Almir José Ferreira

**ANÁLISE E ANOTAÇÃO DO GENOMA DE *Epicoccum*
nigrum E METABOLISMO SECUNDÁRIO**

Tese apresentada ao Programa de Pós-Graduação de Microbiologia, do Instituto de Ciências Biomédicas da Universidade de São Paulo, para obtenção do Título de Doutor em Ciências.

São Paulo
2016

Almir José Ferreira

**ANÁLISE E ANOTAÇÃO DO GENOMA DE *Epicoccum*
nigrum E METABOLISMO SECUNDÁRIO**

Tese apresentada ao Programa de Pós-Graduação de Microbiologia, do Instituto de Ciências Biomédicas da Universidade de São Paulo, para obtenção do Título de Doutor em Ciências.

Área de concentração: Microbiologia

Orientador:
Prof. Dr. Welington Luiz de Araújo

Versão Original

São Paulo
2016

DADOS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
Serviço de Biblioteca e Informação Biomédica do
Instituto de Ciências Biomédicas da Universidade de São Paulo

© reprodução total

Ferreira, Almir José.

Análise e anotação do genoma de *Epicoccum nigrum* e metabolismo secundário / Almir José Ferreira. -- São Paulo, 2016.

Orientador: Prof. Dr. Welington Luiz de Araújo.

Tese (Doutorado) – Universidade de São Paulo. Instituto de Ciências Biomédicas. Departamento de Microbiologia. Área de concentração: Microbiologia. Linha de pesquisa: Genética de micro-organismos.

Versão do título para o inglês: Analysis and annotation of *Epicoccum nigrum* genome and secondary metabolism.

1. *Epicoccum nigrum* 2. Bioinformática 3. Metabólitos secundários
4. Anotação 5. Genoma 6. Transcriptoma I. Araújo, Prof. Dr.
Welington Luiz de II. Universidade de São Paulo. Instituto de Ciências
Biomédicas. Programa de Pós-Graduação em Microbiologia III. Título.

ICB/SBIB068/2016

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS

Candidato(a): Almir José Ferreira.

Título da Tese: Análise e anotação do genoma de *Epicoccum nigrum* e metabolismo secundário.

Orientador(a): Prof. Dr. Welington Luiz de Araújo.

A Comissão Julgadora dos trabalhos de Defesa da Tese de Doutorado, em sessão pública realizada a/...../....., considerou

() Aprovado(a)

() Reprovado(a)

Examinador(a): Assinatura:
 Nome:
 Instituição:

Examinador(a): Assinatura:
 Nome:
 Instituição:

Examinador(a): Assinatura:
 Nome:
 Instituição:

Examinador(a): Assinatura:
 Nome:
 Instituição:

Presidente: Assinatura:
 Nome:
 Instituição:

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS

Cidade Universitária "Armando de Salles Oliveira"
Av. Prof. Lineu Prestes, 2415 – CEP. 05508-000 São Paulo, SP – Brasil
Telefone : (55) (11) 3091-7733 - telefax : (55) (11) 3091-8405
e-mail: cep@icb.usp.br

Comissão de Ética em Pesquisa

CERTIFICADO DE ISENÇÃO

Certificamos que o Protocolo CEP-ICB Nº 710/15 referente ao projeto intitulado: "*Análise e anotação de genes de *Epicoccum nigrum* associados ao metabolismo secundário*" sob a responsabilidade de **Almir José Ferreira**, foi analisado na presente data pela CEUA - COMISSÃO DE ÉTICA NO USO DE ANIMAIS e pela CEPSh- COMISSÃO DE ÉTICA EM PESQUISA COM SERES HUMANOS, tendo sido deliberado que o referido projeto não utilizará animais que estejam sob a égide da lei 11.794 de 8 de outubro de 2008, nem envolverá procedimentos regulados pela Resolução CONEP nº466 de 2012.

São Paulo, 24 de fevereiro de 2015.

PROF. DR. WOTHAN TAVARES DE LIMA
Coordenador da CEUA - ICB/USP

PROF. DR. PAOLO M.A ZANOTTO
Coordenador da CEPsh - ICB/USP

DEDICO

A minha mãe e irmãs,

Pelo amor, carinho e o altruísmo para eu chegasse aqui

OFEREÇO

*Aos meus amigos,
principalmente do laboratório que
me ajudaram a seguir em frente
nestes nove anos*

OBRIGADO!

AGRADECIMENTOS

Ao Prof. Dr. Welington Luiz de Araújo, pela oportunidade dada, pela confiança depositada em meu trabalho, os ensinamentos e por me ajudar nos momentos mais difíceis, atuando muitas vezes até como um pai para mim.

Aos meus queridos amigos que convívio dentro e fora do laboratório os quais contribuíram não só para meu trabalho, mas também para o meu engrandecimento tanto profissional como humano:

À Aline Neves pela confiança inicial no meu trabalho e auxílio nos primeiros passos.

À Carol pelas risadas e bom humor contagiante;

À Daiene pela amizade e carinho e risadas desde os tempos da graduação;

À Emy Mano pela inspiração, força e cumplicidade;

À Jennifer pelo carinho e força inspiradora;

À Léia Fávaro pelos conselhos e apoio;

À Eliane pela amizade, prestatividade e bom coração;

À Lina pela amizade, coragem e senso de justiça;

À Mabel pela amizade e doçura;

À Manuela pelos conselhos, carinho e inspiração;

À Priscila pela inexplicável afinidade, cumplicidade e tantos momentos;

À Raissa pela amizade e dividir comigo as delicias e dores do projeto;

À Roberta pela amizade e cuidados;

À Thais Prado pelo apoio nas atividades iniciais no laboratório e amizade.

Ao Titi pela amizade e piadas;

Ao Fábio Moura pela força nos momentos que precisei;

Ao Felipe pela amizade e estímulo para fazer o melhor dentro e fora do laboratório.

Ao Ricardo pela amizade e risadas;

Ao Alan pelo apoio, acolhimento e ensinamentos;

Ao Arthur pela prestatividade, comprometimento, esforços para realização do trabalho e conselhos;

Aos meus queridos amigos que sempre me apoiaram, mesmo nos momentos mais complicados:

Danilo, Evelyne, Juninho, Jefferson, Paulinha, Tânia, Solange, Agnaldo, Mariote Ramón e Geovanny.

À minha família que fez o melhor por mim se esforçando ao máximo para me dar estrutura, amor e consolo nas horas que precisei.

À Universidade de São Paulo, ICB e ao programa de microbiologia pela oportunidade de fazer parte do corpo estudantil e estrutura oferecida.

À CAPES pela bolsa de estudos que permitiu minha dedicação em tempo integral.

À FAPESP pelo financiamento do projeto de pesquisa.

Muito Obrigado!

*“Aprender é a única coisa de que a mente
nunca se cansa, nunca tem medo e nunca se
arrepende”.*

Leonardo da Vinci

RESUMO

FERREIRA, A. J. **Análise e anotação do genoma de *Epicoccum nigrum* e metabolismo secundário**, 2016. 135 f. Tese (Doutorado em Microbiologia) – Instituto de Ciências Biomédicas, Universidade de São Paulo, São Paulo, 2016.

Atualmente tem aumentado o interesse pelos fungos endofíticos como uma fonte prolífica por causa de seu potencial no controle biológico e produção de metabólitos com atividades biológicas. Apesar das crescentes descobertas de novos compostos de fungos endofíticos, existe uma lacuna no conhecimento a respeito das vias biossintéticas responsáveis pela produção de metabólitos bioativos. Neste contexto, *Epicoccum nigrum* é um fungo endofítico que tem sido usado no biocontrole de fitopatógenos em diferentes plantas hospedeiras. Além disso, produz uma série de compostos com interesse biotecnológico, incluindo antimicrobianos. Neste trabalho, foi utilizado o isolado *E. nigrum* P16, sendo confirmada por ensaios específicos sua capacidade de produzir metabólitos secundários com atividade antimicrobiana. Foi utilizado um conjunto de sequências genômicas previamente montadas. As sequências foram anotadas para compreensão funcional utilizando a plataforma EGene2 incluindo alguns componentes específicos desenvolvido neste trabalho. Foram estimados 10.320 genes codificadores de proteínas, além de tRNAs, rRNAs e ncRNAs. As proteínas preditas foram comparadas a proteomas de outros fungos, classificadas em grupos ortólogos e comparadas filogeneticamente. Além disso, foi desenvolvido Synteny Clusters, um programa capaz de identificar *clusters* de genes associados a metabólitos secundários, comparando-os entre diferentes organismos. Com base nas características estruturais, *E. nigrum* apresenta grande similaridade com outras espécies relacionadas de fungos filamentosos, incluindo fungos de estilo de vida distinto. O fungo e *E. nigrum* também de apresentou maior proximidade filogenética com *Didymella exigua*, um fitopatógeno com perfil de *clusters* de genes relacionados a metabólitos secundários distinto, demonstrando alta variabilidade de desses *clusters* de genes apesar da proximidade. Finalmente foram identificados três *clusters* de genes relatados como envolvidos na ocorrência de doenças que estão presentes em fungos fitopatogênicos e ausentes em *E. nigrum*. Dados da literatura corroboram a importância deste gene no caráter patogênico de alguns fungos. Assim, estes resultados sugerem que as diferenças observadas entre os estilos de vida de endófitos e fitopatógenos pode estar relacionada a um pequeno número de genes.

Palavras-chave: *Epicoccum nigrum*. Bioinformática. metabólitos secundários. Anotação. Genoma. Transcriptoma.

ABSTRACT

FERREIRA, A. J. **Analysis and annotation of *Epicoccum nigrum* genome and secondary metabolism**, 2016. 135 p. Ph.D. Thesis (Microbiology) – Instituto de Ciências Biomédicas, Universidade de São Paulo, São Paulo, 2016.

In recent years, there has been a great interest in endophytic fungi as a prolific source of new species because of its potential use in biological control and production of metabolites with different biological activities. Despite the increasing discovery of new compounds from endophytic fungi, there is a lack of knowledge on biosynthetic pathways responsible for the production of bioactive metabolites. In this context, *Epicoccum nigrum* is an endophytic fungus that has been used for plant pathogen biocontrol in different host plants, since it produces a series of secondary metabolites of biotechnological interest, including antimicrobials. In this regard, we have previously determined the *E. nigrum* isolate P16 produces secondary metabolites with antimicrobial activity and this production have been confirmed in this work by specific assays. An assembled 454 sequencing dataset was used to perform a comprehensive functional annotation using the EGene2 platform, including some specific components developed in this work. We found 10,320 protein coding genes as well as tRNAs, rRNAs and ncRNAs. The predicted proteins were compared to proteomes of other fungi and classified into orthologous groups, as well as used in phylogenetic analyses. In addition, we developed Synteny Clusters, a program that compares gene *clusters* associated to secondary metabolite biosynthesis. Based on structural features, *E. nigrum* presents a genome very similar to closely related filamentous fungi, including some fungi with distinct lifestyles. In addition, the analysis showed close phylogenetic relationship between *E. nigrum* and *Didymella exigua*, a phytopathogenic fungus. This phylogenetic proximity is apparently not directly related to the metabolic profile, which can be highly variable among fungi with similar lifestyles despite the proximity. Finally, we identified three disease-related gene *clusters* that are absent in *E. nigrum* and present in most plant pathogenic fungi. Data from the literature seems to corroborate the importance of some of these genes in the pathogenic character of some fungi. This result suggests that the lifestyle differences observed between endophytic and pathogenic fungi may rely on a relatively low number of genes.

Keywords: *Epicoccum nigrum*. Bioinformatics. Secondary metabolites. Annotation, Genome. Transcriptome.

LISTA DE FIGURAS

	página
Figura 1 - Etapas de módulos da plataforma EGene2 para a anotação de <i>E. nigrum</i>	45.
Figura 2 - Teste de produção de antimicrobiano com 7 e 60 dias.....	47.
Figura 3 - Teste de produção de antimicrobiano com 10, 20 e 30 dias.....	50.
Figura 4 - Anotações de genes exibidas em Artemis.....	54.
Figura 5 - Anotação de genes em relatório em formato <i>feature table</i> gerado pela plataforma EGene.....	54.
Figura 6 - Interface Synteny Clusters e passos para executar a análise.....	60.
Figura 7 - Página em formato web exibindo resultados de Synteny Clusters.....	63.
Figura 8 - Número de sequências de acordo com predição.....	67.
Figura 9 - Tamanho médio (pares de bases) de sequências de acordo com Predição.....	68.
Figura 10 - Proteínas com resultados positivos na anotação em relação ao total.....	72.
Figura 11 - Proporção da anotação em relação ao tamanho do genoma em pares de bases....	74.
Figura 12 - Categorias gerais das proteínas de <i>E. nigrum</i> por eggNOG.....	76.
Figura 13 - Categorias de proteínas de <i>E. nigrum</i> por eggNOG.....	77.
Figura 14 - Distribuição de proteínas transmembranares.....	81.
Figura 15 - Filogenia e divergência dos tempos estimados de Dikarya com base em sequência de sete genes codificadores de proteínas.....	83.
Figura 16 - Alterações no número de famílias de proteínas visualizadas na árvore filogenética.....	88.
Figura 17 - Estrutura molecular presuntivade compostos secundários produzidos por <i>E. nigrum</i> , baseado na análise do genoma pelo programa AntiSmash.....	93.
Figura 18 - Estrutura parcial de metabólitos secundários envolvidos em processos fitopatogênicos.....	98.
Figura 19 - Dendrogramas com clusters compartilhados entre diferentes espécies.....	99.
Figura 20 - Diagramas de Venn com clusters compartilhados entre diferentes espécies.....	100.

LISTA DE TABELAS

	página
Tabela 1 - Características do sequenciamento do genoma do isolado <i>E. nigrum</i> P16.....	36.
Tabela 2 - Espécies de <i>Colletotrichum</i> usadas para análise de Synteny Clusters.....	40.
Tabela 3 - Dothideomycetes com proteomas utilizados para comparação de homologia no pipeline do MAKER.....	41.
Tabela 4 - Fases de anotação e componentes utilizados.....	43.
Tabela 5 - Características do sequenciamento do transcriptoma do isolado <i>E. nigrum</i> P16...	51.
Tabela 6 - Componentes utilizados para teste de integração.....	56.
Tabela 7 - Testes de validação dos componentes para EGene2.....	57.
Tabela 8 - Programas e bancos de dados utilizados no pipeline.....	58.
Tabela 9 - Teste de acurácia de conjunto de treinamento.....	65.
Tabela 10 - Estimativa de diferentes preditores e evidências com diferentes conjuntos de treinamentos.....	69.
Tabela 11 - Índices de AED indicando a similaridade entre proteínas de Dothideomycetes e os genes inteiros preditos a partir de MAKER combinando Augustus com SNAP.....	70.
Tabela 12 - Proporção de categorias de proteínas e tamanho do genoma de <i>E. nigrum</i>	75.
Tabela 13 - Número e proporção de proteínas do genoma de <i>E. nigrum</i> dedicada a diferentes funções.....	78.
Tabela 14 - Diversidade de genes de RNAs transportadores em <i>E. nigrum</i>	80.
Tabela 15 - Identificação dos sete genes melhores genes usados para filogenia com a descrição dos seus principais domínios.....	82.
Tabela 16 - Proteomas de espécies de fungos utilizados para a análise de MCL.....	85.
Tabela 17 - Alteração no número de famílias de proteínas de acordo com os nós da árvore filogenética.....	89.
Tabela 18 - Clusters de genes de metabólitos secundários presentes em <i>E. nigrum</i>	91.
Tabela 19 - Identidade de genes pertencentes à categoria Outros de clusters de metabólitos secundários de <i>E. nigrum</i> de acordo com Antismash.....	92.
Tabela 20 - Clusters compartilhados por mais espécies de fungos.....	94.
Tabela 21 - Similaridade às proteínas com relação de patogenicidade do banco de dados PHI-base.....	97.

LISTA DE SIGLAS E ABREVIATURAS

GO - Gene Ontology

snRNA - small RNA

snoRNA - small nucleolar RNA

snRNA - RNAs nucleares

siRNA - small interfering RNA

exRNA - extracellular RNA

MCL - Markov Cluster

SUMÁRIO

	página
1 INTRODUÇÃO.....	17.
2 REVISÃO DE LITERATURA.....	18.
2.1 Importância do estudo do metabolismo secundário de fungos filamentosos.....	18.
<i>2.1.1 Principais classes de metabólitos secundários em fungos filamentosos.....</i>	<i>19.</i>
<i>2.1.2 Metabolismo secundário de fungos endofíticos.....</i>	<i>20.</i>
2.2 O fungo <i>Epicoccum nigrum</i>.....	21.
2.3 Abordagens moleculares para o estudo do metabolismo secundário de fungos.....	21.
2.4 Análise de Transcriptoma.....	23.
2.5 Plataforma EGene.....	24.
2.6 Predição de Genes.....	25.
2.7 O Programa Augustus.....	27.
2.8 Conjunto de Treinamento.....	27.
2.9 Regiões UTR.....	28.
2.10 RepeatMasker.....	28.
2.11 O pacote MAKER.....	29.
2.12 O programa SNAP.....	30.
2.13 Inferência de divergência na ancestralidade.....	30.
2.14 Estratégia MCL.....	31.
2.15 Protocolo MIRLO.....	31.
2.16 Calibração de datas e árvores filogenéticas.....	32.
2.17 Efetores e Predição.....	33.
3 OBJETIVOS.....	35.
3.1 Objetivos gerais.....	35.
3.2 Objetivos específicos.....	35.
4 MATERIAIS E MÉTODOS.....	36.

4.1 Sequenciamento do genoma de <i>E. nigrum</i>	36.
4.2 Condições de cultivo de <i>E. nigrum</i> para obtenção de extratos e RNA mensageiro....	36.
4.3 Obtenção do mRNA e extratos brutos da cultura.....	37.
4.4 Fracionamento para obtenção dos metabólitos da cultura.....	37.
4.5 Testes de eficiência dos antimicrobianos.....	37.
4.6 Sequenciamento do transcriptoma de <i>E. nigrum</i>	38.
4.7 Análise do transcriptoma de <i>E. nigrum</i>	38.
4.8 Desenvolvimento de ferramentas de Bioinformática.....	38.
4.9 Testes e validação das ferramentas de Bioinformática.....	39.
4.9.1 Testes e validação de Componentes.....	39.
4.9.2 Testes e validação de Synteny Clusters.....	39.
4.10 Conjunto de treinamento.....	40.
4.11 MAKER.....	40.
4.12 Anotação do genoma de <i>E. nigrum</i>	41.
4.2 Relação Filogenética de famílias de genes de <i>E. nigrum</i>	46.
4.2.1 Escolha de genomas.....	46.
4.2.2 Protocolo MIRLO.....	46.
4.2.3 Calibração das Árvores.....	46.
4.2.4 Avaliação de modificações nas famílias de genes.....	47.
4.3 Avaliação de clusters de metabólitos secundários.....	47.
5 RESULTADOS.....	48.
5.1 Extração do RNA mensageiro.....	48.
5.2 Obtenção dos extratos e eficiência dos antimicrobianos.....	48.
5.3 Sequenciamento do transcriptoma de <i>E. nigrum</i>	51.
5.4 Análise do transcriptoma de <i>E. nigrum</i>	52.
5.5 Desenvolvimento de ferramentas de Bioinformática.....	52.
5.5.1 Desenvolvimento de scripts de manipulação de dados.....	53.

5.5.2 Componentes de anotação plataforma EGene2.....	53.
5.5.3 Synteny Clusters.....	54.
<u>5.5.3.1 Execução.....</u>	54.
<u>5.5.3.2 Funcionamento.....</u>	61.
<u>5.5.3.3 Interpretação do relatório.....</u>	61.
<u>5.5.3.4 Validação do Synteny Clusters.....</u>	64.
5.6 Relações filogenéticas.....	64.
5.6.1 Conjunto de treinamento.....	64.
5.6.2 Predição de genes.....	65.
5.7 Anotação genoma <i>E. nigrum</i>	70.
5.7.1 Características Gerais.....	70.
5.7.2 Fração do genoma de <i>E. nigrum</i> dedicada a diferentes funções.....	73.
5.7.3 Proporção de proteínas do genoma de <i>E. nigrum</i> dedicada a diferentes funções.....	76.
5.7.4 RNAs não codificantes.....	79.
5.7.5 Domínios transmembranares e peptídeos sinal.....	80.
5.8 Relações filogenéticas.....	81.
5.8.1 Determinação das relações filogenéticas de <i>E. nigrum</i>	81.
5.8.2 Alterações nas famílias de proteínas.....	87.
5.9 Clusters de genes de metabólitos secundários.....	89.
5.9.1 Cluster de genes de metabólitos secundários de <i>E. nigrum</i>	89.
5.9.2 Comparação entre clusters de metabólitos secundários.....	93.
6 DISCUSSÃO.....	101.
7 CONSIDERAÇÕES FINAIS.....	112.
8 CONCLUSÕES.....	114.
REFERÊNCIAS.....	115.

1 INTRODUÇÃO

A tendência global para que a atividade agrícola e a produção de alimentos sejam realizadas de maneira sustentável exige a introdução de novos modos de controle de pragas e patógenos. Por questões ambientais e de segurança para a população humana, muitos agroquímicos sintéticos têm sido removidos do mercado e a legislação que controla a sua liberação tem se tornado mais severa, criando a necessidade de soluções alternativas de controle de pragas e doenças de plantas. Na área médica, a demanda por novos fármacos é crescente em consequência do surgimento de bactérias multirresistentes e da necessidade de antimicóticos mais eficientes. Isso se deve ao maior número de infecções fúngicas como resultado do aumento de casos de doenças emergentes e do número de pacientes transplantados, os quais fazem uso de medicamentos que deprimem o sistema imunológico. Além disso, há a necessidade do controle de doenças negligenciadas como malária, leishmaniose, tripanossomíase e filariose. Nesse contexto, metabólitos produzidos por micro-organismos são uma fonte importante de compostos para uso humano e na agricultura, seja para uso direto ou servindo como modelo para a síntese parcial ou completa de substâncias bioativas.

Nos últimos anos algumas empresas farmacêuticas têm diminuído ou extinguido suas pesquisas sobre produtos naturais. Entretanto, a análise genômica demonstra que o potencial biossintético de bactérias e fungos é bastante superior ao número de compostos conhecidos de qualquer espécie em particular. Isso se deve ao fato de que a maioria dos metabólitos não é produzida em condições padrões de fermentação. Para acessar estes compostos crípticos estão sendo ampliadas as estratégias na busca de evidências dos mesmo em meio a dados genômicos. Tal abordagem tem levado à descoberta de metabólitos bioativos novos, mesmo para espécies bastante estudadas do ponto de vista genético e bioquímico como *Aspergillus nidulans* e *Streptomyces* spp.

Os fungos são conhecidos como produtores de antibióticos, agentes imunossupressores e redutores de colesterol. Estimativas demonstram que uma pequena parcela das espécies de fungos foi descrita. Ademais, poucos destes fungos descritos foram efetivamente investigados quanto à produção de metabólitos bioativos. Desse modo, somente uma pequena parcela dos metabólitos economicamente importantes de fungos tem sido descoberta. Neste contexto, fungos endofíticos habitam o interior das plantas sem causar prejuízos e são encontrados em todas as espécies vegetais. Esses fungos têm sido considerados uma fonte prolífica de novos produtos naturais. De fato, muitos compostos bioativos de plantas (por exemplo, taxol,

camptotecina, vinblastina e hipericina) são produzidos por fungos que colonizam o interior de tecidos vegetais. Dessa forma, esta comunidade microbiana do interior da planta hospedeira tem sido considerada com uma importante fonte de micro-organismos de espécies ainda não descritas e consequentemente novos metabólitos.

2 REVISÃO DE LITERATURA

2.1 Importância do estudo do metabolismo secundário de fungos filamentosos

Os fungos são uma fonte importante de moléculas que podem servir como modelo para síntese de compostos para uso humano (HARVEY, 2008) e na agricultura (ANKE; THINES, 2007). Alguns metabólitos secundários de fungos podem ser tóxicos, mas outros são benéficos, como o antibiótico penicilina, o agente redutor de colesterol lovastatina (KELLER et al., 2005) e as ciclosporinas (BRAKHAGE, 2013). A importância dos fungos como fonte de produtos naturais é ilustrada pelo fato de que 6 dos 20 fármacos mais vendidos em 1995 foram de origem fúngica (SEYMOUR et al., 2004).

Estudos de ecologia química visando entender as funções de pequenas moléculas nas interações entre organismos estão no centro das descobertas de novos compostos (CAPORALE, 1995), e a aplicação de conceitos de ecologia de fungos para descoberta de metabólitos tem sido uma estratégia bastante efetiva (GLOER, 1997). É pertinente mencionar que somente gêneros tais como *Aspergillus* e *Penicillium* têm sido rigorosamente investigados quanto à produção de compostos bioativos. Conforme revisado por Bérdy (2005), de 6.450 metabólitos bioativos de microfungos, mais de 30% foram obtidos a partir destes gêneros, evidenciando a necessidade de estudos com outros gêneros de fungos, bem como com genótipos divergentes de espécies já estudadas. Um aspecto importante é que somente uma pequena parcela do número estimado de espécies de fungos (1,5 milhões) é conhecida (HAWKSWORTH, 2004) e poucas têm sido avaliadas para produção de compostos bioativos. Uma estratégia para a inferência de compostos secundários é a mineração de genomas, sendo detectada a produção de diferentes compostos identificada pela presença de *clusters* de genes destes compostos (PUSZTAHELYI, 2015).

2.1.1 Principais classes de metabólitos secundários em fungos filamentosos

Os metabólitos secundários de fungos são derivados de quatro classes de enzimas: poliketídeosintases (PKS), peptídeo sintases não ribossômicas (NRPS), terpeno sintases (TS) e dimetilalil difosfato triptofanosintases (DMATS). Cada classe utiliza metabólitos simples como substrato, os quais são rearranjados ou condensados em moléculas mais complexas, como poliketídeos, peptídeos não-ribossômicos, terpenos e alcalóides. Os substratos para as PKS e NRPS são moléculas de acil coenzima A e aminoácidos; os substratos da TS são unidades de dimetilalil difosfato (ou isoprenos), e os alcalóides são produzidos a partir do triptofano. Estas moléculas podem sofrer modificações como oxigenação, ciclização, isomerização e/ou condensação com outros compostos formando metabólitos altamente variáveis quanto à estrutura química e atividade biológica (KELLER et al., 2005).

Os terpenos de fungos incluem fitotoxinas, carotenóides, giberelinas, micotoxinas e indol terpenos, estes últimos sendo conhecidos por seus efeitos tóxicos. (HOFFMEISTER; KELLER, 2007). Já os alcalóides do ergot são tóxicos para animais, mas também são importantes agentes terapêuticos, e sua biossíntese é bem conhecida a nível molecular. Eles incluem as amidas do ácido lisérgico, famosas pelos efeitos alucinógenos e produzidas pela família Clavicipitaceae, e clavinas, produzidas por Trichocomaceae (HOFFMEISTER; KELLER, 2007).

Os poliketídeos são produzidos pelas PKs e são agrupados em três tipos de acordo com sua estrutura e modo de ação (COX, 2007). Em fungos, as PKs (tipo I) são multifuncionais e codificadas por um único gene, podendo apresentar até oito domínios. As PKs (tipo II) de bactérias e plantas são complexos multienzimáticos onde cada domínio é representado por uma cadeia polipeptídica separada. Enzimas do tipo III (chalcona-sintases) são constituídas de uma cadeia polipeptídica, ocorrem em plantas e sintetizam flavonóides. Recentemente foram descritas em bactérias e fungos, ampliando ainda mais o potencial metabólico microbiano (HERTWECK, 2009). Os peptídeos não ribossômicos são produzidos pelas NRPSs. Em fungos as NRPSs produzem penicilinas, cefalosporinas, ciclosporinas, fitotoxinas e sideróforos (HOFFMEISTER; KELLER, 2007). Comparações filogenômicas agruparam as NRPS em dois clados contendo diferentes famílias de NRPS (BUSHLEY; TURGEON, 2010).

Uma característica dos genes do metabolismo secundário é a tendência de serem agrupados (*gene clusters*) e localizados em regiões próximas dos telômeros. A natureza modular das PKs e NRPSs facilita a identificação desses genes (TOYOMASU et al., 2009).

Comparações genômicas mostram que genes biossintéticos são mais abundantes em Euascomicota do que em Basidiomicota, e são escassos em Chitridiomicota, Zigomicota, Schizosaccharomicota e Hemiascomicota (BUSHLEY; TURGEON, 2010; COLLEMARE et al., 2008).

2.1.2 *Metabolismo secundário de fungos endofíticos*

Fungos endofíticos são definidos como aqueles que podem ou não crescer *in vitro* e que habitam o interior da planta sem causar prejuízos e sem produzir estruturas externas (AZEVEDO; ARAÚJO, 2007). A relação endófito-planta é designada como um antagonismo balanceado que é finamente regulado e dependente das respostas de defesa da planta, da demanda por nutrientes do fungo, e das condições ambientais que permeiam a interação (SCHULZ; BOYLE, 2005). Os efeitos bioprotetores conferidos por estes organismos, tais como promoção de crescimento, tolerância a herbivoria e a diferentes estresses bióticos e abióticos são bem conhecidos para os endófitos clavicipitáceos, os quais colonizam algumas plantas (subfamília Pooideae) de clima temperado (KULDAU; BACON, 2008).

O fungo clavicipitáceo *Epichloefestuae* (Hypocreales: Ascomycota) é o único endófito com genoma sequenciado até o momento (<http://www.genome.ou.edu/fungi.html>). Para os demais endófitos, pouco é conhecido sobre a função de metabólitos na interação e se os compostos que são produzidos *in vitro* são também produzidos *in planta*, ou se conferem vantagem direta à planta (SCHULZ; BOYLE, 2005). Dentre os metabólitos novos está a descoberta por Stierle et al., (1993) do antitumoral diterpenóidetaxol produzido pelo endófito *Taxomyces andreanea*.

Os metabólitos de fungos endofíticos têm sido avaliados quanto à atividade antimicrobiana, antioxidante, antitumoral, antidiabetes, imunossupressora, antiolesterolêmica, herbicida, contra fitopatógenos e contra agentes de doenças tropicais como *Plasmodium*, *Leishmania* e *Trypanosoma* (BACKMAN; SIKORA, 2008; CAMPOS et al., 2008). Promoção do crescimento vegetal, tolerância a estresses bióticos e abióticos e biotransformação de compostos químicos também têm sido relatadas em fungos endofíticos (BORGES et al., 2009; RODRIGUEZ; REDMAN, 2008), mas embora se conheça muitos metabólitos destes micro-organismos, muitos outros são ainda desconhecidos, por desconhecimento da sua via de biossíntese e/ou condição de síntese.

2.2 O fungo *Epicoccum nigrum*

Epicoccum purpurascens Ehrenb. exSchlecht. (syn. *E. nigrum* Link) é um fungo cosmopolita encontrado em solos e em vários habitats onde atua como decompositor de tecidos vegetais. Este fungo é habitante comum da comunidade epifítica, mas pode exibir um estilo de vida endofítico, sendo considerado um endófito ubíquo comumente isolado de várias espécies de plantas (ARNOLD, 2007; SCHULZ; BOYLE, 2005). No entanto, seus conídios são encontrados no ar e causam desordens respiratórias em humanos (KUKREJA et al., 2009). *E. nigrum* é um ascomiceto mitospórico (Dotideomicetos). Cerca de 70 espécies de *Epicoccum* já foram descritas, mas a maioria foi reduzida à sinonímia. De acordo com Hawksworth et al., (1995) duas espécies são aceitas, *E. nigrum* e *E. andropogonis*. Recentemente, por meio de análise polifásica da diversidade intraespecífica de isolados endofíticos, foi proposta a divisão de *E. nigrum* em duas espécies (FÁVARO, 2009).

A utilização de *E. nigrum* no biocontrole de fitopatógenos é bem conhecida, além disso, o endófito é capaz de prover o crescimento do sistema radicular (FÁVARO et al., 2012). Este fungo controla *Monilinia* spp. em pêssegos e nectarinas (DE CAL et al., 2009), *Sclerotinia sclerotiorum* em girassol (PIECKENSTAIN et al., 2001) e *Pythium* em algodão (HASHEM; ALI, 2004). Compostos antimicrobianos, antioxidantes, antiretrovirais, antitumorais, inibidores de protease e telomerase e sideróforos têm sido caracterizados em *Epicoccum* (FÁVARO, 2009; FÁVARO et al., 2012). Metabólitos com aplicação biotecnológica incluem vários pigmentos para a indústria de corantes alimentícios (MAPARI et al., 2010) e o composto fluorescente epicocconona (BELL; KARUSO, 2003) o qual é um novo fluoróforo para coloração de células e detecção de proteínas em géis de eletroforese (disponível como LavaCellTM - ActiveMotif e DeepPurpleTM Total ProteinStain - GE Healthcare). Além disso, capacidade de biotransformação de compostos (ANDRADE et al., 2004) e de produção de beta-glicanas (SCHMID et al., 2006).

2.3 Abordagens moleculares para o estudo do metabolismo secundário de fungos

A descoberta de produtos naturais microbianos envolve a coleta e o cultivo das linhagens, extração, isolamento guiado por bioensaio e elucidação da estrutura química. No entanto, esta abordagem pode ser pouco produtiva porque resulta em uma alta taxa de detecção de compostos já conhecidos. Além disso, a produção de metabólitos por fungos é influenciada pelo meio de cultura, período de cultivo, pH, luminosidade, temperatura e

aeração (SCHERLACH; HERTWECK, 2009). Para obter maior diversidade de compostos e evitar a redescoberta de moléculas, o uso de fermentação em estado sólido, co-cultivo com micro-organismos que habitam o mesmo nicho ou extratos de plantas hospedeiras tem sido relatadas (SCHERLACH; HERTWECK, 2009). De fato, a estratégia denominada OSMAC (One Strain MAny Compounds) tem revelado a diversidade química e mostrado que, dependendo das condições de cultivo, micro-organismos produzem uma variedade de produtos naturais diferentes (BODE et al., 2002).

Com a disponibilidade de genomas microbianos, genes biossintéticos podem ser localizados pelo uso de ferramentas de bioinformática. Esta informação pode ser explorada na pesquisa de novas moléculas, abordagem denominada *genome mining* (ZERIKLY; CHALLIS, 2009). Em *A. nidulans* a deleção e a super expressão do gene *LaeA*, que codifica uma proteína reguladora, e a comparação do transcriptoma e do metaboloma entre a linhagem selvagem e mutante, levou à descoberta do novo antitumoral terrequinona A (BOK et al., 2006). A descoberta de *LaeA* em *Aspergillus* spp. tem levado à investigação de ortólogos em outros fungos, contribuindo para o entendimento da regulação epigenética do metabolismo secundário (CICHEWICZ, 2010). Abordagens de *genome mining* baseadas em inativação gênica ou expressão heteróloga só são viáveis quando o metabólito é constitutivamente (ou sob certas condições) produzido pela linhagem (CHIANG et al., 2008). Por exemplo, Chiang et al., (2007) descobriram no genoma de *A. nidulans* um *cluster* híbrido (PKS-NRPS), porém nenhum metabólito foi detectado mesmo após fermentação em diversas condições, demonstrando a natureza críptica do *cluster*. Um fator de transcrição dentro do *cluster* foi expresso sob controle do promotor álcool desidrogenase de *A. nidulans*. Sob condições indutoras, o *cluster* foi expresso e produziu novas aspiridonas. Esta abordagem é considerada promissora para descoberta de novos fármacos em fungos.

