OSVALDO REIS JUNIOR

Identificação e padrões de expressão de transcritos derivados de RNA-Seq: caracterização molecular do fungo *Moniliophthora perniciosa*, causador da vassoura-de-bruxa do cacaueiro

Campinas

2014

SECRETARIA
DE
PÓS-GRADUAÇÃO
I. B.

UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE BIOLOGIA

OSVALDO REIS JUNIOR

"Identificação e padrões de expressão de transcritos derivados de RNA-Seq: caracterização molecular do fungo Moniliophthora perniciosa, causador da vassoura-de-bruxa do cacaueiro"

Este exemplar corresponde à redação final da DISSERTAÇÃO defendida pelo candidato
OSVALDO REIS JUNIOR
e aprovada pela Comissão Julgadora.

DISSERTAÇÃO apresentada ao Instituto de Biologia da UNICAMP para obtenção do Título de Doutor em Genética e Biologia Molecular, na área de Bioinformática.

Orientador: Prof. Dr. Gonçalo Amarante Guimarães Pereira

CAMPINAS,
2014

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Biologia
Mara Janaina de Oliveira - CRB 8/6972

R277i

Reis Junior, Osvaldo, 1986-
Identificação e padrões de expressão de transcritos derivados de RNA-Seq : caracterização molecular do fungo *Moniliophthora perniciosa*, causador da vassoura-de-bruxa do cacaueiro / Osvaldo Reis Junior. – Campinas, SP : [s.n.], 2014.

Orientador: Gonçalo Amarante Guimarães Pereira.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. RNA-Seq. 2. Transcriptoma. 3. Vassoura-de-bruxa (Fitopatologia). I. Pereira, Gonçalo Amarante Guimarães,1964-. II. Universidade Estadual de Campinas. Instituto de Biologia. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Identification and expression pattern of RNA-Seq-derived transcripts : a molecular characterization of the fungal pathogen *Moniliophthora perniciosa*, which causes witches' broom disease of cocoa tree
Palavras-chave em inglês:
RNA-Seq
Transcriptome
Witches' broom disease
Área de concentração: Bioinformática
Titulação: Mestre em Genética e Biologia Molecular
Banca examinadora:
Gonçalo Amarante Guimarães Pereira [Orientador]
Michel Eduardo Beleza Yamagishi
Marcelo Mendes Brandão
Data de defesa: 25-02-2014
Programa de Pós-Graduação: Genética e Biologia Molecular

Campinas, **25** de **fevereiro** de **2014**

BANCA EXAMINADORA

Prof. Dr. Gonçalo Amarante Guimarães Pereira (orientador) _____
Assinatura

Dr. Michel Eduardo Beleza Yamagishi _____
Assinatura

Prof. Dr. Marcelo Mendes Brandão _____
Assinatura

Dr. Francisco Pereira Lobo _____
Assinatura

Dra. Katlin Massirer _____
Assinatura

RESUMO

As tecnologias de sequenciamento de segunda geração geram um grande número de sequências em um curto espaço de tempo e com baixo custo. O sequenciamento em larga escala de cDNA, conhecido como RNA-seq, tem permitido análises precisas de transcriptomas inteiros sem a necessidade de conhecimento prévio sobre sequências genômicas, conforme exigido pela tecnologia de microarranjos. No entanto, existem muitas discussões sobre os métodos utilizados para o alinhamento de *reads*, identificação de genes diferencialmente expressos e montagem de transcriptomas. Desde 2000, o Laboratório de Genômica e Expressão (LGE), tem estudado a doença da vassoura-de-bruxa do cacaueiro (*Theobroma cacao*), que é causada pelo fungo basidiomiceto *Moniliophthora perniciosa*. Recentemente, 54 bibliotecas RNA-seq da interação planta-patógeno foram sequenciadas pelo nosso grupo, a fim de ajudar na elucidação da complexa biologia da doença. Desta forma, este trabalho tem como objetivo gerar informações sobre o perfil de transcrição do *M. perniciosa* e do *T. cacao* durante a doença vassoura-de-bruxa. Os resultados estão divididos em três seções principais, sendo que na primeira apresentamos uma análise de alinhamento e expressão gênica nas condições e tecidos amostrados. Na segunda, analisamos o estágio inicial da doença, conhecido como vassoura-verde, onde através da análise de expressão diferencial identificamos genes relacionados aos mecanismos de interação entre o cacaueiro e o fungo *M. perniciosa*. Na ultima seção, desenvolvemos uma estratégia para identificar sequências de possíveis RNAs não codificantes longos (lncRNA) e aplicamos esta estratégia no fungo *M. perniciosa*. De uma maneira geral estes dados apresentam um importante avanço no estudo desta doença e poderão ser utilizados em trabalhos futuros.

ABSTRACT

Next-generation sequencing technologies generate large numbers of sequences in a short time and at low cost. The high-throughput sequencing of cDNA, known as RNA-seq, has allowed precise analyses of entire transcriptomes without the need of previous knowledge on genomic sequences, as required by microarrays. However, there are many discussions about the methods used for read alignment, identification of differentially expressed genes and assembly of transcriptomes. Since 2000, the Genomics and Expression Laboratory (LGE), has been studying the Witches' Broom Disease (WBD) of cacao (*Theobroma cacao*), which is caused by the basidiomycete fungus *Moniliophthora perniciosa*. Recently, 54 RNA-seq libraries of this plant-pathogen interaction were sequenced by our group in order to help in the elucidation of the complex biology of the disease. Thus, this work aims to generate information about the transcription profile of *M. perniciosa* and *T. cacao* during the Witches' Broom Disease (WBD). The results are divided into three main sections, the first is an analysis of alignment and gene expression under the conditions and sampled tissues. In the second section, we analyze the initial stage of the disease, known as green broom, where through differential expression analysis we could identify genes related to mechanisms of interaction between cacao and the fungus *M. perniciosa*. In the last section, we developed a strategy to identify possible sequences of long noncoding RNAs (lncRNA) and applied this strategy in the fungus *M. perniciosa*. In general these data present an important advance in the study of this disease and may be used in future work.

SUMÁRIO

Dedico este trabalho aos meus pais Osvaldo e Irene e a minha irmã Analy.

AGRADECIMENTOS

A minha querida mãe Irene e o meu querido pai Osvaldo, que sempre despertaram em mim a sede pelo conhecimento e não pouparam esforços para que eu concretizasse meus sonhos.

A minha querida irmã Analy, que sempre elevou minha autoestima dizendo a todos que se orgulhava de ter um irmão como eu.

Ao time de bioinformática do LGE, por ter me passado todo o conhecimento necessário para concretizar esse trabalho.

Ao Dr. Paulo José, pela colaboração durante esse trabalho.

Aos amigos Herai, Ramon, Gustavo, Leandro, Melline e Marcelo, pela amizade durante os últimos anos.

Agradeço ao meu orientador Gonçalo Pereira pela confiança.

Agradeço à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) por financiar este trabalho (processo 2011/03050-3).

A todos o meu muito obrigado.

1 INTRODUÇÃO

Transcriptoma é o conjunto de todas as moléculas de RNA presentes em um organismo, tecido ou célula num determinado momento ou fase de desenvolvimento. Através da análise de transcriptomas, é possível determinar quais genes são expressos e em quais concentrações em diversos tipos de células ou tecidos, submetidos a diversos tratamentos, ou estresses bióticos e abióticos, entre outras aplicações. Em geral, a análise de transcriptomas inclui: (i) identificar os diversos tipos de transcritos como RNAs mensageiros (mRNA) e RNAs não codificantes (ncRNA); (ii) determinar a estrutura dos genes (início, fim, posição dos éxons e íntrons); (iii) quantificar a concentração de cada transcrito em diversas condições (Wang, et al., 2009).

Diversas ferramentas para análise de transcriptomas estão disponíveis há bastante tempo como o PCR de transcrição reversa (RT-PCR), etiquetas de sequências expressas (ESTs), e análise serial de expressão gênica (SAGE). Entretanto, a análise de transcriptomas em larga escala só foi possível a partir do surgimento da tecnologia de microarranjos (Schena, et al., 1995).

Os microarranjos para expressão gênica são *chips* onde são colocadas milhares de pequenas sequências de DNA, chamadas de sondas, complementares aos transcritos conhecidos de um genoma. Com as tecnologias atuais, é possível colocar em um único chip sondas para todos os transcritos conhecidos de um organismo. Para análise de expressão, os transcritos são extraídos das amostras, convertidos em fitas de DNA complementar (cDNA) e, finalmente marcados com marcadores fluorescentes e espalhados sobre o *chip*. Os cDNAs complementares às sequências sondas hibridizam-se e emitem fluorescência, que é medida por um *scanner*, sendo possível inferir o nível de expressão do transcrito correspondente a cada sonda de acordo com a intensidade do sinal de fluorescência emitido por cada região do chip. Apesar de representar um grande avanço no estudo de transcriptomas os microarranjos possuem diversos vieses que levam a uma variação técnica alta, mesmo dentro da mesma plataforma (Shi, et al., 2006). Entretanto, atualmente estes vieses já são conhecidos e diversos métodos foram desenvolvidos para lidar com eles.

Recentemente, o sequenciamento em larga escala de cDNA (RNA-Seq) surgiu como uma tecnologia poderosa para análises de transcriptomas e possui diversas vantagens sobre os microarranjos. Primeiramente, ao contrário dos microarranjos que são limitados a um número finito de sondas de sequências conhecidas, o RNA-seq não depende de um conhecimento prévio do genoma ou dos transcritos de um organismo. Ademais, o RNA-Seq, devido ao grande tamanho da amostragem, é capaz de detectar quase todos os transcritos de uma célula em uma determinada condição, sendo considerada uma tecnologia bem atrativa para organismos que não tem a sequência genômica determinada (Strickler, et al., 2012). A segunda vantagem do RNA-seq é que ele é altamente sensível à variação de expressão, sendo capaz de identificar transcritos pouco abundantes e detectar pequenas variações de expressão (Marguerat and Bähler, 2010). Estas vantagens do RNA-seq têm possibilitado mostrar a complexidade dos transcriptomas, tanto em procariotos (van Vliet, 2010) como em eucariotos (Wang, et al., 2009).

Contudo, apesar de ser uma tecnologia poderosa, o RNA-Seq não dispensa o uso de réplicas biológicas para análises de expressão diferencial. Além disso, ainda existem desafios a serem vencidos, pois são geradas grandes quantidades de sequências (*reads*), que normalmente são curtas e contêm erros. Isto faz com que novos métodos computacionais tenham que ser implementados para conseguir aproveitar ao máximo os dados gerados por RNA-Seq (Chen, et al., 2011).

1.1 Mapeamento dos *reads*

Uma das etapas mais importantes na análise de RNA-seq é o correto mapeamento dos *reads* na sequência de referência de forma a identificar o local correto de onde o *read* foi originado (Horner, et al., 2010). Esta etapa encontra diversos desafios, dentre eles temos o fato de normalmente serem gerados milhões, ou mesmo bilhões, de *reads* e os algoritmos têm que ser eficientes para conseguir realizar esta tarefa rapidamente. O segundo é que os *reads* podem conter erros de sequenciamento. Logo os algoritmos devem levar isso em consideração e permitir alinhamentos contendo *mismatches* e *gaps*. Além disto, podem existir *reads* oriundos de genes parálogos ou de

diferentes isoformas de um gene, que irão mapear em mais de um lugar na referência, tornando difícil encontrar sua verdadeira origem (Trapnell and Salzberg, 2009).

Por se tratar de um processo crucial, diversos algoritmos e programas foram desenvolvidos e estão em constante processo de aperfeiçoamento a fim de se adaptar às constantes mudanças no perfil das sequências produzidas. Quanto a dados de RNA-Seq, podemos dividir os programas de mapeamento de *reads* em dois grupos: (i) os que permitem alinhamento com *splicing,* como Tophat (Trapnell, et al., 2009), STAR (Dobin, et al., 2013) e GSNAP (Wu and Nacu, 2010) e os que não permitem, como Bowtie (Langmead, et al., 2009), BWA (Li and Durbin, 2009) e Soap2 (Li, et al., 2009). Os programas do grupo i são utilizados quando a referência é a sequência genômica, permitindo assim que existam *gaps* do tamanho de íntrons para o correto alinhamento em junções éxon-éxon (Figura 1). Os programas do grupo ii são utilizados quando um transcriptoma de referência, previamente identificado, está disponível.

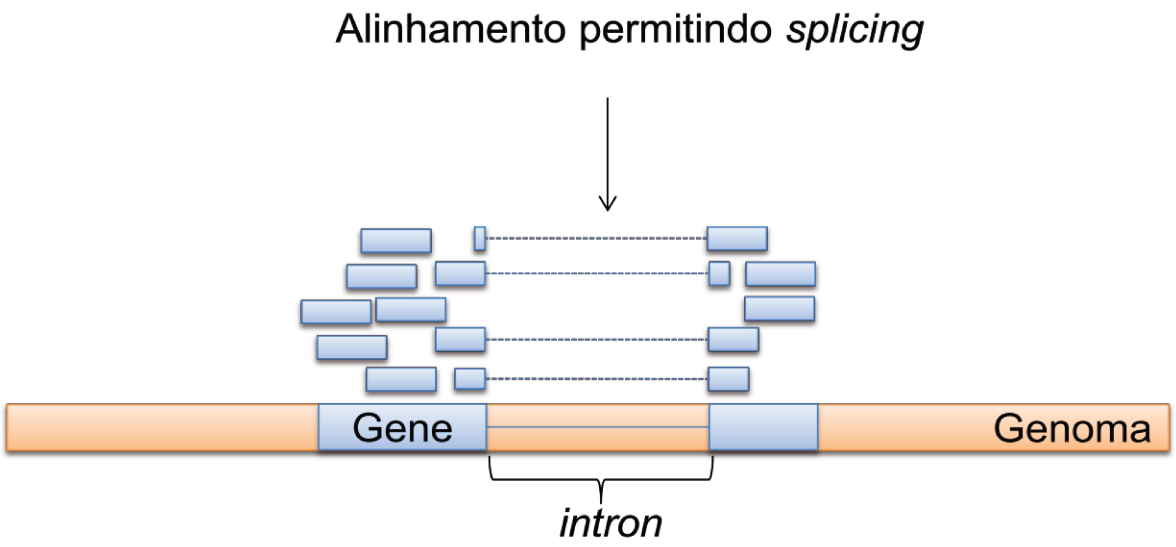

Figura 1 - Alinhamento permitindo *splicing* em regiões de *íntrons*.

1.2 Análise de expressão diferencial

Uma das principais aplicações da técnica RNA-seq é estudar a diferença de expressão dos genes entre amostras de organismos em diferentes condições de estresse, fase de desenvolvimento ou em diferentes tecidos (Camarena, et al., 2010). O

número de *reads* que mapeiam em cada gene, de cada amostra, é esperado ser diretamente proporcional aos seus níveis de expressão. Porém, antes de comparar estes valores de expressão, é necessário normalizá-los, a fim de corrigir alguns vieses, como a variação na quantidade de *reads* produzidos em diferentes sequenciamentos e a variação no tamanho dos transcritos. O tamanho dos transcritos deve ser considerado na normalização, pois transcritos de diferentes tamanhos, mesmo que tenham o mesmo nível de expressão (mesma densidade de *reads* mapeados), terão um número absoluto diferente de *reads* mapeados Um processo de normalização bastante utilizado é o RPKM (*reads per kilobase of exon model per million mapped reads*) (Mortazavi, et al., 2008). A fórmula do RPKM é a seguinte:

$$RPKM = \frac{10^9 * C}{N * L}$$

Onde:

C = número de genes mapeando em um transcrito.

N = número total de *reads* mapeando no experimento.

L = tamanho do transcrito em pares de base (pb);

A análise de expressão diferencial consiste basicamente em usar estatística para decidir se a diferença no valor de expressão de um gene é significativa ou não. A princípio a distribuição de Poisson foi utilizada para modelar os dados (Wang, et al., 2010). A distribuição de Poisson possui apenas um parâmetro que é determinado pela média e uma propriedade importante desta distribuição é que a variância é sempre igual à média. Contudo, a variância observada experimentalmente entre réplicas biológicas frequentemente é maior do que a média, logo esta variância é subestimada. Portanto, testes de hipótese para decidir se dois genes são diferencialmente expressos com base na distribuição de Poisson geram muitos falsos positivos nos seus resultados (Robinson and Smyth, 2007).

Para lidar com esta grande variância, foi proposto modelar os dados utilizando a distribuição binomial negativa, que possuí dois parâmetros, a média e a variância. Atualmente, esta abordagem é utilizada pelos principais programas que fazem testes de

hipótese para expressão diferencial como DESeq (Anders and Huber), edgeR (Robinson, et al., 2010) e baySeq (Hardcastle and Kelly, 2010), sendo a principal diferença entre eles a forma como eles modelam a relação entre a variância e média.

1.3 Identificação de novos transcritos

Diversos estudos analisam a expressão gênica mapeando os *reads* de RNA-seq nos genes previamente anotados (Mortazavi, et al., 2008; Wang, et al., 2008). Contudo, esta abordagem exclui organismos que ainda não tiveram seu genoma sequenciado. Mesmo para organismos com genoma sequenciado, esta abordagem exige que tenha sido realizada uma predição de genes de alta qualidade. Para aproveitar ao máximo os dados de RNA-seq para análise de expressão, é necessário ser capaz de reconstruir os transcritos e medir sua expressão sem utilizar a anotação do genoma como referência (Haas and Zody, 2010).

Existem três estratégias para reconstruir transcritos utilizando RNA-seq. A primeira delas é a montagem *de novo* do transcriptoma (Figura 2), que é altamente útil quando não há um genoma de referência ou o genoma está incompleto. Este método também pode ser utilizado para encontrar variantes de *splicing* e RNAs que não codificam proteínas. Entretanto, a montagem de transcriptomas utilizando RNA-seq é difícil e sua abordagem é diferente da montagem de genomas. Enquanto na montagem de genomas é esperada uma cobertura uniforme, na montagem de transcriptomas a cobertura é variável devido as diferentes expressões dos genes e as suas diferentes isoformas. Existem diversos trabalhos que utilizam diferentes técnicas e montadores para realizar esta tarefa (Birol, et al., 2009; Crawford, et al., 2010; Garg, et al., 2011; Grabherr, et al., 2011; Mizrachi, et al., 2010; Schulz, et al., 2012).

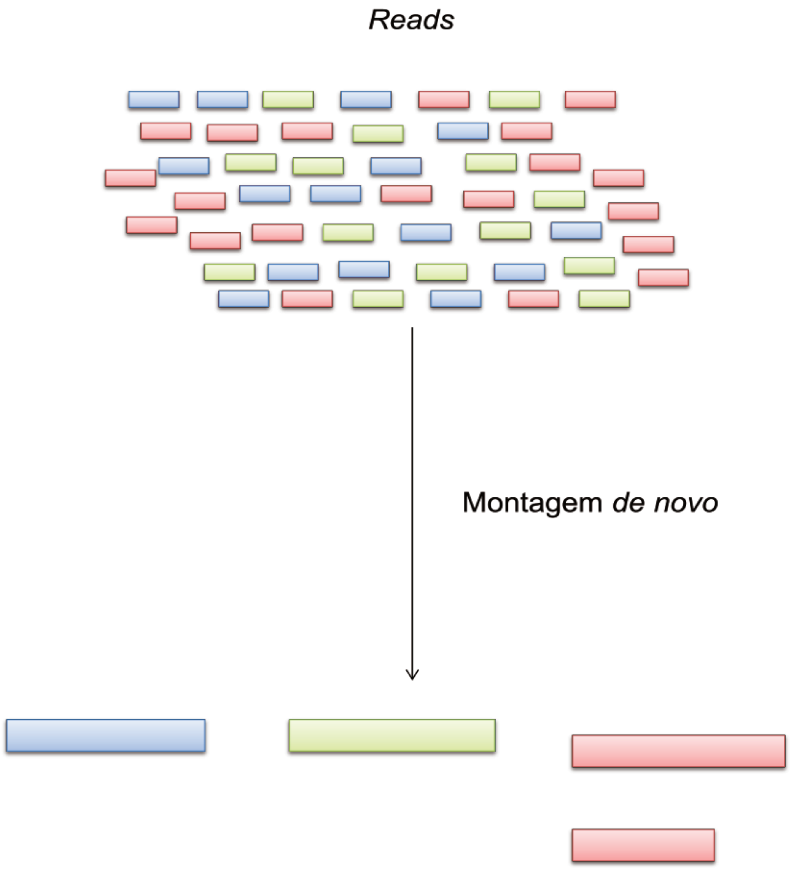

Figura 2 – Está estratégia monta os transcritos *de novo* (diretamente dos *reads* de RNA-Seq).

A segunda estratégia é utilizada quando se conhece o genoma de referência do organismo (Figura 3). É o caso do programa Cufflinks (Trapnell, et al., 2012) e Scripture (Guttman, et al., 2010), que utilizam o mapeamento dos *reads* no genoma para reconstruir os transcritos expressos. Esta segunda estratégia é mais sensível para montar transcritos pouco expressos, porém é altamente dependente do correto alinhamento dos *reads* e da qualidade da montagem do genoma.

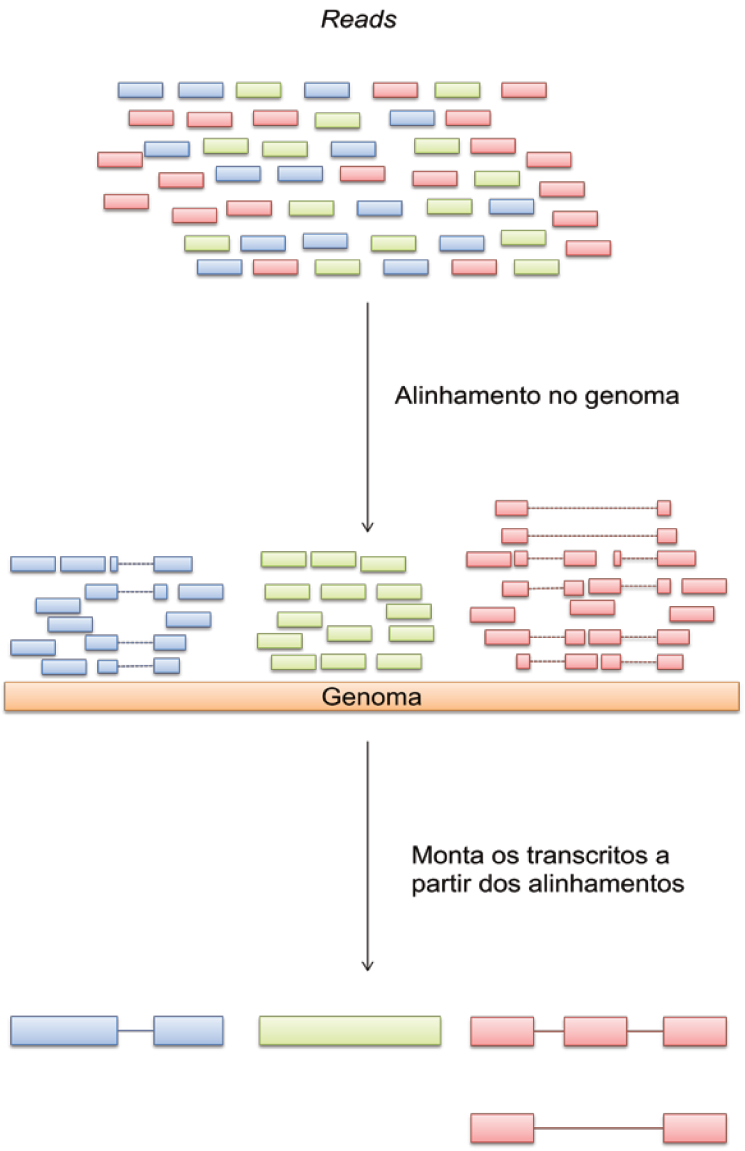

Figura 3 – Esta estratégia primeiramente alinha os *reads* de RNA-Seq no genoma permitindo *splicing* e a partir dos alinhamentos os transcritos são montados.

A terceira é uma estratégia híbrida descrita recentemente em (http://trinityrnaseq.sourceforge.net/genome_guided_trinity.html), que utiliza uma combinação de alinhamento de *reads* de RNA-Seq no genoma, com montagem *de novo* de *reads* de RNA-Seq e montagem baseada no alinhamento de transcritos no genoma. (Figura 4). Está abordagem envolve quatro etapas: (i) primeiro os *reads* são alinhados no genoma utilizando programas que permitem alinhamento com *splicing*; (ii) os *reads* alinhados são divididos em partições de acordo com a cobertura dos alinhamentos e cada partição é montada utilizando a estratégia *de novo*; (iii) os transcritos montados

no passo ii são alinhados no genoma; (iv) é feita a montagem da estrutura correta dos genes a partir dos alinhamentos dos transcritos.

