Int*. 2014;2014:10. Available from: https://doi.org/10.1155/2014/254863
17. **Schneider WDH, Gonçalves TA, Uchima CA, Couger MB, Prade R, Squina FM, et al.** *Penicillium echinulatum* secretome analysis reveals the fungi potential for degradation of lignocellulosic biomass. *Biotechnol Biofuels*. 2016 Mar 17;9(1):66. Available from: https://doi.org/10.1186/s13068-016-0476-3
18. **Schneider WDH, Gonçalves TA, Uchima CA, Reis L dos, Fontana RC, Squina FM, et al.** Comparison of the production of enzymes to cell wall hydrolysis using different carbon sources by *Penicillium echinulatum* strains and its hydrolysis potential

for lignocelullosic biomass. *Process Biochem*. 2018;66:162–70. Available from: https://doi.org/10.1016/j.procbio.2017.11.004

19. **Schneider WDH, Fontana RC, Baudel HM, de Siqueira FG, Rencoret J, Gutiérrez A, et al.** Lignin degradation and detoxification of eucalyptus wastes by on-site manufacturing fungal enzymes to enhance second-generation ethanol yield. *Appl Energy*. 2020;262:114493. Available from: https://doi.org/10.1016/j.apenergy.2020.114493
20. **Visagie CM, Houbraken J, Frisvad JC, Hong SB, Klaassen CHW, Perrone G, et al.** Identification and nomenclature of the genus *Penicillium*. *Stud Mycol*. 2014;78(1):343–71. Available from: https://doi.org/10.1016/j.simyco.2014.09.001
21. **Vega F, Blackwell M.** *Insect-Fungal Associations: Ecology and Evolution*. Oxford University Press; 2005.
22. **Adams DJ.** Fungal cell wall chitinases and glucanases. *Microbiology*. 2004;150(7):2029–35. Available from: https://doi.org/10.1099/mic.0.26980-0
23. **Gow NAR, Latge J-P, Munro CA.** The Fungal Cell Wall: Structure, Biosynthesis, and Function. *Microbiol Spectr*. 2017;5(3). Available from: https://doi.org/10.1128/microbiolspec.FUNK-0035-2016
24. **Free SJ.** Fungal Cell Wall Organization and Biosynthesis. In: Friedmann T, Dunlap JC, Goodwin SF, editors. *Advances in Genetics*. Academic Press; 2013. p. 33–82. Available from: https://doi.org/10.1016/B978-0-12-407677-8.00002-6
25. **Eisenman HC, Casadevall A.** Synthesis and assembly of fungal melanin. *Appl Microbiol Biotechnol*. 2012 Feb;93(3):931–40. Available from: https://doi.org/10.1007/s00253-011-3777-2
26. **Langfelder K, Streibel M, Jahn B, Haase G, Brakhage AA.** Biosynthesis of fungal melanins and their importance for human pathogenic fungi. *Fungal Genet Biol*. 2003;38(2):143–58. Available from: https://doi.org/10.1016/S1087-1845(02)00526-1
27. **Pralea IE, Moldovan RC, Petrache AM, Ilieș M, Hegheș SC, Ielciu I, et al.** From extraction to advanced analytical methods: The challenges of melanin analysis. *Int J Mol Sci*. 2019 Aug 13;20(16):3943. Available from: https://doi.org/10.3390/ijms20163943
28. **Visagie CM, Renaud JB, Burgess KMN, Malloch DW, Clark D, Ketch L, et al.** Fifteen new species of *Penicillium*. *Persoonia Mol Phylogeny Evol Fungi*. 2016 Jun;36:247–80. Available from: https://doi.org/10.3767/003158516X691627
29. **Barbosa RN, Bezerra JDP, Souza-Motta CM, Frisvad JC, Samson RA, Oliveira NT, et al.** New *Penicillium* and *Talaromyces* species from honey, pollen and nests of stingless bees. *Antonie van Leeuwenhoek, Int J Gen Mol Microbiol*. 2018;111(10):1883–912. Available from: https://doi.org/10.1007/s10482-018-1081-1
30. **Diao YZ, Chen Q, Jiang XZ, Houbraken J, Barbosa RN, Cai L, et al.** *Penicillium* section *Lanata-divaricata* from acidic soil. *Cladistics*. 2019 Oct 1;35(5):514–49. Available from: https://doi.org/10.1111/cla.12365
31. **Kubátová A, Hujslová M, Frisvad JC, Chudíčková M, Kolařík M.** Taxonomic revision of the biotechnologically important species *Penicillium oxalicum* with the description of two new species from acidic and saline soils. *Mycol Prog*. 2019;18(1–2):215–28. Available from: https://doi.org/10.1007/s11557-018-1420-7
32. **Villesen P.** FaBox: An online toolbox for FASTA sequences. *Mol Ecol Notes.* 2007;7(6):965–8. Available from: https://doi.org/10.1111/j.1471-8286.2007.01821.x
33. **Katoh K, Rozewicki J, Yamada KD.** MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform*. 2018 Sep 6;20(4):1160–6. Available from: https://doi.org/10.1093/bib/bbx108

34. **Larsson A.** AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*. 2014 Aug 5;30(22):3276–8. Available from: https://doi.org/10.1093/bioinformatics/btu531
35. **Darriba D, Taboada GL, Doallo R, Posada D.** JModelTest 2: More models, new heuristics and parallel computing. *Nat Methods*. 2012 Jul 30;9(8):772. Available from: https://doi.org/10.1038/nmeth.2109
36. **Brewer MJ, Butler A, Cooksley SL.** The relative performance of AIC, AICC and BIC in the presence of unobserved heterogeneity. *Methods Ecol Evol.* 2016;7(6):679–92. Available from: https://doi.org/10.1111/2041-210X.12541
37. **Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, et al.** Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*. 2012 May;61(3):539–42. Available from: https://doi.org/10.1093/sysbio/sys029
38. **Stamatakis A.** RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014 Jan 21;30(9):1312–3. Available from: https://doi.org/10.1093/bioinformatics/btu033
39. **Miller MA, Pfeiffer W, Schwartz T.** Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *Gatew Comput Environ Work GCE*. 2010;1–8. Available from: https://doi.org/10.1109/GCE.2010.5676129
40. **Rambaut A.** FigTree: Computer program and documentation distributed by the author. [Internet]. 2018. Available from: http://tree.bio.ed.ac.uk
41. **Inkscape Project** Computer program and documentation distributed by the author [Internet]. 2019. Available from: https://inkscape.org
42. **Green MR, Sambrook JF.** *Molecular Cloning: A Laboratory Manual*. 4th ed. Press CSHL, editor. Vol. 1, Cold Springs Harbour Press. 2012.
43. **Illumina**. TruSeq DNA Sample Preparation Guide, Part 15026486 Rev. C [Internet]. Sample Prep Guide. 2012. Available from: https://support.illumina.com/
44. **Andrews S, Krueger F, Seconds-Pichon A, Biggins F, Wingett S.** FastQC. A quality control tool for high throughput sequence data. Babraham Bioinformatics [Internet]. Vol. 1, Babraham Institute. 2015. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
45. **Krueger F.** Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files [Internet]. Babraham Institute. 2015. Available from: https://github.com/FelixKrueger/TrimGalore
46. **Chikhi R, Medvedev P.** Informed and automated k-mer size selection for genome assembly. *Bioinformatics*. 2014 Jun 3;30(1):31–7. Available from: https://doi.org/10.1093/bioinformatics/btt310
47. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al.** SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012 May;19(5):455–77. Available from: https://doi.org/10.1089/cmb.2012.0021
48. **Gurevich A, Saveliev V, Vyahhi N, Tesler G.** QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*. 2013 Feb 19;29(8):1072–5. Available from: https://doi.org/10.1093/bioinformatics/btt086
49. **Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, et al.** BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol*. 2018 Dec 6;35(3):543–8. Available from: https://doi.org/10.1093/molbev/msx319
50. **Kriventseva E V., Tegenfeldt F, Petty TJ, Waterhouse RM, Simão FA, Pozdnyakov IA, et al.** OrthoDB v8: Update of the hierarchical catalog of orthologs

and the underlying free software. *Nucleic Acids Res*. 2015 Nov 26;43(D1):D250–6. Available from: https://doi.org/10.1093/nar/gku1220

51. **Illumina**. TruSeq RNA Sample Preparation v2 Guide, Part 15026495, Rev. F [Internet]. Manual. 2014. Available from: https://support.illumina.com/
52. **Kim D, Langmead B, Salzberg SL.** HISAT: A fast spliced aligner with low memory requirements. *Nat Methods*. 2015;12(4):357–60. Available from: https://doi.org/10.1038/nmeth.3317
53. **Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al.** Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011;29(7):644–52. Available from: https://doi.org/10.1038/nbt.1883
54. **Smit A, Hubley R, Green P.** RepeatMasker Open-4.0 [Internet]. 2013. Available from: http://www.repeatmasker.org
55. **Bao W, Kojima KK, Kohany O.** Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*. 2015;6(1):11. Available from: https://doi.org/10.1186/s13100-015-0041-9
56. **Stanke M, Schöffmann O, Morgenstern B, Waack S.** Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*. 2006;7(1):62. Available from: https://doi.org/10.1186/1471-2105-7-62
57. **Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M.** Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res*. 2008;18(12):1979–90. Available from: https://doi.org/10.1101/gr.081612.108
58. **Kent WJ.** BLAT---The BLAST-Like Alignment Tool. *Genome Res*. 2002;12(4):656–64. Available from: https://doi.org/10.1101/gr.229202
59. **Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ.** GMAP and GSNAP for genomic sequence alignment: Enhancements to speed, accuracy, and functionality. In: Mathé E, Davis S, editors. *Methods in Molecular Biology*. New York, NY: Springer New York; 2016. p. 283–334. Available from: https://doi.org/10.1007/978-1-4939-3578-9_15
60. **Slater GSC, Birney E.** Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. 2005;6(1):31. Available from: https://doi.org/10.1186/1471-2105-6-31
61. **The UniProt Consortium**. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*. 2019 Nov 5;47(D1):D506–15. Available from: https://doi.org/10.1093/nar/gky1049
62. **Liu G, Zhang L, Wei X, Zou G, Qin Y, Ma L, et al.** Genomic and Secretomic Analyses Reveal Unique Features of the Lignocellulolytic Enzyme System of *Penicillium decumbens*. *PLoS One*. 2013;8(2):e55185. Available from: https://doi.org/10.1371/journal.pone.0055185
63. **Lowe TM, Chan PP.** tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res*. 2016 May 12;44(W1):W54–7. Available from: https://doi.org/10.1093/nar/gkw413
64. **Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al.** Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol*. 2008;9(1):R7. Available from: https://doi.org/10.1186/gb-2008-9-1-r7
65. **Slater GSC, Birney E.** Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. 2005;6(1):31. Available from: https://doi.org/10.1186/1471-2105-6-31

66. **Birney E, Clamp M, Durbin R.** GeneWise and Genomewise. *Genome Res*. 2004 May;14(5):988–95. Available from: https://doi.org/10.1101/gr.1865504
67. **Solovyev V.** *Statistical Approaches in Eukaryotic Gene Prediction*. Handb Stat Genet Third Ed. 2008 Aug 24;1:97–159. Available from: https://doi.org/10.1002/9780470061619.ch4
68. **McDonnell E, Strasser K, Tsang A.** Manual gene curation and functional annotation. In: de Vries RP, Tsang A, Grigoriev I V, editors. *Methods in Molecular Biology*. New York, NY: Springer New York; 2018. p. 185–208. Available from: https://doi.org/10.1007/978-1-4939-7804-5_16
69. **Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, et al.** SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol*. 2019;37(4):420–3. Available from: https://doi.org/10.1038/s41587-019-0036-z
70. **Moller S, Croning MDR, Apweiler R.** Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*. 2002;18(1):218–218. Available from: https://doi.org/10.1093/bioinformatics/18.1.218
71. **Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al.** InterProScan 5: Genome-scale protein function classification. *Bioinformatics*. 2014 Jan 29;30(9):1236–40. Available from: https://doi.org/10.1093/bioinformatics/btu031
72. **Johnson LS, Eddy SR, Portugaly E.** Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*. 2010;11(1):431. Available from: https://doi.org/10.1186/1471-2105-11-431
73. **El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al.** The Pfam protein families database in 2019. *Nucleic Acids Res*. 2019 Oct 24;47(D1):D427–32. Available from: https://doi.org/10.1093/nar/gky995
74. **Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M.** KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res*. 2016 Oct 17;44(D1):D457–62. Available from: https://doi.org/10.1093/nar/gkv1070
75. **Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M.** KAAS: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res*. 2007 Jul;35(SUPPL.2):W182–5. Available from: https://doi.org/10.1093/nar/gkm321
76. **Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, Von Mering C, et al.** Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol Biol Evol*. 2017 Apr 29;34(8):2115–22. Available from: https://doi.org/10.1093/molbev/msx148
77. **Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al.** EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res*. 2019 Nov 12;47(D1):D309–14. Available from: https://doi.org/10.1093/nar/gky1085
78. **Blin K, Medema MH, Kottmann R, Lee SY, Weber T.** The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res*. 2017 Oct 24;45(D1):D555–9. Available from: https://doi.org/10.1093/nar/gkw960
79. **Epstein SC, Charkoudian LK, Medema MH.** A standardized workflow for submitting data to the Minimum Information about a Biosynthetic Gene cluster (MIBiG) repository: Prospects for research-based educational experiences. *Stand Genomic Sci*. 2018;13(1):16. Available from: https://doi.org/10.1186/s40793-018-0318-y

