UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOTECNOLOGIA

LORENA SILVA DE OLIVEIRA

MÉTODO DE IDENTIFICAÇÃO DE GENES TAXONOMICAMENTE RESTRITOS EM DADOS DE RNA-SEQ EM ORGANISMO NÃO MODELO

BELÉM - PA
SETEMBRO/2015

LORENA SILVA DE OLIVEIRA

MÉTODO DE IDENTIFICAÇÃO DE GENES TAXONOMICAMENTE RESTRITOS EM DADOS DE RNA-SEQ EM ORGANISMO NÃO MODELO

Dissertação submetida ao Programa de Pós-Graduação em Biotecnologia da Universidade Federal do Pará, como requisito parcial para obtenção do grau de Mestre em Biotecnologia.

Orientador: Profº. Dr. Sylvain Henri Darnet

BELÉM - PA

SETEMBRO/2015

Dados Internacionais de Catalogação-na-Publicação (CIP)
Sistema de Bibliotecas da UFPA

Oliveira, Lorena Silva de, 1992-
Método de identificação de genes taxonomicamente restritos em dados de RNA-seq em organismo não modelo / Lorena Silva de Oliveira. - 2015.

Orientadora: Sylvain Henri Darnet.
Dissertação (Mestrado) - Universidade Federal do Pará, Instituto de Ciências Biológicas, Programa de Pós-Graduação em Biotecnologia, Belém, 2015.

1. Biotecnologia - Piperaceae. 2. Pimenta-do-reino - Genética. 3. Transcriptoma. I. Título.

CDD 22. ed. 660.6

LORENA SILVA DE OLIVEIRA

MÉTODO DE IDENTIFICAÇÃO DE GENES TAXONOMICAMENTE RESTRITOS EM DADOS DE RNA-SEQ EM ORGANISMO NÃO MODELO

Dissertação submetida ao Programa de Pós-Graduação em Biotecnologia da Universidade Federal do Pará, como requisito parcial para obtenção do grau de Mestre em Biotecnologia.

Orientador: Profº. Dr. Sylvain Henri Darnet

Avaliado em: ___/___/_____

Banca Examinadora

_____

Prof. Dr. Igor Schneider

_____

Prof. Drª. Emmanuelle Lautié

_____

Prof. Dr. Rommel Thiago Jucá Ramos

INSTITUIÇÃO E FONTES FINANCIADORAS

Este trabalho foi realizado no Laboratório de Biotecnologia Vegetal, do Instituto de Ciências Biológicas, da Universidade Federal do Pará.

Teve como fonte financiadora a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

AGRADECIMENTOS

A Universidade Federal do Pará e ao Programa de Pós-Graduação em Biotecnologia pela oportunidade e disposição de toda a infraestrutura necessária para a realização do curso.

A CAPES pela concessão da bolsa.

Ao Prof. Dr. Sylvain Henri Darnet pela paciência, orientação, pelos ensinamentos e incentivo que tornaram possível a conclusão deste trabalho.

A todos do Laboratório de Biotecnologia Vegetal pelo companheirismo, amizade, pelos ensinamentos e pelas contribuições para o desenvolvimento deste trabalho.

Ao meu namorado, Hélio Oliveira, que de forma especial e carinhosa me deu força e coragem, me apoiando nos momentos de dificuldades.

A toda minha família e amigos que estiveram comigo em todos os momentos, agradeço pelo carinho, a compreensão e empenho que a muito me dedicaram.

Ao Prof. Dr. Igor Schneider e a Profª. Drª. Emmanuelle Lautié pela disponibilidade e contribuições na pré-banca.

Agradeço a todos que direta ou indiretamente contribuíram para a conclusão deste trabalho.

SUMÁRIO

RESUMO

A pimenta do reino (*Piper nigrum*) é uma planta de alto interesse biotecnológico, alvo de pesquisas tanto para a exploração do metabolismo quanto para o melhoramento relacionado a problemas fitopatológicos, além do entendimento da evolução das angiospermas basais, grupo ancestral ao qual ela pertence. O avanço nos métodos de sequenciamento de nova geração proporcionou o acesso ao patrimônio genético de plantas não modelo possibilitando a abertura de novas perspectivas biotecnológicas. A identificação de genes não homólogos restritos a certas espécies, denominados genes taxonomicamente restritos (GTRs), é um alvo biotecnológico prioritário, especialmente em espécies e grupos divergentes e ancestrais. Este trabalho tem por objetivo estabelecer um método de identificação de GTRs a partir de dados de RNA-seq e de validar a abordagem num conjunto de dados de pimenta do reino. O método consiste na filtragem de transcritos em várias etapas, de forma que os transcritos anotados e os falsos positivos são retirados, e os dados restantes sem informações moleculares são classificados como potenciais GTRs. A aplicação da abordagem em dados de transcriptoma de pimenta do reino (35.631 transcritos) resultou em 22.661 transcritos anotados por similaridade. Os transcritos não anotados nessa primeira análise foram processados na ferramenta TRAPID, obtendo 12.895 transcritos não anotados. A avaliação dos transcritos para detecção de falsos positivos resultou em 245 transcritos verdadeiros, que foram analisados quando a presença de RNA não codificante, sendo encontrados 204 transcritos sem identificação. Ao final da aplicação do método restaram 71 transcritos não anotados com regiões codificantes de proteínas, assinalados como potenciais GTRs. A caracterização desses potenciais GTRs em pimenta do reino pode fornecer novas informações sobre o mecanismo molecular dessa espécie e talvez elucidar vias para o estabelecimento de cultivares tolerantes a doenças.

Palavras-chave: Pimenta do reino, planta não modelo, transcriptoma, Genes Taxonomicamente Restritos.

ABSTRACT

Black pepper (*Piper nigrum*) is a biotechnologically interesting plant, research target for both metabolism exploration and improvement related to phytopathological problems, in addition to understanding the evolution of basal angiosperms, ancestral group to which it belongs. With the technological revolution, the next generation sequencing offered access to genetic heritage of non model plants enabling the opening of new biotechnological perspective. The identification of non homologous genes restricted to certain species, called taxonomically restricted genes (TRGs), is a primary biotechnological target, especially in species and groups that are divergent and ancestral. This study aims to establish a method for TRGs identification from RNA-seq data and to validate the approach a dataset for black pepper. The method consists in filtering the transcripts in several stages, so that the annotated transcripts and false positives are removed, and the remaining data without molecular information are classified as potential TRGs. The application of this approach to a black pepper transcriptome dataset (35,631 transcripts) resulted in 22,661 transcripts annotated by similarity. The transcripts that were not annotated in this first analysis were processed in the TRAPID tool, resulting in 12,895 transcripts not annotated. The evaluation of transcripts for false positive detection resulted in 245 true transcripts that were analyzed for the presence of non-coding RNA, resulting in 204 unidentified transcripts. At the end of the method application 71 non annotated transcripts remained with coding regions of protein, indicating potential TRGs. The characterization of these potential TRGs in black pepper can provide new information about the molecular mechanism of this specie and perhaps elucidate pathways for the establishment of cultivars tolerant to disease.

Keywords: Black pepper, non model plant, transcriptome, Taxonomically Restricted Genes (TRGs).

LISTA DE ILUSTRAÇÕES

1. INTRODUÇÃO

A pimenta do reino (*Piper nigrum* L.) é uma planta originária da Índia, que se destaca pelo alto teor de compostos bioativos, como a piperina, apreciado tanto na culinária, quanto na medicina tradicional (JARAMILLO; MANOS, 2001; LEMOS, 2003). Estima-se que em 2013 foram produzidos mundialmente mais de 400 mil toneladas de pimenta do reino, dos quais 79,4% foram produzidos na Ásia, 16,2% na Américas e 4,4% na África, segundo dados da *Food and Agriculture Organization of the United Nations* (FAO, 2015).

No Brasil, a pipericultura possui uma importância econômica e social (LOURINHO *et al.*, 2014). O cultivo é fortemente relacionado à agricultura familiar e fonte de renda para milhares de famílias na Amazônia (LOURINHO *et al*., 2014). O melhoramento dos cultivares brasileiros é importante no âmbito de melhorar a produção nacional e sustentar muitas famílias em regiões pouco desenvolvidas e isoladas (MAGEVSKI *et al*., 2011; LOURINHO *et al*., 2014).

1.1. GÊNERO *PIPER*: ANGIOSPERMAS BASAIS DE INTERESSE BIOTECNOLÓGICO

1.1.1. Interesse econômico

A pimenta do reino merece destaque, por se tratar de uma especiaria amplamente comercializada no mercado global (AHMAD *et al*., 2012). Em nível de produção mundial, é estimado que só em 2013 tenham sido produzidos mais de 400 mil toneladas de pimenta do reino, sendo o Vietnã, Indonésia, Índia, Brasil e China os principais países produtores (FAO, 2015). No Brasil, segundo dados do IBGE (2015), a produção em 2012 contabilizou 43 mil toneladas, com o estado do Pará dominando a produção (75%) seguido do Espírito Santo (15%) e Bahia (9%).

A valorização do cultivo da pimenta do reino varia devido a oscilação dos preços no mercado internacional e limitações na cultura associadas a baixa variabilidade genética da espécie ocasionando vulnerabilidade a doenças como a fusariose que gera, por sua vez, grandes perdas na produção (ALBUQUERQUE, 1964; VANAJA *et al*, 2007; LEMOS *et al*., 2011; GORDO *et al*., 2012).

A fusariose foi descrita pela primeira vez em 1964 e até hoje nenhum cultivar ou sistema de cultivo foi encontrado para limitar ou impedir a ação do patógeno responsável pela doença (ALBUQUERQUE, 1964; LEMOS *et al*., 2011). Nos últimos

anos, com os grandes avanços realizados na genética e genômica vegetal, as abordagens biotecnológicas parecem ser promissoras para o melhoramento da pimenta do reino (LEMOS *et al*., 2011; GORDO *et al*., 2012; JOY*et al*., 2013; HU *et al*., 2015).

1.1.2. Interesse biotecnológico

O interesse de desenvolver estudos biotecnológicos da pimenta do reino é duplo: a) melhorar os cultivar de pimenta do reino para atender as demandas na agricultura, b) caracterizar vias de biossíntese de compostos com alto valor (LEMOS *et al*., 2011; GUTIERREZ *et al*., 2013).

O gênero *Piper* possui a piperina como principal composto bioativo, o qual pode ser utilizado para diversas finalidades, principalmente produção de fármacos (AHMAD *et al*., 2012; GUTIERREZ *et al*., 2013). Em função disso, a química de suas espécies tem sido amplamente investigada e já resultou no conhecimento de uma grande diversidade de metabólitos bioativos (PARMAR *et al*., 1997).

A pimenta do reino é citada para diversas finalidades como, por exemplo, antioxidantes, antiapoptóticos, antibactericida, antifúngico entre outros (PARMAR *et al*., 1997; JUSTO; SILVA, 2008; AHMAD *et al*., 2012; BARBOSA *et al*., 2013).

Poucos cultivares de pimenta do reino estão disponíveis e muitos problemas fitopatológicos e agronômicos ainda não são resolvidos. Um dele é as grandes perdas de produção ocasionadas pelos fungos patogênicos *Fusarium solani*. f.sp. *piperis* e *Phytophora capsici* (ALBUQUERQUE, 1964; BENCHIMOL *et al*., 2000). A obtenção de linhagens de plantas tolerantes e/ou resistentes aos patógenos é ainda um grande desafio (PHILIP *et al*., 1992; BENCHIMOL *et al*., 2000; LEMOS, 2003).

1.1.3. A pimenta-do-reino na era da agrigenômica

A dificuldade de desenvolver estratégias mais eficazes de controle ao patógeno está relacionada a escassa informação a nível molecular e genético sobre a espécie. Sabe-se que a pimenta do reino e uma espécie poliploide, com número de cromossomos 2n=52 (JOSE; SHARMA, 1984; NAIR *et al*., 1993) e com valor estimado do transcriptoma de raiz 3,6Gb (GORDO *et al*., 2012).

Nesse contexto, tecnologias de sequenciamento de nova geração (NGS) se tornam ferramentas atrativas para a obtenção de informações genéticas (EKBLOM; GALINDO, 2011; STRICKLER *et al*., 2012). Em espécies não modelo, com genoma

poliploide e de grande tamanho, a primeira abordagem desenvolvida é a caracterização dos genes expressos, pelo transcriptoma, numa condição determinada, como por exemplo, durante a interação planta-patógeno (LEMOS, 2003; CORRE; CHALLIS, 2007).

A caracterização do transcriptoma de uma planta não modelo, como a pimenta do reino, pode levar a identificação de transcritos de genes homólogos, já identificados e caracterizados funcionalmente em outras plantas, mas também o reconhecimento de transcritos sem semelhança com genes ou transcritos descritos previamente. Estes transcritos correspondam à genes taxonomicamente restritos (GTRs) (DOMAZET-LOZO; TAUTZ, 2003; HORAN *et al*., 2008; YANG *et al*., 2013).