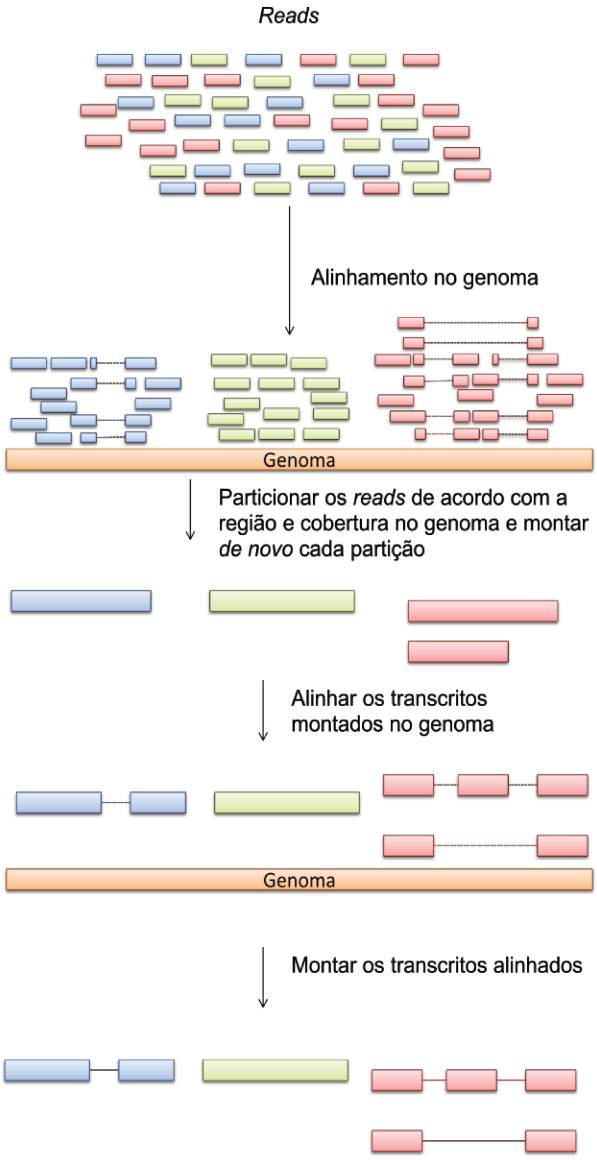

Figura 4 – Estratégia híbrida para montagem de transcritos, onde primeiramente os *reads* são alinhados no genoma permitindo *splicing*. Grupos de *reads* são então particionados de acordo com a cobertura no genoma e cada partição é montada separadamente de forma *de novo*. Os transcritos montados são alinhados no genoma e esses alinhamentos são montados a fim de encontrar a estrutura correta dos genes, assim como variantes de splicing.

1.4 RNAs não codificantes longos

As sequências de RNA-Seq também podem ser utilizadas para identificação de RNAs que não codificam para proteínas (ncRNA). Os ncRNAs compreendem um diverso grupo de transcritos que incluem os ncRNA de manutenção celular, os ncRNAs regulatórios e diversos outros tipos de ncRNAs que ainda não foram caracterizados (Pauli, et al., 2011). Dentre os ncRNAs de manutenção celular podemos citar os RNAs ribossomais (rRNA), RNAs transportadores (tRNA), pequenos RNAs nucleares (snRNA) e pequenos RNAs nucleolares (snoRNA). Os ncRNAs regulatórios são classificados quanto ao seu tamanho, sendo os menores que 200pb chamados pequenos RNAs não codificantes (sncRNA) e os com ao menos 200pb chamados longos RNAs não codificantes (lncRNA). Os lncRNA podem ser classificados em longos RNAs não codificantes intergênicos (lincRNA), que não possuem sobreposição com exons de genes que codificam proteínas e RNAs não codificantes antissenso (ancRNA), que são transcritos a partir da fita oposta de um gene que codifica proteína e tem potencial de formar uma dupla fita com o mRNA maduro (Nam and Bartel, 2012).

Os lncRNAs apresentam pouca conservação em sequência e baixos níveis de expressão quando comparados com os mRNAs. Isso fez com que estes transcritos fossem inicialmente considerados apenas um ruído transcricional (Struhl, 2007). Entretanto, a função dos lncRNAs não necessariamente depende da sua conservação em sequência, sendo possível que a conservação da sua estrutura secundária tenha preferência (Pang, et al., 2006).

A função da maioria dos lncRNAs é desconhecida, porém um tema que tem se destacado é o seu papel na regulação de genes (Pauli, et al., 2012). Esta regulação pode se dar em três níveis: epigenética, regulação transcricional e regulação pós-transcricional (Figura 5).

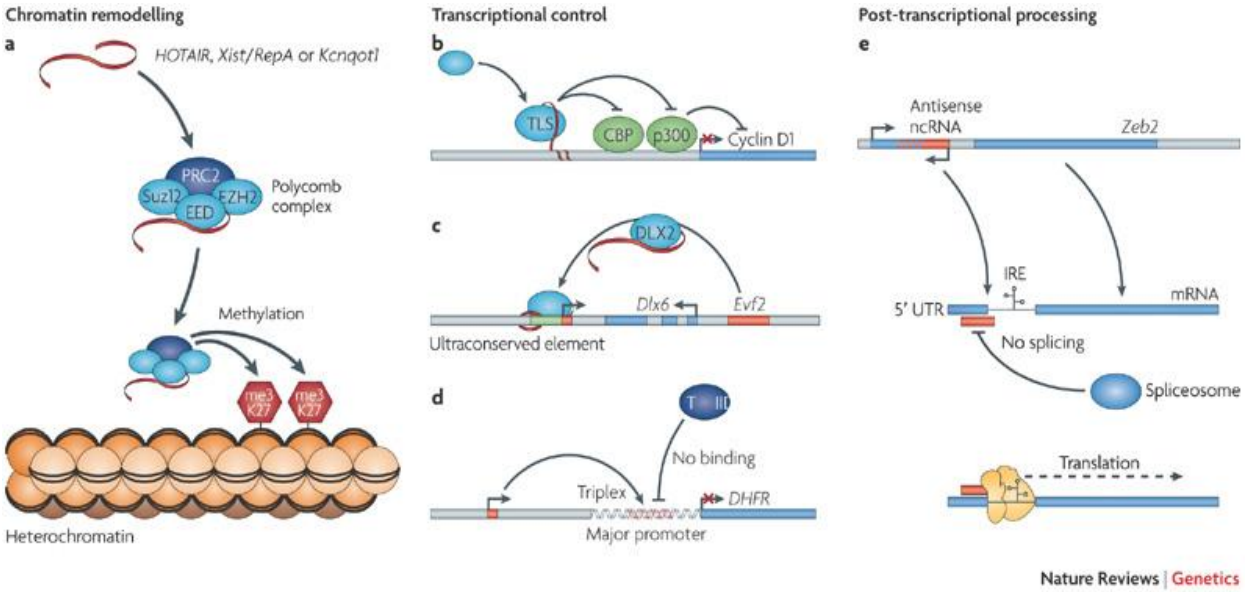

Figura 5 - Mecanismos pelos quais os lncRNAs regulam a expressão dos mRNAs a nível epigenético, transcricional e pós-transcricional. a) lncRNAs podem recrutar complexos modificadores de cromatina para um lócus genômico para transmitir a sua atividade catalítica. Neste caso, os lncRNAs HOTAIR, (Xist e RepA) e *Kcnqot1* recrutam o complexo Polycomb para o lócus HoxD, cromossomo X e o domínio KCNQ1, respectivamente, onde eles induzem a formação de heterocromatina e reprimem a expressão dos genes. b) Os lncRNAs podem regular o processo de transcrição através de uma série de mecanismos. Um lncRNA ligado ao promotor do gene cyclin D1 recruta a proteína de ligação a RNA TLS para inibir a atividade das proteínas CBP e p300 e assim reprimir a transcrição do gene cyclin D1. c) o lncRNA, Evf2, age como um co-ativador para fator de transcrição DLX2, para regular a transcrição do gene Dlx6. d) Um lncRNA transcrito a partir do promotor secundário do gene DHFR em humanos pode formar um triplex com o promotor principal para impedir a ligação do fator de transcrição TFIID e assim silenciar a expressão do gene DHFR. e) Um ncRNA antissenso pode se ligar ao sítio de splicing de um íntron situado na região não traduzida 5' do gene Zeb2 e fazer com que esse íntron seja retido. Este íntron contém um *internal ribosome entry site* (IRES) necessário para a tradução do gene Zeb2. Fonte: (Mercer, et al., 2009)

1.5 Doença vassoura-de-bruxa

A doença vassoura-de-bruxa do cacaueiro (*Theobroma cacao*) é causada pelo fungo basidiomiceto *Moniliophthora perniciosa* (Griffith, et al., 2003). Este fungo possui um ciclo de vida hemibiotrófico (Figura 6), com duas fases bem distintas: a primeira, biotrófica, quando infecta tecidos vivos e a segunda necrotrófica, quando coloniza tecidos mortos (Evans, 1980). Na fase biotrófica, o fungo apresenta micélios mononucleados, sem grampos de conexão. Neste momento ele é encontrado em baixas densidades entre as células do hospedeiro e se alimenta de tecidos vivos da planta. Esta fase é conhecida como vassoura-verde e os sintomas do cacaueiro são hipertrofia, hiperplasia e perda de dominância apical (Scarpari, et al., 2005).

Após cerca de 5 a 9 semanas do início da infecção, devido a mecanismos e sinais desconhecidos, o fungo passa para a fase necrotrófica. Neste estágio o patógeno apresenta micélios binucleados com grampos de conexão e cresce intensamente. Nesta fase o fungo causa necrose nos tecidos da planta, levando ao apodrecimento e morte destes tecidos (vassoura-seca) (Scarpari, et al., 2005).

A partir do micélio necrotrófico, com condições favoráveis de umidade, o fungo inicia a produção de basidiomas. Estas estruturas produzem basidiósporos que são liberados e levados por vento e água, podendo infectar outras plantas reiniciando o ciclo de vida do fungo (Purdy and Schmidt, 1996).

O fungo *M. perniciosa* pode infectar não somente ramos, como também frutos jovens e flores. A infecção de flores gera frutos pequenos e partenocárpicos, enquanto que a infecção de frutos causa necrose e apodrecimento (Evans, 1980).

A doença vassoura-de-bruxa atinge a América Latina e as ilhas do Caribe. No Brasil as primeiras áreas afetadas foram plantações de cacau nos estados de Rondônia e Acre na década de 1970 (Rudgard, 1986).

No sul da Bahia, maior região produtora de cacau do Brasil, a doença foi vista pela primeira vez em 1989. Pela alta densidade de cacaueiros nessa região e por não haver períodos de seca a doença vassoura-de-bruxa foi mais devastadora que em qualquer outro lugar (Pereira, et al., 1990).

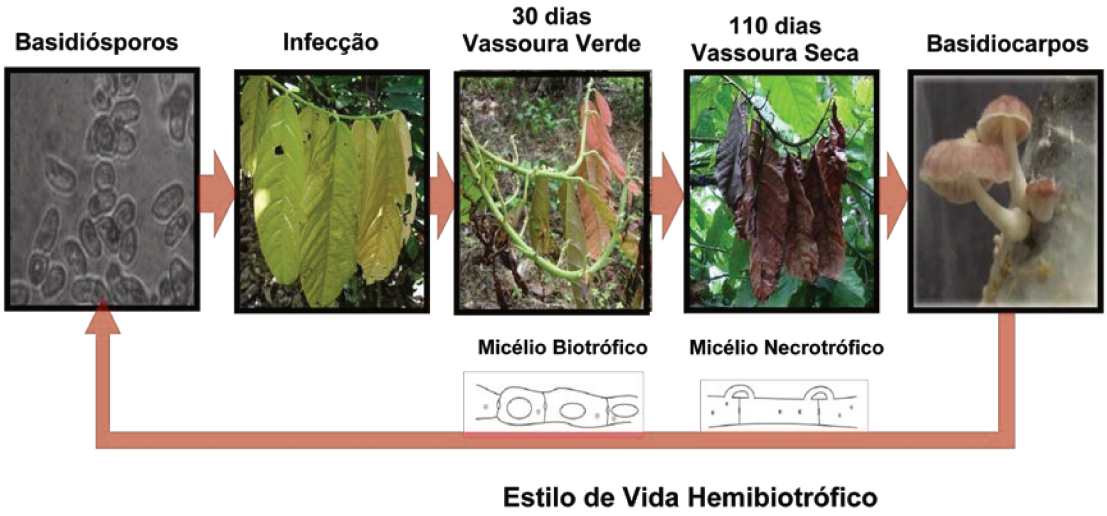

Figura 6 – Ciclo de vida do fungo *M. perniciosa*. Os basidiósporos infectam os ramos do cacaueiro que após aproximadamente 30 dias apresentam os sintomas como hipertrofia, hiperplasia e perda de dominância apical, caracterizando o estágio da doença conhecido como vassoura-verde. Após aproximadamente 110 dias o ramo do cacaueiro já está morto, caracterizando o estágio conhecido como vassoura-seca. A partir da vassoura-seca são formados os basidiocarpos que lançam basidiósporos que irão infectar outras plantas reiniciando o ciclo da doença. Figura cedida pelo Dr. Paulo José P. L. Teixeira.

1.6 O Projeto Genoma da Vassoura-de-Bruxa

No ano 2000 foi iniciado, no Laboratório de Genômica e Expressão (LGE), o Projeto Genoma vassoura-de-bruxa (http://www.lge.ibi.unicamp.br/vassoura) sob a coordenação do Prof. Dr. Gonçalo Amarante Guimarães Pereira. Este projeto envolveu diversos centros de pesquisa federais e estaduais como o LGE, a Universidade Estadual Santa Cruz (UESC), a Universidade Federal da Bahia (UFBA), a Comissão Executiva do Plano da Lavoura Cacaueira (CEPLAC), o CENARGEN-EMBRAPA e a Universidade Estadual de Feira de Santana (UEFS). O projeto tem como objetivo esclarecer aspectos da biologia do fungo e da sua interação com o cacaueiro, sendo que uma das principais contribuições desse projeto foi o sequenciamento e montagem do genoma do fungo *M. perniciosa*.

Recentemente, na tentativa de aumentar o conhecimento sobre os mecanismos moleculares durante a evolução da doença, iniciou-se o projeto Atlas Transcriptômico da Vassoura-de-Bruxa. Através de uma parceria entre o LGE e o Centro de Ciências Genômicas da Universidade da Carolina do Norte (CCGS), foram sequenciadas 54

bibliotecas de RNA-seq pelo Dr. Paulo José P. L. Teixeira como parte de sua tese de doutorado. Os sequenciamentos foram feitos utilizando sequenciadores Illumina Genome Analyzer IIx e Hiseq2000. As bibliotecas correspondem à interação planta-fungo em diversas fases da doença e em diferentes tecidos. Também foram construídas bibliotecas explorando o fungo em diferentes meios de cultura.

2 OBJETIVO

Com a grande quantidade de dados de RNA-Seq sequenciados pelo projeto Atlas Transcriptômico da Vassoura-de-Bruxa, este trabalho tem como objetivo gerar informações sobre o perfil de transcrição do *M. perniciosa* e do *T. cacao* durante a doença vassoura-de-bruxa, a fim de aumentar o entendimento sobre os mecanismos moleculares dessa interação.

3 MATERIAIS E MÉTODOS

3.1 Descrição dos dados utilizados

Para o desenvolvimento do objetivo deste trabalho, foram utilizadas 54 bibliotecas de RNA-Seq do projeto Atlas Transcriptômico da Vassoura-de-Bruxa. Estas bibliotecas correspondem a experimentos tanto do *M. perniciosa* em diversas condições de desenvolvimento e tratamento *in vitro*, como também da sua interação com os ramos e frutos do cacaueiro *in planta*. Com as bibliotecas de interação planta-patógeno é possível medir o nível de expressão de transcritos do fungo e da planta simultaneamente. A seguir são descritos os experimentos e suas amostras.

Interação em vassouras - Este experimento é um seguimento ao longo do tempo da progressão da doença vassoura-de-bruxa em meristemas, que foi realizado em casa de vegetação. Cada um dos quatro pontos da doença tem seu respectivo controle sadio, que são plantas da mesma idade que a infectada (Tabela 1). Como a fase biotrófica do fungo *M. perniciosa*, conhecida como vassoura-verde, é atípica em relação a outros fungos hemibiotróficos, foram sequenciadas cinco bibliotecas desta fase com o objetivo aprofundar as análises desse estágio. Estas cinco bibliotecas foram colhidas em dois experimentos diferentes, sendo duas delas de um primeiro experimento e as outras três de um segundo experimento.

Tabela 1 – Bibliotecas da interação em vassouras.

Bibliotecas	**Tamanho dos *reads***	**Tipo**
Vassoura-verde (5 réplicas)	50 pb	*paired-end*
Necrose inicial	36 pb	*single-end*
Necrose avançada	36 pb	*single-end*
Vassoura-seca	36 pb	*single-end*
Controle sadio (vassoura-verde – 5 réplicas)	50 pb	*paired-end*
Controle sadio (necrose inicial)	36 pb	*single-end*
Controle sadio (necrose avançada)	36 pb	*single-end*
Controle sadio (vassoura-seca)	36 pb	*single-end*

Interação em frutos - Este experimento foi realizado com sementes e cascas de frutos coletados no campo. Da semente foram sequenciadas quatro amostras, sendo uma de fruto sadio e três de frutos infectados em diferentes estágios da doença. Da casca foram sequenciadas cinco amostras, uma de fruto sadio e quatro de frutos infectados em diferentes estágios da doença (Tabela 2-3).

SEMENTES

Tabela 2 – Bibliotecas da interação em frutos (semente).

Bibliotecas	**Tamanho dos *reads***	**Tipo**
Fruto sadio	36 pb	*single-end*
Fruto no início da doença	36 pb	*single-end*
Fruto necrosando	36 pb	*single-end*
Fruto podre	36 pb	*single-end*

CASCAS

Tabela 3 – Bibliotecas da interação em frutos (casca).

Bibliotecas	**Tamanho dos *reads***	**Tipo**
Fruto sadio	36 pb	*single-end*
Fruto no início da doença	36 pb	*single-end*
Fruto necrosando (casca viva)	36 pb	*single-end*
Fruto necrosando (casca morta)	36 pb	*single-end*
Fruto podre	36 pb	*single-end*

Germinação de esporos *in vitro* - Este experimento tem por objetivo a identificação de mecanismos relacionados à patogenicidade e ao desenvolvimento do *M. perniciosa* durante o processo de germinação dos esporos (Tabela 4).

Tabela 4 – Bibliotecas de esporo germinando *in vitro*.

Bibliotecas	**Tamanho dos *reads***	**Tipo**
Esporos germinando	36 pb	*single-end*
Esporos não germinados	36 pb	*single-end*

Senescência em *M. perniciosa in vitro* – Foram sequenciadas amostras de micélios necrotróficos jovens e senescentes do *M. perniciosa* (Tabela 5).

Tabela 5 – Bibliotecas do fungo *M. perniciosa* em senescência *in vitro.*

Bibliotecas	**Tamanho dos *reads***	**Tipo**
Fungo velho	36 pb	*single-end*
Fungo novo	36 pb	*single-end*

Desenvolvimento do *M. perniciosa in vitro* – Neste experimento, foram sequenciadas amostras de micélios em três tempos de interesse: após sete dias de crescimento, representando o micélio monocariótico (biotrófico); após quatorze dias de crescimento, representando a transição de fase; e após vinte e oito dias de crescimento, representando micélio dicariótico (Tabela 6).

Tabela 6 – Bibliotecas do *M. perniciosa* se desenvolvendo *in vitro.*

Bibliotecas	**Tamanho dos *reads***	**Tipo**
Micélio biotrófico (7 dias)	36 pb	*single-end*
Micélio necrotrófico (7 dias)	36 pb	*single-end*
Micélio biotrófico (14 dias)	36 pb	*single-end*
Micélio necrotrófico (14 dias)	36 pb	*single-end*
Micélio biotrófico (28 dias)	36 pb	*single-end*
Micélio necrotrófico (28 dias)	36 pb	*single-end*

Efeito de inibidores em *M. perniciosa in vitro* - O objetivo deste experimento é identificar alterações no metabolismo do *M. perniciosa* após tratamento com inibidores específicos da cadeia respiratória (Tabela 7).

Tabela 7 – Bibliotecas do *M. perniciosa* sobre efeito de inibidores *in vitro.*

Bibliotecas	**Tamanho dos *reads***	**Tipo**
Tratamento com Sham	36 pb	*single-end*
Tratamento com Azoxistrobina	36 pb	*single-end*
Tratamento com Sham + Azoxistrobina	36 pb	*single-end*
Tratamento com Etanol	36 pb	*single-end*
Tratamento com Metanol	36 pb	*single-end*
Tratamento com Etanol + Metanol	36 pb	*single-end*

Formação das estruturas reprodutivas do *M. perniciosa* - O objetivo deste experimento é identificar genes possivelmente relacionados à produção dos basidiomas em *M. perniciosa* (Tabela 8).

Tabela 8 – Bibliotecas da formação de estruturas reprodutivas do *M. perniciosa in vitro*.

Bibliotecas	**Tamanho dos *reads***	**Tipo**
Basidioma	36 pb	*single-end*
Primórdios de basidioma	36 pb	*single-end*

RNA-seq direcional - Estas bibliotecas foram preparadas de forma que os *reads* correspondam à fita de DNA transcrita. São úteis para se caracterizar ORFS (*Open Read Frame*) desconhecidas, identificar sobreposição de genes e auxiliar na predição gênica. A busca por RNA antissenso com função regulatória também é uma aplicação importante (Tabela 9).

Tabela 9 – Bibliotecas de RNA-Seq direcional do *M. perniciosa in vitro*.

Bibliotecas	**Tamanho dos *reads***	**Tipo**
Micélio biotrófico	36 pb	*single-end*
Micélio necrotrófico	36 pb	*single-end*
Basidioma	36 pb	*single-end*
Vassoura-verde	36 pb	*single-end*
Controle sadio (vassoura-verde)	36 pb	*single-end*

Indução de *M. perniciosa* com diferentes fontes de carbono *in vitro* - O experimento foi realizado com o micélio necrotrófico do *M. perniciosa* crescido por 7 dias. Após este período, o micélio foi filtrado, lavado e transferido para um novo "meio biotrófico", por um período de 24 horas, variando-se as fontes de carbono (Tabela 10).

Tabela 10 – Bibliotecas do *M. perniciosa* crescendo em diferentes fontes de carbono *in vitro*.

Bibliotecas	**Tamanho dos *reads***	**Tipo**
Glicerol 1%	36 pb	*single-end*
Glicose 1%	36 pb	*single-end*
Metanol 1%	36 pb	*single-end*
Ácido poligalacturônico 1%	36 pb	*single-end*
Pectina 1%	36 pb	*single-end*
Extrato de cacau 1%	50 pb	*paired-end*

Neste trabalho também foram utilizadas as montagens do genoma do *M. perniciosa* e do *T. cacao*. O genoma do *M. perniciosa* foi sequenciado e montado pela equipe do projeto genoma da vassoura-de-bruxa (Mondego, et al., 2008). A montagem mais recente tem cerca 44 Mpb divididos em 1600 *scaffolds* e 17008 genes preditos. A montagem do genoma do *Theobroma cacao* (Motamayor, et al., 2013) tem cerca 350 Mbp divididos em 1782 *scaffolds* e contém 34997 genes preditos.

3.2 Análise do perfil de transcrição

Primeiramente, os *reads* de todas as bibliotecas sequenciadas foram mapeados contra uma referência única, contendo os genes do *M. perniciosa*, os genes do *T. cacao* e as sequências ribossomais dos dois organismos, utilizando o software Bowtie (Langmead, et al., 2009). Nas bibliotecas que possuem *reads* de 50pb, foram permitidos até 2 *mismatchs*, enquanto que nas bibliotecas que possuem *reads* de 36pb, foi permitido apenas 1 *mismatch*. Nas bibliotecas que são *paired-end* foram exigidos que ambos os *reads* do par fossem mapeados para que o alinhamento fosse válido. Os *reads* que mapearam em mais de um lugar na referência foram excluídos. Com o resultado dos alinhamentos montamos uma tabela com o número de *reads* mapeando

em cada gene, que finalmente foi normalizado utilizando o método do RPKM (*reads per kilobase of exon model per million mapped reads*), para que os valores de expressão dos genes possam ser comparáveis, tanto entre genes, quanto entre bibliotecas (Mortazavi, et al., 2008).