Em outro caso, Schroeckh et al., (2009) co-cultivaram *A. nidulans* com actinomicetos que compartilham o mesmo habitat e verificaram que uma linhagem de *Streptomyceshygroscopicus* induziu a expressão de *clusters* crípticos de *A. nidulans*, analisados por microarranjos. A expressão só foi verificada quando ocorreu interação física entre as células bacterianas e as hifas do fungo, e não devido a compostos difusíveis oriundos da bactéria. Foram descobertos metabólitos até então somente descritos em líquens e compostos novos com atividade antiosteoporose. Estes resultados contribuem para o entendimento da comunicação química entre micro-organismos de diferentes domínios.

A descoberta da atividade de metiltransferase de histonas de *LaeA* em *Aspergillus* spp. tem revelado a importância de mecanismos epigenéticos na regulação do metabolismo de

fungos. Tanto a inativação de genes de histona deacetilases em *A. nidulans* quanto o tratamento de *A. alternata* e *P. expansum* com inibidores destas enzimas resultaram na superprodução de metabólitos secundários (SHWABET al. 2007). A manipulação epigenética do metaboloma de fungos, por meio do tratamento com inibidores de DNA metiltransferases e histona deacetilases, tem sido considerada uma estratégia promissora para a descoberta de novas moléculas em espécies pouco estudadas geneticamente e sua incorporação em programas de *screening* de metabólitos de fungos (CICHEWICZ, 2010).

Estes exemplos ilustram estratégias baseadas em *genome mining* utilizadas para o estudo do metabolismo secundário de fungos e serão implementadas à medida que novas sequências genômicas tornam-se disponíveis (ZERIKLY; CHALLIS, 2009). Como visto anteriormente, mesmo para espécies bastante estudadas como *A. nidulans*, novos compostos têm sido descobertos. É importante que tais estratégias sejam aplicadas aos fungos endofíticos, valorizando-os como fonte de novas moléculas. Em todo caso, na ausência de sequências genômicas, abordagens como clonagem, análise funcional e expressão heteróloga são úteis para descoberta de genes biossintéticos de fungos (SCHUMANN; HERTWECK, 2006).

O conhecimento dos genes biossintéticos e sua regulação podem fornecer informações importantes sobre as funções ecológicas de metabólitos secundários de fungos endofíticos. Além disso, investigações sobre os genes de metabólitos secundários e sua expressão fornecem a base para o melhoramento da produção e para engenharia de vias biossintéticas com o objetivo de sintetizar novas moléculas bioativas.

2.4 Análise de Transcriptoma

O avanço nas técnicas de análises de RNA permite utilizar e identificar os genes expressos nos organismos em resposta a determinadas situações às quais as células são submetidas. Além de permitir a caracterização da estrutura primária de proteínas, o emprego de programas (GenomeScan; GenScan - <http://genes.mit.edu/genomescan.html> ESTScan - ISELI et al., 1999) que utilizam transcriptomas também auxiliam na anotação de genomas, identificando os *exons* e *introns* (PROSDOCIMI et al., 2002). Esta abordagem também facilita construção de mapas físicos, facilitando a predição de sequências estruturais como telômeros, centrômeros, heterocromatina, satélites, além de sequências repetitivas como mini e microssatélites, *transposons*, dentre outras. A comparação também permite a detecção de sequências reguladoras como promotores e *enhancers*. Uma das maiores dificuldades no

processo de anotação de genes eucarióticos é a delimitação das fronteiras *exon-intron*. O mapeamento de leituras de um transcriptoma permite identificar de forma mais segura essas fronteiras, bem como a ocorrência de variantes de *splicing* (SANTOS; ORTEGA, 2003).

A utilização de transcriptomas facilita grandemente o processo de anotação genômica. A anotação genômica é a determinação da função de diferentes regiões do genoma. Assim, encontrar genes homólogos é o principal objetivo deste tipo de anotação. Mesmo encontrando a ORF capaz de codificar uma proteína, o gene ainda pode ser hipotético, havendo a necessidade de determinar sua função. Nessa fase muitos genes hipotéticos podem ser desconsiderados e outros antes despercebidos, podem revelar proteínas com papel funcional já determinado (SANTOS; ORTEGA, 2003). Além disso, os bancos de dados constituem importantes ferramentas para a anotação genômica (BORÉM; SANTOS, 2001), sendo o GenBank (<http://www.ncbi.nlm.nih.gov/genbank/>) a maior referência nos trabalhos com dados genômicos. Bancos como o SWISS-PROT (<http://expasy.org/sprot/>), Pfam (pfam.xfam.org/) e eggNOG (<http://eggNOGdb.embl.de/>) por exemplo, possuem informações de proteínas associadas às características como função, domínios funcionais, proteínas homólogas, dentre outras. Outro banco de dados como o KEGG (Kyoto Encyclopedia of Genes and Genomes - <http://www.genome.jp/kegg/>) disponibiliza informações e buscas em mapas metabólicos, hierarquia funcional, detecção de grupos ortólogos, reações bioquímicas, drogas e doenças relacionadas a vários genomas de procariotos e eucariotos (PROSDOCIMI et al., 2002).

2.5 Plataforma EGene

EGene é uma plataforma modular e genérica para a construção de *pipelines* de processamento e anotação de sequências biológicas. A plataforma utiliza um conjunto de componentes que realizam o processamento de forma independente ou acoplado a diferentes ferramentas de bioinformática amplamente utilizados pela comunidade científica. Seu caráter modular permite flexibilidade, ajustando às diferentes necessidades dos usuários de acordo com a complexidade do projeto e organismo. Outra característica é a simplicidade, desenvolvido para invocação e integração de vários programas em um único processo seriado sem a necessidade de conhecimento em programação. Entretanto, caso o usuário deseja integrar um novo programa à anotação, é possível criar outro componente e integrá-lo à plataforma graças ao código aberto.

O EGene foi desenvolvido em sua primeira versão (DURHAM et al., 2005) com o objetivo de processar sequências para pré-anotação. Entretanto, na sua segunda versão (EGene2), atualmente em desenvolvimento, a plataforma foi incrementada com vários novos componentes, bem como teve os antigos componentes atualizados, sendo aplicada em recentes trabalhos para anotação de genomas virais (ALVES et al., 2016; BRITO et al., 2012) transcriptoma de eucariotos (NOVAES et al., 2015) e de um genoma bacteriano (*Photorhabdus luminescens* MN7 – Prof. Dr. Carlos Eduardo Winter – não publicado).

A plataforma permite executar vários programas que se integram independentemente da linguagem na qual foram escritos e do ambiente de execução como, por exemplo, os formatos de entrada e saída de dados. A pré-anotação envolve etapas como avaliação de qualidade de sequência, filtragem de contaminantes, corte das extremidades, mascaramento de elementos transponíveis e vetores. No processo de anotação, EGene contribui em todas as etapas, primeiramente na predição de genes codificadores de proteínas, tRNAs, rRNAs, ncRNAs, elementos transferidos horizontalmente ou transponíveis, repetições seriadas, pontos de clivagem peptídicas, dentre outros. Após a determinação da natureza das sequências, o EGene dispõe de módulos para caracterizá-la por meio da busca de regiões similares em outras sequências biológicas, características funcionais, presença de domínios protéicos, atuação em vias metabólicas, localização subcelular ou extracelular, presença de regiões de clivagem ou peptídeos sinal, dentre outros. Todo o processo gera grande quantidade de informação, o que muitas vezes pode prejudicar o usuário na interpretação correta dos resultados. Dessa forma, visando melhorar o desempenho do projeto, facilitar o acesso e compartilhamento dessas informações, o EGene conta com módulos para exibição dos resultados. Dentre os módulos estão vários tipos de relatórios, desde individuais para cada módulo até os globais, onde há uma reunião das informações relevantes obtidas por cada componente em relatórios (*features table*/GFF3) e geração de sites *web*.

2.6 Predição de Genes

A anotação de um genoma pode ter profundidade variada de acordo com as etapas adotadas no processo. Apesar de requer várias fases, dois fatores são fundamentais: a qualidade da montagem/mapeamento das sequências (1) e a qualidade da predição dos genes (2). A predição, neste contexto, é uma estimativa da estrutura gênica, a qual pode também variar quanto ao seu detalhamento. Dentre as estruturas gênicas básicas estão os *exons*, códon de iniciação (start) e terminação (stop). Entretanto, os genes, principalmente os de eucariotos,

possuem estrutura gênica com maior complexidade, sendo seus *exons* compostos por UTR5', UTR3' e CDSs. Além disso, em genes compostos por mais de um *exon* existem os *introns*, estruturas não consideradas codificantes. Mais à margem dos genes existem elementos responsáveis pela modulação de sua transcrição como os promotores. Todos estes elementos estruturais são identificados a partir de sequências específicas em meio ao genoma. Apesar de alguns grupos de bases serem mais fáceis de identificar como os códon de iniciação e terminação, muitas vezes em uma análise mais criteriosa é possível encontrar muitos resultados considerados falso positivos, uma vez que os genes estão repletos de, por exemplo, trincas de bases codificantes para metionina (códon de iniciação), pois além de ser um códon de iniciação, o *codon* ATG, pode estar presente tanto em meio ao gene como fora dele, não sendo portanto, considerado o codon de iniciação. O mesmo ocorre com outros códigos de partes estruturais de genes. Dessa forma, não adianta determinar os códigos individualmente sem levar em consideração as outras estruturas que compõe um gene. Assim, a predição gênica não necessita apenas de determinados códigos genéticos das estruturas, mas também a relação e frequência entre os elementos. Para estimar a possibilidade de determinado trecho de sequência ser capaz de codificar uma proteína ou função biológica os preditores gênicos necessitam de algoritmos que tenham um bom suporte estatístico. Os resultados dessas estimativas podem ser validados, mas, a análise experimental é custosa, por isso desde a década de 1980 tem ocorrido a expansão de programas de computadores capazes de realizar predições gênicas (ALLEN, et al. 2004; KORF, 2004). Atualmente, destacam-se os programas de predição gênica Augustus (STANKE; WAACK, 2003), FgeneSH (SALAMOV; SOLOVYEV, 2000), Genscan (BURGE; KARLIN, 1997), GlimmerHMM (MAJOROS et al., 2004), Genemark.hmm (LUKASHIN; BORODOVSKY, 1998), GeneId (GUIGO, 1998; PARRA et al., 2000), GeneMark.hmm-ES (TER-HOVHANNISYAN et al., 2008) e SNAP (KORF, 2004). Estas plataformas variam grandemente quanto à sensibilidade e especificidade, e são importantes na escolha do melhor método para obter resultados mais próximos da realidade, visto que apesar da ampla gama de preditores, os resultados não são completamente satisfatórios (GOODSWEN et al. 2012; STANKE; MORGENSTERN, 2005). Dentre os programas de predição de genes eucarióticos, destaca-se o Augustus pela maior frequência de acertos nas predições como demonstrado em trabalhos comparando diferentes preditores (CANTAREL et al., 2008; KORF, 2004; STANKE et al. 2008; STANKE; WAACK, 2003). A qualidade da predição pode ser verificada de duas formas: Utilização experimental utilizando sequências de RNA-seq para reconstrução do gene (1) e predições

completas suportadas por alinhamentos de alta qualidade de transcritos e proteínas (2) (HASS et al. 2011).

2.7 O Programa Augustus

Augustus é um programa de predição de genes eucarióticos que utiliza modelos ocultos de Markov para determinar a probabilidade de a sequência nucleotídica pertencer a uma estrutura gênica. O programa permite configurar parâmetros como posição específica do sítio de processamento (*splicing*), códon de iniciação (start) e terminação (stop), posição de *exons* e *introns* (STANKE; MORGENSTERN, 2005). O programa permite melhorar a acurácia na predição de genomas grandes, uma vez que eucariotos possuem *introns* grandes, o que dificulta a predição e cria vieses como falsos *exons* (STANKE; WAACK, 2003). Além disso, o Augustus tem a possibilidade de treinamento, esta capacidade é importante, pois permite modular os parâmetros estatísticos por meio de evidências. Estas evidências podem ser sequências expressas (EST) e resultados de sequências protéicas alinhadas satisfatoriamente com o banco de dados. Estes alinhamentos de sequências protéicas, assim como ESTs podem ser utilizados para treinar preditores de genes, mesmo em casos de ausência de informações a respeito do modelo gênico (YANDELL; ENCE, 2012). O preditor Augustus apresenta desempenho superior em relação aos preditores Genscan, GENIE e GENEID em *Drosophila* (KORF, 2004). Sua taxa de acerto é superior a 77% quando comparado com evidências experimentais como sequências de ESTs (STANKE et al., 2008).

2.8 Conjunto de Treinamento

O Conjunto de Treinamento possibilita ajustar os parâmetros estatísticos. Este ajuste permite maior acurácia, uma vez que os genes de eucariotos como fungos são diferentes estruturalmente de genes humanos, por exemplo. Os genes desses organismos podem variar estruturalmente em incidência de bases, distribuição dos códons, frequência de processamento, tamanho dos genes, *exons*, *introns*, além da distância dos promotores e etc. (HASS et al. 2011). O treinamento é a utilização de genes conhecidos estruturalmente, os mesmos são utilizados como métricas para os cálculos dos preditor. Os conjuntos de treinamento são mais eficientes à medida que se utiliza uma diversidade maior de genes de organismos mais relacionados filogeneticamente.

2.9 Regiões UTR

As regiões não traduzidas de RNA mensageiro, ou UTRs, são importantes para a expressão gênica. Essas regiões atuam na expressão dos genes após a tradução e sua regulação pode ocorrer por meio da modulação do transporte de mRNA do núcleo para o citoplasma, assim como na eficiência da tradução, localização subcelular e estabilidade do mRNA (GRILLO et al. 2010). Os preditores geralmente buscam motivos padrões, mas, nestas regiões devido à complexidade estrutural isso nem sempre é possível por meio de alinhamentos simples (DOYLE et al. 2008). Dessa forma, com o advento do sequenciamento de RNA, além de anotar as sequências do organismo de estudo, a anotação dessas sequências pode enriquecer os bancos de dados dessas regiões e auxiliar a expandir o conhecimento dessa área. Estes bancos de dados servem como fonte de motivos para treinar e avaliar programas de identificação de UTRs, melhorando o entendimento de suas variantes de *splicing*, além de caracterizar suas funções biológicas na modulação da expressão gênica. Dentre os bancos de dados para UTR estão UTRdb (GRILLO et al., 2010) e TUTR (DOYLE et al., 2008).

As regiões UTRs têm a capacidade de atuar na expressão devido à existência de elementos *cis* localizados nas extremidades das UTRs (3' e 5'), assim como sítios de interação com miRNA no UTR 3' (FLYNT; LAI, 2008). Estes sítios de ligação aos miRNAs são muito degenerados, tolerando ausência de complementaridade, *gaps*, pareamentos G-U, pareamentos deslocados de 6 a 8 nucleotídeos com a fita 5' dos miRNA alvo. Além disso, os sítios de interação podem estar na fita complementar, dificultando a predição dessas sequências como sítios para proteínas ligantes de RNA ou miRNA (FAGHIHI; WAHLESTEDT, 2009). Dessa forma, para identificação de uma UTR é importante a caracterização de seus domínios funcionais, os quais são sequências curtas de oligonucleotídeos que desempenham atividade biológica por meio da combinação de sua estrutura primária com uma secundária específica. Assim, estes sítios podem atuar reconhecendo fatores de ligação ao RNA ou mesmo à maquinaria de tradução.

2.10 RepeatMasker

O RepeatMasker faz uma busca na sequência de DNA por elementos repetitivos e de baixa complexidade. As sequências repetitivas são mascaradas com a substituição das letras representativas (A,T,G ou C) pela letra "N", "X" ou letras minúsculas. O mascaramento melhora a performance de vários programas de processamento e interpretação de sequências,

pois simplifica a busca. Este processo evita o processamento de sequências que poderiam gerar falsos positivos ou consumir tempo computacional sem necessidade. Este tempo pode ser reduzido grandemente em diversos genomas, uma vez que boa parte dos genomas pode conter elementos repetitivos. O genoma humano, por exemplo, possui aproximadamente 56% de elementos repetitivos identificáveis por RepeatMasker (TARAILO-GRAOVAC; CHEN, 2009).

2.11 O pacote MAKER

O MAKER é um pacote que permite integrar um conjunto de ferramentas para a predição e anotação de genes. A predição pode ser otimizada utilizando sequências de proteínas, sequências de genes de RNA codificantes ou não. A estratégia utiliza as métricas das sequências de proteínas e RNAs mapeando e determinando o melhor arranjo gênico possível (CAMPBELL et al., 2014). O MAKER foi inicialmente desenvolvido como uma ferramenta para anotação *de novo* de organismos. Entretanto essa ferramenta foi expandida para a anotação e curadoria de genomas (HOLT; YANDELL, 2011), mas atualmente o programa anota e mascara sequências repetitivas do genoma, alinhando sequências de proteínas e RNA. Esta abordagem permite detectar sítios de *splicing*, facilitando o processo de curadoria. O *pipeline* do MAKER simplifica a integração dos preditores Augustus e SNAP usando como evidências sequências de mRNA e proteínas, as quais são reunidas por alinhamentos (YANDELL; ENCE, 2012). O processo é iniciado com uma predição *ab initio* com preditores gênicos como Augustus (STANKE; WAACK, 2003), permitindo a comparação. Após esta predição inicial há uma otimização nas regiões de processamento e extremidade do gene utilizando o SNAP (KORF, 2004). A comparação dos diferentes métodos pode ser realizada por meio da métrica de qualidade AED (Annotation Edit Distance), a qual compara a precisão das sequências preditas com as proteínas candidatas homólogas e ESTs. O índice de AED abrange a faixa de 0 a 1, com 0 indicando plena similaridade entre as coordenadas de *exon-intron* da predição e o alinhamento das evidências de proteínas e ESTs (EILBECK et al., 2009).

2.12 O programa SNAP

A ausência de dados experimentais ou mesmo a pequena quantidade é um desafio para a qualidade dos modelos gênicos propostos por preditores. Vários preditores têm sido desenvolvidos com o objetivo de contornar este desafio. Entre os preditores, destaca-se o SNAP (KORF, 2004). Assim como Augustus e Genscan, SNAP utiliza modelos ocultos de Markov, mas seu desempenho sofre menos interferência dos conjuntos de treinamento. Assim, o SNAP é capaz de realizar predições *ab initio* com conjuntos de treinamento mais genéricos. Uma das características responsáveis pelo sucesso na predição utilizando conjuntos de treinamentos mais generalizados é seu ajuste aos parâmetros padrões, como as matrizes de tamanho das regiões de aceptores e doadores nas regiões de fronteira *exon/intron*, regiões adjacentes *start/stop* códons, número de possíveis combinações arquitetonicas. Estes padrões de arquitetura e tamanho de regiões variam de acordo com o organismo utilizado para treinar. O SNAP apresenta desempenho similar a Augustus e GenScan em *Drosophila melanogaster* (KORF, 2004).

2.13 Inferência de divergência na ancestralidade

A utilização de uma grande quantidade de genes para análises filogenéticas tem se tornado uma prática comum (DUNN et al. 2008; SCHIERWATER, et al. 2009). Entretanto, a inserção de genes com um sinal filogenético inconsistente pode levar a incongruências na topologia das árvores, o que pode explicar algumas das divergências frequentemente observadas entre estudos filogenéticos. Nestes casos, a exclusão de táxons divergentes pode ser uma solução, porém nem sempre satisfatória (TALAVERA; CASTRESANA, 2007). Assim, inferência de divergências em ancestrais requer genes com fortes sinais filogenéticos. Contudo a seleção de genes com esta característica constitui-se um desafio, sendo necessário rever individualmente e testar a sua história evolutiva. Técnicas tradicionais como o uso exclusivo de apenas concatenação não apresentam bons resultados, sendo necessário selecionar genes com fortes sinais filogenéticos, excluindo incongruências significativas dos nós para reconstruir as divergências antigas entre os grupos (SALICHOS; ROKAS, 2013). Esta seleção tem como objetivo evitar o risco de incongruências devido à inserção de genes parálogos na análise ou genes transferidos horizontalmente (SLOT; ROKAS, 2010). Estes genes também devem ser selecionados quanto ao alto valor de suporte para *bootstrap* consenso em seus nós, além de possuírem inter-nós com alto suporte. Estes cuidados

permitem uma melhora significativa da inferência devido à utilização de genes ou entre nós com alto suporte (SALICHOS; ROKAS, 2013).

A obtenção de um bom número de genes considerados candidatos ideais para as inferências filogenéticas necessita de uma triagem entre muitos genes. Dessa forma, novas técnicas como o MCL (Markov Cluster) (VAN DONGEN, 2000) facilitam o processo, permitindo analisar grandes quantidades de dados evitando erros de julgamento de pesquisadores quanto à origem dos genes.

2.14 Estratégia MCL

A expansão do sequenciamento de genomas permite realizar comparações mais extensas dos táxons, sendo possível buscar sinais filogenéticos em todo genoma. Estas novas fontes de informações ampliam a busca além dos tradicionais marcadores como 18S, ITS, beta-tubulina, etc. Por meio de abordagens que agrupam os genes de acordo com a similaridade das proteínas, e não sua função ou identidade, é possível determinar os prováveis genes homólogos. O agrupamento de genes homólogos com MCL permite observar seu compartilhamento entre os diferentes táxons, bem como os casos de diferenciação, duplicação ou perda de determinados genes ao longo da história evolutiva destes grupos. Este conjunto de informações pode ajudar a entender porque certos grupos são mais distantes e encontrar genes chave que mesmo em menor quantidade podem ser cruciais para diferenciação de grupos taxonômicos. Além disso, também podem ser determinadas as possíveis adaptações fisiológicas e estratégia para prosperar em ambientes e hospedeiros (EMMS; KELLY, 2015).

O programa MCL (<http://micans.org/mcl/>) realiza o agrupamento a partir do arquivo do resultado de BLASTx entre as proteínas de estudo. Estes resultados, quando submetidos ao algoritmo de Markov Cluster (MCL), são agrupados de acordo com a similaridade entre as proteínas e o valor de erro. Neste processo, os genes responsáveis pela codificação das proteínas são classificados como famílias de genes candidatos a homólogos (ENRIGHT et al., 2002).

2.15 Protocolo MIRLO

A seleção de genes com forte sinal filogenético e a construção de árvores para inferir a divergência de ancestrais a partir do genoma envolve várias fases. Para facilitar este processo, Thon (2015) desenvolveu o protocolo MIRLO, o qual é uma série de procedimentos que utiliza um conjunto de *scripts* e programas para análise filogenética com base nas sequências

protéicas dos organismos. Dentre os programas utilizados estão o MCL, ProtTest 3.4 (<https://code.google.com/p/prottest3/>), PhyML (GUINDON et al., 2005), MAFFT (KATO et al., 2002) e BioPython (http://biopython.org/wiki/Main_Page). O processo é robusto devido à comparação de todos os proteomas dos organismos analisados. A partir dos dados dos proteomas é possível identificar cópias simples ou múltiplas de determinadas famílias de genes que são utilizadas para inferir sinais filogenéticos. O protocolo possui várias fases, sendo as mais importantes a detecção das proteínas similares (BLASTx), o agrupamento em famílias (MCL), o alinhamento das sequências dessas famílias separadamente (MAFT), a construção de árvores filogenéticas (PhyML) e avaliação das melhores árvores (SH-like). O programa PhyML é capaz de estimar de forma rápida e acurada a filogenia a partir de sequências de DNA ou proteínas por máxima verossimilhança (GUINDON et al., 2005).

2.16 Calibração de datas e árvores filogenéticas

A determinação das datas nas quais ocorreram as divergências entre os ancestrais de fungos envolve a integração de diferentes informações de eventos como movimentos de placas tectônicas, paleontologia ou mesmo dados a respeito da evolução dos relógios moleculares de animais e plantas. Apesar de imperfeitos, os relógios moleculares podem fornecer informações importantes a respeito dos tempos geológicos. Estas informações a respeito da idade dos modelos moleculares dos organismos em filogenia são incertas quanto à precisão, divergindo de outros trabalhos por diferenças de centenas de milhões de anos (BERBEE; TAYLOR, 2010). Estas escalas de tempo são muito grandes, sendo proporcionais à divergência entre os nós de ancestrais. Táxons como Ascomycota e Basidiomycota divergiram há cerca de 1,2 bilhões de anos, tornando maiores as imprecisões métricas nas idades dos ancestrais (HECKMAN et al., 2001). Entretanto, estas divergências nas idades dos modelos são reduzidas na medida em que são descobertos novos fósseis de fungos e a origem dos táxons é revisada graças ao incremento de novas informações e parâmetros (BERBEE; TAYLOR, 2010). Infelizmente, há poucos fósseis de fungos disponíveis para se inferir a idade dos membros do grupo, sendo a mesma estimada em 1.600 Ma (milhões de anos). As principais hipóteses quanto à idade dos fungos são apoiadas a partir de medições construídas a partir do ponto de separação entre aves e mamíferos há cerca de 310 milhões de anos o que gera imprecisões (WANG et al., 1999).

O programa R8s (SANDERSON, 2015) é capaz de estimar taxas absolutas de evolução molecular e tempo de divergência em uma árvore filogenética. R8s emprega

métodos como máxima verossimilhança para calcular relógios moleculares globais ou locais, métodos semi-paramétricos e não paramétricos. Estes métodos aceitam um relaxamento no rigor do cálculo dos tempos de acordo com a atribuição de dados experimentais. O processo de nivelamento dos tempos é realizado a partir da árvore filogenética, na qual são inseridas possíveis evidências experimentais do tempo de um determinado nó da árvore (pontos de calibração). Estas evidências são mensuradas em anos e norteiam o número de substituições ao longo de cada ramo. A partir desta métrica pode ser realizado o calculado do número de substituições e extrapolado para outros nós. Podem ser adicionados a esta estimativa um ou mais pontos de calibração, dando mais eficiência à análise com a melhora do dimensionamento. O algoritmo de R8s pode aceitar como evidências para calibração idades fixas para o nó ou, o mais indicado, faixas de tempo com o número mínimo e o máximo de anos (SANDERSON, 2015).

2.17 Efetores e Predição

Os fungos patogênicos podem utilizar como estratégia de infecção as proteínas efetoras. Estas proteínas podem atuar no espaço extracelular (efetores apoplásticos) e/ou no citoplasma do hospedeiro (efetores citoplasmáticos). Funcionalmente atuam de forma avirulenta a tóxica, atacando a planta e causando morte celular. Entretanto, os efetores não atuam livremente, sendo reconhecidos pela defesa da planta. Este reconhecimento se dá pela coevolução dos organismos, consequentemente há uma imensa variação nas estruturas dessas proteínas a fim de burlar a defesa da planta que por sua vez devem evoluir concomitantemente para reconhecer as novas estruturas. Todo este processo evolutivo reflete nos estudos de proteínas efetoras, sendo iniciais como poucas proteínas caracterizadas até o momento. Além disso, essa baixa quantidade de proteínas efetoras fúngicas descritas limita a construção de modelos computacionais para o desenvolvimento de programas de predição (SPERSCHNEIDER et al., 2015).

No domínio bactéria existe um maior sucesso na predição em domínios de proteínas efetoras, principalmente as secretadas pelo sistema de secreção do tipo III (SPERSCHNEIDER et al., 2015). Na busca por efetores bacterianos, uma ferramenta útil é o EffectiveDB (<http://effectivedb.org/>). Os efetores não costumam possuir sequências de domínios com significante similaridade entre diferentes organismos, sendo encontrados muito diversificados, mesmo dentro da mesma espécie. Entretanto, existem casos nos quais são encontrados efetores similares mesmo entre espécies diferentes. O efector Slp1 (*Magnaporthe*

oryzae) e o efetor Ecp6 (*Cladosporium fulvum*) exibem semelhanças devido à presença do domínio LysM (BOLTON et al., 2008). Contudo, o baixo número de sequências semelhantes entre proteínas efetoras de diferentes organismos faz com que outras características, mais amplas, sejam usadas para a caracterização, tais como sinal de secreção, pequeno tamanho e riqueza em cisteína e pontes de sulfeto (STERGIOPOULOS; DE WIT, 2009). Este critério não é uma regra para todos os efetores, uma vez que nem todas as proteínas pequenas secretadas são efetoras. Além disso, também existem outras proteínas que mesmo sendo pobres em cisteína e com tamanho superior são efetoras (GOUT et al., 2006). Assim, além de tamanho, riqueza em cisteína e sinal de secreção são necessários outros critérios mais específicos quanto à funcionalidade comparando com outros presentes em fungos (SPERSCHNEIDER et al., 2015).

3 OBJETIVOS

3.1 Objetivos gerais

Realizar a anotação do genoma e identificar genes relacionados à biossíntese de metabólitos secundários do fungo endofítico *E. nigrum*.

3.2 Objetivos específicos

- Desenvolver módulos em Perl capazes de integrar à plataforma EGene2 e facilitar o processo de anotação dos dados de sequenciamento de *E. nigrum*, bem como de outros organismos;
- Desenvolver um conjunto de treinamento específico de *E. nigrum* para a predição de genes;
- Analisar o transcriptoma de *E. nigrum*;
- Melhorar a predição dos genes de *E. nigrum* utilizando dados do transcriptoma;
- Anotar o genoma de *E. nigrum*;
- Comparar o perfil metabólico dos fungos de *E. nigrum* em diferentes condições de cultivo;
- Identificar grupos de *clusters* de genes relacionados a compostos secundários de *E. nigrum* e sua relação com outros fungos;
- Estabelecer as relações filogenéticas de famílias de genes de *E. nigrum* e outros fungos.

4 MATERIAIS E MÉTODOS

4.1 Sequenciamento do genoma de *E. nigrum*

O DNA genômico de *E. nigrum* foi extraído com DNeasy® Plant Mini Kit (Qiagen 69.104, Qiagen Sciences, USA). O produto da extração de DNA genômico (50 µg de DNA genômico na concentração de 500 ng/uL com razão A260/280 igual a 1,8) foi enviado para a MACROGEN (Gasan-dong, Geumchun-gu, Seoul, Korea) onde foi sequenciado pelo método de pirosequenciamento 454 (Roche). A análise preliminar de bioinformática incluindo a análise de qualidade e montagem do genoma foram realizadas pela empresa contratada (tabela 1).

Tabela 1 - Características do sequenciamento do genoma do isolado *E. nigrum* P16.

Dados Primários	pb
Número de leituras	3.006.244
Número de bases	1.285.687.655
Número de leituras montados	2.930.138 (97%)
Número de <i>contigs</i>	670
Número total de bases de <i>contigs</i>	3.731.056
Tamanho médio dos <i>contigs</i>	51.837
N50 [*]	209.454
<i>Contig</i> de maior tamanho	648.904
Cobertura	32 vezes

^{*} N50: significa que metade de todas as bases residem em *contigs* desse tamanho ou maior.

4.2 Condições de cultivo de *E. nigrum* para obtenção de extratos e RNA mensageiro

O fungo *E. nigrum* foi cultivado em triplicata no meio BD (Batata Dextrose) em dois tempos diferentes com o objetivo de obter o micélio e metabólitos de colônias jovens (7 dias) e na fase de senescência (60 dias). A cultura foi realizada em volume de 500 mL em frascos de 1 litro, os quais foram mantidos a 28 °C sem agitação.

As amostras do fungo *E. nigrum* foram cultivados em triplicata no meio BD (Batata Dextrose) e em meio Sabouraud em três períodos diferentes como objetivo de obter o micélio em uma maior variação de condições de cultivo. As colônias foram mantidas por 10 (colônia

jovem), 20 (intermediária) e 30 dias (senescente) nos dois meios de cultura. As culturas foram realizadas em volume de 1 litro em frascos de 2 litros, os quais foram mantidos estaticamente a 28 °C.

4.3 Obtenção do mRNA e extratos brutos da cultura

Após o crescimento fúngico (7, 10, 20, 30 e 60 dias) o micélio das culturas foi separado do meio de cultura líquido por meio de filtração em papel filtro com o auxílio de bomba a vácuo. O meio de cultura (extrato bruto) foi separado para o fracionamento em solvente, enquanto o micélio obtido foi congelado imediatamente em nitrogênio líquido e estocados em freezer a -80 °C por até 30 dias. Posteriormente, o RNA mensageiro foi extraído utilizando o kit *Dynabeads® mRNA DIRECT™* (Invitrogen, USA) seguindo as recomendações do fabricante.

4.4 Fracionamento para obtenção dos metabólitos da cultura

Os extratos obtidos das culturas de 7 e 60 dias crescidas em meio BD e as culturas mantidas por 10, 20 e 30 dias em meio BD e Sabouraud foram fracionadas em solvente utilizando o funil de separação seguida pela concentração em evaporador rotativo. As culturas foram submetidas ao fracionamento no solvente nas proporções de 1:3, seguidas por 1:10 e novamente 1:10. Primeiramente foi realizado procedimento com hexano, seguido por butanol, acetato de etila e clorofórmio, respeitando uma ordem de polaridade crescente. Em seguida os solventes contendo os compostos secundários foram colocados no evaporador rotativo e o extrato bruto obtido foi estocado na concentração de 100 mg.mL⁻¹ a -80 °C.

4.5 Testes de eficiência dos antimicrobianos

Para a realização dos testes de produção de antimicrobianos foram utilizadas as bactérias *E. coli*, *S. aureus* e *Bacillus* sp. As mesmas foram cultivadas por 24 horas em TSB (Tryptic Soy Broth - Oxoid Thermo Scientific) e inoculadas superficialmente em placas com TSA (Tryptic Soy Agar - Oxoid Thermo Scientific). Em seguida, foram colocados discos de papel filtro esterilizado com 5mm de diâmetro, aos quais foram adicionados 10 µL de extrato do fungo. Em seguida, as culturas foram incubadas a 37 °C por 24 horas e após este período foi medido o halo de inibição das amostras. As análises foram realizadas em triplicata e como controle negativo, foi adicionado apenas o solvente no disco de papel.

4.6 Sequenciamento do transcriptoma de *E. nigrum*

O RNA mensageiro obtido de duas amostras do tratamento em 7 dias de cultivo em meio BD foi sequenciado por RNAseq (SOLiD 5500xl) pelo CEFAP-USP. Foram construídas bibliotecas *paired-end* de 35 e 75pb.

4.7 Análise do transcriptoma de *E. nigrum*

Os dados provenientes do sequenciamento SOLiD 5500xl em formato XSQ foram convertidos via XSQ Tools para *color space* e seu correspondente arquivo de qualidade. Em seguida, as bases e sequências com pontuação Phred inferior a 20 foram excluídas. Além disso as sequências correspondentes à porção F3 foi reduzida na sua extremidade 3' de 75 bases para 60 bases. Utilizando o programa Bowtie (LANGMEAD et al., 2009) e bancos de sequências de rRNA e tRNA, foram identificadas e descartadas as sequências positivas. As sequências resultantes foram utilizadas para o mapeamento com o programa Bowtie novamente, mas, utilizando como referência o genoma de *E. nigrum*. Além disso, foi utilizada uma abordagem de montagem “*de novo*” via VelvetG e VelvetH, do pacote Velvet 1.2.10 (<http://www.ebi.ac.uk/~zerbino/velvet/>) com o auxílio do programa Oases 0.2.08 (<https://www.ebi.ac.uk/~zerbino/oases/>), visando determinar as melhores isoformas. A partir das sequências de mRNA, foram realizadas buscas de similaridade com o programa BLASTx do pacote BLAST+. Os melhores alinhamentos obtidos com BLASTx foram selecionados por um *script* em Perl.