80. **Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al.** BLAST+: Architecture and applications. *BMC Bioinformatics*. 2009;10(1):421. Available from: https://doi.org/10.1186/1471-2105-10-421
81. **Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, et al**. DbCAN2: A meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res*. 2018 May 16;46(W1):W95–101. Available from: https://doi.org/10.1093/nar/gky418
82. **Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD.** The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res*. 2018 Nov 14;46(D1):D624–32. Available from: https://doi.org/10.1093/nar/gkx1134
83. **Peng M, Aguilar-Pontes M V., de Vries RP, Mäkelä MR.** In silico analysis of putative sugar transporter genes in *Aspergillus niger* using phylogeny and comparative transcriptomics. *Front Microbiol*. 2018;9(MAY):1045. Available from: https://doi.org/10.3389/fmicb.2018.01045
84. **Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al.** Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*. 2014 Sep 11;158(6):1431–43. Available from: https://doi.org/10.1016/j.cell.2014.08.009
85. **Darling AE, Mau B, Perna NT.** Progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS One*. 2010 Jun 25;5(6):e11147. Available from: https://doi.org/10.1371/journal.pone.0011147
86. **Teixeira MM, Moreno LF, Stielow BJ, Muszewska A, Hainaut M, Gonzaga L, et al.** Exploring the genomic diversity of black yeasts and relatives (*Chaetothyriales*, *Ascomycota*). *Stud Mycol*. 2017 Mar;86:1–28. Available from: https://doi.org/10.1016/j.simyco.2017.01.001
87. **Woo PCY, Tam EWT, Chong KTK, Cai JJ, Tung ETK, Ngan AHY, et al.** High diversity of polyketide synthase genes and the melanin biosynthesis gene cluster in *Penicillium marneffei*. *FEBS J*. 2010 Sep;277(18):3750–8. Available from: https://doi.org/10.1111/j.1742-4658.2010.07776.x
88. **Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska SJ.** Proteinortho: Detection of (Co-)orthologs in large-scale analysis. *BMC Bioinformatics*. 2011 Apr 28;12:124. Available from: https://doi.org/10.1186/1471-2105-12-124
89. The GIMP Development Team. GIMP. 2020. Available from: https://www.gimp.org
90. **Medlar AJ, Törönen P, Holm L.** AAI-profiler: Fast proteome-wide exploratory analysis reveals taxonomic identity, misclassification and contamination. *Nucleic Acids Res*. 2018 May 14;46(W1):W479–85. Available from: https://doi.org/10.1093/nar/gky359
91. **Posada D.** ModelTest Server: A web-based tool for the statistical selection of models of nucleotide substitution online. *Nucleic Acids Res*. 2006 Jul 1;34(WEB. SERV. ISS.):W700–3. Available from: http://dx.doi.org/10.1093/nar/gkl042
92. **Letunic I, Bork P.** Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res*. 2019 Jul 2;47(W1):W256–9. Available from: https://doi.org/10.1093/nar/gkz239
93. **Bernard M, Latgé JP.** *Aspergillus fumigatus* cell wall: Composition and biosynthesis. *Med Mycol Suppl*. 2001;39(1):9–17. Available from: https://doi.org/10.1080/mmy.39.1.9.17
94. **Patel PK, Free SJ.** The Genetics and Biochemistry of Cell Wall Structure and Synthesis in *Neurospora crassa*, a Model Filamentous Fungus. Vol. 10, *Frontiers in Microbiology*. 2019. p. 2294. Available from: https://doi.org/10.3389/fmicb.2019.02294

95. **Park MS, Oh SY, Lee S, Eimes JA, Lim YW.** Fungal diversity and enzyme activity associated with sailfin sandfish egg masses in Korea. *Fungal Ecol*. 2018;34:1–9. Available from: https://doi.org/10.1016/j.funeco.2018.03.004
96. **Aguilar-Pontes M V., Brandl J, McDonnell E, Strasser K, Nguyen TTM, Riley R, et al.** The gold-standard genome of *Aspergillus niger* NRRL 3 enables a detailed view of the diversity of sugar catabolism in fungi. *Stud Mycol*. 2018 Sep;91:61–78. Available from: https://doi.org/10.1016/j.simyco.2018.10.001
97. **Shida Y, Yamaguchi K, Nitta M, Nakamura A, Takahashi M, Kidokoro SI, et al.** The impact of a single-nucleotide mutation of bgl2 on cellulase induction in a *Trichoderma reesei* mutant. *Biotechnol Biofuels*. 2015;8(1):230. Available from: https://doi.org/10.1186/s13068-015-0420-y
98. **Li Z, Yao G, Wu R, Gao L, Kan Q, Liu M, et al.** Synergistic and Dose-Controlled Regulation of Cellulase Gene Expression in *Penicillium oxalicum*. *PLoS Genet*. 2015 Sep 11;11(9):e1005509–e1005509. Available from: https://doi.org/10.1371/journal.pgen.1005509
99. **Yao G, Wu R, Kan Q, Gao L, Liu M, Yang P, et al.** Production of a high-efficiency cellulase complex via β-glucosidase engineering in *Penicillium oxalicum*. *Biotechnol Biofuels*. 2016 Mar 31;9(1):78. Available from: https://doi.org/10.1186/s13068-016-0491-4
100. **Tang H, Thomas PD.** PANTHER-PSEP: Predicting disease-causing genetic variants using position-specific evolutionary preservation. *Bioinformatics*. 2016 May 18;32(14):2230–2. Available from: https://doi.org/10.1093/bioinformatics/btw222
101. **Van Munster JM, Nitsche BM, Akeroyd M, Dijkhuizen L, Van Der Maarel MJEC, Ram AFJ.** Systems approaches to predict the functions of glycoside hydrolases during the life cycle of *Aspergillus niger* using developmental mutants Δ*brlA* and Δ*flbA*. *PLoS One*. 2015 Jan 28;10(1):e0116269. Available from: https://doi.org/10.1371/journal.pone.0116269
102. **Pal AK, Gajjar DU, Vasavada AR.** DOPA and DHN pathway orchestrate melanin synthesis in *Aspergillus* species. *Med Mycol*. 2014 Sep 2;52(1):10–8. Available from: https://doi.org/10.3109/13693786.2013.826879
103. **Tsai HF, Wheeler MH, Chang YC, Kwon-Chung KJ.** A developmentally regulated gene cluster involved in conidial pigment biosynthesis in *Aspergillus fumigatus*. *J Bacteriol*. 1999 Oct;181(20):6469–77. Available from: https://doi.org/10.1128/jb.181.20.6469-6477.1999
104. **Nygaard S, Hu H, Li C, Schiøtt M, Chen Z, Yang Z, et al.** Reciprocal genomic evolution in the ant-fungus agricultural symbiosis. *Nat Commun*. 2016;7(1):12233. Available from: https://doi.org/10.1038/ncomms12233
105. **Albalat R, Cañestro C.** Evolution by gene loss. *Nat Rev Genet*. 2016;17(7):379–91. Available from: https://doi.org/10.1038/nrg.2016.39
106. **Lenz AR, Galán-Vasquez E, Balbinot E, Abreu FP, Souza de Oliveira N, Rosa LO, de Avila e Silva S, Camassola M, Dillon, AJP, Perez-Rueda E.** Gene regulatory networks of *Penicillium echinulatum* 2HH and *Penicillium oxalicum* 114-2 inferred by a computational biology approach. Vol. 11, *Frontiers in Microbiology*. 2020. Available from: https://doi.org/10.3389/fmicb.2020.588263

5 DISCUSSÃO GERAL

Os resultados desta tese representam um marco histórico para os estudos de *P. echinulatum* 2HH, cumprindo todos os objetivos traçados. A seção 2.3.2 da revisão bibliográfica desta tese foi publicada em forma de capítulo do livro Bioinformática: Contexto Computacional e Aplicações, ISBN 978-65-5807-001-6. Além disso, foram escritos três manuscritos de artigos científicos, sendo que o primeiro artigo foi publicado na revista *Frontiers in Microbiology* sob DOI: 10.3389/fmicb.2020.588263. Já o segundo manuscrito foi submetido à revista *Fungal Genetics & Biology* e está em fase de revisão. Já o último manuscrito depende de uma investigação colaborativa que não faz parte do escopo desta tese, compreendendo a caracterização morfológica e do perfil de extrólitos da nova espécie, para complementação da análise molecular realizada nesta tese. Ressalta-se que a classificação *Penicillium ucsensis* **sp. nov.** será válida somente após a caracterização da nova espécie e depósito no MycoBank. Essa investigação está em andamento, sendo realizada em parceria com o PhD. Jos Houbraken, Westerdijk Fungal Biodiversity Institute, Utrecht, Holanda.

Os genomas *draft* da linhagem selvagem 2HH e do mutante S1M29 de *P. echinulatum* foram depositados nas bases de dados públicas DDBJ, ENA e GenBank, possibilitando a ampliação do entendimento molecular desse microrganismo e permitindo a obtenção de linhagens a partir de técnicas avançadas de melhoramento genético.

Foi realizada a identificação molecular da nova espécie e seu reposicionamento na Série *Oxalica*, resultados essenciais para a realização de estudos comparativos com outros microrganismos. Nesse sentido, destaca-se o fato do isolado selvagem 2HH pertencer à seção *Lanata-Divaricata*, e estar localizado no clado de *P. diatomitis* e *P. soosanum*, ambos isolados na República Tcheca e *Penicillium* sp. SFC101850, isolado na República da Coreia. O reposicionamento de *P. echinulatum* 2HH na seção *Lanata-Divaricata* esclarece a sua alta produtividade de celulases, comum em fungos da Série *Oxalica*, como *P. oxalicum* amplamente utilizado para produção comercial de sistemas celulolíticos na China.

A comparação genômica das linhagens mutante e selvagem destacou um amplo conjunto de mutações, sugerindo que BGL2 provavelmente contém a principal mutação envolvida na hiperprodução de enzimas celulolíticas pelo mutante S1M29. No entanto, outras evidências experimentais também sugerem que a mutação no fator de transcrição FlbA também pode ter impactado positivamente a hiperprodução de celulases. Apesar de FlbA ser um TF pouco explorado em fungos modelo, na GRN de *P. echinulatum* foi observado que esse TF é regulado por COL-26, FF-7 e AmyR, demostrando sua conexão com importantes módulos regulatórios de enzimas celulolíticas. Assim, destacamos BGL2 e FlbA como importantes genes-alvo para projetar cepas hiperprodutoras de celulases a partir do mutante S1M29.

O estudo comparativo de ortologia de proteínas relacionadas à melanina fúngica, juntamente com observações macromorfológicas da linhagem selvagem 2HH e do mutante S1M29, revelaram que o fenótipo de albinismo do mutante S1M29 resultou de uma única substituição de aminoácido na enzima ALB1, precursora da biossíntese de DHN-melanina. As análises demonstraram a perda completa da cor dos conidiósporos no mutante S1M29, confirmando que a via DHN-melanina é responsável pela produção de melanina em *P. echinulatum*. Ainda foi verificado que a organização dos genes responsáveis pela biossíntese de DHN-melanina em formato de *cluster* não foi encontrada em *P. echinulatum* 2HH e *P. oxalicum* 114-2. A melanina fúngica apresenta características de interesse biotecnológico como, por exemplo, sua capacidade fotovoltaica. Filmes de melanina fúngica podem ser utilizados para produção de energia solar e até para produção de tecidos com proteção à radiação ultravioleta.

As análises comparativas e evolutivas de *P. echinulatum* 2HH e *P. oxalicum* 114-2 englobaram dois conjuntos de genes, o primeiro conjunto relacionado à composição de proteínas associadas à parede celular e o segundo conjunto relacionado às enzimas relacionadas à degradação de madeira. Nossos resultados mostram diferenças consideráveis no número de proteínas em *P. echinulatum* 2HH. Sugerimos que essas diferenças encontradas entre *P. echinulatum* 2HH e *P. oxalicum* 114-2 poderiam ser explicadas por interações ambiente-específicas resultantes de uma simbiose mutualística potencial a longo prazo entre *P. echinulatum* 2HH e *A. punctatum*. A hipótese da simbiose mutualística a longo prazo foi reforçada pelos nossos resultados, embora permaneça como uma hipótese.

A caracterização do CAZyoma demonstra o repertório de enzimas de *P. echinulatum* 2HH envolvidas na degradação da biomassa lignocelulolítica. Os genes que constituem o sistema enzimático celulolítico são predominantemente ortólogos aos genes do sistema enzimático celulolítico de *P. oxalicum* 114-2. Ambos os sistemas celulolíticos incluem uma enzima do tipo LPMO da família AA16, que atua sobre a celulose com clivagem oxidativa na posição C1 da unidade de glicose, descrita pela primeira vez nesses fungos. Apesar das enzimas celulolíticas compreenderem a única aplicação biotecnológica estudada para *P. echinulatum*, sabe-se que este fungo tem potencial para produção de inúmeras outras enzimas de interesse biotecnológico como, por exemplo, glicoamilases. Além da produção enzimática, fungos filamentosos podem ser vistos como potenciais biofábricas, capazes de produzir biomoléculas de interesse biotecnológico. Destaca-se o potencial para produção de metabólitos secundários que podem ser utilizados como fármacos, cosméticos, inseticidas, etc.