3.3 Análise na vassoura-verde

Realizamos a identificação dos genes diferencialmente expressos do cacaueiro no início da infecção dos ramos (vassoura-verde), em relação às plantas controles. Para esta análise utilizamos o número de *reads* mapeados em cada gene de cacaueiro e testamos três dos *softwares* comumente utilizados para a análise de expressão diferencial, são eles DESeq (Anders and Huber, 2010), EdgeR (Robinson, et al., 2010) e Bayseq (Hardcastle and Kelly, 2010). Para os três programas, consideramos como diferencialmente expressos os genes com um *False Discovery Rate* (FDR) menor que 1%. Os genes classificados como diferencialmente expressos foram submetidos à análise de enriquecimento de termos do Gene Ontology (GO), utilizando o programa Blast2GO (Conesa, et al., 2005).

Para análise do transcriptoma do *M. perniciosa* nesta fase (vassoura-verde) fizemos um agrupamento (*clustering*) hierárquico de todos os genes, utilizando o logaritmo na base 2 dos RPKMs dos genes em todas as 54 bibliotecas deste trabalho. Um subgrupo deste *clustering*, que correspondia a genes preferencialmente expressos na fase biotrófica da doença foi selecionado e submetido à análise de enriquecimento de termos GO utilizando o programa Blast2GO.

3.4 Identificação de novos transcritos

Os transcritos foram montados seguindo o *pipeline* de montagem guiada pelo genoma do *software* Trinity (disponível em http://trinityrnaseq.sourceforge.net/genome_guided_trinity.html). Este processo consiste em quatro passos: (i) alinhar os *reads* no genoma utilizando o programa GSNAP (Wu and Nacu, 2010) e então particionar estes *reads* de acordo com a região e cobertura no

genoma; (ii) o montador Trinity (Grabherr, et al., 2011) é utilizado para montar cada partição de *reads* separadamente; (iii) os transcritos montados pelo Trinity são realinhados no genoma utilizando o alinhador GMAP (Wu and Watanabe, 2005); (iv) Os transcritos alinhados são então montados em estruturas completas utilizando o montador PASA (Haas, et al., 2003).

4 RESULTADOS E DISCUSSÃO

Os resultados encontrados foram divididos em três partes, sendo a primeira delas uma análise de expressão global de todas as condições e tecidos sequenciados. Na segunda apresentamos a análise de expressão diferencial do início da infecção, conhecida como vassoura-verde. E na quarta parte fornecemos uma análise inicial da identificação de lncRNAs no genoma do *M. perniciosa*.

4.1 Alinhamento dos *reads*

Um total de 54 bibliotecas de RNA-Seq foram analisadas, sendo que 11 delas são exclusivamente de cacaueiro, 27 são exclusivamente do fungo *M. perniciosa* e 16 são da interação entre os dois organismos. Somando todas as bibliotecas, foram gerados 2.244.229.024 *reads*. De forma a analisar o perfil de expressão, o programa Bowtie (Langmead, et al., 2009) foi utilizado para alinhar todos os *reads* contra uma referência única contendo as sequências codantes dos genes do *T. cacao* e *M. perniciosa*, além das sequências de rRNA dos dois organismos. Do total de *reads* utilizados no alinhamento, 76,13% foram alinhados unicamente, 2,2% tiveram alinhamentos múltiplos e 21,66% não puderam ser alinhados (Figura 7).

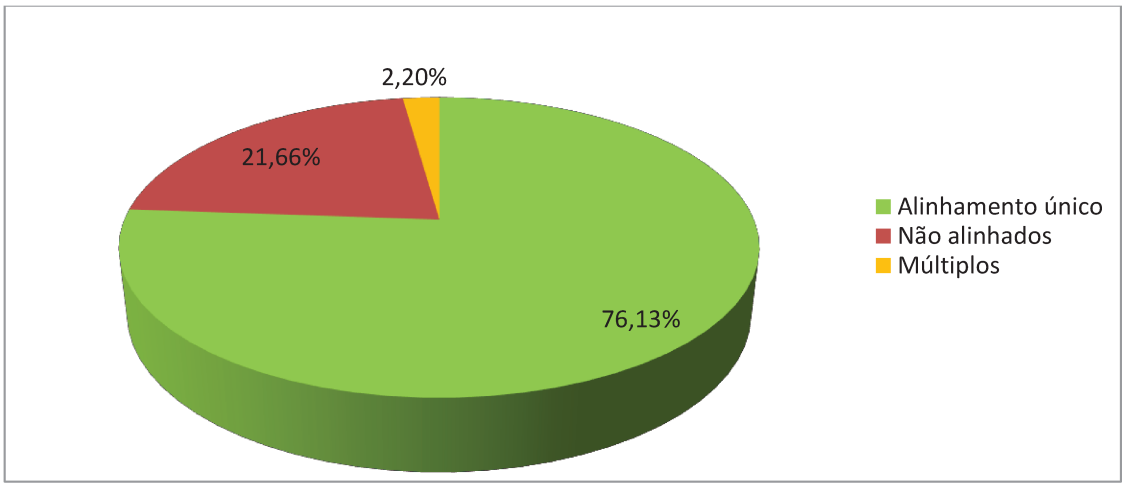

Figura 7 - Porcentagem de *reads* com alinhamento único, alinhamento múltiplo e não alinhados nos genes do *M. perniciosa*, *T. cacao* e nas sequências de RNA ribossomal dos dois organismos.

Das sequências alinhadas unicamente, 33,67% correspondem aos genes do *M. perniciosa*, 61,66% correspondem aos genes do *T. cacao* e 4,68% correspondem as sequências de rRNA dos dois organismos (Figura 8).

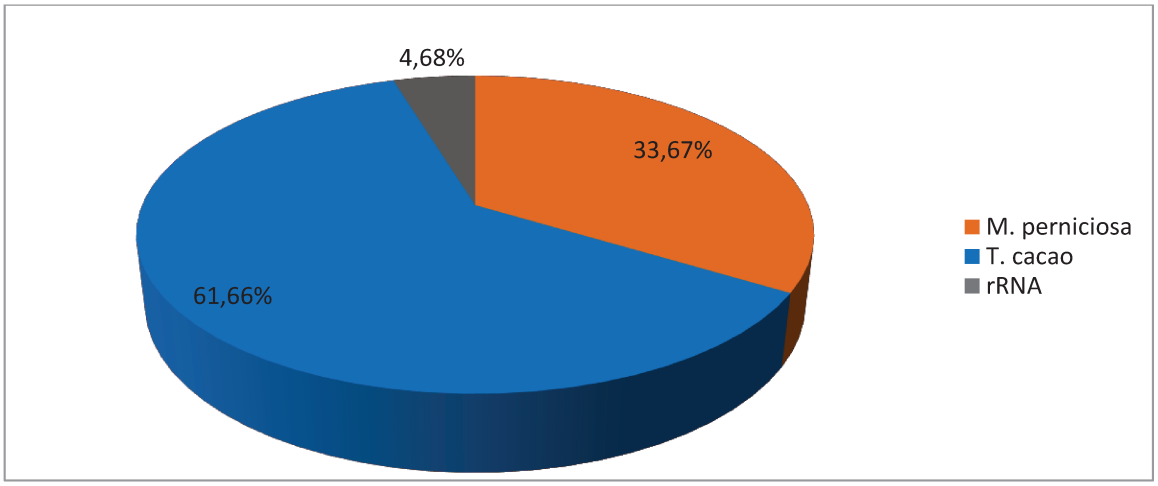

Figura 8 - Porcentagem de *reads* mapeando com *M. perniciosa*, *T. cacao* e sequência de RNA ribossomal dos dois organismos.

Os resultados dos alinhamentos dos *reads* de cada biblioteca são mostrados na Tabela 11.

Tabela 11 – Mapeamento dos *reads* ao longo de todas as bibliotecas.

Experimentos	Bibliotecas	Código	Número de *reads*	Alinhados com *M. perniciosa*	Alinhados com *T. cacao*	Alinhados com rRNA	Não alinhados	Múltiplos
Interação em vassouras	Vassoura-verde	VV0E	207.441.756	904.741 (0,44%)	166.890.211 (80,45%)	3.474.637 (1,67%)	33.733.070 (16,26%)	2.439.097 (1,18%)
	Vassoura-verde	VV3F	90.591.723	186.840 (0,21%)	71.381.534 (78,79%)	2.249.540 (2,48%)	15.695.360 (17,33%)	1.078.449 (1,19%)
	Vassoura-verde	VV4C	103.196.353	158.024 (0,15%)	82.082.732 (79,54%)	994.987 (0,96%)	18.773.763 (18,19%)	1.186.847 (1,15%)
	Vassoura-verde	VV1D	83.953.292	196.093 (0,23%)	67.776.159 (80,73%)	1.999.993 (2,38%)	13.032.324 (15,52%)	948.723 (1,13%)
	Vassoura-verde	NZ3	77.308.238	265.253 (0,34%)	56.746.425 (73,40%)	3.151.209 (4,08%)	16.313.536 (21,10%)	831.815 (1,08%)
	Necrose inicial	NA4	30.140.571	255.161 (0,85%)	24.593.595 (81,60%)	409.873 (1,36%)	4.160.647 (13,80%)	721.295 (2,39%)
	Necrose avançada	NB1	26.499.917	549.744 (2,07%)	9.424.626 (35,56%)	286.192 (1,08%)	15.804.200 (59,64%)	435.155 (1,64%)
	Vassoura-seca	DB1	23.646.179	445.095 (1,88%)	98.109 (0,41%)	104.881 (0,44%)	22.860.672 (96,68%)	137.422 (0,58%)
	Controle sadio (vassoura-verde)	CS2E	82.560.659	634 (0,00%)	68.519.668 (82,99%)	868.416 (1,05%)	12.000.100 (14,53%)	1.171.841 (1,42%)
	Controle sadio (vassoura-verde)	CS4B	80.950.589	96 (0,00%)	65.487.193 (80,90%)	670.051 (0,83%)	14.036.009 (17,34%)	757.240 (0,94%)
	Controle sadio (vassoura-verde)	CS0D	89.659.731	27 (0,00%)	73.888.882 (82,41%)	881.745 (0,98%)	13.201.829 (14,72%)	1.687.248 (1,88%)

Experimentos	Bibliotecas	Código	Número de *reads*	Alinhados com *M. perniciosa*	Alinhados com *T. cacao*	Alinhados com rRNA	Não alinhados	Múltiplos
	Controle sadio (vassoura-verde)	CS3D	86.454.964	213 (0,00%)	68.901.656 (79,70%)	1.368.631 (1,58%)	15.095.300 (17,46%)	1.089.164 (1,26%)
	Controle sadio (vassoura-verde)	HA2	96.856.054	14 (0,00%)	73.806.211 (76,20%)	2.611.610 (2,70%)	19.019.355 (19,64%)	1.418.864 (1,46%)
	Controle sadio (necrose inicial)	HB3	28.411.022	61 (0,00%)	22.872.797 (80,51%)	678.578 (2,39%)	3.926.765 (13,82%)	932.821 (3,28%)
	Controle sadio (necrose avançada)	HC1	28.726.748	98 (0,00%)	24.138.820 (84,03%)	406.823 (1,42%)	3.566.357 (12,41%)	614.650 (2,14%)
	Controle sadio (vassoura-seca)	HD2	23.400.070	70 (0,00%)	20.196.130 (86,31%)	354.882 (1,52%)	2.347.226 (10,03%)	501.762 (2,14%)
Interação em frutos (semente)	Fruto sadio	HPB2	22.285.402	1.546 (0,01%)	18.327.689 (82,24%)	1.496.825 (6,72%)	1.754.440 (7,87%)	704.902 (3,16%)
	Fruto no início da doença	GPB1	26.588.983	7.671.392 (28,85%)	11.799.478 (44,38%)	1.106.724 (4,16%)	5.294.051 (19,91%)	717.338 (2,70%)
	Fruto necrosando	DPB1	24.623.147	5.326.535 (21,63%)	12.797.877 (51,97%)	1.975.189 (8,02%)	3.700.880 (15,03%)	822.666 (3,34%)
	Fruto podre	BPB1	28.007.214	19.985.345 (71,36%)	83.225 (0,30%)	1.057.553 (3,78%)	6.207.883 (22,17%)	673.208 (2,40%)
Interação em frutos (casca)	Fruto sadio	HS2	27.677.832	5.018.864 (18,13%)	15.437.825 (55,78%)	2.365.548 (8,55%)	3.889.714 (14,05%)	965.881 (3,49%)
	Fruto no início da doença	GS3	28.219.072	11.706.344 (41,48%)	8.869.310 (31,43%)	1.467.819 (5,20%)	5.248.566 (18,60%)	927.033 (3,29%)

Experimentos	Bibliotecas	Código	Número de *reads*	Alinhados com *M. perniciosa*	Alinhados com *T. cacao*	Alinhados com rRNA	Não alinhados	Múltiplos
	Fruto necrosando (casca viva)	DG2	28.435.732	8.459.466 (29,75%)	12.683.297 (44,60%)	2.578.931 (9,07%)	3.559.006 (12,52%)	1.155.032 (4,06%)
	Fruto necrosando (casca morta)	DR2	26.387.255	5.205.928 (19,73%)	32.114 (0,12%)	11.353.621 (43,03%)	4.644.034 (17,60%)	5.151.558 (19,52%)
	Fruto podre	BS5	16.184.686	8.841.600 (54,63%)	2.924 (0,02%)	1.108.495 (6,85%)	5.616.465 (34,70%)	615.202 (3,80%)
Germinação de esporos *in vitro*	Esporos germinando	GE1	27.753.555	17.316.049 (62,39%)	295 (0,00%)	1.744.095 (6,28%)	7.605.645 (27,40%)	1.087.471 (3,92%)
	Esporos não germinados	NGS1	30.251.513	17.479.364 (57,78%)	229 (0,00%)	1.807.730 (5,98%)	9.970.504 (32,96%)	993.686 (3,28%)
Senescência do *M. perniciosa in vitro*	Fungo velho (BP10)	BN1	28.377.071	17.601.529 (62,03%)	224 (0,00%)	3.202.986 (11,29%)	5.655.011 (19,93%)	1.917.321 (6,76%)
	Fungo novo (BP10)	BV3	29.261.911	17.305.126 (59,14%)	245 (0,00%)	3.558.151 (12,16%)	6.265.552 (21,41%)	2.132.837 (7,29%)
Desenvolvimento do *M. perniciosa in vitro*	Micélio biotrófico (7 dias)	BIO7	28.631.853	19.060.329 (66,57%)	680 (0,00%)	943.192 (3,29%)	7.906.973 (27,62%)	720.679 (2,52%)
	Micélio necrotrófico (7 dias)	SAP7	31.235.935	21.466.571 (68,72%)	1.675 (0,01%)	1.046.073 (3,35%)	7.924.698 (25,37%)	796.918 (2,55%)
	Micélio biotrófico (14 dias)	PTA1	31.267.280	21.242.117 (67,94%)	5.259 (0,02%)	908.668 (2,91%)	8.427.854 (26,95%)	683.382 (2,19%)
	Micélio necrotrófico (14	NEC14	25.707.455	16.421.420 (63,88%)	949 (0,00%)	1.434.769 (5,58%)	6.911.231 (26,88%)	939.086 (3,65%)

Experimentos	Bibliotecas	Código	Número de *reads*	Alinhados com *M. perniciosa*	Alinhados com *T. cacao*	Alinhados com rRNA	Não alinhados	Múltiplos
	dias)							
	Micélio biotrófico (28 dias)	SB1	24.030.203	14.509.447 (60,38%)	1.701 (0,01%)	1.114.827 (4,64%)	7.719.447 (32,12%)	684.781 (2,85%)
	Micélio necrotrófico (28 dias)	NS28	22.731.804	13.845.796 (60,91%)	1.852 (0,01%)	1.372.439 (6,04%)	6.699.060 (29,47%)	812.657 (3,57%)
Efeito de inibidores no *M. perniciosa in vitro*	Tratamento com Sham	DPT1	27.402.510	19.312.645 (70,48%)	281 (0,00%)	320.566 (1,17%)	7.201.266 (26,28%)	567.752 (2,07%)
	Tratamento com Azoxistrobina	DPT2	24.559.919	18.247.250 (74,30%)	200 (0,00%)	254.573 (1,04%)	5.676.852 (23,11%)	381.044 (1,55%)
	Tratamento com Sham + Azoxistrobina	DPT3	25.839.171	19.052.814 (73,74%)	225 (0,00%)	196.732 (0,76%)	6.273.990 (24,28%)	315.410 (1,22%)
	Tratamento com Etanol	DPT4	28.144.082	20.639.222 (73,33%)	126 (0,00%)	573.815 (2,04%)	6.434.165 (22,86%)	496.754 (1,77%)
	Tratamento com Metanol	DPT5	25.714.105	18.022.505 (70,09%)	150 (0,00%)	1.272.477 (4,95%)	5.682.416 (22,10%)	736.557 (2,86%)
	Tratamento com Etanol + Metanol	DPT6	27.061.209	19.757.426 (73,01%)	154 (0,00%)	443.052 (1,64%)	6.445.158 (23,82%)	415.419 (1,54%)
Formação de estruturas reprodutivas do *M. perniciosa in vitro*	Basidioma	BDA4	30.384.732	22.050.675 (72,57%)	163 (0,00%)	888.917 (2,93%)	6.874.550 (22,63%)	570.427 (1,88%)
	Primórdios de basidioma	PRD1	29.033.649	20.904.890 (72,00%)	270 (0,00%)	1.519.054 (5,23%)	5.758.568 (19,83%)	850.867 (2,93%)

Experimentos	Bibliotecas	Código	Número de *reads*	Alinhados com *M. perniciosa*	Alinhados com *T. cacao*	Alinhados com rRNA	Não alinhados	Múltiplos
RNA-Seq Direcional	Micélio biotrófico	DIBIO1	19.097.253	13.057.768 (68,38%)	687 (0,00%)	246.682 (1,29%)	5.584.471 (29,24%)	207.645 (1,09%)
	Micélio necrotrófico	DISAP1	9.290.622	5.416.695 (58,30%)	776 (0,01%)	328.117 (3,53%)	3.390.102 (36,49%)	154.932 (1,67%)
	Basidioma	DIBDA4	32.104.741	13.973.079 (43,52%)	665 (0,00%)	591.917 (1,84%)	17.264.303 (53,77%)	274.777 (0,86%)
	Vassoura-verde	DIGB2	32.968.672	60.542 (0,18%)	20.960.239 (63,58%)	302.096 (0,92%)	11.101.307 (33,67%)	544.488 (1,65%)
	Controle sadio (vassoura-verde)	DIHCA1	31.258.985	306 (0,00%)	22.397.374 (71,65%)	456.254 (1,46%)	7.846.114 (25,10%)	558.937 (1,79%)
Indução de *M. perniciosa* com diferentes fontes de carbono	Glicerol 1%	OLP1	27.044.195	19.586.862 (72,43%)	142 (0,00%)	1.479.144 (5,47%)	5.056.309 (18,70%)	921.738 (3,41%)
	Glicose 1%	GliP1	24.732.386	17.699.544 (71,56%)	107 (0,00%)	1.287.722 (5,21%)	4.963.371 (20,07%)	781.642 (3,16%)
	Metanol 1%	METP1	25.290.296	18.864.455 (74,59%)	115 (0,00%)	972.035 (3,84%)	4.758.933 (18,82%)	694.758 (2,75%)
	Ácido poligalacturônico 1%	POP1	28.165.627	21.403.087 (75,99%)	68 (0,00%)	1.050.153 (3,73%)	5.091.884 (18,08%)	620.435 (2,20%)
	Pectina 1%	PECP1	31.776.484	23.308.870 (73,35%)	166 (0,00%)	683.353 (2,15%)	7.238.284 (22,78%)	545.811 (1,72%)
	Extrato de cacau 1%	CAP1	55.556.098	32.359.142 (58,25%)	16 (0,00%)	3.955.062 (7,12%)	18.968.997 (34,14%)	272.881 (0,49%)

As bibliotecas de controle sadio do cacaueiro apresentaram uma quantidade insignificante (em média 0,001%) de *reads* alinhando com os genes do *M. perniciosa*, e da mesma forma, as bibliotecas do fungo *in vitro*, apresentaram uma quantidade insignificante (em média 0,003%) de *reads* alinhando com os genes do cacaueiro. A existência destes alinhamentos inespecíficos pode ser atribuída ao pequeno tamanho dos *reads* (36pb) e por eles serem *single-end*. Uma proporção bem menor pode ser vista nos controles sadios das vassouras-verdes, onde os *reads* são *paired-end* e possuem 50pb.

Nas bibliotecas de ramo infectado, durante a progressão da doença a quantidade de *reads* do cacaueiro permanece constante nas bibliotecas de controle sadio, enquanto nas bibliotecas do cacaueiro infectado este número vai diminuindo, chegando a um mínimo de 0,45% na fase de "vassoura-seca" (Figura 9). Já a quantidade de *reads* do fungo aumenta durante a progressão da doença, porém chega a um máximo de apenas 1,95% (Figura 10).

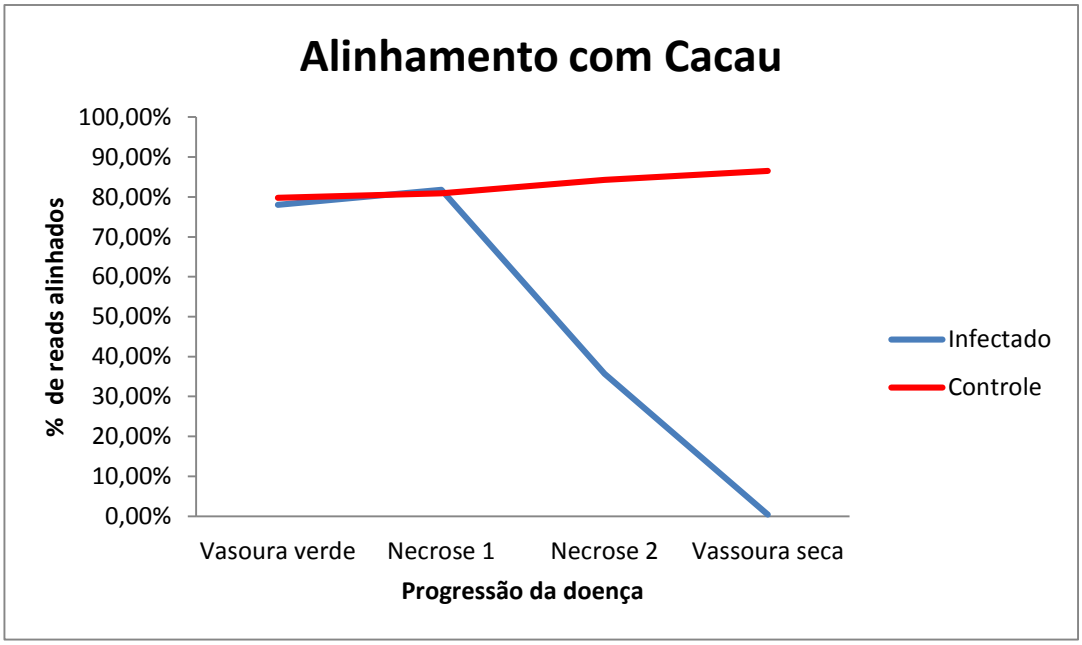

Figura 9 - Quantidade de *reads* do cacaueiro durante a progressão da doença nas bibliotecas de cacaueiro infectado e no respectivo controle sadio.