1.2. ANGIOSPERMAS BASAIS E GENES TAXONOMICAMENTES RESTRITOS (GTRs)

1.2.1. Definição de GTRs

Os GTRs compreendem aproximadamente 10 a 20% das regiões codificantes do genoma e não apresentam homologia com genes descritos em outros organismos (KHALTURIN *et al*., 2009; JOHNSON; TSUTSUI, 2011). Os GTRs podem apresentar uma função biológica específica e restrita à uma espécie ou linhagem e são considerados intimamente associados à respostas adaptativas (WILSON *et al*., 2005; JOHNSON; TSUTSUI, 2011; KHALTURIN *et al*., 2009; LIN *et al*., 2010; YANG *et al*., 2013).

Dependendo dos autores, os GTRs podem ser denominados de genes órfãos ou genes de linhagem especifica (WILSON *et al*., 2005, 2007; KHALTURIN *et al*., 2009; YANG *et al*., 2013). O termo gene órfão foi primeiramente utilizado no início de 1990, quando foi realizado o sequenciamento do cromossomo III de levedura, para caracterizar os genes que não foram anotados funcionalmente (OLIVER *et al*., 1992; KHALTURIN *et al*., 2009).

Porém o sequenciamento de genomas bacterianos, demostrou que algumas espécies estreitamente relacionadas possuem genes órfãos, não encontrados em outros grupos taxonômicos. Resultando na origem do termo genes de linhagem específica ou de genes taxonomicamente restritos (GTRs) (WILSON *et al*., 2005; KHALTURIN *et al*., 2009; TAUTZ; DOMAZET-LOSO, 2011; YANG *et al*., 2013).

A definição de GTR é em relação a origem desses novos genes, os fenômenos evolutivos que levam ao surgimento de sequências gênicas e, possivelmente, de novas funções. Os GTRs podem ser relacionados à novos fenótipos celulares, morfológicos, fisiológicos e comportamentais observados em alguns organismos ou em linhagens restritas (DOMAZET-LOSO; TAUTZ, 2010; KAESSMANN, 2010).

1.2.2. Origem dos GTRs

A origem genômica possível dos GTRs pode ser baseada em diversos fenômenos: 1) a duplicação de genes – foi o primeiro mecanismo proposto para a origem do gene e refere-se à duplicação de segmentos cromossômicos, contendo genes inteiros ou fragmentos de genes (evento comum) ou a duplicação completa do genoma, que causa um maior impacto genômico (evento raro); 2) embaralhamento de éxons – processo que gera novos arranjos de éxons e pode levar a criação de novos genes quiméricos; 3) fusão ou fissão – processos relacionados à recombinação gênica que resulta em proteínas com novas modalidades de domínios, sendo as fusões quatro vezes mais frequentes que as fissões; 4) elementos de transposição – promovem ou controlam rearranjos cromossômicos e expressão gênica; 5) transferência lateral ou horizontal de genes – consiste na transferência de material genético entre organismos diferentes e 6) origem de genes *de novo* - refere-se a novos genes que se originam diretamente de regiões de DNA não codificante (KAESSMANN, 2010; RAZQUIN, 2013; WU; ZHANG, 2013).

A origem dos GTRs é complexa e não bem descrita. A identificação dos GTRs é mais baseada em métodos de bioinformática, como a predição de sequências codificantes ou a busca de homologia entre sequências (DOMAZET-LOSO; TAUTZ, 2010; KAESSMANN, 2010; YANG *et al*., 2013).

1.2.3. Identificação bioinformática dos GTRs nos dados genômicos

1.2.3.1. Fontes de informação biológica para detectar os GTRs

O NGS (Next-Generation Sequencing) está sendo cada vez mais utilizado em biologia e foi observado um crescimento exponencial dos dados de genomas e transcriptomas numa grande quantidade de espécies, modelo ou não modelo (LISTER *et al*., 2009; MOROZOVA *et al*., 2009; GRABHERR *et al*., 2011).

A criação de bancos de dados que disponibiliza dados biológicos de diversas espécies e proporciona a anotação funcional de transcritos, porém ainda são necessários muitos recursos para caracterizar genes que ainda se encontram desconhecidos, e assim poder identificar a real informação biológica que estes genes possuem (BRENT, 2005; GRABHERR *et al*., 2011; LI *et al*., 2011).

1.2.3.2. Métodos extrínsecos de detecção

A predição de função por homologia é a base de todos os métodos bioinformáticos (RUEPP; MEWES, 2006). Esta abordagem é baseada na comparação de sequências à nível proteico ou nucleotídico, usando ferramentas bioinformáticas e matrizes evolutivas, afim de definir quais os graus de similaridade entre sequências (WHISSTOCK; LESK, 2003; RUEPP; MEWES, 2006; HORAN *et al*., 2008). Se o nível de similaridade é suficiente, pode ser definido que as duas sequências são homólogas e apresentam uma conservação estrutural e funcional (WHISSTOCK; LESK, 2003; HORAN *et al*., 2008).

No caso de detecção de GTRs, as sequências de interesse são as sequências que apresentaram nenhuma similaridade com outra sequência conhecida em outro organismo, obtendo então, sequências sem homologia (DOMAZET-LOSO; TAUTZ, 2003; WILSON *et al*., 2007; KHALTURIN *et al*., 2009; MILDE *et al*., 2009; YANG *et al*., 2013). A identificação de GTRs proporciona a obtenção de genes com função desconhecidas, podendo ser caracterizados como genes restritos à uma espécie, ou a um grupo de organismos, e possivelmente relacionados a evolução destas espécies (WILSON *et al*., 2005, 2007; JOHNSON; TSUTSUI, 2011).

A avaliação por homologia também é um processo primordial para avaliação de genes que ainda não possuem suas funções descritas (STUDER; ROBINSON-RECHAVI, 2009; TRACHANA *et al*., 2011). Os genes homólogos podem ser classificados em ortólogos e parálogos, dependendo se eles surgiram por especiação ou duplicação, respectivamente (ALEXEYENKO *et al*., 2006; STUDER; ROBINSON-RECHAVI, 2009; TRACHANA *et al*., 2011).

1.2.3.3. Métodos intrínsecos de detecção

Os métodos intrínsecos de detecção de GTRs são baseados nas informações contidas na própria sequência, sem comparação com dados externos, como outras

sequências de outros organismos. Assim a identificação de um gene pode ser baseada na detecção de quadros de leitura aberta, ou também chamado de ORFs (*Open Read Frames*) (ALIMI *et al*., 2000; BRENT, 2005).

Outros parâmetros usados para detecção de genes são ligados à estrutura da sequência genica, como quantidade de aminoácidos e nucleotídeos, percentual de GC, além de identificar se a sequencia que codifica o gene é completa ou incompleta (MILDE *et al*., 2009; GRABHERR *et al*., 2011). A caracterização de um transcrito é primordial para identificar se o mesmo é um transcrito verdadeiro, ou se trata de um falso positivo, podendo ser um artefato de bioinformática (WHISSTOCK; LESK, 2003; MOROZOVA *et al*., 2009; LI *et al*., 2011).

1.2.4. Ferramentas de identificação de genes taxonomicamente restritos (GTRs)

1.2.4.1. Baseado em similaridade entre sequencias

Para cada grupo taxonômico em estudo, existe um método de identificação de GTRs, porém o processo de comparação entre genomas utiliza-se da ferramenta BLAST, sendo o método mais comum entre eles, buscando funções de proteínas a partir da similaridade entre sequencias (ALTSCHUL *et al*., 1990; DOMAZET-LOSO; TAUTZ, 2003; WILSON *et al*., 2007; MILDE *et al*., 2009; YANG *et al*., 2013).

1.2.4.2. Baseado em homologia

As ferramentas baseadas na detecção de genes homólogos envolve, principalmente, a análise por ortologia que possui duas abordagens diferentes, uma baseada em árvores filogenéticas e outra em gráfos (STUDER; ROBINSON-RECHAVI, 2009; TRACHANA *et al*., 2011). A primeira consiste em definir famílias de genes homólogos e construir uma árvore para cada um dos genes, atribuindo incongruências a eventos de especiação, duplicação ou perda de gene. Enquanto que a segunda, baseia-se em semelhança entre sequências, sendo o método baseado em gráficos o mais utilizado, devido maior eficiência, capacidade de manipulação de um grande número de dados e custo mais acessível (VAN DER HEIJDEN *et al*., 2007; TRACHANA *et al*., 2011; RAZQUIN, 2013).

Existem diversos bancos de dados de análise ortológica. O método baseado em árvore inclui bancos de dados como TreeFam, Ensembl Compara, PhylomeDB, Panther

e HOGENOM, enquanto que os bancos OrthoDB, EggNOG, OMA, InParanoid e OrthoMCL são baseados em gráficos (VAN DER HEIJDEN *et al*., 2007; TRACHANA *et al*., 2011; RAZQUIN, 2013).

1.2.4.3. Baseados nos domínios e motivos proteicos

A comparação com motivos, domínios conservados em famílias de proteínas também é uma forma de caracterizar transcritos não anotados (WILSON *et al*., 2007). O uso de bases de dados como PFAM (FINN *et al*., 2014) e INTERPRO (MITCHELL *et al*., 2015), disponibilizam um BLAST *online* para que seja realizada a identificação e a caracterização de grupos de famílias proteicas que não foram reconhecidos pelas avaliações por similaridade e homologia (FINN *et al*., 2014; MITCHELL *et al*., 2015).

1.2.4.4. Detecção de transcritos artefatos de montagem de novo de transcriptoma

No conjunto de dados obtidos por RNA-seq, a presença de transcritos artefatos é observado, e podem ter sido gerados devido a viéses de análises de bioinformática (SIMS *et al*., 2014). A análise do conjunto de transcritos através da avaliação da cobertura para cada *read* é uma forma de identificar falsos positivos e obter transcritos com qualidade de cobertura alta, a qual o nível de cobertura é relativo ao tamanho da sequencia (WANG *et* al., 2011; SIMS *et al*., 2014).

1.2.4.5. Predição de ORFs

As tecnologias de RNA-seq geram um vasto número de *reads* curtos, resultando na necessidade de identificar quadros de leitura aberta (ORFs) dos transcritos montados (TANG *et al*., 2015), o que define a importância de identificar a funcionalidade de um transcrito obter um produto proteico, permitindo, assim, a caracterização de um GTRs (LI, 2012; TANG *et al*., 2015). Utilizando a ferramenta GeneMarkS é possível identificar regiões codificantes de proteínas em transcritos de RNA (BESEMER *et al*., 2001; TANG *et al*., 2015).

Diante de todas as considerações acerca dos GTRs o presente estudo tem por finalidade identificar os prováveis GTRs presentes no transcriptoma de *P. nigrum*, a partir da comparação com o proteoma predito de plantas superiores e das angiospermas basais, na qual perspectivas futuras podem caracterizar a função destes genes e

determinar o quão específico será para a linhagem, a fim de promover recursos para o fomento do melhoramento genético da espécie.

2. OBJETIVOS

2.1. OBJETIVO GERAL

- Desenvolver um método de identificação de genes taxonomicamente restritos em dados de RNA-seq em organismo não modelo.

2.2. OBJETIVOS ESPECÍFICOS

- Realizar o levantamento de métodos de identificação de genes taxonomicamente restritos;
- Comparar métodos de identificação de genes taxonomicamente restritos;
- Comparar o proteoma de transcrito de pimenta do reino com proteoma predito de monocotiledôneas e dicotiledôneas;
- Equiparar o transcriptoma de pimenta do reino com o proteoma predito das demais espécies pertencentes ao grupo das angiospermas basais;
- Identificar, através de análise por ortologia, outros eucariotos que possuem homologia com o transcriptoma de pimenta do reino;
- Caracterizar transcritos quanto a regiões que apresentam domínios de proteínas descritos;
- Identificar e caracterizar os transcritos com regiões codificadoras de proteínas;
- Definir os transcritos identificados como prováveis genes taxonomicamente restritos (GTRs).

3. MATERIAL E MÉTODOS

3.1. CONJUNTO DE DADOS DE SEQUÊNCIAS

No presente estudo foram utilizados dados de transcriptoma de pimenta do reino (*Piper nigrum*) e proteomas preditos das espécies *Arabidopsis thaliana*, *Oryza* sativa, *Amborella trichopoda*, *Aristolochia fimbriata*, *Nuphar adverna*, *Persea americana* e *Zamia furfuracea*. Cada conjunto de dados de proteoma encontra-se disponível para download em bancos de dados como Ensembl Plants e Ancestral Angiosperm Genome Project (AAGP), além disso, as espécies utilizadas foram separadas em 3 grupos, sendo eles dicotiledônea, monocotiledônea e angiospermas basais (Tabela 1).

Os dados de transcriptoma de pimenta do reino encontram-se disponíveis para download em Sequence Read Archive (SRA) no banco de dados do National Center for Biotechnology Information (NCBI), estes dados são resultado dos trabalhos de Gordo *et al*. (2012) e Joy *et al*. (2013), sendo um conjunto de dados de raiz de pimenta do reino desenvolvido com a plataforma SOLiD e o outro de folha de pimenta do reino desenvolvido com a plataforma Illumina, respectivamente.

