

UNIVERSIDADE FEDERAL DE SÃO CARLOS
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA EVOLUTIVA E BIOLOGIA
MOLECULAR
DEPARTAMENTO DE GENÉTICA E EVOLUÇÃO

ALUNO:
ANDRÉS SANTIAGO ORTIZ MORAZÁN

**ANÁLISES BIOINFORMÁTICAS DE PROTEÍNAS DE RESISTÊNCIA DO TIPO
NUCLEOTIDE BINDING SITE-LEUCINE RICH REPEAT (NBS-LRR) DE CAFÉ**

São Carlos

2019

UNIVERSIDADE FEDERAL DE SÃO CARLOS
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA EVOLUTIVA E BIOLOGIA
MOLECULAR
DEPARTAMENTO DE GENÉTICA E EVOLUÇÃO

ALUNO:
ANDRÉS SANTIAGO ORTIZ MORAZÁN

**ANÁLISES BIOINFORMÁTICAS DE PROTEÍNAS DE RESISTÊNCIA DO TIPO
NUCLEOTIDE BINDING SITE-LEUCINE RICH REPEAT (NBS-LRR) DE CAFÉ**

Dissertação apresentada ao Programa de Pós-Graduação em Genética Evolutiva e Biologia Molecular para obtenção do título de Mestre em Genética Evolutiva e Biologia Molecular.

Orientação: Prof. Dr. Flávio Henrique Silva

São Carlos

2019

Ortiz Morazan, Andres Santiago

Análises Bioinformáticas de Proteínas de Resistência do Tipo Nucleotide Binding Site-Leucine Rich Repeat (NBS-LRR) de Café / Andres Santiago Ortiz Morazan. -- 2019.

68 f. : 30 cm.

Dissertação (mestrado)-Universidade Federal de São Carlos, campus São Carlos, São Carlos

Orientador: Flavio Henrique Silva

Banca examinadora: Caio Cesar de Melo Freire, Celio Dias Santos Junior

Bibliografia

1. Café. 2. Proteínas de resistencias. 3. Bioinformática. I. Orientador. II. Universidade Federal de São Carlos. III. Título.

Ficha catalográfica elaborada pelo Programa de Geração Automática da Secretaria Geral de Informática (SIn).

DADOS FORNECIDOS PELO(A) AUTOR(A)

Bibliotecário(a) Responsável: Ronildo Santos Prado – CRB/8 7325



UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Biológicas e da Saúde
Programa de Pós-Graduação em Genética Evolutiva e Biologia Molecular

Folha de Aprovação

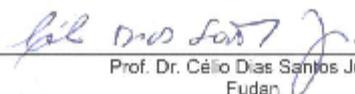
Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Andrés Santiago Ortiz Morazan, realizada em 15/02/2019:



Prof. Dr. Flavio Henrique da Silva
UFSCar



Prof. Dr. Caio Cesar de Melo Freire
UFSCar



Prof. Dr. Célio Dias Santos Júnior
Fudan

DEDICATÓRIA

A minha querida família.

Aquele que têm uma razão para viver pode se enfrentar todos os "como".

Friedrich Nietzsche.

AGRADECIMENTO

Ao deus de Spinoza se por acaso existisse.

Ao meu orientador o Prof. Dr. Flávio Henrique Silva que abriu as portas do seu laboratório e me orientou excelentemente, sempre me dando os melhores conselhos para o desenvolvimento do meu projeto com sucesso.

A Universidade Federal de São Carlos (UFSCar).

Ao Programa de Pós-graduação em Genética Evolutiva e Biologia Molecular (PPGGEv).

Ao Departamento de Genética e Bioquímica

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela assinatura da minha bolsa de mestrado.

À Organização dos Estados Americanos (OEA), que participou na seleção e assinatura da minha bolsa de mestrado.

Ao Instituto Hondurenho do Café (IHCAFE), quem contribuiu com o capital econômico para o desenvolvimento do meu projeto de pesquisa.

À Universidad Nacional Autónoma de Honduras (UNAH) quem contribuiu economicamente com o desenvolvimento do meu projeto.

Ao pessoal do Laboratório de Biologia Molecular (LBM).

A todos os meus amigos e parentes.

Muito obrigado!

RESUMO

Os fatores moleculares envolvidos na resistência a ferrugem do café e outras doenças ainda não são totalmente conhecidos, mas os projetos relacionados com o tema vêm sendo feitos desde o século passado. Um dos principais problemas que dificultam a identificação destes fatores é a falta de uma versão estável do genoma do *Coffea arabica*. Recentemente, o consórcio *World Coffee Research* (WCR) liberou a sua versão do genoma de *C. arabica* feito com tecnologia de Sequenciamento de Nova Geração (NGS). É conhecido que as relações patogênicas da Ferrugem e o Café têm proteínas efetoras que interagem com proteínas do tipo NBS-LRR da planta. Estas ativam a resposta imunitária da planta, o que resulta normalmente em Reações de Hipersensibilidade (HR). O gene de resistência de café S_{H3} codifica uma proteína similar à classe *Coiled Coil*-NBS-LRR e, sabe-se que as plantas que contêm esse gene apresentam resistência vertical à ferrugem. A procura de fatores similares ao gene S_{H3} poderia prover novos fatores de resistência que possam ser utilizados no desenvolvimento de novas variedades de café com resistência vertical para ferrugem. Neste estudo, foi possível identificar de proteínas presentes exclusivamente em genomas de plantas do gênero *Coffea*, assim como motivos presente exclusivamente em sequências de café arábico e robusta. Também foi possível identificar a localização cromossômica das sequencias identificadas. Finalmente, os resultados deste estudo conseguiram classificar e fazer anotações de todas as proteínas candidatas com o objetivo de utiliza-las em programas de melhoramento genético de café.

Palavras-chave: *Coffea arabica*. NBS-LRR. Genes de Resistência. Análise bioinformática.

ABSTRACT

The molecular factors involved in the coffee's resistance to rust and other diseases have not been completely known, but related projects have been developed since the last century. One of the main problems that make difficult to identify these factors is the lack of a stable version of the *Coffea arabica* genome. The organization World Coffee Research (WCR) recently made available its version of the *C. arabica* genome made out of Next Generation Sequencing (NGS) technology. It is known that the pathogenic relations of rust and coffee include effector proteins interacting with the plant's NBS-LRR proteins, which activate its immune response, normally resulting on a Hypersensitivity Response (HR). The coffee's resistance gene S_{H3} codifies a protein similar to the class Coiled Coil-NBS-LRR and the plants that include that gene show a vertical resistance to rust. Looking for factors similar to gene S_{H3} could lead to new resistance factors able to be used in the development of new coffee varieties with vertical resistance to rust. Through this research it was possible to identify sequences of protein exclusively present in genomes of the *Coffea* plants, as well as features differentially present in sequences of Arabica and robusta coffee. It was also possible to identify the chromosome localization of the identified sequences. Finally, the results of this research allowed classifying and systematizing all the candidate proteins to use them in coffee genetic improvement programs.