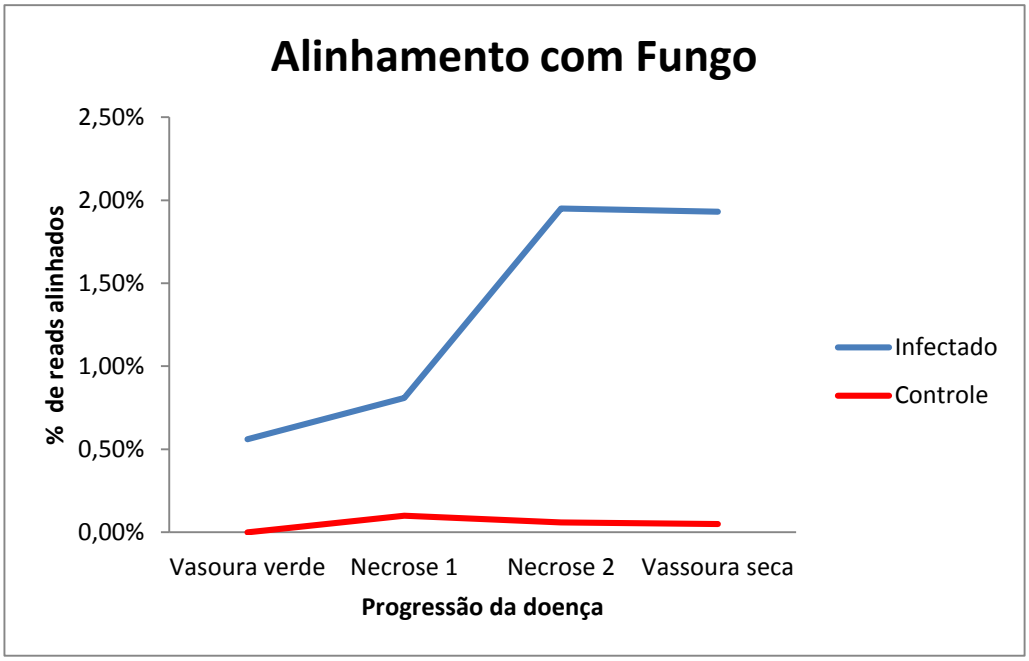

Figura 10 – Quantidade de *reads* do *M. perniciosa* durante a progressão da doença nas bibliotecas de cacaueiro infectado e no respectivo controle sadio.

Como durante o último estágio da doença 97,62% dos *reads* não alinharam nem contra os genes do cacaueiro nem contra os genes do *M. perniciosa*, o *pipeline* descrito na Figura 11 foi construído de forma a identificar a origem destes *reads*.

Inicialmente montamos *de novo*, os 97,62% *reads* que não alinharam com cacau nem com *M. perniciosa*, utilizando o *software* Trinity, gerando 41.275 *contigs* maiores que 200 pb. Estes *contigs* foram alinhados, utilizando o software Blastx, contra o banco de dados não redundante de proteínas do NCBI (NR), sendo que, 26479 dos *contigs* obtiveram *hits* com pelo menos uma proteína, usando-se um *e-value* de corte de $1x10^{-5}$. Verificando as espécies correspondentes a estes *hits*, identificamos que os *reads* não alinhados eram principalmente de fungos do grupo dos Ascomicetos, sugerindo a presença de outros fungos, possivelmente oportunistas, aproveitando a fragilidade da planta neste ultimo estágio da doença.

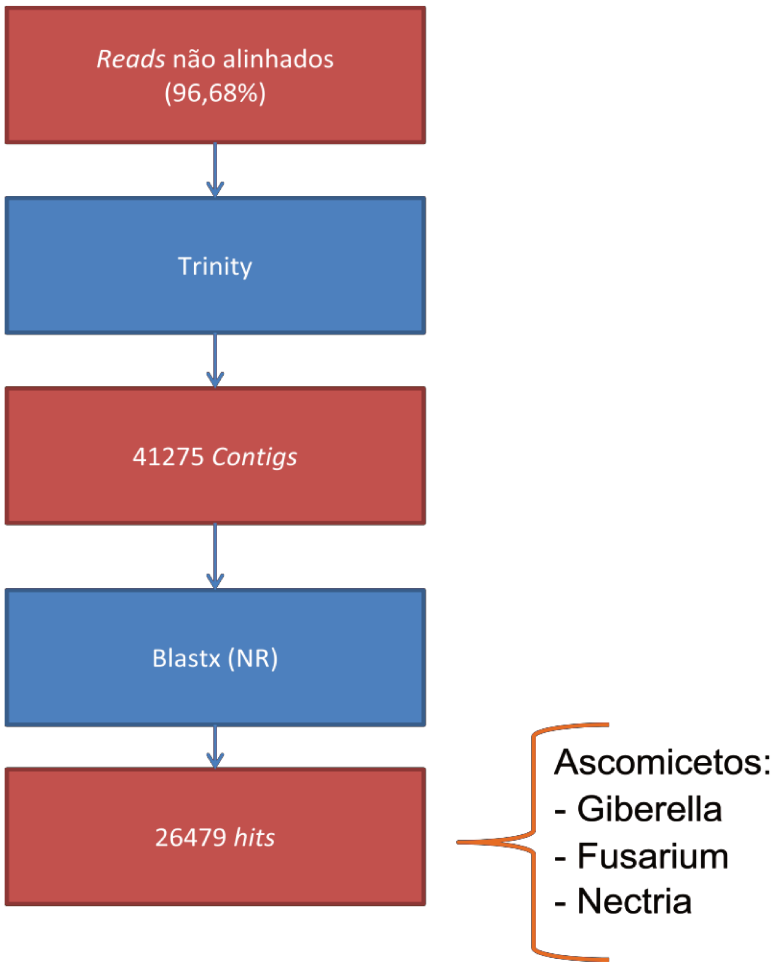

Figura 11 – Pipeline de análise dos *reads* não alinhados com cacaueiro nem com *M. perniciosa* no último estágio da doença.

Uma observação inesperada foi a presença de cerca de 18% de transcritos do *M. perniciosa* em uma biblioteca do controle sadio do fruto no experimento da progressão da doença nas cascas dos frutos. Porém, recentemente, novos sequenciamentos desta amostra e de outras réplicas biológicas (dados não apresentados), mostraram que não há fungo naquele tecido, indicando que a quantidade de fungo vista no sequenciamento anteriormente provavelmente foi proveniente de alguma contaminação na manipulação das amostras.

Utilizando o método de normalização por RPKM (Mortazavi, et al., 2008), calculamos o valor de expressão dos genes dos dois organismos, para cada uma das condições. Utilizando um limiar arbitrário de RPKM >= 1 em pelo menos uma das bibliotecas para considerar um gene como sendo expresso, nós conseguimos identificar

a expressão 15.521 genes do *M. perniciosa* e 21.321 genes do *T. cacao*. Os dados de expressão gerados a partir desta grande variedade de condições sequenciadas fornece uma ferramenta poderosa para a investigação e levantamento de hipóteses sobre a interação entre o *M. perniciosa* e o cacaueiro.

Como um dos resultados deste trabalho, estes dados fizeram parte de uma publicação do nosso grupo no qual foram feitas a caracterização molecular de 11 proteínas *PR-1* (*MpPR-1a, MpPR-1b, MpPR-1c, MpPR-1d, MpPR-1e, MpPR-1f, MpPR-1g, MpPR-1h, MpPR-1i, MpPR-1j* e *MpPR-1k*) do fungo *M perniciosa* durante sua interação com o cacaueiro (Teixeira, et al., 2012) (Anexo I). Em plantas, estas proteínas estão relacionadas à resposta de defesa (van Loon, et al., 2006), contudo seu papel em fungos durante a interação planta-patógeno ainda é desconhecida. Neste estudo foi mostrado que cinco *MpPR-1s* (*MpPR-1c, MpPR-1f, MpPR-1g, MpPR-1h* e *MpPR-1k*) são altamente expressas durante a fase biotrófica da doença vassoura-de-bruxa *in planta*, o que sugere que elas desempenham um papel importante na patogenicidade do *M. perniciosa*. Ainda duas destas *MpPR-1* (*MpPR-1f,* e *MpPR-1h*) também são expressas durante a germinação dos esporos, que é um estágio crucial para o estabelecimento da doença (Figura 12-16). Adicionalmente também foi visto que a *MpPR-1j* é notavelmente mais expressa nas amostras de basidioma e primórdios de basidioma, sugerindo sua importância neste estágio (Figura 17).

Figura 12 – Expressão do gene PR-1c ao longo de todas as bibliotecas.

Figura 13 – Expressão do gene PR-1f ao longo de todas as bibliotecas.

Figura 14 – Expressão do gene PR-1g ao longo de todas as bibliotecas.

Figura 15 – Expressão do gene PR-1h ao longo de todas as bibliotecas.

Figura 16 – Expressão do gene PR-1k ao longo de todas as bibliotecas.

Figura 17 – Expressão do gene PR-1j ao longo de todas as bibliotecas.

4.2 Vassoura-verde

Diferentemente de outros fungos hemibiotróficos que possuem uma fase biotrófica curta e assintomática, o fungo *M. perniciosa* possui uma fase biotrófica longa podendo durar até 9 semanas e causando sintomas bem definidos no cacaueiro, como hipertrofia, hiperplasia e perda de dominância apical (Scarpari, et al., 2005). Isto torna este estágio da doença, chamado de vassoura-verde, um modelo de estudo muito interessante. A fim de entender os mecanismos que permeiam essa interação durante a vassoura-verde, foram analisadas dez bibliotecas, sendo cinco réplicas biológicas de plantas infectadas e cinco réplicas biológicas de plantas controles.

Na análise do transcriptoma do cacaueiro, em ambas as condições, foram alinhados, em média, cerca de 80% dos *reads* nos genes do cacaueiro. Com estes alinhamentos conseguimos identificar a expressão de 17.013 genes com RPKM >=1 em todas as réplicas de pelo menos uma das condições. Na primeira etapa desta análise comparamos o perfil de expressão dos genes do cacaueiro nas amostras através de uma análise de componente principal (Figura 18).

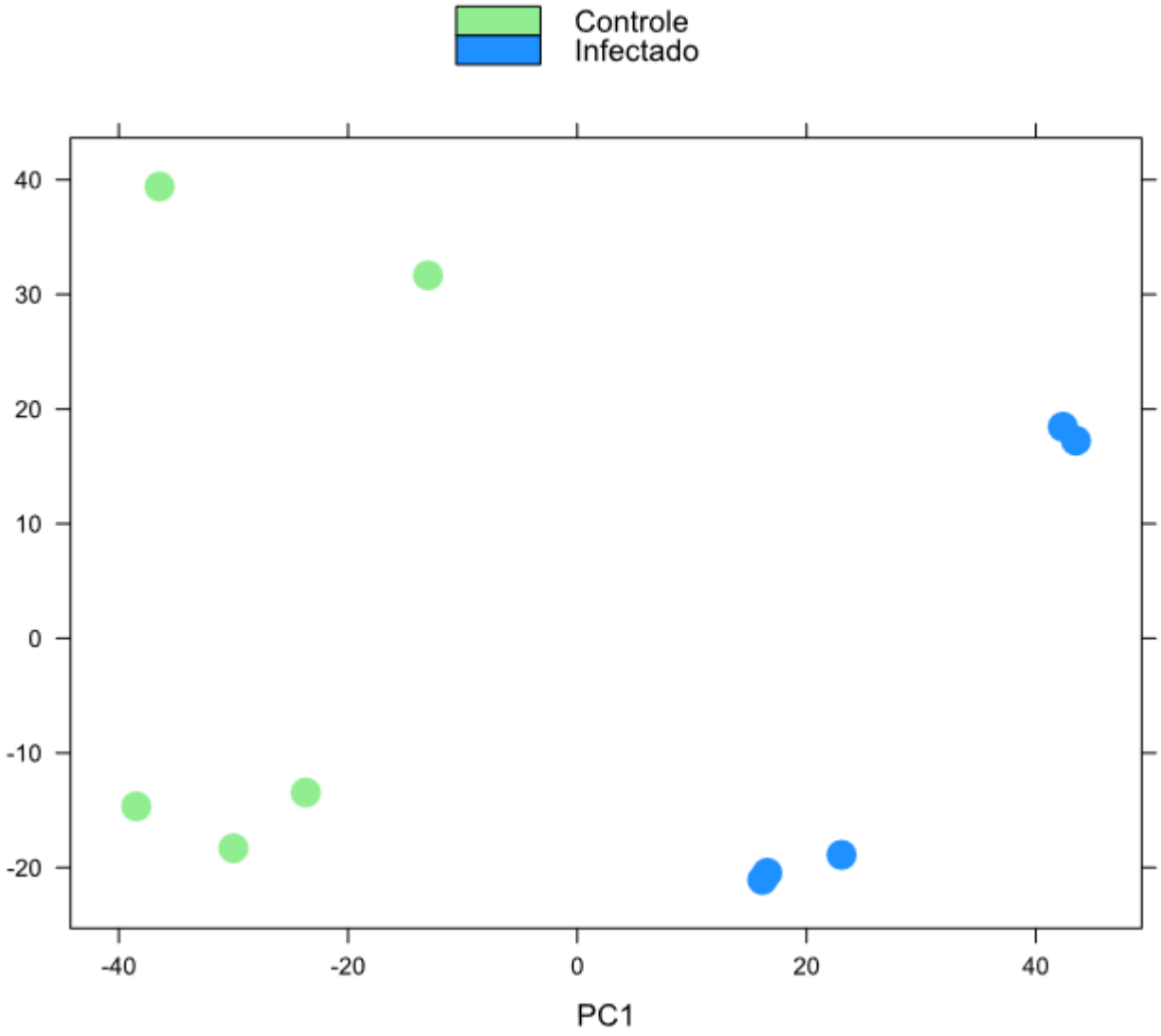

Figura 18 – Análise de componente principal das bibliotecas infectadas e controles do estágio “vassoura-verde”. O PC1 é a dimensão que melhor separa as amostras e o PC2 e a segunda dimensão que melhor separa as amostras. Os pontos em verde representam as bibliotecas controle e as azuis às infectadas.

Nesta análise o PC1 explica 47,9% da variação entre as amostras e separa claramente as plantas infectadas das não infectadas, indicando que a maior fonte de variação nestas amostras é causada pela doença. Pela dimensão PC2 também conseguimos identificar uma separação entre as amostras dentro de uma mesma condição, entretanto isso se deve ao fato de as amostras terem sido coletadas de dois experimentos que ocorreram tempos diferentes. Finalmente esta análise também é

importante para mostrar que nenhuma das réplicas representa um *outlier* e todas podem ser utilizadas para análise de expressão diferencial.

Na análise de expressão diferencial testamos os programas DESeq, edgeR, baySeq, que são os mais utilizado para análise de expressão diferencial. Consideramos como diferencialmente expressos os genes que tiveram um FDR < 1%. A Tabela 12 mostra o número de genes diferencialmente expressos encontrados por cada programa.

Tabela 12 – Número de genes diferencialmente expressos encontrados em cada programa.

Programas	**Número de genes diferencialmente expressos**
baySeq	953
DESeq	632
edgeR	1967
Total de genes únicos	**2184**

O edgeR parece apresentar o resultado menos conservador entre os programas testados, classificando cerca de duas vezes mais genes como diferencialmente expressos do que o baySeq e cerca de três vezes mais do que o DESeq. A Figura 19 mostra o diagrama de *Venn* com a comparação entre os resultados dos quatro *softwares*. Praticamente todos os genes considerados como diferencialmente expressos pelo DESeq também o foram pelo edgeR. Já o baySeq possui cerca de 70% dos genes identificados como diferencialmente expressos em concordância com o edgeR.

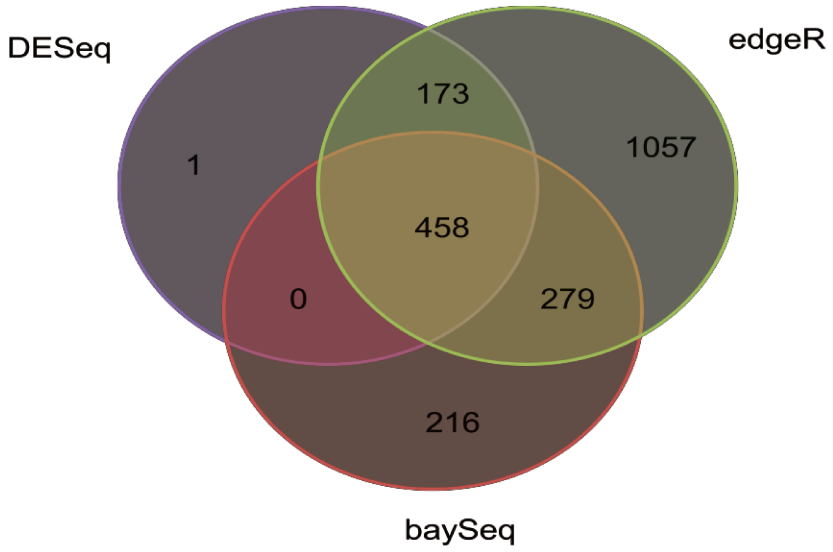

Figura 19 – Diagrama de Venn comparando os genes diferencialmente expressos encontrados pelos programas DESeq, edgeR e baySeq.

A Figura 20 apresenta a distribuição de frequências dos genes considerados expressos por cada um dos programas. Apesar da grande diferença no número de genes encontrados por cada programa, é possível observar que parece não existir uma preferência em nenhum dos programas por genes muito expressos ou pouco expressos.

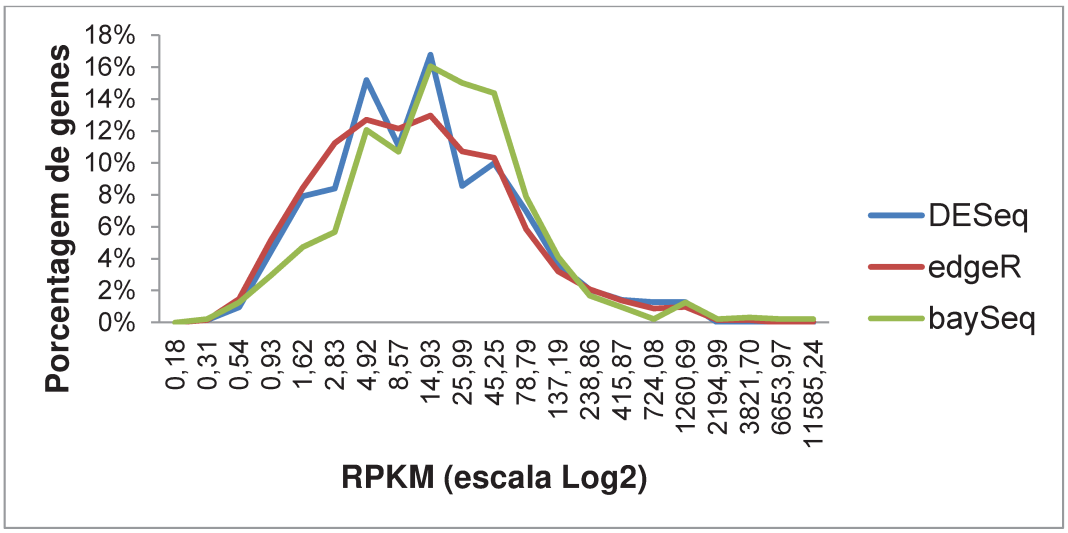

Figura 20 – Distribuição de RPKM dos genes considerados diferencialmente expressos pelos programas DESeq, edgeR e baySeq.

Contudo, analisando a distribuição de *fold-change* destes mesmos genes (Figura 21), podemos observar que o baySeq encontrou mais genes com um f*old-change* mais baixos que os demais programas. A fim de confirmar se estes genes são os específicos encontrados por cada programa, realizamos está mesma análise para os genes únicos de cada programa e os genes que foram encontrados por todos os programas, que aqui chamamos de *Core*. Excluímos desta parte da análise o resultado do programa DESeq, pois ele apresenta apenas 1 gene específico, sendo que os demais genes foram todos encontrados pelo edgeR. A distribuição de expressão dos genes (Figura 22), mostra que entre os genes específicos encontrados pelo baySeq, a maioria tem expressão alta, enquanto que os específicos do edgeR e os do *Core* estão mais distribuídos ao longo dos níveis de expressão. A Figura 23 também mostra que dentre estes genes específicos do baySeq, a maioria tem um *fold-change* mais baixo em relação ao edgeR ou *Core*.

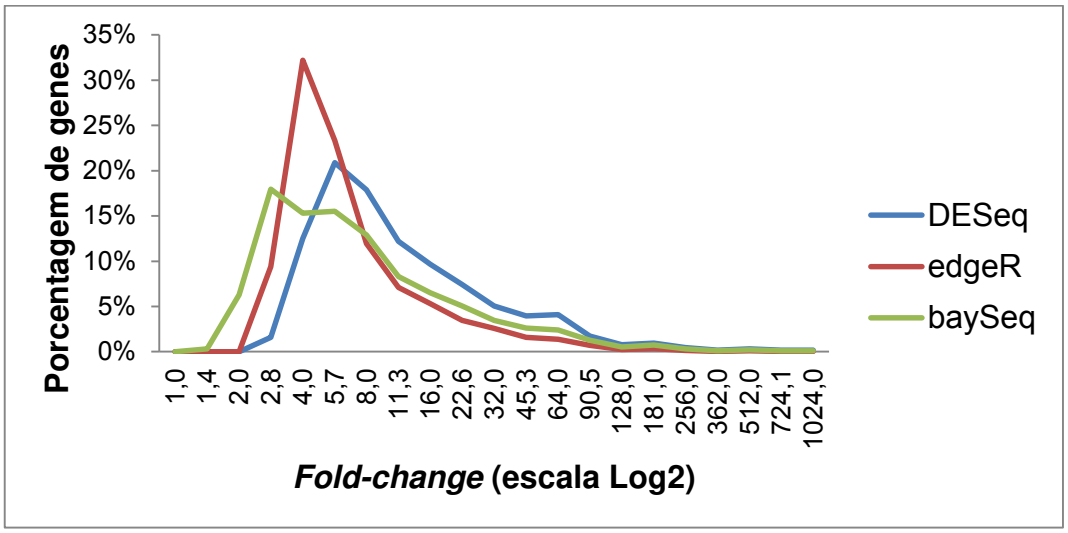

Figura 21 – Distribuição de *fold-change* dos genes considerados diferencialmente expressos pelos programas DESeq, edgeR e baySeq.

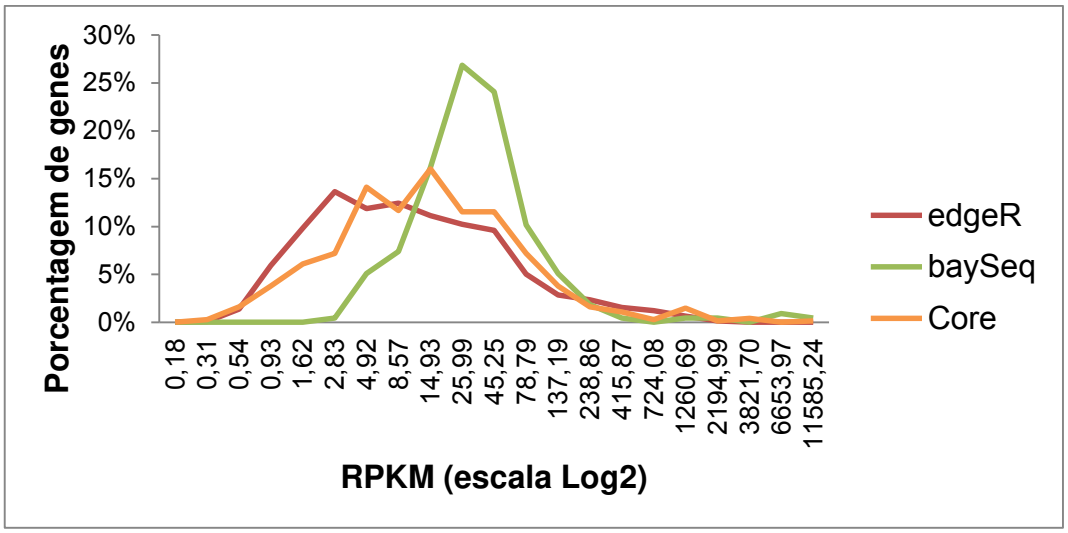

Figura 22 – Distribuição de RPKM dos genes específicos considerados diferencialmente expressos pelos programas edgeR, baySeq e pelo *Core*.