4.8 Desenvolvimento de ferramentas de Bioinformática

Foram desenvolvidos componentes em linguagem Perl capazes de invocar e integrar a execução de programas e seus relatórios em um único processo seriado na plataforma de anotação EGene2, derivada da versão original do EGene (DURHAM et al., 2005). Estes componentes foram desenvolvidos em parceria com os professores Dr. Alan Mitchell Durham (IME-USP) e Dr. Arthur Gruber (ICB-USP) com o objetivo de possibilitar a anotação de *E. nigrum*, bem como outros organismos eucariotos. Foram desenvolvidos neste trabalho um total de cinco componentes: `annotation_augustus.pl`, `annotation_myop.pl`, `report_fasta.pl`, `upload_gtf.pl` e `annotation_psort.pl`. Os componentes `annotation_augustus.pl` e `annotation_myop.pl` têm a função de invocar os preditores de genes Augustus (STANKE et al., 2004) e Myop (KASHIWABARA, 2011), respectivamente.

O componente `upload_gtf.pl` foi desenvolvido para inserir anotações diretamente a partir do formatos GTF ou GFF3. O `annotation_psort.pl` possui a função de detectar e anotar sinais de endereçamento de proteínas invocando e integrando os programas Psortb (YU *et al.*, 2010), WoLFPSORT (HORTON *et al.*, 2006) e iPSORT (BANNAI *et al.*, 2002). O componente `report_fasta.pl` gera arquivos em formato FASTA, de acordo com o componente de anotação utilizado (aminoácidos, mRNA, tRNA, genes, CDs, DNA, etc.). Além do desenvolvimento de novos componentes, outros já existentes foram adaptados (`report_feature_table_artemis.pl` e `report_feature_table_submission.pl`) para incorporar os dados de anotação gerados pelos componentes descritos acima.

O programa Synteny Clusters trabalha com *clusters* de metabólitos secundários preditos por AntiSmash (WEBER *et al.*, 2015). Synteny Clusters foi desenvolvido neste trabalho com a finalidade de identificar *clusters* de genes presentes em genomas de diferentes organismos. O programa foi escrito nas linguagens de programação Perl e Java. O programa gera como relatório principal uma página *web* construída em formato HTML com figuras em formato SVG.

4.9 Testes e validação das ferramentas de Bioinformática

4.9.1 Testes e validação de Componentes

Os componentes para EGene foram testados separadamente utilizando-se inicialmente um trecho de sequência (cerca de 32 kb) do *contig0001* de *E. nigrum* e sua integração com outros componentes foi testada com o *contig00193* (cerca de 36 kb). O componente `annotation_psort.pl` foi testado para procariotos usando um *contig* de bactéria de aproximadamente 23 kb.

4.9.2 Testes e validação de Synteny Clusters

Para os testes com Synteny Clusters foram utilizados os resultados da análise de compostos secundários com o programa AntiSmash. Foram utilizados genomas do gênero *Colletotrichum* em razão da proximidade entre os táxons (tabela 2).

Tabela 2 - Espécies de *Colletotrichum* usadas para análise de Synteny Clusters.

Espécie	Tamanho	Referência
<i>Colletotrichum graminicola</i>	51,60 Mb	(O'CONNELL et al., 2012).
<i>Colletotrichum higginsianum</i>	49,08 Mb	(O'CONNELL et al., 2012).
<i>Colletotrichum somersetensis</i>	53,67 Mb	JGI - <i>Colletotrichum somersetensis</i> , 2016.
<i>Colletotrichum sublineola</i>	46,92 Mb	JGI - <i>Colletotrichum sublineola</i> , 2016.
<i>Colletotrichum zoysiae</i>	46,53 Mb	JGI - <i>Colletotrichum zoysiae</i> , 2016.

(Fonte: FERREIRA, A.J., 2016).

4.10 Conjunto de treinamento

O conjunto de treinamento foi criado a partir do próprio genoma de *Epicoccum nigrum*. Os genes de *E. nigrum* utilizados foram obtidos inicialmente por predição com um conjunto de treinamento de *Aspergillus nidulans* fornecidos por Augustus. Por curadoria, os genes com maior probabilidade de serem transcritos no modelo proposto pelo preditor foram validados. A curadoria foi realizada comparando os genes com as sequências dos bancos Swissprot, eggNOG e NCBI com o auxílio da plataforma EGene 2 (DURHAM et al., 2005). Os genes selecionados como modelo para construção do conjunto de treinamento foram submetidos aos *scripts* indicados pelo protocolo do Augustus (STANKE et al., 2008). Após a criação do conjunto de treinamento de *E. nigrum*, sua acurácia foi avaliada por predições e comparações por amostragens sucessivas do fracionamento dos genes utilizados como modelo.

4.11 MAKER

O *pipeline* MAKER foi realizado utilizando-se duas abordagens. Na primeira, utilizou-se como conjunto de treinamento o modelo de *Aspergillus nidulans* fornecido com o programa Augustus. A segunda abordagem teve como base o conjunto de treinamento construído a partir dos genes selecionados do próprio *Epicoccum nigrum*. Ambas as predições foram iniciadas com a entrada do genoma de *E. nigrum*, o qual é automaticamente submetido e processado ao RepeatMasker (SMIT et al., 2015). Em seguida, os resultados do RepeatMasker foram usados como entrada para o programa de predição gênica Augustus utilizando os respectivos conjuntos de treinamento. Além disso, os genes resultantes foram comparados por BLAST com as sequências protéicas de Dothideomycetes (tabela 3) obtidas no banco de dados JGI (<http://jgi.doe.gov/>). A comparação por BLAST também foi realizada com as sequências de RNA mensageiro obtidos do transcriptoma de *E. nigrum*, as quais foram utilizadas para inferir predições. Após a primeira predição, os resultados foram obtidos em

dois estágios de processamento com SNAP para aperfeiçoar as predições e corrigir os possíveis erros nas fronteiras *exon/intron*, bem como falsos positivos. A qualidade da predição foi calculada a partir dos índices de AED presentes no arquivo de saída em formato GFF3 calculados pelo MAKER.

Tabela 3 - Dothideomycetes com proteomas utilizados para comparação de homologia no *pipeline* do MAKER.

Espécie	# Genes	Referência
<i>Alternaria brassicicola</i>	10.688	(OHM et al., 2012)
<i>Aureobasidium pullulans</i> var. <i>melanogenum</i> CBS 110374	10.594	(GOSTINCAR et al., 2014)
<i>Aureobasidium pullulans</i> var. <i>namibiae</i> CBS 147.97	10.266	(GOSTINCAR et al., 2014)
<i>Aureobasidium pullulans</i> var. <i>pullulans</i> EXF-150	11.866	(GOSTINCAR et al., 2014)
<i>Aureobasidium pullulans</i> var. <i>subglaciale</i> EXF-2481	10.809	(GOSTINCAR et al., 2014)
<i>Baudoinia compniacensis</i> UAMH 10762 (4089826) v1.0	10.513	(OHM, et al., 2012)
<i>Cladosporium fulvum</i> v1.0	14.127	(DE WIT, et al., 2012)
<i>Cochliobolus carbonum</i> 26-R-13 v1.0	12.857	(CONDON, et al., 2013)
<i>Cochliobolus heterostrophus</i> C4 v1.0	12.72	(OHM et al., 2012)
<i>Cochliobolus heterostrophus</i> C5 v2.0	13.336	(OHM et al., 2012)
<i>Cochliobolus miyabeanus</i> ATCC 44560 v1.0	12.007	(CONDON et al., 2013)
<i>Cochliobolus sativus</i> ND90Pr v1.0	12.25	(OHM et al., 2012)
<i>Cochliobolus victoriae</i> FI3 v1.0	12.894	(CONDON et al., 2013)
<i>Diplodia seriata</i> DS831	9.343	(MORALES-CRUZ et al., 2015)
<i>Dothistroma septosporum</i> NZE10 v1.0	12.58	(DE WIT et al., 2012)
<i>Hysterium pulicare</i>	12.352	(OHM et al., 2012)
<i>Leptosphaeria maculans</i>	12.469	(ROUXEL et al., 2011)
<i>Macrophomina phaseolina</i> MS6	13.806	(ISLAM et al., 2012)
<i>Mycosphaerella graminicola</i> v2.0	10.933	(GOODWIN et al., 2011)
<i>Neofusicoccum parvum</i> UCRNP2	10.366	(BLANCO-ULATE et al., 2013)
<i>Pyrenophora teres</i> f. <i>teres</i>	11.799	(ELLWOOD et al., 2010)
<i>Pyrenophora tritici-repentis</i>	12.141	(MANNING et al., 2013)
<i>Rhizidhysterium rufulum</i>	12.117	(OHM et al., 2012)
<i>Septoria musiva</i> SO2202 v1.0	10.233	(OHM et al., 2012)
<i>Septoria populicola</i> v1.0	9.739	(OHM et al., 2012)
<i>Setosphaeria turcica</i> Et28A v1.0	11.702	(OHM et al., 2012)
<i>Stagonospora nodorum</i> SN15 v2.0	12.38	(HANE et al., 2007)
<i>Venturia pirina</i>	11.89	(COOKE et al., 2014)
<i>Zymoseptoria ardabiliae</i> STIR04_1.1.1	10.788	(STUKENBROCK et al., 2012)
<i>Zymoseptoria pseudotritici</i> STIR04_2.2.1	11.044	(STUKENBROCK et al., 2012)

(Fonte: FERREIRA, A.J., 2016).

4.12 Anotação do genoma de *E. nigrum*

Foi construído um *pipeline* na plataforma EGene2 com todos os módulos necessários para se realizar a anotação do genoma (tabela 4). Com o módulo de entrada de GFF/GTF (`upload_gtf.pl`) foi inserido o arquivo em formato GFF contendo a predição de genes codificadores de proteínas realizada com MAKER (Augustus + SNAP + mRNA + conjunto

de treinamento de *E. nigrum*). Além da predição de genes codificadores de proteínas, foram identificadas sequências possivelmente transferidas horizontalmente com o programa Alien_hunter. Os RNAs ribossômicos foram preditos com o RNAmmer, enquanto RNAs transportadores foram preditos com o tRNAscan-SE e os RNAs não codificadores de proteínas com o Infernal (Inference of RNA Alignments). Sequências repetidas seriadamente foram preditas com TRF (Tandem Repeats Finder).

As sequências protéicas obtidas pelos preditores de genes (ver item 4.11) foram caracterizadas e anotadas utilizando-se BLAST (identidade por similaridade em banco nr e Swiss-Prot), big-PI/PredGPI (pontos de clivagem para âncoras de GPI), eggNOG (similaridade com genes ortólogos), Interpro (análise funcional e características de proteínas), KEGG pathway mapping (mapeamento em vias do KEGG), Phobius (busca de domínios transmembranares e peptídeo sinal), RPS-BLAST (busca de domínios conservados), SignalP (peptídeo sinal), TMHMM (busca por hélices transmembranares) e TCDB (banco de proteínas de transporte).

As predições e anotações foram organizadas em relatórios específicos para cada programa utilizando-se os módulos do EGene2 (figura 1). Estes relatórios foram organizados em páginas HMTL para facilitar a organização, visualização e consulta das informações com o módulo `report_html.pl`.

Tabela 4 - Fases de anotação e componentes utilizados.

Fase	Etapas de processamento	Função	Componente
Predição	Alien_hunter	identifica sequências transferidas horizontalmente	annotation_alienhunter.pl
	TRF	identifica sequências de repetições seriadas	annotation_trf.pl
	Infernal	identifica sequências de RNAs não codificadores	annotation_infernal.pl
	RNAmmer	identifica sequências de rRNA (5S/5.8S, 16S/18S,23S/28S)	annotation_rnammer.pl
	TRNAscan-SE	identifica sequências de tRNAs	annotation_trna.pl
Anotação	BLASTp x SwissProt	busca regiões similares em sequências de proteínas curadas do banco SwissProt	annotation_BLAST.pl
	BLASTp x nr	busca regiões similares em sequências de proteínas não redundantes na base nr	annotation_BLAST.pl
	big-PI	detecta regiões de clivagem para ligação em âncoras de GPI	annotation_bigpi.pl
	PredGPI	detecta regiões de clivagem para ligação em âncoras de GPI	annotation_predgpi.pl
	eggNOG	banco de dados de grupos de genes ortólogos classificados pela característica funcional	annotation_orthology.pl
	InterproScan	detecta motivos proteicos na base InterPro	annotation_interpro.pl
	KEGG pathway mapping	mapeamento de proteínas em vias da base KEGG	annotation_pathways.pl
	Phobius	busca de domínios transmembranares e peptídeo sinal	annotation_phobius.pl
	RPS-BLAST	compara sequências com bibliotecas de domínios conservados	annotation_rpsBLAST.pl
	SignalP	busca de peptídeo sinal	annotation_signalP.pl
	TMHMM	busca de domínios transmembranares	annotation_tmhmm.pl
	TCDB	classificação de proteínas de transporte	annotation_tcdb.pl
entrada dados	EGene2	entrada de sequência do genoma em formato FASTA	upload_fasta.pl
		entrada da predição de genes codificadores de proteína em formato GFF3	upload_gtf.pl
		entrada de arquivo em formato XML gerado pelo EGene2	upload_xml.pl

(continua)

Tabela 4 - Fases de anotação e componentes utilizados (continuação).

Fase	Etapas de processamento	Função	Componente
Geração de relatórios	EGene2	atribuição de identificador de locus	assign_locus_tags.pl
		organização das informações da anotação	report_conclusion.pl
		gravação de dados em formato XML dos dados	outsave.pl
		relatório de sequências (aminoácidos, genes, CDS e mRNA) em formato FASTA	report_fasta.pl
		relatório da anotação em formato <i>feature table</i> compatível com Artemis	report_feature_table_artemis.pl
		relatório da anotação em formato <i>feature table</i> compatível com o GenBank	report_feature_table_submission.pl
		relatório da anotação em formato GFF3	report_gff3.pl
		relatório da anotação dos termos GO (Gene Ontology)	report_go_mapping.pl
		relatório da anotação dos resultados de eggNOG	report_orthology.pl
		relatório da anotação dos resultados de KEGG	report_pathways.pl
		constrói página HTML com os relatórios	report_html.pl

(Fonte: FERREIRA, A.J., 2016).

Maker	
-------	--

4.2 Relação Filogenética de famílias de genes de *E. nigrum*

4.2.1 Escolha de genomas

Para realização do protocolo MIRLO, além do proteoma de *E. nigrum* foram escolhidos 41 proteomas (tabela 20), sendo 32 proteomas de espécies de fungos patogênicos, saprófagos e endofíticos, além de outros 10 para dar robustez à análise (todos pertencentes ao Sub-reino Dikarya).

4.2.2 Protocolo MIRLO

Para a análise, seguindo-se os passos do protocolo MIRLO (THON, 2015), primeiramente foi realizada uma busca de similaridade com BLASTx “todos contra todos” das sequências dos proteomas dos 41 fungos escolhidos juntamente com *E. nigrum*. O resultado da saída foi formatado com os *scripts* fornecidos pelo MIRLO e submetidos à análise de agrupamento MCL. A partir da saída de MCL, foi possível se analisar as proteínas com grande similaridade e presentes em todos os organismos. Estas sequências foram alinhadas com MAFFT (KATO et al., 2002) e usadas para construir árvores PhyML (máxima verossimilhança). As melhores árvores do resultado foram avaliadas por meio do tamanho médio de seus ramos com suporte *SH-like*.

Os proteomas das espécies dos ramos mais próximos foram selecionados e submetidos novamente ao protocolo MIRLO para avaliação do número de famílias de proteínas.

4.2.3 Calibração das Árvores

Os melhores resultados de proteínas codificadas pelos genes selecionados tiveram suas sequências alinhadas (MAFFT), concatenadas e uma árvore filogenética ultramétrica foi construída com PhyML. O tamanho dos ramos foi calibrado com o programa R8s (SANDERSON, 2003), assumindo como prováveis idades dos ramos de Sordariomycetes (182 a 316 milhões de anos) e Dothideomycetes (247 a 427 milhões de anos) os dados de Beimforde et al., 2014.

4.2.4 Avaliação de modificações nas famílias de genes

A árvore construída e calibrada com R8s foi utilizada como modelo de evolução de aquisição de novos genes, perdas, e duplicações dos genes existentes por meio do programa CAFE (DE BIE et al., 2006).

4.3 Avaliação de *clusters* de metabólitos secundários

Os *clusters* de genes relacionados a metabólitos secundários foram preditos pelo programa AntiSmash para todos os proteomas de Dothideomycetes mais próximos de *E. nigrum*, de acordo com as análises de filogenia. Estes resultados foram inseridos no Synteny Clusters onde, por meio de comparações, foram selecionados os genes compartilhados pelos patógenos.

Os genes de *clusters* de metabólitos secundários foram comparados por BLASTP ao banco de dados PHI-base (WINNENBURG et al., 2007) e determinados aqueles com maiores probabilidades de estarem relacionados ao caráter fitopatogênico dos fungos. A estrutura dos metabólitos secundários identificados pelo AntiSmash foi comparada com a base de dados do ChemSpider (<http://www.chemspider.com/>) para verificar a estrutura dos compostos.

5 RESULTADOS

5.1 Extração do RNA mensageiro

A quantidade reduzida de micélio nas culturas com 7 dias não interferiu na extração do mRNA, mas nas amostras com 60 dias não foi possível obter RNA mensageiro em quantidade e qualidade suficientes para o sequenciamento via SOLiD.

5.2 Obtenção dos extratos e eficiência dos antimicrobianos

As amostras geraram quantidades relevantes de extrato, sendo a maior quantidade obtida proporcional ao tempo do tratamento. Os solventes que geraram uma maior quantidade de metabólitos foram respectivamente butanol, acetato de etila, clorofórmio e hexano. Foi identificada produção de metabólitos com propriedades antimicrobianas a partir de 7 dias nas amostras obtidas de clorofórmio contra *Bacillus* sp., hexano contra *E. coli* e acetato de etila contra *S. aureus*. Entretanto, os metabólitos com afinidade por butanol apresentaram atividade inibidora contra *Bacillus* sp. apenas nas amostras com 60 dias (figura 2).

Os extratos obtidos das culturas de 10, 20 e 30 dias nos meios BD e Sabouraud apresentaram maiores diferenças entre o controle e as amostras (figura 3). Nos tratamentos utilizando o meio BD foi observada atividade significativamente maior a partir de 20 dias com os metabólitos com afinidade por acetato de etila e butanol em relação a 10 dias. A partir de 20 dias o tratamento com clorofórmio também contra *E. coli* e *S. aureus*. Entretanto, a atividade apresenta redução no período de 30 dias em *E. coli*. Além disso, os testes com hexano apresentaram significância apenas em 20 dias com *Bacillus* sp. e *S. aureus*.

Nos tratamentos utilizando os extratos obtidos das culturas em meio Sabouraud, a partir de 20 dias, foi observada atividade antimicrobiana (contra *Bacillus* sp. e *S. aureus*) significativamente maior em acetato de etila. As amostras obtidas de hexano apresentaram atividade antimicrobiana (contra *Bacillus* sp. e *S. aureus*) significativamente maior a partir de amostras de 20 dias em relação aos tratamentos com 10 e 30 dias.

Figura 2 - Teste de produção de antimicrobiano com 7 e 60 dias.

Teste de produção de antimicrobiano usando o próprio solvente como controle (centro da placa). Extratos de acetato de etila, butanol, clorofórmio e hexano obtidos de culturas de *Epicoccum nigrum* com 7 e 60 dias (* $p < 0,05$). (Fonte: FERREIRA, A.J., 2016).

Os extratos de compostos com afinidade por diferentes solventes geraram uma grande quantidade de tratamentos, sendo possível observar quais tratamentos foram melhores para comparar com os dados de transcriptoma. Os resultados dos testes de eficiência dos antimicrobianos extraídos em acetato de etila e butanol indicam que a comparação dos tempos 10 e 20 dias dos transcriptomas do fungo, podem indicar os genes envolvidos no metabolismo destes compostos antimicrobianos. Os testes com hexano mostraram grande atividade apenas após 20 dias de cultivo, indicando uma redução da atividade antimicrobiana dos compostos produzidos pelo fungo em meio Sabouraud.

Figura 3 - Teste de produção de antimicrobiano com 10, 20 e 30 dias.

Extratos de acetato de etila, butanol, clorofórmio e hexano obtidos de culturas de *E. nigrum* com 10, 20 e 30 dias (* $p < 0,05$). Em A, extrato a partir de fungo cultivado em BD contra *Bacillus* sp. B, extrato a partir de meio Sabouraud contra *Bacillus* sp. C, extrato a partir de BD contra *S. aureus*. D, extrato a partir de Sabouraud contra *S. aureus*. E, extrato a partir de BD contra *E. coli*. F, extrato a partir de Sabouraud contra *E. coli*. B10 – crescido em meio por BD com 10 dias, B20 – meio BD por 20 dias, B30 – meio BD por 30 dias. S10 – crescido em meio Sabouraud por 10 dias, S20 – meio Sabouraud por 20 dias, S30 – meio Sabouraud por 30 dias. (Fonte: FERREIRA, A.J., 2016).

5.3 Sequenciamento do transcriptoma de *E. nigrum*

O sequenciamento do RNA mensageiro obtido de duas amostras do tratamento em 7 dias de cultivo em meio BD por RNAseq resultou em aproximadamente 124 milhões de leituras *paired-end* (tabela 5). A amostra 7(1) rendeu aproximadamente 35 milhões de sequências com 75 pb e 50 milhões com 35 pb. Enquanto a amostra 7(2) rendeu aproximadamente 6,3 milhões de leituras com 75 pb e 31 milhões com 35 pb.

O sequenciamento do transcriptoma foi bem-sucedido, mas uma grande quantidade de sequências foi removida pelos critérios de estringência. Apesar de mapeados transcritos de tamanho inferior à maioria das proteínas (N50 de 188), foi obtida uma grande quantidade de *contigs* (77 mil). A análise por BLASTx dos transcritos contra a base de dados do GenBank indicou coerência na montagem *de novo* realizada com o Velvet, visto que ocorreu predominância de *hits* provenientes de fungos, principalmente de grupos próximos como *Leptosphaeria maculans* e *Pyrenophora teres* (FÁVARO, 2009; FÁVARO et al., 2011).

Tabela 5 - Características do sequenciamento do transcriptoma do isolado *E. nigrum* P16.

Dados Primários	Total (# de bases)	Total (# leituras)	7 (1)	7 (2)
leituras (F3) cruas	28,20 Gb	376 M	102 M	174 M
leituras (F5) cruas	13,16 Gb	376 M	102 M	174 M
leituras (F3)	3,24 Gb	43,18 M	35,5 M	6,3 M
leituras (F5)	2,84 Gb	81,2 M	50,2 M	31 M
leituras (Após filtro Q20 – F3)	3,08 Gb	41,1 M	34,9 M	6,2 M
leituras (Após filtro Q20 – F5)	1,81 Gb	51,8 M	43 M	8,8 M
<i>Bowtie</i> - leituras mapeados a tRNA e rRNA (7(1)+7(2), F3+F5)		6 M		
<i>Bowtie</i> - leituras mapeados (7(1)+7(2), F3+F5)		23 M		
<i>Bowtie</i> - <i>contigs</i> do genoma mapeados totais		562		
<i>Bowtie</i> - <i>contigs</i> do genoma mapeados com mRNA		561		
<i>Bowtie</i> - <i>contigs</i> mRNA	9,45Mb	77,8 K		
<i>Bowtie</i> N50	188			
<i>Velvet</i> - <i>contigs</i> mRNA	3,4 Mb	20,8 K		
<i>Velvet</i> - N50	161 b		154 b	184 b

Siglas: K (mil), M (milhões), Mb (milhões de bases), Gb (bilhões de bases), Q20 – índice de qualidade da base de no mínimo 20, N50 (metade de todas as bases residem em *contigs* desse tamanho ou maior). (Fonte: FERREIRA, A.J., 2016).

5.4 Análise do transcriptoma de *E. nigrum*

No processo de filtragem das sequências de acordo com o índice de qualidade a amostra 7(1) foi reduzida em aproximadamente 61%, enquanto a amostra 7(2) foi reduzida em 95%. A melhor abordagem para o mapeamento das sequências resultantes foi a que utilizou o programa Bowtie (LANGMEAD et al., 2009) com sequências no formato “*color space*”. Enquanto a montagem *ab initio*, foi bem-sucedida utilizando os softwares do pacote Velvet.

No Bowtie foram alinhadas ao genoma de *E. nigrum* um total de aproximadamente 23 M (milhões), 23% do total de sequências de transcritos após a exclusão de sequências de rRNA e tRNA (6 milhões aprox.). Na abordagem de montagem “*ab initio*” após a curadoria, a montagem do transcriptoma da amostra 7(1) utilizando o programa Velvet com o auxílio do Oasis resultou em 20 milhões *isotigs*, tendo o maior deles 1.021 pb, um N50 de 154 pb e 49% de conteúdo CG. Apesar da amostra 7(2) ter apresentado um número menor de *isotigs* (4.583) seu N50 foi maior (184 pb). Além disso, a amostra 7(2) apresentou 44% de CG e o maior *isotig* com 823 pb.

As sequências de *isotigs* reconhecidas como pertencentes à RNA mensageiro pelo montador foram comparadas contra a base de sequências de proteínas do NCBI. A análise indicou um total de 837 *hits* diferentes de proteínas resultantes de 1.757 sequências utilizadas como *query*. Os resultados mais frequentes foram genes de codificação de proteínas da subunidade ribossômica 60S, 6-fosfogluconato desidrogenase, dessaturase acil-CoA, álcool desidrogenase, aldeído desidrogenase, ATP sintase, citocromo c oxidase, citocromo P450, glicosilhidrolase, NADP-dependente manitol desidrogenase, fosfoglucomutase, fosfoquetolase, piruvato decarboxilase, superóxido dismutase e ubiquitina.

Os resultados do BLASTx dos transcritos mais frequentes nas amostras com 7 dias indicaram predominância de produtos envolvidos em vias de compostos primários como pentoses, glicólise e base de síntese de proteínas como proteínas das subunidades 60S e 40S. Outras enzimas estavam envolvidas em vias diversas como álcool desidrogenase, fosfoglucomutase aldeído desidrogenase.

5.5 Desenvolvimento de ferramentas de Bioinformática

5.5.1 Desenvolvimento de scripts de manipulação de dados

O grande volume de dados oriundos da saída do programa BLAST+ requereu a construção de um programa em Perl capaz receber o arquivo e gerar relatórios simplificados.

A ferramenta obteve sucesso em gerar relatórios discriminando sequências que não apresentaram *hits*, sequências com informações insuficientes (*hypothetical protein* e *predicted protein*). Além disso, a ferramenta também foi construída para gerar um relatório com apenas os *hits* mais relevantes e selecionar automaticamente o melhor *hit* para cada sequência.

5.5.2 Componentes de anotação plataforma EGene2

Os componentes desenvolvidos foram o `annotation_augustus.pl`, `annotation_myop.pl`, `annotation_psort.pl`, `upload_gtf.pl` e `report_fasta.pl`. Além disso, foram realizadas modificações para entrada correta dos dados em componentes pré-existent que fornecem relatórios para abertura em Artemis (CARVER et al., 2012) e outro capaz de emitir uma saída em formato compatível com o de submissão ao NCBI.

Os componentes foram testados em “*pipeline*” montado utilizando apenas como entrada o arquivo fasta contendo o maior *contig* (*contig* 1) de *E. nigrum* ou GTF. O componente `annotation_augustus.pl` foi capaz de invocar o programa Augustus (STANKE et al., 2004), executá-lo, obter o seu arquivo de saída e integrar suas informações em um único processo seriado. Foram preditos 23 genes, os quais tiveram extraídas suas informações importantes a respeito da posição no genoma de *introns*, *exons*, códons de iniciação e terminação, além da sequência protéica. Estas informações por sua vez foram repassadas aos outros componentes sendo possível realizar consulta nos bancos de dados e programas de busca de regiões genômicas, bem como, domínios protéicos das sequências traduzidas e a possível localização celular dessas proteínas. Assim, estas informações puderam ser integradas, permitindo de forma satisfatória a geração de relatórios em formato Artemis (CARVER et al., 2012) (figura 4) e arquivos no formato de submissão requerido pelo NCBI (figura 5). O mesmo desempenho pode ser realizado com `annotation_myop.pl`, o componente invocou Myop corretamente, permitindo a integração de suas informações de coordenadas de genes, CDS, e códons de iniciação e terminação à plataforma EGene2. Foram preditos 26 genes e 144 CDS os quais puderam ser visualizados no formato *feature table* compatíveis com Artemis e GenBank. Outro componente de inserção de anotação foi `upload_gtf.pl`. O relatório em formato Artemis indicou a correta extração das informações do arquivo em formato GFF utilizado e integração correta dos 93 genes e 306 CDSs utilizados para validação.

Figura 4 - Anotações de genes exibidas em Artemis.

Anotação em relatório visualizada em Artemis. A - trecho anotação em eucarioto. B – trecho de anotação em bactéria. (Fonte: FERREIRA, A.J., 2016).

Figura 5 - Anotação de genes em relatório em formato *feature table* gerado pela plataforma EGene

Amostra de anotação de gene em relatório de submissão gerado automaticamente pelo *pipeline* da plataforma EGene utilizando componentes integrados. A - anotação de eucarioto com destaque para upload_gtf.pl e WoLFPSORT. B - anotação de procarioto com destaque para Psortb. (Fonte: FERREIRA, A.J., 2016).

O componente `annotation_psort.pl` foi testado tanto com os genomas de procarioto quanto eucarioto com sucesso. Seu modo de operação é diferente dos componentes descritos acima, pois envolve a invocação do gerenciador de informações do EGene e a extração das coordenadas de genes e CDS previamente descritos, estas informações foram convertidas em sequências de proteínas as quais foram submetidas a Psortb (procariotos) ou WoLFPSORT e iPSORT (eucariotos). Os relatórios a respeito das informações de possível localização subcelular das proteínas foram integrados pelo próprio componente às características da proteína, sendo visualizadas nos relatórios de anotação. Além dos componentes de anotação, o componente de relatório `report_fasta.pl` foi testado com uma anotação prévia com o objetivo de obter arquivos em formato fasta contendo sequências de aminoácidos, gene, CDS e mRNA.

Os testes de integração dos componentes foram realizados para eucariotos com o *contig* 193 de *E. nigrum* e utilizou 34 componentes para realizar a anotação. Enquanto os testes de integração dos componentes para procariotos utilizaram 35 componentes (tabela 6). Os testes de validação (tabela 7) estão presentes no endereço <https://drive.google.com/open?id=0B0P5-T-6tCtxSk9zNDhDNTNUOGM>).

Tabela 6 - Componentes utilizados para teste de integração.

Componente	Procarioto	Eucarioto
annotation_alienhunter.pl	X	X
annotation_bigpi.pl	X	X
annotation_BLAST.pl	X	X
annotation_dgpi.pl	X	X
annotation_glimmer3.pl	X	
annotation_infernal.pl	X	X
annotation_interpro.pl	X	X
annotation_mreps.pl	X	
annotation_orthology.pl	X	X
annotation_pathways.pl	X	X
annotation_phobius.pl	X	X
annotation_predgpi.pl	X	X
annotation_psort.pl	X	X
annotation_rbsfinder.pl	X	X
annotation_rnammer.pl	X	X
annotation_rpsBLAST.pl	X	X
annotation_signalP.pl	X	X
annotation_string.pl	X	
annotation_tcdb.pl	X	X
annotation_tmhmm.pl	X	X
annotation_transterm.pl	X	X
annotation_trf.pl	X	X
annotation_trna.pl	X	X
assign_locus_tags.pl	X	X
outsave.pl	X	X
report_conclusion.pl	X	X
report_fasta.pl		X
report_feature_table_artemis.pl	X	X
report_feature_table_submission.pl	X	X
report_gff3.pl	X	X
report_go_mapping.pl	X	X
report_orthology.pl	X	X
report_pathways.pl	X	X
upload_fasta.pl	X	X
upload_gtf.pl		X
upload_xml.pl	X	X

(Fonte: FERREIRA, A.J., 2016).

Tabela 7 - Testes de validação dos componentes para EGene2.

Diretório	Componente	Objetivo
AUGUSTUS	annotation_augustus.pl	Invocação e inserção de predição do programa Augustus
BACTERIA	annotation_psort e report_fasta.pl	Integração entre componentes (tabela 8).
EUCARIOTO	annotation_augustus.pl, annotation_psort.pl, report_fasta.pl	Integração entre componentes (tabela 8).
MYOP	annotation_myop.pl	Invocação e inserção de predição do programa Myop
PSORT	annotation_psort.pl	Invocação e inserção de predição dos programas WoLFPSORT e iPSORT
REPORT_FASTA	annotation_augustus.pl, report_fasta.pl	Obter sequências de aminoácidos, CDS, genes e mRNA de predições do Augustus
UPLOAD_GTF	upload_gtf.pl	Inserir uma anotação de gene predição prévia ou nova sem a utilização de preditor

(Fonte: FERREIRA, A.J., 2016).

Tabela 8 - Programas e bancos de dados utilizados no *pipeline*.

Programa	Função	Referência
Alien_hunter	Localiza prováveis genes transferidos horizontalmente	VERNIKOS; PARKHILL, 2006.
Augustus	Predição de genes eucarióticos	STANKE et al., 2004.
BLAST	Busca regiões similares em sequências biológicas	ALTSCHUL et al., 1990.
DGPI	Detecta regiões de clivagem para ligação em âncoras de GPI	PIERLEONE et al., 2008.
eggNOG	Banco de dados de grupos de genes ortólogos anotados de acordo com a característica funcional	POWELL et al., 2014.
Glimmer	Predição de genes microbianos (principalmente bactérias e arqueias)	DELCHER et al., 2007.
Infernal	Localiza sequências de RNA não codificador	NAWROCKI; Eddy, 2013.
Interpro	Análise funcional, classificação de famílias de proteínas e sítios e domínios importantes.	JONES et al., 2014.
iPSORT	Deteção de peptídeo sinal em proteínas eucarióticas	BANNAI et al., 2002.
KEGG pathway mapping	Busca e representação de enzimas em mapas metabólicos	KANEHISA et al., 2012.
MREPS	Detecta repetições seriadas	KOLPAKOV et al., 2003.
Myop	Predição de genes eucarióticos	KASHIWABARA, 2011.
Phobius	predição de domínios transmembranae sinalpeptídeos	KÄLL et al., 2007.
Psortb	Localização subcelular de proteínas em procariotos	YU <i>et al.</i> , 2010
RNAmmer	Localiza sequências de rRNA (5s/5,8s, 16s/18s,23s/28s)	LAGESEN et al., 2007.
RPS-BLAST	Compara uma sequência com bibliotecas de domínios conservados	MARCHLER-BAUER et al., 2002.
SignalP	Predição dos pontos de clivagem em sequências peptídicas.	PETERSEN et al., 2011.
String	Detecta repetições seriadas	PARISI et al., 2003.
TMHMM	Localiza hélices transmembrana	KROGH et al., 2001.
TrantermHP	Base de dados para tradução de mRNA	JACOBS et al., 2002.
TRF	Localiza sequências seriadas	BENSON, 1999.
TRNAscan-SE	Localiza sequências de tRNA e snoRNAs	SCHATTNER et al., 2005.
WoLFPSORT	Localização subcelular de proteínas em eucariotos	HORTON et al., 2006

(Fonte: FERREIRA, A.J., 2016).