Key words: *Coffea arabica* . NBS-LRR. Genes of Resistance. Bioinformatics analysis

LISTA DE FIGURAS

FIGURA 1: Mapa dos países produtores de café por tipo de café produzido. -----	17
FIGURA 2: Mapa das regiões produtoras de café de Honduras -----	18
FIGURA 3: Planta de café robusta (<i>C. canephora</i>). -----	20
FIGURA 4: Ilustração de Café Arábicos -----	21
FIGURA 5: Distribuição mundial da Ferrugem do café (<i>Hemileia vastatrix</i>) -----	22
FIGURA 6: Modelo de zig-zag da resposta quantitativa do sistema imune das plantas ----- -----	24
FIGURA 7: Representação esquemática da estrutura e principais domínios das NBS-LRR proteínas. -----	30
FIGURA 8: Diagrama de fluxo para a identificação de RG's em sequências públicas de Café -----	34
FIGURA 9: Sub-etapas do primer e segundo passo do diagrama de fluxo de trabalho para a identificação de RG's em café. -----	38
FIGURA 10: Diagrama das sub-etapas do terceiro e quarto passo do diagrama de fluxo de trabalho para a identificação de RG's de café. -----	40
FIGURA 11: Exemplo dos resultados da interface gráfica do software MEME. -----	45
FIGURA 12: Localização do motivo Ploop em sequências de café robusta. -----	46
FIGURA 13: Exemplo da interface de resultado do Pfam. -----	47
FIGURA 14: Motivos conservados utilizados para o desenho de primers. -----	48
FIGURA 15: Árvore filogenética das sequências candidatas a CNL de <i>C. arabica</i> – WCR. -----	50
FIGURA 16: Árvore filogenética das sequências candidatas a CNL de <i>C. canephora</i> –CGH. -----	51
FIGURA 17: Localização cromossômica das sequências candidatas para CNL de <i>C.</i> <i>arabica</i> – WCR. -----	52

LISTA DE TABELAS

TABELA 1: Resultados do processo de Data Mining. -----	42
TABELA 2: Número de seqüências recuperadas nos por busca de palavras-chaves nas bases de dados do café e de RG -----	43
TABELA 3: Número de seqüências candidatas a RG's do café sem redundâncias. -----	44

LISTA DE ABREVIATURAS E SIGLAS

ABNT = Associação Brasileira de Normas Técnicas

AVR= Avirulência

BAK1 = Receptor-like Cytoplasmic Kinase Botrytis-Induced Kinase 1

BCH = Banco Central de Honduras

CC= Coiled Coil

CDNA= DNA complementar

CSP= Proteína de shock de frio.

DAMP= Padrões Moleculares Ativados por Danos

DP= Primer Degenerado

EF-Tu= Fator de Elongação Tu

ETS= Susceptibilidade Ativada por Efeitores

ETI = Imunidade Ativada por Efeitores

ELRR= Repetições Ricas em Leucina Extracelular

HMM = Modelos de Markov Ocultos

ICO = Organização Internacional do Café

IHCAFE = Instituto Hondurenho do Café

FAO = Organização das Nações Unidas para Agricultura e Alimentação

KIN=Domínio Quinase

LRR= Repetições Ricas em Leucina

MAPK= Proteínas Quinase Ativada por Mitógenos

NBS= Sítios de Ligação ao Nucleotídeo

NGS= Sequenciamento de Nova Geração

NLS= Sinal de Localização Nuclear

PAMP= Padrões Moleculares Associados a Patógenos

PCR= Reação em Cadeia da Polimerase

PIB = Produto Interno Bruto

PIBA= Produto Interno Bruto Agrícola

PRR = Receptores de Reconhecimento de PAMP

PTI = Imunidade Ativada por PAMP

RG = Genes de Resistência

RGA = Genes de Resistência Análogos

RNS = Espécies Reativas de Nitrogênio

RP = Proteína de Resistência

ROS = Espécies Reativas de Oxigênio

STK = Serine/Treonine Quinase

TLR = Receptor Similar a Toll

TNL = TIR-NBS-LRR

TIR = Proteína Similar ao Receptor Toll/Interlecina 1

TRD = Dominio Transmembrana

UFSCAR = Universidade Federal de São Carlos.

SUMÁRIO

DEDICATÓRIA	V
AGRADECIMENTO	VII
RESUMO	VIII
ABSTRACT	IX
1 INTRODUÇÃO.....	16
1.1. BREVE REVISÃO HISTÓRICA DO CAFÉ.....	16
1.2. DADOS COMERCIAIS DO CAFÉ.....	17
1.3. O GÊNERO <i>Coffea</i>	18
1.3.1 CARACTERÍSTICAS BOTÂNICAS DO GÊNERO <i>Coffea</i>	18
1.4 A FERRUGEM DO CAFÉ	22
1.4.1 A ORIGEM E DISPERSÃO DA FERRUGEM DO CAFÉ	22
1.4.2 A FERRUGEM NO CONTINENTE AMERICANO.....	23
1.4.3 CICLO PATOLÓGICO DA FERRUGEM DO CAFÉ	23
1.5 VISÃO GERAL DO SISTEMA IMUNE DE PLANTAS.....	24
1.5.1 PADRÕES MOLECULARES ASSOCIADOS A PATÓGENOS OU PAMP's	25
1.5.2 RECEPTORES DE RECONHECIMENTO DE PAMP's ou PRR	26
1.5.3 MECANISMOS DE SINALIZAÇÃO DA PTI.....	26
1.5.4 IMUNIDADE ATIVADA POR EFETORES OU ETI.....	28
1.5.5 GENES DE RESISTÊNCIA ANÁLOGOS (RGA's).....	28
1.5.6 RG's DE CAFÉ	30
1.6 IDENTIFICAÇÃO DE NBS-LRR's	31
1.6.1 PROCEDIMENTO PARA A IDENTIFICAÇÃO DE RGA's.....	32
3.MATERIAIS E MÉTODOS	34
3.1. DESENHO DO DIAGRAMA DE FLUXO DE TRABALHO	34
3.2. DATA MINING	35
3.2.1 BASES DE DADOS DE CAFÉ	35
3.2.2 BASES DE DADOS DE RG's.....	36
3.2.3 BUSCA POR PALAVRAS-CHAVES	37
3.3. ALINHAMENTOS E FILTRAGEM.....	37
3.4.1 BUSCA DE PADRÕES DE MOTIVOS	38
3.4.2 PROCURA DE DOMÍNIOS DE PROTEÍNAS	39
3.5. DESENHO DE PRIMERS DEGENERADOS	39
3.7 ANÁLISES FILOGENÉTICAS	40
3.8 LOCALIZAÇÃO CROMOSSÔMICA	41