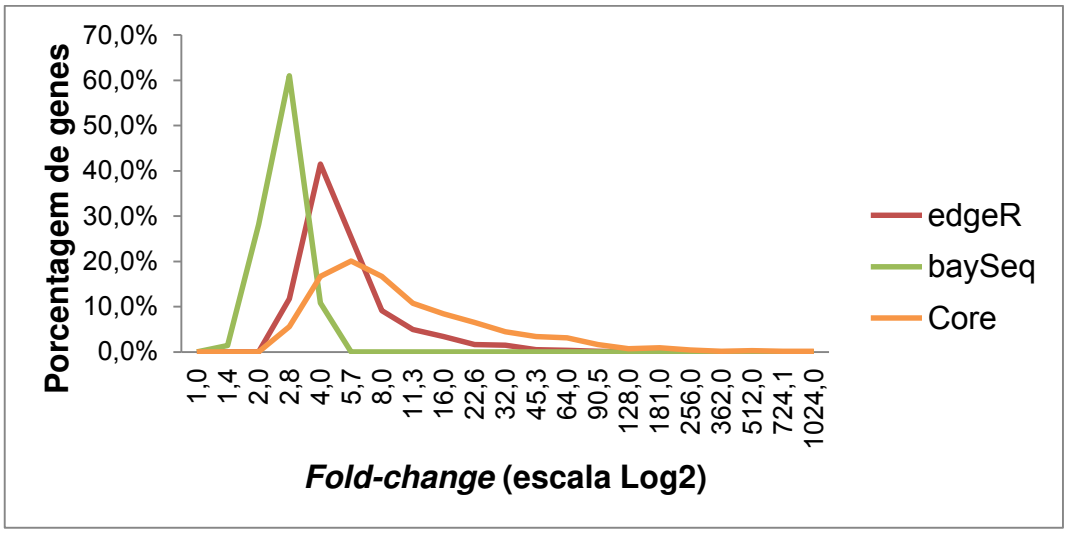

Figura 23 – Distribuição de *fold-change* dos genes específicos considerados diferencialmente expressos pelos programas edgeR, baySeq e pelo *Core*.

Outros trabalhos comparativos já testaram estes programas em diversas situações e concluíram que nenhum dos métodos é ótimo em todas as circunstâncias e que a escolha do método depende das condições experimentais (Rapaport, et al., 2013; Soneson and Delorenzi, 2013). Em nossa análise o programa edgeR encontrou mais genes diferencialmente expressos que os demais programas, o que já foi observado em (Nookaew, et al., 2012). O fato de o programa DESeq ter achado o menor número de genes diferencialmente expressos está consistente com o fato de ele ser bastante

conservador (Soneson and Delorenzi, 2013). O baySeq normalmente tem uma performance comparável com o edgeR e DESeq (Kvam, et al., 2012), contudo em nossa análise ele apresenta, entre seus genes específicos, muitos genes com expressão alta, porém com baixos valores de *fold-change.* Como este trabalho inicial tem um caráter exploratório e pouco se sabe sobre os mecanismos dessa interação, optamos por utilizar os resultados obtidos pelo edgeR, que apresenta um resultado menos conservador que o DESeq e mais consistente em termos de *fold-change* através dos vários níveis de expressão do que o baySeq.

Dos 34.997 genes do *T. cacao*, 1.967 foram identificados como diferencialmente expressos, sendo 1.269 estavam induzidos e 698 reprimidos. Uma análise de enriquecimento de termos GO foi feita utilizando os genes induzidos (Tabela 13) e reprimidos (Tabela 14).

Tabela 13 – Termos GO enriquecidos nos genes do cacaueiro induzidos durante vassoura-verde

GO-ID	Termo	Categoria	FDR	Genes induzidos	Referência
GO:0006950	response to stress	P	$2{,}21\text{x}10^{-44}$	377	3825
GO:0009719	response to endogenous stimulus	P	$7{,}08\text{x}10^{-31}$	229	2046
GO:0009607	response to biotic stimulus	P	$8{,}21\text{x}10^{-26}$	142	1035
GO:0009628	response to abiotic stimulus	P	$5{,}15\text{x}10^{-13}$	188	2210
GO:0019748	secondary metabolic process	P	$1{,}19\text{x}10^{-11}$	97	893
GO:0005975	carbohydrate metabolic process	P	$5{,}79\text{x}10^{-10}$	97	969
GO:0009653	anatomical structure morphogenesis	P	$8{,}65\text{x}10^{-10}$	147	1753
GO:0009991	response to extracellular stimulus	P	$9{,}59\text{x}10^{-07}$	37	277
GO:0007165	signal transduction	P	$8{,}72\text{x}10^{-06}$	142	1978
GO:0009908	flower development	P	$3{,}83\text{x}10^{-4}$	43	469
GO:0006629	lipid metabolic process	P	$2{,}91\text{x}10^{-3}$	82	1180
GO:0016049	cell growth	P	$4{,}86\text{x}10^{-3}$	33	379
GO:0009058	biosynthetic process	P	$5{,}15\text{x}10^{-3}$	147	2401

Tabela 14 - Termos GO enriquecidos nos genes do cacaueiro reprimidos durante vassoura-verde.

GO-ID	Termo	Categoria	FDR	Genes reprimidos	Referência
GO:0006091	generation of precursor metabolites and energy	P	$9{,}98\text{x}10^{-18}$	36	203
GO:0015979	Photosynthesis	P	$1{,}46\text{x}10^{-15}$	13	8
GO:0009058	biosynthetic process	P	$2{,}50\text{x}10^{-3}$	87	2461
GO:0009628	response to abiotic stimulus	P	$1{,}97\text{x}10^{-2}$	78	2320

Entre os processos que possuem genes induzidos destacamos a resposta a estresse, resposta a estimulo biótico, metabolismo secundário, metabolismo de carboidratos e metabolismo de lipídios. Já entre os processos que tem genes reprimidos, destacamos a fotossíntese.

A indução de genes relacionados à reposta a estimulo biótico e metabolismo secundário sugere que durante esta interação a planta ativa respostas de defesa. Já a indução de genes associados metabolismo de carboidratos e lipídeos, juntamente com a repressão de genes associados à fotossíntese é uma evidência de que, neste estágio, o tecido infectado deixa de ser uma fonte e passa a ser um dreno de nutrientes, o que potencialmente pode afetar toda a planta.

Analisando o transcriptoma do fungo nas bibliotecas deste experimento, identificamos a expressão 8617 genes com RPKM >=1 em todas as réplicas de planta infectada, apesar da baixa quantidade de *reads* alinhados (em média 0,3% nas bibliotecas de planta infectada). A baixa quantidade de *reads* reflete a densidade de células do fungo, que neste estágio vive apenas no espaço intercelular da planta.

A fim de identificar genes do fungo expressos preferencialmente durante a interação biotrófica com o cacaueiro, foi realizada uma análise de *clustering* dos padrões de expressão dos genes do *M. perniciosa* ao longo das bibliotecas. Foi selecionado um subgrupo deste *clustering* contendo 214 genes preferencialmente expressos durante a interação biotrófica (Figura 24). Estes genes foram submetidos à análise de enriquecimento de termos GO (Tabela 15).

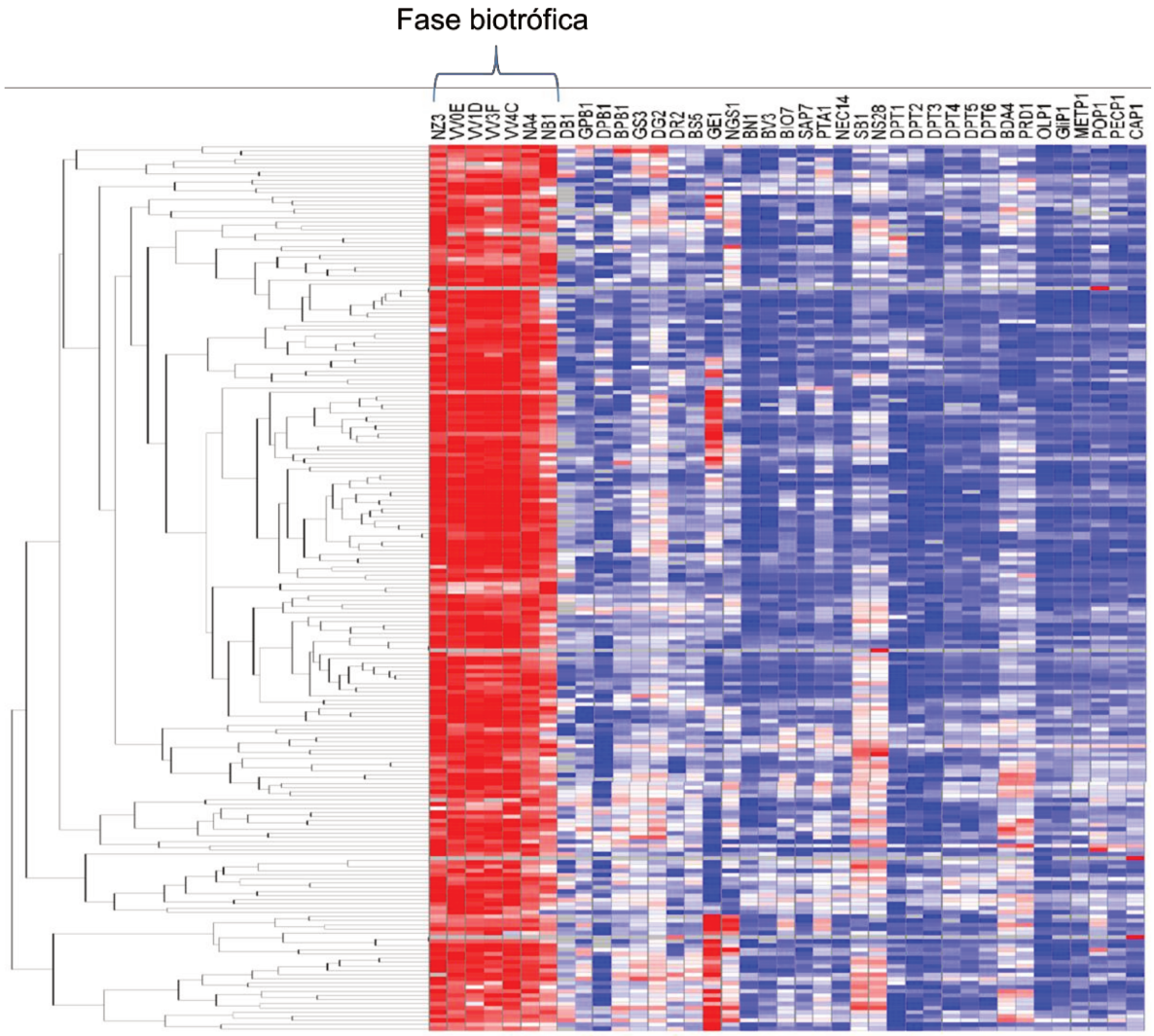

Figura 24 – Identificação de genes preferencialmente expressos pelo *M. perniciosa* durante a interação biotrófica com o cacaueiro. A cor vermelha representa uma alta expressão dos genes enquanto a cor azul representa uma baixa expressão.

Tabela 15 - Termos GO enriquecidos nos genes do *M. perniciosa* preferencialmente expressos durante vassoura-verde

GO-ID	Termo	Categoria	FDR	Genes específicos	Referência
GO:0055085	transmembrane transport	P	2,00E-02	19	456
GO:0006528	asparagine metabolic process	P	3,73E-02	3	4
GO:0005985	sucrose metabolic process	P	3,73E-02	6	51
GO:0005982	starch metabolic process	P	3,73E-02	6	51

Somente quatro processos estão enriquecidos, sendo que um é de transporte transmembrana, que pode ser relacionado à obtenção de nutrientes e os outros três

são de processos metabólicos, porém nenhum termo relacionado à patogenicidade do fungo foi encontrado. Muito provavelmente isso se deva ao fato muitos genes do *M. perniciosa* ainda não terem anotação funcional. Dos 214 genes analisados 84 não possuem termo GO associado.

4.3 Identificação de possíveis lncRNAs

Além da análise de expressão, outra utilidade do RNA-Seq é na descoberta de novos transcritos. Neste trabalho, nós construímos um *pipeline* (Figura 25) para anotação de novos transcritos, porém focamos na descoberta de lincRNA e ancRNA.

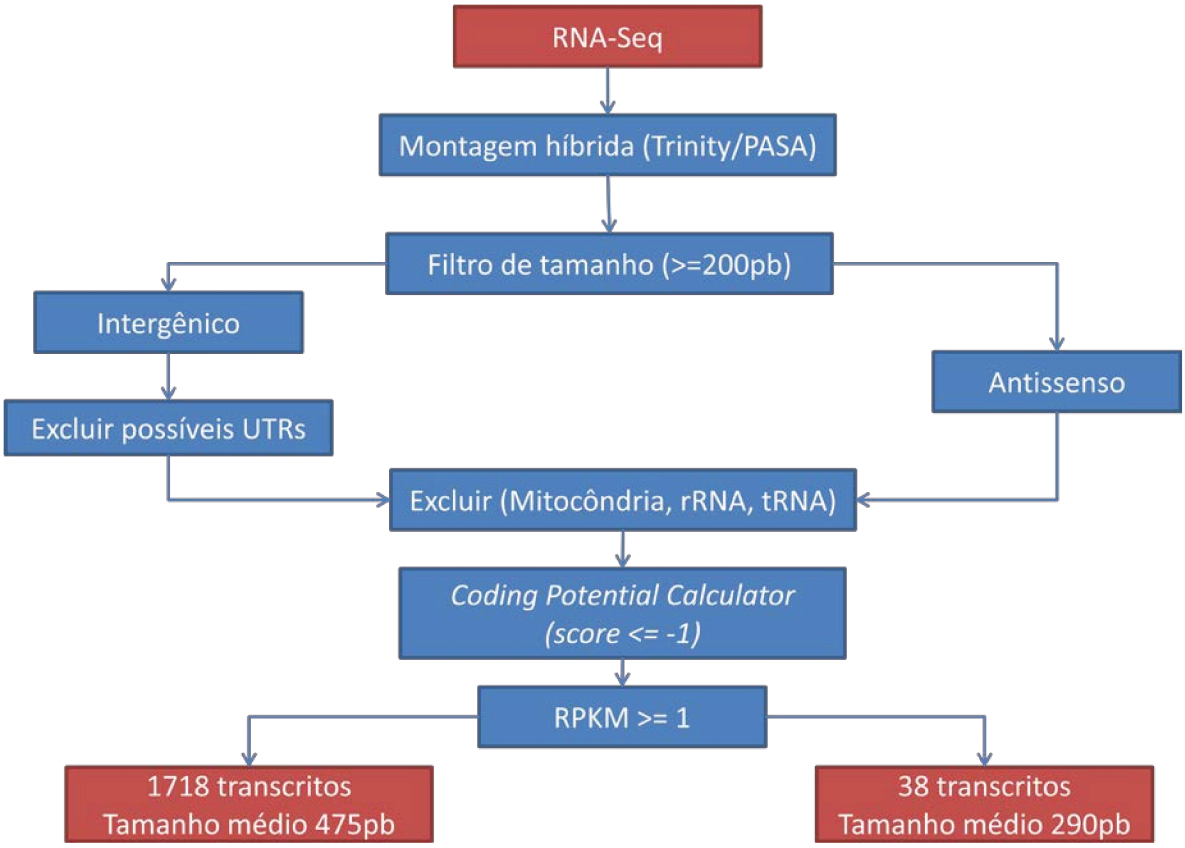

Figura 25 – *Pipeline* de identificação de lincRNA e ancRNA.

Os transcritos foram montados seguindo o *pipeline* de montagem guiada do *software* Trinity (disponível em http://trinityrnaseq.sourceforge.net/genome_guided_trinity.html). As bibliotecas

direcionais e não direcionais foram montadas separadamente. Os transcritos montados pelas bibliotecas direcionais foram utilizados para buscar ancRNA, enquanto os transcritos montados pelas bibliotecas não direcionais foram utilizados para encontrar os lincRNAs. Na busca dos ancRNAs, foram utilizados apenas os transcritos que tinham sobreposição na fita antissenso com algum dos genes já preditos, enquanto que na busca dos lincRNA, foram utilizados apenas os transcritos que estavam a pelo menos 500pb de distância de qualquer gene já predito. O conjunto de transcritos restantes destes dois processos foi submetido, primeiramente à exclusão de possíveis transcritos provenientes do genoma mitocondrial e sequências de rRNA utilizando o *software* Blastn (Altschul, et al., 1990) com um *e-value* de corte de $1x10^{-5}$, e em seguida a exclusão de possíveis tRNAs através do *software* tRNAScan (Schattner, et al., 2005). Os transcritos que sobraram foram submetidos a uma análise de potencial de codificação utilizando o *software* CPC (Kong, et al., 2007) e foram mantidos apenas os que obtiveram um *score* <= -1. A fim de evitar transcritos provenientes de ruídos, mantivemos apenas os que possuíssem uma expressão >= 1 RPKM em pelo menos uma das bibliotecas. No final do processo foram identificados 1718 candidatos a lincRNAs com tamanho médio de 475pb e 38 candidatos ancRNA com tamanho médio de 290pb.

Dentre os possíveis lncRNAs identificados, destacamos um ancRNA que está sendo expresso na fita oposta ao gene MP01336. Este gene faz parte de um *transposon* que está localizado entre os genes HD1 e HD2 do lócus A do *mating-type* (Figura 26).

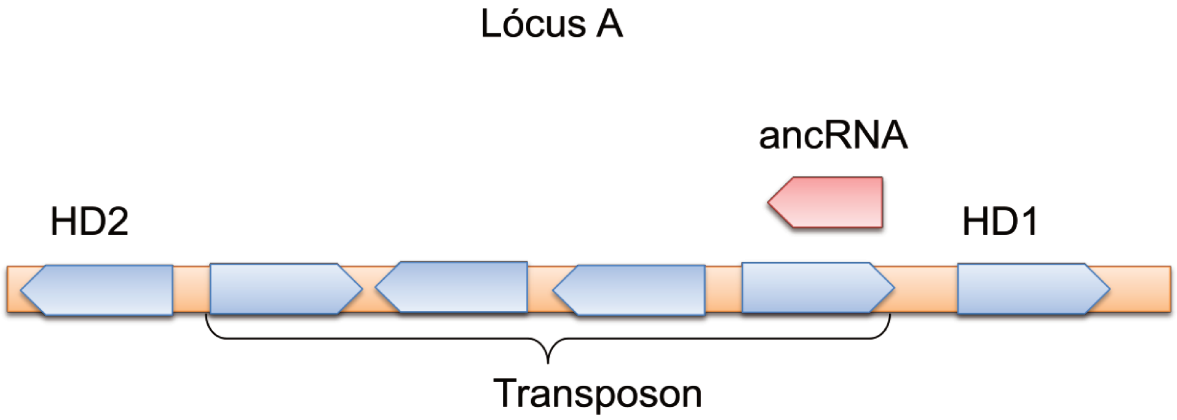

Figura 26 – Estrutura gênica do lócus A do *mating–type* no fungo *M. perniciosa*.

Nas três bibliotecas direcionais o ancRNA é notavelmente mais expresso que o gene na fita oposta (Figura 27).

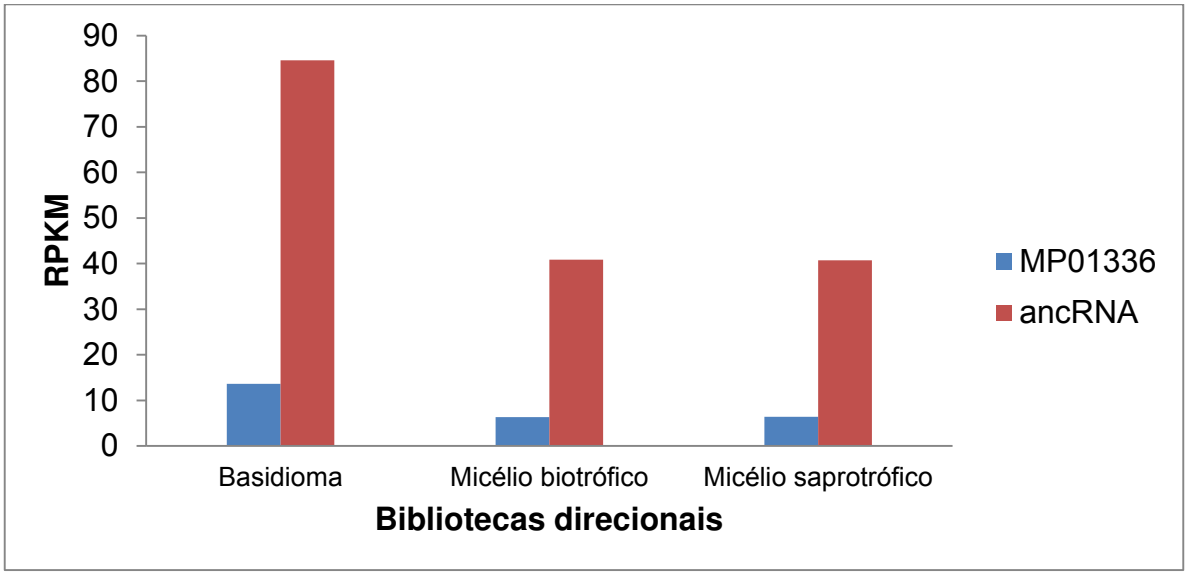

Figura 27 – Expressão do gene MP01336 e do ancRNA na fita oposta.

O alinhamento dos *reads* das bibliotecas direcionais no genoma também mostra um número maior de *reads* alinhando com a fita oposta ao *transposon* (Figura 28).

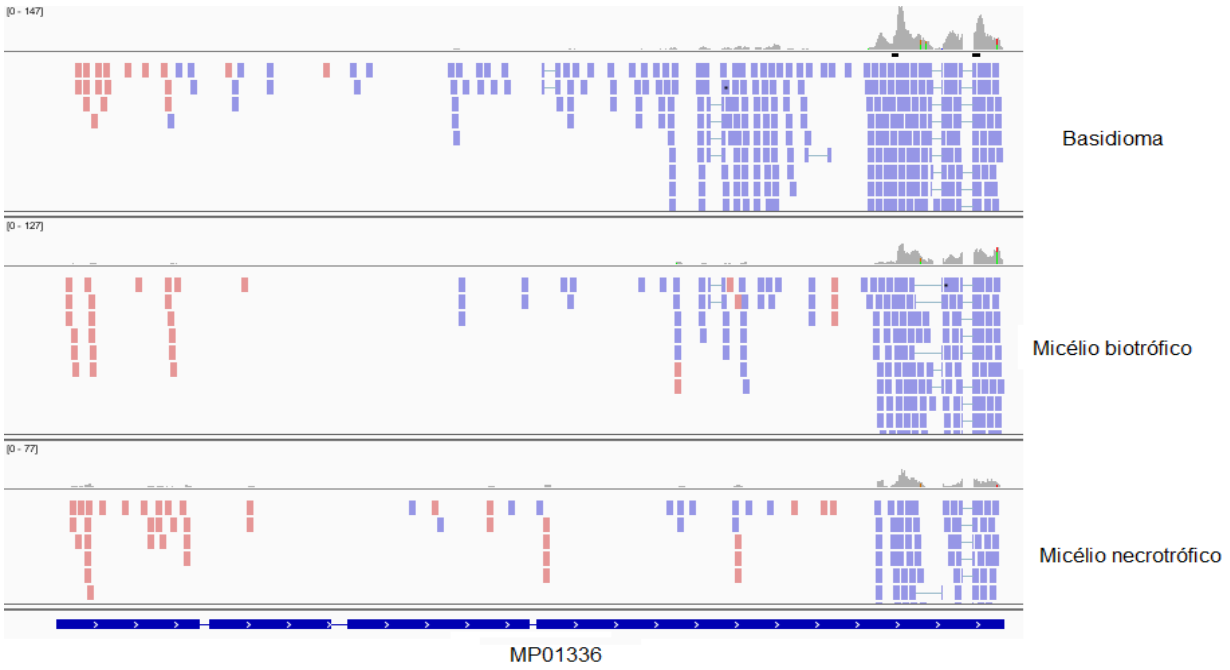

Figura 28 – Alinhamento dos *reads no* gene MP01336. Os traços de cor rosa representam os *reads* alinhando na fita mais, enquanto que os traços azuis representam os *reads* alinhando na fita menos.

Futuras análises de bioinformática para tentar inferir funções a estes lncRNAs incluem análise de co-expressão dos lncRNA juntamente com os genes que codificam proteínas e predição de interação entre lncRNAs e proteínas. Estas análises permitirão encontrar candidatos para serem explorados.

5 CONCLUSÃO

Neste trabalho nós realizamos as análises de RNA-Seq do Atlas Transcriptômico da doença Vassoura–de-Bruxa. A tecnologia de RNA-Seq provou ser uma ferramenta poderosa para análise de transcriptomas, sendo muito útil para identificar tanto a expressão de RNAs codificantes como de RNAs não codificantes, como os lncRNAs. Esta tecnologia também é muito útil para análise de interação entre organismo, principalmente por apresentar uma alta sensibilidade em detectar transcritos quando um dos organismos está em baixa concentração, como no caso do estágio da vassoura-verde. Contudo, existem diversas ferramentas e *pipelines* para estas análises e nenhum deles parece ser ótimo em todas as situações. Na análise de expressão diferencial nos dados da vassoura-verde testamos três programas comumente utilizado, que apesar de serem baseados em princípios similares, apresentaram resultados distintos.