5.5.3 Synteny Clusters

5.5.3.1 Execução

O Programa Synteny Clusters foi desenvolvido para comparação de *clusters* de metabólitos secundários de diferentes organismos. Sua parte de processamento lógico foi desenvolvida em linguagem Perl e com linguagem Java foi criada uma interface mais intuitiva para os usuários (figura 6). O programa foi desenvolvido em Linux, tendo os pré-requisitos instalados (Perl 5 e Java 7) e não necessita de instalação. Juntamente com o manual estão as

instruções para sua instalação e utilização. Na janela principal, o usuário pode abrir um *link* de ajuda em caso de dúvidas no botão indicado por um ponto de interrogação (figura 6 - 1H). Para utilização, o usuário primeiramente deve ter realizado a análise de compostos secundários com AntiSmash e realizado o *download* dos dados. Os diretórios contendo os resultados de AntiSmash são selecionados para a comparação de seus *clusters* com o botão “add directories” (figura 6 - 1A), o qual aciona uma janela de seleção de múltiplos diretórios (figura 6 - 2C). Após a seleção, o usuário pode selecionar o destino onde deseja que os resultados sejam gravados clicando no botão “folder destiny” (figura 6 - 1B) onde abrirá uma caixa de seleção similar à anterior. Após a seleção dos diretórios de entrada e destino o usuário avança (figura 6 - 1C), abrindo uma janela para o preenchimento com o nome dos organismos que deseja que sejam exibidos nos resultados (figura 6 - 3A). Após o preenchimento, o usuário segue (figura 6 - 3B) e executa (figura 6 - 4A).

A identificação dos genes é baseada nos relatórios de BLAST do AntiSmash, assim sua identificação está sujeita ao nome que as sequências foram depositadas no GenBank. Uma vez que as sequências de genes similares são frequentemente depositadas com nomes diversos ou sinônimos como no caso de enzimas, a comparação dos genes de diferentes *clusters* necessita de nomes similares para identificação do gene entre *clusters*. Para contornar este problema, Synteny Clusters conta com dicionário de sinônimos e simplificação dos termos mais frequentes em *clusters*. Devido à grande variedade destes termos é possível que o usuário encontre algum resultado redundante com um gene com duas nomenclaturas similares. Dessa forma, Synteny Clusters também conta com uma seção de substituição de termos. O usuário acessa esta sessão após a ter realizado a análise primeiramente indicando o destino onde foram salvos os arquivos de resultado da análise anterior (figura 6 1-B). Depois é só clicar em “New Terms” (figura 6 - N1) e abrirá uma janela com um botão “Load all terms” (figura 6 - N2) com todos os nomes de genes utilizados. Similarmente à sessão de preenchimento de nomes, esta seção contém duas colunas (figura 6 - N3), tendo a primeira os nomes dos genes utilizados e separados por conjunto de *clusters* usados. Basta o usuário preencher na segunda coluna apenas os nomes que deseja substituir na primeira e seguir clicando em “next” (figura 6 N-4). Na nova janela é só iniciar a análise clicando em “run” (figura 6 – 4A) e os resultados antigos serão substituídos pelos novos.

Figura 6 - Interface Synteny Clusters e passos para executar a análise.

Interface do programa Synteny Clusters com passos. 1H - Botão de ajuda. 1A - selecionar diretórios de resultados de AntiSmash. 1B - selecionar diretório de destino. 2C - caixa de seleção de diretório. 3A - campo de preenchimento com o nome do táxon. 3B - Botão para avançar após terminar preenchimento. 4A - execução de análises. N1- abre janela de correção de termos. N2 – carrega termos utilizados. N3 – campo de substituição de termos. N4 – botão para avançar após o preenchimento dos termos desejados. (Fonte: FERREIRA, A.J., 2016).

5.5.3.2 Funcionamento

O programa Synteny Clusters foi desenvolvido primeiramente para identificar quais *clusters* de metabólitos secundários são similares entre si em diferentes organismos. Na análise de AntiSmash não há um padrão de numeração de *clusters*, sendo estes apenas ordenados de acordo com a ordem de descoberta para cada sequência do organismo. Isto significa, por exemplo, que o *cluster* chamado de “1” por AntiSmash de determinado organismo pode ser similar ao *cluster* 30 de outro, este por sua vez pode ser similar ao *cluster* de número 13 de um terceiro e assim por diante. Synteny Cluster é capaz de verificar quais *clusters* são similares por meio de comparação com resultados de bancos de dados emitidos por AntiSmash. Estes “hits” são comparados e apenas os *clusters* com similaridade razoável e presentes em todos os organismos são agrupados como um conjunto. A partir da detecção do conjunto de *clusters* similares é realizada uma análise dos genes de cada *cluster* que são similares por meio de seus resultados de BLAST. Esta etapa consiste na detecção de genes equivalentes entre os *clusters* de diferentes organismos e uma padronização de suas cores (antes não relacionáveis) para facilitar a visualização do usuário. Estes resultados podem ser visualizados em página *web* por meio de imagens em formato SVG dos *clusters* agrupados. Assim, é possível realizar a sentença dos *clusters* em organismos diferentes e construir hipóteses para identificação de outros genes, assim como a evolução e importância de determinado *cluster* para aquele grupo de organismos.

5.5.3.3 Interpretação do relatório

Os resultados da análise de Synteny Clusters são encontrados no diretório de nome padrão “Result_Synteny Clusters” localizado no destino de escolha do usuário. Dentro do diretório criado o usuário encontrará dentre os arquivos e diretórios de configuração um chamado “ResultSyntenyClusters.html”. Este, quando aberto em Firefox (Mozilla), exibe uma página em formato *web* com os resultados (figura 7). Os *clusters* considerados homólogos são mostrados separadamente em conjuntos. Cada conjunto é selecionado através de um botão seletor (figura 7 - A), expondo os *clusters* que estão organizados por organismo de origem com o nome descrito pelo usuário na seção de preenchimento (figura 6 - 3B). Os *clusters* são identificados por nome do organismo de origem, número do *cluster* descrito por AntiSmash e a categoria de metabólito ao qual o *cluster* está relacionado (figura 7 - B). Os genes dentro dos *clusters* dos organismos seguem as características encontradas no genoma, seguindo ordem e tamanho proporcional. Além disso, suas cores seguem de acordo com a possível

homologia, entre os genes dos outros *clusters*. Abaixo do conjunto de *clusters* está a legenda com todos os genes (figura 7 - C). A legenda inferior é composta pela cor dos genes, nome do *locus*, nome do organismo de origem, o número do *cluster*, sua categoria e a melhor identificação do gene por BLAST, seguida pelos atributos da identificação. Para cada gene também há legendas dinâmicas (figura 7 - D) com detalhes contendo, nome do gene, origem, seu *locus* e link para abertura da proteína codificada em outra página para investigações (figura 7 - E).

Figura 7 - Página em formato *web* exibindo resultados de Synteny Clusters.

Resultados de sintenia de *clusters* realizada para fungos do gênero *Colletotrichum*. Em A, botão seletor de conjunto de *clusters* homólogos. Em B, *clusters* com genes coloridos de acordo com sua identificação. Em C, legenda seguindo as cores, nome, origem do gene e identificação seguindo os resultados de BLAST. Em D, legenda com nome do gene acima, abaixo link para sequência protéica e localização. Em E, detalhe com sequência protéica que abre a partir do link. (Fonte: FERREIRA, A.J., 2016).

5.5.3.4 Validação do Synteny Clusters

A análise do Synteny Cluster foi realizada utilizando 5 espécies de *Colletotrichum* (tabela 2) por causa da proximidade entre os organismos. Foram encontrados 4 *clusters* similares entre todas as espécies. O primeiro conjunto de *clusters* está relacionado a terpenos, o segundo a peptídeos não ribossômicos, o terceiro a outra categoria e o quarto a PKS do tipo 1. Os quatro conjuntos de *clusters* apresentaram genes com identificação suficiente para relacioná-los entre *clusters*, sendo observados poucos não relacionáveis (sem resultado de BLAST ou hipotéticos). Além disso, os genes foram marcados corretamente com as cores correspondentes, sendo possível a fácil visualização. Ainda analisando funcionalmente os resultados do programa, as legendas dinâmicas apresentaram corretamente as informações e o link para abertura foi capaz de dar acesso às proteínas de forma correta (análise está disponível no endereço <https://drive.google.com/open?id=0B0P5-T-6tCtxSk9zNDhDNTNUOGM>).

5.6 Relações filogenéticas

5.6.1 Conjunto de treinamento

A predição inicial realizada com o programa Augustus utilizou o conjunto de treinamento de *Aspergillus nidulans*, resultando 11.412 genes inteiros e parciais (tabela 10). A partir destes, foi feita uma curadoria manual que levou à seleção de 420 genes de *E. nigrum* potencialmente expressos, os quais apresentaram alta cobertura de *exons* do início ao fim quando comparados aos bancos de dados. Estes genes foram utilizados como modelo de entrada para os scripts do programa Augustus para a construção do conjunto de treinamento. A qualidade do conjunto de treinamento foi avaliada por meio das medições da qualidade da predição dos genes em meio ao genoma de *E. nigrum*. A predição apresentou uma especificidade de 86,3% e uma sensibilidade de 68,8% exatidão dos *exons* preditos (tabela 9).

Tabela 9 - Teste de acurácia de conjunto de treinamento.

Total genes	Sensibilidade	Especificidade	Falso positivo			Falso negativo		
			parcial	sobreposição	erro	parcial	sobreposição	erro
420	86,30%	68,80%	28	7	96	28	3	15

(Fonte: FERREIRA, A.J., 2016).

5.6.2 Predição de genes

A predição apenas com o programa Augustus utilizando como conjunto de treinamento os genes de *A. nidulans* estimou a presença de 11.412 genes (figura 8 A) com um tamanho médio de 1762 pb (figura 9 A). O número de CDSs foi estimado em 33.217 CDSs (figura 8 B) com um tamanho médio de 546 pb (figura 9 B). Entretanto, a mesma estratégia de predição utilizando como conjunto de treinamento os genes de *E. nigrum* estimou um número menor de genes (7.899) com um tamanho médio também ligeiramente menor (1.517 pb). Seguindo a mesma tendência o preditor calculou 21.700 de CDSs (~34% menor) com um tamanho médio de 491 pb (tabela 10).

Com o *pipeline* do programa MAKER o qual reuniu a predição do Augustus juntamente com as evidências dos proteomas de Dothideomycetes (tabela 3) e sequências de mRNA de *E. nigrum* foi observada uma mudança no perfil estrutural dos genes. Neste perfil estrutural foi verificado um aumento no número de genes preditos com o conjunto de treinamento de *E. nigrum*. Contrariamente, o número de genes preditos utilizando o conjunto de treinamento de *A. nidulans* diminuiu, equiparando em ambos os casos em 10.605 genes (figura 8 A). O mesmo evento foi observado com o número de CDSs que aumentou no conjunto de treinamento de *E. nigrum* (27.935) e reduziu em *A. nidulans* (29.381) (figura 8 B) (tabela 10). A comparação das análises indicou uma estabilidade no tamanho médio das CDSs preditas com o conjunto de treinamento de *E. nigrum* (figura 9 B), enquanto foi observado um leve aumento no tamanho das mesmas com o conjunto de treinamento de *A. nidulans*.

A utilização de SNAP combinado ao Augustus, juntamente com as evidências, não expandiu a detecção de genes e CDSs. Entretanto, foi observado um decréscimo no tamanho médio dos mesmos quando avaliado os genes preditos com o conjunto de treinamento de *A. nidulans* (1.582,1), assim como os CDSs (486,8) (figura 9 B).

Além da predição de genes e seus CDSs, com a adição das sequências de mRNA foi possível ao MAKER realizar a predição das regiões 5'UTR e 3'UTR. Inicialmente MAKER

foi capaz de prever 151 sequências de 5'UTRs e 68 3'UTRs (figura 8 C e D), enquanto seu tamanho em média foi de 162 e 246,4 pares de bases, respectivamente, para o tratamento utilizando o conjunto de treinamento de *A. nidulans* (figura 9 C e D). O tratamento utilizando o conjunto de treinamento de *E. nigrum* apresentou quantidade e tamanho menor de UTRs, dos quais 64 5'UTRs, 44 3'UTRs com tamanhos médios respectivos de 109 pb para 5'UTR e 180 pb para a 3'UTR. Assim como para CDSs, a quantidade de UTRs e seu tamanho médio diminuíram no tratamento envolvendo o conjunto de treinamento de *A. nidulans* em relação ao tratamento com o conjunto de treinamento de *E. nigrum* quando submetidos ao programa SNAP. Essa combinação de SNAP gerou 74 5UTR's com 118 pb em média e 49 5UTR's com 187 pb em média no conjunto de treinamento de *A. nidulans*. Enquanto, os 5'UTRs se mantiveram estáveis com 65 5'UTRs com 109 pb em média e 45 3UTR's de 176 pb em média.

A abordagem de anotação utilizando apenas genes inteiros revelou uma grande similaridade entre os genes, CDSs e UTRs das duas predições com distintos conjuntos de treinamento. Os números das sequências preditas foram extremamente próximos, mesmo com diferentes conjuntos de treinamento (figura 8 A - D). O número de genes preditos (10.320) foi o mesmo para ambos os tratamentos. Além dos genes, os CDSs (27.826) e os 3'UTRs (62) também foram preditos na mesma quantidade. Apenas a predição utilizando *E. nigrum* que encontrou uma sequência de 5'UTR a mais (265) em relação a *A. nidulans* (264). O tamanho médio das sequências preditas também apresentou grande similaridade entre os tratamentos considerando apenas genes inteiros (figura 9 A - D). Nestes casos os tratamentos apresentaram genes, CDSs, 5'UTRs e 3'UTRs com tamanho médio de 1.591 pb, 489 pb, 92 pb e 162 pb, respectivamente (tabela 10).

Figura 8 - Número de seqüências de acordo com predição.

A. Número de seqüências de genes. B. Número de seqüências de CDS. C Número de seqüências de RNA 3' UTR. D. Número de seqüências de RNA 5' UTR. Tipo de predição: AugParc - Preditor Augustus incluindo seqüências parciais, MarkAugParc - MAKER com Augustus incluindo genes parciais. MarkAugSnParc - MAKER com Augustus e SNAP incluindo genes parciais. MarkAugSnIn - MAKER com Augustus e SNAP apenas genes inteiros. Conjunto de Treinamento: An - *Aspergillus nidulans*. En - *Epicoccum nigrum*. (Fonte: FERREIRA, A.J., 2016).

Figura 9 - Tamanho médio (pares de bases) de sequências de acordo com Predição.

A - Tamanho médio (pares de bases) de sequências de genes. B - Tamanho médio (pares de bases) de sequências de CDS. C - Tamanho médio (pares de bases) de sequências de RNA 3' UTR. D - Tamanho médio (pares de bases) de sequências de RNA 5' UTR. Tipo de predição: AugParc - Preditor Augustus incluindo sequências parciais, MarkAugParc - MAKER com Augustus incluindo genes parciais. MarkAugSnParc - MAKER com Augustus e SNAP incluindo genes parciais. MarkAugSnIn - MAKER com Augustus e SNAP apenas genes inteiros. Conjunto de Treinamento: An - *Aspergillus nidulans*. En - *Epicoccum nigrum*. (Fonte: FERREIRA, A.J., 2016).

Tabela 10 - Estimativa de diferentes preditores e evidências com diferentes conjuntos de treinamentos.

Preditor	Evidências	CT	Número de sequências				Tamanho médio (pb)			
			Gene	CDS	5'UTR	3'UTR	Gene	CDS	5'UTR	3'UTR
Augustus (gene parcial)	CT	<i>A. nidulans</i>	11.412	33.217	-	-	1.761,5	545,6	-	-
Augustus (gene parcial)	CT	<i>E. nigrum</i>	7.899	21.700	-	-	1.517,3	490,7	-	-
MAKER (Augustus - gene parcial)	CT, R, P	<i>A. nidulans</i>	10.605	29.381	151	68	1.697,4	563,3	162,0	246,4
MAKER (Augustus - gene parcial)	CT, R, P	<i>E. nigrum</i>	10.605	27.935	64	44	1.521,0	488,3	109,1	179,6
MAKER (Augustus + SNAP - gene parcial)	CT, R, P	<i>A. nidulans</i>	10.339	27.901	74	49	1.582,1	486,8	118,2	187,2
MAKER (Augustus + SNAP - gene parcial)	CT, R, P	<i>E. nigrum</i>	10.606	27.936	65	45	1.521,7	488,3	108,5	176,2
MAKER (Augustus + SNAP - gene inteiro)	CT, R, P	<i>A. nidulans</i>	10.320	27.826	264	62	1.591,0	489,3	92,8	161,9
MAKER (Augustus + SNAP - gene inteiro)	CT, R, P	<i>E. nigrum</i>	10.320	27.826	265	62	1.591,0	489,3	92,5	161,9

Siglas: CT. - Conjunto de treinamento. P - Proteoma de Dothideomycetes, R - mRNA.
(Fonte: FERREIRA, A.J., 2016).

A avaliação da qualidade das predições revelou grande similaridade entre os tratamentos com os conjuntos de treinamento construídos a partir de genes de *E. nigrum* e *A. nidulans* (tabela 11). A avaliação realizada utilizando os índices AED (Annotation Edit Distance) considerou os tratamentos com apenas genes inteiros preditos com MAKER (combinando Augustus e SNAP juntamente com as evidências de mRNA e Proteínas). O tratamento com os conjuntos de treinamento de genes de *A. nidulans* apresentou uma média 0,152101, enquanto, o conjunto de treinamento de genes de *E. nigrum* apresentou uma média de 0,1521. A análise da quantidade de genes por categoria indicou 698 genes idênticos (índice 0) às proteínas de Dothideomycetes nos dois tratamentos. A categoria envolvendo genes de 90 a 99% similares (índice 0-9) ao banco de dados indicou 5.012 genes para o conjunto de treinamento de *E. nigrum* e 5011 para *A. nidulans*. A categoria de 80 a 90% similares (índice de 0,8-0,9) também apresentou valores muito similares, com 1.576 para os conjuntos de *E.*

nigrum e 1.577 *A. nidulans*. Nenhuma proteína foi totalmente diferente das proteínas de Dothideomycetes (índice de 0).

Tabela 11 - Índices de AED indicando a similaridade entre proteínas de Dothideomycetes e os genes inteiros preditos a partir de MAKER combinando Augustus com SNAP.

Conj. Treinamento	Média AED	#proteínas preditas	Número de proteínas de acordo com AED						
			0	0-0,1	0,1-0,2	0,2-0,5	0,5-0,9	0,9-1	1
<i>E. nigrum</i>	0,152100	10320	698	5012	1576	2825	204	5	0
<i>A. nidulans</i>	0,152101	10320	698	5011	1577	2825	204	5	0

(Fonte: FERREIRA, A.J., 2016).

5.7 Anotação genoma *E. nigrum*

5.7.1 Características Gerais

O *pipeline* do MAKER realizou a predição de 10.320 genes codificadores de proteínas a partir da combinação dos resultados de Augustus utilizando como conjunto de treinamento modelos construídos a partir dos genes do próprio *E. nigrum*. As predições de Augustus foram combinadas aos ESTs do transcriptoma e melhoradas com SNAP. Estas predições gênicas após serem inseridas à plataforma EGene2 foram anotadas com sucesso em XML (arquivo de armazenamento de todas informações relacionadas à predição do genoma), permitindo compará-las a diferentes bancos de dados (figura 10). Em uma busca preliminar por similaridade com BLASTp no banco de dados de proteínas não redundantes do GenBank e BLASTp no banco de dados de Swiss-Prot do Uniprot revelou 5.394 resultados diversos e 4.926 classificadas como proteínas hipotéticas. A investigação quanto à ortologia nos bancos de dados de eggNOG retornou resultados para 9.393 proteínas, dos quais 5.718 foram puderam ser relacionadas a termos GO (Gene Ontology).

Estas 9.393 proteínas putativas foram investigadas via EGene2 em bancos de dados mais restritos, revelando 7.653 proteínas classificadas em famílias com domínios funcionais importantes presentes no Interpro. Estas proteínas apresentaram um total de 21.765 sítios e domínios importantes. Além da busca por domínios conservados em Interpro, também foi realizada busca RPS-BLAST com 342 domínios conservados. A busca por domínios transmembranares e de peptídeo sinal com Phobius demonstrou possíveis ocorrências em 3.296 proteínas. Similarmente, também foi realizada busca por hélices transmembranares com a TMHMM, revelando a presença destas hélices em 2.077 proteínas. Além dos domínios transmembranares, também foram encontrados possíveis pontos de clivagem em sequências

peptídicas em 883 proteínas com SignalP. Também foram encontrados 368 pontos de clivagem específicos para realizar a ligação de âncoras de proteína GPI com big-GPI e nenhum detectado com o algoritmo específico de PredGPI. Dentre as proteínas de *E. nigrum* foi verificado quais enzimas estão em contexto de mapas metabólicos com KEGG, sendo encontradas 3.299 possíveis resultados positivos. Outra classe funcional investigada foi a presença de proteínas de transporte com TCDB, às quais apresentaram possíveis evidências em 164 peptídeos.

Além das predições de genes codificadores de proteínas, também foram preditas 2.525 sequências de relacionadas RNAs pelos programas Infernal (2.306), RNAmmer (41) e tRNAscan-SE (178).

Figura 10 - Proteínas com resultados positivos na anotação em relação ao total.

Proteínas confirmadas positivamente de acordo com a análise. A - resultados de BLASTP x nr + BLASTP x Swiss-Prot. B - resultados de ortologia em eggNOG. C - resultados de classificação em GO terms. D - resultados de domínios em Interpro. E - resultados de domínios conservados em RPS-BLAST. F - resultados de domínios transmembranares e sinais peptídicos em Phobius. G - resultados de hélices transmembranares em TMHMM. H - resultados de pontos de clivagem para ancoras de GPI em big-PI. I - resultados de peptídeos com pontos de clivagem com SignalP. J - resultados de enzimas presentes em mapas metabólicos em KEGG. K - resultados de proteínas de transporte em TCDB. (Fonte: FERREIRA, A.J., 2016).

A busca por sequências repetidas seriadamente com TRF revelou a presença de 27.299 destes elementos em *E. nigrum*. Em 245 *contigs* do genoma também foram detectadas 12.700 sequências possivelmente transferidas horizontalmente com Alien_hunter.

5.7.2 Fração do genoma de *E. nigrum* dedicada a diferentes funções

O sequenciamento do genoma de *E. nigrum* revelou que seu tamanho é de aproximadamente 35 Mb. Dessa forma, de acordo com a anotação do genoma a maior parte é composta por genes (42%), dos quais 35% são sequências codificantes (CDSs) e 7% *introns* (figura 11). As sequências em repetição seriada corresponderam a aproximadamente 1% do genoma. Os RNAs não mensageiros corresponderam por menos de 1%, (0,4% - ncRNA, 0,01% - rRNA e 0,04% tRNA). Outros 56% do genoma não foram preditos em nenhuma destas classificações. Entretanto, dentre estas sequências, uma grande parcela do genoma (27,8%) foi identificada por meio do programa Alien_hunter como sendo possivelmente adquirida por horizontal.

Aproximadamente 30% do genoma é composto de proteínas que puderam ser relacionadas ao banco de grupos de ortologia eggNOG. Aproximadamente 12% das proteínas possuem função desconhecida (categoria S) e outros 7% pertencem a funções gerais (categoria R). O terceiro maior valor indicou que aproximadamente 1,3% do genoma é composto por proteínas desempenhando funções em múltiplas categorias. Outra fração do genoma (1,3%) é correspondente a proteínas envolvidas na biossíntese de metabólitos secundários, seu transporte e catabolismo. Em 1% das bases do genoma pertencem a proteínas de sinais de tradução (categoria T) e 0,8 % estão relacionadas ao metabolismo e transporte de lipídios. Outras 19 categorias num total de 4,9% do genoma correspondem a menos de 0,8% do genoma cada. Estas categorias correspondem em termos mais amplos a 1% voltado para armazenamento de informação e processamento (5 categorias), 1,86% metabolismo (6 categorias), e 2,86% dedicados a processos celulares e sinalização (8 categorias) (tabela 12).

Figura 11 - Proporção da anotação em relação ao tamanho do genoma em pares de bases.

Representação do tamanho do genoma de *E. nigrum* em círculo cinza mais interno. Camada intermediária composta pela soma do tamanho das sequências previstas de CDSs, *introns*, sequências repetitivas e RNAs não codificantes. A camada mais externa representa a soma dos tamanhos das sequências gênicas das proteínas de acordo com sua categorização em eggNOG. Categorias: S - proteínas com funções desconhecidas. R - proteínas que possuem funções gerais. Múltiplas categorias - proteínas classificadas em mais de uma categoria. Q - biossíntese de metabólitos secundários, transporte e catabolismo. T - sinais de tradução. I - transporte de lipídio e metabolismo. Outras 19 categorias com menos de 0,8% em relação ao genoma foram relacionadas ao processamento e estoque de informação (A, B, J, K e L); metabolismo (C, E, F, G, H e P); processo e sinalização celular (D, M, N, O, U, W, Y e Z). (Fonte: FERREIRA, A.J., 2016).

Tabela 12 - Proporção de categorias de proteínas e tamanho do genoma de *E. nigrum*.

Categoria geral	Categoria funcional	Código	Tamanho total (pb)	Proporção
armazenamento de informação e processamento	modificação e processamento de RNA	A	75362	0,21%
armazenamento de informação e processamento	dinâmica estrutural e cromatina	B	13917	0,039%
armazenamento de informação e processamento	tradução, estrutura ribossômica e biogênese	J	74436	0,207%
armazenamento de informação e processamento	transcrição	K	90392	0,252%
armazenamento de informação e processamento	replicação, recombinação e reparo	L	105065	0,293%
Metabolismo	produção e conversão de energia	C	109304	0,305%
Metabolismo	transporte de aminoácido e metabolismo	E	170725	0,476%
Metabolismo	transporte de nucleotídeo e metabolismo	F	30496	0,085%
Metabolismo	transporte de carboidrato e metabolismo	G	241870	0,674%
Metabolismo	transporte de coenzima e metabolismo	H	22446	0,063%
Metabolismo	transporte de lipídio e metabolismo	I	276481	0,77%
Metabolismo	transporte de íon inorgânico e metabolismo	P	94456	0,263%
Metabolismo	biossíntese de metabólitos secundários, transporte, catabolismo	Q	463076	1,29%
Pouco caracterizado	funções gerais	R	2770954	7,721%
Pouco caracterizado	função desconhecida	S	4302032	11,987%
processos celulares e sinalização	controle de ciclo celular, divisão celular e particionamento cromossômico	D	46464	0,129%
processos celulares e sinalização	parede celular\membrana/biogênese de envelope	M	24862	0,069%
processos celulares e sinalização	motilidade celular	N	124856	0,348%
processos celulares e sinalização	modificação pós tradução, chaperoninas	O	245677	0,685%
processos celulares e sinalização	mecanismos de sinais de tradução	T	383499	1,069%
processos celulares e sinalização	tráfego celular, secreção e transporte vesicular	U	217909	0,607%
processos celulares e sinalização	mecanismos de defesa	V	170712	0,475%
processos celulares e sinalização	estruturas celulares	W	1515	0,004%
processos celulares e sinalização	estrutura nuclear	Y	3480	0,01%
processos celulares e sinalização	Citoesqueleto	Z	63522	0,177%
Múltiplas categorias	presente em mais de uma categoria	Múltiplas	468605	1,306%

Sigla: pb - pares de bases. (Fonte: FERREIRA, A.J., 2016).

5.7.3 Proporção de proteínas do genoma de *E. nigrum* dedicada a diferentes funções

Além da proporção da ocupação de proteínas em relação ao tamanho do genoma, também foi analisada a proporção em relação ao número de proteínas de acordo com suas funções atribuídas por eggNOG. Diferentemente da análise de ocupação no genoma, aqui as proteínas presentes em mais de uma categoria foram representadas em suas respectivas categorias simultaneamente. Dessa forma, o perfil das análises e proporção de proteínas difere do perfil de análise da proporção em relação à sua ocupação no genoma.

A análise das categorias mais amplas indica que 32% das proteínas do genoma possuem função desconhecida e 17% funções gerais. Aproximadamente 23% das proteínas do genoma de *E. nigrum* são dedicadas ao metabolismo, seguido por 16% dedicado a processos celulares e sinalização. Uma menor fração de 12% é dedicada ao armazenamento de informação e processamento (figura 12).

Figura 12 - Categorias gerais das proteínas de *E. nigrum* por eggNOG.

Proporção de proteínas de acordo com as categorias gerais por eggNOG (Fonte: FERREIRA, A.J., 2016).

A análise abrangendo a quantidade de proteínas seguindo categorias mais específicas indicou que além das categorias de proteínas de funções desconhecidas (32%) ou funções gerais (17%) a terceira maior classificada é composta por proteínas relacionadas à modificação após a tradução e chaperoninas (5,3%) (figura 13). A quarta categoria está relacionada ao metabolismo e transporte de carboidratos (4,62%). Outra grande parcela está envolvida com o metabolismo protéico, como a tradução e regulação de genes relacionados aos mecanismos de sinais de tradução (3,85%), o próprio processo de tradução, estrutura

ribossômica e biogênese (3,78%) e o transporte e metabolismo de aminoácidos (3,45%). Em proporção similar a quantidade de proteínas relacionadas à biossíntese de metabólitos secundários, assim como seu transporte e catabolismo (3,6%) seguida pela produção e conversão de energia (3,43%).

As proteínas relacionadas ao tráfico celular, secreção e transporte vesicular são de aproximadamente 3,39%. O transporte de lipídio e seu metabolismo de 3,24% e a transcrição de 3,05%. Outras 13 categorias com representação inferior 2,5% estão presentes na tabela 13.

Figura 13 - Categorias de proteínas de *E. nigrum* por eggNOG.

Categorias de proteínas: A - modificação e processamento de RNA; B - dinâmica estrutural e cromatina; C - produção e conversão de energia; D - controle de ciclo celular, divisão celular e particionamento cromossômico; E - transporte de aminoácido e metabolismo; F - transporte de nucleotídeo e metabolismo; G - transporte de carboidrato e metabolismo; H - transporte de coenzima e metabolismo; I - transporte de lipídio e metabolismo; J - tradução, estrutura ribossômica e biogênese; K - transcrição; L - replicação, recombinação e reparo; M - parede celular/membrana/biogênese de envelope; N - motilidade celular; O - modificação pós tradução, chaperoninas; P - transporte de íon inorgânico e metabolismo; Q - biossíntese de metabólitos secundários, transporte, catabolismo; R - funções gerais; S - função desconhecida; T - mecanismos de sinais de tradução; U - tráfico celular, secreção e transporte vesicular; V - mecanismos de defesa; W - estruturas celulares; Y - estrutura nuclear; Z - citoesqueleto. (Fonte: FERREIRA, A.J., 2016).

Tabela 13 - Número e proporção de proteínas do genoma de *E. nigrum* dedicada a diferentes funções.

Categoria geral	Categoria funcional	Código	# proteínas	Proporção
armazenamento de informação e processamento	modificação e processamento de RNA	A	1436	2,48%
armazenamento de informação e processamento	dinâmica estrutural e cromatina	B	473	0,82%
metabolismo	produção e conversão de energia	C	1983	3,43%
processos celulares e sinalização	controle de ciclo celular, divisão celular e particionamento cromossômico	D	748	1,29%
metabolismo	transporte de aminoácido e metabolismo	E	1995	3,45%
metabolismo	transporte de nucleotídeo e metabolismo	F	651	1,12%
metabolismo	transporte de carboidrato e metabolismo	G	2675	4,62%
metabolismo	transporte de coenzima e metabolismo	H	621	1,07%
metabolismo	transporte de lipídio e metabolismo	I	1875	3,24%
armazenamento de informação e processamento	tradução, estrutura ribossômica e biogênese	J	2190	3,78%
armazenamento de informação e processamento	transcrição	K	1764	3,05%
armazenamento de informação e processamento	replicação, recombinação e reparo	L	1404	2,43%
processos celulares e sinalização	parede celular\membrana/biogênese de envelope	M	355	0,61%
processos celulares e sinalização	motilidade celular	N	19	0,03%
processos celulares e sinalização	modificação pós tradução, chaperoninas	O	3069	5,30%
metabolismo	transporte de íon inorgânico e metabolismo	P	1184	2,05%
metabolismo	biossíntese de metabólitos secundários, transporte, catabolismo	Q	2085	3,60%
Pouco caracterizado	funções gerais	R	9636	16,65%
Pouco caracterizado	função desconhecida	S	18440	31,86%
processos celulares e sinalização	mecanismos de sinais de tradução	T	2228	3,85%
processos celulares e sinalização	tráfego celular, secreção e transporte vesicular	U	1962	3,39%
processos celulares e sinalização	mecanismos de defesa	V	193	0,33%
processos celulares e sinalização	estruturas celulares	W	23	0,04%
processos celulares e sinalização	estrutura nuclear	Y	124	0,21%
processos celulares e sinalização	citoesqueleto	Z	745	1,29%

proteínas – número de proteínas. (Fonte: FERREIRA, A.J., 2016).

5.7.4 RNAs não codificantes

A busca por RNAs não codificantes revelou 2.036 sequências correspondentes a RNAs diferentes de mensageiros com o preditor Infernal. Dentre estes, a maior parte (1.041) é composta de sequências de tipos variados como similaridade a motivos e regiões conservadas de RNA (motivo radC RNA, motivo mini-ykkC RNA, proteína ribossômica L19, região conservada Six3os1 3, região conservada ST7 antisense RNA 2, região conservada isrG Hfq RNA, etc.), seguida por 861 pequenos RNAs nucleolares (snoRNA), 361 micro RNAs (miRNA), 22 pequenos RNAs nucleares (snRNA), 15 RNAs fita complementar e 6 ribozimas. Além disso, a análise com RNAmmer identificou 41 sequências correspondentes a rRNA. Finalmente, a busca por genes de RNAs transportadores realizada com tRNAscan-SE revelou 178 genes (tabela 14). Os genes para tRNAs apresentaram muitas cópias chegando a 10 cópias para metionina mesmo apesar da pouca variedade de molécula (único anticódon). Os outros tipos de tRNAs somando os diferentes anticódons demonstraram os correspondentes a leucina (17), valina (16) e arginina (15) como os mais abundantes. Entretanto, alguns genes foram menos abundantes como os de tRNAs responsáveis por glutamina (2), prolina (1) e histidina (1).

Tabela 14 - Diversidade de genes de RNAs transportadores em *E. nigrum*.

Aminoácido	anticódon	# genes	aminoácido	anticódon	# genes
Alanina	AGC	7	Leucina	AAG	6
Alanina	CGC	2	Leucina	CAA	3
Alanina	TGC	4	Leucina	CAG	3
Arginina	ACG	7	Leucina	TAA	3
Arginina	CCG	1	Leucina	TAG	2
Arginina	CCT	2	Lisina	CTT	8
Arginina	TCG	3	Lisina	TTT	2
Arginina	TCT	2	Metionina	CAT	10
Asparagina	GTT	7	Fenilalanina	GAA	8
Ácido Aspartico	GTC	9	Prolina	CGG	1
Cisteína	GCA	3	Serina	AGA	5
Ácido Glutâmico	CTC	9	Serina	CGA	2
Ácido Glutâmico	TTC	3	Serina	GCT	3
Glutamina	TTG	2	Serina	TGA	1
Glicina	CCC	1	Treonina	AGT	5
Glicina	GCC	9	Treonina	CGT	2
Glicina	TCC	3	Treonina	TGT	2
Histidina	GTG	1	Triptofano	CCA	4
Isoleucina	AAT	7	Tirosina	GTA	6
Isoleucina	GAT	1	Valina	AAC	8
Isoleucina	TAT	2	Valina	CAC	4
Outro	TCA	1	Valina	TAC	4

genes - número de genes. (Fonte: FERREIRA, A.J., 2016).