4 RESULTADOS E DISCUSSÃO	42
4.1. DATA MINING	42
4.1.1 BASES DE DADOS DO CAFÉ.....	42
4.1.2 BASES DE DADOS DE RGA's.....	42
4.2 ALINHAMENTOS E FILTRAGEM.....	43
4.3 BUSCA DE MOTIVOS	44
4.4. PROCURA DE DOMÍNIOS E ANOTAÇÕES	46
4.6 ANÁLISES FILOGENÉTICAS	49
4.7 LOCALIZAÇÃO CROMOSSÔMICA	52
5. CONSIDERAÇÕES FINAIS E CONCLUSÕES	53
6. REFERENCIAS	54
7. ANEXOS	61
ANEXO 1: <i>Primers</i> degenerados de <i>C. arabica</i>	61
ANEXO 2: <i>Primers</i> degenerados de <i>C. arabica</i>	63
ANEXO 3: Principais proteínas de café relacionadas como proteínas de resistência de outras plantas relatadas com antecedência	64
ANEXO 4: Motivos identificados exclusivamente em sequencias de <i>Coffea arabica</i> e <i>Coffea canephora</i>	67
ANEXO 5: Árvore filogenética das sequências candidatas a CNL de <i>C. canephora</i> –CGH, <i>C. arabica</i> -WCR e sequências de referência PRGdb.....	68

1 INTRODUÇÃO

1.1. BREVE REVISÃO HISTÓRICA DO CAFÉ

O café é um dos *commodities* mais importantes assim como a bebida não alcoólica mais consumida no mundo (FAO, 2013). A palavra “café” vem provavelmente do árabe “*Qahwah*”, traduzida como “algo que têm gosto amargo”, mas também poderia estar relacionado com a sua região de origem na província de *Kaffa* em Etiópia onde foi localizado o centro Vavilov dos cafés (HARLAN, 1969). Wellman (1961) e Purseglove (1968) falam do consumo de uma pasta de grãos de café muito parecidos ao chocolate pelas tribos da África oriental na era pré-islâmica com o objetivo de aumentar a força dos guerreiros (TEKETAY, 1999). Outros autores falam de possíveis experimentos feitos com grãos de café nos mosteiros do Oriente Médio, dando assim, provavelmente, origem ao processo de torrefação dos grãos e à preparação das primeiras infusões de café (GONZALES, 2012). Foram os árabes quem levaram a planta fora da África, primeiro à península arábica onde seu consumo se tornou popular como um substituto do álcool e mais tarde, através de rotas comerciais, chegou a Turquia e, Egito e de lá ingressou na Europa, provavelmente levado por comerciantes italianos (TEKETAY, 1999).

Crawford (1852) relata que os árabes monopolizaram o comércio de café até meados do século XVII, quando peregrinos muçulmanos levaram sementes férteis a Índia, Java e Ceilão, onde foram estabelecidas as primeiras culturas de café fora do império árabe. O médico e botânico alemão Léonard Rauwolf (1583 apud AVELINO et al., 2013) fez um dos primeiros relatos ocidentais da preparação, consumo e efeitos do café na saúde. Nos séculos XVIII e XIX a sua produção foi localizada predominante nas ilhas Ceilão e Indonésia sob domínio do Império Britânico até a primeira epidemia de ferrugem da planta em 1890. Rayner (1972) relatou o impacto econômico dessa epidemia e a necessidade de trocar as culturas de café altamente susceptíveis à doença pelo chá, o que causou importantes mudanças culturais dos britânicos até então amantes da bebida (AVELINO et al., 2013).

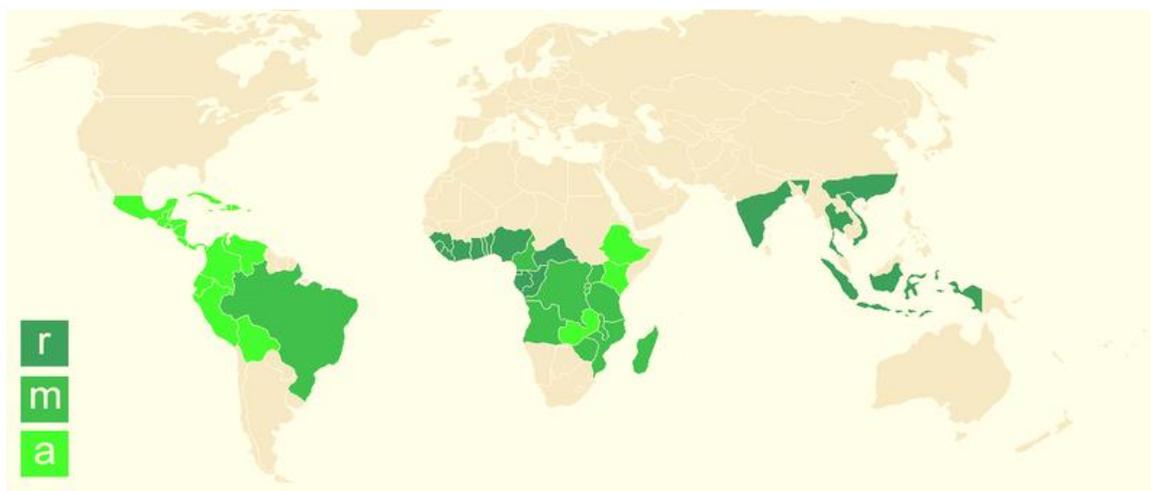
As primeiras culturas de café no “novo mundo” foram estabelecidas no Caribe provavelmente pelos colonos holandeses e franceses que habitaram as ilhas dessa região desde o século XVIII, Relatos da época falam da introdução do café no Haiti e Martinica nos primeiros anos de 1700 (TEKETAY, 1999). Mas foi em 1727 que o café chegou ao Brasil no estado de Pará. As primeiras plantas de café no Brasil foram trazidas das Antilhas Francesa pelo oficial português Francisco de Mello Palheta (ABIC, 2018). Em 1785, o café chegou a Colômbia vindo das Antilhas Holandesas e para o século XIX foi introduzido na América Central e México pelos espanhóis. Um dos relatos mais antigos da cultura na região foi feito por Ramón de

Anguiano em 1801, governador da província de Honduras durante a era colonial. Ele fez um relatório para o rei da Espanha em que conta nele os detalhes da cultura de café na região de “Manto”, hoje parte do departamento de Olancho na atual República de Honduras (GONZALES, 2012).

1.2. DADOS COMERCIAIS DO CAFÉ

Economicamente, o café é uma das culturas mais importantes do mundo, sendo produzido por mais de 60 países na faixa tropical e subtropical do planeta (Figura 1). *Coffea arabica* (café arábico) e *Coffea canephora* (café robusta) são as duas espécies de café com valor comercial, as quais representam 60% e 40% da produção mundial, respectivamente (ICO, 2018a). Em 2017, mais de 159 milhões de sacas de café foram produzidas pelos países exportadores do grão. O Brasil é o maior exportador de café do mundo, produzindo mais de 52 milhões de sacas o que representa 33% da produção mundial de 2017. Vietnã (18,5%), Colômbia (8,8%), Indonésia (6,8%) e Honduras (4,8%) são os seguintes na lista dos maiores produtores e exportadores do grão no mundo (ICO, 2017a). Entre os principais países importadores de café em grão estão a União Europeia (64%), os Estados Unidos (22%), Japão (5,8), a Federação Russa (4,1%), Suíça (2,4%), Noruega (0,5%) e Tunísia (0,4%) (ICO, 2018b).