O conjunto de bibliotecas geradas neste projeto é grande e diversificado, oferecendo assim uma ferramenta extremamente útil para investigação e levantamento de hipóteses sobre os mecanismos de interação entre os dois organismos, como por exemplo, a hipótese de que alguns genes PR-1 estão relacionados à patogenicidade do *M. perniciosa*. A interação biotrófica entre o *M. perniciosa* e o cacaueiro, diferentemente da maioria dos fungos, é longa e apresenta sintomas bem definidos, sendo assim um alvo interessante de estudo. As nossas análises das bibliotecas da vassoura-verde estão possibilitando a criação de um modelo desta interação.

A área de estudo de lncRNAs ainda está no começo e poucas destas sequências tem função atribuída, porém pesquisas indicam que elas possuem um papel importante na regulação de genes. A identificação de possíveis lncRNAs apresentada aqui representa apenas o primeiro passo no estudo destas sequências.

REFERÊNCIAS

Altschul, S.F.*, et al.* (1990) Basic local alignment search tool, *J Mol Biol*, **215**, 403-410.

Anders, S. and Huber, W. Differential expression analysis for sequence count data, *Nature Precedings*.

Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data, *Genome Biol*, **11**, R106.

Birol, I.*, et al.* (2009) De novo transcriptome assembly with ABySS, *Bioinformatics*, **25**, 2872-2877.

Camarena, L.*, et al.* (2010) Molecular mechanisms of ethanol-induced pathogenesis revealed by RNA-sequencing, *PLoS Pathog*, **6**, e1000834.

Chen, G., Wang, C. and Shi, T. (2011) Overview of available methods for diverse RNA-Seq data analyses, *Sci China Life Sci*, **54**, 1121-1128.

Conesa, A.*, et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, *Bioinformatics*, **21**, 3674-3676.

Crawford, J.E.*, et al.* (2010) De Novo Transcriptome Sequencing in Anopheles funestus Using Illumina RNA-Seq Technology, *PLoS One*, **5**, -.

Dobin, A.*, et al.* (2013) STAR: ultrafast universal RNA-seq aligner, *Bioinformatics*, **29**, 15-21.

Evans, H.C. (1980) Pleomorphism in Crinipellis-Perniciosa, Causal Agent of Witches Broom Disease of Cocoa, *T Brit Mycol Soc*, **74**, 515-523.

Garg, R.*, et al.* (2011) De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification, *DNA Res*, **18**, 53-63.

Grabherr, M.G.*, et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome, *Nat Biotechnol*, **29**, 644-652.

Griffith, G.W.*, et al.* (2003) Witches' brooms and frosty pods: Two major pathogens of cacao, *Journal of Botany*, **41**, 423-435.

Guttman, M.*, et al.* (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs, *Nat Biotechnol*, **28**, 503-510.

Haas, B. and Zody, M. (2010) Advancing RNA-Seq analysis, *Nature Biotechnology*, **28**, 421-423.

Haas, B.J., *et al.* (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies, *Nucleic Acids Res*, **31**, 5654-5666.

Hardcastle, T. and Kelly, K. (2010) baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data, *BMC Bioinformatics*, **11**, 422.

Horner, D.S., *et al.* (2010) Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing, *Brief Bioinform*, **11**, 181-197.

Kong, L., *et al.* (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine, *Nucleic Acids Res*, **35**, W345-349.

Kvam, V.M., Liu, P. and Si, Y. (2012) A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data, *Am J Bot*, **99**, 248-256.

Langmead, B., *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome Biol*, **10**, R25.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, **25**, 1754-1760.

Li, R., *et al.* (2009) SOAP2: an improved ultrafast tool for short read alignment, *Bioinformatics*, **25**, 1966-1967.

Marguerat, S. and Bähler, J. (2010) RNA-seq: from technology to biology, *Cellular and molecular life sciences : CMLS*, **67**, 569-579.

Mercer, T.R., Dinger, M.E. and Mattick, J.S. (2009) Long non-coding RNAs: insights into functions, *Nature Reviews Genetics*, **10**, 155-159.

Mizrachi, E., *et al.* (2010) De novo assembled expressed gene catalog of a fast-growing Eucalyptus tree produced by Illumina mRNA-Seq, *BMC Genomics*, **11**, 681.

Mondego, J.M., *et al.* (2008) A genome survey of Moniliophthora perniciosa gives new insights into Witches' Broom Disease of cacao, *BMC Genomics*, **9**, 548.

Mortazavi, A., *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq, *Nat Methods*, **5**, 621-628.

Motamayor, J.C., *et al.* (2013) The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color, *Genome Biol*, **14**, r53.

Nam, J.W. and Bartel, D.P. (2012) Long noncoding RNAs in C. elegans, *Genome Research*, **22**, 2529-2540.

Nookaew, I., *et al.* (2012) A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison

with microarrays: a case study in Saccharomyces cerevisiae, *Nucleic Acids Res*, **40**, 10084-10097.

Pang, K.C., Frith, M.C. and Mattick, J.S. (2006) Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function, *Trends Genet*, **22**, 1-5.

Pauli, A., Rinn, J.L. and Schier, A.F. (2011) Non-coding RNAs as regulators of embryogenesis, *Nature Reviews Genetics*, **12**, 136-149.

Pauli, A.*, et al.* (2012) Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis, *Genome Res*, **22**, 577-591.

Pereira, J.L.*, et al.* (1990) First occurrence of witches' broom disease in the principal cocoa-growing region of Brazil, *Tropical Agriculture*, **67**, 188-189.

Purdy, L.H. and Schmidt, R.A. (1996) STATUS OF CACAO WITCHES' BROOM: biology, epidemiology, and management, *Annu Rev Phytopathol*, **34**, 573-594.

Rapaport, F.*, et al.* (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data, *Genome Biol*, **14**, R95.

Robinson, M., McCarthy, D. and Smyth, G. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics (Oxford, England)*, **26**, 139-140.

Robinson, M.D. and Smyth, G.K. (2007) Moderated statistical tests for assessing differences in tag abundance, *Bioinformatics*, **23**, 2881-2887.

Rudgard, S.A. (1986) Witches' broom disease on cocoa in Rondonia, Brazil: basidiocarp production on detached brooms in the field, *Plant Pathology*, **35**, 434-442.

Scarpari, L.M.*, et al.* (2005) Biochemical changes during the development of witches' broom: the most important disease of cocoa in Brazil caused by Crinipellis perniciosa, *J. Exp. Bot.*, **56**, 865-877.

Schattner, P., Brooks, A.N. and Lowe, T.M. (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs, *Nucleic Acids Res*, **33**, W686-689.

Schena, M.*, et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, **270**, 467-470.

Schulz, M.H.*, et al.* (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels, *Bioinformatics*, **28**, 1086-1092.

Shi, L.*, et al.* (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements, *Nat Biotechnol*, **24**, 1151-1161.

Soneson, C. and Delorenzi, M. (2013) A comparison of methods for differential expression analysis of RNA-seq data, *BMC Bioinformatics*, **14**, 91.

Strickler, S.R., Bombarely, A. and Mueller, L.A. (2012) Designing a transcriptome next-generation sequencing project for a nonmodel plant species, *Am J Bot*, **99**, 257-266.

Struhl, K. (2007) Transcriptional noise and the fidelity of initiation by RNA polymerase II, *Nat Struct Mol Biol*, **14**, 103-105.

Teixeira, P.J.*, et al.* (2012) The fungal pathogen Moniliophthora perniciosa has genes similar to plant PR-1 that are highly expressed during its interaction with cacao, *PLoS ONE*, **7**, e45929.

Trapnell, C., Pachter, L. and Salzberg, S. (2009) TopHat: discovering splice junctions with RNA-Seq, *Bioinformatics (Oxford, England)*, **25**, 1105-1111.

Trapnell, C.*, et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, *Nat Protoc*, **7**, 562-578.

Trapnell, C. and Salzberg, S.L. (2009) How to map billions of short reads onto genomes, *Nat Biotechnol*, **27**, 455-457.

van Loon, L.C., Rep, M. and Pieterse, C.M. (2006) Significance of inducible defense-related proteins in infected plants, *Annu Rev Phytopathol*, **44**, 135-162.

van Vliet, A.H. (2010) Next generation sequencing of microbial transcriptomes: challenges and opportunities, *FEMS Microbiol Lett*, **302**, 1-7.

Wang, E.T.*, et al.* (2008) Alternative isoform regulation in human tissue transcriptomes, *Nature*, **456**, 470-476.

Wang, L.*, et al.* (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data, *Bioinformatics*, **26**, 136-138.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics, *Nature Reviews Genetics*, **10**, 57-63.

Wu, T. and Nacu, S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads, *Bioinformatics*, **26**, 873-881.

Wu, T.D. and Nacu, S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads, *Bioinformatics*, **26**, 873-881.

Wu, T.D. and Watanabe, C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences, *Bioinformatics*, **21**, 1859-1875.

ANEXOS

Anexo I

The fungal pathogen *Moniliophthora perniciosa* has genes similar to plant PR-1 that are highly expressed during its interaction with cacao

Paulo José P.L. Teixeira*, Daniela P.T. Thomazella*, Ramon O. Vidal, Paula F.V. do Prado, Osvaldo Reis, Renata M. Baroni, Sulamita F. Franco, Piotr Mieczkowski, Gonçalo A.G. Pereira, Jorge M.C. Mondego

* Autores com igual contribuição

Trabalho publicado na revista *PLoS ONE*, Setembro de 2012, Vol. 9: e45929.

OPEN ACCESS Freely available online

PLOS ONE

The Fungal Pathogen *Moniliophthora perniciosa* Has Genes Similar to Plant PR-1 That Are Highly Expressed during Its Interaction with Cacao

Paulo J.P.L. Teixeira[1☯], Daniela P.T. Thomazella[1☯], Ramon O. Vidal[1,2], Paula F.V. do Prado[1], Osvaldo Reis[1], Renata M. Baroni[1,3], Sulamita F. Franco[1], Piotr Mieczkowski[4], Gonçalo A.G. Pereira[1]*, Jorge M.C. Mondego[3]

1 Departamento de Genética, Evolução e Bioagentes, Universidade Estadual de Campinas, Campinas, São Paulo, Brazil, **2** Laboratório Nacional de Biociências, Campinas, São Paulo, Brazil, **3** Centro de Pesquisa e Desenvolvimento em Recursos Genéticos Vegetais, Instituto Agronômico de Campinas, Campinas, São Paulo, Brazil, **4** Department of Genetics, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America

Abstract

The widespread SCP/TAPS superfamily (SCP/Tpx-1/Ag5/PR-1/Sc7) has multiple biological functions, including roles in the immune response of plants and animals, development of male reproductive tract in mammals, venom activity in insects and reptiles and host invasion by parasitic worms. Plant Pathogenesis Related 1 (PR-1) proteins belong to this superfamily and have been characterized as markers of induced defense against pathogens. This work presents the characterization of eleven genes homologous to plant *PR-1* genes, designated as *MpPR-1*, which were identified in the genome of *Moniliophthora perniciosa*, a basidiomycete fungus responsible for causing the devastating witches' broom disease in cacao. We describe gene structure, protein alignment and modeling analyses of the MpPR-1 family. Additionally, the expression profiles of *MpPR-1* genes were assessed by qPCR in different stages throughout the fungal life cycle. A specific expression pattern was verified for each member of the *MpPR-1* family in the conditions analyzed. Interestingly, some of them were highly and specifically expressed during the interaction of the fungus with cacao, suggesting a role for the MpPR-1 proteins in the infective process of this pathogen. Hypothetical functions assigned to members of the *MpPR-1* family include neutralization of plant defenses, antimicrobial activity to avoid competitors and fruiting body physiology. This study provides strong evidence on the importance of *PR-1-like* genes for fungal virulence on plants.

Citation: Teixeira PJPL, Thomazella DPT, Vidal RO, Prado PFVdo, Reis O, et al. (2012) The Fungal Pathogen *Moniliophthora perniciosa* Has Genes Similar to Plant PR-1 That Are Highly Expressed during Its Interaction with Cacao. PLoS ONE 7(9): e45929. doi:10.1371/journal.pone.0045929

Editor: Dee A. Carter, University of Sydney, Australia

Received March 21, 2012; **Accepted** August 27, 2012; **Published** September 20, 2012

Copyright: © 2012 Teixeira et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This study was supported by funds from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, 2006/53553-3, 2007/50262-0, 2009/51018-1 and 2009/50119-9). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: goncalo@unicamp.br

☯ These authors contributed equally to this work.

Introduction

The basidiomycete fungus *Moniliophthora perniciosa* is the causative agent of witches' broom disease (WBD) in cacao. This devastating disease is responsible for large losses in cacao plantations in the Americas and is a potential threat to other cacao-growing areas throughout the world [1,2]. *M. perniciosa* displays a hemibiotrophic lifestyle, with sequential biotrophic (infective) and necrotrophic stages in the plant. These two mycelial stages are morphologically distinct: whereas the biotrophic mycelium is monokaryotic, the necrotrophic stage is dikaryotic and presents clamp connections for nuclei transfer.

The disease cycle initiates when fungal basidiospores infect meristematic tissues of cacao – such as shoots, fruits and floral cushions – where they germinate and develop as biotrophic monokaryotic hyphae. *M. perniciosa* does not use any specialized infection structure to enter the plant (i.e. apressorium), as observed for the majority of biotrophic and hemibiotrophic fungi [3]. This fungus enters the host tissues through stomata or wounds and colonizes the plant apoplast as thick monokaryotic hyphae. In this stage of the disease, the parasitic fungus causes drastic morpho-physiological alterations in the host, resulting in the formation of hyperplastic and hypertrophic stems, known as green brooms. During the disease progression, the pathogen switches to its necrotrophic dikaryotic stage, which parallels the death of the infected plant tissue. In this phase of WBD, known as dry broom, *M. perniciosa* colonizes the dead plant and can be found in the intracellular spaces of cacao. After alternating wet and dry periods, the fungus produces basidiomata that release basidiospores, reinitiating the disease cycle [1,2].

During recent years, efforts have been directed to develop a solution to control this disease. In 2000, the WBD genome initiative (www.lge.ibi.unicamp.br/vassoura) was launched and, since then, it has supported several molecular and biochemical studies involving both the pathogen and the plant [4–9]. With the recent technological advances in the area of DNA sequencing, transcriptomes representing a variety of growth and developmental conditions of *M. perniciosa* – including transcriptomes of the fungus developing *in planta* – were sequenced using the RNA-seq technology. As a result, a comprehensive database named WBD

Transcriptome Atlas has been constructed, and has contributed important information on the molecular basis of the *M. perniciosa*-cacao interaction (Teixeira *et al.*, manuscript in preparation).

The establishment of a disease process depends on the ability of the pathogen to overcome or neutralize plant defenses and then initiate a parasitic relationship with its host. However, to halt pathogenic colonization, plants have developed an arsenal of defense responses, which include induction of pathogenesis-related (PR) genes [10], production of secondary metabolites as well as reinforcement of cell walls. Also, usually triggered by the recognition of a pathogen attack, plants produce highly toxic radicals, such as nitric oxide and reactive oxygen species, which can lead to the establishment of a local cell death (the hypersensitive response, HR). Among the induced pathogenesis-related genes, PR-1s have been frequently identified and used as markers of plant defense responses [10]. Notably, they were shown to have microbicide activity against oomycetes and fungi [11–14].

PR-1 proteins are members of a superfamily named SCP/TAPS (Sperm-Coating Protein/Tpx-1/Ag5/PR-1/Sc7) or CAP (Cysteine-rich secretory proteins, Antigen 5, and Pathogenesis-related 1). This superfamily has members throughout the eukaryotic kingdom, suggesting an important role for these proteins in the biology of eukaryotes [15,16]. Thus far, only a single report has shown the existence of enzymatic activity for a SCP/TAPS protein [17]. The protein Tex31 of the predatory marine mollusk *Conus textile* showed serine-proteolytic activity against a specific pro-peptide precursor of a venom toxin [17]. In addition, structural analyses indicated that four highly conserved amino acids (two histidines and two glutamates) form the putative catalytic site of SCP/TAPS proteins [16–22]. Although the existence of biochemical activity has not been shown for any other SCP/TAPS proteins, they are associated with various biological processes, such as male reproductive tract development [23,24], immune responses in plants and animals [25], venom activity of reptiles and insects [26–29] and host invasion by parasites [30–32].

In fungal species, SCP/TAPS proteins have been studied in *Saccharomyces cerevisiae*, in which they are highly expressed under nutrient starvation conditions [33]. In the basidiomycete *Schizophyllum commune*, SCP/TAPS proteins have been associated with fruiting body formation [34]. Interestingly, in the ascomycetes *Candida albicans* and *Fusarium oxysporum*, deletion of a SCP/TAPS gene impaired virulence on animals, indicating a role for this class of genes in fungal pathogenicity [35,36]. Considering that PR-1 proteins are widespread markers of the induced defense response in plants, what would be the function of their homologs in a plant pathogenic fungus, such as *M. perniciosa*?

This article describes the identification of a SCP/TAPS family in the *M. perniciosa* genome, the analysis of structural features of these genes, and their expression profile throughout *M. perniciosa* development. *M. perniciosa* SCP/TAPS proteins were modeled, and some structural differences were revealed among them. Based on these results, we present a hypothetical model in which SCP/TAPS proteins play a role in *M. perniciosa* pathogenicity by interfering with the defense system of cacao plants.

Results

Characterization of the *PR-1* gene family in *M. perniciosa*

Annotation of a genome draft of *M. perniciosa* [7], and inspection of fungal EST libraries [6,8] identified four *PR-1*-like genes in this pathogen (*MpPR-1a* to *MpPR-1d*). Later, improvements in the genome assembly obtained with next generation sequencing data (unpublished data) allowed the identification of seven additional members of the *MpPR-1* family (*MpPR-1e* to *MpPR-1k*), totaling eleven *PR-1*-like genes in *M. perniciosa*. These members are very heterogeneous in size and gene structure, with coding sequences (CDS) ranging from 447 to 1,152 nucleotides and intron composition varying between two to five introns. Three of these genes (*MpPR-1c*, *MpPR-1d* and *MpPR-1j*) are organized *in tandem*. Sequence details of these eleven genes are shown in Table 1, and their respective structures (exon-intron positions) are depicted in figure S1.

Hydrophobic signal peptide sequences predicted with the TargetP program [37] were identified in all eleven *MpPR-1* sequences (NN score >0.80), strongly suggesting that these proteins are secreted. Additionally, all MpPR-1 proteins showed a single SCP/TAPS domain (InterPro ID IPR014044), as predicted by the InterProScan server [38]. These domains were approximately 130 amino acids in length and ranged between 34% and 82% of the total amino acid sequence of an individual MpPR-1 protein (Fig. 1A). In addition to the SCP/TAPS domain, MpPR-1b and MpPR-1g present N-terminal and C-terminal extensions, respectively. No other InterProScan predicted domain was identified in these extensions or in any of the MpPR-1 proteins described. Interestingly, a careful manual inspection revealed that the C-terminal extension of MpPR-1g is rich in residues of lysine (K) and glutamic acid (E) that are mostly organized in alternating positions, resulting in the formation of a "KEKE" motif [39] (Fig. 2). The MpPR-1b N-terminal extension is also a low complexity region, being rich in serine, threonine and proline residues. However, no described motif could be recognized.

Alignment of the amino acid sequences encoded by all *MpPR-1* genes revealed significant similarity only over the SCP/TAPS domains (Fig. 1B). The amino acids proposed to form the putative catalytic site of SCP/TAPS proteins (two histidines and two glutamic acids, shown in red) were identified in six MpPR-1s (MpPR-1b, MpPR-1c, MpPR-1d, MpPR-1e, MpPR-1h and MpPR-1j). In contrast, MpPR-1a presented two, MpPR-1f had only one and MpPR-1g, MpPR-1i and MpPR-1k had none of the four conserved residues (Fig. 1B).

Protein structure modeling

To explore the tertiary structural characteristics within the SCP/TAPS domains of the MpPR-1 family members, we created homology models using the fold prediction metaserver I-TASSER (Fig. 3). The derived models indicated that the MpPR-1 SCP/TAPS domains adopt the α-β-α sandwich conformation, which is common to all superfamily members across the species studied [16,18] (Fig. 3A). Furthermore, all MpPR-1 proteins possess the large cleft proposed to constitute the SCP/TAPS active site. As mentioned above, six MpPR-1 proteins (MpPR-1b, MpPR-1c, MpPR-1d, MpPR-1e, MpPR-1h and MpPR-1j) have the four conserved residues of the putative catalytic site. These residues are localized within this cleft (Fig. 3B) and are found in the same direction of the orthologous amino acids identified in previous SCP/TAPS crystal structures [18,21,22,25]. These residues are lacking (MpPR-1g, MpPR-1i and MpPR-1k) or partially absent (MpPR-1a and MpPR-1f) in other members of the MpPR-1 family (Fig. 3C), indicating some diversification in their mode of action.

Remarkably, MpPR-1b and MpPR-1g contain two modules: the SCP/TAPS domain and either N-terminal (MpPR-1b) or C-terminal (MpPR-1g) extensions. Protein models indicate that such extensions are structurally organized and have α-helix conformations (Fig. 3A). These additional regions possibly confer a different activity or regulation to these proteins.

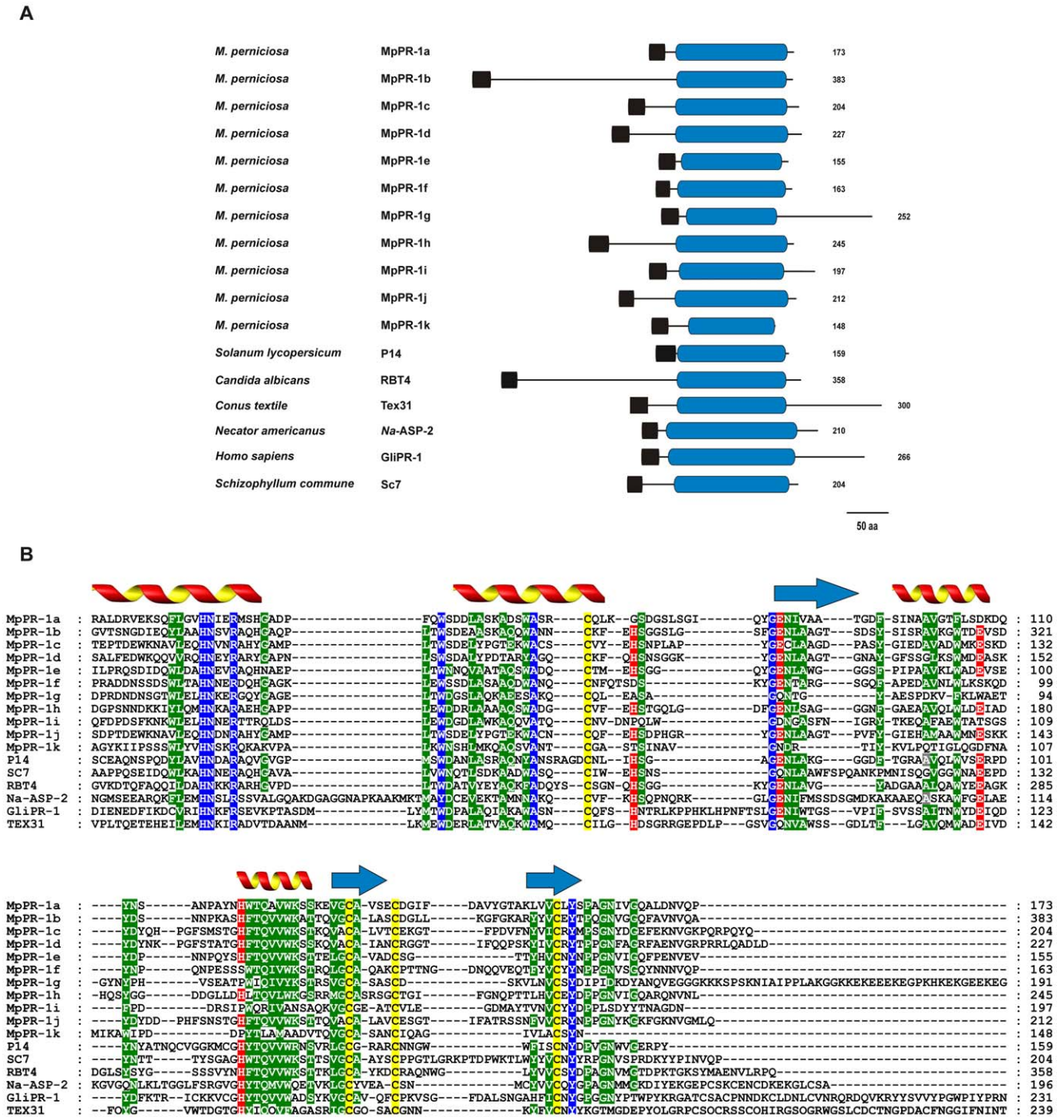

Figure 1. Comparison of MpPR-1 and SCP/TAPS proteins of representative organisms. (A) Domain arrangement of SCP/TAPS proteins. Hydrophobic signal peptides are shown in black and SCP/TAPS domains are represented in blue. The numbers on the right show the size of each protein. Large N-terminal and C-terminal expansions are observed in MpPR-1b and MpPR-1g, respectively. (B) Alignment of the conserved domain of SCP/TAPS proteins. In general, the SCP/TAPS superfamily members show similarities only over the SCP/TAPS domain. Conserved residues (100% of identity) are shown in blue and semi-conserved residues (at least 60% of identity) in green. Putative active site residues are highlighted in red and cysteines in yellow. Secondary structure elements are shown above the alignment (arrow: β-sheets; helix: α-helixes). P14, tomato PR-1 (GenBank P04284); RBT4, repressed by TUP1 from *Candida albicans* (GenBank AAG09789); Tex31, SCP/TAPS from the mollusk *Conus textile* (GenBank CAD36507); Na-ASP-2, *Necator americanus* secreted protein (GenBank AAP41952); GliPR-1, human glioma PR-1 protein (GenBank P48060); SC7, SCP/TAPS from the basidiomycete *Schizophyllum commune* (GenBank P35794).
doi:10.1371/journal.pone.0045929.g001

Table 1. Characteristics of the eleven *MpPR-1* genes identified in the *M. perniciosa* genome.