Tabela 1 – Dados dos transcritos das espécies utilizadas no estudo.

Espécies	Classificação	Nº de transcritos	Fonte
Arabidopsis thaliana	Dicotiledônea	35386	Ensembl Plants
Oryza sativa	Monocotiledônea	42132	Ensembl Plants
Piper nigrum	Angiosperma basal	35631	SRA NCBI
Amborella trichopoda	Angiosperma basal	27313	Ensembl Plants
Aristolochia fimbriata	Angiosperma basal	37643	Ancestral Angiosperm Genome Project (AAGP)
Nuphar adverna	Angiosperma basal	49422	Ancestral Angiosperm Genome Project (AAGP)
Persea americana	Angiosperma basal	61548	Ancestral Angiosperm Genome Project (AAGP)
Zamia furfuraceae	Angiosperma basal	30585	Ancestral Angiosperm Genome Project (AAGP)

3.2. FERRAMENTAS DE BIOINFORMÁTICA

As ferramentas utilizadas para o processo de identificação e caracterização de GTRs consistem nas ferramentas BLAST, TRAPID, CLC Workbech e GenMarkS, enquanto que os bancos de dados utilizados foram OrthoMCL, PFAM e RFAM.

Com a ferramenta BLAST foram utilizada s as variações BLASTx e BLASTn, afim de realizar a anotação por similaridade dos transcritos de pimenta do reino com as

espécies de análise descritas no tópico anterior (ALTSCHUL *et al*., 1990; GISH; STATES, 1993).

O sistema TRAPID foi associado ao banco de dados OrthoMCL DB, versão 5, e teve por finalidade identificar a homologia dos transcritos de pimenta do reino com o proteoma predito de outros eucariotos (LI *et al*., 2003; FISCHER *et al*., 2011; VAN BEL *et al*., 2013).

A ferramenta CLC Workbench, versão 7.0, foi utilizada para mapear *reads* de baixa e alta cobertura (WANG *et al*., 2011). Enquanto que o GenMarkS teve a finalidade de detectar regiões codificantes de proteínas e inferir os potenciais GTRs (BESEMER *et al*., 2001).

A partir de um BLAST *online* foram utilizados os bancos de dados PFAM (FINN *et al*., 2014) e RFAM (NAWROCKI *et al*., 2015), para realizar a identificação de famílias de proteínas e RNA não codificante (ncRNA), respectivamente.

3.3. APLICAÇÃO DA ABORDAGEM

O processo de identificação e caracterização de GTRs (Figura 1) foi elaborado em 7 etapas. Na etapa 1 (Figura 1A), foi realizado a comparação entre proteomas preditos, alinhando inicialmente o transcriptoma de *P. nigrum* com o proteoma de *A. thaliana* e *Oryza sativa*, e em seguida o transcriptoma de pimenta do reino com as demais espécies pertencentes ao grupo das angiospermas basais, utilizando a ferramenta BLASTx (e-value < $1e^{-10}$).

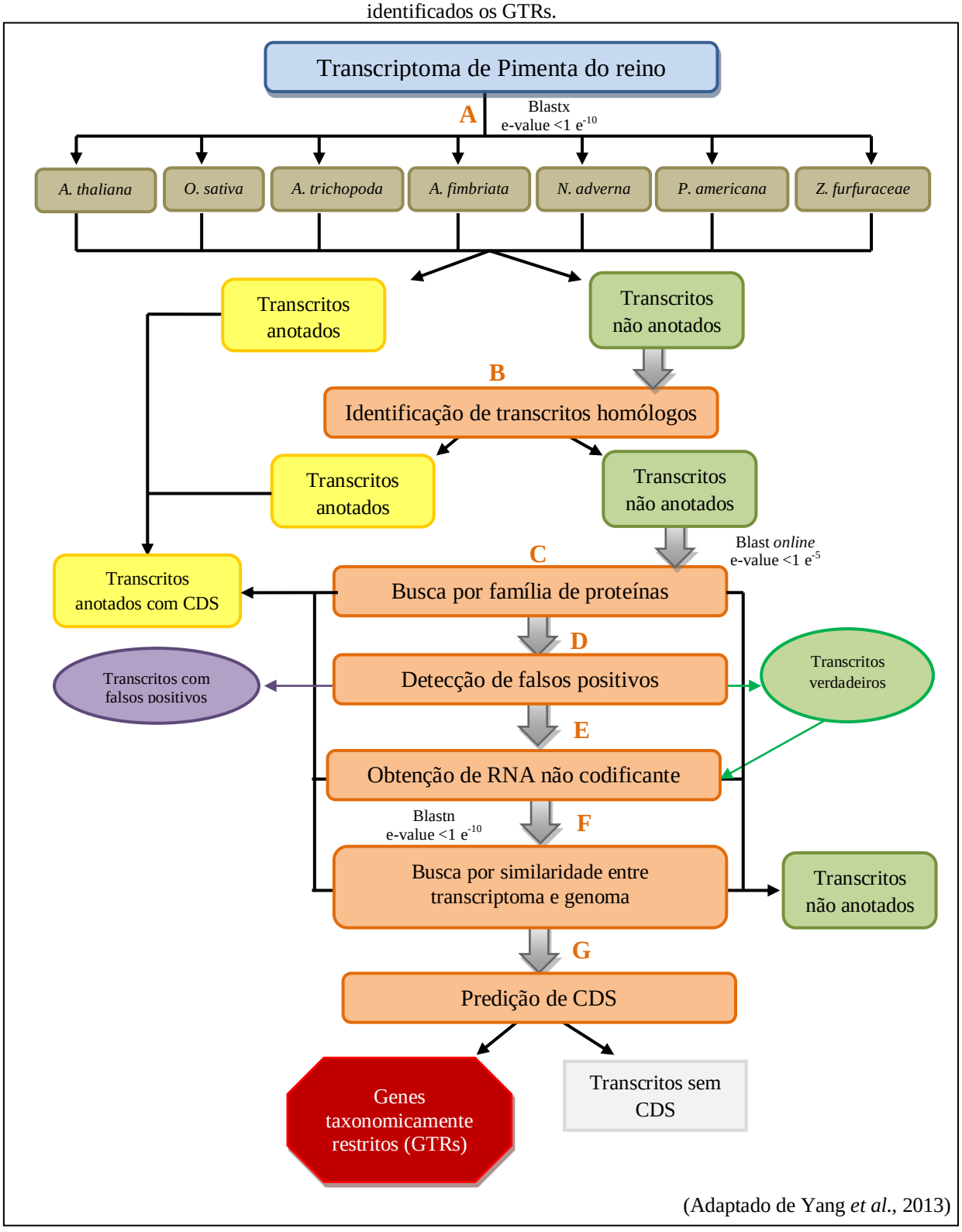
Na etapa 2, (Figura 1B), os dados resultantes do alinhamento que não apresentaram similaridade foram analisados com a ferramenta TRAPID associada ao banco de dados OrthoMCL. Em seguida, na etapa 3 (Figura 1C), os transcritos de pimenta do reino sem homologia foram avaliados por família de proteínas, utilizando o banco de dados PFAM .

O conjunto de dados resultante sem família de proteínas identificada foi avaliado quanto a presença de transcritos com falsos positivos, a partir da avaliação da cobertura para cada *read*, esta é caracterizada como a etapa 4 (Figura 1D). No procedimento seguinte, etapa 5 (Figura 1E), os transcritos identificados como verdadeiros, foram analisados quanto a famílias de ncRNA, com auxilio do banco de dados RFAM.

Os dados sem identificação de ncRNA, foram avaliados na etapa 6 (Figura 1F) para identificar a existência de transcritos com similaridade entre os transcritos de pimenta do reino e espécies com genoma descrito do grupo das angiospermas basais. Dessa forma, foi realizado um BLASTn (e-value < $1e^{-10}$) entre pimenta do reino e as espécies *Amborella trichopoda* e *Aristolochia fimbriata*.

A etapa 7 (Figura 1G), resultou no carregamento dos transcritos não anotados da etapa anterior, a ferramenta GeneMarkS, A partir desta última etapa de avaliação dos transcritos de pimenta do reino foi obtido os potenciais GTRs.

Figura 1 – Procedimento de identificação e caracterização de gene taxonomicamente restrito. (A) Etapa de comparação entre transcriptoma de pimenta do reino e os proteomas preditos das 7 espécies selecionadas; (B) Etapa de identificação de transcritos homólogos; (C) Etapa de busca por família de proteínas; (D) Detecção de falsos positivos nos transcritos de pimenta do reino; (E) Etapa de utilização dos transcritos caracterizados como verdadeiros para busca por similaridade entre pimenta do reino e genomas descritos de espécies próximas filogeneticamente; (F) Etapa de identificação dos transcritos em ncRNA; (G) Etapa de predição de ORFs dos transcritos que não foram caracterizados com ncRNA, sendo identificados os GTRs.



(Adaptado de Yang *et al.*, 2013)

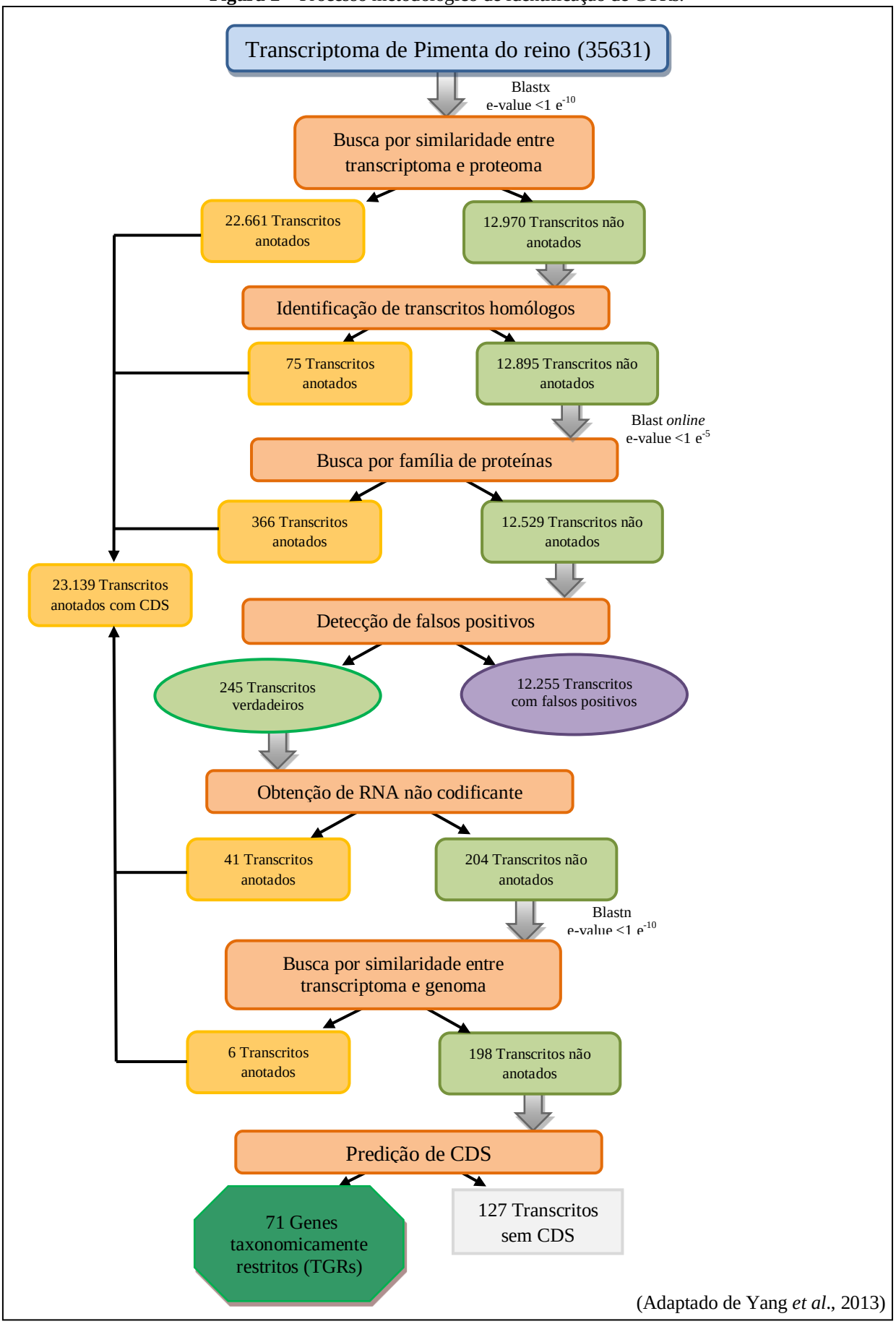
4. RESULTADOS

4.1. MÉTODO DE IDENTIFICAÇÃO DE GTRs

A metodologia de identificação de GTRs é caracterizada de subtrativa, consistindo na filtragem de transcritos de acordo com as etapas descritas na Figura 2. Cada etapa permite a filtragem de uma parte dos transcritos, uma denominada de anotada é removida, se for identificada como artefato de sequenciamento, o restante que passou pelo filtro foi usado na próxima etapa.