5.7.5 Domínios transmembranares e peptídeos sinal

A análise de domínios transmembranares com TMHMM indicou a presença destes domínios em 2.077 proteínas. Similarmente, a análise de proteínas com Phobius apontou para 2.060 proteínas contendo possíveis hélices transmembranares. Deste total, 681 delas teriam pelo menos uma hélice e 3 poderiam conter até 20 destas estruturas (figura 14).

Figura 14 - Distribuição de proteínas transmembranares.

Relação da quantidade de proteínas e a quantidade de hélices transmembranares.

Além das proteínas com possíveis domínios transmembranares também foram analisadas as proteínas com peptídeo sinal. A análise com Phobius indicou a presença de 1.236 proteínas contendo sinal peptídico. Entretanto, o número de proteínas preditas com SignalP foi menor (883).

5.8 Relações filogenéticas

5.8.1 Determinação das relações filogenéticas de *E. nigrum*

O teste de suporte *SH-like* indicou que das 40 árvores filogenéticas obtidas pelo protocolo MIRLO, 7 apresentaram maiores valores médios em seus braços, sendo consideradas bons sinais filogenéticos. A identificação por meio de BLASTx dos genes de cada uma dessas árvores revelou que possuem similaridade principalmente com genes de metabolismo primário de fungos (tabela 15).

Tabela 15 - Identificação dos sete genes melhores genes usados para filogenia com a descrição dos seus principais domínios.

Descrição	Cob.	E-value	Identif.	Acesso NCBI	Principal domínio	Acesso Pfam
proteína de controle de divisão celular	100%	0.0	71%	XP_003000093.1	Proteína quinase	pfam00069
helicase do C terminal	97%	0.0	73%	XP_009655311.1	Helicase conservada	pfam00271
histona-lisina N-metiltransferase	96%	0.0	70%	XP_003000946.1	Lisina metiltransferase.	pfam00856
família metalopeptidase M24	100%	0.0	72%	EQB51975.1	Metalopeptidases	pfam00557
metilenotetrahidrofolato redutase 1	99%	0.0	83%	ENH80777.1	Metilenotetrahidrofolato redutase	pfam02219
proteína da família mrs7	98%	0.0	74%	XP_007276203.1	Proteína LETM1(membrana mitocondrial interna)	pfam07766
canal de cloreto	99%	0.0	72%	EQB55442.1	Canais de íons com 10 ou 12 hélices transmembranares	pfam00654

Síglas: Cob. - cobertura, Identif. – identificação. (Fonte: FERREIRA, A.J., 2016).

O filo Ascomycota compreende uma grande diversidade de fungos, sendo esta diversidade grande principalmente em Dothideomycetes, Eurotiomycetes e Sordariomycetes. A utilização de sete genes concatenados utilizados em uma árvore filogenética para as 42 espécies permitiu relacionar *E. nigrum* com outros fungos (figura 15) (tabela 16). Estas análises de relação filogenética puderam ser realizadas em um contexto mais amplo evolutivamente por meio de uma busca mais profunda dos genomas inteiros e não apenas genes tradicionalmente utilizados. Na árvore pôde ser observado que as estimativas de tempo de divergência são indicadas pelas faixas etárias representadas em blocos cinza. A idade destes grupos taxonômicos relevantes foi calculada de acordo com registros fósseis. Os resultados indicam Dothideomycetes como um dos grupos que mais divergiram recentemente na história evolutiva dos fungos neste estudo, tendo *E. nigrum* divergido há aproximadamente 100 milhões de anos. Apesar de comportamento endofítico, o fungo *E. nigrum* está na ordem Pleosporales juntamente com diversos fungos patogênicos, não formando grupos de acordo com o estilo de vida.

Figura 15 - Filogenia e divergência dos tempos estimados de Dikarya com base em sequência de sete genes codificadores de proteínas.

Espécies de fungos de pertencentes à Dikarya com genoma sequenciado. As faixas verticais cinzas indicam as idades mínimas e máximas para Dothideomycetes e Sordariomycetes. Em alguns casos, as linhas horizontais relevantes foram estilisticamente estendidas para destacar rótulos de nodos. Os principais níveis taxonômicos das espécies estão destacados. Símbolos ao lado das espécies indicam seu principal estilo de vida: losango, patogênicos; triângulo, micorrízicos; quadrado, saprofíticos; círculo, endofíticos; pentágono,

endossimbiótico. Siglas: Aar (*Aaosphaeria arxii*), Aal (*Acremonium alcalophilum*), Aat (*Alternaria alternata*), Amn (*Amanita muscaria*), Aol (*Arthrobotrys oligospora*), Asa (*Ascocoryne sarcoides*), And (*Aspergillus nidulans*), Ang (*Aspergillus niger*), Apu (*Aureobasidium pullulans*), Cat (*Candida tenuis*), Cca (*Cochliobolus carbonum*), Che (*C. heterostrophus*), Cmi (*C. miyabeanus*), Cvi (*C. victoriae*), Cgr (*C. graminicola*), Chi (*C. higginsianum*), Dex (*Didymella exigua*), Dse (*Dothistroma septosporum*), Eco (*Epichloe coenophiala*), Efe (*Epichloe festucae*), Eni (*Epicoccum nigrum*), Ela (*Eutypa lata*), Fox (*Fusarium oxysporum*), Hor (*Harpophora oryzae*), Lbi (*Laccaria bicolor*), Lma (*Leptosphaeria maculans*), Mha (*Monacrosporium haptotylum*), Mco (*Morchella conica*), Ncr (*Neurospora crassa*), Oma (*Oidiodendron maius*), Pbr (*Paracoccidioides brasiliensis*), Pgl (*Penicillium glabrum*), Pst (*Pichia stipitis*), Pin (*Piriformospora indica*), Pte (*Pyrenophora teres*), Rru (*Rhystidhysterion rufulum*), Sve (*Sebacina vermifera*), Stu (*Setosphaeria turcica*), Sno (*Stagonospora nodorum*), Tat (*Trichoderma atroviride*), Tvi (*Trichoderma virens*) e Tme (*Tuber melanosporum*). (Fonte: FERREIRA, A.J., 2016).

Os membros da classe Dothideomycetes teriam compartilhado ancestrais em comum há no máximo 427 milhões de anos. Os resultados também indicam que os genomas da classe sofreram diversificação na ordem Pleosporales, a qual *E. nigrum* pertence. Os membros pertencentes a esta ordem, teriam se originado há aproximadamente 200 milhões de anos.

Tabela 16 - Proteomas de espécies de fungos utilizados para a análise de MCL.

Código	Espécie	Principal estilo de vida	Observação	Referência
Aar	<i>Aaosphaeria arxii</i>	desconhecido	isolado inicialmente de milho	(JGI - Aaosphaeria arxii, 2016)
Aal	<i>Acremonium alcalophilum</i>	Saprofítico	isolado a partir de dejetos de suínos	(JGI - Acremonium alcalophilum, 2016)
Aat	<i>Alternaria alternata</i>	Patogênico	patogênico para varias plantas	(JGI - Alternaria alternata , 2016)
Amn	<i>Amanita muscaria</i>	Saprofítico	amplamente distribuído, crescendo sobre madeiras e coníferas	(KOHLE et al., 2015)
Aol	<i>Arthrobotrys oligospora</i>	Saprofítico	saprofítico ou infectante de nematóides	(YANG et al., 2011)
Asa	<i>Ascocoryne sarcoides</i>	Endofítico	endófito de <i>Eucryphia cordifolia</i> saprofítico responsável pela degradação de vários tipos de substâncias	(GIANOULIS et al., 2012)
And	<i>Aspergillus nidulans</i>	Saprofítico	degradação de lignina	(ARNAUD et al., 2012)
Ang	<i>Aspergillus niger</i>	Saprofítico	degradação de lignina	(ANDERSEN et al., 2011)
Apu	<i>Aureobasidium pullulans</i>	Endofítico	biocontrole e infecções oportunistas em humanos eficiente fermentador de açúcares como xilose, podendo atuar como simbiote	(GOSTINCAR et al., 2014)
Cat	<i>Candida tenuis</i>	Saprofítico	simbiote	(JAAFAR; PETTIT et al., 1992)
Cca	<i>Cochliobolus carbonum</i>	Patogênico	patógeno de vários tipos de cereais	(CONDON et al., 2013)
Che	<i>Cochliobolus heterostrophus</i>	Patogênico	causador da doença da folha do milho (SCLB)	(OHM et al., 2012)
Cmi	<i>Cochliobolus miyabeanus</i>	Patogênico	utiliza fitotoxinas para desencadear a morte celular hospedeira	(CONDON et al., 2013)
Cvi	<i>Cochliobolus victoriae</i>	Patogênico	gênero pertencente a várias espécies de patógenos	(CONDON et al., 2013)
Cgr	<i>Colletotrichum graminicola</i>	Patogênico	causa antracnose doença na folha do milho	(O'CONNELL et al., 2012)
Chi	<i>Colletotrichum higginsianum</i>	Patogênico	causa antracnose em Brassica e Raphanus	(O'CONNELL et al., 2012)
Dex	<i>Didymella exigua</i>	Patogênico	patógeno de várias plantas	(AVESKAMP et al., 2010)
Dse	<i>Dothistroma septosporum</i>	Patogênico	patógeno de pinus	DE WIT et al., 2012.
Eco	<i>Epichloe coenophiala</i>	Endofítico	associado a doenças de citrus	(YOUNG, et al.; 2014)
Efe	<i>Epichloe festucae</i>	Endofítico	proteção de herbivoria	(SAIKIA et al., 2012)
Eni	<i>Epicoccum nigrum</i>	Endofítico	isolado de varias plantas, dentre elas cana de açúcar	(FAVARO et al., 2012)
Ela	<i>Eutypa lata</i>	Patogênico	causa morte progressiva de videiras	(BLANCO-ULATE et al., 2010)
Fox	<i>Fusarium oxysporum</i>	Patogênico	fitopatógeno comum em mais de 100 cultivares	(MA et al., 2010)
Hor	<i>Harpophora oryzae</i>	Endofítico	endofítico de arroz	(XU. et al., 2014)
Lbi	<i>Laccaria bicolor</i>	Micorrízico	simbiote de árvores de álamo	(MARTIN et al., 2008)

continua...

Tabela 16 - Proteomas de espécies de fungos utilizados para a análise de MCL. (continuação)

Código	Espécie	Principal estilo de vida	Observação	Referência
Lma	<i>Leptosphaeria maculans</i>	patogênico	patógeno de cultura de biocombustíveis (<i>Brassica napus</i>)	ROUXEL et al., 2011.
Mha	<i>Monacrosporium haptotylum</i>	saprofítico	saprofítico ou infectante de nematóides	(MEERUPATI et al., 2013)
Mco	<i>Morchella cônica</i>	saprofítico	atua em simbiose com diversas micorrizas	(JGI - <i>Morchella conica</i> , 2016)
Ncr	<i>Neurospora crassa</i>	saprofítico	saprófito amplamente distribuído, possuído de muitos genes em comum com patógenos	(GALAGAN et al., 2003)
Oma	<i>Oidiodendron maius</i>	micorrízico	micorrízicos de plantas ericáceas (<i>Vaccinium myrtillus</i> , <i>Calluna vulgaris</i>)	(KOHLENER et al., 2015)
Pbr	<i>Paracoccidioides brasiliensis</i>	patogênico	fungo patogênico causador de infecção persistente em humanos	(DESJARDINS et al., 2011)
Pgl	<i>Penicillium glabrum</i>	saprofítico	causa podridão pós-colheita de frutas e vegetais	(BASÍLIO et al., 2006)
Pst	<i>Pichia stipitis</i>	endossimbionte	endossimbionte de besouros	(JEFFRIES et al., 2007)
Pin	<i>Piriformospora indica</i>	endofítico	coloniza as células do córtex da raiz de uma ampla gama de espécies de plantas	(ZUCCARO et al., 2011)
Pte	<i>Pyrenophora teres</i>	patogênico	fungo necrotrófico de cevada	(ELLWOOD et al., 2010)
Rru	<i>Rhystidhysterion rufulum</i>	patogênico	associado a doenças de citros	(OHM et al., 2012.
Sve	<i>Sebacina vermifera</i>	micorrízico	ectomicorrízica amplamente distribuída em briófitas e angiospermas	(KOHLENER et al., 2015)
Stu	<i>Setosphaeria túrcica</i>	patogênico	patógeno de monocotiledôneas como o milho	(CONDON et al., 2013)
Sno	<i>Stagonospora nodorum</i>	patogênico	patógeno de trigo e outros cereais	(HANE et al., 2009)
Tat	<i>Trichoderma atroviride</i>	endofítico	biocontrole de <i>Rhizoctonia solani</i> e <i>Botrytis cinerea</i>	(KUBICEK et al., 2011)
Tvi	<i>Trichoderma virens</i>	endofítico	Biocontrole	(KUBICEK et al., 2011)
Tme	<i>Tuber melanosporum</i>	micorrízico	ectomicorrízica de carvalhos e avelã	(MARTIN et al., 2010)

(Fonte: FERREIRA, A.J., 2016).

5.8.2 Alterações nas famílias de proteínas

As análises da estimativa de alterações evolutivas no tamanho das famílias de proteínas usando CAFE foram realizadas com Pleosporales. Observando apenas as famílias de proteínas compartilhadas por pelo menos dois táxons foi observado um número total de 12.146 famílias no genoma de *E. nigrum* com uma média de 1,37 por família de proteínas. Tamanho similar a *Didymella exigua* (1,37), organismo mais próximo. Estes valores também foram as maiores médias para os 12 táxons analisados. Além disso, *Rhystidhysterion rufulum* foi o táxon que obteve a menor média (0,96) (tabela 17 e figura 16). Os resultados da análise com CAFE indicam que 11.662 famílias de genes de *E. nigrum* foram mantidas, enquanto ocorreu o ganho de 484 novas famílias e 963 sofreram contração, deixando de existir no táxon.

Figura 16 - Alterações no número de famílias de proteínas visualizadas na árvore filogenética.

Árvore filogenética construída com PhyML usando os 7 melhores genes concatenados. Os números indicam a porcentagem de famílias de genes mantidos, adquiridos e perdidos durante o processo evolutivo para cada táxon ou ancestral (nó). Siglas: Aat (*Alternaria alternata*), Cca (*Cochliobolus carbonum*), Che (*C. heterostrophus*), Cmi (*C. miyabeanus*), Cvi (*C. victoriae*), Dex (*Didymella exigua*), Eni (*Epicoccum nigrum*), Lma (*Leptosphaeria maculans*), Pte (*Pyrenophora teres*), Rru (*Rhystidhysterum rufulum*), Stu (*Setosphaeria turcica*) e Sno (*Stagonospora nodorum*). (Fonte: FERREIRA, A.J., 2016).

Tabela 17 - Alteração no número de famílias de proteínas de acordo com os nós da árvore filogenética.

Nó	Expansão média	# expansão	# mantidas	# decréscimo	# proteínas compartilhadas	Média fam./prot.
Aat	0,049203	930	11553	626	9818	1,27144
Pte	-0,031276	497	11107	1505	9043	1,283202
Cca	-0,015562	152	12584	370	10625	1,198682
Cvi	0,002212	254	12574	281	10696	1,199327
Cmi	-0,080555	187	11692	1230	10092	1,177071
Che	-0,045465	403	11625	1081	10357	1,16134
Stu	-0,142116	439	10161	2509	9228	1,148678
Lma	-0,351362	444	7722	4943	8047	1,014788
Sno	-0,261347	631	8336	4142	8737	1,026325
Dex	-0,002899	409	11800	899	8757	1,394199
Eni	-0,023648	484	11662	963	8822	1,376785
Rru	-0,296132	1086	6528	5495	7873	0,967103
NOD0	-0,197193	1286	7379	4444	-	-
NOD19	-0,022656	3	12815	290	-	-
NOD3	-0,158746	36	10955	2118	-	-
NOD13	-0,063773	100	12064	945	-	-
NOD7	-0,014646	142	12627	338	-	-
NOD9	-0,013731	5	12919	184	-	-
NOD11	-0,00595	292	12395	421	-	-
NOD5	-0,020825	97	12617	395	-	-
NOD15	0	0	13109	0	-	-
NOD17	-0,027309	3	12758	348	-	-
NOD21	-0,244946	201	9367	3541	-	-
NOD1	-0,004577	1	13047	61	-	-

Siglas: fam. família, prot. proteína, Aat (*Alternaria alternata*), Cca (*Cochliobolus carbonum*), Che (*C. heterostrophus*), Cmi (*C. miyabeanus*), Cvi (*C. victoriae*), Dex (*Didymella exigua*), Eni (*Epicoccum nigrum*), Lma (*Leptosphaeria maculans*), Pte (*Pyrenophora teres*), Rru (*Rhystidhysterion rufulum*), Stu (*Setosphaeria turcica*) e Sno (*Stagonospora nodorum*). (Fonte: FERREIRA, A.J., 2016).

5.9 Clusters de genes de metabólitos secundários

5.9.1 Cluster de genes de metabólitos secundários de *E. nigrum*

A análise de metabólitos secundários realizada por meio do Antismash indicou a presença de 38 *clusters* (tabela 18). Dentre eles 13 envolvidos na produção de peptídeos não ribossômicos (NRPS), 5 de terpenos, 2 envolvidos na produção de indol, 1 envolvido tanto na produção de indol quanto de NRPS (Indol-NRPS), 2 envolvidos tanto na produção de policetídeos tipo 1 quanto de NRPS (T1PKS-NRPS), 9 de policetídeos tipo 1 (T1PKS), 1 de policetídeos tipo 3 (T3PKS), 1 de Lantipeptídeo e 4 de outros metabólitos (*cluster* contendo uma proteína relacionada com metabólito secundário que não se encaixa em nenhuma outra

categoria). A categoria de *clusters* “outros” significa que os *clusters* contém alguma proteína relacionada a metabólito secundário, não se encaixando em nenhuma categoria específica. Dentre os *clusters* presentes nesta categoria não foram encontrados genes que pudessem ser relacionados diretamente à metabólitos que podem ser identificados. Alguns genes desses *clusters* foram identificados apenas como pertencentes a proteínas transportadoras e de transferência de grupos químicos como metiltransferases, dehidrogenases, monofosfatases, sintetases, oxidoredutases, permeases, dentre outras (tabela 19). Outros *clusters* (6, 29, 30, 32, 29, 33, 34, 35, 36 e 38) apesar de incompletos puderam ser classificados, sendo a estrutura estimada realizada parcialmente. Foi possível determinar a estrutura dos *clusters* 6, 30, 33 e 34 de NRPS (figura 17). Parte das estruturas dos metabólitos relacionados aos *clusters* 5, 26 30 e 33 foi idêntica. Além disso, os metabólitos dos *clusters* de NRPS foram os que tiveram mais estruturas inferidas, enquanto nenhum de terpeno pode ser determinado e somente dois de policetídeo tipo1 foi determinado (*clusters* 16 e 25).

As estruturas foram comparadas com a base de dados do ChemSpider (<http://www.chemspider.com/>), porém a baixa similaridade com os compostos químicos não permitiu determinar a identidade quando comparado a compostos conhecido.

Tabela 18 - *Clusters* de genes de metabólitos secundários presentes em *E. nigrum*.

Cluster	Tipo	posição (# de bases)	# genes	Contig
Cluster 1	Outro	326619	369762	50
Cluster 2	Lantipeptídeo	423895	467747	52
Cluster 3	T1pks	34468	83596	9
Cluster 4	Terpeno	238065	259145	14
Cluster 5	NRPS	174654	236332	55
Cluster 6	T1PKS-NRPS	1	40041	27
Cluster 7	T3PKS	288229	329361	28
Cluster 8	T1PKS	243105	290847	55
Cluster 9	T1PKS	110813	155798	44
Cluster 10	Outro	510	43372	52
Cluster 11	T1PKS	129527	175621	13
Cluster 12	Indol	169449	190609	31
Cluster 13	NRPS-T1PKS	219666	265101	42
Cluster 14	T1PKS	19370	68905	47
Cluster 15	NRPS	39856	82900	54
Cluster 16	T1PKS	66724	122707	49
Cluster 17	Indol	145908	171916	33
Cluster 18	Outro	84402	128526	53
Cluster 19	T1PKS	83023	125825	41
Cluster 20	Terpeno	60477	82337	25
Cluster 21	NRPS	43885	100347	52
Cluster 22	Terpeno	22244	43040	24
Cluster 23	NRPS	25657	90620	66
Cluster 24	T1PKS	25742	72070	13
Cluster 25	T1PKS	17511	69477	54
Cluster 26	NRPS	32300	74881	49
Cluster 27	Outro	8416	54514	41
Cluster 28	Indol-NRPS	4434	53416	15
Cluster 29	NRPS	1	26317	55
Cluster 30	NRPS	1	25899	25
Cluster 31	Terpeno	5580	24895	19
Cluster 32	Terpeno	1	21602	35
Cluster 33	NRPS	1	21501	11
Cluster 34	NRPS	1	7922	1
Cluster 35	NRPS	1	5348	5
Cluster 36	NRPS	1	2353	1
Cluster 37	NRPS	1	2081	1
Cluster 38	NRPS	1	1657	1

(Fonte: FERREIRA, A.J., 2016).

Tabela 19 - Identidade de genes pertencentes à categoria Outros de *clusters* de metabólitos secundários de *E. nigrum* de acordo com Antismash.

Código COG	Descrição	Gene	Cluster
SMCOG1089	Metiltransferase	ctg4_395	cluster 1
SMCOG1001	dehidrogenase de cadeia curta/redutase SDR	ctg4_396	cluster 1
SMCOG1040	álcool dehidrogenase	ctg4_405	cluster 1
SMCOG1002	sintetase e ligase dependente de AMP	ctg4_420	cluster 1
SMCOG1002	sintetase e ligase dependente de AMP	ctg4_421	cluster 1
SMCOG1072	Dehidrogenase	ctg24_16	cluster 10
SMCOG1002	sintetase e ligase dependente de AMP	ctg24_31	cluster 10
SMCOG1002	sintetase e ligase dependente de AMP	ctg24_32	cluster 10
SMCOG1277	Proteína com domínio SAM	ctg24_45	cluster 10
SMCOG1169	Proteína transportadora de carboidrato	ctg24_49	cluster 10
SMCOG1169	Proteína transportadora de carboidrato	ctg24_50	cluster 10
SMCOG1169	Proteína transportadora de carboidrato	ctg24_51	cluster 10
SMCOG1169	Proteína transportadora de carboidrato	ctg24_52	cluster 10
SMCOG1175	oxidoreductase disulfato nucleotídeo piridina	ctg46_50	cluster 15
SMCOG1002	sintetase e ligase dependente de AMP	ctg46_68	cluster 15
SMCOG1002	sintetase e ligase dependente de AMP	ctg46_69	cluster 15
SMCOG1207	inositol monofosfatase	ctg46_78	cluster 15
SMCOG1002	sintetase e ligase dependente de AMP	ctg71_124	cluster 18
SMCOG1173	Proteína contendo repetição WD-40	ctg71_126	cluster 18
SMCOG1038	permease específica de fenilalanina	ctg71_144	cluster 18
SMCOG1038	permease específica de fenilalanina	ctg71_146	cluster 18
SMCOG1030	serina/proteína quinase treonina	ctg71_150	cluster 18
SMCOG1002	sintetase e ligase dependente de AMP	ctg149_30	cluster 27
SMCOG1002	sintetase e ligase dependente de AMP	ctg149_31	cluster 27
SMCOG1002	sintetase e ligase dependente de AMP	ctg149_36	cluster 27
SMCOG1106	maior facilitador de transporte	ctg149_39	cluster 27
SMCOG1106	maior facilitador de transporte	ctg149_40	cluster 27

(Fonte: FERREIRA, A.J., 2016).

Figura 17 - Estrutura molecular presuntivada compostos secundários produzidos por *E. nigrum*, baseado na análise do genoma pelo programa AntiSmash.

A- cluster 5, (aromático hidrofóbico). B - cluster 6. C - cluster 13. D - cluster 21 (hidrofílico). E - cluster 23 (hidrofóbico alifático). F - cluster 26 (hidrofóbico alifático). G - cluster 30 (hidrofóbico alifático). H - cluster 33 (hidrofóbico alifático). I - cluster 34 (hidrofóbico alifático). J cluster 28. K - cluster 16, PKsT1. L - cluster 25, PKsT1.(Fonte: FERREIRA, A.J., 2016).

5.9.2 Comparação entre clusters de metabólitos secundários

A análise de comparação entre clusters de metabólitos secundários utilizou o programa Synteny Clusters. Foi realizada a comparação entre os clusters de *E. nigrum* e outras 12 espécies de fungos próximos (*Aaosphaeria arxii*, *Alternaria alternata*, *Cochliobolus carbonum*, *C. heterostrophus*, *C. miyabeanus*, *C. victoriae*, *Didymella exigua*, *Leptosphaeria maculans*, *Pyrenophora teres*, *Rhystidhysterion rufulum*, *Setosphaeria turcica* e *Stagonospora*

nodorum). A comparação das 13 espécies permitiu um alto número de combinações de *clusters* (434). Os *clusters* de metabólitos secundários similares entre si foram agrupados e considerados como uma unidade no conjunto. Similarmente, os *clusters* que não puderam ser agrupados com nenhum outro (exclusivos) foram considerados outras unidades (cada um com um *cluster*). Os resultados indicaram que dos 434 conjuntos formados, 67 estão compostos por apenas um *cluster* cada. Entre os conjuntos que foram compartilhados por mais espécies, seis foram compostos por até dez espécies. Todos estes dez *clusters* estão presentes também em *E. nigrum* (tabela 20). Entretanto, entre alguns *clusters* largamente compartilhados (8 e 9 espécies), oito (Conj_cluster_53, Conj_cluster_91, Conj_cluster_164, Conj_cluster_174, Conj_cluster_313, Conj_cluster_389, Conj_cluster_394 e Conj_cluster_432) não foram detectados em *E. nigrum*, sugerindo que os metabólitos resultantes destes *clusters* possam participar de processos patogênicos. Dessa forma, os genes desses oito conjuntos de *clusters* foram comparados a PHI-base (tabela 21).

Tabela 20 - *Clusters* compartilhados por mais espécies de fungos.

Conjunto	Espécies presentes													Total
	Aar	Aal	Cca	Che	Cmi	Cvi	Dex	Eni	Lma	Pte	Rru	Stu	Sno	
Conj_cluster_123														10
Conj_cluster_205														10
Conj_cluster_319														10
Conj_cluster_337														10
Conj_cluster_341														10
Conj_cluster_43														10
Conj_cluster_164														9
Conj_cluster_248														9
Conj_cluster_266														9
Conj_cluster_321														9
Conj_cluster_362														9
Conj_cluster_394														9
Conj_cluster_53														9
Conj_cluster_114														8
Conj_cluster_134														8
Conj_cluster_137														8
Conj_cluster_163														8
Conj_cluster_173														8
Conj_cluster_174														8
Conj_cluster_199														8

Continua

Tabela 20 - *Clusters* compartilhados por mais espécies de fungos. (*continuação*)

Conjunto	Espécies presentes													Total
	Aar	Aal	Cca	Che	Cmi	Cvi	Dex	Eni	Lma	Pte	Rru	Stu	Sno	
Conj_cluster_200														8
Conj_cluster_214														8
Conj_cluster_219														8
Conj_cluster_221														8
Conj_cluster_234														8
Conj_cluster_249														8
Conj_cluster_252														8
Conj_cluster_261														8
Conj_cluster_272														8
Conj_cluster_278														8
Conj_cluster_300														8
Conj_cluster_31														8
Conj_cluster_313														8
Conj_cluster_345														8
Conj_cluster_371														8
Conj_cluster_377														8
Conj_cluster_389														8
Conj_cluster_39														8
Conj_cluster_411														8
Conj_cluster_420														8
Conj_cluster_425														8
Conj_cluster_432														8
Conj_cluster_5														8
Conj_cluster_57														8
Conj_cluster_61														8
Conj_cluster_73														8
Conj_cluster_8														8
Conj_cluster_91														8

Siglas: Aar, *Aaosphaeria arxii*; Aal, *Alternaria alternata*; Cca, *Cochliobolus carbonum*; Che, *C. heterostrophus*; Cmi, *Cochliobolus miyabeanus*; Cvi, *Cochliobolus victoriae*; Dex, *Didymella exigua*; Eni, *Epicoccum nigrum*; Lma, *Leptosphaeria maculans*; Pte, *Pyrenophora teres*; Rru, *Rhystidhysterion rufulum*; Stu, *Setosphaeria turcica*; Sno, *Stagonospora nodorum*. Os *Clusters* com genes envolvidos em processos fitopatogênicos estão em destaque com cor mais escura (Fonte: FERREIRA, A.J., 2016).

A análise revelou que dos oito *clusters*, sete codificam proteínas envolvidas com fitopatogênese. Entretanto, apenas algumas estruturas dos compostos secundários puderam ser inferidas. Os metabólitos secundários Conj_cluster_313, Conj_cluster_174 e Conj_cluster_53 estão envolvidos na produção de metabólitos secundários PKS tipo 1, porém sua estrutura não pôde ser determinada. Além disso, não pôde ser inferida a estrutura dos metabólitos secundários relacionados ao Conj_cluster_389 (tipo indol). Alguns outros puderam ser

relacionados a estruturas parciais de metabólitos secundários (figura 18). A molécula de *clusters* do conjunto Conj_cluster_91 possui maior similaridade com o peptídeo não ribossômico Brevianamida F (Norine ID: NOR00303). Estes compostos pertencem à família da brevianamida e em fungos são descritos como precursores de compostos com atividade tóxica (fumitremorginas A, B e C e tryprostatina B) (MAIYA et al., 2006). Outro metabólito resultante a partir dos *clusters* do Conj_cluster_394 demonstrou maior similaridade com o peptídeo não ribossômico himenamida D (Norine ID: NOR00540). Estes compostos pertencem à família da himenamida e possuem atividade antibiótica. O metabólito relacionado aos *clusters* do conjunto Conj_cluster_432 (origem NRPS-T1PKS) também está relacionado à atividade de antibiótico. A molécula apresentou maior similaridade com o tripeptídeo fosfinotricina (Norine ID: NOR00670). Dentre as proteínas que apresentaram maior similaridade e estrutura, destacam-se as presentes nos *clusters* compartilhados pelo conjunto Conj_cluster_432. O conjunto Conj_cluster_432 ligado à produção do antibiótico tripeptídeo fosfinotricina foi o único presente no fitopatógeno *D. exigua* e ausente no endófito *E. nigrum*.

Tabela 21 - Similaridade às proteínas com relação de patogenicidade do banco de dados PHI-base.

Grupo	Prot. Rep.	Cluster	Proteína PHI	Ident.	E-value	Score	Relação
Conj_cluster_53	ctg36_orf8	cluster013	PHI_255_FUM1	55%	1E-159	503	Patogenicidade mantida
Conj_cluster_53	ctg36_orf9	cluster013	PHI_255_FUM1	43%	2E-171	544	Patogenicidade mantida
Conj_cluster_91	ctg61_allorf103	cluster041	PHI_2511_Pes1	29%	5E-140	472	Redução de virulência
Conj_cluster_91	ctg61_orf003	cluster041	PHI_2511_Pes1	34%	7E-75	273	Redução de virulência
Conj_cluster_174	ctg12_orf94	cluster016	PHI_55_PKS1	31%	0	1016	Redução de virulência
Conj_cluster_313	ctg34_orf15	cluster031	PHI_1046_CTB5	28%	9E-33	125	Redução de virulência
Conj_cluster_313	ctg34_orf24	cluster031	PHI_3162_G4MXJ1	52%	3E-34	134	Perda da patogenicidade
Conj_cluster_389	ctg29_orf98	cluster028	PHI_922_um03615	26%	1E-32	128	Patogenicidade mantida
Conj_cluster_389	ctg29_orf100	cluster028	PHI_2378_DEP4	32%	8E-77	254	Resultado misto
Conj_cluster_394	ctg31_orf10	cluster029	PHI_267_MLT1	38%	7E-31	127	Redução de virulência
Conj_cluster_394	ctg31_orf11	cluster029	PHI_2511_Pes1	40%	0	1867	Redução de virulência
Conj_cluster_394	ctg31_orf15	cluster029	PHI_438_BcBOT1	37%	2E-24	99,4	Redução de virulência
Conj_cluster_432	ctg8_orf174	cluster014	PHI_1047_CTB6	36%	1E-38	140	Redução de virulência
Conj_cluster_432	ctg8_orf176	cluster014	PHI_255_FUM1	65%	0	1295	Patogenicidade mantida

Siglas: Prot. Rep., Proteína Representante; Ident., Identidade. (Fonte: FERREIRA, A.J., 2016).

Figura 18 - Estrutura parcial de metabólitos secundários envolvidos em processos fitopatogênicos.

Em A. *clusters* tipo NRPS presentes em Conj_cluster_91, similares à Brevianamida F, possuem atividade tóxica. Em B. *clusters* tipo NRPS presentes em Conj_cluster_394, similares à himenamida D, possuem atividade antibiótica. Em C. *clusters* tipo NRPS e PKS tipo 1 presentes em Conj_cluster_432, similares ao tripeptídeo fosfinotricina, possui atividade antibiótica.

Os fungos *E. nigrum* e *Didymella exigua* compartilham 8 *clusters* de metabólitos secundários. Entretanto, apesar de próximas filogeneticamente foi observada uma grande distância entre quanto ao compartilhamento de seus *clusters* (figura 19). A análise indicou que *E. nigrum* compartilha a maior parte de seus *clusters* com espécies patogênicas. O fungo *E. nigrum* compartilha mais *clusters* com *Stagonospora nodorum* (14 *clusters*) (figura 20 – A). A análise também indica que o número de compartilhamento entre *E. nigrum* e *D. exigua* ou *E. nigrum* e *S. nodorum* aumenta ou reduz de acordo com a proximidade das espécies utilizadas (figura 20 – B - C). A análise utilizando 3 espécies resalta que *E. nigrum* possui perfil metabólico mais similar a *S. nodorum*, sendo intermediário entre *S. nodorum* e *D. exigua* (figura 20 – E).

Figura 19 - Dendrogramas com *clusters* compartilhados entre diferentes espécies.

Proximidade de espécies de acordo com a presença de *clusters* de metabólitos secundários compartilhados.

Figura 20 - Diagramas de Venn com *clusters* compartilhados entre diferentes espécies.

Aar - *Aaosphaeria arxii*, Eni - *Epicoccum nigrum*, Rru - *Rhystidhysterum rufulum*, Sno - *Stagonospora nodorum*, Dex - *Didymella exigua*. **A** - compartilhamento entre *A. arxii*, *D. exigua*, *E. nigrum* e *S. nodorum*. **B** - *A. arxii*, *D. exigua*, *E. nigrum* e *R. rufulum*. **C** - *A. arxii*, *E. nigrum*, *R. rufulum* e *S. nodorum*. **D** - *A. arxii*, *D. exigua* e *E. nigrum*. **E** - *D. exigua*, *E. nigrum* e *S. nodorum*. **F** - *A. arxii*, *E. nigrum* e *S. nodorum*. (Fonte: FERREIRA, A.J., 2016).