Figura 1 - Mapa dos países produtores de café por tipo de café produzido.



Fonte: OIC, 2017

r = Robusta (Região tropical da Ásia e a África Ocidental), m = Misto (Regiões da África Central, e Brasil), a = Arábica (América com exceção do Brasil, e Etiópia Quênia e Zâmbia na África).

Em Honduras, quinze dos dezoito estados do país são produtores de café. As culturas são feitas estritamente apenas com variedades de *C. arabica*. A cultura de café robusta é restrita à pesquisa científica e seu cultivo comercial está proibido, visto que o país dedica-se à produção

de cafés suaves de alta qualidade (IHCAFE, 2017) (Figura 2). O grão contribui com 10% do produto interno bruto (PIB) e 35% do produto interno bruto agrícola (PIBA) do país (BCH, 2017). Aproximadamente 1 milhão de pessoas são empregadas em atividades relacionadas com a cadeia produtiva do café, o que representa quase 25% dos empregos do país (IHCAFE, 2017). Mais do 80% do café de Honduras é produzido por pequenos produtores (pessoas que possuem apenas 1-5 hectares de cultura) e é um importante fator de desenvolvimento e seguridade alimentar nas regiões mais pobres do país (IHCAFE, 2017). Honduras é o primeiro exportador de café da região de América Central e México, o terceiro do continente depois de Brasil e a Colômbia, e o quinto em nível mundial, exportando cafés para mais de 60 países diferentes (IHCAFE, 2017).

Figura 2 - Mapa das regiões produtoras de café em Honduras



(Fonte: Café de Honduras, 2010)

Agalta: Inclui parte dos estados de Colón, Atlántida, Yoro, Cortes e Olancho. El Paraíso: inclui parte dos estados de El Paraíso, Francisco Morazán, Comayagua y Choluteca. Comayagua: inclui parte dos estados de Comayagua y Francisco Morazán. Motecillos: inclui parte dos estados de La Paz, e Santa Bárbara. Opalaca: Inclui parte dos estados de Santa Bárbara, Intibucá e Lempira. Copán: Inclui parte dos estados de Copán, Santa Bárbara, Lempira e Intibucá.

1.3. O GÊNERO *Coffea*

1.3.1 CARACTERÍSTICAS BOTÂNICAS DO GÊNERO *Coffea*

Os cafés são plantas da Família das Rubiáceas que inclui mais de 500 gêneros e cerca de 13.000 espécies. No gênero *Coffea* existem mais de 100 espécies, todas nativas da África tropical (MEKURIA et al., 2004, ROVA et al., 2002. ROJO-JIMENEZ, 2014). Chevelier (1947 apud

MAURIN et al., 2007) e Coste (1955 apud MAURIN et al., 2007) classificaram formalmente os cafés pela primeira vez em três grupos botânicos: *Eucoffea* em que se classificam plantas que provem da África ocidental e Central; *Mozambicoffea* que agrupa espécies da África oriental; e *Mascarocoffea* que agrupa espécies que provem de Madagascari e as ilhas Mascarenas. Mais recentemente, Bridson (1988) e Davis (2005) classificaram finalmente o gênero *Coffea* em 103 espécies agrupados em dois subgêneros: *Coffea* subgênero *Coffea* que agrupa a maioria de espécies (95 espécies) entre elas as principais com valor comercial (MAURIN et al., 2007, DE KOCHKO et al., 2010). O segundo é *Coffea*, subgênero *Baracoffea* que inclui oito espécies que, de acordo com Leroy (1982), provém do ocidente de Madagascar, do nordeste do Quênia e do sudeste da Somália (DAVIS et al., 2005, MAURIN et al., 2007).

Os cafés são plantas lenhosas que podem medir entre 5 e 10 metros de altura, possuem folhas elípticas que aparecem em pares ao longo do ramo e apresentam pequenas estípulas e pecíolos curtos. Os tons de cores das folhas vão desde o verde obscuro até roxo, as flores são brancas, pequenas e aparecem em inflorescências. Os frutos são tipo drupa, com epicarpo carnoso e dupla semente (WALLER; BIGGER; HILLOCS, 2007). Além de *C. arabica* e *C. canephora* outras, espécies de importância são *Coffea liberica*, que têm sido utilizadas em programas de melhoramento de café arábicos; *Coffea excelsa* e *Coffea dewevreri*, as quais são produzidas comercialmente, mas sem impacto no comércio mundial de café (ROJO-JIMENEZ, 2014)

1.3.1.1 O CAFÉ ROBUSTA (*Coffea canephora*)

Reino: Plantae

Divisão: Magnoliophyta

Classe: Magnoliopsida

Ordem: Gentianales

Família: Rubiaceae

Gênero: *Coffea*

Espécie: *C. canephora*

C. canephora é uma planta nativa da floresta equatorial da África ocidental e têm sua origem, provavelmente, no Congo ou Angola (ROJO-JIMENEZ, 2014). Os cafés robusta, como são chamados, são árvores que podem medir até 10 metros de altura e tem a raiz pouco

profunda. Os frutos são redondos e as sementes são pequenas e alongadas. As folhas são maiores que as folhas de *C. arabica* e de aparência mais robusta ou primitiva (Figura 3). *C. canephora* caracteriza-se pela sua alta produtividade e resistência a doenças, como a ferrugem, e secas. No entanto, sua qualidade de xícara é menor frente a dos cafés arábicos (DE KOCHKO et al, 2010).

Figura 3 - Planta de café robusta (*C. canephora*).



Fonte: DE KOCHKO et al. 2010, p. 25

A: Cafeeiro com frutos em diferentes estádios de maturação; B: Grupos de flores brancas e alguns frutos de cor vermelho; C: Frutos maduros em grupos com diferentes estádios de maturação.

1.3.1.2 O CAFÉ ARÁBICO (*Coffea arabica*)

Reino: Plantae

Divisão: Magnoliophyta

Classe: Magnoliopsida

Ordem: Gentianales

Família: Rubiaceae

Gênero: *Coffea*

Espécie: *C. arábica*

O café arábico, descrito por Linneo em 1753 é uma árvore de 5 a 10 metros de altura, possui folas ovaladas, de tons verde até bronze, dependendo da variedade, e são distribuídos

em pares no ramo. As flores são brancas, pequenas, possuem um cheiro suave e doce, são distribuídas em grupos e são capazes de autopolinização. Os frutos são ovalados de cor verde até vermelho, dependendo do nível de maturação, e possuem normalmente duas sementes na forma de meia lua (CLARKE; MACRAE, 1989) (Figura 4).

Figura 4: Ilustração de Café Arábicos



Fonte: KURT STUEBER, 2003, p. 106.

Apresenta a distribuição dos frutos e as flores no ramo, assim como a estrutura das flores e frutos em diferentes etapas de desenvolvimento. Publicado em *Kohler's Medical Plants*, 1887.