Como resultado de todo o processo de análise dos dados foi possível obter uma metodologia satisfatória, que atende as necessidades questionadas pelo estudo, além de descrever um método de identificação de GTRs apropriado para outros estudos envolvendo análise de RNA-seq em organismos não modelo.

Figura 2 – Processo metodológico de identificação de GTRs.

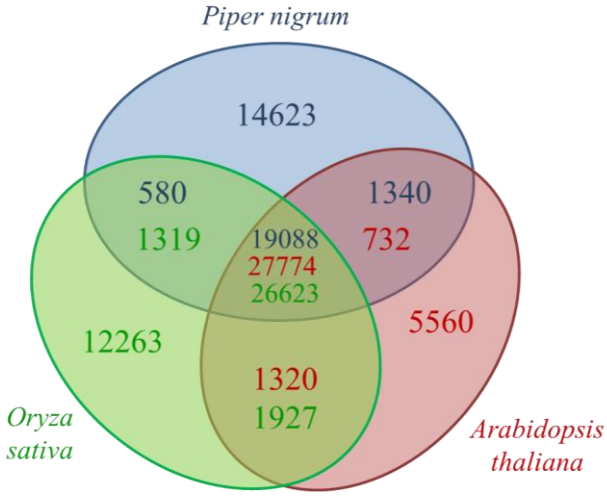


(Adaptado de Yang *et al*., 2013)

4.2. APLICAÇÃO DA ABORDAGEM EM CONJUNTO DE RNA-SEQ EM ORGANISMO NÃO MODELO

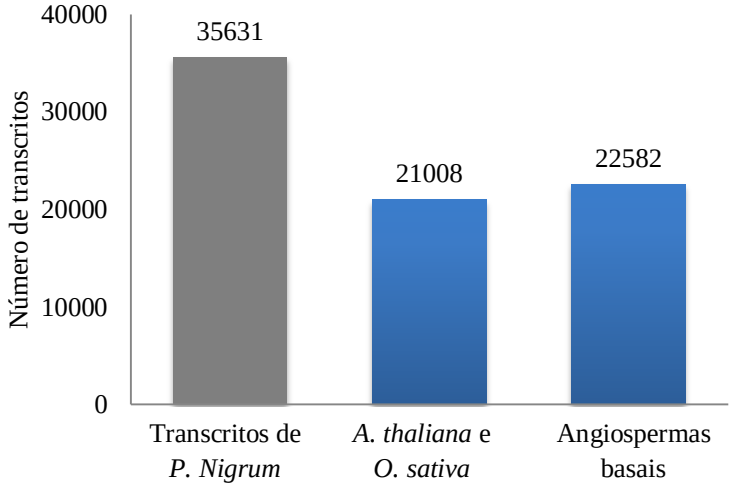
A primeira etapa de identificação e caracterização de GTRs consistiu em comparação do transcriptoma total de pimenta do reino, contendo 35.631 transcritos, com proteoma das espécies *A. thaliana* e *O. sativa*, neste procedimento obteve-se a anotação de 21.008 transcritos, dos quais 19.088 transcritos de pimenta do reino são comuns as três espécies (Figura 3). Uma vez que, 14.623 transcritos são únicos de pimenta do reino, ou seja, não compartilham similaridade com nenhuma das espécies analisadas.

Figura 3 – Diagrama de Venn descrevendo a correlação de similaridade entre as espécies, representando os transcritos compartilhados e únicos entre as espécies.



Na etapa seguinte foi realizada a comparação do conjunto total transcritos de pimenta do reino com o proteoma das demais angiospermas basais obteve-se o ganho de 1653 transcritos, visto que, este processo resultou em 22582 transcritos anotados (Figura 4).

Figura 4 – Variação da quantidade de transcritos de pimenta do reino anotados por similaridade entre o grupo das espécies *A. thaliana* e *O. sativa* e o grupo das angiospermas basais.



A comparação entre os transcritos de pimenta do reino e o proteoma de *A. fimbriata* obteve-se 502 transcritos exclusivos, ou seja, corresponde a transcritos que só foram anotados a partir da comparação entre pimenta do reino com uma única espécie, neste caso *A. fimbriata*. Em contrapartida, com o proteoma de *Z. furfuracea* resultou em 32 transcritos exclusivos (Tabela 2).

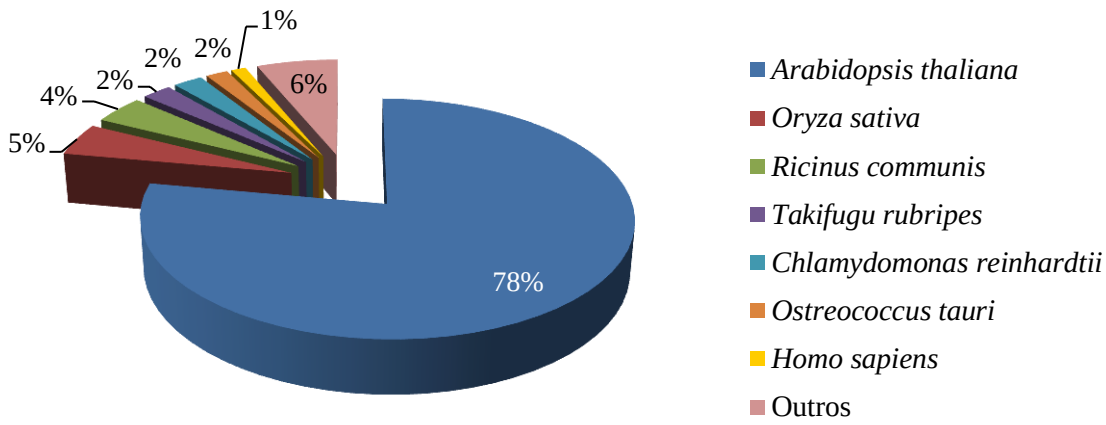
Tabela 2 – Quantidade de transcritos de pimenta do reino que compartilham similaridade com apenas uma espécie e transcritos que compartilham similaridade entre todas as espécies do grupo das angiospermas basais.

Espécies	**Nº de transcritos**
Trancriptoma de *Piper nigrum*	35631
Transcritos comuns a todas as espécies	16063
Aristolochia fimbriata	502
Persea americana	215
Amborella tricophoda	168
Nuphar adverna	51
Zamia furfuraceae	32

Em resumo, a partir da análise por similaridade entre os transcritos de pimenta do reino e o proteoma predito das espécies *A. thaliana, O. sativa* e o grupo das angiospermas basais resultou em 22661 transcritos anotados, restando 12970 transcritos que foram utilizados para a etapa seguinte de avaliação.

A avaliação dos transcritos de pimenta do reino por homologia, através da ferramenta TRAPID, identificou 75 transcritos como ortólogos entre diversos grupos de eucariotos (Figura 5). Observa-se que 78% dos transcritos foram identificados como homólogo a espécie *A. thaliana*, enquanto que a espécie *H. sapiens* apresentou o equivalente a 1% dos transcritos.

Figura 5 – Transcritos de pimenta do reino identificados com homologia em diversos grupos de eucariotos.



Utilizando o banco de dados PFAM foram encontrados, dos 112.895 transcritos de pimenta do reino, 366 transcritos identificados como motivos funcionais, domínios e famílias proteicas distintas. Obteve-se 2 motivos funcionais, 2 famílias e 3 domínios de proteínas como os mais encontrados dentre os transcritos identificados (Tabela 3).

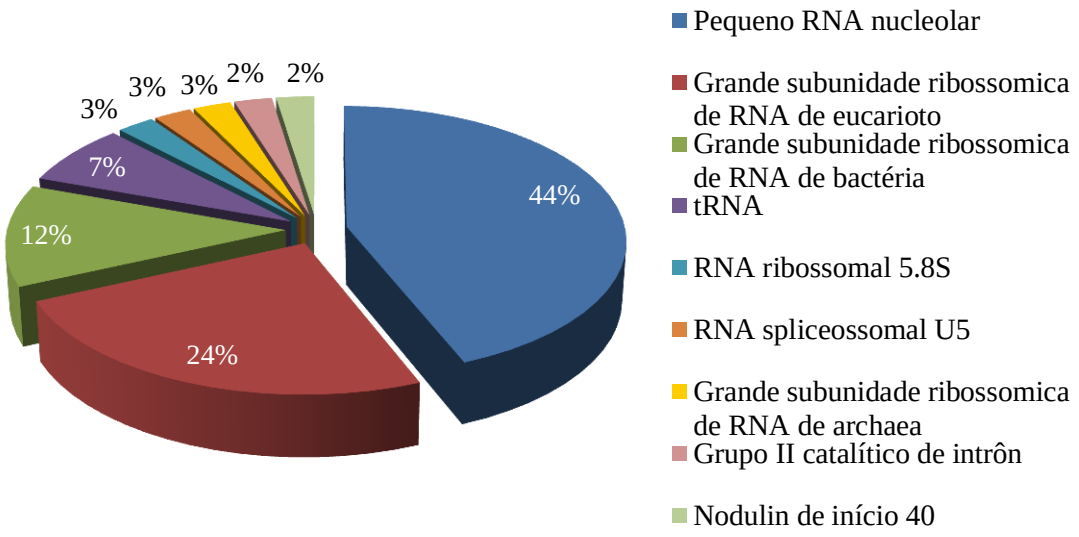
Tabela 3 – Motivos funcionais, famílias e domínios proteicos mais identificados pelo banco de dados PFAM nos transcritos de pimenta do reino.

ID do Transcrito	ID PFAM	Descrição	Classificação
UN00831	PF03478.14	DUF295 Family	Família de proteínas
UN10329			
UN26089			
UN26763			
UN29680			
UN09568	PF07172.7	Proteína Rica em Glicina	Família de proteínas
UN10287			
UN19343			
UN30659			
UN31435			
UN06471	PF00646.29	Proteína F-box (CL0271)	Domínio de proteínas
UN07827			
UN10259			
UN11326			
UN31650			
UN31951			
UN32898			
UN17651	PF00069.21	Proteína Quinase	Domínio de proteínas
UN21493			
UN26759			
UN06368	PF00538.15	Histona	Domínio de proteínas
UN14898			
UN18954			
UN14713	PF05678.10	VQ Motif	Motivo funcional
UN32102			
UN32108			
UN20434	PF06203.10	CCT Motif	Motivo funcional
UN25899			
UN26752			

A detecção de transcritos com falsos positivos foi realizada a partir da análise de cobertura, utilizando a ferramenta CLC Workbench, versão 7.5. Dessa forma, foram detectados 245 transcritos classificados com alta cobertura, de forma que os demais transcritos obtiveram a cobertura por *reads* igual a zero.

A identificação de regiões não codificantes realizado pelo banco de dados RFAM, resultou em 41 transcritos de pimenta do reino com ncRNA identificadas, dentre eles as funções de pequeno RNA nucleolar (44%) e grande subunidade ribossômica de RNA de eucarioto foram as mais encontradas entre os transcritos identificados (Figura 6).

Figura 6 – Função dos ncRNA identificados no transcriptoma de pimenta do reino.



Ao final do processo de identificação de ncRNA restaram 204 transcritos sem nenhuma caracterização. Para a avaliação destes transcritos restantes foi desenvolvido um BLASTn contra o genoma das espécies *A. trichopoda* e *A. fimbriata*, que são as espécies do grupo das angiospermas basais que possuem seu genoma descrito. Dessa forma, foram identificados 6 transcritos de pimenta do reino como similaridade apenas com a espécie *A. fimbriata*.

Para caracterizar um transcrito como provável GTRs é necessário identificar se o mesmo apresenta região codificante de proteína, dessa maneira foi utilizada a ferramenta GenMarkS para avaliar os 197 transcritos de pimenta do reino que não apresentaram informações sobre função proteica, a partir das etapas metodológicas anteriores. Como resultado obteve-se 71 transcritos identificados como GTRs que possuem regiões codificantes de proteínas.

Os 71 transcritos identificados como GTRs são caracterizados por terem um valor médio de 143 aminoácidos entre os transcritos, possuindo sequencias entre 28 a 1.247 aminoácidos, o valor médio de nucleotídeos foi calculado em 438, contendo sequencias de 87 a 3744 nucleotídeos. Além disso, o percentual de GC nestes transcritos é de 54,71 %, sendo identificados 40 transcritos 3' *full*, 1 transcrito 5' *full* e o 1 transcrito *full lenght*.

O transcrito de pimenta do reino de identificador UN01868 é caracterizado como *full length* devido possuir códon de iniciação e de parada para o processo de transcrição. Este transcrito possui 507 nucleotídeos e 169 aminoácidos, além de percentual de GC de 59,76%, dessa forma podemos classifica-lo como um GTRs em potencial para futuros estudos.

Os 71 transcritos de pimenta do reino caracterizados como potenciais GTRs serão disponibilizados em um banco de dado de pimenta do reino (Apêndice 1), contendo o protocolo executado para a obtenção destes dados.