6 DISCUSSÃO

A criação de componentes integrados à plataforma envolveu uma vasta coordenação na execução de programas e reunião de informações. A diversidade de processos teve que ser adaptada aos diferentes formatos de entrada e saída, alguns dos quais constantemente renovados. Estes programas possuem algoritmos e abordagens variadas de acordo com seu objetivo, indo desde predição de genes codificadores de proteínas (Glimmer - DELCHER et al., 2007; Augustus - STANKE et al., 2004; FGENESH - SOLOVYEV et al., 2006), preditores de tRNA e RNAs não codificadores (tRNAscan-SE - SCHATTNER et al., 2005; promotores Promoter (KNUDSEN, 1999), dentre outras regiões. Além disso, estes modelos preditivos são confirmados e relacionados aos bancos de dados (NCBI - <http://www.ncbi.nlm.nih.gov/>; KEGG - NAKAYA et al., 2013; eggNOG - POWELL et al., 2014) com bases em expansão a cada dia. Apesar do crescente aumento do número de ferramentas para caracterizar e identificar as sequências de DNA há um grande trabalho que requer tempo para o processamento e interpretação dos dados. Esta diversidade de programas emite arquivos nos mais variados formatos de saída, havendo a necessidade de processamento manual ou utilização de ferramentas de conversão e comparação. Entretanto, muitas dessas ferramentas não são dominadas por todos usuários.

A anotação manual é bem-sucedido em análises de genes pontuais, porém sua aplicação em projetos genômicos e transcriptômicos torna-se inviável. Assim, a anotação manual é mais prática na curadoria das informações levantadas pela anotação automática. Neste contexto, os *pipelines* são plataformas poderosas capazes de interligar diferentes programas de predição e bancos de dados, reduzindo enormemente o tempo de anotação de um genoma. Neste trabalho foram desenvolvidos componentes em linguagem Perl capazes de se integrar a plataforma de anotação EGene2 (versão atual do EGene - DURHAM et. al., 2005). Dentre os componentes desenvolvidos, o componente para o Augustus mostrou-se eficaz, sendo este software considerado um dos melhores preditores descritos para a predição de genes eucarióticos. A partir da confecção do componente do Augustus foi possível integrar as predições de *E. nigrum* e integrar ao ambiente EGene2 de forma automática e permitindo acesso a outras ferramentas pré-existentes poderosas, assim como bases de dados utilizadas pela plataforma. Dessa forma, a reunião dos componentes permitiu uma anotação ajustada às necessidades de do projeto *E. nigrum*.

Uma plataforma de anotação requer a entrada de dados com qualidade para iniciar o processo. Assim, a determinação das fronteiras dos genes é influenciada pela utilização de

conjuntos de treinamento. Os conjuntos de treinamento requerem 200 genes no mínimo para alcançarem resultados com melhores desempenhos. Entretanto esta quantidade pode chegar a 1000 genes dependendo da complexidade do organismo. Além disso, a quantidade de genes com mais de um *exon* também é importante, pois adiciona variações de modelos que são importantes para criação das matrizes de cálculo de *introns*. Os modelos toleram certa inexatidão, mas, é necessário que os genes estejam inteiros (STANKE, 2015). Neste trabalho foram utilizados 420 genes curados, valor dobro do mínimo recomendado. Além disso, os testes para medir a qualidade da predição foram bem-sucedidos, pois o conjunto de treinamento construído a partir dos genes de *E. nigrum* apresentou uma acurácia 3 vezes superior aos valores mínimos exigidos. Para a consideração do uso de um conjunto de treinamento esta acurácia deve ser superior a 20%. (STANKE, 2015). A confecção do conjunto de treinamento específico para *E. nigrum* permitiu uma melhor predição gênica com o programa Augustus dentro da plataforma MAKER. Esta plataforma melhora a predição, consequentemente as anotações, pois é fortalecida com as evidências experimentais do organismo (CAMPBELL et al., 2014; LAW et al., 2014). Além do transcriptoma de *E. nigrum*, os proteomas de organismos relacionados foram utilizados como evidências. Neste aspecto, o mapeamento foi importante para melhorar a predição dos genes, visto que a predição “*in silico*” pode apresentar vieses (HYATT et al., 2010). A correção destes vieses pelo treinamento dos preditores melhora fronteiras (intergênico/UTRs/*exon/intron*) e permite a detecção de genes não preditos em algumas regiões (HAAS et al., 2003; STANKE et al., 2008). Finalmente, o conhecimento da filogenia dos organismos relacionados é fundamental para reunir evidências para o preditor (CAMPBELL et al., 2014). Neste trabalho a utilização de Augustus para treinamento e predição inicial combinado ao SNAP permitiram que MAKER, juntamente com ExonerateMAKER (CANTAREL et al., 2008) utilizassem ferramentas de combinação de alinhamentos de mRNA e proteínas de organismos relacionados.

A análise comparando as diferentes predições indicou que o conjunto treinamento construído a partir dos genes de *E. nigrum* subestima a quantidade e tamanho médio dos genes e CDSs presentes no genoma. Por outro lado, a quantidade e tamanho médio de genes e CDSs foi superestimada quando utilizado apenas o conjunto de treinamento de *A. nidulans*. Esta hipótese baseada em experimentos subsequentes nos quais utilizam como evidências de proteínas de organismos próximos e as sequências de RNA mensageiro do próprio organismo. Uma possível explicação é a maior densidade de genes em genomas de fungos quando comparados a outros eucariotos como humanos. Sem evidências ou parâmetros de correção os

preditores tendem a fundir genes e consequentemente seus genes também se tornam maiores (HOLT; YANDELL, 2011). Apesar da tendência no aumento no número e tamanho de genes com o conjunto de treinamento de *E. nigrum*, a comparação das CDSs nas análises indicou uma estabilidade no tamanho médio das CDSs com conjunto de treinamento quando adicionadas das evidências. Os dados indicam o conjunto de treinamento construído com os genes de *E. nigrum* possui bom desempenho na detecção das fronteiras de *exon-intron*, visto que o tamanho se manteve estável. Esta estabilidade foi mantida, mesmo após implementação de SNAP, enquanto o conjunto de treinamento construído com os genes de *A. nidulans* necessitou o auxílio de SNAP para tais correções. Dessa forma, os CDSs e fronteiras (*exon-intron*) de *E. nigrum* são diferentes de *A. nidulans*. Em casos de não utilização de um conjunto de treinamento próprio é importante a correção com evidências e o programa SNAP.

Apesar da menor quantidade, as predições das regiões de UTRs também apresentaram a mesma tendência das predições de CDS e genes. Assim, a utilização de conjuntos de treinamento confeccionados a partir de genes de organismos mais próximos também é importante para predição destas regiões. Além disso, a predição considerando apenas genes inteiros apresentou melhores resultados, provavelmente por causa de imprecisões na fronteira na detecção das fronteiras de *start* e *stop codon*. A determinação dessas UTRs é mais bem-sucedida com preditores mais especializados capazes de combinar a estrutura primária com a secundária (FAGHIHI; WAHLESTEDT, 2009). O conhecimento das regiões UTRs também desempenha um papel na expressão alternativa devido a variação nos sítios de *start* e *stop codon* (KIM et al., 2008). Estas hipóteses alternativas na transcrição do gene podem alterar a forma como um gene pode ser anotado, consequentemente explicando algumas diferenças entre as predições e os dados experimentais de sequências protéicas presentes em bancos de dados. Assim, quando avaliados os resultados dos índices de AED, pôde ser considerada normal uma pequena diferença entre as evidências de proteínas e os genes preditos. Os índices de AED melhoram o controle de qualidade da anotação e influenciam os pesquisadores na tomada de decisões (HOLT; YANDELL, 2011).

Os resultados da anotação mostraram que o genoma de *E. nigrum* possui aproximadamente 42% de suas bases pertencentes aos genes (compostos por 35% de CDSs e 7% de *introns*). A anotação também revelou que a densidade de genes no genoma é similar à maioria dos fungos filamentosos (250 a 300 genes por Mb), ficando entre bactérias (500 a 1000 genes por Mb) e mamíferos como humanos (7 a 11 genes por Mb) (CHAE et al., 2015; KUO; OCHMAN, 2010; VOGEL; MORAN, 2013). Além disso, a anotação indicou que as características gerais do genoma de *E. nigrum* seguem tendências comuns na organização para

o metabolismo celular fúngico. Neste contexto, a regulação e produção de proteínas relacionadas ao transporte e metabolismo de carboidratos correspondem à maior porção do genoma. Característica esperada por organismos como os fungos, seres heterótrofos com sua evolução bem-sucedida apoiada na degradação e modificação de variados tipos de carboidratos. A anotação também demonstrou pouco mais de 27 mil sequências repetitivas, correspondendo a 1,2% do genoma de *E. nigrum*. Esta fração pode ser considerada pequena, não ultrapassando 5% em ascomicetos (WÖSTEMEYER; KREIBICH, 2002). Os níveis de elementos repetitivos são baixos provavelmente devido aos mecanismos de defesa RIP (repeat-induced point mutations) que protegem os genomas de sequências altamente repetitivas inibindo o acúmulo de elementos transponíveis (HOOD et al., 2005).

Aproximadamente 55% das sequências biológicas pode ser relacionada à função (GO terms). Este número é baixo principalmente devido às anotações incompletas e/ou incorretas nos bancos de dados (KHATRI; DRAGHICI, 2005). A carência dessas informações se reflete em bancos específicos e dificulta a anotação de outros genomas. Assim, o sequenciamento e a anotação de *E. nigrum* auxiliam no preenchimento de lacunas no entendimento do funcionamento de Dothideomycetes e no comportamento dos endófitos.

As informações a respeito da predição foram utilizadas para o estudo das relações filogenéticas de *E. nigrum*. A árvore foi datada com o auxílio de informações de diferentes pontos de calibração com fósseis. Esta abordagem permitiu dar tamanhos diferentes aos ramos, pois os diferentes grupos não evoluem em taxas homogêneas (PETERSON et al., 2004). Esta calibração nas idades dos grupos na árvore filogenética envolve certa margem de erro, mesmo utilizando dados de fungos fósseis. Essas diferenças são aceitáveis, pois a probabilidade de fossilização dos primeiros novos membros dos grupos de fungos é muito pequena (BERBEE; TAYLOR, 2010). Dessa forma, é compreensível que as escalas de tempo dos fósseis utilizadas para determinar as idades de Sordariomycetes e Dothideomycetes tenham grande margem. A mensuração dessas taxas de variação é tida como uma importante característica da evolução para os relógios moleculares (LEPAGE et al., 2007; WELCH; BROMHAM, 2005). Neste contexto, o R8s é uma importante ferramenta para realizar a calibração e extrapolação das taxas de divergência. Neste trabalho foram utilizadas evidências para pontos de calibração como o recomendado pelos desenvolvedores de R8s, utilizando mais de um ponto de calibração em determinada faixa de tempo (SANDERSON, 2003). Estas características permitiram maior robustez na análise, indicando resultados como o agrupamento do endófito *E. nigrum* com outros grupos de fungos fitopatogênicos como *Didymella exigua* e *Stagonospora nodorum* é uma consequência normal da sua evolução. Seu

estilo de vida poderia ter sido adquirido recentemente em função do ganho ou perda de genes relacionados à patogenicidade. Os fungos endofíticos costumam compartilhar genes com fitopatógenos uma vez que possuem necessidades similares para o estabelecimento dentro da planta. Ambos devem ser capazes de adentrar a planta, interagir com as moléculas da sua hospedeira, interferir em suas repostas biológicas e nutrir-se. Assim, patogenicidade ou simbiose se deflagram pela maneira como o fungo interage com a planta hospedeira. Os fungos podem se nutrir por estratégias biotróficas, necrotróficas ou hemibiotróficas. Os fungos biotróficos (patogênicos e mutualistas) mantêm seus hospedeiros vivos sem causar grandes danos mesmo metabolizando ativamente. Os necrotróficos (patogênicos) crescem e matam células hospedeiras para obter nutrientes das células mortas. Os hemibiotróficos mantêm as células hospedeiras vivas por um tempo, entretanto, após a fase de colonização atuam como necrotróficos. Neste contexto, a ausência de alguns genes pode influenciar no estilo de vida dos fungos. Um exemplo é o caso do endofítico *Piriformospora indica*, o qual não possui genes envolvidos no metabolismo do azoto, limitando seu crescimento e consequentemente não causando danos à sua hospedeira. Além disso, o mesmo não possui genes relacionados a metabólitos secundários tóxicos e peptídeos cíclicos. Entretanto, em seu genoma podem ser encontrados genes fracamente expressos relacionados às enzimas de saprotróficos e hemibiotróficos. Estas características o colocam como um fungo intermediário entre decompositores e mutualistas biotróficos (ZUCCARO et al., 2011). No trabalho de Xu et al. (2014), *Harpophora oryzae*, um endófito fungo relativamente próximo ao patógeno *Magnaporthe oryzae*, indicou que eventos no número de genes podem ter sido responsáveis pela evolução do patógeno ancestral comum a ambos em um fungo endofítico. Estes eventos no número de genes podem ser deleções, duplicações, assim como expansões da família de genes e translocações de elementos de transponíveis.

Um terço das famílias de enzimas encontradas em fungos patogênicos e ausentes em *E. nigrum* estavam relacionadas ao metabolismo de carboidratos. A possível explicação poderia ser a necessidade de produção de grande variedade de enzimas de degradação de polissacarídeos pelos fungos para facilitar a infecção ou a nutrição. Assim um maior arsenal dessas enzimas em relação a *E. nigrum* possivelmente está relacionado ao estilo de vida destes organismos (ZHAO et al., 2014). Dentre as enzimas presentes em fitopatógenos e ausentes em *E. nigrum* está uma família de álcool desidrogenase com domínio GroES-like. Em estudo com *P. infestans*, Seidl et al., (2011) identificaram os genes relacionados sendo expressos significativamente durante a infecção do hospedeiro. Os domínios de enzimas possuem papel importante na relação patógeno-hospedeiro. A presença destes domínios pode interferir em

processos biológicos nas fases de interação como a absorção de nutrientes, supressão e eliciação de respostas antagônicas das plantas (SEIDL et al., 2011). Além do gene que codifica a enzima álcool desidrogenase com domínio GroES-like, não foi identificado o gene da enzima Frutose-bisfosfato aldolase. A enzima é conhecida principalmente pela sua atuação na via glicolítica, sendo necessária em todos os organismos dependentes da mesma. Entretanto uma família distinta não foi encontrada em *E. nigrum*, sendo detectada nos fitopatógenos estudados. Atualmente há evidências que a Frutose-1,6 bisfosfato aldolase apesar de participar da via glicolítica, em *Paracoccidioides* a mesma é secretada. A enzima liga-se ao plasminogênio, uma proteína presente no plasma sanguíneo. A ligação ao plasminogênio melhora a capacidade fibrinolítica do fungo e consequentemente sua patogênese em humanos (CHAVES et al., 2015). Apesar do contexto analisado neste trabalho ser em fitopatologia, a Frutose-bisfosfato aldolase de *Paracoccidioides* marcada para secreção chama a atenção para o papel diferenciado da enzima quando na superfície do fungo. Assim esta enzima se secretada poderia relacionar-se à nutrição dos fungos patogênicos, uma vez que existem relações fortes entre enzimas secretadas e o estilo de vida dos fungos (FERNANDES et al., 2014; MEINHARDT et al., 2014). Suplantar a defesa da planta hospedeira é outra necessidade de fungos patogênicos e endofíticos. O fungo *E. nigrum* não apresentou algumas das enzimas ligadas à Superfamília dos Facilitadores Majoritários (MFS). Em micro-organismos as proteínas de transportadores estão divididas principalmente em duas famílias, as ABC e a Superfamília dos Facilitadores Majoritários (MFS). As MFS são carreadoras secundárias capazes de transportar carboidratos, peptídeos, drogas, além de íons orgânicos e inorgânicos (ROOHPARVAR et al., 2007). Estes transportadores estão presentes na membrana plasmática e em fungos filamentosos sua função mais comum é a proteção contra compostos tóxicos como antibióticos e outras toxinas produzidos por outros micro-organismos. Assim as MFS de fungos patogênicos atuam no processo de virulência fornecendo proteção contra compostos de defesa da planta ou mediando à secreção de toxinas específicas (STERGIOPOULOS et al., 2002). Também foram detectados genes similares à família de glioxalases apenas em fungos fitopatogênicos. Estes genes estão ligados à resistência à bleomicina, um antibiótico produzido por actinomicetos. A glioxalase é produzida também pelos actinomicetos a fim de evitar a ação do antibiótico no próprio organismo, uma vez que a mesma atua clivando DNA. Além dos actinomicetos, a glioxalase também já foi descrita em vários fungos (ZHENG et al., 2011). A presença de glioxalases poderia indicar a grande capacidade competitiva destes fungos para nutrirem-se dos tecidos.

A secreção de proteínas eucarióticas pode acontecer pela via clássica ou não (PRUDOVSKY et al., 2008). As proteínas secretadas pela via clássica possuem um sinal peptídico na porção N-terminal que indica que a proteína deve fazer a rota por meio do retículo endoplasmático e aparelho de Golgi antes da secreção (JONES; DANGL, 2006). Os genes de proteínas secretadas são regulados a fim de atuar na fase biotrófica principalmente na quebra de matriz extracelular, modificação das hifas e mascarar as defesas da planta (MEINHARDT et al., 2014). Estas proteínas secretadas por fitopatógenos podem ser de caráter funcional hidrolítico, atacando a parede celular da hospedeira, degradando outras fontes de carbono ou participando em reações de azoto. Algumas dessas proteínas são pequenas e não aparentam domínio funcional aparente, sendo chamadas de efetores. Acredita-se que os efetores sejam responsáveis por suprimir ou neutralizar as defesas da planta (JONES; DANGL, 2006). Assim, estudos do perfil funcional do genoma e sua expressão têm indicado que a produção de proteínas efetoras nessa fase é a resposta para caracterizar o processo de desenvolvimento dos sintomas (MORAIS DO AMARAL et al., 2012). O transporte de substâncias na célula também possui papel importante nos mais variados processos de interação celular. Muitos genes ligados ao transporte são frequentemente descritos como diferencialmente expressos na maioria das respostas celulares às mudanças no ambiente. Naturalmente a presença ou a ausência destes genes envolvidos no transporte também está ligada ao estilo de vida dos fungos. Neste trabalho não foram encontradas em *E. nigrum* algumas famílias de genes ligadas ao transporte de substâncias da célula. Assim, estes genes de fungos fitopatogênicos merecem atenção, pois podem atuar na infecção. Dentre os genes encontrados em patógenos e ausentes em *E. nigrum*, os codificadores de proteínas tipo SNARE estão envolvidos na fusão de vesículas. Proteínas tipo SNARE são importantes no estudo de rotas secretoras, sendo encontradas em fungos filamentosos em uma variedade de compartimentos. Sua presença em locais variados de acordo com o tipo de proteína SNARE está ligada às rotas específicas (KUBICEK et al., 2014). Em *Magnaporthe oryzae*, fitopatógeno de arroz, existem evidências de duas rotas distintas de secreção para entrega de efetores que atuam durante a infecção. Estes efetores ao acumularem-se na face apoplástica da célula hospedeira. Os outros efetores que se destinam a atuar no citoplasma da célula do arroz tomam uma via distinta, atuando o complexo interfacial biotrófico (GIRALDO et al., 2013). Além disso, as proteínas tipo SNARE também são foco para a resistência de plantas, sendo a razão da resistência de algumas espécies de plantas a patógenos (COLLINS et al., 2003). Assim, a ausência de alguns tipos de proteína SNARE em *E. nigrum* pode estar associada à não necessidade deste tipo de recurso para burlar as defesas do hospedeiro devido a uma

maior tolerância do hospedeiro ao endófito. Outra família ligada não encontrada em *E. nigrum* e presente em fitopatógenos próximos que pode atuar como um efector é glicosiltransferase. Estas enzimas são capazes de transferir monossacarídeos para substratos como proteínas, lipídeos, carboidratos, dentre outros. Existem vários tipos de glicosiltransferases de procariotos a mamíferos, as quais muitos de seus subtipos não existem em grupos distintos de organismos (KLUTTS et al., 2006). O gene da enzima da família da glicosiltransferase foi descrito como um candidato potencial no processo de patogênese. Em mutantes de *Colletotrichum gloeosporioides* deficientes para glicosiltransferase observou-se o não desenvolvimento de antracnose em seringueira (CAI et al., 2013). Após ser observada como um efector potencialmente secretado, a enzima da família glicosiltransferase 2 de *Rhizoctonia solani* foi inoculada em folhas de arroz, milho e soja, causando morte celular e reforçando seu papel no desenvolvimento de doenças em plantas (ZHENG et al., 2013).

A expressão gênica dos organismos está sujeita a complexas regulações. Devido a esta complexidade, o cultivo de micro-organismos em laboratório para a produção de metabólitos específicos nas mesmas condições ambientais torna-se impraticável. Assim foi necessário cultivar *E. nigrum* em diferentes condições, buscando diferentes fases do seu ciclo de vida. Inicialmente foram realizados os tratamentos de 7 e 60 dias em meio BD, avaliando apenas os efeitos do tempo sobre o metabolismo do fungo. Entretanto, amostras com 60 dias apresentaram pouca atividade antimicrobiana. Possivelmente por causa de períodos longos nos quais o fungo tende a acumular compostos polifenólicos. Além disso, estes compostos polifenólicos se ligam aos ácidos nucleicos de forma irreversível, dificultando a extração de RNA (RUBIO-PIÑA; ZAPATA-PÉREZ, 2011). Dessa forma, a redução de 60 dias para 30 dias o período máximo de avaliação foi importante para avaliar a produção de metabólitos antes da queda nos níveis de metabólitos antimicrobianos. Também foi adicionado um tempo intermediário para acompanhar a produção de metabólitos secundários (20 dias) e elevado para 10 dias o tempo inferior. No meio BD existe uma tendência de o fungo aumentar a produção de metabólitos, dentre os quais pigmentos e compostos fenólicos, o que prejudica a extração de RNA e atividade de antimicrobianos. Assim, o cultivo em meio Sabouraud foi considerado, pois não estimula excessivamente a produção de compostos fenólicos. Ademais, o aumento do número de tratamentos pode favorecer de forma diferencial a expressão genes em diferentes *clusters* envolvidos no metabolismo secundário.

A variedade de resultados de acordo com o tempo, meio, solvente e micro-organismos indica a presença de mais de uma substância com propriedades antimicrobianas. Além disso, *E. nigrum* é uma espécie conhecida por produzir uma ampla gama de diferentes compostos

com atividade antimicrobiana (FÁVARO et al., 2012) como epicorazinas (BAUTE et al., 1978), epirodinas (IKAWA et al., 1978), flavipinas (BAMFORD et al., 1961), epicoccinas (ZHANG et al., 2007), epipiridonas, epicocarines (WANGUN; HERTWECK, 2007) e epicolactona (ARAÚJO et al., 2012).

A análise dos transcritos de culturas com 7 dias inferida por BLAST contra as sequências do nr do GenBank revelou que compostos primários, corroborando com a fase de vida do fungo. Hifas jovens necessitam sintetizar principalmente compostos primários para seu estabelecimento. Além disso, nos períodos de 30 e 60 dias a análise de inibição de *E. coli*, *S. aureus* e *Bacillus* sp. não foi observada intensa atividade antimicrobiana. Isto sugere que nestes tempos não deve ser esperada uma alta frequência de transcritos envolvidos na síntese de compostos secundários com atividade antimicrobiana. A fase de desenvolvimento dos fungos caracteriza-se por um dos principais fatores que influenciam na produção de metabólitos secundários, desencadeando os mais variados processos como esporulação, produção de micotoxinas tais como a esterigmatocistina presente em *Aspergillus nidulans* (CALVO et al., 2002). A produção de metabólitos primários ou secundários está relacionada à regulação, pois a síntese de metabólitos secundários pode inibir vias metabólicas de compostos primários (BROWN et al., 1996), desviando acetil-coA, por exemplo. Esta molécula é precursora de um dos maiores grupos de metabólitos secundários, os policetídeos (PKs). Os PKs assim como os outros compostos secundários (terpenóides, peptídeos ribossômicos e não ribossômicos) são produzidos por meio de diversas atividades enzimáticas de modificação de polímeros, as quais devem ser reguladas por mecanismos complexos (ANDERSEN et al., 2013). Neste contexto, torna-se mais eficaz a busca pelos genes de compostos secundários que estão organizados em *clusters*, nos quais as enzimas primárias são codificadas (BRAKHAGE, 2013). Neste trabalho a busca por indícios de sinais de expressão dos doze *clusters* de PKs encontrados anteriormente no genoma de *E. nigrum* mostrou que apenas o *cluster* 1 teve nível de expressão detectável, sendo necessários transcritos de outras fases de desenvolvimento do fungo para comparações.

A análise destes *clusters* de metabólitos secundários do ponto de vista genético detectou que *E. nigrum* possui poucos *clusters* homólogos compartilhados com outros fungos. A baixa quantidade de *clusters* homólogos pode se decorrente de três fatores. Baixa sensibilidade de AntiSmash (1), uma vez que aumentando a sensibilidade com Synteny Clusters foi ampliada e não detectou mais *clusters* relacionados. Esta baixa sensibilidade pode ser por causa da grande diversidade de genes presentes em *clusters* com a mesma finalidade (2). Assim, *clusters* apesar de relacionados, possuem genes com grandes diferenças em seus

tamanhos, domínios, *introns*, duplicações e deleções. Estas diferenças geram dificuldades de relacionar com outros já existentes, processo ao qual o AntiSmash utiliza para detecção e consequentemente Synteny Clusters. Outra explicação poderia ser simplesmente a baixa frequência de *clusters* homólogos entre os táxons (3). Esta baixa frequência é possível até entre espécies do mesmo gênero como observado entre espécies de *Colletotrichum*, onde *C. graminicola* possui mais que o dobro de *clusters* que *C. sublineola*. Assim, estes fatores podem contribuir tanto isoladamente como em conjunto para explicar a baixa frequência de *clusters* homólogos relacionados. Outra dificuldade que pode ser contornada é a presença de *clusters* incompletos devido a possíveis falhas no processo de sequenciamento. Estas lacunas geralmente tendem a dificultar análises de comparação. Entretanto, o método de Synteny Clusters determina a identidade primeiramente para depois realizar as comparações, permitindo estender as análises de compartilhamento para estes *clusters* mesmo que presentes em *contigs* fragmentados. A análise deste compartilhamento de *clusters* de metabólitos secundários é importante, pois possibilita indicar o quanto uma espécie é metabolicamente próxima à outra. O fungo *D. exigua* mesmo sendo filogeneticamente mais próximo de *E. nigrum* em relação aos outros possui um perfil diferente. Possuindo uma menor quantidade de *clusters* de metabólitos, *D. exigua* demonstra um arsenal mais compacto de metabólitos secundários em relação a *E. nigrum*, dificultando relações diretas entre os grupos. Entretanto, a comparação com outros grupos de fungos demonstra que o compartilhamento pode ir além do estilo de vida do fungo. Estes resultados reforçam as teorias de que a mudança de poucos genes pode ser o suficiente para alterar o estilo de vida. O fungo *E. nigrum* apesar de ser mais próximo filogeneticamente de *D. exigua*, porém compartilha menos *clusters* com *E. nigrum*. O endófito *E. nigrum* por sua vez, compartilha mais *clusters* com outras espécies de patógenos, indicando que a presença de certos *clusters* está mais relacionada às adaptações ambientais do que o caráter filogenético. Alguns *clusters* relacionados aos policetídeos além de serem importantes compostos farmacêuticos são conhecidos como fatores de virulência em várias espécies de fungos (KROKEN et al., 2003). A ausência destes *clusters* em *E. nigrum* pode ser uma das causas do estilo de vida diferente de espécies próximas. A ausência do *cluster* relacionado à produção do tripeptídeo fosfinotricina pode ser uma das razões para o estilo de vida endofítico. O tripeptídeo fosfinotricina é descrito como um herbicida produzido por *Streptomyces viridochromogenes* (SCHWARTZ et al., 2004). *Clusters* possivelmente envolvidos na produção deste composto estão presentes na maior partes dos fungos patogênicos estudados neste trabalho, mesmo *D. exigua*. Esta espécie possui poucos *clusters* de metabólitos secundários, porém é fitopatogênica para várias plantas (AVESKAMP et al.,

2010). Dessa forma, o estudo do compartilhamento de *clusters* de metabólitos secundários entre endófitos e patógenos como realizado neste trabalho é importante para detecção de genes envolvidos na patogenicidade. A investigação de genes ausentes em endófitos e presentes em uma gama de outros patógenos facilita o processo de triagem *in silico*. Entretanto há necessidade constante de ferramentas de bioinformática para facilitar este processo, assim como o aumento de bases de dados importantes como PHI-base. Esta base de dados contém sequências de proteínas com informações curadas de genes comprovadamente capazes de afetar as interações patógeno-hospedeiro (WINNENBURG et al., 2007).

Além da homologia, a identificação dos compostos produzidos é fundamental. Apesar da dificuldade em determinar a identidade dos compostos secundários relacionados aos *clusters*, foi possível determinar a presença de um antimicrobiano. Dentre os *clusters* de metabólitos secundários foi encontrado um envolvido na produção de lantipeptídeos. Este metabólito não é produzido na via ribossômica (KNERR; VAN DER DONK, 2012). Funcionalmente são conhecidos por sua função antimicrobiana e pode ser produzido tanto por fungos quanto bactérias (BRÖTZ; SAHL, 2012). Este grande *cluster* pode ser responsável pela capacidade de controle biológico de *E. nigrum* por meio da produção de antimicrobianos.

7 CONSIDERAÇÕES FINAIS

Epicoccum nigrum produz compostos antimicrobianos antes de 7 dias quando crescidos em meio de cultura, entretanto a produção torna-se mais evidente entorno de 20 dias. Neste período são produzidos diferentes metabólitos com atividade antimicrobiana, uma vez que foi observada atividade diferente de acordo com o extrato e bactéria analisados. A expressão dos genes relacionados a estes metabólitos é pequena nos períodos iniciais, uma vez que a análise do transcriptoma no período indica que os níveis de expressão de genes dos *clusters* de metabólitos secundários não são praticamente detectados. Neste período a colônia fúngica possui seu metabolismo concentrado na produção de metabólitos primários.

A determinação dos genes de *E. nigrum* foi criteriosa, uma vez que este processo é passível de erros. Dessa forma, a este trabalho buscou realizá-la por diferentes processos que envolveram a comparação com o transcriptoma do próprio fungo, diversos proteomas de Dothideomycetes e a criação do conjunto de treinamento a partir dos próprios genes de *E. nigrum*.

As análises filogenéticas indicam que *E. nigrum* teria um ancestral comum com *D. exigua* e que teria ocorrido divergência há aproximadamente 100 milhões de anos. Estes dois fungos compartilham diversos genes com patógenos, mas apesar de similares em número de famílias de genes, a aquisição ou perda de uma pequena fração influenciaram em estilos de vida distintos.

O fungo *E. nigrum* apresenta um genoma com características típicas de fungos filamentosos em tamanho, quantidade de genes e metabolismo. Sua anotação o estabelece como um dos poucos endofíticos sequenciados e anotados, sendo o primeiro Dothideomycetes. Estes resultados foram proporcionados graças aos componentes desenvolvidos que foram capazes de invocar programas e integrar a outros associados à plataforma EGene2. Esta integração ao ambiente da plataforma permite a reprodutibilidade rápida dos dados, bem como a possibilidade de utilização por futuros usuários com outros organismos. Assim, estas ferramentas permitiram detectar e anotar importantes características do genoma de *E. nigrum*, estabelecendo-o o fungo endofítico melhor caracterizado da atualidade. Além disso, dados a respeito da anotação do genoma de *E. nigrum* juntamente com outros grupos permitiu uma maior compreensão das relações filogenéticas de *E. nigrum* na maior escala observada para o grupo. A análise dos dados também permitiu a detecção de *clusters* de genes de metabólitos secundários que auxiliam *E. nigrum* no controle biológico

por meio de metabólitos secundários como lantipeptídeos. Este trabalho com os *clusters* de genes de metabólitos secundários só foi possível ser realizada de forma rápida e com tantos dados graças ao desenvolvimento do programa Synteny Clusters. A abordagem utilizando Synteny Clusters foi a maior análise de comparação de genes de metabólitos secundários entre os grupos de Dothideomycetes. Essa análise comparativa dos *clusters* entre espécies indicou importantes *clusters* de metabólitos secundários ausentes em *E. nigrum*. A perda destes *clusters* envolvidos na fitopatogenicidade pode ser alguns dos fatores responsáveis pela adoção do novo estilo de vida e ajudar a reconhecer processos similares em outros fungos.

8 CONCLUSÕES

A construção do conjunto de treinamento específico para *E. nigrum* permitiu melhores resultados na predição dos genes em relação ao conjunto de treinamento de *A. nidulans*;

A análise do transcriptoma de *E. nigrum* indicou que culturas com 7 dias apresentam baixa expressão de compostos secundários. O transcriptoma também foi útil para curadoria na construção do conjunto de treinamento e na correção da predição de genes.

Os módulos desenvolvidos em Perl integraram-se à plataforma EGene2, invocando programas (Augustus, Psortb, iPSORT, WoLFPSORT e MYOP), extraindo informações (report_fasta.pl) e disponibilizando as novas informações para módulos existentes, alguns dos quais adaptados (report_feature_table_artemis.pl e report_feature_table_submission.pl).

A plataforma EGene2 possibilitou uma rápida e eficiente anotação do genoma de *E. nigrum*, relatando muitas informações a respeito do arsenal genético e sua funcionalidade.

A análise dos metabólitos por ação de antimicrobianos indicou que culturas crescidas em torno de 20 dias possuem maior produção de antimicrobianos e ocorre um decréscimo em torno de 30 dias. No genoma foram encontrados 38 *clusters* de metabólitos secundários, alguns dos quais relacionados ao efeito antimicrobiano, dentre os quais lantipeptídeos.

O fungo *E. nigrum* possui maior proximidade filogenética com *Didymella exigua*, *Aaosphaeria arxii* e *Stagonospora nodorum*. Entretanto, seus *cluster* de metabólitos secundários não seguem a mesma relação. Existindo um compartilhamento de boa parte de seus *clusters* com outros fungos, mesmo fitopatogênicos. Entretanto, alguns contendo genes envolvidos em processos de fitopatogênese estão ausentes em *E. nigrum*.

REFERÊNCIAS*

- ALLEN, J.E.; PERTEA, M.; SALZBERG, S.L. Computational gene prediction using multiple sources of evidence. **Genome Research**. v 14, p 142-148, 2004.
- ALTSCHUL, S.F.; GISH, W.; MILLER, W.; MYERS, E.W.; LIPMAN, D.J. Basic local alignment search tool. **Journal Molecular Biology**. v 215, p 403-410, 1990.
- ALVES, J.M.P.; OLIVEIRA, A.L.; SANDBERG, T.O.M.; MORENO-GALLEGO, J.L.; TOLEDO, M.A.F.; MOURA, E.M.M.; OLIVEIRA, L.S.; DURHAM, A.M.; MEHNERT, D.U.; ZANOTTO, P.M.A.; REYES, A.; GRUBER, A. GenSeed-HMM: A Tool for Progressive Assembly Using Profile HMMs as Seeds and its Application in *Alpavirinae* Viral Discovery from Metagenomic Data. **Frontiers in Microbiology**. v 7, 2016.
- ANDERSEN, M.; JAKOB, B.; NIELSEN, A.; KLITGAARD, L.M.; PETERSEN, M.; TILDE, J.H.; LENE, H.B. Accurate prediction of secondary metabolite gene *clusters* in filamentous fungi. **Proceedings of the National Academy of Sciences**. v. 110 p. E99-E107, 2013.
- ANDERSEN, MR; SALAZAR, MP.; SCHAAP, PJ.; VAN DE VONDERVOORT, PJ.; CULLEY, D.; THYKAER, J.; FRISVAD, JC.; NIELSEN, KF.; ALBANG, R.; ALBERMANN, K.; BERKA, RM.; BRAUS, GH.; BRAUS-STROMEYER, SA.; CORROCHANO, LM.; DAI, Z.; VAN DIJCK, PW.; HOFMANN, G.; LASURE, LL.; MAGNUSON, JK.; MENKE, H.; MEIJER, M.; MEIJER, SL.; NIELSEN, JB.; NIELSEN, ML.; VAN OUYEN, AJ.; PEL, HJ.; POULSEN, L.; SAMSON, RA.; STAM, H.; TSANG, A.; VAN DEN BRINK, JM.; ATKINS, A.; AERTS, A.; SHAPIRO, H.; PANGILINAN, J.; SALAMOV, A.; LOU, Y.; LINDQUIST, E.; LUCAS, S.; GRIMWOOD, J.; GRIGORIEV, IV.; KUBICEK, CP.; MARTINEZ, D.; VAN PEIJ, NN.; ROUBOS, JA.; NIELSEN, J.; BAKER, SE.; Comparative genomics of citric-acid-producing *Aspergillus niger* ATCC 1015 versus enzyme-producing CBS 513.88. **Genome Research**. v 21, p 157-170, 2011.
- ANDRADE, L.H.; KEPPLERA, F.; SCHOENLEIN-CRUSIUS, I.H.; PORTO, A.L.M.; COMASSETO, J.V. Evaluation of acetophenone monooxygenase and alcohol dehydrogenase activities in different fungal strains by biotransformation of acetophenone derivatives. **Journal Molecular Catalysis Enzymatic**. v. 31 p. 129-135, 2004.
- ANKE, T; THINES, E. **Fungal metabolites as lead structures for agriculture**. In: Robson GD, Van WEST P, GADD GM (eds), *Exploitation of Fungi*. Cambridge University Press: Cambridge, pp. 45–58. 2007.