Charrier e Berthaud (1985) identificaram uma das principais particularidades dessa espécie. Eles relataram que *C. arabica* possui um genoma Tetraploide ($2n=4x=44$), sendo a única espécie poliploide no gênero *Coffea* (LASHERMES et al., 1999. NOIR et al. 2004). *C. arabica* possui um genoma anfidiplóide que provem da hibridação entre *C. eugenioides* (E^{a_1}) e *C. canephora* (C^a) ou eco-tipos deste último (LASHERMES et al., 1999). Harlan e DeWet (1975 apud LASHERMES et al., 1999) descreveram que provavelmente as primeiras plantas poliploides férteis deste cruzamento vieram dos híbridos diploides interespecíficos que

duplicaram os seus cromossomas, ou de retrocruzamentos de triploides espontâneos (LASHERMES et al., 1999).

1.4 A FERRUGEM DO CAFÉ

Reino: Fungi

Divisão: Basidiomycota

Classe: Pucciniomycetes

Orden: Pucciniales

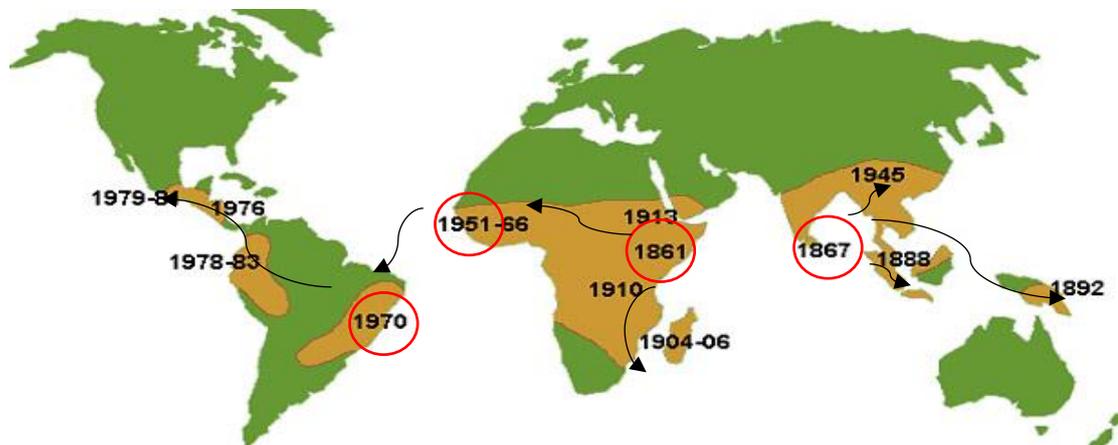
Género: *Hemileia*

Espécie: *H. vastatrix*

1.4.1 A ORIGEM E DISPERSÃO DA FERRUGEM DO CAFÉ

A ferrugem do café foi reportada pela primeira vez na Ilha de Ceilão atual Sri Lanka. O cientista micólogo Berkeley reportou a doença em 1869 assim como o fungo *Hemileia vastatrix*, por ela responsável. No entanto, sua origem foi objeto de discrepâncias pela distância dos centros de origem do café na África (AVELINO e RIVAS, 2013). Até 1960, alguns países de África Ocidental ainda não tinham reportado a doença, e foi na década de 70 que ela chegou ao continente americano, especificamente no estado da Bahia no Brasil (VILLEGAS; BAEZA, 1990). Mais tarde, atingiu a América Central, sendo relatada pela primeira vez na Nicarágua em 1976 e gradualmente se dispersou pela região (Figura 5). Atualmente só Hawaii e Austrália não têm a presença da ferrugem em suas culturas (AVELINO; RIVAS, 2014).

Figura 5: Distribuição mundial da Ferrugem do café (*Hemileia vastatrix*)



Fonte: AMESON, 2000, p.1

Distribuição mundial da Ferrugem do café e anos de primeiros relato por regiões, os círculos vermelhos amostram as regiões de origem das primeiras epidemias relatadas. As setas pretas mostram as possíveis rotas de dispersão de *Hemileia vastatrix*.

1.4.2 A FERRUGEM NO CONTINENTE AMERICANO

Desde sua chegada ao continente, a ferrugem do café representou uma ameaça importante à produção da região. As estratégias mais promovidas durante a primeira epidemia de ferrugem na América foi o desenvolvimento de programas de melhoramento genético (AVELINO; RIVAS, 2013). Esta estratégia junto com a utilização de controladores químicos, biológicos e boas práticas agrícolas conseguiram controlar parcial ou totalmente o ataque da doença. Desde 1980 até finais dos anos 90 Brasil, Colômbia e Honduras conseguiram desenvolver variedades com resistência vertical à diferentes raças de ferrugem.

Estas variedades conseguiram resistir a doença por quase 20 anos até a epidemia do 2008-2012, que afetou os principais produtores de variedades desenvolvidas utilizando o Híbrido de Timor (TH) como o principal progenitor doador de genes de resistências para ferrugem utilizado (CRISTANCHO et al., 2012). Atualmente, novas variedades do Híbrido de Timor e cruzamentos do TH x Robusta têm sido utilizados como doadores de genes de resistências para variedades de café arábico (SERA et al., 2007). *C. liberica* também têm sido identificadas como um importante doador de genes de resistência não só para ferrugem, senão também para outras doenças com impacto econômico importante como o Doença do Fruto do Café ou CBD (*Coffee Berry Disease*) (PRAKASH, et al., 2004).

1.4.3 CICLO PATOLÓGICO DA FERRUGEM DO CAFÉ

O ciclo patológico da ferrugem do café se inicia com a dispersão dos uredósporos, que é a forma infectante do fungo, sendo esta etapa dividida em duas partes: a primeira é a liberação, na qual os esporos saem do esporóforo por efeito do vento, chuva etc. A segunda é a deposição, em que os esporos conseguem se transportar da lesão de origem a outras áreas da folha ou outras folhas e plantas (AVELINO; WILLOCQUET; SAVARY, 2004). A etapa seguinte é a germinação, que é dependente da radiação solar, humidade e temperatura. Nela se inicia a colonização do tecido folhar: os esporos geram um tubo germinativo, que normalmente, entra na planta pelos estomas levando à terceira etapa, a penetração. E é assim que começa o processo de infecção (AVELINO; WILLOCQUET; SAVARY, 2004).