A caracterização do transportoma de açúcares de *P. echinulatum* 2HH demonstrou a diversidade e especificidade de STs, incluindo oito famílias principais de STs com especificidade para diferentes grupos de açúcares. A classificação filogenética dos STs ajuda a esclarecer seus papéis em relação à fonte de carbono preferida de *P. echinulatum* 2HH. O metabolismo flexível notável deste fungo sugere uma aptidão para degradar diversos materiais, permitindo estudos direcionados a partir da análise dos genes codificadores de enzimas e STs, visando a obtenção de

novas moléculas de interesse biotecnológico.

A caracterização funcional do TFoma nos genomas de *P. echinulatum* 2HH e *P. oxalicum* 114-2 mostra a existência de um grande conjunto de proteínas dedicadas a regular a expressão gênica em fungos filamentosos, destacando três famílias (Zn2Cys6, C2H2 ZF e RRM) que incluem cerca de 75 % do total de domínios de ligação ao DNA identificados. O conhecimento regulatório apresentado nesta tese é ínfimo, quando comparamos a quantidade de TFs analisados em relações à quantidade total de TFs (478) identificados em *P. echinulatum* 2HH, oportunizando uma série de estudos para compreensão do complexo sistema regulatório deste fungo.

As GRNs construídas para *P. echinulatum* 2HH e *P. oxalicum* 114-2 compreendem recursos valiosos de interações regulatórias, sendo as primeiras GRNs construídas para o gênero *Penicillium*. Foram encontradas propriedades topológicas semelhantes à outras redes biológicas de fungos modelo, destacando a existência de (pelo menos) dez reguladores globais. Apenas poucos TFs foram analisados em detalhe nesta tese, no entanto, as GRNs proporcionam um vasto conjunto de informações regulatórias que podem ser exploradas para obtenção de conhecimento relevante acerca de *P. echinulatum* 2HH e *P. oxalicum* 114-2.

Foram realizadas análises filogenéticas de iBGLs e STs, compreendendo genes-alvo valiosos para entender os mecanismos subjacentes ao mecanismo intracelular de acumulação de celodextrinas. Além disso, as GRNs inferidas revelaram diversos módulos regulatórios (CpcA, COL-26, FF-7, AmyR, ClrB, CreA, XlnR, entre outros) envolvidos na regulação da expressão de enzimas de interesse biotecnológico. A regulação da expressão do sistema celulolítico é bastante complexa e dependente de um vasto conjunto de genes, uma vez que depende da interconexão de inúmeros fatores ambientais como resposta ao stress, iluminação, etc. A compreensão desse complexo sistema de regulação vai muito além de TFs, STs e iBGLs, demandando estudos subsequentes para identificar novos componentes regulatórios.

Sugere-se o uso de CRISPR-CAS9 para edição gênica de *P. echinulatum*. Os genes-alvo identificados, a partir do conhecimento regulatório do sistema celulolítico apresentado nos artigos, fornecem uma base para estudos subsequentes, visando o projeto de cepas de *P. echinulatum* hipersecretoras de enzimas. Pode ser feita a disrupção ou super expressão de i) TFs identificados a partir da inferência da GRN; ii) iBGLs das famílias GH1 e GH3 identificadas a partir da filogenia; iii) STs com especificidade para celodextrinas identificados a partir da filogenia; e iv) BGL2 e FlbA identificados a partir das mutações encontradas no mutante S1M29. Sendo estes últimos dois genes, os mais relevantes para os estudos subsequentes.

A revisão da literatura e a proximidade filogenética de *P. echinulatum* 2HH e *P. oxalicum* 114-2 sugerem o emprego de cepas de *P. echinulatum* para produção comercial de enzimas celulolíticas e consequentemente seu uso em plantas de etanol 2G. Estudos de viabilidade podem ser conduzidos a partir da comparação da produção enzimática de cepas comerciais de *P. oxalicum* e *T. reesei* com as novas cepas de *P. echinulatum* que serão obtidas a partir dos genes-alvo identificados nesta tese.

6 CONCLUSÕES

Nossos resultados contribuem significativamente para construção de um alicerce de compreensão molecular de *P. echinulatum* 2HH, revelando características surpreendentes relacionadas à produção de enzimas lignocelulolíticas, captação de açúcares, produção de melanina, composição da parede celular e regulação da expressão gênica desse fungo. O conhecimento de vias regulatórias, aliado à caracterização de CAZymes e STs fornecem instrumentos valiosos para concepção de cepas comerciais de *P. echinulatum* 2HH para a produção de etanol 2G. Finalmente, destacamos os fungos da série *Oxalica* devido ao seu potencial diferenciado para produção de enzimas celulolíticas e, assim, essas espécies de *Penicillium* podem ser vistas como importantes aliados biotecnológicos para a produção de biocombustíveis lignocelulósicos, desempenhando um papel de destaque na transição energética global.

7 PERSPECTIVAS FUTURAS

As perspectivas futuras englobam uma série de estudos relevantes, dos quais destacam-se: i) sequenciamento completo do genoma da linhagem 2HH, incluindo leituras de tamanho longo para permitir a montagem completa, contígua e sem bases indefinidas; ii) RNA-Seq para comparação de expressão gênica entre as linhagens 2HH e S1M29 em diferentes fontes de carbono; iii) revisão e curadoria da anotação dos genes que não foram inspecionados nesta tese; iv) caracterização funcional de outros grupos de genes de interesse; v) identificação de potenciais moléculas de interesse biotecnológico além das enzimas lignocelulolíticas; vi) construção e avaliação de mutantes a partir dos genes-alvo identificados nesta tese; vii) incorporação de novas fontes regulatórias à GRN para elevar a confiabilidade das interações regulatórias; ix) construção de um sistema de inferência de GRNs baseado em ortologia para fungos; x) construção de um sistema de inferência de GRNs baseado em ortologia para bactérias; xi) construir um banco de dados para espécies de *Penicillium*, tal como SGD (*Saccharomyces* Genome Database) e AspGD (*Aspergillus* Genome Database).

REFERÊNCIAS

ABDULLAH, B.; MUHAMMAD, S. A. F. S.; SHOKRAVI, Z.; ISMAIL, S.; KASSIM, K. A.; MAHMOOD, A. N.; AZIZ, M. M. A. Fourth generation biofuel: A review on risks and mitigation strategies. **Renewable and Sustainable Energy Reviews**, v. 107, p. 37–50, 2019. ISSN 1364-0321. Disponível em: <https://doi.org/10.1016/j.rser.2019.02.018>. Citado na página 63.

ABEEL, T.; PEER, Y. Van de; SAEYS, Y. Toward a gold standard for promoter prediction evaluation. **Bioinformatics**, v. 25, n. 12, p. i313–i320, 2009. ISSN 13674803. Disponível em: <https://doi.org/10.1093/bioinformatics/btp191>. Citado na página 51.

ABEEL, T.; SAEYS, Y.; BONNET, E.; ROUZÉ, P.; PEER, Y. V. D. Generic eukaryotic core promoter prediction using structural features of DNA. **Genome Research**, Cold Spring Harbor Laboratory Press, v. 18, n. 2, p. 310–323, 11 2008. ISSN 10889051. Disponível em: <https://doi.org/10.1101/gr.6991408>. Citado na página 51.

ADAMS, D. J. Fungal cell wall chitinases and glucanases. **Microbiology**, Microbiology Society, v. 150, n. 7, p. 2029–2035, 7 2004. ISSN 13500872. Disponível em: <https://doi.org/10.1099/mic.0.26980-0>. Citado na página 47.

AGUILAR-PONTES, M. V.; BRANDL, J.; MCDONNELL, E.; STRASSER, K.; NGUYEN, T. T.; RILEY, R.; MONDO, S.; SALAMOV, A.; NYBO, J. L.; VESTH, T. C.; GRIGORIEV, I. V.; ANDERSEN, M. R.; TSANG, A.; VRIES, R. P. de. The gold-standard genome of *Aspergillus niger* NRRL 3 enables a detailed view of the diversity of sugar catabolism in fungi. **Studies in Mycology**, Elsevier, v. 91, p. 61–78, 9 2018. ISSN 01660616. Disponível em: <https://doi.org/10.1016/j.simyco.2018.10.001>. Citado na página 37.

AJANOVIC, A.; HAAS, R. Electric vehicles: solution or new problem? **Environment, Development and Sustainability**, v. 20, n. 1, p. 7–22, 2018. ISSN 1573-2975. Disponível em: <https://doi.org/10.1007/s10668-018-0190-3>. Citado na página 65.

AKINOSHO, H.; YEE, K.; CLOSE, D.; RAGAUSKAS, A. The emergence of *Clostridium thermocellum* as a high utility candidate for consolidated bioprocessing applications. **Frontiers in Chemistry**, Frontiers Media S.A., v. 2, n. AUG, p. 66–, 7 2014. ISSN 22962646. Disponível em: <https://doi.org/10.3389/fchem.2014.00066>. Citado na página 70.

ALLEN, J. E.; PERTEA, M.; SALZBERG, S. L. Computational gene prediction using multiple sources of evidence. **Genome Research**, Cold Spring Harbor Laboratory Press, v. 14, n. 1, p. 142–148, 1 2004. ISSN 10889051. Disponível em: <https://doi.org/10.1101/gr.1562804>. Citado na página 41.

AMORE, A.; GIACOBBE, S.; FARACO, V. Regulation of Cellulase and Hemicellulase Gene Expression in Fungi. **Current Genomics**, Bentham Science Publishers, v. 14, n. 4, p. 230–249, 4 2013. ISSN 13892029. Disponível em: <https://doi.org/10.2174/1389202911314040002>. Citado na página 29.

ANDREWS, S.; KRUEGER, F.; SECONDS-PICHON, A.; BIGGINS, F.; WINGETT, S. **FastQC. A quality control tool for high throughput sequence data. Babraham Bioinformatics**. 2015. 1 p. Disponível em: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Citado na página 36.

ANIL, M.; SAYGIN, D.; MIKETA, A.; GIELEN, D.; NICHOLAS, W. **The True Cost of Fossil Fuels : Saving on the Externalities of Air Pollution And Climate Change**. Abu Dhabi, United Arab Emirates: [s.n.], 2016. 1–12 p. ISBN 978-92-95111-87-5. Disponível em: <https://www.irena.org/publications/2016/May/The-True-Cost-of-Fossil-Fuels-Saving-on-the-Externalities-of-Air-Pollution-and-Climate-Change>. Citado na página 63.

ANP. **Anuário estatístico brasileiro do petróleo, gás natural e biocombustíveis 2019**. Brasília, 2019. 264 p. Disponível em: <http://www.anp.gov.br/arquivos/central-conteudos/anuario-estatistico/2019/>. Citado 2 vezes nas páginas 65 e 66.

ARO, N.; PAKULA, T.; PENTTILÄ, M. Transcriptional regulation of plant cell wall degradation by filamentous fungi. **FEMS Microbiology Reviews**, v. 29, n. 4, p. 719–739, 9 2005. ISSN 01686445. Disponível em: <https://doi.org/10.1016/j.femsre.2004.11.006>. Citado na página 61.

BALAT, M.; BALAT, H.; ÖZ, C. Progress in bioethanol processing. **Progress in Energy and Combustion Science**, v. 34, n. 5, p. 551–573, 10 2008. ISSN 03601285. Disponível em: <https://doi.org/10.1016/j.pecs.2007.11.001>. Citado na página 69.

BALDAUF, S. L. Phylogeny for the faint of heart: A tutorial. **Trends in Genetics**, v. 19, n. 6, p. 345–351, 6 2003. ISSN 01689525. Citado na página 46.

BAO, W.; KOJIMA, K. K.; KOHANY, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. **Mobile DNA**, v. 6, n. 1, p. 11, 2015. ISSN 17598753. Disponível em: <https://doi.org/10.1186/s13100-015-0041-9>. Citado na página 38.

BAYRAKCI, A. G.; KOÇAR, G. Second-generation bioethanol production from water hyacinth and duckweed in Izmir: A case study. **Renewable and Sustainable Energy Reviews**, v. 30, p. 306–316, 2 2014. ISSN 13640321. Disponível em: <https://doi.org/10.1016/j.rser.2013.10.011>. Citado na página 64.

BERLEMONT, R. Distribution and diversity of enzymes for polysaccharide degradation in fungi. **Scientific Reports**, v. 7, n. 1, p. 222, 2017. ISSN 20452322. Disponível em: <https://doi.org/10.1038/s41598-017-00258-w>. Citado na página 59.

BISCHOF, R. H.; RAMONI, J.; SEIBOTH, B. Cellulases and beyond: The first 70 years of the enzyme producer *Trichoderma reesei*. **Microbial Cell Factories**, BioMed Central, v. 15, n. 1, p. 106, 6 2016. ISSN 14752859. Disponível em: <https://doi.org/10.1186/s12934-016-0507-6>. Citado na página 54.

BPSTATS. BP Statistical Review of World Energy Statistical Review of World, 68th edition. **The Editor BP Statistical Review of World Energy**, p. 1–69, 2019. Disponível em: <https://www.bp.com/content/dam/bp/business-sites/en/global/corporate/pdfs/energy-economics/statistical-review/bp-stats-review-2019-full-report.pdf>. Citado na página 63.