5. DISCUSSÃO

5.1. MÉTODO DE DETECÇÃO DE GTRs

Até a presente data, não se conhece nenhum método descrito de identificação de GTRs relacionada a transcriptoma, a maioria dos trabalhos associados a esta temática desenvolve a caracterização destes por comparação entre genomas, e é descrito por diversos grupos taxonômicos, como bactérias (ALIMI *et al*., 2000; WILSON *et al*., 2007), plantas (LIN *et al*., 2010), cnidários (MILDE *et al*., 2009), teleósteos (YANG *et al*., 2013), metazoários (DOMAZET-LOSO; TAUTZ, 2003; JOHNSON; TSUTSUI, 2011; WISSLER *et al*., 2013) , primatas (TOLL-RIERA *et al*., 2009), entre outros.

O método subtrativo empregado neste trabalho se mostrou adequado a identificação e caracterização de GTRs em dados de RNA-seq de organismo não modelo. A importância do desenvolvimento deste método é associada a identificação de transcritos que sejam potenciais de respostas adaptativas, que possam ser utilizados na elaboração de cultivares, neste caso pimenta do reino, resistentes à doenças ou que desempenhem o papel funcional que auxilie no aumento produtivo de uma espécie economicamente importante.

A partir do método de identificação de GTRs podem ser desenvolvidos estudos em outros organismos não modelo, com a finalidade de obter transcritos de interesse e/ou intensificar informações sobre características moleculares de uma espécie com pouca oferta de informações a nível molecular e biológico.

5.2. MÉTODO APLICADO À INFORMAÇÃO BIOLÓGICA

A identificação de GTRs em dados de RNA-seq de pimenta do reino foi obtida através de métodos extrínsecos, desenvolvendo a caracterização funcional dos transcritos, e métodos intrínsecos proporcionando informações acerca dos potenciais GTRs identificados.

Dentre os métodos extrínsecos de avaliação, a análise por similaridade permitiu identificar uma maior anotação de transcritos entre os transcritos de pimenta do reino e o proteoma das espécies pertencentes ao grupo das angiospermas basais houve uma maior anotação dos transcritos em comparado com as espécies *A. thaliana* e *O. sativa*, isso evidencia que análise entre espécies taxonomicamente mais próximas é mais eficaz em relação as espécies que são caracterizadas como referência, mas isso não diminui a importância da utilização de genomas de referência para a analises por comparação,

pois são estes que possuem uma maior quantidade de informações a nível molecular e biológico.

Uma observação pertinente com relação a similaridade entre sequencia de *P. nigrum* e *A. fimbriata* está relacionada as duas espécies pertencerem a mesma ordem Piperales, o que proporcionou uma maior quantidade de transcritos únicos de *A. fimbriata* em comparado com as outras espécies do grupo das angiospermas basais (GORDO *et* al., 2012; BLISS *et al*., 2013).

A análise do conjunto de dados de pimenta do reino por homologia, não obteve muitos transcritos anotados comparado com a análise por similaridade entre sequencias, isso demonstra uma baixa frequência de genes ortólogos no transcriptoma de pimenta do reino. A relação quantitativa entre os dados obtidos por similaridade e homologia reflete na relação entre as sequencias que são estreitamente relacionadas e entre as principais linhagens filogenéticas, o que demonstra a importância da utilização de conjuntos de referência (TAUTZ; DOMAZET-LOSO, 2011; TRACHANA *et al*., 2011).

Segundo Trachana et al. (2011), a dificuldade da identificação de transcritos homólogos pode estar relacionada a qualidade do sequenciamento, devido a obtenção de *reads* com baixa cobertura, ou a complexidade da família de proteínas. Dessa maneira, a caracterização quando a presença de transcritos associados a família de proteínas auxiliou na detecção destes.

A baixa cobertura das sequencias equivale a *reads* pouco mapeados, que podem ser interpretados de maneira errônea, gerando assim falsos positivos (WANG *et* al., 2011; SIMS *et al*., 2014). A possibilidade de recuperação destes transcritos é nula, devido ser gerada por falhas no sequenciamento (LI *et al*., 2011; SIMS *et al*., 2014). A análise de falsos positivos permitiu identificar muitos transcritos com baixa cobertura, o que demonstra a importância da qualidade de sequenciamento na interpretação dos resultados.

A identificação de transcritos de pimenta do reino associados a ncRNA permitiu a identificação de genes de interesse, em regiões de transcriptoma que não são codificantes, mas que estão associado a modelação da expressão genica e/ou genes candidatos para elaboração de espécies tolerantes à doenças, de maneira a impulsionar estudos sobre as variações genicas. Porém mesmo desempenhando um papel especifico os ncRNA não caracterizam os GTRs, pois estes genes são responsáveis por desempenhar funções de uma espécie taxonomicamente restrita, diferentemente do

ncRNA que na maioria dos casos são originados de regiões de *splicing* (WILSON *et al*., 2005; MATTICK; MAKUNIN, 2006; KHALTURIN *et al*., 2009).

Para caracterizar um transcrito como GTRs é necessário, identificar se este transcrito possui a capacidade de codificar uma proteína, pois caso contrário não pode ser assim classificado (ARENDSEE *et al*., 2014). Isso indica a importância da detecção de ORFs para caracterizar um transcrito como GTRs, permitindo identificar 71 transcritos de pimenta do reino caracterizados como potenciais GTRs, ou seja, sequencias codificantes que não apresentam homologia com outras espécies (WILSON et al., 2005; ARENDSEE *et al*., 2014). Esta etapa está associada ao método intrínseco de detecção de GTRs.

A classificação de transcritos em GTRs envolve organização de dados e a busca exaustiva por funções proteicas, porém não possibilita a identificação dos GTRs quanto a sua função. Uma forma de identificar a função dos GTRs é avaliar os transcritos quanto a sua idade evolutiva (ARENDSEE *et al*., 2014).

A filoestratigrafia é uma técnica conhecida por traçar a origem de genes modernos de volta a seus fundadores específicos (DOMAZET-LOSO *et* al., 2007; DOMAZET-LOSO; TAUTZ, 2010; ARENDSEE *et al*., 2014). De maneira geral, esta técnica consiste em selecionar grupos taxonômicos da espécie foco por hierarquia, e para cada gene encontrar o táxon mais antigo ao qual ele possui homologia, ao reconhecer a característica mais antiga é possível reconhecer a função proteica deste gene (DOMAZET-LOSO *et* al., 2007; DOMAZET-LOSO; TAUTZ, 2010; ARENDSEE *et al*., 2014).

A técnica empregada para identificar a função de um GTRs corrobora com inúmeros estudos que descrevem que estes genes possuem uma relação evolutiva, podendo possuir funções relacionadas a respostas adaptativas específicas (WILSON *et al*., 2005, 2007; KHALTURIN *et al*., 2009; MILDE *et al*., 2009; DUARTE *et al*., 2013; YANG *et al*., 2013; ARENDSEE *et al*., 2014; TAUTZ; DOMAZET-LOSO, 2011).

A identificação do transcrito de pimenta do reino, de identificador UN01868, é um transcrito capaz de desenvolver uma função específica da espécie, pois possui características peculiares de genes que codificam proteínas. Apesar de ser um transcritos de apenas 507 nucleotídeos, sua função pode estar associada a processos de regulação genica importante. Silveira *et al*. (2013), descreve o gene Qua-Quine Amido (QQS) de *A. thaliana*, que possui comprimento de sequencia com 24 nucleotídeos, é resultado de uma variação epigenética, a qual alterações hereditária no processo de metilação do

DNA afetam a expressão do gene. De maneira que estas alterações podem ser herdadas por várias gerações.

O comprimento da sequencia de um transcrito pode influenciar o mesmo em ser caracterizado como codificante ou não, porém existe a possibilidade que transcritos bem curtos participem de processos de regulação genica e variações epigenéticas, que estão intimamente relacionadas a repostas evolutivas e adaptativas de uma espécie (WILSON *et al*., 2005; JOHNSON; TSUTSUI, 2011; SILVEIRA *et al*., 2013; YANG *et al*., 2013).

Acredita-se que com a identificação funcional dos GRTs sejam obtidos genes candidatos que possam ser utilizados na engenharia de cultivares valiosos e até mesmo na criação de fármacos e pesticidas (ARENDSEE *et al*., 2014), além de proporcionar um avanço sobre informações moleculares de espécies de interesse econômico e biotecnológico, como é o caso da pimenta do reino.

6. CONCLUSÃO

A eficácia do método de identificação de GTRs, aqui descrito, colabora para a caracterização destes genes, de forma que pode ser empregado em dados de RNA-seq em outros organismos não modelo. Sendo que o processo de comparação entre genomas é a etapa comum a todas as metodologias descritas na literatura sobre a identificação de GTRs.

Dentre os 35631 transcritos de pimenta do reino foram anotados 22661 transcritos a partir da comparação entre os proteomas de *A. thaliana*, *O. sativa* e as demais espécies do grupo das angiospermas basais. Por ortologia obteve-se 75 transcritos anotados, além de 12895 identificados com falsos positivos, 40 caracterizados como ncRNA, 6 identificados por similaridade com transcritos de *A. fimbriata*.

Foram identificados 71 transcritos com regiões codificadoras de proteínas, os quais são caracterizados como potenciais GTRs. Dentre eles, um transcrito apenas possui a possibilidade de estar associado a uma função proteica restrita a espécie.

Genes órfãos, genes taxonomicamente restritos (GTRs) e genes de linhagem específica são terminologias adotadas a um mesmo conceito, de modo que vale ressaltar a necessidade de uma caracterização universal, a fim de possibilitar a elaboração de estratégias que auxiliem de maneira precisa na identificação destes genes.

Estudos futuros podem identificar as funções proteicas destes genes, possibilitando o reconhecimento de genes candidatos que podem ser utilizados na produção de cultivares economicamente importantes, e até mesmo na produção de fármacos e pesticidas.

REFERÊNCIAS BIBLIOGRÁFICAS

AHMAD, N. H. *et al*. Biological role of *Piper nigrum* L. (Black pepper): A review. **Asian Pacific Journal of Tropical Biomedicine**, v.2, nº. 3, p. S1945-S1953, 2012.

ALBUQUERQUE, F. C. Podridão das raízes e do pé da pimenta do reino. **Circular do Instituto de Pesquisa e Experimentação Agropecuárias do Norte**, nº.8, 1964.

ALEXEYENKO, A. *et al*. Overview and comparison of ortholog databases. **Drug Discovery Today: Technologies**, v.3, nº. 2, p. 137-143, 2006.

ALIMI, J. P. *et al*. Reverse Transcriptase-Polymerase Chain Reaction Validation of 25 "Orphan" Genes from *Escherichia coli* K-12 MG1655. **Genome Research**, v.10, nº 7, p. 959-966, 2000.

ALTSCHUL, S. F. *et al*. Basic local alignment search tool. **Journal of Molecular Biology**, v.215, nº. 3, p. 403-410, 1990.

ARENDSEE, Z. W. *et al*. Coming of age: orphan genes in plants. **Trends in Plant Science**, v.19, nº. 11, p. 698-708, 2014.

BENCHIMOL, R. L. *et al*. Controle da fusariose em plantas de pimenta-do-reino com bactérias endofíticas: Sobrevivência e respostas morfofisiológicas. **Pesquisa Agropecuária Brasileira**, v.35, p. 1343-1348, 2000.

BESEMER, J. *et al*. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. **Nucleic Acids Research**, v.29, nº.12, p. 2607-2618, 2001.

BLISS, B. J. *et al*. Characterization of the basal angiosperm *Aristolochia fimbriata*: a potential experimental system for genetic studies. **BMC Plant Biology**, v.13, p. 13, 2013.

BRENT, M. R. Genome annotation past, present, and future: how to define an ORF at each locus. **Genome Research**, v.15, nº.12, p. 1777-1786, 2005.

CORRE, C.; CHALLIS, G. L. Heavy tools for genome mining. **Chemistry & Biology**, v.14, nº.1, p. 7-9. 2007.

DOMAZET-LOSO, T. *et al*. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. **Trends in Genetics**, v.23, nº.11, p. 533-539, 2007.

DOMAZET-LOSO, T.; TAUTZ, D. An evolutionary analysis of orphan genes in Drosophila. **Genome Research**, v.13, p. 2213-2219, 2003.

DOMAZET-LOSO, T.; TAUTZ, D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. **Nature**, v.468, nº. 7325, p. 815-818, 2010.

DUARTE, K. E. *et al*. Identificação e caracterização de genes órfãos ("no hits") de café (*Coffea Canephora*), envolvidos na resposta à seca. In: Simpósio de Pesquisa dos cafés do Brasil, 8., 2013, Salvador, BA. **Anais…** Brasília: Consócio Pesquisa Café, 2013.

EKBLOM, R.; GALINDO, J. Applications of next generation sequencing in molecular ecology of non-model organisms. **Heredity (Edinb)**, v.107, nº.1, p. 1-15, 2011.

FAO, Food Agriculture Organization of the United Nations: Statistics of Agricultural Production. Rome: FAO; 2015.

FINN, R. D. *et al*. Pfam: the protein families database. **Nucleic Acids Research**, v.42, Database issue, p. D222-230, 2014.