*De acordo com:

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 6023**: informação e documentação: referências: elaboração, Rio de Janeiro, 2002.

ARAÚJO, F.D.S.; FÁVARO, L.C.L.; ARAÚJO, W.L.; OLIVEIRA, F.L.; APARICIO, R.; MARSAIOLI, A.J. Epicolactone – Natural Product Isolated from the Sugarcane Endophytic Fungus *Epicoccum nigrum*. **European Journal of Organic Chemistry**. v 27, 2012.

ARNAUD, MB.; CERQUEIRA, GC.; INGLIS, DO.; SKRZYPEK, MS.; BINKLEY, J.; CHIBUCOS, MC.; CRABTREE, J.; HOWARTH, C.; ORVIS, J.; SHAH, P.; WYMORE, F.; BINKLEY, G.; MIYASATO, SR.; SIMISON, M.; SHERLOCK, G.; WORTMAN, JR.; The Aspergillus Genome Database (AspGD): recent developments in comprehensive multispecies curation, comparative genomics and community resources. **Nucleic Acids Research**. v 40, p D653-659, 2012.

AVESKAMP, M.; GRUYTER, H.; WOUTENBERG, J.; VERKLEY, G.; CROUS, PW. Highlights of the Didymellaceae: A polyphasic approach to characterise Phoma and related pleosporalean genera. **Studies in Mycology**. v 65, p 1–60, 2010.

AZEVEDO J.L.; ARAÚJO, W.L. **Diversity and applications of endophytic fungi isolated from tropical plants**. In: Ganguli BN, Deshmukh SK (Ed.). Fungi: multifaceted microbes. Boca Raton: CRC Press, 2007. cap. 6, p. 189-207.

BACKMAN PA, SIKORA, R.A. Endophytes: an emerging tool for biological control. **Biologic Control**, v 46, p 1–3, 2008.

BAMFORD, P.C.; NORRIS, G.L.F.; WARD, G. Flavipin production by *Epicoccum* spp. **Transactions of the British Mycological Society**. 44:354–356, 1961.

BANNAI, H.; TAMADA, Y.; MARUYAMA, O.; NAKAI, K.; MIYANO, S. "Extensive feature detection of N-terminal protein sorting signals". **Bioinformatics**, v 18, p 298-305, 2002.

BASÍLIO, MC.; GASPAR, R.; SILVA PEREIRA, C.; SAN ROMÃO, MV.; Penicillium glabrum cork-colonising isolates - preliminary analysis of their genomic similarity. **Revista Iberoamericana de Micología**. v 23 , p 151-154, 2006.

BAUTE, M.A, DEFFIEUX, G.; BAUTE, R.; NEVE, U.A. New antibiotics from the fungus *Epicoccum nigrum*. I. Fermentation, isolation and antibacterial properties. **Journal of Antibiotics**. v 31, p 1099–1101, 1978.

BEIMFORDE, C; FELDBERG, K; NYLINDER, S; RIKKINEN, J; TUOVILA, H.; DÖRFELT, H.; GUBE, M.; JACKSON, DJ.; REITNER, J.; SEYFULLAH, LJ.; SCHMIDT, AR. Estimating the Phanerozoic history of the Ascomycota lineages: combining fossil and molecular data. **Molecular Phylogenetics and Evolution**. v 78, p 386-398, 2014.

BENSON, G. Tandem repeats finder: a program to analyze DNA sequences. **Nucleic Acids Research**. v 27, p 573-580, 1999.

BERBEE, M.L.; TAYLOR, J.W. Dating the molecular clock in fungi - how close are we? **Fungal Biology Reviews**. v 24, p1-16, 2010.

BÉRDY, J. Bioactive microbial metabolites: a personal view. **The Journal of Antibiotics**. v 58, p 1–26, 2005.

BLANCO-ULATE, B.; ROLSHAUSEN, PE.; CANTU, D.; Draft Genome Sequence of the Grapevine Dieback Fungus *Eutypa lata* UCR-EL1. **Genome announcements**. v 1, p , 2013.

BODE, HB.; BETHE B.; HÖFS, R.; ZEECK, A. Big effects from small changes, p possible ways to explore nature's chemical diversity. **ChemBioChem**. v 3, p 619-627, 2002.

BOK, JW.; HOFFMEISTER, D.; MAGGIO-HALL, LA.; MURILLO R.; GLASNER, JD.; KELLER, NP. Genomic mining for *Aspergillus* natural products. **Chemical Biology**, v 13, p 31–37, 2006.

BOLTON, M.D.; VAN ESSE, H.P.; VOSSEN, J.H.; DE JONGE, R.; STERGIOPOULOS, I.; STULEMEIJER, I.J.; VAN DEN BERG, G.C.; BORRÁS-HIDALGO, O.; DEKKER, H.L.; DE KOSTER, C.G.; DE WIT, P.J.; JOOSTEN, M.H.; THOMMA, B.P. The novel *Cladosporium fulvum* lysin motif effector Ecp6 is a virulence factor with orthologues in other fungal species. **Molecular Microbiology**. v 69, p 119-136, 2008.

BORÉM, A.; SANTOS F. R. **Biotecnologia Simplificada**. Editora Suprema. Viçosa, MG, 2001.

BORGES, W.S.; BORGES, K.B.; BONATO, P.S.; SAID, S.; PUPO, M.T. Endophytic fungi: natural products, enzymes and biotransformation reactions. **Current Organic Chemistry**. v 13, p 1137-1163, 2009.

BRAKHAGE, A.A. Regulation of fungal secondary metabolism. **Nature Reviews Microbiology**. v 11, p 21-32, 2013.

BRITO, A.F.; BRACONI, C.T.; WEIDMANN, M.; DILCHER, M.; ALVES, J.M.; GRUBER, A.; ZANOTTO, P.M. The Pangenome of the *Anticarsia gemmatilis* Multiple Nucleopolyhedrovirus (AgMNPV). **Genome Biology Evolution**. v 27, p 94-108, 2015.

BRÖTZ, H.; SAHL, H.G. New insights into the mechanism of action of lantibiotics-diverse biological effects by binding to the same molecular target. **J Antimicrob Chemother**. v 46, p 1-6, 2000.

BROWN, D.W.; ADAMS, T.H.; KELLER, N.P. *Aspergillus* has distinct fatty acid synthases for primary and secondary metabolism. **Proceedings of the National Academy of Sciences**. v 93, p 14873-14877, 1996.

BURGE, C.; KARLIN, S. Prediction of complete gene structures in human genomic DNA. **J Mol Biol**. v 268, p 78 -94, 1997.

BUSHLEY, K.E.; TURGEON, B.G. Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships. **BMC Evolutionary Biology**. v 10, p 26. doi:10.1186/1471-2148-10-26, 2010.

- CAI, Z.; LI, G.; LIN, C.; SHI, T.; ZHAI, L.; CHEN, Y.; HUANG, G. Identifying pathogenicity genes in the rubber tree anthracnose fungus *Colletotrichum gloeosporioides* through random insertional mutagenesis. **Microbiological Research**. v 168, p 340-350, 2013.
- CALVO, A. M.; WILSON, R.A.; BOK, J.W.; KELLER, N.P. Relationship between secondary metabolism and fungal development. **Microbiology Molecular Biology Review**. v 66, p 447-459, 2002.
- CAMPBELL, M.S.; HOLT, C.; MOORE, B.; YANDELL, M. Genome Annotation and Curation Using MAKER and MAKER-P. **Current Protocol Bioinformatics**. v 48, p 4.11.1-4.11.39, 2014.
- CAMPOS, F.F.; ROSA, L.H.; COTA, B.B.; CALIGIORNE, R.B.; TELESRABELLO, A.L.; ALVES T.M.A.; ROSA, C.A.; ZANIC, L. Leishmanicidal metabolites from *Cochliobolus* sp., an endophytic fungus isolated from *Piptadenia adiantoides* (Fabaceae). **PLoS Neglected Tropical Diseases** v 2 p 12:e348.doi:10.1371/journal.pntd.0000348, 2008.
- CANTAREL, B.L.; KORF, I.; ROBB, S.M.; PARRA, G.; ROSS, E.; MOORE, B.; HOLT, C.; SANCHEZ ALVARADO, A.; YANDELL, M. MAKER: an easy-to-use annotation *pipeline* designed for emerging model organism genomes. **Genome Research**. v 18 , p 188–196, 2008.
- CAPORALE, L.H. Chemical Ecology: a view from the pharmaceutical industry. **PNAS**, v 92, p 75-82, 1995.
- CHAE, M; DANKO, C.G.; KRAUS, W.L. groHMM: a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data. **BMC Bioinformatics**. v 16, 2015
- CHAVES, E.G.; WEBER, S.S.; BÁO, S.N.; PEREIRA, L.A.; BAILÃO, A.M.; BORGES, C.L.; SOARES, C.M. Analysis of *Paracoccidioides* secreted proteins reveals fructose 1,6-bisphosphate aldolase as a plasminogen-binding protein. **BMC Microbiology**. v 15, 2015.
- CHIANG, YM.; SZEWCZYK, E.; NAYAK, T.; DAVIDSON, AD.; SANCHEZ JF.; LO, HC.; HO, WY.; SIMITYAN, H.; KUO, E.; PRASEUTH, A.; WATANABE, K.; OAKLEY, BR.; WANG, CC. Genomics-driven discovery of PKS-NRPS hybrid metabolites from *Aspergillus nidulans*. **Nature Chemical Biology**. v 3, p 213–217, 2007.
- CHIANG, YM.; SZEWCZYK, E.; NAYAK, T.; DAVIDSON, AD.; SANCHEZ, JF.; LO, HC.; HO, WY.; SIMITYAN, H.; KUO, E.; PRASEUTH, A.; WATANABE, K.; OAKLEY, BR.; WANG, CC. Molecular genetic mining of the *Aspergillus* secondary metabolome, p Discovery of the emericellamide biosynthetic pathway. **Chemical Biology**, v 15, p 527–532, 2008.
- CICHEWICZ, RH. Epigenome manipulation as a pathway to new natural product scaffolds and their congeners. **Natural Product Reports**, v 27, p 11-22, 2010.

COLLEMARE, J.; BILLARD, A.; BÖHNERH, U.; LEBRUN, M. Biosynthesis of secondary metabolites in the rice BLAST fungus *Magnaporthe grisea*: the role of hybrid PKS-NRPS in pathogenicity. **Mycology Research**. v 112, p 207–215, 2008.

COLLINS, N.C.; THORDAL-CHRISTENSEN, H.; LIPKA, V.; BAU, S.; KOMBRINK, E.; QIU, J.L.; HÜCKELHOVEN, R.; STEIN, M.; FREIALDENHOVEN, A.; SOMERVILLE, S.C.; SCHULZE-LEFERT, P. SNARE-protein-mediated disease resistance at the plant cell wall. **Nature**. v 425, p 973-977, 2003.

CONDON, B.J.; LENG, Y.; WU, D.; BUSHLEY, K.E.; OHM, R.A.; OTILLAR, R.; MARTIN, J.; SCHACKWITZ, W.; GRIMWOOD, J.; MOHDZAINUDIN, N.; XUE, C.; WANG, R.; MANNING, V.A.; DHILLON, B.; TU, Z.J.; STEFFENSON, B.J.; SALAMOV, A.; SUN, H.; LOWRY, S.; LABUTTI, K.; HAN, J.; COPELAND, A.; LINDQUIST, E.; BARRY, K.; SCHMUTZ, J.; BAKER, S.E.; CIUFFETTI, L.M.; GRIGORIEV, I.V.; ZHONG, S.; TURGEON, B.G.; Comparative genome structure, secondary metabolite, and effector coding capacity across *Cochliobolus* pathogens. **PLoS Genetics**. v 9, 2013. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/23357949>> Acesso em: sep. 2015.

CONDON, B.J.; LENG, Y.; WU, D.; BUSHLEY, K.E.; OHM, R.A.; OTILLAR, R.; MARTIN, J.; SCHACKWITZ, W.; GRIMWOOD, J.; MOHDZAINUDIN, N.; XUE, C.; WANG, R.; MANNING, V.A.; DHILLON, B.; TU, Z.J.; STEFFENSON, B.J.; SALAMOV, A.; SUN, H.; LOWRY, S.; LABUTTI, K.; HAN, J.; COPELAND, A.; LINDQUIST, E.; BARRY, K.; SCHMUTZ, J.; BAKER, S.E.; CIUFFETTI, L.M.; GRIGORIEV, I.V.; ZHONG, S.; TURGEON, B.G.; Comparative genome structure, secondary metabolite, and effector coding capacity across *Cochliobolus* pathogens. **PLoS Genetics**. v 9, p , 2013;

COOKE, I.R.; JONES, D.; BOWEN, J.K.; DENG, C.; FAOU, P.; HALL, N.E.; JAYACHANDRAN, V.; LIEM, M.; TARANTO, A.P.; PLUMMER, K.M.; MATHIVANAN, S.; Proteogenomic analysis of the *Venturia pirina* (Pear Scab Fungus) secretome reveals potential effectors. **Journal of Proteome Research**. v 13, p 3635-44, 2014.

COX, R.J. Polyketides, proteins and genes in fungi: programmed nano-machines begin to reveal their secrets. **Organic & Biomolecular Chemistry**. v 5, p 2010-2026, 2007.

DE BIE, T.; CRISTIANINI, N.; DEMUTH, J.P.; HAHN, M.W.; CAFE: a computational tool for the study of gene family evolution. **Bioinformatics**. v 22, p 1269-71, 2006.

DE CAL, M.A.; LARENAI, L.; TORRES, R.; LAMARCAN, U.; DOMENICHINIP, B.A.; MELGAREJOP, E. Population dynamics of *Epicoccum nigrum*, a biocontrol agent against brown rot in stone fruit. **Journal of Applied Microbiology**. v 106, p 592–605, 2009.

DE WIT, P.J.; VAN DER BURGT, A.; ÖKMEN, B.; STERGIOPOULOS, I.; ABDEL-SALAM, K.A.; AERTS, A.L.; BAHKALI, A.H.; BEENEN, H.G.; CHETTRI, P.; COX, M.P.; DATEMA, E.; DE VRIES, R.P.; DHILLON, B.; GANLEY, A.R.; GRIFFITHS, S.A.; GUO, Y.; HAMELIN, R.C.; HENRISSAT, B.; KABIR, M.S.; JASHNI, M.K.; KEMA, G.; KLAUBAUF, S.; LAPIDUS, A.; LEVASSEUR, A.; LINDQUIST, E.; MEHRABI, R.; OHM,

RA.; OWEN, TJ.; SALAMOV, A.; SCHWELM, A.; SCHIJLEN, E.; SUN, H.; VAN DEN BURG, HA.; VAN HAM, RC.; ZHANG, S.; GOODWIN, SB.; GRIGORIEV, IV.; COLLEMARE, J.; BRADSHAW, RE. The genomes of the fungal plant pathogens *Cladosporium fulvum* and *Dothistroma septosporium* reveal adaptation to different hosts and lifestyles but also signatures of common ancestry. **PLoS Genetics**. v 8, 2012. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/23209441> Acesso em: sep. 2015.

DELCHER, A.L.; BRATKE, K.A.; POWERS, E.C.; SALZBERG, S.L. Identifying bacterial genes and endosymbiont DNA with Glimmer. **Bioinformatics**. v 23, p 673-679, 2007.

DESJARDINS, CA.; CHAMPION, MD.; HOLDER, JW.; MUSZEWSKA, A.; GOLDBERG, J.; BAILÃO, AM.; BRIGIDO, MM.; FERREIRA, ME.; GARCIA, AM.; GRYNBERG, M.; GUJJA, S.; HEIMAN, DI.; HENN, MR.; KODIRA, CD.; LEÓN-NARVÁEZ, H.; LONGO, LV.; MA, LJ.; MALAVAZI, I.; MATSUO, AL.; MORAIS, FV.; PEREIRA, M.; RODRÍGUEZ-BRITO, S.; SAKTHIKUMAR, S.; SALEM-IZACC, SM.; SYKES, SM.; TEIXEIRA, MM.; VALLEJO, MC.; WALTER, ME.; YANDAVA, C.; YOUNG, S.; ZENG, Q.; ZUCKER, J.; FELIPE, MS.; GOLDMAN, GH.; HAAS, BJ.; MCEWEN, JG.; NINO-VEGA, G.; PUCCIA, R.; SAN-BLAS, G.; SOARES, CM.; BIRREN, BW.; CUOMO, CA.; Comparative genomic analysis of human fungal pathogens causing paracoccidioidomycosis. **PLoS Genetics**. v 7, p , 2011.

DOYLE, F.; ZALESKI, C.; GEORGE, A.D.; STENSON, E;K; RICCIARDI, A.; TENENBAUM, S.A. Bioinformatic tools for studying post-transcriptional gene regulation : The UAlbany TUTOR collection and other informatic resources. **Methods in Molecular Biology**. v 419, p 39-52, 2008.

DUNN, CW; HEJNOL, A; MATUS, DQ.; PANG, K.; BROWNE, WE.; SMITH, SA.; SEAVER, E.; ROUSE, GW.; OBST, M.; EDGEcombe, GD.; SØRENSEN, MV.; HADDOCK, SH.; SCHMIDT-RHAESA, A.; OKUSU, A.; KRISTENSEN, RM.; WHEELER, WC.; MARTINDALE, MQ.; GIRIBET, G. Broad phylogenomic sampling improves resolution of the animal tree of life. **Nature**. v 452, p 745-749, 2008.

DURHAM, A.M.; Kashiwabara, A.Y.; Matsunaga, F.T.; Ahagon, P.H.; Rainone, F.; Varuzza, L.; Gruber, A. EGene: a configurable *pipeline* generation system for automated sequence analysis. **Bioinformatics**. v 21, p 2812-2813, 2005.

EILBECK, K.; MOORE, B.; HOLT, C.; YANDELL, M. Quantitative measures for the management and comparison of annotated genomes. **BMC Bioinformatics**. v 10, 2009.

ELLWOOD, SR.; LIU, Z.; SYME, RA.; LAI, Z.; HANE, JK.; KEIPER, F.; MOFFAT, CS.; OLIVER, RP.; FRIESEN, TL. A first genome assembly of the barley fungal pathogen *Pyrenophora teres* f. *teres*. **Genome Biology**. v 11, p R109, 2010.

EMMS, D.M.; KELLY, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. **Genome Biology**. v 16, 2015.

ENRIGHT, A.J.; VAN DONGEN, S.; OUZOUNIS, C.A. An efficient algorithm for large-scale detection of protein families. **Nucleic Acids Research**. v 30 p 1575-1584, 2002.

FAGHIHI, M.A.; WAHLESTEDT, C. Regulatory roles of natural antisense transcripts. **Nature Reviews Molecular Cell Biology**. v 10, p 637–643, 2009.

FÁVARO L.C.L. (2009) Diversidade e interação de *Epicoccum* spp. com cana-de-açúcar (*Saccharum officinarum*, L.). 291p. Tese (Doutorado em Genética e Melhoramento de Plantas) – Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba.

FÁVARO, LC; SEBASTIANES, FL.; ARAÚJO, WL.; *Epicoccum nigrum* P16, a sugarcane endophyte, produces antifungal compounds and induces root growth. **PLoS One**. v 7, 2012.

FERNANDES, I.; ALVES, A.; CORREIA, A.; DEVREESE, B.; ESTEVES, A.C. Secretome analysis identifies potential virulence factors of *Diplodia corticola*, a fungal pathogen involved in cork oak (*Quercus suber*) decline. **Fungal Biology**. v 118, p 516-523, 2014.

FLYNT, A.S.; LAI, E.C. Biological principles of microRNA-mediated regulation: shared themes amid diversity. **Nature Reviews Genetics**. v 9, p 831-842, 2008.

GALAGAN, JE.; CALVO, SE.; BORKOVICH, KA.; SELKER, EU.; READ, ND.; JAFFE, D.; FITZHUGH, W.; MA, LJ.; SMIRNOV, S.; PURCELL, S.; REHMAN, B.; ELKINS, T.; ENGELS, R.; WANG, S.; NIELSEN, CB.; BUTLER, J.; ENDRIZZI, M.; QUI, D.; IANAKIEV, P.; BELL-PEDERSEN, D.; NELSON, MA.; WERNER-WASHBURNE, M.; SELITRENNIKOFF, CP.; KINSEY, JA.; BRAUN, EL.; ZELTER, A.; SCHULTE, U.; KOTHE, GO.; JEDD, G.; MEWES, W.; STABEN, C.; MARCOTTE, E.; GREENBERG, D.; ROY, A.; FOLEY, K.; NAYLOR, J.; STANGE-THOMANN, N.; BARRETT, R.; GNERRE, S.; KAMAL, M.; KAMVYSSELIS, M.; MAUCELI, E.; BIELKE, C.; RUDD, S.; FRISHMAN, D.; KRYSTOFOVA, S.; RASMUSSEN, C.; METZENBERG, RL.; PERKINS, DD.; KROKEN, S.; COGONI, C.; MACINO, G.; CATCHESIDE, D.; LI, W.; PRATT, RJ.; OSMANI, SA.; DESOUSA, CP.; GLASS, L.; ORBACH, MJ.; BERGLUND, JA.; VOELKER, R.; YARDEN, O.; PLAMANN, M.; SEILER, S.; DUNLAP, J.; RADFORD, A.; ARAMAYO, R.; NATVIG, DO.; ALEX, LA.; MANNHAUPT, G.; EBBOLE, DJ.; FREITAG, M.; PAULSEN, I.; SACHS, MS.; LANDER, ES.; NUSBAUM, C.; BIRREN, B.; The genome sequence of the filamentous fungus *Neurospora crassa*. **Nature**. v 422, p 859-68, 2003.

GIANOULIS, TA; GRIFFIN, MA.; SPAKOWICZ, DJ.; DUNICAN, BF.; ALPHA, CJ.; SBONER, A.; SISMOUR, AM.; KODIRA, C.; EGHOLM, M.; CHURCH, GM.; GERSTEIN, MB.; STROBEL, SA.; Genomic analysis of the hydrocarbon-producing, cellulolytic, endophytic fungus *Ascocoryne sarcoides*. **PLoS Genetics**. v 8, p , 2012;

GIRALDO, M.C.; DAGDAS, Y.F.; GUPTA, Y.K.; MENTLAK, T.A.; YI, M.; MARTINEZ-ROCHA, A.L.; SAITOH, H.; TERAUCHI, R.; TALBOT, N.J.; VALENT, B. Two distinct

secretion systems facilitate tissue invasion by the rice BLAST fungus *Magnaporthe oryzae*. **Nature Community**. v 4, 2013.

GLOER, J.B. Applications of fungal ecology in the search of new bioactive natural products. In: Wicklow DT, Soderstrom B, eds. *The Mycota IV: Environmental and Microbial Relationships*, pp. 249–268. Berlin, Germany: Springer-Verlag, 1997.

GOODSWEN, S.J.; KENNEDY, P.J.; ELLIS, J.T. Evaluating High-Throughput *Ab Initio* Gene Finders to Discover Proteins Encoded in Eukaryotic Pathogen Genomes Missed by Laboratory Techniques. **PLoS ONE** v 7, 2012, e50609. doi:10.1371/journal.pone.0050609.

GOODWIN, SB.; M'BAREK, SB.; DHILLON, B.; WITTENBERG, AH.; CRANE, CF.; HANE, JK.; FOSTER, AJ.; VAN DER LEE, TA.; GRIMWOOD, J.; AERTS, A.; ANTONIW, J.; BAILEY, A.; BLUHM, B.; BOWLER, J.; BRISTOW, J.; VAN DER BURGT, A.; CANTO-CANCHÉ, B.; CHURCHILL, AC.; CONDE-FERRÀEZ, L.; COOLS, HJ.; COUTINHO, PM.; CSUKAI, M.; DEHAL, P.; DE WIT, P.; DONZELLI, B.; VAN DE GEEST, HC.; VAN HAM, RC.; HAMMOND-KOSACK, KE.; HENRISSAT, B.; KILIAN, A.; KOBAYASHI, AK.; KOOPMANN, E.; KOURMPETIS, Y.; KUZNIAR, A.; LINDQUIST, E.; LOMBARD, V.; MALIEPAARD, C.; MARTINS, N.; MEHRABI, R.; NAP, JP.; PONOMARENKO, A.; RUDD, JJ.; SALAMOV, A.; SCHMUTZ, J.; SCHOUTEN, HJ.; SHAPIRO, H.; STERGIPOULOS, I.; TORRIANI, SF.; TU, H.; DE VRIES, RP.; WAALWIJK, C.; WARE, SB.; WIEBENGA, A.; ZWIERS, LH.; OLIVER, RP.; GRIGORIEV, IV.; KEMA, GH.; Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. **PLoS Genetics**. v 7, p , 2011.

GOSTINČAR, C.; OHM, RA.; KOGEJ, T.; SONJAK, S.; TURK, M.; ZAJC, J.; ZALAR, P.; GRUBE, M.; SUN, H.; HAN, J.; SHARMA, A.; CHINIQUY, J.; NGAN, CY.; LIPZEN, A.; BARRY, K.; GRIGORIEV, IV.; GUNDE-CIMERMAN, N.; Genome sequencing of four *Aureobasidium pullulans* varieties: biotechnological potential, stress tolerance, and description of new species. **BMC Genomics**. v 15, 2014. Disponível em: <<http://bmcbgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-549>> Acesso em: sep. 2015.

GOUT, L.; FUDAL, I.; KUHN, M.L.; BLAISE, F.; ECKERT, M.; CATTOLICO, L.; BALESDENT, M.H.; ROUXEL, T. Lost in the middle of nowhere: the AvrLm1 avirulence gene of the Dothideomycete *Leptosphaeria maculans*. **Molecular Microbiology**. v 60, p 67-80, 2006.

GRILLO, G.; TURI, A.; LICCIULLI, F.; MIGNONE, F. LIUNI, S.; BANFI, S.; GENNARINO, V.A.; HORNER, D.S. PAVESI, G.; PICARDI, E.; PESOLE, G. UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. **Nucleic Acids Research**. v D75–D80, 2010.

GUIGO, R. Assembling genes from predicted *exons* in linear time with dynamic programming. **Journal of Computational Biology**. v 5, p 681–702, 1998.

GUINDON, S.; LETHIEC, F.; DUROUX, P.; GASCUEL, O.; PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. **Nucleic Acids Research**. v 33, p w557-w559, 2005.

HAAS, B.J.; DELCHER, A.L.; MOUNT, S.M.; WORTMAN, J.R.; SMITH JR, R.K.; HANNICK, L.I.; MAITI, R.; RONNING, C.M.; RUSCH, D.B.; TOWN, C.D.; SALZBERG, S.L.; WHITE, O. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. **Nucleic Acids Research**. v 31, p 5654-5666, 2003.

HANE, JK.; LOWE, RG.; SOLOMON, PS.; TAN, KC.; SCHOCH, CL.; SPATAFORA, JW.; CROUS, PW.; KODIRA, C.; BIRREN, BW.; GALAGAN, JE.; TORRIANI, SF.; MCDONALD, BA.; OLIVER, RP.; Dothideomycete plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum*. **Plant Cell**. v , p , 2007.

HARVEY A.L. Natural products in drug discovery. **Drug Discovery Today**, v 13 p 894-901, 2008.

HASHEM M, ALI, E.H. *Epicoccum nigrumas* biocontrol agent of *Pythium* damping-off and root rot of cotton seedlings. **Archives of Phytopathology and Plant Protection**. v 37, p 283-297, 2004.

HAWKSWORTH, D.L. Fungal diversity and its implications for genetic resource collections. **Studies of Mycology**. v 50, p 9–18, 2004.

HAWKSWORTH, D.L.; KIRK, P.M.; SUTTON, B.C.; PEGLER, D.N. Ainsworth & Bisby's Dictionary of the Fungi (eighth ed.), CAB International, Wallingford, 1995.

HECKMAN, D.S.; GEISER, D.M. EIDELL, B.R.; STAUFFER, R.L.; KARDOS, N.L.; HEDGES, S.B. Molecular evidence for the early colonization of land by fungi and plants. **Science**, v 293, p 1129-1133, 2001.

HERTWECK, C. The biosynthetic logic of polyketide diversity. **Angewandte Chemie International Edition** v 48, p 4688-4716, 2009.

HOFFMEISTER, D.; KELLER, N.P. Natural products of filamentous fungi: enzymes, genes, and their regulation. **Natural Product Reports**. v 24, p 393-416, 2007.

HOLT, C.; YANDELL, M. MAKER2: an annotation *pipeline* and genome-database management tool for second-generation genome projects. **BMC Bioinformatics**, v 12, p 491, 2011.

HOOD, M.E.; KATAWCZIK, M.; GIRAUD, T. Repeat-induced point mutation and the population structure of transposable elements in *Microbotryum violaceum*. **Genetics**. v 170, p 1081-1089, 2005.

HORTON, P.; PARK, K.-J.; OBAYASHI, T.; FUJITA, N.; HARADA, H.; ADAMS-COLLIER, C. J.; NAKAI, K. WoLF PSORT: protein localization predictor. **Nucleic Acids Research**. v 35, p W585-587, 2007.

HYATT, D.; Chen, G.L.; Locascio, P.F.; Land, M.L.; Larimer, F.W.; Hauser, L.J. Prodigal: prokaryotic gene recognition and translation initiation site identification. **BMC Bioinformatics**. v 11, p 119, 2010.

IKAWA M, MCGRATTAN, C.J.; BURGE, W.R.; IANNITELLI, R.C.; UEBEL, J.J. Epirocin, a polyene antibiotic from mold *Epicoccum nigrum*. **Journal of Antibiotics**. v 31, p 159–160, 1978.

ISELI, C; JONGENEEL, CV; BUCHER, P.; ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. **Proceedings of the International Conference on Intelligent Systems for Molecular**. p 138-48, 1999.

ISLAM, MS.; HAQUE, MS.; ISLAM, MM.; EMDAD, EM.; HALIM, A.; HOSSEN, QM.; HOSSAIN, MZ.; AHMED, B.; RAHIM, S.; RAHMAN, MS.; ALAM, MM.; HOU, S.; WAN, X.; SAITO, JA.; ALAM, M.; Tools to kill: genome of one of the most destructive plant pathogenic fungi *Macrophomina phaseolina*. **BMC Genomics**. v 13, p , 2012.

JAAFAR, R.; PETTIT, JH.; *Candida albicans*--saprophyte or pathogen? **International Journal of Dermatology**. v 11 , p 783-785 , 1992.

JACOBS, GH; RACKHAM, O; STOCKWELL, PA; TATE, W; BROWN, CM. Transterm: a database of mRNAs and translational control elements. **Nucleic Acids Research**. v 30, p 310-311, 2002.

JEFFRIES, TW.; GRIGORIEV, IV.; GRIMWOOD, J.; LAPLAZA, JM.; AERTS, A.; SALAMOV, A.; SCHMUTZ, J.; LINDQUIST, E.; DEHAL, P.; SHAPIRO, H.; JIN, Y.; PASSOTH, P.; RICHARDSON, PM. Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*. **Nature Biotechnology**. v 25, p 319 - 326, 2007.

JGI - *Colletotrichum somersetensis*. Disponível em: <http://genome.jgi.doe.gov/Colso1/Colso1.info.html>. Acesso em: Janeiro de 2016.

JGI - *Colletotrichum sublineola*. Disponível em: <http://genome.jgi.doe.gov/Colsu1/Colsu1.info.html>. Acesso em: Janeiro de 2016.

JGI - *Colletotrichum zoysiae*. Disponível em: <http://genome.jgi.doe.gov/Colzo1/Colzo1.home.html>. Acesso em: Janeiro de 2016.

JGI. *Aaosphaeria arxii*. Disponível em: <http://genome.jgi.doe.gov/Aaoar1/Aaoar1.home.html> Acesso em: jan. 2016.

JGI. *Acremonium alcalophilum* v2.0. Disponível em: <http://genome.jgi.doe.gov/Acral2/Acral2.home.html> Acesso em: jan. 2016.

JGI. *Alternaria alternata* SRC1lrK2f v1.0. Disponível em: <<http://genome.jgi.doe.gov/Altal1/Altal1.home.html>> Acesso em: jan. 2016.

JGI. *Morchella conica* CCBAS932 v1.0. Disponível em: <<http://genome.jgi.doe.gov/Morco1/Morco1.home.html>> Acesso em: jan. 2016.

JONES, J.D.; DANGL, J.L. The plant immune system. **Nature**. v 444, p 323-329, 2006.

JONES, P.; BINNS, D.; CHANG, H.; FRASER, M.; LI, W.; MCANULLA, C.; MCWILLIAM, H.; MASLEN, J.; MITCHELL, A.; NUKA, G.; PESSEAT, S.; QUINN, AF; SANGRADOR-VEGAS, A.; SCHEREMETJEW, M.; YONG, S.; LOPEZ, R.; HUNTER, S. InterProScan 5: genome-scale protein function classification. **Bioinformatics**. v 30, p 1236-1240, 2014.

KÄLL, L.; KROGH, A.; SONNHAMMER, E.L.L. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. **Nucleic Acids Research**. v 35, p W429-32, 2007.

KANEHISA, M.; GOTO, S.; SATO, Y.; Furumichi, M.; Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. **Nucleic Acids Research**. v 40, p D109-114, 2012.

KAZUTAKA, K.; MISAWA, K.; KUMA, K.; MIYATAA, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. **Nucleic Acids Research**. v 30, p 3059–3066, 2002

KELLER, N.P.; TURNER, G.; BENNETT, J.W. Fungal secondary metabolism – from biochemistry to genomics. **Nature Review Microbiology**. v 3, p 937–947, 2005.

KHATRI, P; DRAGHICI, S. Ontological analysis of gene expression data: current tools, limitations, and open problems. **Bioinformatics**. v 21, p 3587–3595, 2005.

KIM, E.; GOREN, A.; AST, G. Alternative splicing: current perspectives. **Bioessays**. v 30, p 38–47, 2008.