BRETHAUER, S.; STUDER, M. H. Consolidated bioprocessing of lignocellulose by a microbial consortium. **Energy and Environmental Science**, The Royal Society of Chemistry, v. 7, n. 4, p. 1446–1453, 2014. ISSN 17545706. Disponível em: <http://dx.doi.org/10.1039/C3EE41753K>. Citado na página 70.

BRODEUR, G.; YAU, E.; BADAL, K.; COLLIER, J.; RAMACHANDRAN, K. B.; RAMAKRISHNAN, S. Chemical and physicochemical pretreatment of lignocellulosic biomass: A review. **Enzyme Research**, SAGE-Hindawi Access to Research, v. 2011, n. 1, p. 787532–, 3 2011. ISSN 20900406. Disponível em: <https://doi.org/10.4061/2011/787532>. Citado 3 vezes nas páginas 56, 57 e 70.

CAI, P.; WANG, B.; JI, J.; JIANG, Y.; WAN, L.; TIAN, C.; MA, Y. The putative cellodextrin transporter-like protein CLP1 is involved in cellulase induction in *Neurospora crassa*. **Journal of Biological Chemistry**, American Society for Biochemistry and Molecular Biology, v. 290, n. 2, p. 788–796, 1 2015. ISSN 1083351X. Disponível em: <https://doi.org/10.1074/jbc.M114.609875>. Citado 2 vezes nas páginas 61 e 62.

CAIRNS, T. C.; NAI, C.; MEYER, V. How a fungus shapes biotechnology: 100 years of *Aspergillus niger* research. **Fungal Biology and Biotechnology**, v. 5, n. 1, p. 13, 2018. ISSN 2054-3085. Disponível em: <https://doi.org/10.1186/s40694-018-0054-5>. Citado 2 vezes nas páginas 54 e 56.

CAMASSOLA, M.; BITTENCOURT, L. R. D.; SHENEM, N. T.; ANDREAUS, J.; DILLON, A. J. P. Characterization of the Cellulase Complex of *Penicillium echinulatum*. **Biocatalysis and Biotransformation**, Taylor & Francis, v. 22, n. 5-6, p. 391–396, 12 2004. ISSN 1024-2422. Disponível em: <https://doi.org/10.1080/10242420400024532>. Citado na página 29.

CAMASSOLA, M.; DILLON, A. J. P. Production of cellulases and hemicellulases by *Penicillium echinulatum* grown on pretreated sugar cane bagasse and wheat bran in solid-state fermentation. **Journal of Applied Microbiology**, John Wiley & Sons, Ltd, v. 103, n. 6, p. 2196–2204, 12 2007. ISSN 1364-5072. Disponível em: <https://doi.org/10.1111/j.1365-2672.2007.03458.x>. Citado na página 30.

CARDONA, E.; RIOS, J.; PEÑA, J.; RIOS, L. Effects of the pretreatment method on enzymatic hydrolysis and ethanol fermentability of the cellulosic fraction from elephant grass. **Fuel**, v. 118, p. 41–47, 2 2014. ISSN 00162361. Disponível em: <https://doi.org/10.1016/j.fuel.2013.10.055>. Citado na página 64.

CARRAU, J. L.; DILLON, A. J. P.; RIBEIRO, R. T. S.; LEYGUE-ALBA, N. M. R.; AZEVEDO, J. L. Produção de enzimas celulolíticas por microrganismos. In: **Simpósio Internacional de Engenharia Genética**. Piracicaba, SP: [s.n.], 1981. p. 39. Citado na página 28.

CARTER, H.; HOFREE, M.; IDEKER, T. Genotype to phenotype via network analysis. **Current Opinion in Genetics and Development**, v. 23, n. 6, p. 611–621, 12 2013. ISSN 0959437X. Disponível em: <https://doi.org/10.1016/j.gde.2013.10.003>. Citado na página 45.

CHEN, M.; QIN, Y.; CAO, Q.; LIU, G.; LI, J.; LI, Z.; ZHAO, J.; QU, Y. Promotion of extracellular lignocellulolytic enzymes production by restraining the intracellular β-glucosidase in *Penicillium decumbens*. **Bioresource Technology**, Elsevier, v. 137, p. 33–40, 6 2013. ISSN 18732976. Disponível em: <https://doi.org/10.1016/j.biortech.2013.03.099>. Citado na página 62.

CONESA, A.; MADRIGAL, P.; TARAZONA, S.; GOMEZ-CABRERO, D.; CERVERA, A.; MCPHERSON, A.; SZCZEŚNIAK, M. W.; GAFFNEY, D. J.; ELO, L. L.; ZHANG, X.; MORTAZAVI, A. Erratum: A survey of best practices for RNA-seq data analysis [Genome Biol. (2016), 17, 13] doi: 10.1186/s13059-016-0881-8. **Genome Biology**, v. 17, n. 1, p. 1–19, 2016. ISSN 1474760X. Disponível em: <http://dx.doi.org/10.1186/s13059-016-0881-8>. Citado na página 40.

CONSORTIUM, T. U. UniProt: a worldwide hub of protein knowledge. **Nucleic Acids Research**, v. 47, n. D1, p. D506–D515, 11 2019. ISSN 0305-1048. Disponível em: <https://doi.org/10.1093/nar/gky1049>. Citado 2 vezes nas páginas 44 e 46.

DALENA, F.; SENATORE, A.; IULIANELLI, A.; PAOLA, L. D.; BASILE, M.; BASILE, A. Ethanol From Biomass. In: BASILE, A.; IULIANELLI, A.; DALENA, F.; VEZIROĞLU, T. N. B. T. E. (Ed.). **Ethanol**. Elsevier, 2019. p. 25–59. ISBN 978-0-12-811458-2. Disponível em: <https://doi.org/10.1016/b978-0-12-811458-2.00002-x>. Citado 3 vezes nas páginas 27, 56 e 62.

DILLON, A.; PAESI-TORESAN, S.; BARP, L. Isolation of cellulase-producing mutants from *Penicillium* sp. strains denominated 3MUV3424. **Revista Brasileira de Genética**, v. 15, p. 491–498, 1992. Citado na página 29.

DILLON, A. J.; BETTIO, M.; POZZAN, F. G.; ANDRIGHETTI, T.; CAMASSOLA, M. A new *Penicillium echinulatum* strain with faster cellulase secretion obtained using hydrogen peroxide mutagenesis and screening with 2-deoxyglucose. **Journal of Applied Microbiology**, John Wiley & Sons, Ltd (10.1111), v. 111, n. 1, p. 48–53, 7 2011. ISSN 13645072. Disponível em: <https://doi.org/10.1111/j.1365-2672.2011.05026.x>. Citado na página 30.

DILLON, A. J.; ZORGI, C.; CAMASSOLA, M.; HENRIQUES, J. A. P. Use of 2-deoxyglucose in liquid media for the selection of mutant strains of *Penicillium echinulatum* producing increased cellulase and β-glucosidase activities. **Applied Microbiology and Biotechnology**, v. 70, n. 6, p. 740–746, 2006. ISSN 01757598. Disponível em: <http://dx.doi.org/10.1007/s00253-005-0122-7>. Citado na página 30.

DRUZHININA, I. S.; KUBICEK, C. P. Genetic engineering of *Trichoderma reesei* cellulases and their production. **Microbial Biotechnology**, John Wiley and Sons Inc., v. 10, n. 6, p. 1485–1499, 11 2017. ISSN 17517915. Disponível em: <https://doi.org/10.1111/1751-7915.12726>. Citado 2 vezes nas páginas 56 e 59.

DUTTA, K.; DAVEREY, A.; LIN, J.-G. Evolution retrospective for alternative fuels: First to fourth generation. **Renewable Energy**, v. 69, p. 114–122, 2014. ISSN 0960-1481. Disponível em: <https://doi.org/10.1016/j.renene.2014.02.044>. Citado na página 64.

EISENMAN, H. C.; CASADEVALL, A. Synthesis and assembly of fungal melanin. **Applied Microbiology and Biotechnology**, v. 93, n. 3, p. 931–940, 2 2012. ISSN 01757598. Disponível em: <https://doi.org/10.1007/s00253-011-3777-2>. Citado na página 47.

EKBLOM, R.; WOLF, J. B. A field guide to whole-genome sequencing, assembly and annotation. **Evolutionary Applications**, BlackWell Publishing Ltd, Oxford, UK, v. 7, n. 9, p. 1026–1042, 5 2014. ISSN 17524571. Disponível em: <https://doi.org/10.1111/eva.12178>. Citado 8 vezes nas páginas 34, 35, 36, 37, 38, 41, 42 e 72.

EL-GEBALI, S.; MISTRY, J.; BATEMAN, A.; EDDY, S. R.; LUCIANI, A.; POTTER, S. C.; QURESHI, M.; RICHARDSON, L. J.; SALAZAR, G. A.; SMART, A.; SONNHAMMER, E. L.; HIRSH, L.; PALADIN, L.; PIOVESAN, D.; TOSATTO, S. C.; FINN, R. D. The Pfam protein families database in 2019. **Nucleic Acids Research**, v. 47, n. D1, p. D427–D432, 10 2019. ISSN 13624962. Disponível em: <https://doi.org/10.1093/nar/gky995>. Citado 2 vezes nas páginas 44 e 53.

EPE. Análise de Conjuntura dos Biocombustíveis: Ano 2018. **EPE - Empresa de Pesquisa Energética**, p. 64, 2019. Disponível em: <http://www.epe.gov.br/Petroleo/Documents/AnálisedeConjunturadosBiocombustíveis-boletinsperiódicos/AnálisedeConjuntura-Ano2013.pdf>. Citado 3 vezes nas páginas 66, 67 e 69.

____. Balanço energético nacional: Ano base 2018. **EPE - Empresa de Pesquisa Energética**, p. 67, 2019. Citado 3 vezes nas páginas 65, 66 e 68.

FERNANDEZ-VALVERDE, S. L.; AGUILERA, F.; RAMOS-DÍAZ, R. A. Inference of Developmental Gene Regulatory Networks Beyond Classical Model Systems: New Approaches in the Post-genomic Era. **Integrative and Comparative Biology**, v. 58, n. 4, p. 640–653, 6 2018. ISSN 1540-7063. Disponível em: <https://doi.org/10.1093/icb/icy061>. Citado na página 49.

FILIATRAULT-CHASTEL, C.; NAVARRO, D.; HAON, M.; GRISEL, S.; HERPOËL-GIMBERT, I.; CHEVRET, D.; FANUEL, M.; HENRISSAT, B.; HEISS-BLANQUET, S.; MARGEOT, A.; BERRIN, J. G. AA16, a new lytic polysaccharide monooxygenase family identified in fungal secretomes. **Biotechnology for Biofuels**, v. 12, n. 1, p. 55, 2019. ISSN 17546834. Disponível em: <https://doi.org/10.1186/s13068-019-1394-y>. Citado na página 59.

FISCHER, S.; BRUNK, B. P.; CHEN, F.; GAO, X.; HARB, O. S.; IODICE, J. B.; SHANMUGAM, D.; ROOS, D. S.; STOECKERT, C. J. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. **Current Protocols in Bioinformatics**, Chapter 6, n. SUPPL.35, p. Unit–6.12.19, 9 2011. ISSN 1934340X. Disponível em: <https://doi.org/10.1002/0471250953.bi0612s35>. Citado na página 46.

FREE, S. J. Fungal Cell Wall Organization and Biosynthesis. In: FRIEDMANN, T.; DUNLAP, J. C.; GOODWIN, S. F. (Ed.). **Advances in Genetics**. Academic Press, 2013. v. 81, p. 33–82. ISBN 9780124076778. Disponível em: <https://doi.org/10.1016/B978-0-12-407677-8.00002-6>. Citado na página 47.

GABALDÓN, T.; KOONIN, E. V. Functional and evolutionary implications of gene orthology. **Nature Reviews Genetics**, v. 14, n. 5, p. 360–366, 5 2013. ISSN 14710056. Disponível em: <https://doi.org/10.1038/nrg3456>. Citado 5 vezes nas páginas 40, 44, 45, 46 e 49.

GASCH, A. P.; MOSES, A. M.; CHIANG, D. Y.; FRASER, H. B.; BERARDINI, M.; EISEN, M. B. Conservation and evolution of cis-regulatory systems in ascomycete fungi. **PLoS Biology**, Public Library of Science, v. 2, n. 12, p. e398–, 11 2004. ISSN 15449173. Disponível em: <https://doi.org/10.1371/journal.pbio.0020398>. Citado na página 53.

GE, Y.; LI, L. System-level energy consumption modeling and optimization for cellulosic biofuel production. **Applied Energy**, v. 226, p. 935–946, 2018. ISSN 0306-2619. Disponível em: <https://doi.org/10.1016/j.apenergy.2018.06.020>. Citado 2 vezes nas páginas 64 e 70.

GIELEN, D.; BOSHELL, F.; SAYGIN, D.; BAZILIAN, M. D.; WAGNER, N.; GORINI, R. The role of renewable energy in the global energy transformation. **Energy Strategy Reviews**, v. 24, p. 38–50, 2019. ISSN 2211-467X. Disponível em: <https://doi.org/10.1016/j.esr.2019.01.006>. Citado na página 64.