Gene name	GenBank accession number	CDS size (bp)	Number of introns	Protein size (aa)	Signal peptide	BlastX First Hit (Swissprot)*	BlastP First Hit (NCBI-NR)*
MpPR-1a	JN620340	522	2	173	Yes (NN score = 0.894)	SC14 (1e-19) *Schizophyllum commune*	XP_001876569.1- predicted protein (1e-54) *Laccaria bicolor*
MpPR-1b	JN620341	1152	5	383	Yes (NN score = 0.844)	SC7 (3e-25) *Saccharomyces cerevisiae*	XP_001873270.1- *predicted protein* (2e-63) *Laccaria bicolor*
MpPR-1c	JN620342	615	4	204	Yes (NN score = 0.898)	PRY1 (1e-28) *Saccharomyces cerevisiae*	XP_001889714.1- predicted protein (2e-48) *Laccaria bicolor*
MpPR-1d	JN620343	684	5	227	Yes (NN score = 0.951)	PRY1 (3e-31) *Saccharomyces cerevisiae*	XP_001889714.1- predicted protein (8e-66) *Laccaria bicolor*
MpPR-1e	JN620344	468	3	155	Yes (NN score = 0.919)	PR-1C (6e-35) *Nicotiana tabacum*	XP_003038868.1- hypothetical protein (2e-45) *Schizophyllum commune*
MpPR-1f	JN620345	492	3	163	Yes (NN score = 0.947)	SC7 (4e-25) *Schizophyllum commune*	CCA68148 – related to PRY1 (8e-42) *Piriformospora indica*
MpPR-1g	JN620346	759	4	252	Yes (NN score = 0.905)	SC7 (4e-13) *Schizophyllum commune*	EFY95292.1 – hypothetical protein (4e-15) *Metarhizium anisopliae*
MpPR-1h	JN620347	738	5	245	Yes (NN score = 0.804)	PR-1B (5e-24) *Nicotiana tabacum*	XP_001828886.1- hypothetical protein (3e-41) *Coprinopsis cinerea*
MpPR-1i	JN620348	498	3	165	Yes (NN score = 0.940)	SC7 (4e-05) *Schizophyllum commune*	EGO02028.1- hypothetical protein (3e-10) *Serpula lacrymans*
MpPR-1j	JN620349	639	4	212	Yes (NN score = 0.918)	PRY3 (8e-30) *Saccharomyces cerevisiae*	XP_001889714.1- predicted protein (2e-54) *Laccaria bicolor*
MpPR-1k	JN620350	447	2	148	Yes (NN score = 0.944)	P14 (2e-03) *Solanum lycopersicum*	XP_002578075.1- venom allergen 21 (4e-09) *Schistosoma mansoni*

*e-values are shown in parentheses.

doi:10.1371/journal.pone.0045929.t001

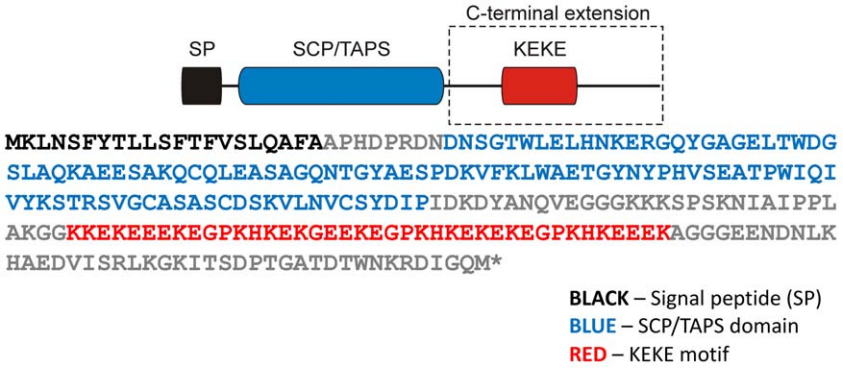

Figure 2. Domains identified in the MpPR-1g protein. In addition to the SCP/TAPS domain, this protein has a KEKE motif in its C-terminal extension. This motif is known to mediate the interaction with other proteins or ions.
doi:10.1371/journal.pone.0045929.g002

Expression profile of *MpPR-1* genes throughout *M. perniciosa* development

The expression pattern of *MpPR-1* family members was assessed by quantitative real time PCR (qPCR) throughout the different developmental stages of *M. perniciosa*. As shown in Figure 4, each *MpPR-1* gene showed a specific expression profile: expression of *MpPR-1a*, *MpPR-1b*, *MpPR-1d* and *MpPR-1i* genes was relatively uniform, with nearly similar expression values for most conditions analyzed. In contrast, *MpPR-1e* was up-regulated in the dikaryotic hyphae, whereas *MpPR-1j* was exclusively expressed in mushrooms (basidiomata). Strikingly, five *MpPR-1* genes (*MpPR-1c*, *MpPR-1f*, *MpPR-1g*, *MpPR-1h*, and *MpPR-1k*) were highly expressed during the biotrophic interaction of the fungus with the cacao plant (green broom stage of WBD). These genes were poorly expressed in the necrotrophic hyphae (dry broom stage of

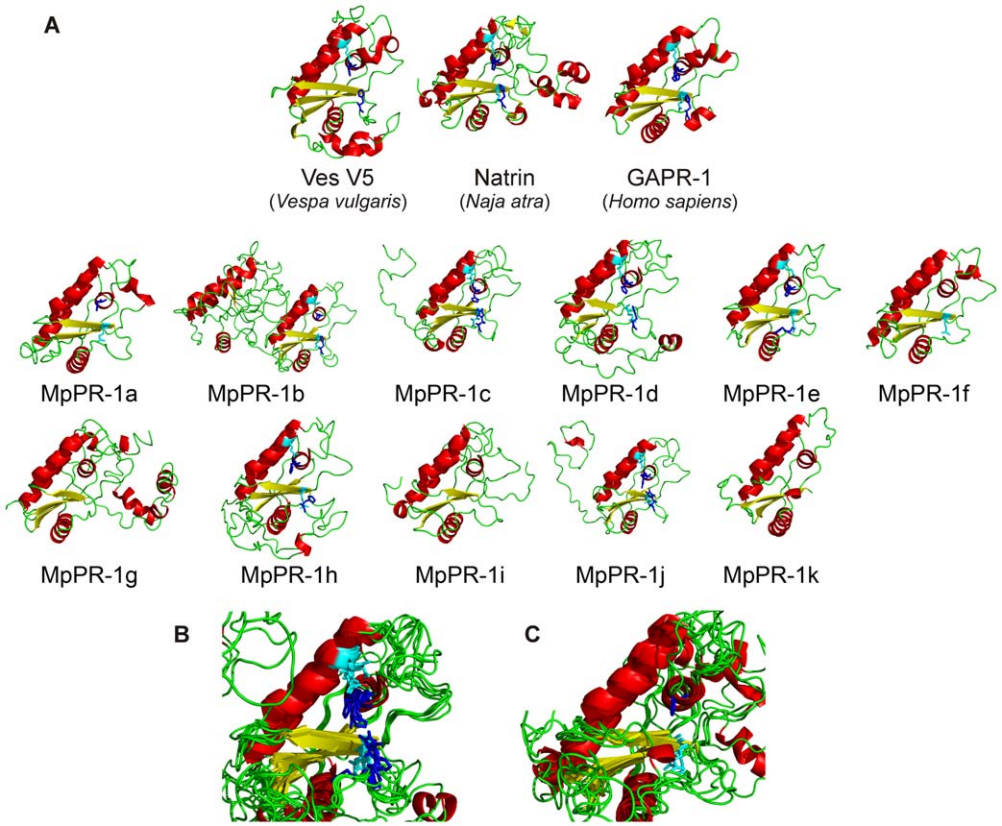

Figure 3. Homology modeling of MpPR-1 proteins. (A) Ribbon stick representation showing the folding of eleven MpPR-1 proteins and three SCP/TAPS proteins used to obtain these models. The putative residues forming the catalytic site are highlighted in dark blue (histidines) and light blue (glutamic acids). Note the presence of an additional protein module in MpPR-1b and MpPR-1g. These modules respectively correspond to the N-terminal and C-terminal extensions observed in these proteins. (B) MpPR-1b, MpPR-1c, MpPR-1d, MpPR-1e, MpPR-1h and MpPR-1j have the four putative active site residues of the SCP/TAPS domain. (C) These residues are partially or completely absent in MpPR-1a, MpPR-1f, MpPR-1g, MpPR-1i and MpPR-1k.
doi:10.1371/journal.pone.0045929.g003

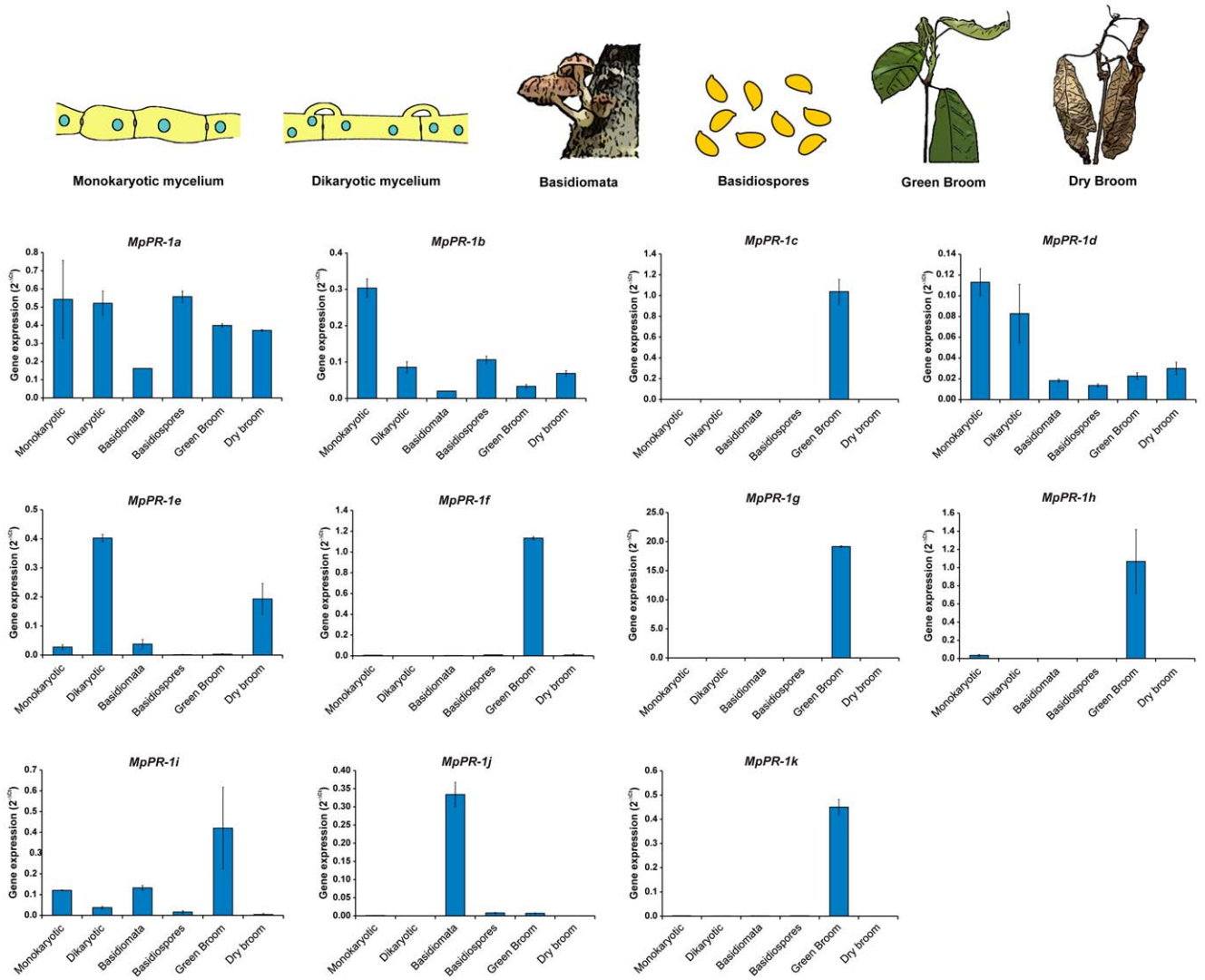

Figure 4. Transcriptional profile of *MpPR-1* family members throughout the *M. perniciosa* life cycle. Each *MpPR-1* gene has a distinct expression profile during fungal development. "Monokaryotic" and "Dikaryotic" hyphae represent the two mycelial stages (biotrophic and necrotrophic) grown under *in vitro* conditions. "Green broom" and "dry broom" correspond to the biotrophic and necrotrophic stages of *M. perniciosa*, respectively, during its interaction with cacao. Analyses were performed by qPCR and the *M. perniciosa β-actin* gene was used as endogenous control to normalize data. Error bars represent standard deviations determined with two biological replicates. Representative drawings of the conditions analyzed are shown on the top.
doi:10.1371/journal.pone.0045929.g004

WBD) and in the *ex planta* conditions, suggesting a specific role for the encoded proteins in fungal pathogenicity.

Characterization of an *MpPR-1* cluster

As mentioned above, genes *MpPR-1c*, *MpPR-1d* and *MpPR-1j* are arranged *in tandem* over a region of approximately 5 kbp in the *M. perniciosa* genome (Fig. 5). This gene cluster points to the occurrence of gene duplication events during the evolution of the *MpPR-1* family. Indeed, the proteins encoded by these three genes are more similar to each other than to other MpPR-1 members (data not shown). In accordance, these three genes have very similar structures, with a minor difference in *MpPR-1d*, which has an additional intron and a mini-exon following exon 2 (Fig. S1). Importantly, genes *MpPR-1c* and *MpPR-1j* are also highly similar at the nucleotide level (84% identity), indicating a recent event of gene duplication. Despite their similarity, these genes have distinct expression profiles: whereas *MpPR-1j* is highly expressed in basidiomata, *MpPR-1c* is mainly expressed during cacao infection (green broom stage of WBD) (Figs. 4 and 5).

Discussion

In this study, we identified a family of genes encoding proteins of the SCP/TAPS superfamily in the plant pathogen *Moniliophthora perniciosa*. SCP/TAPS proteins are found in a vast number of organisms, including plants, insects, mammals, fungi, mollusks and worms. Plant Pathogenesis-related proteins (PR-1) belong to this superfamily and are known to accumulate after pathogen invasion [10]. Despite being broadly spread, evidence for the importance of fungal SCP/TAPSs in plant-pathogen interactions has not yet been described.

M. perniciosa has a larger number (11) of *SCP/TAPS* genes encoding PR-1-like secreted proteins than other fungal species analyzed to date (Table S1). Although these proteins have a single

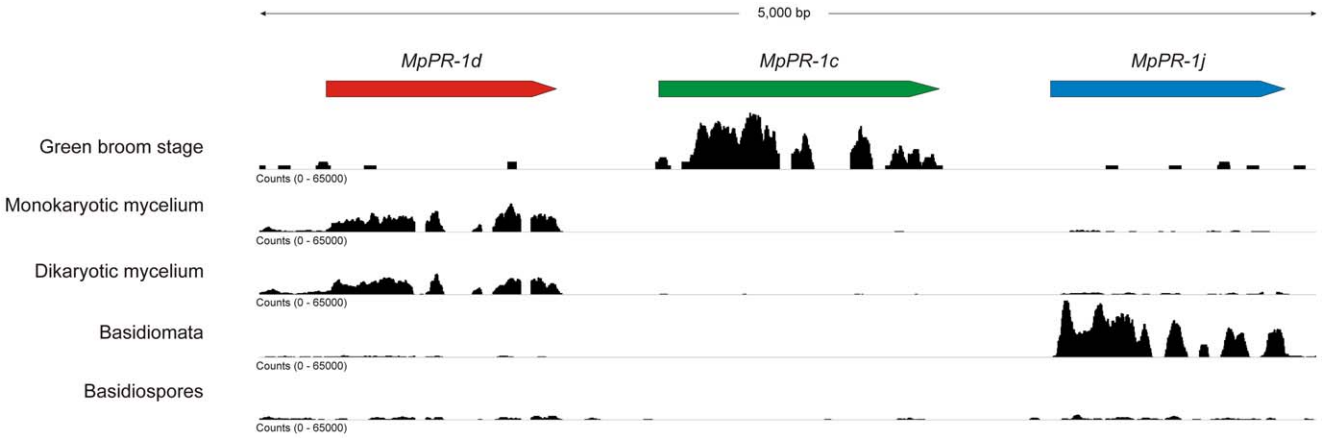

Figure 5. Genomic organization and transcriptional profile of the *MpPR-1* gene cluster found in *M. perniciosa*. The *MpPR-1c*, *MpPR-1d* and *MpPR-1j* genes are arranged *in tandem* over a region of approximately 5 kbp. Analysis of the WBD RNA-seq Atlas shows the expression profile of these *MpPR-1* genes in different conditions (green broom – *in planta* development of the biotrophic monokaryotic hyphae; monokaryotic mycelium; dikaryotic mycelium; basidiomata and basidiospores). Data were visualized using the Integrative Genomics Viewer [62]. The black coverage plot shows cumulative RNA-seq read coverage along the transcripts in all different conditions. Note that these genes were named according to the order they were identified in the fungal genome, and the nomenclature does not necessarily reflect their relative localization in the genome.
doi:10.1371/journal.pone.0045929.g005

SCP/TAPS domain, they are very divergent in sequence (Fig. 1B). Some of them show extensions in their N-terminal (MpPR-1b) or C-terminal (MpPR-1g) regions (Fig. 1A and Fig. 3A). Also, six MpPR1s (MpPR-1b, MpPR-1c, MpPR-1d, MpPR-1e, MpPR-1h and MpPR-1j) have all four conserved amino acids (two histidines and two glutamates) of the active site proposed for SCP/TAPS proteins, whereas the other five proteins (MpPR-1a, MpPR-1f, MpPR-1g, MpPR-1i and MpPR-1k) lack the catalytic tetrad (Fig. 1B and Fig. 3). However, the basic signatures of SCP/TAPS domains remain conserved.

The high number of genes and the variation in protein sequence of MpPR-1 members may reflect a process of diversification of this gene family in *M. perniciosa*. In accordance, the cluster including three *MpPR-1* members (Fig. 5) indicates the occurrence of recent events of gene duplication, which are central for the generation of genetic variability. In this regard, functional diversity may occur within this family, and different MpPR-1s likely play a particular role in *M. perniciosa* biology.

Although widely distributed among several species, the functions of SCP/TAPS proteins are still uncertain. Their broad distribution indicates that they play a role in a plethora of biological processes. In the mushroom-forming basidiomycete *Schizophyllum commune*, SCP/TAPS proteins were identified in the basidiomata, being involved in the formation of pseudo-parenchymous tissue of this reproductive structure [34]. In a similar way, *MpPR-1j* is exclusively expressed in the basidiomata of *M. perniciosa*, suggesting a role for this isoform in the physiology/metabolism of fruiting bodies. In contrast, we also identified *SCP/TAPS* genes in the genomes of some basidiomycetes that do not produce mushrooms (e.g., *Ustilago maydis*, *Puccinia graminis*, *Melampsora spp* and *Cryptococcus spp*) (data not shown). Therefore, it is likely that SCP/TAPS proteins have functions in basidiomycetes other than fruiting body development/physiology.

In response to pathogen invasion, plants typically produce PR-1 proteins, which have antimicrobial activity and, consequently, are able to inhibit the development of fungi and oomycetes [11–13]. Considering that, we hypothesize that some MpPR-1s could play a role in limiting the growth of other microbial competitors (e.g., oomycetes from the genus *Phytophthora*, responsible for causing black pod rot in cacao), thus favoring *M. perniciosa* colonization during WBD progression. Functional experiments are needed to confirm the existence of antimicrobial activity in any of the MpPR-1 proteins identified.

Gene expression analyses revealed that five *MpPR-1* genes (*MpPR-1c*, *MpPR-1f*, *MpPR-1g*, *MpPR-1h* and *MpPR-1k*) are strikingly up-regulated in the green broom stage of WBD, when the fungus grows biotrophically within the plant tissues (Fig. 4). Remarkably, inspection of the WBD RNA-seq Transcriptome Atlas (Teixeira *et al.*, manuscript in preparation) revealed that *MpPR-1g* and *MpPR-1h* are among the most highly expressed genes of *M. perniciosa* during its biotrophic interaction with cacao. Moreover, in addition to the green broom stage, *MpPR-1f* and *MpPR-1h* are notably expressed in germinating basidiospores (Fig. S2), a critical stage for the establishment of infection. None of these genes were significantly expressed in non-germinating spores or in the dry broom stage of WBD (Fig. 4 and Fig. S2), strongly indicating a major role for the encoded proteins in the infective (biotrophic) stage of *M. perniciosa*. Similarly, *SCP/TAPS* genes identified in some animal parasitic worms (*Schistosoma mansoni*, *Brugia malayi*, *Necator americanus* and *Ancylostoma caninum*) are highly expressed in the infective stage and are considered important pathogenicity factors [30,31,40–42]. In these parasites, SCP/TAPSs are supposed to contribute to their virulence by modulating the host immune response [15,16].

Recently, a SCP/TAPS protein in the plant-parasitic nematode *Globodera rostochiensis* (Gr-VAP1) was shown to function as an effector by interacting with the tomato cysteine protease Rcr3, which is also a target of the Avr2 effector from the fungus *Cladosporium fulvum* [43]. In addition, SCP/TAPS proteins have been identified in other plant infecting nematodes (e.g. *Heterodera glycines*, *Meloidogyne incognita* and *Bursaphelenchus xylophilus*), and these are thought to be required for the establishment of parasitism [44–48]. Considering the expression pattern of some *SCP/TAPS* genes in *M. perniciosa* and the functions ascribed for the encoded proteins in other pathogenic organisms, it is plausible that some MpPR-1s play a role in the *M. perniciosa*-cacao interaction and may be candidate effectors of this fungal pathogen. Accordingly, a recent study that aimed at the identification of putative effectors in *Melampsora larici-populina* and *Puccinia graminis* reported the enrichment of SCP/TAPSs in the predicted secretome of these rust fungi

[49]. A fungal *SCP/TAPS* gene was also identified in EST libraries produced from rye infected with the ascomycete pathogen *Claviceps purpurea* [50], suggesting that these genes might also be important in other plant-fungus interactions.