FISCHER, S. *et al*. Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. **Current Protocols in Bioinformatics**, capítulo 6, Unit 6.12, p. 11-19, 2011.

GISH, W.; D. J. STATES, Identification of protein coding regions by database similarity search. **Nature Genetics**, v.3, nº. 3, p. 266-272, 1993.

GORDO, S. M. *et al*. High-throughput sequencing of black pepper root transcriptome. **BMC Plant Biology**, v.12, nº. 168, 2012.

GRABHERR, M. G. *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. **Nature Biotechnology**, v. 29, nº.7, p. 644-652, 2011.

GUTIERREZ, R. M. *et al*. Alkaloids from Piper: A Review of its Phytochemistry and Pharmacology. **Mini reviews in medicinal chemistry**, v.13, nº.2, p. 163-193, 2013.

HORAN, K. *et al*. Annotating genes of known and unknown function by large-scale coexpression analysis. **Plant Physiology**, v.147, nº.1, p.41-57. 2008.

HU, L. *et al*. De novo assembly and characterization of fruit transcriptome in black pepper (*Piper nigrum*). **PLoS One**, v.10, nº.6, p. e0129822, 2015.

IBGE, Instituto Brasileiro de Geografia e Estatística. **Banco de dados**. Disponível em:<http://www.sidra.ibge.gov.br>. Acesso em: 2 de julho de 2015.

JARAMILLO, M. A.; MANOS, P. S. Phylogeny and patterns of floral diversity in the genus *Piper* (Piperaceae). **American journal of botany**, v.88, nº.4, p. 706-716, 2001.

JOHNSON, B. R.; TSUTSUI, N. D. Taxonomically restricted genes are associated with the evolution of sociality in the honey bee. **BMC Genomics**, v.12, nº. 164, 2011.

JOSE, J.; SHARMA, A. K. Chromosome studies in the genus *Piper* L. **Journal Indian Bot. Soc.** v.63, pág. 313–319, 1984.

JOY, N. *et al*. De novo transcriptome sequencing reveals a considerable bias in the incidence of simple sequence repeats towards the downstream of 'Pre-miRNAs' of black pepper. **PLoS One**, v.8, nº.3, p. e56694, 2013.

JUSTO, S. C.; SILVA, C. M. *Piper methysticum* G. Foster (Kava-kava): Uma abordagem geral. **Revista Eletrônica de Farmácia**, v.1, pág. 73-82, 2008.

KAESSMANN, H. Origins, evolution, and phenotypic impact of new genes. **Genome Research**, v. 20, nº.10, p. 1313-1326, 2010.

KHALTURIN, K. *et al*. More than just orphans: are taxonomically-restricted genes important in evolution? **Trends in Genetics**, v.25, nº.9, p. 404-413, 2009.

LEMOS, O. F. D., **Mutagênese e tecnologia in vitro no melhoramento genético da pimenta-do-reino (*Piper nigrum* L.).** 2003, 182 f., Tese de Doutorado. Piracicaba - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo. São Paulo, 2003.

LEMOS, O. F. D., Conservação e melhoramento genético da pimenteira-do-reino (Piper nigrum L.) associado às técnicas de biotecnologia. **Documentos 375**, Embrapa Amazônia Oriental, 2011.

LI, L. *et al*. OrthoMCL: identification of ortholog groups for eukaryotic genomes. **Genome Research**, v.13, nº.9, p. 2178-2189, 2003.

LI, W. ORF phage display to identify cellular proteins with different functions. **Methods**, v.58, nº.1, p. 2-9, 2012.

LI, Y. *et al*. Low-coverage sequencing: implications for design of complex trait association studies. **Genome Research**, v.21, nº.6, p. 940-951, 2011.

LIN, H., *et al*. Comparative analyses reveal distinct sets of lineage-specific genes within Arabidopsis thaliana. **BMC Evolutionary Biology**, v.10, nº. 41, 2010.

LISTER, R. *et al*. Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. **Current Opinion in Plant Biology**, v.12, nº.2, p. 107-118, 2009.

LOURINHO, M. P. *et al*. Conjuntura da pimenta-do-reino no mercado nacional e na região norte do Brasil. **Enciclopédia Biosfera**, v.10, nº.8, p. 1016-1031, 2014.

MAGEVSKI, G. *et al*. Propagação vegetativa de espécies silvestres do gênero *Piper*, com potencial para uso como porta enxertos em pimenta-do-reino (*Piper nigrun*). **Rev. Bras. Pl. Med.**, v.13, p. 559-563, 2011.

MATTICK, J. S.; MAKUNIN, I. V. Non-coding RNA. **Human Molecular Genetics**, v. 15 especial, nº. 1, p. R17-29, 2006.

MILDE, S. *et al*. Characterization of taxonomically restricted genes in a phylum-restricted cell type. **Genome Biology**, v.10, nº.1, p. R8, 2009.

MITCHELL, A. *et al*. The InterPro protein families database: the classification resource after 15 years. **Nucleic Acids Research**, v.43, Database issue, p. D213-221, 2015.

MOROZOVA, O. *et al*. Applications of new sequencing technologies for transcriptome analysis. **Annu Rev Genomics Hum Genet**, v.10, p. 135-151, 2009.

NAIR, R. R. *et al*. Polyploid in a cultivar od black pepper (Piper nigrum L.) and its open plollinated progenies. **Cytologia**, v.58, p. 27-31, 1993.

NAWROCKI, E. P. *et al*. Rfam 12.0: updates to the RNA families database. **Nucleic Acids Research**, v.43, Database issue, pág. D130-137. 2015.

OLIVER, S. G. *et al*. The complete DNA sequence of yeast chromosome III. **Nature**, v.357, nº. 6373, p. 38-46, 1992.

PARMAR, V. S. *et al*. Phytochemistry of the genus *Piper*. **Phytochemistry**, v.46, nº.4, p. 597-673, 1997.

PHILIP, V. J., *et al*. Micropropagation of black pepper (*Piper nigrum* Linn.) through shoot tip cultures. **Plant Cell Reports**, v.12, p. 41-44, 1992.

RAZQUIN, A. C. **Origin of genes with unresolved ancestry: Analysis of orphan genes in *H. sapiens, D. melanogaster* and *S. cerevisiae***. 2013. 56 f., Tese de Mestrado em Proteômica e Bioinformática. Université de Genève. 2013.

RUEPP, A.; MEWES, H. W. Prediction and classification of protein functions. **Drug Discovery Today: Technologies**, v.3, nº.2, p. 145-151, 2006.

SILVEIRA, A. B. *et al*. Extensive natural epigenetic variation at a de novo originated gene. **PLoS Genetics**, v.9, nº.4, p. e1003437, 2013.

SIMS, D. *et al*. Sequencing depth and coverage: key considerations in genomic analyses. **Nature Reviews Genetics**, v.15, nº.2, p. 121-132, 2014.

STRICKLER, S. R. *et al*. Designing a transcriptome next-generation sequencing project for a nonmodel plant species. **American Journal of Botany**, v.99, nº.2, p. 257-266, 2012.

STUDER, R. A.; ROBINSON-RECHAVI, M. How confident can we be that orthologs are similar, but paralogs differ?, **Trends Genetics**, v.25, nº.5, p. 210-216, 2009.

TANG, S. *et al*. Identification of protein coding regions in RNA transcripts. **Nucleic Acids Research**, v.43, nº.12, p. e78, 2015.

TAUTZ, D.; DOMAZET-LOSO, E T. The evolutionary origin of orphan genes. **Nature Reviews Genetics**, v.12, nº.10, p. 692-702, 2011.

TOLL-RIERA, M. *et al*. Origin of primate orphan genes: a comparative genomics approach. **Molecular Biology and Evolution**, v.26, nº.3, p. 603-61, 2009.

TRACHANA, K. *et al*. Orthology prediction methods: a quality assessment using curated protein families. **Bioessays**, v.33, nº.10, p. 769-780, 2011.

VAN BEL, M. *et al*. TRAPID: an efficient online tool for the functional and comparative analysis of de novo RNA-Seq transcriptomes. **Genome Biology**, v.14, nº.12, p. R134, 2013.

VAN DER HEIJDEN, R. T. *et al*. Orthology prediction at scalable resolution by phylogenetic tree analysis. **BMC Bioinformatics**, v.8, nº. 83, 2007.

VANAJA, T. *et al.* Development of a promising interspecific hybrid in black pepper (*Piper nigrum* L.) for Phytophthora foot rot resistance. **Euphytica**, v.161, nº.3, p. 437-445, 2007.

WANG, Y. *et al*. Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens. **BMC Bioinformatics**, v.12 Suppl 10, p. S5, 2011.

WHISSTOCK, J. C.; LESK, A. M. Prediction of protein function from protein sequence and structure. **Quarterly Reviews of Biophysics**, v.36, nº.3, p. 307-340, 2003.

WILSON, G. A. *et al*. Orphans as taxonomically restricted and ecologically important genes. **Microbiology**, v.151, Part.8, p. 2499-2501, 2005.

WILSON, G. A. *et al*. Large-scale comparative genomic ranking of taxonomically restricted genes (TRGs) in bacterial and archaeal genomes. **PLoS One**, v.2, nº.3, p. e324, 2007.

WISSLER, L. *et al.* Mechanisms and dynamics of orphan gene emergence in insect genomes. **Genome Biology and Evolution**, v.5, nº.2, p. 439-455, 2013.

WU, D. D.; ZHANG, Y. P. Evolution and function of de novo originated genes. **Molecular Phylogenetics and Evolution**, v.67, nº.2, p. 541-545, 2013.

YANG, L. *et al*. Genome-wide identification, characterization, and expression analysis of lineage-specific genes within zebrafish. **BMC Genomics**, v.14, nº. 65, 2013.

ZHANG, J., Evolution by gene duplication: an update. **Trends in Ecology & Evolution**, v.18, nº.6, p. 292-298, 2003.

APÊNDICE

APÊNDICE 1 – Artigo de banco de dados de pimenta do reino.

BPD: um banco de dados de pimento do reino

RESUMO

O Black Pepper Database (BPD) é uma base de dados científica devidamente curada, que visa o acesso de forma eficiente a dados de transcriptoma de raiz e folha de pimenta do reino (*Piper nigrum* L.). O banco é baseado em dados de raiz e folha que se encontram disponíveis para download em Sequence Read Archive (SRA) no banco de dados do National Center for Biotechnology Information (NCBI). Os dados foram gerados, a partir de montagem *de novo*, utilizando os dois conjuntos, sendo um utilizando dados da plataforma SOLiD e o outro utilizando a combinação de dados das plataformas SOLiD e Illumina. O BPD fornece como ferramenta a Basic Local Alignment Search Tool (BLAST), além de informações sobre proteínas preditas, dados de microssatélites, informações sobre banco de germoplasma no Brasil, publicações relacionadas a espécie de estudo e download dos dados. O foco do BPD é a integração de dados que busquem fornecer informações sobre a base molecular e funcional dos transcritos, de maneira a estabelecer estratégias de melhoramento da espécie.

Palavras – chave: Black pepper, transcriptoma, banco de dados, melhoramento genético.

1. INTRODUÇÃO

Dentro da família Piperaceae, o gênero *Piper*, é o mais diverso apresentando distribuição pantropical de suas espécies (JARAMILLO; MANOS, 2001). Estudos filogenéticos classificam o gênero como um integrante do grupo das angiospermas basais (JARAMILLO; MANOS, 2001).

A pimenta do reino (*Piper nigrum* L.) é uma das espécies mais importante do gênero *Piper* devido se tratar de uma das especiarias mais comercializadas no mundo (JOY *et al*., 2007; AHMAD *et al*., 2012). No nível de produção mundial estima-se que sejam produzidos mais de 400 mil toneladas ao ano, dos quais 79,4% são produzidos na Ásia e 16,2% nas Américas, sendo o Vietnã, Indonésia, Índia, Brasil e China os maiores produtores (FAO, 2015).

A valorização da pimenta do reino no mercado internacional varia devido a oscilação dos preços, podendo desestimular o cultivo (LEMOS *et al*., 2011). Entretanto, o que ocasiona sérios prejuízos na produção dos cultivares é a pouca variabilidade genética entre as espécies, que está associado ao método de propagação vegetativa que

muitas vezes é utilizado, o que implica na produção de cultivares vulneráveis à doenças (LEMOS, 2003; LEMOS *et al*., 2011; GORDO *et al*., 2012).

No Brasil, os patógenos *Fusarium solani* f. sp. *piperis* e *Phytophtora capsici* é uma das patologias que mais acomete os cultivares de pimenta do reino, as estratégias de melhoramento clássico não se mostraram eficientes para resolver a vulnerabilidade a estas doenças (CHU *et al*., 2006; MAJU; SONIYA, 2012).

Na busca por cultivares resistentes, a utilização da biotecnologia vegetal surge como estratégia, visando a ampliação do conhecimento e tendo como papel principal o ganho da produtividade dos cultivares de pimenta do reino (MAJU; SONIYA, 2012; GORDO *et al*., 2012).