GLASS, N. L.; SCHMOLL, M.; CATE, J. H.; CORADETTI, S. Plant Cell Wall Deconstruction by Ascomycete Fungi. **Annual Review of Microbiology**, v. 67, n. 1, p. 477–498, 2013. ISSN 0066-4227. Disponível em: <https://doi.org/10.1146/annurev-micro-092611-150044>. Citado 4 vezes nas páginas 28, 57, 59 e 61.

GOW, N. A. R.; LATGE, J.-P.; MUNRO, C. A. The Fungal Cell Wall: Structure, Biosynthesis, and Function. **Microbiology Spectrum**, v. 5, n. 3, 2017. ISSN 2165-0497. Disponível em: <https://doi.org/10.1128/microbiolspec.FUNK-0035-2016>. Citado na página 47.

GUPTA, V. K.; STEINDORFF, A. S.; PAULA, R. G. de; SILVA-ROCHA, R.; MACH-AIGNER, A. R.; MACH, R. L.; SILVA, R. N. The Post-genomic Era of *Trichoderma reesei*: What's Next? **Trends in Biotechnology**, Elsevier Current Trends, v. 34, n. 12, p. 970–982, 12 2016. ISSN 18793096. Disponível em: <https://doi.org/10.1016/j.tibtech.2016.06.003>. Citado 2 vezes nas páginas 54 e 56.

GUREVICH, A.; SAVELIEV, V.; VYAHHI, N.; TESLER, G. QUAST: Quality assessment tool for genome assemblies. **Bioinformatics**, v. 29, n. 8, p. 1072–1075, 2 2013. ISSN 13674803. Disponível em: <https://doi.org/10.1093/bioinformatics/btt086>. Citado na página 36.

GUSAKOV, A. V. Alternatives to *Trichoderma reesei* in biofuel production. **Trends in Biotechnology**, Elsevier Current Trends, v. 29, n. 9, p. 419–425, 9 2011. ISSN 01677799. Disponível em: <https://doi.org/10.1016/j.tibtech.2011.04.004>. Citado na página 29.

GUSAKOV, A. V.; SINITSYN, A. P. Cellulases from *Penicillium* species for producing fuels from biomass. **Biofuels**, Taylor & Francis, v. 3, n. 4, p. 463–477, 7 2012. ISSN 17597269. Disponível em: <https://doi.org/10.4155/bfs.12.41>. Citado na página 29.

GUSELLA, J. F.; WEXLER, N. S.; CONNEALLY, P. M.; NAYLOR, S. L.; ANDERSON, M. A.; TANZI, R. E.; WATKINS, P. C.; OTTINA, K.; WALLACE, M. R.; SAKAGUCHI, A. Y.; YOUNG, A. B.; SHOULSON, I.; BONILLA, E.; MARTIN, J. B. A polymorphic DNA marker genetically linked to Huntington's disease. **Nature**, v. 306, n. 5940, p. 234–238, 1983. ISSN 00280836. Disponível em: <https://doi.org/10.1038/306234a0>. Citado na página 34.

HAAS, B. J.; ZENG, Q.; PEARSON, M. D.; CUOMO, C. A.; WORTMAN, J. R. Approaches to fungal genome annotation. **Mycology**, v. 2, n. 3, p. 118–141, 10 2011. ISSN 21501211. Disponível em: <https://doi.org/10.1080/21501203.2011.606851>. Citado 5 vezes nas páginas 40, 41, 42, 43 e 44.

HAAS, M.; HIMMELBACH, A.; MASCHER, M. The contribution of cis- and trans-acting variants to gene regulation in wild and domesticated barley under cold stress and control conditions. **Journal of Experimental Botany**, 1 2020. ISSN 0022-0957. Disponível em: <https://doi.org/10.1093/jxb/eraa036>. Citado na página 53.

HALL, B. K. **Evolution: Principles and Processes**. Jones & Bartlett Learning, 2011. Disponível em: <https://books.google.es/books?id=V24EHUgEl5EC>. Citado na página 52.

HARIDAS, S.; SALAMOV, A.; GRIGORIEV, I. V. Fungal genome annotation. In: VRIES, R. P. de; TSANG, A.; GRIGORIEV, I. V. (Ed.). **Methods in Molecular Biology**. New York, NY: Springer New York, 2018. v. 1775, p. 171–184. ISBN 978-1-4939-7804-5. Disponível em: <https://doi.org/10.1007/978-1-4939-7804-5_15>. Citado 2 vezes nas páginas 37 e 40.

HASSANI-PAK, K.; RAWLINGS, C. Knowledge Discovery in Biological Databases for Revealing Candidate Genes Linked to Complex Phenotypes. **Journal of integrative bioinformatics**, De Gruyter, v. 14, n. 1, p. 20160002, 6 2017. ISSN 16134516. Disponível em: <https://doi.org/10.1515/jib-2016-0002>. Citado na página 49.

HO, C. Y.; CHANG, J. J.; LEE, S. C.; CHIN, T. Y.; SHIH, M. C.; LI, W. H.; HUANG, C. C. Development of cellulosic ethanol production process via co-culturing of artificial cellulosomal *Bacillus* and kefir yeast. **Applied Energy**, v. 100, p. 27–32, 12 2012. ISSN 03062619. Disponível em: <https://doi.org/10.1016/j.apenergy.2012.03.016>. Citado na página 69.

HOUBRAKEN, J.; VRIES, R. P. de; SAMSON, R. A. Modern taxonomy of biotechnologically important *Aspergillus* and *Penicillium* species. **Advances in Applied Microbiology**, v. 86, p. 199–249, 2014. ISSN 00652164. Disponível em: <https://doi.org/10.1016/B978-0-12-800262-9.00004-4>. Citado na página 32.

HU, J.; CHEN, C.; HUANG, K.; MITCHELL, T. K. A distribution pattern assisted method of transcription factor binding site discovery for both yeast and filamentous fungi. **Advances in Bioscience and Biotechnology**, v. 04, n. 04, p. 509–517, 2013. ISSN 2156-8456. Disponível em: <https://doi.org/10.4236/abb.2013.44067>. Citado 2 vezes nas páginas 53 e 54.

HU, J.; TIAN, D.; RENNECKAR, S.; SADDLER, J. N. **Enzyme mediated nanofibrillation of cellulose by the synergistic actions of an endoglucanase, lytic polysaccharide monooxygenase (LPMO) and xylanase**. Department of Wood Science, Forest Products Biotechnology/Bioenergy Group, Faculty of Forestry, University of British Columbia, 2424 Main Mall, Vancouver, British Columbia, V6T 1Z4, Canada.: [s.n.], 2018. 3195 p. Disponível em: <https://doi.org/10.1038/s41598-018-21016-6>. Citado 2 vezes nas páginas 57 e 59.

HU, Y.; QIN, Y.; LIU, G. Collection and Curation of Transcriptional Regulatory Interactions in *Aspergillus nidulans* and *Neurospora crassa* Reveal Structural and Evolutionary Features of the Regulatory Networks. **Frontiers in Microbiology**, v. 9, p. 27, 2018. ISSN 1664-302X. Disponível em: <https://doi.org/10.3389/fmicb.2018.02713>. Citado na página 49.

HUBER, W.; CAREY, V. J.; LONG, L.; FALCON, S.; GENTLEMAN, R. Graphs in molecular biology. **BMC Bioinformatics**, BioMed Central, v. 8, n. SUPPL. 6, p. S8–S8, 9 2007. ISSN 14712105. Disponível em: <https://doi.org/10.1186/1471-2105-8-S6-S8>. Citado na página 45.

HYDE, K. D.; XU, J.; RAPIOR, S.; JEEWON, R.; LUMYONG, S.; NIEGO, A. G. T.; ABEYWICKRAMA, P. D.; ALUTHMUHANDIRAM, J. V.; BRAHAMANAGE, R. S.; BROOKS, S.; CHAIYASEN, A.; CHETHANA, K. W.; CHOMNUNTI, P.; CHEPKIRUI, C.; CHUANKID, B.; SILVA, N. I. de; DOILOM, M.; FAULDS, C.; GENTEKAKI, E.; GOPALAN, V.; KAKUMYAN, P.; HARISHCHANDRA, D.; HEMACHANDRAN, H.; HONGSANAN, S.; KARUNARATHNA, A.; KARUNARATHNA, S. C.; KHAN, S.; KUMLA, J.; JAYAWARDENA, R. S.; LIU, J. K.; LIU, N.; LUANGHARN, T.; MACABEO, A. P. G.; MARASINGHE, D. S.; MEEKS, D.; MORTIMER, P. E.; MUELLER, P.; NADIR, S.; NATARAJA, K. N.; NONTACHAIYAPOOM, S.; O'BRIEN, M.; PENKHRUE, W.; PHUKHAMSAKDA, C.; RAMANAN, U. S.; RATHNAYAKA, A. R.; SADABA, R. B.; SANDARGO, B.; SAMARAKOON, B. C.; TENNAKOON, D. S.; SIVA, R.; SRIPROM, W.; SURYANARAYANAN, T. S.; SUJARIT, K.; SUWANNARACH, N.; SUWUNWONG, T.; THONGBAI, B.; THONGKLANG, N.; WEI, D.; WIJESINGHE, S. N.; WINISKI, J.; YAN, J.; YASANTHIKA, E.; STADLER, M. The amazing potential of fungi: 50 ways we can exploit fungi industrially. **Fungal Diversity**, v. 97, n. 1, p. 1–136, 2019. ISSN 18789129. Disponível em: <https://doi.org/10.1007/s13225-019-00430-9>. Citado 5 vezes nas páginas 27, 54, 55, 56 e 61.

IEA. **Energy and Climate Change, World Energy Outlook Special Report**. Paris, France, 2015. 1–200 p. Disponível em: <https://www.iea.org/reports/energy-and-climate-change>. Citado 2 vezes nas páginas 67 e 68.

____. **Renewables 2019**. Paris, France: [s.n.], 2019. Disponível em: <https://www.iea.org/reports/renewables-2019>. Citado na página 67.

IRENA. **Road transport: the cost of renewable solutions**. Abu Dhabi, United Arab Emirates, 2013. 83 p. Disponível em: <https://www.irena.org/publications/2013/Jul/Road-Transport-The-Cost-of-Renewable-Solutions>. Citado na página 68.

JABLONSKI, A.; LEWERA, A. Improving the efficiency of a direct ethanol fuel cell by a periodic load change. **Chinese Journal of Catalysis**, v. 36, n. 4, p. 496–501, 2015. ISSN 1872-2067. Disponível em: <https://doi.org/10.1016/S1872-2067(14)60226-6>. Citado na página 65.

JACKSON, C. A.; CASTRO, D. M.; SALDI, G. A.; BONNEAU, R.; GRESHAM, D. Gene regulatory network reconstruction using single-cell rna sequencing of barcoded genotypes in diverse environments. **eLife**, eLife Sciences Publications, Ltd, v. 9, p. e51254, 1 2020. ISSN 2050084X. Disponível em: <https://doi.org/10.7554/eLife.51254>. Citado 2 vezes nas páginas 48 e 49.

JOUZANI, G. S.; TAHERZADEH, M. J. Advances in consolidated bioprocessing systems for bioethanol and butanol production from biomass: A comprehensive review. **Biofuel Research Journal**, v. 2, n. 1, p. 152–195, 3 2015. ISSN 22928782. Disponível em: <https://doi.org/10.18331/BRJ2015.2.1.4>. Citado 5 vezes nas páginas 59, 64, 68, 69 e 70.

KARLEBACH, G.; SHAMIR, R. Modelling and analysis of gene regulatory networks. **Nature Reviews Molecular Cell Biology**, v. 9, n. 10, p. 770–780, 2008. ISSN 1471-0080. Disponível em: <https://doi.org/10.1038/nrm2503>. Citado na página 48.

KEILWAGEN, J.; HARTUNG, F.; PAULINI, M.; TWARDZIOK, S. O.; GRAU, J. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. **BMC Bioinformatics**, BioMed Central, v. 19, n. 1, p. 189, 5 2018. ISSN 14712105. Disponível em: <https://doi.org/10.1186/s12859-018-2203-5>. Citado na página 41.

KOEHLER, N.; MCCAHERTY, J.; WILSON, C.; COOPER, G.; SCHWARCK, R.; KEMMET, N.; BAKER, R.; MCAFEE, E.; DROOK, R.; MARKHAM, S.; SITZMANN, C.; FRIEDBERG, J.; RICKETTS, M.; CHRISTENSEN, S.; BOYLE, P.; HARDER, S.; KEISER, K.; WOODSIDE, C.; ROE, S.; WILSON, C. **2020 Ethanol Industry Outlook**. Washington, DC, USA, 2020. 40 p. Disponível em: <https://ethanolrfa.org/wp-content/uploads/2020/02/2020-Outlook-Final-for-Website.pdf>. Citado na página 67.

KRACHER, D.; LUDWIG, R. Cellobiose dehydrogenase: An essential enzyme for lignocellulose degradation in nature – A review / Cellobiosedehydrogenase: Ein essentielles Enzym für den Lignozelluloseabbau in der Natur – Eine Übersicht. **Die Bodenkultur: Journal of Land Management, Food and Environment**, Sciendo, Berlin, v. 67, n. 3, p. 145–163, 2016. Disponível em: <https://doi.org/10.1515/boku-2016-0013>. Citado na página 59.