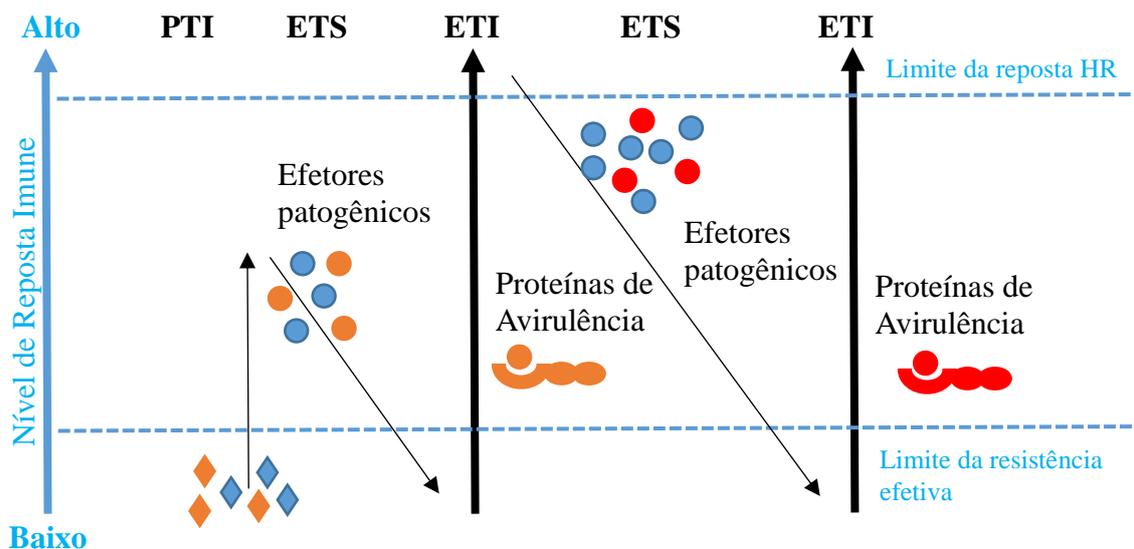
A colonização dos tecidos se inicia após a penetração, e é possível observar os primeiros sintomas da doença nas folhas da planta. Nessa etapa, tanto fatores ambientais como a fisiologia da planta favorecem a doença. Nutrição deficiente, pH ácido do solo, radiação solar direta e

secas são fatores que favorecem a severidade da enfermidade. Finalmente, o ciclo da ferrugem é fechado com a produção de novos esporos infectantes, que ao se dispersarem, dão início a um novo ciclo (AVELINO; RIVAS, 2013). Na última epidemia de ferrugem acontecida no ano de 2012, pelo menos 31% de produção da região de América Central e México viu-se afetada pela ferrugem e, 3 países da região (Honduras, Guatemala e Costa Rica) declararam-se em estado de emergência (AVELINO; RIVAS, 2013).

1.5 VISÃO GERAL DO SISTEMA IMUNE DE PLANTAS

No meio ambiente, as plantas são atacadas por diferentes patógenos (bactérias, fungos e insetos) capazes de causar doenças. Diferentemente dos animais, elas não possuem células móveis relacionadas com defesa e imunidade (OJITO-RAMOS e PORTAL, 2010). Elas restringem a sua defesa quase que exclusivamente à utilização de barreiras físicas e à expressão local de fatores que inibem o crescimento e colonização dos patógenos (ZIPFEL, 2008). O sistema imune das plantas consiste basicamente de duas formas ligadas de imunidade inata. Jones e Dangl (2006) descreveram um modelo em zig-zag que demonstra etapas da imunidade das plantas, os fatores moleculares envolvidos e o nível de resposta imune desencadeado por elas em cada etapa (NÜRNBERGER e KEMMERLING, 2009) (Figura 6).

Figura 6 – Modelo de zig-zag da resposta quantitativa do sistema imune das plantas



Fonte: baseado no modelo de JONES e DANGL, 2006

A seta azul amostra a amplitude da resposta imune, as linhas pontilhadas de cor azul mostram o nível de resistência efetiva e a resposta de hipersensibilidade as setas pretas, o fluxo do modelo de resposta imune, os diamantes representam os PAMP's, as bolinha pretas, vermelhas e azuis representam os efetores patogênicos e as figuras em forma de meia lua vermelha e azul, as proteínas de resistência das plantas. PAMP= Padrões Moleculares Associados a Patógenos, PTI = Imunidade Ativada por PAMP, ETS= Susceptibilidade Ativada por Efetores, ETI= Imunidade Ativada por Efetores. Avr-R= Proteínas de Avirulência ou Resistência.

Na Fase I do modelo de imunidade de plantas, os Padrões Moleculares Associados a Patógenos ou PAMP (*Pathogen-Associated Molecular Pattern*) são reconhecidos pelos Receptores de Reconhecimento de Padrões de plantas ou PRR (*Pattern Recognition-Receptor*), o que resulta na ativação da Imunidade Ativada por PAMP ou PTI (*PAMP-Triggered Immunity*), também conhecida como Imunidade Ativada por Padrões por poder se ativar por outros padrões além dos PAMP (JONES; DANGL, 2006). Na Fase II, os patógenos que não foram controlados pela PTI ativam fatores de virulência conhecidos como Efetores que inibem os fatores de resistência das plantas. Este tipo de interação é conhecido como Susceptibilidade Ativada por Efetores ou ETS (*Effector-Triggered Immunity*) (JONES; DANGL, 2006, BIGEARD; COLCOMBET; HIRT, 2014).

Na fase III, o efector microbiano é reconhecido por proteínas específicas de resistência da planta, que, normalmente são do tipo *Nucleotide Binding-Site-Leucine Rich Repeat* ou NBS-LRR. Se a proteína de resistência é capaz de controlar o efeito do efector, é possível falar de Imunidade Ativada por Efetores ou ETI (*Effector-Triggered Immunity*) (DANGL; JONES, 2001, JONES; DANGL, 2006). Em geral, tanto a PTI quanto a ETI promovem Reações de Hipersensibilidade (HR - *Hypersensitivity Reactions*) o que resulta em morte celular programada no sítio da infecção. É conhecido que a ETI produz uma resposta imune mais forte e rápida que a PTI (JONES; DANGL, 2006, BOLLER; FELIX, 2009). Finalmente, na Fase IV, a seleção natural promove tanto a eliminação, diversificação e o desenvolvimento de novos efetores patogênicos capazes de inibir o efeito das proteínas de resistência da planta quanto novos fatores de resistência que possam ativar novamente a ETI na planta (JONES; DANGL 2006, BOLLER; FELIX, 2009).

1.5.1 PADRÕES MOLECULARES ASSOCIADOS A PATÓGENOS OU PAMP'S

Os Padrões Moleculares Associados a Patógenos ou "PAMP" são epítomos invariáveis de moléculas conservadas e indispensáveis para o microrganismo patógeno, que estão ausentes nas plantas e são facilmente identificáveis por diferentes hospedeiros (SCHWESSINGER et al., 2008). No caso dos fungos, os PAMP mais comuns são componentes derivados da membrana celular tais como: α -glucano, ergosterol e quitina (HE et al., 2007, RYAN et al., 2007).

São numerosos nas bactérias, porém os mais comuns são os lipopolissacarídeos (LPS), principalmente o LPS A, o peptidoglicano e a flagelina (SCHWESSINGER et al., 2008, HEESE et al., 2007, ZIPFEL et al., 2008). Embora a maioria dos PAMP sejam componentes

extracelulares, existem PAMP's como as Proteínas de Choque de Frio (CSP *Cold Shock Protein*) e o Fator de Elongação de Tradução (*Elongation Factor-Tu*) que são proteínas intracelulares capazes de serem reconhecidas pelos PRR e ativar a PTI (HE et al., 2007).