O ganho de produtividade dos cultivares de pimenta do reino é importante para a economia e agronomia, porém o entendimento limitado da via da interação entre *P. nigrum* e seu patógeno tem restringido a produção de cultivar com maiores níveis de tolerância (LEMOS *et al*., 2011; MAJU; SONIYA, 2012). A utilização de tecnologia de sequenciamento de nova geração (NGS) é uma eficiente ferramenta para a obtenção de informações genéticas de espécies não modelo (EKBLOM; GALINDO, 2011).

A pimenta do reino é uma espécie vegetal que não possui seu genoma completamente descrito, denominando-a espécie não modelo (GORDO *et al*., 2012; JOY *et al*., 2013). O sequenciamento de espécies não modelo a partir do transcriptoma é uma maneira eficiente de obter dados de tecidos específicos (STRICKLER *et al*., 2012).

A partir da obtenção do transcriptoma é possível caracterizar os constituintes moleculares e funcionais de células e tecidos, proporcionando a catalogação de transcritos, a determinação da estrutura de transcrição de genes e quantificação das mudanças nos níveis de expressão de cada transcrição durante o desenvolvimento e sob diferentes condições (WANG *et al*., 2009; STRICKLER *et al*., 2012). No caso de espécies não modelo o sequenciamento *de novo* é mais utilizado, devido não possui dados de referencia adequados para análise (STRICKLER *et al*., 2012).

Estudos de transcriptoma de pimenta do reino tem se mostrado eficientes para aumentar a caracterização genética desta espécie (GORDO *et al*., 2012; JOY *et al*., 2013; HU *et al*., 2015). Para integrar e divulgar dados de transcriptoma de pimenta do reino foi construído um banco de dados de pimenta do reino com o objetivo de servir como plataforma de fonte de informações genéticas que poderão ser exploradas para descrição molecular ou acelerar a pesquisa promissora de melhoramento biotecnológico, abordando a caracterização funcional de genes específicos em pimenta do reino.

O Black pepper Database (BPD) é uma base de dados científica devidamente curada, que visa o acesso de forma eficiente a dados de transcriptoma de raiz e folha de pimenta do reino. Dois conjuntos de dados sequenciados foram gerados, a partir de montagem *de novo*, utilizando as plataformas SOLiD (GORDO et al., 2012) e Illumina (JOY et al., 2013), obtidos a partir da espécie *P. nigrum*.

O foco principal do BPD é a integração de dados que busquem fornecer informações sobre a base molecular, de maneira a estabelecer estratégias de melhoramento genético da espécie.

2. RECURSOS DE DADOS

2.1. SEQUENCIAMENTO DO TRANSCRIPTOMA E MONTAGEM *DE NOVO*

Os dados de transcriptoma utilizados foram são descritos nos estudos de Gordo *et al*. (2012), possuindo 3.6Gb (número de acesso SRX104901) e Joy *et al*. (2013), contendo 5Gb (número de acesso SRX119532), referente a transcriptoma de raiz e folha, respectivamente. Os conjuntos foram gerados utilizando dados das plataformas SOLiD e pela abordagem híbrida, utilizando dados das plataformas SOLiD e Illumina, respectivamente.

Para o primeiro conjunto de dados (Dados I) foi utilizado a plataforma SOLiD foi desenvolvido através do método de múltiplos *k-mers* utilizando as ferramentas Velvet e Oases (GORDO *et al*., 2012).

Para o segundo conjunto (Dados II) foi realizado uma combinação de dados entre as plataformas SOLiD e Illumina. A integração de dados de diferentes plataformas de NGS otimiza e aumenta a montagem de *contigs*, que por conseguinte, aumenta a quantidade de transcritos anotados (SALMELA, 2010; WANG *et al*., 2012). Neste processo foi realizada uma abordagem clássica de interação, proposta por Wang *et al*. (2012), onde as montagens são feitas separadamente e agrupadas em níveis de contigs e a nossa abordagem, onde a montagem é feita em nível de *reads* com todos os dados já agrupados, todos em *color-space*.

No caso das plataformas SOLiD e Illumina o resultado do sequenciamento são diferentes, sendo *color-space* e *base-space*, respectivamente (SURGET-GROBA; MONTOYA-BURGOS, 2010; WANG *et al*., 2012).

A integração de dados de diferentes plataformas de NGS otimizar aumentar a montagem de *contigs* e por conseguinte ampliar a quantidade de transcritos anotados.

Para este conjunto de dados foi utilizado as metodologias de abordagem clássica (WANG *et al*., 2012).

A metodologia de abordagem clássica e montagem das *reads* de Illumina e SOLiD são realizadas separadamente, na qual as etapas de pré e pós processamento das *reads* da plataforma SOLiD foi utilizado os *scripts* "denovo_preprocessor_solid.pl" e "denovo_postprocessor_solid.pl", respectivamente .

Para o procedimento de montagem foram usadas as ferramentas Velvet, versão 1.2.10, e Oases, version 0.8.08., utilizando o método de múltiplos *k-mers* (SURGET-GROBA; MONTOYA-BURGOS, 2010), em seguida, os dados foram combinados na ferramenta CD-HIT-EST, versão 4.6. (LI; GODZIK, 2006).

O método de abordagem híbrida utilizando diferentes plataformas de NGS possui o intuito de melhorar o conjunto do transcriptoma, a fim de amplificar a predição de genes no transcriptoma. Através da combinação dos dados de diferentes plataformas de sequenciamento estima-se que haja a correção de viéses de cada plataforma, além do aumento na cobertura e profundidade do sequenciamento (WANG *et al*., 2012). Entretanto o resultado do sequenciamento das plataformas SOLiD e Illumina são diferentes, dessa forma, como descrito por Salmella (2010), a interação do conjunto de dados de SOLiD com a plataforma Illumina é mais eficiente em *color-space*, permitindo um aumento da confiabilidade da análise.

No procedimento de pré processamento a montagem dos transcritos os dados de Illumina foram convertidos para o formato *color-space* e posteriormente agrupados com os dados de SOLiD, convertendo ambos para o formato double-encoded, utilizando o *script* caseiro "denovo_preprocessor_illumina.pl". Em seguida, as *reads* de SOLiD foram convertidas em formato double-encoded, neste processo foi aplicado o *script* "denovo_preprocessor_solid.pl", da em presa Life Technologies. Após as conversões as leituras de SOLiD e Illumina, ambas em formato *color-space*, foram combinadas e submetidas as ferramentas Velvet, e Oases, com método de múltiplos *k-mers*, o resultado deste procedimento foi convertido em *base-space* a partir do *script* "denovo_postprocessor_solid.pl", posteriormente os dados foram submetidos a ferramenta CD-HIT-EST.

Após o processo de montagem, os dois conjuntos de dados, foram combinados em um único conjunto com a ferramenta CAP3, versão 8.6.13, de forma a remover redundâncias e aumentar o tamanho do transcrito, restando um único conjunto de dados.

Como resultado final obteve-se dois conjuntos de dados, o primeiro (Dados I) de transcriptoma de raiz utilizando dados da ferramenta SOLiD, proveniente do trabalho de Gordo et al. (2012) e o segundo (Dados II) de transcriptoma de raiz e folha unindo dados de SOLiD e Illumina em um único conjunto, resultante da abordagem híbrida (Tabela 1).

Tabela 1 – Dados obtidos do sequenciamento utilizando a plataforma SOLiD e utilizando as plataformas SOLiD e Illumina.

Dados brutos		
	Dados I (SOLiD)	**Dados II (Illumina e SOLiD)**
Montagem do transcrito	Multiplos *k-mers*	Multiplos *k-mers*
Total do nº de read	13300000	68372366
Total do nº de contigs	22363	233109
Total do nº de unigenes	10338	60645
Tamanho médio do contig (bp)	1314	1172
N50 (pb)	168	1653
Proteínas preditas	4472	60107

pb: pares de base.

2.2. PROTEÍNAS PREDITAS NO TRANSCRIPTOMA DE PIMENTO DO REINO

A identificação das proteínas preditas para do conjunto de dados I foi realizada utilizando a ferramenta FrameD e BLASTP (E-value $1e^{-05}$), para busca por homologia com as plantas encontradas no banco de dados PlantGBD e nr-viridiplantae (NCBI), obteve-se 4472 proteínas preditas, sendo aproximadamente 52% das proteínas preditas são homólogas a sequencias proteicas de *Arabidopsis thaliana*, do banco de dados do NCBI. As espécies *Popullus trichocarpa* (54.38%), *Aristolochia fimbriata* (54.02%) e *Vitis vinífera* (53.93%) obtiveram maior percentual de proteínas preditas por homologia com os dados de P. nigrum em relação ao banco de dados de outras espécies (Tabela 2).

Além da predição de proteínas foi realizada a anotação funcional dos dados I baseado na ferramenta BLASTX contra o banco de dados de nr-viriplantae, na qual o programa BLAST2GO foi utilizado para identificar as funções putativas dos transcritos, sendo que por comparação, 3055 unigenes foram anotados funcionalmente com o banco de dados do Gene Ontology (GO) e 9664 unigenes anotados com banco de dados de *A. thaliana* (GORDO et al., 2012).

O processo de predição de proteínas do conjunto de dados II foi realizado utilizando a ferramenta EvitentialGene, para remoção de redundâncias usando as

informações dos aminoácidos para estabelecer as melhores sequencias codificantes de proteínas, além de executar um BLAST para identificar as sequencias homólogas (NAKASUGI et al., 2014), o que resultou em 60107 proteínas preditas.

Para identificar a homologia dos transcritos do conjunto de dados II foi realizado a comparação utilizando a ferramenta BLASTX (e-value $1e^{-5}$) com proteomas preditos das espécies *Arabidopsis thaliana*, *Oryza sativa*, *Aristolochia fimbriata*, *Popullus trichocarpa* e *Vitis vinifera* (Tabela 2).

Tabela 2 – Percentagem de proteínas preditas por homologia para cada conjuntos de dados.

Protein database	**% with homology**	
	Data I	Data II
Arabidopsis thaliana	51.57	72.55
Aristolochia fimbriata	54.02	73.70
Popullus trichocarpa	54.38	74.54
Vitis vinifera	53.93	74.55

2.3. DETECÇÃO DE MICROSSATÉLITES

A detecção de microssatélites foi realizada a partir dos dados I utilizando o script MISA, para identificar sequencias de repetição simples (SSRs). A partir do número de *contigs* (22363) foram identificados 168 repetições de di-, tri- e tetranucleotídeo (GORDO et al., 2012).

3. CONSTRUÇÃO DO BANCO DE DADOS

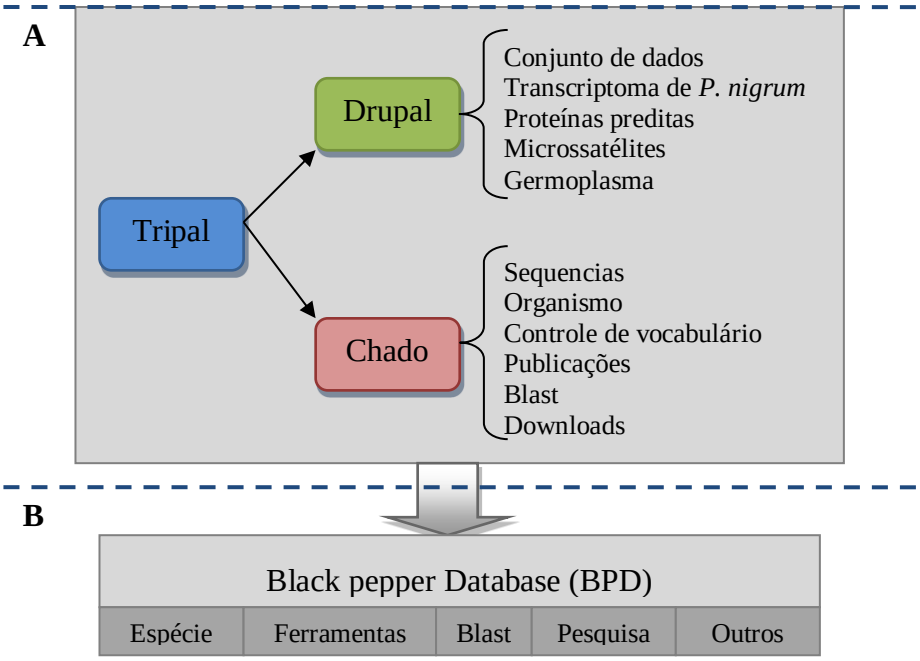
3.1. DESCRIÇÃO DO SISTEMA

O banco de dados foi construído com o auxilio do Tripal, que consiste em uma ferramenta que proporciona a criação de um banco de dados *online*, que integra dados biológicos com vários sistemas para o desenvolvimento de análises de genomas de plantas (FICKLIN et al., 2011; SANDERSON et al., 2013). Este sistema também é utilizado por outros bancos de dados como The Banana Genome Hub (DROC et al., 2013) e CottonGen (YU et al., 2014).