In recent years, a role for SCP/TAPSs as virulence factors has emerged in many organisms. It is likely that these proteins converged as pathogenicity mechanisms in distinct pathogens/parasites from either plants or animals. Whereas the function of fungal SCP/TAPSs as virulence factors in plant pathogens remains to be confirmed, previous studies demonstrated that these proteins are required for fungal virulence on animals (e.g. *C. albicans* and *F. oxysporum*) [35,36]. The ascomycete *F. oxysporum* is a multi-host pathogen that is able to infect both plants and animals. Previous work by Prados-Rosales *et al.* verified that *fpr1*, one of the six *SCP/TAPS* genes from this pathogen, is required for fungal virulence on animals but not on plants [36]. Given that the *F. oxysporum* genome contains five other *SCP/TAPS* genes, the absence of a phenotype on plants can be explained by the occurrence of functional redundancy in this gene family. Notably, there is evidence that *fpr1* is part of a gene family that has expanded in *F. oxysporum* and in other plant pathogenic Sordariomycetes [36].

Although the precise activity of SCP/TAPSs is currently unknown, Prados-Rosales *et al.* presented the first genetic evidence for a biological function of the proposed active site of SCP/TAPS proteins [36]. The authors demonstrated that the integrity of the active site is required for *F. oxysporum* virulence on animals. In *M. perniciosa*, six MpPR-1s (MpPR-1b, MpPR-1c, MpPR-1d, MpPR-1e, MpPR-1h and MpPR-1j) contain all four amino acids of the proposed active site. In contrast, the other five proteins do not have the complete catalytic tetrad (Fig. 1B). In this regard, it is possible that *M. perniciosa* PR-1s have distinct mode of actions. For instance, whereas those proteins with the complete catalytic tetrad can function as enzymes, the other PR-1s may act as inhibitors.

Concomitantly to the up-regulation of some *MpPR-1* genes *in planta*, we identified a cacao *PR-1* gene over-expressed in the green broom stage of WBD (Fig. S3). Based on these findings, we suggest that some MpPR-1s could act as competitive inhibitors of the plant PR-1, modulating the cacao immune response. It has already been shown that the SCP/TAPS protein *Na*-ASP-2 of the hookworm *Necator americanus* has a high structural similarity to chemokines, and this protein is proposed to be an antagonistic ligand of receptors that activate the immune system of the vertebrate host [51]. Furthermore, NIF (Neutrophil inhibitory factor), a SCP/TAPS protein from *Ancylostoma caninum*, interferes with the host immune system by interacting with neutrophil receptors [32]. Confirmation of this interesting mechanism in the *M. perniciosa*-cacao interaction may be of primary relevance to the understanding of many other plant diseases and will shed light on our understanding of PR-1 functions.

Among the *MpPR-1* genes that are highly expressed *in planta*, *MpPR-1g* is the only one with a C-terminal extension in addition to the SCP/TAPS domain (Fig. 1A and Fig. 3A). This additional region is rich in lysine (K) and glutamic acid (E) residues, which are mostly organized in alternating positions, resulting in the formation of a "KEKE" motif [39] (Fig. 2). This motif is known to mediate protein-protein associations [39,52] and is also able to bind divalent ions, such as calcium and zinc [53]. Calcium is an important regulator of many cellular processes, including plant defense responses [54]. In this regard, this additional module may be important in determining the mode of action of MpPR-1g. Whether this protein interacts with other proteins, particularly cacao proteins, and/or interferes with the plant calcium signaling during infection should be the object of future studies.

Overall, this study presents important evidence on the role of fungal SCP/TAPSs in the context of a plant-pathogen interaction. Although the precise function of each MpPR-1 family member is currently unknown, the information provided in our study suggests they have potential roles in some important biological processes, such as fruiting body metabolism, spore penetration and modulation of the host defense response. As a consequence, our results may inform the study of the role of PR-1-encoding genes in other organisms, particularly phytopathogens. Further studies concerning the *M. perniciosa PR-1* gene family will focus on the characterization of this interesting family in terms of fungal development and roles in the *M. perniciosa* interaction with cacao.

Materials and Methods

Biological material

Isolate CP02 of *Moniliophthora perniciosa* (Stahel) Aime & Philliphs-Mora [55], was used to perform the experiments. Under *in vitro* conditions, the fungus can only be maintained as a dikaryotic mycelium, and all other developmental stages (basidiomata, basidiospores and monokaryotic mycelium) are obtained from the dikaryotic stage. The reproductive structures (basidiomata) were produced in laboratory according to the protocol described by Pires *et al.* [8]. Fresh basidiomata were used to collect basidiospores according to Frias *et al.* [56].

Basidiospores suspensions were utilized for the *in vitro* production of the monokaryotic mycelia. For this purpose, approximately 3.75×10^5 basidiospores were inoculated in 125 ml Erlenmeyer flasks containing 50 ml liquid medium (LMCpL+), as described by Meinhardt *et al.* [57]. Liquid cultures were maintained at 28°C and incubated under agitation at 120 rpm. Dikaryotic mycelium was inoculated in the same medium and maintained under the same conditions. Both mycelia were collected 7 days post inoculation to perform the experiments.

Theobroma cacao L. cv. "Comum" was used to perform the infection experiments. Three-months-old plantlets were inoculated with 30 µL of a basidiospore suspension (1×10^5 spores mL^{-1}) according to the procedure described by Frias *et al.* [56]. Plantlets were kept in a greenhouse under controlled conditions of temperature (26°C) and humidity (>80%). Green brooms (biotrophic stage) and dry brooms (necrotrophic stage) were collected 30 and 105 days post inoculation, respectively.

Sequence analysis

Inspection of the *M. perniciosa* genome led to the identification of eleven genes encoding proteins similar to plant pathogenesis-related proteins 1 (PR-1). These genes were named *MpPR-1a* to *MpPR-1k* according to the order they were discovered. The complete open reading frames (ORFs) of these genes were predicted using the program Augustus [58] and confirmed by cDNA sequencing. These sequences have been submitted to GenBank with the accession numbers JN620340 to JN620350. Blast searches were performed using the NCBI-NR and Swissprot databases. Domain prediction of the encoded proteins was performed using the InterProScan server [38] and the presence of a signal peptide for secretion was predicted using the software TargetP 1.1 [37].

Total RNA extraction and cDNA synthesis

With the exception of basidiospores, samples were ground to a fine powder in liquid nitrogen using a pestle and mortar. Basidiospores walls were broken by vortexing the sample in RNA extraction buffer (Buffer RLT, RNeasy Plant Mini Kit) and 200 mg glass beads (0.4–0.6 µm, Sigma-Aldrich, St. Louis, MO,

EUA). RNA isolation was performed using the RNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA) according to the manufacturer's instructions. RNA was treated with DNAse I AmpGrade (Invitrogen, Carlsbad, CA, USA) and its concentration was accessed using the ND-1000 spectrophotometer (NanoDrop, Wilmington, DE, USA). cDNA was synthesized from 1 µg total RNA using the SuperScript II Reverse Transcriptase (Invitrogen), according to the manufacturer's instructions.

Gene expression assays

Quantitative real time PCR (qPCR) was performed on a StepOne Plus Real Time PCR System (Applied Biosystems, Foster City, CA, USA) using Sybr Green I dye for the detection of PCR products. Each reaction contained 8 µl SYBR Green PCR Master Mix (Applied Biosystems), 250 nM each primer and 50 ng cDNA template in a final volume of 16 µl. No-template reactions were included as negative controls for each set of primers used. The thermal cycling conditions were 94°C for 10 min, followed by 40 cycles of 94°C for 15 s, 53°C for 30 s and 60°C for 1 min, with fluorescence detection at the end of each cycle. In addition, a melting curve analysis was performed to verify the amplification of a single product per reaction. All reactions were conducted in technical triplicates using two independent biological replicates of each sample. The *M. perniciosa β-actin* gene was used to normalize data and expression levels are presented as $2^{-\Delta Ct}$. Primers used in this assay are shown in Table S2.

Protein structure modeling

The fold recognition-based method was implemented using the I-TASSER server [59], which constructed structure models for each MpPR-1 protein using folds of the most similar proteins deposited in the PDB (Protein Data Bank) database (http://www.rcsb.org/pdb). The main templates were based on the structure of three proteins: i) Natrin (PDB – 1xta), a component of the venom of the snake *Naja atras*; ii) GAPR-1 (PDB – 1smb), a SCP/TAPS protein associated with the membrane of the human Golgi system; and iii) Ves V5 (PDB – 1qnx), present in the venom of the wasp *Vespula vulgaris*. The modeled structures were validated by analyzing the Ramachandran plots generated by PROCHECK [60], and the models were displayed using the software PyMOL [61].

Supporting Information

Figure S1 Structure of the *MpPR-1* genes. Exons are represented by boxes, while introns are shown as lines. Exons are colored to highlight the regions encoding important protein features: predicted signal peptides (black), SCP/TAPS domain (blue) and the remaining ORF (gray).
(TIF)

Figure S2 Expression levels of *MpPR-1* genes in germinating and non-germinating basidiospores. *MpPR-1f* and *MpPR-1h* are highly expressed in germinating basidiospores, supporting a role for the encoded proteins in the establishment of witches' broom disease. Data are part of the WBD Transcriptome Atlas and were obtained by RNA-seq sequencing. Gene expression values are given in Reads Per Kilobase of exon model per Million mapped reads (RPKM).
(TIF)

Figure S3 Gene expression levels of a cacao *PR-1* (ID CGD0027635) in infected and healthy plants. Similar to some *MpPR-1* genes, a cacao *PR-1* (*TcPR-1*) is up-regulated in the green broom stage of WBD. The analysis was performed by qPCR and the *T. cacao α-tubulin* gene (ID CGD0029727) was used as endogenous control to normalize data. Gene IDs refer to the Cacao Genome Database (http://www.cacaogenomedb.org). The qPCR assay was conducted as described in the Material and Methods section and primers used in the experiment were: TcPR-1_F: 5′ ACCTTATGGCGAGAACCTTG 3′, TcPR-1_R: 5′ GGAGTAATCATAGTCGGCCTTC 3′, TcTub_F: 5′ ACCAATCTTAACCGCCTTGTCT 3′ and TcTub_R: 5′ GTTAGTCTGGAACTCAGTCACAT 3′.
(TIF)

Table S1 Number of *SCP/TAPS* genes in fungal species with different lifestyles. Numbers correspond to the genes coding proteins with the InterPro ID IPR014044.
(DOC)

Table S2 Primers used for quantitative real time PCR analyses of *M. perniciosa* PR-1 genes.
(DOC)

Acknowledgments

We thank Dr. Halley Caixeta de Oliveira (State University of Campinas) for critical reading of the manuscript.

Author Contributions

Conceived and designed the experiments: PJPLT JMCM. Performed the experiments: PJPLT DPTT ROV PFVP OR RMB SFF JMCM . Analyzed the data: PJPLT DPTT ROV JMCM. Contributed reagents/materials/analysis tools: PM. Wrote the paper: PJPLT DPTT JMCM GAGP.

References

1. Meinhardt LW, Rincones J, Bailey BA, Aime MC, Griffith GW, et al. (2008) *Moniliophthora perniciosa*, the causal agent of witches' broom disease of cacao: what's new from this old foe? Mol Plant Pathol 9: 577–588.
2. Purdy LH, Schmidt RA (1996) STATUS OF CACAO WITCHES' BROOM: biology, epidemiology, and management. Annu Rev Phytopathol 34: 573–594.
3. Perfect SE, Green JR (2001) Infection structures of biotrophic and hemibiotrophic fungal plant pathogens. Mol Plant Pathol 2: 101–108.
4. Scarpari LM, Meinhardt LW, Mazzafera P, Pomella AW, Schiavinato MA, et al. (2005) Biochemical changes during the development of witches' broom: the most important disease of cocoa in Brazil caused by *Crinipellis perniciosa*. J Exp Bot 56: 865–877.
5. Garcia O, Macedo JA, Tiburcio R, Zaparoli G, Rincones J, et al. (2007) Characterization of necrosis and ethylene-inducing proteins (NEP) in the basidiomycete *Moniliophthora perniciosa*, the causal agent of witches' broom in *Theobroma cacao*. Mycol Res 111: 443–455.
6. Rincones J, Scarpari LM, Carazzolle MF, Mondego JM, Formighieri EF, et al. (2008) Differential gene expression between the biotrophic-like and saprotrophic mycelia of the witches' broom pathogen *Moniliophthora perniciosa*. Mol Plant Microbe Interact 21: 891–908.
7. Mondego JM, Carazzolle MF, Costa GG, Formighieri EF, Parizzi LP, et al. (2008) A genome survey of *Moniliophthora perniciosa* gives new insights into Witches' Broom Disease of cacao. BMC Genomics 9: 548.
8. Pires AB, Gramacho KP, Silva DC, Goes-Neto A, Silva MM, et al. (2009) Early development of *Moniliophthora perniciosa* basidiomata and developmentally regulated genes. BMC Microbiol 9: 158.
9. Thomazella DP, Teixeira PJ, Oliveira HC, Saviani EE, Rincones J, et al. (2012) The hemibiotrophic cacao pathogen *Moniliophthora perniciosa* depends on a mitochondrial alternative oxidase for biotrophic development. New Phytol 194: 1025–1034.
10. Van Loon LC, Rep M, Pieterse CM (2006) Significance of inducible defense-related proteins in infected plants. Annu Rev Phytopathol 44: 135–162.
11. Rauscher M, Adam AL, Wirtz S, Guggenheim R, Mendgen K, et al. (1999) PR-1 protein inhibits the differentiation of rust infection hyphae in leaves of acquired resistant broad bean. Plant J 19: 625–633.
12. Niderman T, Genetet I, Bruyere T, Gees R, Stintzi A, et al. (1995) Pathogenesis-related PR-1 proteins are antifungal. Isolation and characterization of three 14-kilodalton proteins of tomato and of a basic PR-1 of tobacco with inhibitory activity against *Phytophthora infestans*. Plant Physiol 108: 17–27.

13. Kiba A, Nishihara M, Nakatsuka T, Yamamura S (2007) Pathogenesis-related protein 1 homologue is an antifungal protein in *Wasabia japonica* leaves and confers resistance to *Botrytis cinerea* in transgenic tobacco. PLANT BIOTECHNOLOGY 24: 247–254.
14. Zhu F, Xu M, Wang S, Jia S, Zhang P, et al. (2012) Prokaryotic expression of pathogenesis related protein 1 gene from *Nicotiana benthamiana*: antifungal activity and preparation of its polyclonal antibody. Biotechnol Lett .
15. Cantacessi C, Campbell BE, Visser A, Geldhof P, Nolan MJ, et al. (2009) A portrait of the "SCP/TAPS" proteins of eukaryotes–developing a framework for fundamental research and biotechnological outcomes. Biotechnol Adv 27: 376–388.
16. Gibbs GM, Roelants K, O'Bryan MK (2008) The CAP superfamily: cysteine-rich secretory proteins, antigen 5, and pathogenesis-related 1 proteins–roles in reproduction, cancer, and immune defense. Endocr Rev 29: 865–897.
17. Milne TJ, Abbenante G, Tyndall JD, Halliday J, Lewis RJ (2003) Isolation and characterization of a cone snail protease with homology to CRISP proteins of the pathogenesis-related protein superfamily. J Biol Chem 278: 31105–31110.
18. Fernandez C, Szyperski T, Bruyere T, Ramage P, Mosinger E, et al. (1997) NMR solution structure of the pathogenesis-related protein P14a. J Mol Biol 266: 576–593.
19. Henriksen A, King TP, Mirza O, Monsalve RI, Meno K, et al. (2001) Major venom allergen of yellow jackets, Ves v 5: structural characterization of a pathogenesis-related protein superfamily. Proteins 45: 438–448.
20. Serrano RL, Kuhn A, Hendricks A, Helms JB, Sinning I, et al. (2004) Structural analysis of the human Golgi-associated plant pathogenesis related protein GAPR-1 implicates dimerization as a regulatory mechanism. J Mol Biol 339: 173–183.
21. Guo M, Teng M, Niu L, Liu Q, Huang Q, et al. (2005) Crystal structure of the cysteine-rich secretory protein stecrisp reveals that the cysteine-rich domain has a K^+ channel inhibitor-like fold. J Biol Chem 280: 12405–12412.
22. Shikamoto Y, Suto K, Yamazaki Y, Morita T, Mizuno H (2005) Crystal structure of a CRISP family $Ca2^+$ -channel blocker derived from snake venom. J Mol Biol 350: 735–743.
23. Mizuki N, Sarapata DE, Garcia-Sanz JA, Kasahara M (1992) The mouse male germ cell-specific gene Tpx-1: molecular structure, mode of expression in spermatogenesis, and sequence similarity to two non-mammalian genes. Mamm Genome 3: 274–280.
24. Gibbs GM, Orta G, Reddy T, Koppers AJ, Martinez-Lopez P, et al. (2011) Cysteine-rich secretory protein 4 is an inhibitor of transient receptor potential M8 with a role in establishing sperm function. Proc Natl Acad Sci U S A 108: 7034–7039.
25. Szyperski T, Fernandez C, Mumenthaler C, Wuthrich K (1998) Structure comparison of human glioma pathogenesis-related protein GliPR and the plant pathogenesis-related protein P14a indicates a functional link between the human immune system and a plant defense system. Proc Natl Acad Sci U S A 95: 2262–2266.
26. Lu G, Villalba M, Coscia MR, Hoffman DR, King TP (1993) Sequence analysis and antigenic cross-reactivity of a venom allergen, antigen 5, from hornets, wasps, and yellow jackets. J Immunol 150: 2823–2830.
27. King TP, Spangfort MD (2000) Structure and biology of stinging insect venom allergens. Int Arch Allergy Immunol 123: 99–106.
28. Mochca-Morales J, Martin BM, Possani LD (1990) Isolation and characterization of helothermine, a novel toxin from *Heloderma horridum horridum* (Mexican beaded lizard) venom. Toxicon 28: 299–309.
29. Wang J, Shen B, Guo M, Lou X, Duan Y, et al. (2005) Blocking effect and crystal structure of natrin toxin, a cysteine-rich secretory protein from *Naja atra* venom that targets the BKCa channel. Biochemistry 44: 10145–10152.
30. Hawdon JM, Jones BF, Hoffman DR, Hotez PJ (1996) Cloning and characterization of Ancylostoma-secreted protein. A novel protein associated with the transition to parasitism by infective hookworm larvae. J Biol Chem 271: 6672–6678.
31. Chalmers IW, McArdle AJ, Coulson RM, Wagner MA, Schmid R, et al. (2008) Developmentally regulated expression, alternative splicing and distinct subgroupings in members of the *Schistosoma mansoni* venom allergen-like (SmVAL) gene family. BMC Genomics 9: 89.
32. Moyle M, Foster DL, McGrath DE, Brown SM, Laroche Y, et al. (1994) A hookworm glycoprotein that inhibits neutrophil function is a ligand of the integrin CD11b/CD18. J Biol Chem 269: 10008–10015.
33. Zúñiga S, Boskovic J, Garcia-Cantalejo JM, Jim nA, Ballesta JP, et al. (1999) Deletion of 24 open reading frames from chromosome XI from *Saccharomyces cerevisiae* and phenotypic analysis of the deletants. Gene 233: 141–150.
34. Schuren FH, Asgeirsdottir SA, Kothe EM, Scheer JM, Wessels JG (1993) The Sc7/Sc14 gene family of *Schizophyllum commune* codes for extracellular proteins specifically expressed during fruit-body formation. J Gen Microbiol 139: 2083–2090.
35. Braun BR, Head WS, Wang MX, Johnson AD (2000) Identification and characterization of TUP1-regulated genes in *Candida albicans*. Genetics 156: 31–44.
36. Prados-Rosales RC, Roldan-Rodriguez R, Serena C, Lopez-Berges MS, Guarro J, et al. (2012) A PR-1-like Protein of *Fusarium oxysporum* Functions in Virulence on Mammalian Hosts. J Biol Chem 287: 21970–21979.
37. Emanuelsson O, Brunak S, von HG, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2: 953–971.
38. Zdobnov EM, Apweiler R (2001) InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17: 847–848.
39. Realini C, Rogers SW, Rechsteiner M (1994) KEKE motifs. Proposed roles in protein-protein association and presentation of peptides by MHC class I receptors. FEBS Lett 348: 109–113.
40. Murray J, Gregory WF, Gomez-Escobar N, Atmadja AK, Maizels RM (2001) Expression and immune recognition of *Brugia malayi* VAL-1, a homologue of vespid venom allergens and Ancylostoma secreted proteins. Mol Biochem Parasitol 118: 89–96.
41. Cantacessi C, Hofmann A, Young ND, Broder U, Hall RS, et al. (2012) Insights into SCP/TAPS Proteins of Liver Flukes Based on Large-Scale Bioinformatic Analyses of Sequence Datasets. PLoS One 7: e31164.
42. Del Valle A, Jones BF, Harrison LM, Chadderdon RC, Cappello M (2003) Isolation and molecular cloning of a secreted hookworm platelet inhibitor from adult *Ancylostoma caninum*. Mol Biochem Parasitol 129: 167–177.
43. Lozano-Torres JL, Wilbers RH, Gawronski P, Boshoven JC, Finkers-Tomczak A, et al. (2012) Dual disease resistance mediated by the immune receptor Cf-2 in tomato requires a common virulence target of a fungus and a nematode. Proc Natl Acad Sci U S A 109: 10119–10124.
44. Gao B, Allen R, Maier T, Davis EL, Baum TJ, et al. (2001) Molecular characterisation and expression of two venom allergen-like protein genes in *Heterodera glycines*. Int J Parasitol 31: 1617–1625.
45. Lozano J, Smant G (2011) Survival of Plant-parasitic Nematodes inside the Host. In: Perry RN, Wharton DA, editors. Molecular and Physiological Basis of Nematode Survival. Oxfordshire: CABI. 28–62.
46. Ding X, Shields J, Allen R, Hussey RS (2000) Molecular cloning and characterisation of a venom allergen Ag5-like cDNA from *Meloidogyne incognita*. Int J Parasitol 30: 77–81.
47. Wang X, Li H, Hu Y, Fu P, Xu J (2007) Molecular cloning and analysis of a new venom allergen-like protein gene from the root-knot nematode *Meloidogyne incognita*. Exp Parasitol 117: 133–140.
48. Kang JS, Koh YH, Moon YS, Lee SH (2012) Molecular properties of a venom allergen-like protein suggest a parasitic function in the pinewood nematode *Bursaphelenchus xylophilus*. Int J Parasitol 42: 63–70.
49. Saunders DG, Win J, Cano LM, Szabo LJ, Kamoun S, et al. (2012) Using hierarchical clustering of secreted protein families to classify and rank candidate effectors of rust fungi. PLoS One 7: e29847.
50. Oeser B, Beaussart F, Haarmann T, Lorenz N, Nathues E, et al. (2009) Expressed sequence tags from the flower pathogen *Claviceps purpurea*. Mol Plant Pathol 10: 665–684.
51. Asojo OA, Goud G, Dhar K, Loukas A, Zhan B, et al. (2005) X-ray structure of Na-ASP-2, a pathogenesis-related-1 protein from the nematode parasite, *Necator americanus*, and a vaccine antigen for human hookworm infection. J Mol Biol 346: 801–814.
52. Kobayashi YM, Alseikhan BA, Jones LR (2000) Localization and characterization of the calsequestrin-binding domain of triadin 1. Evidence for a charged beta-strand in mediating the protein-protein interaction. J Biol Chem 275: 17639–17646.
53. Realini C, Rechsteiner M (1995) A proteasome activator subunit binds calcium. J Biol Chem 270: 29664–29667.
54. Reddy AS, Ali GS, Celesnik H, Day IS (2011) Coping with stresses: roles of calcium- and calcium/calmodulin-regulated gene expression. Plant Cell 23: 2010–2032.
55. Aime MC, Phillips-Mora W (2005) The causal agents of witches' broom and frosty pod rot of cacao (chocolate, *Theobroma cacao*) form a new lineage of Marasmiaceae. Mycologia 97: 1012–1022.
56. Frias GA, Purdy LH, Schimidt RA (1995) An inoculation method for evaluating resistance of cacao to *Crinipellis perniciosa*. Plant Disease 79: 787–791.
57. Meinhardt LW, Bellato CM, Rincones J, Azevedo RA, Cascardo JC, et al. (2006) *In vitro* production of biotrophic-like cultures of *Crinipellis perniciosa*, the causal agent of witches' broom disease of *Theobroma cacao*. Curr Microbiol 52: 191–196.
58. Stanke M, Steinkamp R, Waack S, Morgenstern B (2004) AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res 32: W309–W312.
59. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5: 725–738.
60. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) Procheck – A Program to Check the Stereochemical Quality of Protein Structures. Journal of Applied Crystallography 26: 283–291.
61. DeLano WL (2002) The PyMOL Molecular Graphics System. http://www.pymol.org.
62. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, et al. (2011) Integrative genomics viewer. Nat Biotechnol 29: 24–26.