KRISTIANSSON, E.; THORSEN, M.; TAMÁS, M. J.; NERMAN, O. Evolutionary forces act on promoter length: Identification of enriched cis-regulatory elements. **Molecular Biology and Evolution**, v. 26, n. 6, p. 1299–1307, 3 2009. ISSN 07374038. Disponível em: <https://doi.org/10.1093/molbev/msp040>. Citado na página 51.

KRIVENTSEVA, E. V.; TEGENFELDT, F.; PETTY, T. J.; WATERHOUSE, R. M.; SIMÃO, F. A.; POZDNYAKOV, I. A.; IOANNIDIS, P.; ZDOBNOV, E. M. OrthoDB v8: Update of the

hierarchical catalog of orthologs and the underlying free software. **Nucleic Acids Research**, v. 43, n. D1, p. D250–D256, 11 2015. ISSN 13624962. Disponível em: <https://doi.org/10.1093/nar/gku1220>. Citado na página 37.

KRUEGER, F. **Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files**. 2015. https://www.bioinformatics.babraham.ac.uk/projects p. Disponível em: <https://github.com/FelixKrueger/TrimGalore>. Citado na página 36.

KUMAGAI, A.; KAWAMURA, S.; LEE, S. H.; ENDO, T.; RODRIGUEZ, M.; MIELENZ, J. R. Simultaneous saccharification and fermentation and a consolidated bioprocessing for *Hinoki cypress* and *Eucalyptus* after fibrillation by steam and subsequent wet-disk milling. **Bioresource Technology**, v. 162, p. 89–95, 6 2014. ISSN 18732976. Disponível em: <https://doi.org/10.1016/j.biortech.2014.03.110>. Citado na página 69.

LAMBERT, S. A.; YANG, A. W. H.; SASSE, A.; COWLEY, G.; ALBU, M.; CADDICK, M. X.; MORRIS, Q. D.; WEIRAUCH, M. T.; HUGHES, T. R. Similarity regression predicts evolution of transcription factor sequence specificity. **Nature Genetics**, v. 51, n. 6, p. 981–989, 2019. ISSN 1546-1718. Disponível em: <https://doi.org/10.1038/s41588-019-0411-1>. Citado na página 54.

LAMPA, S.; DAHLÖ, M.; OLASON, P. I.; HAGBERG, J.; SPJUTH, O. Lessons learned from implementing a national infrastructure in Sweden for storage and analysis of next-generation sequencing data. **GigaScience**, BioMed Central, v. 2, n. 1, p. 9, 6 2013. ISSN 2047217X. Disponível em: <https://doi.org/10.1186/2047-217X-2-9>. Citado na página 34.

LANGFELDER, K.; STREIBEL, M.; JAHN, B.; HAASE, G.; BRAKHAGE, A. A. Biosynthesis of fungal melanins and their importance for human pathogenic fungi. **Fungal Genetics and Biology**, Academic Press, v. 38, n. 2, p. 143–158, 3 2003. ISSN 10871845. Disponível em: <https://doi.org/10.1016/S1087-1845(02)00526-1>. Citado na página 47.

LANTZ, H.; ANGEL, V. D. D.; HJERDE, E.; STERCK, L.; CAPELLA-GUTIERREZ, S.; NOTREDAME, C.; PETTERSSON, O. V.; AMSELEM, J.; BOURI, L.; BOCS, S.; KLOPP, C.; GIBRAT, J. F.; VLASOVA, A.; LESKOSEK, B. L.; SOLER, L.; BINZER-PANCHAL, M. Ten steps to get started in Genome Assembly and Annotation. **F1000Research**, v. 7, n. 148, 2018. ISSN 1759796X. Disponível em: <https://doi.org/10.12688/f1000research.13598.1>. Citado 10 vezes nas páginas 34, 36, 37, 38, 39, 40, 41, 42, 43 e 44.

LATCHMAN, D. S. Eukaryotic Transcription Factors. In: MALOY, S.; HUGHES, K. B. T. B. E. o. G. S. E. (Ed.). **Brenner's Encyclopedia of Genetics: Second Edition**. San Diego: Academic Press, 2013. p. 537–539. ISBN 9780080961569. Disponível em: <https://doi.org/10.1016/B978-0-12-374984-0.01548-5>. Citado na página 53.

LECHNER, M.; FINDEISS, S.; STEINER, L.; MARZ, M.; STADLER, P. F.; PROHASKA, S. J. Proteinortho: Detection of (Co-)orthologs in large-scale analysis. **BMC Bioinformatics**, BioMed Central, v. 12, p. 124, 4 2011. ISSN 14712105. Disponível em: <https://doi.org/10.1186/1471-2105-12-124>. Citado na página 46.

LI, J.; LIU, G.; CHEN, M.; LI, Z.; QIN, Y.; QU, Y. Cellodextrin transporters play important roles in cellulase induction in the cellulolytic fungus *Penicillium oxalicum*. **Applied Microbiology and Biotechnology**, Germany, v. 97, n. 24, p. 10479–10488, 12 2013. ISSN 01757598. Disponível em: <https://doi.org/10.1007/s00253-013-5301-3>. Citado na página 62.

LIANG, Z.; SONG, L.; DENG, S.; ZHU, Y.; STAVITSKI, E.; ADZIC, R. R.; CHEN, J.; WANG, J. X. Direct 12-Electron Oxidation of Ethanol on a Ternary Au(core)-PtIr(Shell) Electrocatalyst. **Journal of the American Chemical Society**, American Chemical Society, v. 141, n. 24, p. 9629–9636, 6 2019. ISSN 0002-7863. Disponível em: <https://doi.org/10.1021/jacs.9b03474>. Citado na página 65.

LIU, G.; QIN, Y.; LI, Z.; QU, Y. Improving lignocellulolytic enzyme production with *Penicillium*: from strain screening to systems biology. **Biofuels**, Taylor & Francis, v. 4, n. 5, p. 523–534, 9 2013. ISSN 17597269. Disponível em: <https://doi.org/10.4155/bfs.13.38>. Citado na página 29.

LIU, G.; ZHANG, J.; BAO, J. Cost evaluation of cellulase enzyme for industrial-scale cellulosic ethanol production based on rigorous Aspen Plus modeling. **Bioprocess and Biosystems Engineering**, v. 39, n. 1, p. 133–140, 2016. ISSN 1615-7605. Disponível em: <https://doi.org/10.1007/s00449-015-1497-1>. Citado na página 69.

LOMBARD, V.; RAMULU, H. G.; DRULA, E.; COUTINHO, P. M.; HENRISSAT, B. The carbohydrate-active enzymes database (CAZy) in 2013. **Nucleic Acids Research**, v. 42, n. D1, p. D490–D495, 11 2014. ISSN 03051048. Disponível em: <https://doi.org/10.1093/nar/gkt1178>. Citado 2 vezes nas páginas 57 e 58.

LYND, L. R.; WEIMER, P. J.; ZYL, W. H. van; PRETORIUS, I. S. Microbial Cellulose Utilization: Fundamentals and Biotechnology. **Microbiology and Molecular Biology Reviews**, v. 66, n. 4, p. 739–739, 2002. ISSN 1092-2172. Disponível em: <http://dx.doi.org/10.1128/MMBR.66.3.506-577.2002>. Citado 2 vezes nas páginas 64 e 69.

MAEDA, R. N.; SERPA, V. I.; ROCHA, V. A. L.; MESQUITA, R. A. A.; ANNA, L. M. M. S.; CASTRO, A. M. H. D.; DRIEMEIER, C. E.; PEREIRA, N.; POLIKARPOV, I. Enzymatic hydrolysis of pretreated sugar cane bagasse using *Penicillium funiculosum* and *Trichoderma harzianum* cellulases. **Process Biochemistry**, v. 46, n. 5, p. 1196–1201, 5 2011. ISSN 13595113. Disponível em: <https://doi.org/10.1016/j.procbio.2011.01.022>. Citado na página 64.

MÄKELÄ, M. R.; MANSOURI, S.; WIEBENGA, A.; RYTIOJA, J.; VRIES, R. P. de; HILDÉN, K. S. *Penicillium subrubescens* is a promising alternative for *Aspergillus niger* in enzymatic plant biomass saccharification. **New Biotechnology**, Elsevier, v. 33, n. 6, p. 834–841, 12 2016. ISSN 18764347. Disponível em: <https://doi.org/10.1016/j.nbt.2016.07.014>. Citado na página 29.

MARTINS, L. F.; KOLLING, D.; CAMASSOLA, M.; DILLON, A. J. P.; RAMOS, L. P. Comparison of *Penicillium echinulatum* and *Trichoderma reesei* cellulases in relation to their activity against various cellulosic substrates. **Bioresource Technology**, Elsevier, v. 99, n. 5, p. 1417–1424, 3 2008. ISSN 09608524. Disponível em: <https://doi.org/10.1016/j.biortech.2007.01.060>. Citado na página 29.

MASTON, G. A.; EVANS, S. K.; GREEN, M. R. Transcriptional Regulatory Elements in the Human Genome. **Annual Review of Genomics and Human Genetics**, Annual Reviews, v. 7, n. 1, p. 29–59, 9 2006. ISSN 1527-8204. Disponível em: <http://dx.doi.org/10.1146/annurev.genom.7.080505.115623>. Citado na página 52.

MAUSER, W.; KLEPPER, G.; ZABEL, F.; DELZEIT, R.; HANK, T.; PUTZENLECHNER, B.; CALZADILLA, A. Global biomass production potentials exceed expected future demand without the need for cropland expansion. **Nature Communications**, Nature Publishing Group,

a division of Macmillan Publishers Limited. All Rights Reserved., v. 6, p. –, 11 2015. ISSN 20411723. Disponível em: <http://dx.doi.org/10.1038/ncomms9946>. Citado na página 68.

MCDONNELL, E.; STRASSER, K.; TSANG, A. Manual gene curation and functional annotation. In: VRIES, R. P. de; TSANG, A.; GRIGORIEV, I. V. (Ed.). **Methods in Molecular Biology**. New York, NY: Springer New York, 2018. v. 1775, p. 185–208. ISBN 978-1-4939-7804-5. Disponível em: <https://doi.org/10.1007/978-1-4939-7804-5_16>. Citado 2 vezes nas páginas 42 e 44.

MEDLAR, A. J.; TÖRÖNEN, P.; HOLM, L. AAI-profiler: Fast proteome-wide exploratory analysis reveals taxonomic identity, misclassification and contamination. **Nucleic Acids Research**, v. 46, n. W1, p. W479–W485, 5 2018. ISSN 13624962. Disponível em: <https://doi.org/10.1093/nar/gky359>. Citado na página 46.

MENEGOL, D.; FONTANA, R. C.; DILLON, A. J. P.; CAMASSOLA, M. Second-generation ethanol production from elephant grass at high total solids. **Bioresource Technology**, Elsevier, v. 211, p. 280–290, 7 2016. ISSN 18732976. Disponível em: <https://doi.org/10.1016/j.biortech.2016.03.098>. Citado 2 vezes nas páginas 31 e 64.

MENEGOL, D.; SCHOLL, A. L.; FONTANA, R. C.; DILLON, A. J. P.; CAMASSOLA, M. Increased release of fermentable sugars from elephant grass by enzymatic hydrolysis in the presence of surfactants. **Energy Conversion and Management**, v. 88, p. 1252–1256, 12 2014. ISSN 01968904. Disponível em: <https://doi.org/10.1016/j.enconman.2014.02.071>. Citado na página 31.

MENEGOL, D.; SCHOLL, A. L.; FONTANA, R. C.; DILLON, A. J. P.; CAMASSOLA, M. Potential of a *Penicillium echinulatum* enzymatic complex produced in either submerged or solid-state cultures for enzymatic hydrolysis of elephant grass. **Fuel**, v. 133, p. 232–240, 2014. ISSN 00162361. Disponível em: <http://dx.doi.org/10.1016/j.fuel.2014.05.003>. Citado na página 31.

MERCATELLI, D.; SCALAMBRA, L.; TRIBOLI, L.; RAY, F.; GIORGI, F. M. Gene regulatory network inference resources: A practical overview. **Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms**, v. 1863, n. 6, p. 194430, 2020. ISSN 1874-9399. Disponível em: <https://doi.org/10.1016/j.bbagrm.2019.194430>. Citado 2 vezes nas páginas 49 e 50.

MITCHELL, A. L.; ATTWOOD, T. K.; BABBITT, P. C.; BLUM, M.; BORK, P.; BRIDGE, A.; BROWN, S. D.; CHANG, H. Y.; EL-GEBALI, S.; FRASER, M. I.; GOUGH, J.; HAFT, D. R.; HUANG, H.; LETUNIC, I.; LOPEZ, R.; LUCIANI, A.; MADEIRA, F.; MARCHLER-BAUER, A.; MI, H.; NATALE, D. A.; NECCI, M.; NUKA, G.; ORENGO, C.; PANDURANGAN, A. P.; PAYSAN-LAFOSSE, T.; PESSEAT, S.; POTTER, S. C.; QURESHI, M. A.; RAWLINGS, N. D.; REDASCHI, N.; RICHARDSON, L. J.; RIVOIRE, C.; SALAZAR, G. A.; SANGRADOR-VEGAS, A.; SIGRIST, C. J.; SILLITOE, I.; SUTTON, G. G.; THANKI, N.; THOMAS, P. D.; TOSATTO, S. C.; YONG, S. Y.; FINN, R. D. InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. **Nucleic Acids Research**, v. 47, n. D1, p. D351–D360, 11 2019. ISSN 13624962. Disponível em: <https://doi.org/10.1093/nar/gky1100>. Citado 2 vezes nas páginas 44 e 53.