KLUTTS, J.S.; YONEDA, A.; REILLY, M.C.; BOSE, I; DOERING, T.L. Glycosyltransferases and their products: cryptococcal variations on fungal themes. **FEMS Yeast Research**. v 4, p 499-512, 2006.

KNERR, P.J.; VAN DER DONK, W.A. Discovery, biosynthesis, and engineering of lantipeptides. **Annual Review of Biochemistry**. v 81, p 479-505, 2012.

KNUDSEN, S. Promoter 2.0: for the recognition of PolII promoter sequences. **Bioinformatics**. v 15, p 356-361, 1999.

KOHLER, A; KUO, A; NAGY, LG.; MORIN, E.; BARRY, KW.; BUSCOT, F.; CANBÄCK, B.; CHOI, C.; CICHOCKI, N.; CLUM, A.; COLPAERT, J.; COPELAND, A.; COSTA, MD.; DORÉ, J.; FLOUDAS, D.; GAY, G.; GIRLANDA, M.; HENRISSAT, B.; HERRMANN, S.;

HESS, J.; HÖGBERG, N.; JOHANSSON, T.; KHOUJA, HR.; LABUTTI, K.; LAHRMANN, U.; LEVASSEUR, A.; LINDQUIST, EA.; LIPZEN, A.; MARMEISSE, R.; MARTINO, E.; MURAT, C.; NGAN, CY.; NEHLS, U.; PLETT, JM.; PRINGLE, A.; OHM, RA.; PEROTTO, S.; PETER, M.; RILEY, R.; RINEAU, F.; RUYTINX, J.; SALAMOV, A.; SHAH, F.; SUN, H.; TARKKA, M.; TRITT, A.; VENEULT-FOURREY, C.; ZUCCARO A; MYCORRHIZAL GENOMICS INITIATIVE, Consortium.; TUNLID, A.; GRIGORIEV, IV.; HIBBETT, DS.; MARTIN, F.; Convergent losses of decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal mutualists. **Nature Genetics**. v 47, p , 2015.

KOLPAKOV, R; BANA, G; KUCHEROV, G. mreps: Efficient and flexible detection of tandem repeats in DNA. **Nucleic Acids Research**. v 31, p 3672-3678, 2003.

KORF, I. Gene finding in novel genomes. **BMC Bioinformatics**. v 5, p 59, 2004.

KORF, I. Gene finding in novel genomes. **BMC Bioinformatics**. v 5, p S1–S9, 2004.

KROGH, A.; LARSSON, B.; HEIJNE, G.; SONNHAMMER, E.L.L. Predicting Transmembrane Protein Topology with a Hidden MarkovModel: Application to Complete Genomes. *Journal Molecular Biology*. v 305 p 567-580, 2011.

KROKEN, S.; GLASS, N.L.; TAYLOR, J.W.; YODER, O.C.; TURGEON, B.G. Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. **Proceedings of the National Academy of Sciences**. v 100, p 15670-15675, 2003.

KUBICEK, C.P.; STARR, T.L.; GLASS, N.L. Plant cell wall-degrading enzymes and their secretion in plant-pathogenic fungi. **Annual Review of Phytopathology**. v 52, p 427-51, 2014.

KUBICEK, CP; HERRERA-ESTRELLA, A.; SEIDL-SEIBOTH, V.; MARTINEZ, DA.; DRUZHININA, IS.; THON, M.; ZEILINGER, S.; CASAS-FLORES, S.; HORWITZ, BA.; MUKHERJEE, PK.; MUKHERJEE, M.; KREDICS, L.; ALCARAZ, LD.; AERTS, A.; ANTAL, Z.; ATANASOVA, L.; CERVANTES-BADILLO, MG.; CHALLACOMBE, J.; CHERTKOV, O.; MCCLUSKEY, K.; COULPIER, F.; DESHPANDE, N.; VON DÖHREN, H.; EBBOLE, DJ.; ESQUIVEL-NARANJO, EU.; FEKETE, E.; FLIPPHI, M.; GLASER, F.; GÓMEZ-RODRÍGUEZ, EY.; GRUBER, S.; HAN, C.; HENRISSAT, B.; HERMOSA, R.; HERNÁNDEZ-OÑATE, M.; KARAFFA, L.; KOSTI, I.; LE CROM, S.; LINDQUIST, E.; LUCAS, S.; LÜBECK, M.; LÜBECK, PS.; MARGEOT, A.; METZ, B.; MISRA, M.; NEVALAINEN, H.; OMANN, M.; PACKER, N.; PERRONE, G.; URESTI-RIVERA, EE.; SALAMOV, A.; SCHMOLL, M.; SEIBOTH, B.; SHAPIRO, H.; SUKNO, S.; TAMAYO-RAMOS, JA.; TISCH, D.; WIEST, A.; WILKINSON, HH.; ZHANG, M.; COUTINHO, PM.; KENERLEY, CM.; MONTE, E.; BAKER, SE.; GRIGORIEV, IV.; Comparative genome sequence analysis underscores mycoparasitism as the ancestral life style of Trichoderma. **Genome Biology**. v 12, p 56-64, 2011.

KUKREJA, N.; SRIDHARA, S.; SINGH, B.P.; GAUR, S.N.; ARORA, N. Purification and immunological characterization of a12-kDa allergen from *Epicoccum purpurascens*. **FEMS Immunology and Medical Microbiology**. v 56, p 32-40, 2009.

KULDAU, G.; BACON, C. Clavicipitaceous endophytes: Their ability to enhance resistance of grasses to multiple stresses. **Biologic Control**. v 46, p 57–71, 2008.

KUO, C.H.; OCHMAN, H. The extinction dynamics of bacterial pseudogenes. **PLoS Genetics**. v 6, 2010.

LAGESEN, K.; HALLIN, P.F.; RØDLAND, E.; STÆRFELDT, H.H.; ROGNES, T.; USSERY, D.W. RNAmmer: consistent annotation of rRNA genes in genomic sequences. **Nucleic Acids Research**. v 35 p 3100-3108, 2007.

LANGMEAD, B.; TRAPNELL, C.; POP, M.; SALZBERG, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. **Genome Biology**. v 10, p R25, 2009.

LAW, M.; CHILDS, K.L.; CAMPBELL, M.S.; STEIN, J.C.; HOLT, C.; PANCHY, N.; YANDELL, M. Automated update, revision and quality control of the Zea mays genome annotations using MAKER-P improves the B73 RefGen_v3 gene models and identifies new gene models. **Plant Physiology**. v 167, p 25-39, 2014.

LEPAGE, T.; BRYANT, D.; PHILIPPE, H.; LARTILLOT, N. A general comparison of relaxed molecular clock models. **Molecular Biology and Evolution**. v 24, p 2669-2680, 2007.

LUKASHIN, A.V.; BORODOVSKY, M. GeneMark.hmm: new solutions for gene finding. **Nucleic Acids Research**. v 26, p 1107–1115, 1998.

MA, L.J.; VAN DER DOES, H.C.; BORKOVICH, K.A.; COLEMAN, J.J.; DABOUSSI, M.J.; DI PIETRO, A.; DUFRESNE, M.; FREITAG, M.; GRABHERR, M.; HENRISSAT, B.; HOUTERMAN, P.M.; KANG, S.; SHIM, W.B.; WOLOSHUK, C.; XIE, X.; XU, J.R.; ANTONIW, J.; BAKER, S.E.; BLUHM, B.H.; BREAKSPEAR, A.; BROWN, D.W.; BUTCHKO, R.A.; CHAPMAN, S.; COULSON, R.; COUTINHO, P.M.; DANCHIN, E.G.; DIENER, A.; GALE, L.R.; GARDINER, D.M.; GOFF, S.; HAMMOND-KOSACK, K.E.; HILBURN, K.; HUA-VAN, A.; JONKERS, W.; KAZAN, K.; KODIRA, C.D.; KOEHRSEN, M.; KUMAR, L.; LEE, Y.H.; LI, L.; MANNERS, J.M.; MIRANDA-SAAVEDRA, D.; MUKHERJEE, M.; PARK, G.; PARK, J.; PARK, S.Y.; PROCTOR, R.H.; REGEV, A.; RUIZ-ROLDAN, M.C.; SAIN, D.; SAKTHIKUMAR, S.; SYKES, S.; SCHWARTZ, D.C.; TURGEON, B.G.; WAPINSKI, I.; YODER, O.; YOUNG, S.; ZENG, Q.; ZHOU, S.; GALAGAN, J.; CUOMO, C.A.; KISTLER, H.C.; REP, M.; Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. **Nature**. v 464, p 367-373, 2010.

MAIYA, S.; GRUNDMANN, A.; LI, S.M.; TURNER, G. Chembiochem. **European journal of chemical biology**. v 3, p 1062-1069, 2006.

MAJOROS, W.H.; PERTEA, M.; SALZBERG, S.L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. **Bioinformatics**. v 20, p 2878–2879, 2004.

MANNING, VA.; PANDELOVA, I.; DHILLON, B.; WILHELM, LJ.; GOODWIN, SB.; BERLIN, AM.; FIGUEROA, M.; FREITAG, M.; HANE, JK.; HENRISSAT, B.; HOLMAN, WH.; KODIRA, CD.; MARTIN, J.; OLIVER, RP.; ROBBERTSE, B.; SCHACKWITZ, W.; SCHWARTZ, DC.; SPATAFORA, JW.; TURGEON, BG.; YANDAVA, C.; YOUNG, S.; ZHOU, S.; ZENG, Q.; GRIGORIEV, IV.; MA, LJ.; CIUFFETTI, LM.; Comparative genomics of a plant-pathogenic fungus, *Pyrenophora tritici-repentis*, reveals transduplication and the impact of repeat elements on pathogenicity and population divergence. **G3 (Bethesda)**. v 3, p , 2013.

MAPARI, S.A.S; THRANE, U.; MEYER A.S. Fungal polyketide azaphilone pigments as future natural food colorants? Trends in Biotechnology (*In press*), 2010.

MARCHLER-BAUER, A.; PANCHENKO, A.R.; SHOEMAKER, B.A.; THIESSEN, P.A.; GEER, L.Y.; BRYANT, S.H. "CDD: a database of conserved domain alignments with links to domain three-dimensional structure." **Nucleic Acids Research**. v 30 p 281-283, 2002.

MARTIN, F; KOHLER, A.; MURAT, C.; BALESTRINI, R.; COUTINHO, PM.; JAILLON, O.; MONTANINI, B.; MORIN, E.; NOEL, B.; PERCUDANI, R.; PORCEL, B.; RUBINI, A.; AMICUCCI, A.; AMSELEM, J.; ANTHOUARD, V.; ARCIONI, S.; ARTIGUENAVE, F.; AURY, JM.; BALLARIO, P.; BOLCHI, A.; BRENNNA, A.; BRUN, A.; BUÉE, M.; CANTAREL, B.; CHEVALIER, G.; COULOUX, A.; DA SILVA, C.; DENOEUDE, F.; DUPLESSIS, S.; GHIGNONE, S.; HILSELBERGER, B.; IOTTI, M.; MARÇAIS, B.; MELLO, A.; MIRANDA, M.; PACIONI, G.; QUESNEVILLE, H.; RICCONI, C.; RUOTOLO, R.; SPLIVALLO, R.; STOCCHI, V.; TISSERANT, E.; VISCOMI, AR.; ZAMBONELLI, A.; ZAMPIERI, E.; HENRISSAT, B.; LEBRUN, MH.; PAOLOCCI, F.; BONFANTE, P.; OTTONELLO, S.; WINCKER, P.; Périgord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis. **Nature**. v 464, p 1033-8, 2010.

MARTIN, F.; AERTS, A.; AHR, D.; BRUN, A.; DANCHIN, E. G. J.; DUCHAUSSOY, F.; GIBON, J.; KOHLER, A.; LINDQUIST, E.; PEREDA, V.; SALAMOV, A.; SHAPIRO, H. J. ; WUYTS, J.; BLAUDEZ, D. BU, M.; BROKSTEIN, P. CANB, B. ; COHEN, D.; COURTY, P.E.; COUTINHO, P. M.; DELARUELLE, C. ; DETTER, J. C. ; DEVEAU, A.; DIFAZIO, S.; DUPLESSIS, S.; FRAISSINET-TACHET, L.; LUCIC, E.; FREY-KLETT, P.; FOURREY, C.; FEUSSNER, I.; GAY, G.; GRIMWOOD, J.; HOEGGER, P. J.; JAIN, P.; KILARU, S.; LABB, J.; LIN, Y. C.; LEGU, V.; LE TACON, F.; MARMEISSE, R.; MELAYAH, D.; MONTANINI, B.; MURATET, M.; NEHLS, U.; NICULITA-HIRZEL, H.; OUDOT-LE SECQ, M. P. ; PETER, M.; QUESNEVILLE, H.; RAJASHEKAR, B.; REICH, M.; ROUHIER, N.; SCHMUTZ, J.; YIN,T.; CHALOT, J.; HENRISSAT, B.; LUCAS, S.; VAN DE PEER, Y.; PODILA, G. K.; POLLE, A.; PUKKILA, P. J.; RICHARDSON, P. M.; ROUZ, P.; SANDERS, I. R.; STAJICH, J. E.; TUNLID, A.; TUSKAN G.; GRIGORIEV, I. V. The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. **Nature**. v 452, p 88-92, 2008.

MEERUPATI, T.; ANDERSSON, KM.; FRIMAN, E.; KUMAR, D.; TUNLID, A.; AHRÉN, D.; Genomic mechanisms accounting for the adaptation to parasitism in nematode-trapping fungi. **PLoS Genetics**. v 9, p , 2013.

MEINHARDT, L.W.; COSTA, G.G.; THOMAZELLA, D.P.; TEIXEIRA, P.J.; CARAZZOLLE, M.F.; SCHUSTER, S.C.; CARLSON, J.E.; GUILTINAN, M.J.; MIECZKOWSKI, P.; FARMER, A.; RAMARAJ, T.; CROZIER, J.; DAVIS, R.E.; SHAO, J.; MELNICK, R.L.; PEREIRA, G.A.; BAILEY, B.A. Genome and secretome analysis of the hemibiotrophic fungal pathogen, *Moniliophthora roreri*, which causes frosty pod rot disease of cacao: mechanisms of the biotrophic and necrotrophic phases. **BMC Genomics**. v 15, 2014.

MORAIS DO AMARAL, A.; ANTONIW, J.; RUDD, J.J.; HAMMOND-KOSACK, K.E. Defining the Predicted Protein Secretome of the Fungal Wheat Leaf Pathogen *Mycosphaerella graminicola*. **PLoS ONE**. v 12, 2012.

MORALES-CRUZ, A; AMRINE, KC.; BLANCO-ULATE, B.; LAWRENCE, DP.; TRAVADON, R.; ROLSHAUSEN, PE.; BAUMGARTNER, K.; CANTU, D.; Distinctive expansion of gene families associated with plant cell wall degradation, secondary metabolism, and nutrient uptake in the genomes of grapevine trunk pathogens. **BMC Genomics**. v 16, p , 2015.

NAKAYA, A.; KATAYAMA, T.; ITOH, M.; HIRANUKA, K.; KAWASHIMA, S.; MORIYA, Y.; OKUDA, S.; TANAKA, M.; TOKIMATSU, T.; YAMANISHI, Y.; YOSHIZAWA, A.C.; KANEHISA, M.; GOTO, S. KEGG OC: a large-scale automatic construction of taxonomy-based ortholog *clusters*. **Nucleic Acids Research**. v 41, p D353–D357, 2013.

NAWROCKI, E.P.; EDDY, S.R. Infernal 1.1: 100-fold faster RNA homology searches. **Bioinformatics**. v 29, p 2933-2935, 2013.

NOVAES, J.; RANGEL, L.T.; FERRO, M.; ABE, R.Y.; MANHA, A.P.; DE MELLO, J.C.; VARUZZA, L.; DURHAM, A.M.; MADEIRA, A.M.; GRUBER, A. A comparative transcriptome analysis reveals expression. profiles conserved across three *Eimeria* spp. of domestic fowl and associated with multiple developmental stages. **International Journal for Parasitology**. v 42, p 39-48, 2012.

O'CONNELL, RJ.; THON, MR.; HACQUARD, S.; AMYOTTE, SG.; KLEEMANN, J.; TORRES, MF.; DAMM, U.; BUIATE, EA.; EPSTEIN, L.; ALKAN, N.; ALTMÜLLER, J.; ALVARADO-BALDERRAMA, L.; BAUSER, CA.; BECKER, C.; BIRREN, BW.; CHEN, Z.; CHOI, J.; CROUCH, JA.; DUVICK, JP.; FARMAN, MA.; GAN, P.; HEIMAN, D.; HENRISSAT, B.; HOWARD, RJ.; KABBAGE, M.; KOCH, C.; KRACHER, B.; KUBO, Y.; LAW, AD.; LEBRUN, MH.; LEE, YH.; MIYARA, I.; MOORE, N.; NEUMANN, U.; NORDSTRÖM, K.; PANACCIONE, DG.; PANSTRUGA, R.; PLACE, M.; PROCTOR, RH.; PRUSKY, D.; RECH, G.; REINHARDT, R.; ROLLINS, JA.; ROUNSLEY, S.; SCHARDL, CL.; SCHWARTZ, DC.; SHENOY, N.; SHIRASU, K.; SIKHAKOLLI, UR.; STÜBER, K.; SUKNO, SA.; SWEIGARD, JA.; TAKANO, Y.; TAKAHARA, H.; TRAIL, F.; VAN DER

DOES, HC.; VOLL, LM.; WILL, I.; YOUNG, S.; ZENG, Q.; ZHANG, J.; ZHOU, S.; DICKMAN, MB.; SCHULZE-LEFERT, P.; VER LOREN VAN THEMAAT, E.; MA, LJ.; VAILLANCOURT, LJ. Lifestyle transitions in plant pathogenic *Colletotrichum* fungi deciphered by genome and transcriptome analyses. **Nature Genetics**. v 44, p 1060-1065, 2012.

OHM, RA.; FEAU, N.; HENRISSAT, B.; SCHOCH, CL.; HORWITZ, BA.; BARRY, KW.; CONDON, BJ.; COPELAND, AC.; DHILLON, B.; GLASER, F.; HESSE, CN.; KOSTI, I.; LABUTTI, K.; LINDQUIST, EA.; LUCAS, S.; SALAMOV, AA.; BRADSHAW, RE.; CIUFFETTI, L.; HAMELIN, RC.; KEMA, GH.; LAWRENCE, C.; SCOTT, JA.; SPATAFORA, JW.; TURGEON, BG.; DE WIT, PJ.; ZHONG, S.; GOODWIN, SB.; GRIGORIEV, IV.; Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. **PLoS Pathogens**. v 8, 2012. Disponível em: <http://www.plospathogens.org-images-ppat_120x30.png> Acesso em: sep. 2015.

PARISI, V.; FONZO, V.; ALUFFI-PENTINI, F. STRING: finding tandem repeats in DNA sequences. **Bioinformatics**. v 19, p 1733–1738, 2003.

PARRA, G.; BLANCO, E.; GUIGO, R. GeneID in *Drosophila*. **Genome Research**. v 10, p 511–515, 2000.

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

PETERSON, K.J.; LYONS, J.B.; NOWAK, K.S.; TAKACS, C.M.; WARGO, M.J.; MCPEEK, M.A. Estimating metazoan divergence times with a molecular clock. **Proceedings of the National Academy of Sciences of the United States of America**, v 101, p 6536–6541, 2004.

PIECKENSTAIN, F.L.; BAZZALO, M.E.; ROBERTS, A.M.I.; UGALDE, R.A. *Epicoccum purpurascens* for biocontrol of *Sclerotinia* head rot of sunflower. **Mycology Research**. 105, p 77-84, 2001.

PIERLEONI, A.; MARTELLI, P.L.; CASADIO, R. PredGPI: a GPI-anchor predictor. **BMC Bioinformatics**. v 9, p 392, 2008.

POWELL, S.; FORSLUND, K.; SZKLARCZYK, D.; TRACHANA, K.; ROTH, A.; HUERTA-CEPAS, J.; GABALDÓN, T.; RATTEI, T.; CREEVEY, C.; KUHN, M.; JENSEN, L.J.; MERING, C.; BORK, P. eggNOG v4.0: nested orthology inference across 3686 organisms. **Nucleic Acids Research**. v 42, p D231–D239, 2014.

PROSDOCIMI, F.; CERQUEIRA, G.C.; BINNECK, E.; SILVA, A. F.; REIS, A. N.; JUNQUEIRA, A.C.M.; SANTOS, A.C.F.; NHANI JÚNIOR, A.; WUST, C.I.; FILHO F.C.; KESSEDJIAN, J.L.; PETRETSKI, J.H.; CAMARGO, L.P.; FERREIRA, R.G.M.; LIMA, R.P.; PEREIRA, R.M.; JARDIM, S.; SAMPAIO, V.S., FOLGUERAS, A.V. Bioinformática: Manual do Usuário. **Biotecnologia Ciência & Desenvolvimento** (Encarte especial). v 29, p 12-25, 2002.

PUSZTAHELYI, T.; HOLB, I.J.; PÓCSI, I. Secondary metabolites in fungus-plant interactions. **Frontiers in Plant Science**. v 6, 2015.

RODRIGUEZ, R.; REDMAN, R. More than 400 million years of evolution and some plants still can't make it on their own: plant stress tolerance via fungal symbiosis. **The Journal of Experimental Botany**. v 59, p 1109-1114, 2008.

ROOHPARVAR, R.; DE WAARD, M.A.; KEMA, G.H.; ZWIERS, L.H. MgMfs1, a major facilitator superfamily transporter from the fungal wheat pathogen *Mycosphaerella graminicola*, is a strong protectant against natural toxic compounds and fungicides. **Fungal Genetics Biology**. v 44, p 378-388, 2007.

ROUXEL, T; GRANDAUBERT, J; HANE, JK.; HOEDE, C.; VAN DE WOUW, AP.; COULOUX, A.; DOMINGUEZ, V.; ANTHOUARD, V.; BALLY, P.; BOURRAS, S.; COZIJNSEN, AJ.; CIUFFETTI, LM.; DEGRAVE, A.; DILMAGHANI, A.; DURET, L.; FUDAL, I.; GOODWIN, SB.; GOUT, L.; GLASER, N.; LINGLIN, J.; KEMA, GH.; LAPALU, N.; LAWRENCE, CB.; MAY, K.; MEYER, M.; OLLIVIER, B.; POULAIN, J.; SCHOCH, CL.; SIMON, A.; SPATAFORA, JW.; STACHOWIAK, A.; TURGEON, BG.; TYLER, BM.; VINCENT, D.; WEISSENBAACH, J.; AMSELEM, J.; QUESNEVILLE, H.; OLIVER, RP.; WINCKER, P.; BALESDENT, MH.; HOWLETT, BJ.; Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. **Nature Communications**. v 2, p 202, 2011.

RUBIO-PIÑA, J.A. ZAPATA-PÉREZ, O. Isolation of total RNA from tissues rich in polyphenols and polysaccharides of mangrove plants. **Journal of Biotechnology**. 14 <http://dx.doi.org/10.2225/vol14-issue5-fulltext-10>. 2011.

SAIKIA, S.; TAKEMOTO, D.; TAPPER, B. A.; LANE, G. A.; FRASER, K.; SCOTT, B. Functional analysis of an indole-diterpene gene *cluster* for lolitrem B biosynthesis in the grass endosymbiont *Epichloë festucae*. **FEBS Letters**. v 586, p 2563-2569, 2012.

SALAMOV, A.A.; SOLOVYEV, V.V. Ab initio gene finding in *Drosophila* genomic DNA. **Genome Research**. v 10, p 516-522, 2000.

SALICHOS, L.; ROKAS, A.; Inferring ancient divergences requires genes with strong phylogenetic signals. **Nature**. v 497, p 327-31, 2013.

SANDERSON, M.J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. **Bioinformatics**. v 19, p 301-302, 2003.

SANTOS, F.R.; ORTEGA, J.M. Bioinformática aplicada à Genômica. Manuscritos para capítulo do Biowork IV - Melhoramento Genômico. Belo Horizonte, MG. 2003.

SCHATTNER, P.; BROOKS, A.N.; LOWE, T.M. The tRNAscan-SE, snoscan and snoGPS Web servers for the detection of tRNAs and snoRNAs. **Nucleic Acids Research**. v 33, p W686-9, 2005.

SCHERLACH, K.; HERTWECK, C. Triggering cryptic natural product biosynthesis in microorganisms. **Organic & Biomolecular Chemistry**, v 7, p 1753-1760, 2009.

SCHIERWATER, B; EITEL, M; JAKOB, W.; OSIGUS, HJ.; HADRY, H.; DELLAPORTA, SL.; KOLOKOTRONIS, SO.; DESALLE, R. Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. **PLoS Biology**. v 7, p e20, 2009.

SCHMID, F, STONE, B.A.; BROWNLEE, R.T.C.; MCDUGALL, B.M.; SEVIOUR, R.J. Structure and assembly of epiglucan, the extracellular X-glucan produced by the fungus *Epicoccum nigrum* strain F19. **Carbohydrate Research**, v 341, p 365-373, 2006.

SCHULZ, B.; BOYLE, C. The endophytic continuum. **Mycology Research**. v 109, p 661-686, 2005.

SCHUMANN, J.; HERTWECK, C. Advances in cloning.; functional analysis and heterologous expression of fungal polyketide synthase genes. **Journal of Biotechnology**, v 124, p 690-703, 2006.

1 SCHWARTZ, D.; BERGER, S.; HEINZELMANN, E.; MUSCHKO, K.; WELZEL, K.; WOHLLEBEN, W. Biosynthetic Gene Cluster of the Herbicide Phosphinothricin Tripeptide from *Streptomyces viridochromogenes* Tü494. **Applied Environmental Microbiology**. v 70, p 7093-7102, 2004.

SEIDL, M.F.; ACKERVEKEN, G.V.; GOVERS, F.; SNEL, B. A Domain-Centric Analysis of Oomycete Plant Pathogen Genomes Reveals Unique Protein Organization. **Plant Physiology**. v 155, p 628-644, 2011.

SEYMOUR, F.A.; CRESSWELL, J.E.; FISHER, P.J.; LAPPIN-SCOTT, H.M.; HAAG, H.; TALBOT, N.J. The influence of genotypic variation on metabolite diversity in populations of two endophytic fungal species. **Fungal Genetics Biology**. v 41, p 721-734, 2004.

SLOT, J. C.; ROKAS, A. Multiple GAL pathway gene clusters evolved independently and by different mechanisms in fungi. *Proceedings of the National Academy of Sciences*. v 107, p 10136-10141, 2010.

SMIT, A.F.A; HUBLEY, R.; GREEN, P. RepeatMasker Open-4.0. 2013-2015. Conteúdo disponível em: <<http://www.repeatmasker.org>>. Data de acesso 10 de outubro de 2015.

SOLOVYEV, V.; KOSAREV, P.; SELEDOV, I.; VOROBYEV, D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. **Genome Biology**. v 7, p 10.1-10.12, 2006.

SPERSCHNEIDER, J.; DODDS, P.N.; GARDINER, D.M.; MANNERS, J.M.; SINGH, K.B.; TAYLOR, J.M. Advances and Challenges in Computational Prediction of Effectors from Plant Pathogenic Fungi. **PLoS Pathogens**. v 11, 2015.

STANKE, M; BURKHARD, M. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints **Nucleic Acids Research**. v 33, 2005.

STANKE, M. Tutorial on Gene Prediction with AUGUSTUS. Conteúdo disponível em: <http://www.molecularrevolution.org/molevolfiles/exercises/augustus/training.html> Acesso: Janeiro de 2016.

STANKE, M.; DIEKHANS, M.; BAERTSCH, R.; HAUSSLER, D. Sequence analysis using native and syntenically mapped cDNA alignments to improved de novo gene finding. **Bioinformatics**. v 24, p 637-644 2008.

STANKE, M.; STEINKAMP, R.; WAACK, S.; MORGENSTERN, B. AUGUSTUS: a web server for gene finding in eukaryotes. **Nucleic Acids Research**. v 32, p W309-W312, 2004.

STANKE, M.; WAACK, S. Gene prediction with a hidden Markov model and a new *intron* submodel. **Bioinformatics**. v 19, p II215-II225, 2003.

STERGIOPOULOS, I.; DE WIT, P.J...; Fungal effector proteins. **Annual Review of Phytopathology**. v 47, p 233-263, 2009.

STIERLE, A.; STROBEL, G.; STIERLE, D. Taxol and taxane production by *Taxomyces andreanae*, an endophytic fungus of Pacific yew. **Science**. v 260, p 214-216, 1993.

STUKENBROCK, EH; CHRISTIANSEN, FB; HANSEN, TT.; DUTHEIL, JY.; SCHIERUP, MH.; Fusion of two divergent fungal individuals led to the recent emergence of a unique widespread pathogen species. **PNAS USA**. v 109, p 10954-9, 2012.

TALAVERA, G.; CASTRESANA, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. **Systematic Biology**. v 56, p 564-577, 2007.

TARAILO-GRAOVAC, M.; CHEN, N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. **Current Protocol Bioinformatics**. v 25, p 10.1-4.10.14, 2009.

TER-HOVHANNISYAN, V.; LOMSADZE, A.; CHERNOFF, YO.; BORODOVSKY, M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. **Genome Research**. v 18, p 1979-1990, 2008.

THON, M. A set of scripts to help automate the construction of data sets for multi-gene phylogenetic analyses. Conteúdo Disponível: <https://github.com/mthon/mirlo> Acesso em: Nov. 2015.

TOYOMASU, T.; KANEKO, A.; TOKIWANO, T.; KANNO, Y.; KANNO, Y.; NIIDA, R.; MIURA, S.; NISHIOKA, T.; IKEDA, C.; MITSUHASHI, W.; DAIRI, T.; KAWANO, T.; OIKAWA, H.; KATO, N.; SASSA, T. Biosynthetic gene-based secondary metabolite screening: a new diterpene, methyl phomopsenonate, from the fungus *Phomopsis amygdale*. **The Journal of Organic Chemistry**, v 74, p 1541-1548, 2009.

VAN DONGEN, S. "Graph clustering by flow simulation." Ph.D thesis, University of Utrecht, The Netherlands, 2000.

VERNIKOS, G.S.; PARKHILL, J. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. **Bioinformatics**. v 22, p 2196-203, 2006.

VOGEL, K.J.; MORAN, N.A. Functional and Evolutionary Analysis of the Genome of an Obligate Fungal Symbiont. **Genome Biology Evolution**. v 5, 891-904, 2013.

WANG, D.Y.C.; KUMAR, S.; HEDGES, S.B. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. **Proceedings of the Royal Society of London Series B-Biological Sciences**. v 266, p 163-171, 1999.

WANGUN, H.V.K.; Hertweck, C. Epicoccarines A, B and epipyridone: tetramic acids and pyridone alkaloids from an *Epicoccum* sp. associated with the tree fungus *Pholiotas quarrosa*. **Organic & Biomolecular Chemistry**. v 5, p 1702-1705, 2007.

WEBER, T.; BLIN, K.; DUDELA, S.; KRUG, D.; KIM, K.U.; BRUCCOLERI, R.; LEE, S.Y.; FISCHBACH, M.A.; MÜLLER, R.; WOHLLEBEN, W.; BREITLING, R.; TAKANO, E.; MEDEMA, M.H. AntiSMASH 3.0 — a comprehensive resource for the genome mining of biosynthetic gene clusters. **Nucleic Acids Research**. 2015.

WELCH, J.J.; BROMHAM, L. Molecular dating when rates vary. **Trends in Ecology & Evolution**. v 20, pp. 320-327, 2005.

WÖSTEMEYER, J; KREIBICH, A. Repetitive DNA elements in fungi (Mycota): impact on genomic architecture and evolution. **Current Genetics**. v 41, p 189-198, 2002.

XU, XH.; SU, ZZ.; WANG, C.; KUBICEK, CP.; FENG, XX.; MAO, LJ.; WANG, JY.; CHEN, C.; LIN, FC.; ZHANG, CL.; The rice endophyte *Harpophora oryzae* genome reveals evolution from a pathogen to a mutualistic endophyte. **Scientific Reports**. v 4, p , 2014.

YADAV, G.; GOKHALE, R.S.; MOHANTY, D. SEARCHPKS: A program for detection and analysis of polyketide synthase domains. **Nucleic Acids Research**. v 31, p 3654-3658, 2003.

YANDELL, M.; ENCE, E. A beginner's guide to eukaryotic genome annotation. **Nature Reviews Genetics**. v 13, p 329-342, 2012.

YANG, J.; WANG, L.; JI, X.; FENG, Y.; LI, X.; ZOU, C.; XU, J.; REN, Y.; MI, Q.; WU, J.; LIU, S.; LIU, Y.; HUANG, X.; WANG, H.; NIU, X.; LI, J.; LIANG, L.; LUO, Y.; JI, K.; ZHOU, W.; YU, Z.; LI, G.; LIU, Y.; LI, L.; QIAO, M.; FENG, L.; ZHANG, KQ.; Genomic and proteomic analyses of the fungus *Arthrotrichum oligospora* provide insights into nematode-trap formation. **PLoS Pathogens**. v 7, p , 2011.

YOUNG, CA.; CHARLTON, ND.; TAKACH, JE.; SWOBODA, GA.; TRAMMELL, MA.; HUHMAN, DV.; HOPKINS, AA. Characterization of *Epichloë coenophiala* within the US: are all tall fescue endophytes created equal? **Frontiers in Chemistry**. v 2, p , 2014.

YU, N.Y.; WAGNER, J.R.; LAIRD, M.R.; MELLI, G.; REY, S.; LO, R.; BRINKMAN, F.S.L. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. **Bioinformatics**. v 26, p 1608-1615, 2010.

ZERIKLY.; M; CHALLIS.; GL. Strategies for the discovery of new natural products by genome mining. **Chembiochem**. v 10, p 625–633, 2009.

ZHANG, Y.; Liu, S.; Che, Y.; Liu, X. Epicoccins A-D, epipolythiodioxopiperazines from a *Cordyceps*-colonizing isolate of *Epicoccum nigrum*. **Journal of Natural Products**. v 70, p 1522-1525, 2007.

ZHAO, Z.; LIU, H.; WANG, C.; X.U, J.R. Correction: Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. **BMC Genomics**. v 15, 2014.

ZHENG, A.; LIN, R.; ZHANG, D.; QIN, P.; XU, L.; AI, P.; DING, L.; WANG, Y.; CHEN, Y.; LIU, Y.; SUN, Z.; FENG, H.; LIANG, X.; FU, R.; TANG, C.; LI, Q.; ZHANG, J.; XIE, Z.; DENG, Q.; LI, S.; WANG, S.; ZHU, J.; WANG, L.; LIU, H.; LI, P.; The evolution and pathogenic mechanisms of the rice sheath blight pathogen. **Nature Communications**. v 4, 2013.

ZHENG, P.; XIA, Y.; XIAO, G.; XIONG, C.; HU, X.; ZHANG, S.; ZHENG, H.; Huang, Y.; ZHOU, Y.; WANG, S.; ZHAO, G.-P.; LIU, X.; ST., LEGER, R.J.; WANG, C. Genome sequence of the insect pathogenic fungus *Cordyceps militaris*, a valued traditional Chinese medicine. **Genome Biology**. v 12, p R116-R116, 2011.

ZUCCARO, A.; LAHRMANN, U.; GÜLDENER, U.; LANGEN, G.; PFIFFI, S.; BIEDENKOPF, D.; WONG, P.; SAMANS, B.; GRIMM, C.; BASIEWICZ, M.; MURAT, C.; MARTIN, F.; KOGEL, KH.; Endophytic life strategies decoded by genome and transcriptome analyses of the mutualistic root symbiont *Piriformospora indica*. **PLoS Pathogens**. v 7, p , 2011.