1.5.2 RECEPTORES DE RECONHECIMENTO DE PAMP's ou PRR

Os PAMP's são reconhecidos pelo PRR's também chamadas Receptores de Superfície, os quais podem ser de tipo Receptor Kinase (RK) (*Receptor Kinase*), ou Receptores de Proteínas (RP) (*Receptor Protein*) (JONES e DANGL, 2006, DANGL e JONES 2001, LI, 2014). Os PRR's são análogos funcionais e estruturais dos Receptores Similares a Toll ou TLR (*Toll-Like Receptor*) dos animais, e podem reconhecer tanto moléculas microbianas, quanto Padrões Moleculares Associados a Danos ou DAMP (*Damage-Associated Molecules Patterns*) (BOLLER e FELIX, 2009).

Os PRR possuem grande afinidade e especificidade pelos PAMP's e DAMP's, que são proteínas transmembranas que contém ectodomínios com Repetições Ricas em Leucinas ou LRR (*Leucine Rich Repeats*), um domínio transmembrana e um domínio intracelular variável, que entre os mais comuns ficam os do tipo Kinase (HEESE et al., 2007, BOLLER e FELIX, 2009). Os reconhecimentos dos PAMP's ou DAMP's pelo PRR ativa a primeira resposta de defesa ou imunidade basal das plantas, o que normalmente controla os microrganismos não adaptados com potencial patogênico (SCHWESSINGER et al., 2008, BIGEARD; COLCOMBET; HIRT, 2015).

1.5.3 MECANISMOS DE SINALIZAÇÃO DA PTI

A PTI pode ser ativada pela interação PAMP-PRR ou DAMP-PRR. Este tipo de imunidade é também conhecido como resistência basal ou horizontal (YAMAGUCHI et al., 2006). A PTI ativa: a rápida difusão de íons pela membrana citoplasmática (principalmente Ca^{2+}), a ativação da cascata das Proteínas Quinase Ativada por Mitógenos ou MAPK (*Mitogen-activated protein kinase*), a produção de espécies reativas de oxigênio, a expressão de genes de resistência específicos e o reforço da parede celular (HE et al., 2007, ZIPFEL et al., 2008).

Uma das primeiras respostas após a identificação de PAMP's ou DAMP's pelas plantas é o influxo extracelular de Ca^{2+} para o citosol, regulado, possivelmente, por proteínas do tipo Receptor Similar a Quinases Citoplasmática induzida por *Botrytis* 1 ou BAK1 (*Receptor-like Cytoplasmic Kinase Botrytis-Induced Kinase 1*). Sabe-se que o BAK1 e outros complexos

proteicos são requeridos tanto para induzir o influxo de Ca^+ na célula tanto como para ativar a cascata MAPK o que aciona Fatores de Transcrição (BIGEARD; COLCOMBET; HIRT, 2015). O Ca^{2+} induz a abertura de outros transportadores de membrana (H^+ , K^+ , Cl^- e Nitrato) o que conduz à alcalinização do ambiente extracelular e à polarização das membranas. (JEWORUTZKI, et al., 2010, LI, et al., 2014). A alcalinização do ambiente extracelular é inibida pelo ATPase- H^+ de membrana plasmática, que contribui com a entrada de anions no citosol e saída de prótons para o apoplasto (ELMORE; COAKER, 2011).

Outra resposta que acontece rapidamente após do reconhecimento dos patógenos é a produção de Espécies Reativas de Oxigênio ou ROS (*Reactive Oxygen Species*). Esta em *Arabidopsis sp.* foi relacionada com a proteína de membrana RBOHD (*Respiratory Burst Oxidase Homolog D*) a qual é responsável pela produção de ROS induzida pela interação PAMP-PRR (NUHSE, et al., 2007). RBOHD produz superóxidos não permeáveis pela membrana no apoplasto e depois, através do efeito das Superóxido Dismutase ou SOD (*Superoxide Dismutase*), são convertidos em peróxido de hidrogênio (H_2O_2), o qual consegue entrar no citosol e organelas aumentando a concentração de Ca^+ (BIGEARD; COLCOMBET; HIRT, 2015).

Embora o Ca^+ e ROS sejam as moléculas mais importantes e envolvidas na sinalização dos PAMP's, outras moléculas como o Óxido Nítrico ou NO (*Nitric Oxide*) ou derivados do NO conhecidos como Espécies Reativas de Nitrogênio ou RNS (*Reactive Nitrogen Species*) têm sido envolvidas em diferentes etapas de sinalização durante a resposta a patógenos (GROß; DURNER; GAUPELS, 2013). A origem biosintética do NO ainda não é muito clara, mas sabe-se que o influxo de Ca^+ no citosol ativa a proteína Calmodulina (CaM) ou proteínas similares a CaM, que conduzem à indução *downstream* da produção do NO (MA, et al., 2008, GRUPTA, et al., 2011).

Alguns estudos demonstraram que lípidios como o Ácido Fosfatídico ou PA (*Phosphatidic Acid*) e as Ceramidas também podem funcionar como moléculas de sinalização na interação Planta-Patógeno (OKAZAKI; SAITO, 2014). PA pode regular diferentes sinais de transdução de imunidade tais como: produção de ROS, ativação das MAPK, produção de Ácido Jasmonico ou JA (*Jasmonic Acid*) e produção de Etileno ou ET (*Ethylene*) (WANG, et al. 2000, TESTERINK, et al. 2007, ZHANG, et al. 2009). As Proteínas G Heterométricas (*Heterometric G protein*) também têm sido relacionadas com sinais de tradução da resposta imune,

principalmente com a produção de ROS e ativação das MAPK (BIGEARD; COLCOMBET; HIRT, 2015). As Proteínas 14-3-3 (*14-3-3 Proteins*) foram relacionados em diferentes níveis da sinalização imune e estão associadas com a produção de ROS. Até o fechamento dos estomas, sabe-se que estas famílias de proteínas interagem com múltiplas proteínas fosforiladas envolvidas em outras etapas da sinalização imunitária, mas a informação sobre estas ainda é limitada (SEHNKE; DELILLE; FERL, 2002, BIGEARD; COLCOMBET; HIRT, 2015).

1.5.4 IMUNIDADE ATIVADA POR EFETORES OU ETI

As interações plantas-patógeno, especialmente entre patógenos biotróficos, são governadas pela interação dos loci de avirulência (Avr) dos patógenos e pelos alelos do locus que correspondem ao Gene de Resistência (RG) na planta (FLOR, 1971). Os fatores de Avirulência ou Avr (*Avirulence*) dos patógenos podem ser reconhecidos pela Proteínas de Resistência ou RP (*Resistance Protein*) das plantas e ativar a ETI que é um tipo de imunidade mais rápida e forte que a PTI (JONES e DANGL, 2006).

Posteriormente à ativação das RP, desencadeiam-se reações de hipersensibilidade ou HR (*Hypersensitivity Reactions*) na célula infectada, limitando o crescimento e colonização do patógeno (DANGL e JONES, 2001). Finalmente, se as RP's da planta não são capazes de inibir o efeito dos Avr ou não possuem uma RP para esse Avr específico, a planta desenvolve a doença resultando em uma interação ETS (JONES e DANGL, 2006). ETS e ETI promovem processos evolutivos que levam ao desenvolvimento de novos Avr pelos patógenos e RP pelas plantas. Nestes últimos, esse processo é conhecido como Resistência Sistêmica Adquirida ou SAR (*Systemic Acquired Resistance*) (CONRATH, 2006).