MONTEIRO, P. T.; OLIVEIRA, J.; PAIS, P.; ANTUNES, M.; PALMA, M.; CAVALHEIRO, M.; GALOCHA, M.; GODINHO, C. P.; MARTINS, L. C.; BOURBON, N.; MOTA, M. N.; RIBEIRO, R. A.; VIANA, R.; SÁ-CORREIA, I.; TEIXEIRA, M. C. YEASTRACT+: a portal

for cross-species comparative genomics of transcription regulation in yeasts. **Nucleic acids research**, v. 48, n. D1, p. D642–D649, 10 2020. ISSN 13624962. Disponível em: <https://doi.org/10.1093/nar/gkz859>. Citado na página 49.

MUELLER, G. M.; BILLS, G. F. Introduction. In: MUELLER, G. M.; BILLS, G. F.; FOSTER, M. S. B. T. B. o. F. (Ed.). **Biodiversity of Fungi: Inventory and Monitoring Methods**. Burlington: Academic Press, 2004. p. 1–4. ISBN 9780080470269. Disponível em: <https://doi.org/10.1016/B978-012509551-8/50003-9>. Citado na página 47.

MUSZEWSKA, A.; STEPNIEWSKA-DZIUBINSKA, M. M.; STECZKIEWICZ, K.; PAWLOWSKA, J.; DZIEDZIC, A.; GINALSKI, K. Fungal lifestyle reflected in serine protease repertoire. **Scientific Reports**, Nature Publishing Group UK, v. 7, n. 1, p. 9147, 8 2017. ISSN 20452322. Disponível em: <https://doi.org/10.1038/s41598-017-09644-w>. Citado na página 27.

NOVELLO, M.; VILASBOA, J.; SCHNEIDER, W. D. H.; REIS, L. D.; FONTANA, R. C.; CAMASSOLA, M. Enzymes for second generation ethanol: Exploring new strategies for the use of xylose. **RSC Advances**, v. 4, n. 41, p. 21361–21368, 2014. ISSN 20462069. Disponível em: <https://doi.org/10.1039/c4ra00909f>. Citado na página 31.

NYGAARD, S.; HU, H.; LI, C.; SCHIØTT, M.; CHEN, Z.; YANG, Z.; XIE, Q.; MA, C.; DENG, Y.; DIKOW, R. B.; RABELING, C.; NASH, D. R.; WCISLO, W. T.; BRADY, S. G.; SCHULTZ, T. R.; ZHANG, G.; BOOMSMA, J. J. Reciprocal genomic evolution in the ant-fungus agricultural symbiosis. **Nature Communications**, v. 7, n. 1, p. 12233, 2016. ISSN 20411723. Disponível em: <https://doi.org/10.1038/ncomms12233>. Citado na página 48.

PAL, S.; GUPTA, R.; DAVULURI, R. V. Genome-wide mapping of RNA Pol-II promoter usage in mouse tissues by ChIP-seq. **Methods in Molecular Biology**, Oxford University Press, v. 1176, n. 1, p. 1–9, 8 2014. ISSN 10643745. Disponível em: <https://doi.org/10.1007/978-1-4939-992-6_1>. Citado na página 51.

PANAHI, H. K. S.; DEHHAGHI, M.; KINDER, J. E.; EZEJI, T. C. A review on green liquid fuels for the transportation sector: A prospect of microbial solutions to climate change. **Journal of Survey in Fisheries Sciences**, Faculty of Medicine and Health Sciences, Macquarie University, NSW, Australia., v. 6, n. 3, p. 995–1024, 2019. ISSN 23687487. Disponível em: <https://doi.org/10.18331/brj2019.6.3.2>. Citado 2 vezes nas páginas 63 e 68.

PANCHAPAKESAN, A.; SHANKAR, N. Fungal Cellulases: An Overview. **New and Future Developments in Microbial Biotechnology and Bioengineering: Microbial Cellulase System Properties and Applications**, Elsevier, p. 9–18, 1 2016. Disponível em: <https://doi.org/10.1016/B978-0-444-63507-5.00002-2>. Citado na página 58.

PARISUTHAM, V.; KIM, T. H.; LEE, S. K. Feasibilities of consolidated bioprocessing microbes: From pretreatment to biofuel production. **Bioresource Technology**, v. 161, p. 431–440, 6 2014. ISSN 18732976. Disponível em: <https://doi.org/10.1016/j.biortech.2014.03.114>. Citado 2 vezes nas páginas 68 e 70.

PARKIN, E. A. The Digestive Enzymes of Some Wood-Boring Beetle Larvae. **Journal of Experimental Biology**, v. 17, n. 4, p. 364–377, 11 1940. ISSN 0022-0949. Disponível em: <http://jeb.biologists.org/content/17/4/364.abstract>. Citado na página 28.

PAULY, M.; KEEGSTRA, K. Cell-wall carbohydrates and their modification as a resource for biofuels. **Plant Journal**, Blackwell Publishing Ltd, v. 54, n. 4, p. 559–568, 2008. ISSN 09607412. Disponível em: <http://dx.doi.org/10.1111/j.1365-313X.2008.03463.x>. Citado na página 68.

PÉREZ, J.; MUÑOZ-DORADO, J.; RUBIA, T. D. L.; MARTÍNEZ, J. Biodegradation and biological treatments of cellulose, hemicellulose and lignin: An overview. **International Microbiology**, v. 5, n. 2, p. 53–63, 2002. ISSN 16181905. Disponível em: <http://dx.doi.org/10.1007/s10123-002-0062-3>. Citado 4 vezes nas páginas 56, 57, 61 e 69.

PRALEA, I. E.; MOLDOVAN, R. C.; PETRACHE, A. M.; ILIEș, M.; HEGHEș, S. C.; IELCIU, I.; NICOARĂ, R.; MOLDOVAN, M.; ENE, M.; RADU, M.; UIFĂLEAN, A.; IUGA, C. A. From extraction to advanced analytical methods: The challenges of melanin analysis. **International Journal of Molecular Sciences**, MDPI, v. 20, n. 16, p. 3943, 8 2019. ISSN 14220067. Disponível em: <https://doi.org/10.3390/ijms20163943>. Citado na página 48.

PUNT, P. J.; BIEZEN, N. V.; CONESA, A.; ALBERS, A.; MANGNUS, J.; HONDEL, C. V. D. Filamentous fungi as cell factories for heterologous protein production. **Trends in Biotechnology**, Elsevier, v. 20, n. 5, p. 200–206, 2002. ISSN 01677799. Disponível em: <http://dx.doi.org/10.1016/S0167-7799(02)01933-9>. Citado na página 59.

RAMOS, L. P.; SILVA, L. da; BALLEM, A. C.; PITARELO, A. P.; CHIARELLO, L. M.; SILVEIRA, M. H. L. Enzymatic hydrolysis of steam-exploded sugarcane bagasse using high total solids and low enzyme loadings. **Bioresource Technology**, v. 175, p. 195–202, 2015. ISSN 18732976. Disponível em: <http://dx.doi.org/10.1016/j.biortech.2014.10.087>. Citado na página 64.

RAN, Y.; WANG, Y. Z.; LIAO, Q.; ZHU, X.; CHEN, R.; LEE, D. J.; WANG, Y. M. Effects of operation conditions on enzymatic hydrolysis of high-solid rice straw. **International Journal of Hydrogen Energy**, v. 37, n. 18, p. 13660–13666, 9 2012. ISSN 03603199. Disponível em: <https://doi.org/10.1016/j.ijhydene.2012.02.080>. Citado na página 64.

RAU, M.; HEIDEMANN, C.; PASCOALIN, A. M.; FILHO, E. X. F.; CAMASSOLA, M.; DILLON, A. J. P.; Fernandes Das Chagas, C.; ANDREAUS, J. Application of cellulases from *Acrophialophora nainiana* and *Penicillium echinulatum* in textile processing of cellulosic fibres. **Biocatalysis and Biotransformation**, Taylor & Francis, v. 26, n. 5, p. 383–390, jan 2008. ISSN 1024-2422. Disponível em: <https://doi.org/10.1080/10242420802249430>. Citado na página 30.

REIS, L. dos; SCHNEIDER, W. D. H.; FONTANA, R. C.; CAMASSOLA, M.; DILLON, A. J. Cellulase and Xylanase Expression in Response to Different pH Levels of *Penicillium echinulatum* S1M29 Medium. **Bioenergy Research**, v. 7, n. 1, p. 60–67, 2014. ISSN 19391234. Disponível em: <http://dx.doi.org/10.1007/s12155-013-9345-0>. Citado na página 69.

REIS, T. F. D.; LIMA, P. B. A. D.; PARACHIN, N. S.; MINGOSSI, F. B.; OLIVEIRA, J. V. D. C.; RIES, L. N. A.; GOLDMAN, G. H. Identification and characterization of putative xylose and cellobiose transporters in *Aspergillus nidulans*. **Biotechnology for Biofuels**, v. 9, n. 1, p. 204, 2016. ISSN 17546834. Disponível em: <https://doi.org/10.1186/s13068-016-0611-1>. Citado na página 62.

RIBEIRO, D. A.; COTA, J.; ALVAREZ, T. M.; BRÜCHLI, F.; BRAGATO, J.; PEREIRA, B. M.; PAULETTI, B. A.; JACKSON, G.; PIMENTA, M. T.; MURAKAMI, M. T.; CAMASSOLA,

M.; RULLER, R.; DILLON, A. J.; PRADELLA, J. G.; LEME, A. F. P.; SQUINA, F. M. The *Penicillium echinulatum* Secretome on Sugar Cane Bagasse. **PLoS ONE**, Public Library of Science, v. 7, n. 12, p. e50571–e50571, 2012. ISSN 19326203. Disponível em: <https://doi.org/10.1371/journal.pone.0050571>. Citado na página 30.

ROMBAUTS, S.; FLORQUIN, K.; LESCOT, M.; MARCHAL, K.; ROUZÉ, P.; PEER, Y. V. D. Computational approaches to identify promoters and cis-regulatory elements in plant genomes. **Plant Physiology**, The American Society for Plant Biologists, v. 132, n. 3, p. 1162–1176, 3 2003. ISSN 00320889. Disponível em: <https://doi.org/10.1104/pp.102.017715>. Citado 2 vezes nas páginas 52 e 53.

ROY, A. L.; SINGER, D. S. Core promoters in transcription: Old problem, new insights. **Trends in Biochemical Sciences**, v. 40, n. 3, p. 165–171, 2 2015. ISSN 13624326. Disponível em: <https://doi.org/10.1016/j.tibs.2015.01.007>. Citado na página 51.

RUBINI, M. R.; DILLON, A. J.; KYAW, C. M.; FARIA, F. P.; POÇAS-FONSECA, M. J.; SILVA-PEREIRA, I. Cloning, characterization and heterologous expression of the first *Penicillium echinulatum* cellulase gene. **Journal of Applied Microbiology**, John Wiley & Sons, Ltd (10.1111), v. 108, n. 4, p. 1187–1198, 4 2010. ISSN 13645072. Disponível em: <https://doi.org/10.1111/j.1365-2672.2009.04528.x>. Citado na página 30.

RYTIOJA, J.; HILDÉN, K.; YUZON, J.; HATAKKA, A.; VRIES, R. P. de; MÄKELÄ, M. R. Plant-Polysaccharide-Degrading Enzymes from *Basidiomycetes*. **Microbiology and Molecular Biology Reviews**, American Society for Microbiology, 1752 N St., N.W., Washington, DC, v. 78, n. 4, p. 614–649, 12 2014. ISSN 1092-2172. Disponível em: <https://doi.org/10.1128/mmbr.00035-14>. Citado 4 vezes nas páginas 57, 58, 59 e 61.

SANDELIN, A.; CARNINCI, P.; LENHARD, B.; PONJAVIC, J.; HAYASHIZAKI, Y.; HUME, D. A. Mammalian RNA polymerase II core promoters: Insights from genome-wide studies. **Nature Reviews Genetics**, Nature Publishing Group, v. 8, n. 6, p. 424–436, 6 2007. ISSN 14710056. Disponível em: <http://dx.doi.org/10.1038/nrg2026>. Citado na página 51.

SAYERS, E. W.; AGARWALA, R.; BOLTON, E. E.; BRISTER, J.; CANESE, K.; CLARK, K.; CONNOR, R.; FIORINI, N.; FUNK, K.; HEFFERON, T.; HOLMES, J.; KIM, S.; KIMCHI, A.; KITTS, P. A.; LATHROP, S.; LU, Z.; MADDEN, T. L.; MARCHLER-BAUER, A.; PHAN, L.; SCHNEIDER, V. A.; SCHOCH, C. L.; PRUITT, K. D.; OSTELL, J. Database resources of the National Center for Biotechnology Information. **Nucleic Acids Research**, v. 47, n. D1, p. D23–D28, 11 2018. ISSN 0305-1048. Disponível em: <https://doi.org/10.1093/nar/gky1069>. Citado na página 44.