O Tripal é integrado por dois sistemas o Drupal e o Chado, esses sistemas foram desenvolvidos para ajudar no processo de instalação do *website* (Figura 1-A). O sistema Drupal inclui várias extensões que auxiliam na organização dos dados, de forma a simplificar a construção e gestão do *website*. Por sua vez, o sistema Chado auxilia no armazenamento dos dados biológicos, que coordena a construção e a distribuição de

várias ferramentas que suportam o armazenamento, mineração e visualização dos dados genômicos, estando associado com o Chado é um produto associado ao projeto Generic Model Organism Database (GMOD) (FICKLIN et al., 2011; SANDERSON et al., 2013).

Figura 1 – Ilustração esquemática do sistema de BDP. (**A**) Organização do sistema do servidor da *web*. (**B**) Interface da *web* do sistema BPD.



3.1.1. Blast

A ferramenta BLAST fornecer uma interface gráfica de usuário através dos formulários *web* (Figura 1-B), munido de um módulo associado ao Chado que viabiliza o alinhamento de sequencias de pimenta do reino com outras espécies. O NCBI nr e o ExPASy Swiss-Prot são os bancos de dados de referencia para a realização dos blasts. A partir desta ferramenta o usuário pode executar pesquisas de similaridade utilizando as variações de BLAST (BLASTn, BLASTp, BLASTx) (ALTSCHUL et al., 1990; FICKLIN et al., 2011; SANDERSON et al., 2013).

Para executar o BLAST devem ser carregadas as sequencias de nucleotídeo ou aminoácido em formato FASTA, em seguida deve-se escolher o banco de dados que será utilizado como base para o alinhamento, posteriormente é necessário definir os parâmetros do alinhamento. Ao clicar no ícone do BLAST escolhido é executado o processamento e o resultado da pesquisa mostra o alinhamento entre a sequencia fornecida juntamente com a sequencia do transcriptoma de *P. nigrum* produzidos, além

do ordenamento de acordo com o valor esperado (FICKLIN et al., 2011; SANDERSON et al., 2013).

3.1.2. Germoplasma

A espécie *P. nigrum* é uma planta de possui diversos cultivares, dentre os principais são Cingapura, Guajarina, Bragantina, APRA, Kuthiravally, Kottanadan e Iaçará (LEMOS et al., 2011). A elaboração de bancos de germoplasma contribui para pesquisas de melhoramento genético, de maneira a desenvolver espécimes referencia e proporcionar o conhecimento molecular acerca da mesma (ALBUQUERQUE et al., 1997; DE SOUZA et al., 2002).

A EMBRAPA (Empresa Brasileira de Pesquisa Agropecuária) possui um banco de germoplasma de pimenta do reino (http://plataformarg.cenargen.embrapa.br/rede-vegetal/bancos-ativos-degermoplasma/hortalicas/pimenta-do-reino).

3.1.3. Publicações

O banco de dados de pimenta do reino possui informações sobre diversas publicações sobre a espécie *P. nigrum*. Algumas publicações foram importadas automaticamente a partir do banco de dados do PubMed, enquanto que outras foram adicionadas manualmente devido estarem em outros bancos de dados (SANDERSON et al., 2013). Além disso, o banco de dados de pimenta do reino integra informações sobre protocolos de cultivo, trabalhos publicados em anais de congressos, capítulos de livros e teses. O banco de dados possui um total de 400 publicações, com publicações a partir do ano de 1966.

Para realizar a pesquisa por publicações no banco de dados deve-se utilizar de palavras-chave e posteriormente escolher a que campo pode estar relacionado, podendo ser ao título, resumo, autores e/ou nome da revista. Como resultado da pesquisa será obtido informações sobre as publicações, como título do trabalho, resumo, citação e link externo para visualizar o artigo completo (FICKLIN et al., 2011; SANDERSON et al., 2013).

3.1.4. Downloads

Nesta sessão serão disponibilizados os dois conjuntos de dados de transcriptoma de pimenta do reino, marcadores microssatélites, informações sobre as proteínas preditas e anotação funcional dos conjuntos de dados.

4. DISCUSSÃO

Devido ao grande número de projetos de sequenciamento muitos dados são disponibilizados na *web* sem a devida avaliação da qualidade, e como alternativa para esta problemática o BPD disponibiliza dados de transcriptoma de pimenta do reino que foram avaliados de maneira exaustiva, possibilitando o acesso rápido a conjuntos de dados curados.

O BPD além de ser um banco de dados que compõe informações importantes sobre a espécie, seus dados são interligados, possibilitando ao usuário o acesso fácil aos conteúdos, que envolvem são só dados moleculares, mas também informações sobre o cultivo da espécie.

O sistema de dados permite o acesso a dois conjuntos de dados de transcriptoma de pimenta do reino, além de informações sobre predição de proteínas, anotação funcional dos transcritos, marcadores moleculares, do tipo microssatélites, e banco de germoplasma.

O acesso a informações sobre pimenta do reino fornece subsídios para estudos, principalmente, relacionados ao caráter molecular, a fim de aumentar a compreensão sobre o conhecimento da espécie (GORDO et al., 2012; JOY et al., 2013; HU et al., 2015). A homogeneidade genética de cultivares comerciais amplia a vulnerabilidade a estresses bióticos e abióticos, dessa maneira, grande parte dos esforços em conhecer a base genômica desta espécie está aliada a estudos de conservação e melhoramento genético (NASCIMENTO et al., 2009; LEMOS et al., 2011).

As técnicas de biotecnologia são ferramentas promissoras que abrem perspectivas de programas de melhoramento de cultivares de pimenta do reino. Isso inclui compreensão da diversidade genética entre os cultivares (PRADEEPKUMAR et al., 2003; GAIA et al., 2005; SEN et al., 2010), e identificação de genes de interesse que sejam associados a estresses bióticos e abióticos.

A busca por informações moleculares que viabilizem a criação de explantes livres de patógenos, que fornecem subsídios sobre o uso de marcadores moleculares

possibilitando a análise sobre a diversidade dos cultivares (PRADEEPKUMAR et al., 2003; SEN et al., 2010), além de proporcionar a identificação de genes de interesse que sejam responsáveis pela expressão de genótipos que produzam culturas de pimenta do reino tolerantes a impactos ambientais (LEMOS, 2003; LEMOS et al., 2011).

O papel do melhoramento genético está voltado, principalmente, para a produção de culturas tolerantes a efeitos bióticos e abióticos, mas também fomenta a criação de bancos de germoplasma que contenham amostras de espécimes referência que concedem a conservação ao patrimônio genético (LEMOS et al., 2011; DE SOUZA et al., 2002).

5. CONCLUSÕES E PERSPECTIVAS

O BPD disponibiliza dados transcriptômicos e informações sobre pimenta do reino, além de ferramentas de análise que viabilizem a pesquisa por similaridade. Poderão ser incorporados mais dados à medida que essas informações sejam maximizadas, otimizando o crescimento molecular e genético básico e identificação de marcadores microssatélites ou genes de interesse biotecnológico.

O sistema do banco de dados também possibilita a interação do usuário com o administrador, na qual o usuário pode auxiliar no melhoramento das informações disponíveis através de um *feedback*.

Perspectivas futuras, para este banco de dados, estão relacionadas ao aumento de informações de carater molecular e funcional da espécie, possibilitando a sua utilização como plataforma para catalogar famílias gênicas de interesse, estudo evolução e estabelecimento de estratégias de melhoramento genético para minimizar os efeitos associados a estresses bióticos e abióticos em pimenta do reino.

REFERÊNCIAS BIBLIOGRÁFICAS

AHMAD, N. H. *et al.* Biological role of *Piper nigrum* L. (Black pepper): A review. **Asian Pacific Journal of Tropical Biomedicine**, v.2, nº. 3, p. S1945-S1953, 2012.

ALBUQUERQUE, F. C. de *et al*. Comportamento de germoplasma de pimenta-do-reino em relação a produtividade e a resistência a doença em regiões da Amazônia brasileira. **Relatório Técnico Anual do Centro de Pesquisa Agroflorestal da Amazônia Oriental**. Centro de Pesquisa Agroflorestal da Amazônia Oriental, Geração de tecnologia agroindustrial para o desenvolvimento do trópico úmido: síntese dos resultados do projeto, 1997.

ALTSCHUL, S. F. *et al*. Basic local alignment search tool. **Journal of Molecular Biology**, v.215, nº. 3, p. 403-410, 1990.

CHU, E. Y. *et al*. **A cultura da pimenta-do-reino**. 2ª edição, Brasília: DF, Embrapa Amazônia Oriental, 2006.

DE SOUZA, A. d. G. C. *et al*. The cupuaçu genetic improvement program at Embrapa Informação Técnológica. **Crop Breeding and Applied Biotechnology**, v.2, p. 471-478, 2002.

DROC, G. *et al*. The banana genome hub. **Database (Oxford)**, v. 2013, 2013.

EKBLOM, R.; GALINDO, J. Applications of next generation sequencing in molecular ecology of non-model organisms. **Heredity (Edinb)**, v.107, nº.1, p. 1-15, 2011.

FAO, Food Agriculture Organization of the United Nations: Statistics of Agricultural Production. Rome: FAO; 2015.

FICKLIN, S. P. *et al*. Tripal: a construction toolkit for online genome databases. **Database (Oxford)**, v. 2011, 2011.

GAIA, J. M. D. *et al*. Diversidade e similaridade genéticas em clones de pimenta-do-reino. **Horticultura Brasileira**, v. 23, p. 221-227, 2005.

GORDO, S. M. *et al*. High-throughput sequencing of black pepper root transcriptome. **BMC Plant Biology**, v.12, nº.168, 2012.

HU, L. *et al*. De novo assembly and characterization of fruit transcriptome in black pepper (*Piper nigrum*). **PLoS One**, v. 10, nº. 6, 2015.

JARAMILLO, M. A.; MANOS, P. S. Phylogeny and patterns of floral diversity in the genus *Piper* (Piperaceae). **American journal of botany**, v. 88, p. 706-716, 2001.

JOY, N. *et al*. A preliminary assessment of genetic relationships among agronomically important cultivars of black pepper. **BMC Genetics**, v. 8, nº.42, 2007.

JOY, N. *et al*. De novo transcriptome sequencing reveals a considerable bias in the incidence of simple sequence repeats towards the downstream of 'Pre-miRNAs' of black pepper. **PLoS One**, v. 8, nº.3, p. e56694, 2013.

LEMOS, O. F. D., **Mutagênese e tecnologia in vitro no melhoramento genético da pimenta-do-reino (*Piper nigrum* L.).** 2003, 182 f., Tese de Doutorado. Piracicaba - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo. São Paulo, 2003.

LEMOS, O. F. D., Conservação e melhoramento genético da pimenteira-do-reino (*Piper nigrum* L.) associado às técnicas de biotecnologia. **Documentos 375**, Embrapa Amazônia Oriental, 2011.

LI, W.; GODZIK, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. **Bioinformatics**, v. 22, nº. 13, p. 1658-1659, 2006.

MAJU, T. T.; SONIYA, E. V. In vitro regeneration system for multiplication and transformation in Piper nigrum L. **Int. J. Med. Arom. Plants**, v. 2, nº. 1, p. 178-184, 2012.

NAKASUGI, K. *et al*. Combining transcriptome assemblies from multiple de novo assemblers in the allo-tetraploid plant Nicotiana benthamiana. **PLoS One**, v. 9, nº. 3, p. e91776, 2014.

NASCIMENTO, S. *et al*. Identifying sequences potentially related to resistance response of *Piper tuberculatum* to *Fusarium solani* f. sp. *piperis* by suppression subtractive hybridization. **Protein & Peptide Letters**, v. 16, nº. 12, p. 1429-1434, 2009.

PRADEEPKUMAR, J. L. *et al*. Analysis of genetic diversity in *Piper nigrum* L. using RAPD markers. **Genetic Resources and Crop Evolution**, v. 50, nº. 3, p. 469-475, 2003.

SALMELA, L. Correction of sequencing errors in a mixed set of reads. **Bioinformatics**, v. 26, nº. 10, p. 1284-1290, 2010.

SANDERSON, L. A. *et al*. Tripal v1.1: a standards-based toolkit for construction of online genetic and genomic databases. **Database (Oxford)**, v. 2013, 2013.

SEN, S. *et al*. Genetic diversity analysis in *Piper* species (Piperaceae) using RAPD markers. **Mol Biotechnol**, v. 46, nº. 1, p. 72-79, 2010.

STRICKLER, S. R. *et al*. Designing a transcriptome next-generation sequencing project for a non model plant species. **American Journal of Botany**, v.99, nº.2, p. 257-266, 2012.

SURGET-GROBA, Y.; MONTOYA-BURGOS, J. I. Optimization of de novo transcriptome assembly from next-generation sequencing data. **Genome Research**, v. 20, nº. 10, p. 1432-1440, 2010.

WANG, Z. *et al*. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, v. 10, p. 57-63, 2009.

WANG, Y. *et al*. Optimizing hybrid assembly of next-generation sequence data from *Enterococcus faecium*: a microbe with highly divergent genome. **BMC Systems Biology**, v. 6, Suppl 3, 2012.

YU, J. *et al*. CottonGen: a genomics, genetics and breeding database for cotton research. **Nucleic Acids Research**, v. 42, p. D1229-1236, 2014.