SCHNEIDER, W. D. H.; FONTANA, R. C.; BAUDEL, H. M.; SIQUEIRA, F. G. de; RENCORET, J.; GUTIÉRREZ, A.; EUGENIO, L. I. de; PRIETO, A.; MARTÍNEZ, M. J.; MARTÍNEZ, T.; DILLON, A. J. P.; CAMASSOLA, M. Lignin degradation and detoxification of eucalyptus wastes by on-site manufacturing fungal enzymes to enhance second-generation ethanol yield. **Applied Energy**, v. 262, p. 114493, 2020. ISSN 0306-2619. Disponível em: <https://doi.org/10.1016/j.apenergy.2020.114493>. Citado na página 32.

SCHNEIDER, W. D. H.; GONÇALVES, T. A.; UCHIMA, C. A.; COUGER, M. B.; PRADE, R.; SQUINA, F. M.; DILLON, A. J. P.; CAMASSOLA, M. *Penicillium echinulatum* secretome analysis reveals the fungi potential for degradation of lignocellulosic biomass. **Biotechnology for Biofuels**, BioMed Central, v. 9, n. 1, p. 66, 3 2016. ISSN 17546834. Disponível em: <https://doi.org/10.1186/s13068-016-0476-3>. Citado 4 vezes nas páginas 28, 30, 31 e 71.

SCHNEIDER, W. D. H.; GONÇALVES, T. A.; UCHIMA, C. A.; REIS, L. d.; FONTANA, R. C.; SQUINA, F. M.; DILLON, A. J. P.; CAMASSOLA, M. Comparison of the production of enzymes to cell wall hydrolysis using different carbon sources by *Penicillium echinulatum* strains and its hydrolysis potential for lignocelullosic biomass. **Process Biochemistry**, v. 66, p. 162–170, 2018. ISSN 1359-5113. Disponível em: <https://doi.org/10.1016/j.procbio.2017.11.004>. Citado 2 vezes nas páginas 30 e 32.

SCHNEIDER, W. D. H.; REIS, L. D.; CAMASSOLA, M.; DILLON, A. J. P. Morphogenesis and production of enzymes by *Penicillium echinulatum* in response to different carbon sources. **BioMed Research International**, v. 2014, p. 10, 2014. ISSN 23146141. Disponível em: <https://doi.org/10.1155/2014/254863>. Citado na página 31.

SHEN, F.; HU, J.; ZHONG, Y.; LIU, M. L.; SADDLER, J. N.; LIU, R. Ethanol production from steam-pretreated sweet sorghum bagasse with high substrate consistency enzymatic hydrolysis. **Biomass and Bioenergy**, v. 41, p. 157–164, 6 2012. ISSN 09619534. Disponível em: <https://doi.org/10.1016/j.biombioe.2012.02.022>. Citado na página 64.

SHIDA, Y.; YAMAGUCHI, K.; NITTA, M.; NAKAMURA, A.; TAKAHASHI, M.; KIDOKORO, S. I.; MORI, K.; TASHIRO, K.; KUHARA, S.; MATSUZAWA, T.; YAOI, K.; SAKAMOTO, Y.; TANAKA, N.; MORIKAWA, Y.; OGASAWARA, W. The impact of a single-nucleotide mutation of *bgl2* on cellulase induction in a *Trichoderma reesei* mutant. **Biotechnology for Biofuels**, v. 8, n. 1, p. 230, 2015. ISSN 17546834. Disponível em: <https://doi.org/10.1186/s13068-015-0420-y>. Citado na página 61.

SILVA, R. da; CAETANO, R.; OKAMOTO, D.; OLIVEIRA, L. de; BERTOLIN, T.; JULIANO, M.; JULIANO, L.; OLIVEIRA, A. H. de; ROSAE, J.; CABRAL, H. **The Identification and Biochemical Properties of the Catalytic Specificity of a Serine Peptidase Secreted by *Aspergillus fumigatus* Fresenius**. 2014. 663–671 p. Disponível em: <https://doi.org/10.2174/092986652166614040811464 6>. Citado na página 27.

SIMSKE, S. Introduction, overview, and applications. In: SIMSKE, S. B. T. M.-A. (Ed.). **Meta-Analytics: Consensus Approaches and System Patterns for Data Analysis**. Morgan Kaufmann, 2019. p. 1–98. ISBN 978-0-12-814623-1. Disponível em: <https://doi.org/10.1016/b978-0-12-814623-1.00001-0>. Citado na página 45.

SINGHANIA, R. R.; PATEL, A. K.; SUKUMARAN, R. K.; LARROCHE, C.; PANDEY, A. Role and significance of beta-glucosidases in the hydrolysis of cellulose for bioethanol production. **Bioresource Technology**, Elsevier, v. 127, p. 500–507, 1 2013. ISSN 18732976. Disponível em: <https://doi.org/10.1016/j.biortech.2012.09.012>. Citado na página 29.

SINGHANIA, R. R.; SAINI, J. K.; SAINI, R.; ADSUL, M.; MATHUR, A.; GUPTA, R.; TULI, D. K. Bioethanol production from wheat straw via enzymatic route employing *Penicillium janthinellum* cellulases. **Bioresource Technology**, v. 169, p. 490–495, 10 2014. ISSN 18732976. Disponível em: <https://doi.org/10.1016/j.biortech.2014.07.011>. Citado na página 63.

SOHN, J. I.; NAM, J. W. The present and future of de novo whole-genome assembly. **Briefings in Bioinformatics**, v. 19, n. 1, p. 23–40, 10 2018. ISSN 14774054. Disponível em: <https://doi.org/10.1093/bib/bbw096>. Citado 2 vezes nas páginas 35 e 36.

SOUDHAM, V. P.; RAUT, D. G.; ANUGWOM, I.; BRANDBERG, T.; LARSSON, C.; MIKKOLA, J. P. Coupled enzymatic hydrolysis and ethanol fermentation: ionic liquid pretreatment for enhanced yields. **Biotechnology for Biofuels**, BioMed Central, London, v. 8, n. 1, p. 135–,

8 2015. ISSN 17546834. Disponível em: <https://doi.org/10.1186/s13068-015-0310-3>. Citado na página 64.

STAJICH, J. E. Fungal Genomes and Insights into the Evolution of the Kingdom. **Microbiology Spectrum**, v. 5, n. 4, p. 619–633, 7 2017. ISSN 2165-0497. Disponível em: <https://doi.org/10.1128/microbiolspec.funk-0055-2016>. Citado 2 vezes nas páginas 29 e 48.

STEENWYK, J. L.; SHEN, X. X.; LIND, A. L.; GOLDMAN, G. H.; ROKAS, A. A robust phylogenomic time tree for biotechnologically and medically important fungi in the genera *Aspergillus* and *Penicillium*. **mBio**, v. 10, n. 4, p. 00925–19, 8 2019. ISSN 21507511. Disponível em: <https://doi.org/10.1128/mBio.00925-19>. Citado na página 46.

TODD, R. B.; ZHOU, M.; OHM, R. A.; LEEGGANGERS, H. A.; VISSER, L.; VRIES, R. P. de. Prevalence of transcription factors in ascomycete and basidiomycete fungi. **BMC Genomics**, v. 15, n. 1, p. 1–12, 2014. ISSN 14712164. Disponível em: <http://dx.doi.org/10.1186/1471-2164-15-214>. Citado na página 53.

VAISHNAV, N.; SINGH, A.; ADSUL, M.; DIXIT, P.; SANDHU, S. K.; MATHUR, A.; PURI, S. K.; SINGHANIA, R. R. *Penicillium*: The next emerging champion for cellulase production. **Bioresource Technology Reports**, Elsevier, v. 2, p. 131–140, 6 2018. ISSN 2589014X. Disponível em: <https://doi.org/10.101et6/j.biteb.2018.04.003>. Citado 2 vezes nas páginas 29 e 60.

VEGA, F.; BLACKWELL, M. **Insect-Fungal Associations: Ecology and Evolution**. [S.l.]: Oxford University Press, 2005. – p. ISSN 978-0195166521. Citado na página 29.

VISAGIE, C. M.; HOUBRAKEN, J.; FRISVAD, J. C.; HONG, S. B.; KLAASSEN, C. H.; PERRONE, G.; SEIFERT, K. A.; VARGA, J.; YAGUCHI, T.; SAMSON, R. A. Identification and nomenclature of the genus *Penicillium*. **Studies in Mycology**, CBS Fungal Biodiversity Centre, v. 78, n. 1, p. 343–371, 6 2014. ISSN 01660616. Disponível em: <https://doi.org/10.1016/j.simyco.2014.09.001>. Citado 3 vezes nas páginas 28, 32 e 33.

WANG, M.; LI, Z.; FANG, X.; WANG, L.; QU, Y. Cellulolytic Enzyme Production and Enzymatic Hydrolysis for Second-Generation Bioethanol Production. In: BAI, F.-W.; LIU, C.-G.; HUANG, H.; TSAO, G. T. (Ed.). **Biotechnology in China III: Biofuels and Bioenergy**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. p. 1–24. ISBN 978-3-642-28478-6. Disponível em: <https://doi.org/10.1007/10_2011_131>. Citado na página 60.

WATERHOUSE, R. M.; SEPPEY, M.; SIMAO, F. A.; MANNI, M.; IOANNIDIS, P.; KLIOUTCHNIKOV, G.; KRIVENTSEVA, E. V.; ZDOBNOV, E. M. BUSCO applications from quality assessments to gene prediction and phylogenomics. **Molecular Biology and Evolution**, v. 35, n. 3, p. 543–548, 12 2018. ISSN 15371719. Disponível em: <https://doi.org/10.1093/molbev/msx319>. Citado 3 vezes nas páginas 37, 42 e 46.

WHEELER, Q.; CROWSON, R. A. The Biology of the Coleoptera. **Systematic Zoology**, Oxford University Press, Society of Systematic Biologists, Taylor & Francis, Ltd., v. 31, n. 3, p. 342, 4 1982. ISSN 00397989. Disponível em: <https://doi.org/10.2307/2413243>. Citado na página 28.

WOOLFIT, M.; BROMHAM, L. Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. **Molecular Biology and Evolution**, v. 20, n. 9, p. 1545–1555, 9 2003. ISSN 07374038. Disponível em: <https://doi.org/10.1093/molbev/msg167>. Citado na página 29.

XIA, Q.; CHEN, Z.; SHAO, Y.; GONG, X.; WANG, H.; LIU, X.; PARKER, S. F.; HAN, X.; YANG, S.; WANG, Y. Direct hydrodeoxygenation of raw woody biomass into liquid alkanes. **Nature Communications**, Nature Publishing Group, v. 7, p. 11162–, 2 2016. ISSN 20411723. Disponível em: <https://doi.org/10.1038/ncomms11162>. Citado na página 69.

YANDELL, M.; ENCE, D. A beginner's guide to eukaryotic genome annotation. **Nature Reviews Genetics**, Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., v. 13, n. 5, p. 329–342, 5 2012. ISSN 14710056. Disponível em: <http://dx.doi.org/10.1038/nrg3174>. Citado 3 vezes nas páginas 38, 39 e 41.

YAO, G.; WU, R.; KAN, Q.; GAO, L.; LIU, M.; YANG, P.; DU, J.; LI, Z.; QU, Y. Production of a high-efficiency cellulase complex via β-glucosidase engineering in *Penicillium oxalicum*. **Biotechnology for Biofuels**, BioMed Central, v. 9, n. 1, p. 78, 3 2016. ISSN 17546834. Disponível em: <https://doi.org/10.1186/s13068-016-0491-4>. Citado na página 61.

ZAMPIERI, D. **Gene expression and activities of cellulases, β-glucosidases, xylanases and swollenins of *Penicillium echinulatum* S1M29**. 131 p. Tese (Doutorado) — University of Caxias do Sul, Biotechnology Institute, 2015. Disponível em: <https://repositorio.ucs.br/xmlui/handle/11338/1087>. Citado na página 31.

ZAMPIERI, D.; NORA, L. C.; BASSO, V.; CAMASSOLA, M.; DILLON, A. J. Validation of reference genes in *Penicillium echinulatum* to enable gene expression study using real-time quantitative RT-PCR. **Current Genetics**, v. 60, n. 3, p. 231–236, 2014. ISSN 14320983. Disponível em: <http://dx.doi.org/10.1007/s00294-014-0421-6>. Citado na página 31.

ZHANG, W.; KOU, Y.; XU, J.; CAO, Y.; ZHAO, G.; SHAO, J.; WANG, H.; WANG, Z.; BAO, X.; CHEN, G.; LIU, W. Two major facilitator superfamily sugar transporters from *Trichoderma reesei* and their roles in induction of cellulase biosynthesis. **Journal of Biological Chemistry**, v. 288, n. 46, p. 32861–32872, 11 2013. ISSN 00219258. Disponível em: <https://doi.org/10.1074/jbc.M113.505826>. Citado na página 62.

ZHAO, Q. Y.; WANG, Y.; KONG, Y. M.; LUO, D.; LI, X.; HAO, P. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. **BMC bioinformatics**, v. 12 Suppl 1, n. 14, p. S2–, 2011. ISSN 14712105. Disponível em: <https://doi.org/10.1186/1471-2105-12-S14-S2>. Citado na página 40.