1.5.5 GENES DE RESISTÊNCIA ANÁLOGOS (RGA's)

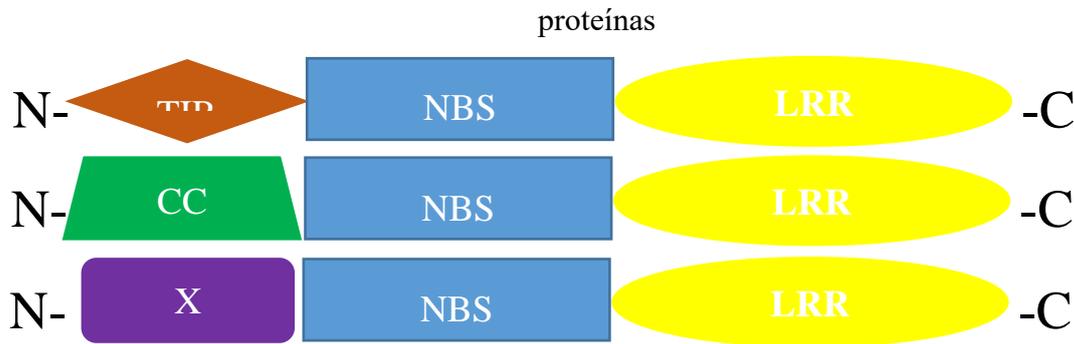
Os fatores moleculares envolvidos nas interações planta-patógeno ainda não são totalmente conhecidos, mas os Genes de Resistência ou RG's (*Resistance Genes*) são os fatores de resistência mais estudados e numerosos (BELKHADIR et al., 2004). As proteínas codificadas pelos RG's se dividem em 8 diferentes classes, as duas principais são: A classe I que inclui proteínas citoplasmáticas conhecidas como TIR-NBS-LRR (TNL) as quais contêm três domínios principais, um domínio *Toll/Interleukin-1-like receptor* (TIR) no N-terminal, um *Nucleotide-Binding Site* (NBS) e uma série de *Leucine Rich Repeats* (LRR) no C-terminal. A classe II são as *Coiled Coil-NBS-LRR* (CNL) ou non-TIR-NBS-LRR que também são proteínas citoplasmáticas e semelhantes à classe I, apenas diferindo no extremo N-terminal onde a classe

II possui um domínio do tipo *Coiled Coil* ou CC (Figura 7). As duas classes têm sido relacionadas com a interação com proteínas efetoras de patógenos e HR, importantes fatores de resistência efetiva a doenças em plantas (GURURANI et al. 2012).

A classe III são as chamadas LRR extra-citoplasmáticas (eLRR), as quais possuem um domínio LRR, um NBS e um domínio Transmembrana (TrD). Estas são conhecidas pelo papel que desenvolvem na resistência efetiva, na identificação dos patógenos e na ativação de outros genes de resistência (JONES; DANGL, 2001). A classe IV são as Serine/Threonine Kinase (STK), que possuem um domínio LRR extracelular, um TrD e um domínio KIN intracelular. As STK interagem com outras proteínas numa ampla gama de processos na resistência a doenças (AFZAL et al. 2007). A classe V são aquelas que contêm um domínio LRR extracelular junto com um domínio PEST (Pro-Glu-Ser-Thr), que é responsável pela degradação de proteínas e de pequenos motivos delas por meio de endocitose dirigida por outros receptores de proteína (GURURANI et al. 2012). Na classe VI estão as proteínas que contêm um domínio TrD e um CC putativo. A função mais conhecida dessas proteínas é a ativação das reações HR, e outros genes de resistência, pela via do ácido salicílico (YINGQIANG, et al., 2009).

A classe VII são as proteínas semelhantes às TNLs, mas que no seus extremos C-terminais contêm uma provável Sinal de Localização Nuclear (NLS - *Nuclear Localization Signal*) e um domínio chamado de WRKY, que é uma região de 60 aminoácidos caracterizada pela presença do motivo WRKYGQK, junto com um motivo de tipo *zinc-finger*. Sua função está relacionada com a ativação de fatores de transcrição (DESLANDES et al., 2002). A última classe de genes é a classe VIII, que inclui proteínas com funções enzimáticas que contêm também domínios NBS e LRR e sua função principal é a inativação de fatores de virulência específicos e produzidos pelos agentes patogênicos (GURURANI, et al., 2012).

Figura 7 – Representação esquemática da estrutura e principais dominós das NBS-LRR



Fonte: baseado no modelo de MCHALE et al., 2006.

TIR= Receptor semelhante a Toll/Inteleukin-1, NBS= Sítio de União ao nucleotídeo, LRR = Repetições ricas em Leucinas, CC = *coiled-coil*, X = dominós variável.

Os TNL e CNL são os RG's mais conhecidos e numerosos (MCHALE et al., 2006). Muitos dos motivos que conformam estas proteínas são utilizados como objetivos de marcadores celulares de tipo efetor ou sinalização descendente (MUKHTAR, et al., 2011). Os domínios LRR no extremo C-terminal são altamente variáveis e estão conformados por séries de folhas β . Acredita-se que são os LRR os relacionados com interação proteína-proteína (SHARMA, et al. 2009).

1.5.6 RG's DE CAFÉ

Noronha-Wagner e Bettencourt (1967) identificaram nove fatores de resistência relacionados com resistência vertical a ferrugem do café, eles chamaram de S_H1 a S_H9 (SERA, et al., 2007). Os fatores de resistência S_H1, S_H2, S_H4 e S_H5 foram identificados em plantas de cafés arábicos principalmente da Etiópia. Por outro lado, o fator S_H3 presume-se que tem sua origem em *C. liberica*, uma espécie de café que vem sendo muito utilizada em programas de melhoramento de café. Finalmente, os fatores S_H6, S_H7, S_H8 e S_H9 têm sua origem em *C. canephora*. (SERA, et al., 2007).

CNL são as RP mais numerosas em plantas (tanto monocotiledôneas quanto dicotiledôneas) e são altamente relacionadas com a resistência efetiva delas e as reações HR (MOFFETT, et al., 2002). No caso do café, é conhecido que o gene S_H3, codifica uma proteína semelhante à CNL e é um importante fator de resistência à ferrugem. Cópias do gene são distribuídos em duas regiões separadas por mais de 160Kb. A primeira região, chamada região "A", é composta por duas ou três cópias do segmento do gene no mesmo sentido. A outra região, chamada de "B", contém uma ou duas cópias do mesmo gene em orientações opostas à região "A" (RIBAS, et al., 2011).

1.6 IDENTIFICAÇÃO DE NBS-LRR's

Recentemente, têm sido identificadas sequências relacionadas com proteínas do tipo NBS-LRR envolvidas em processo de resistência e imunidade em genomas de diferentes plantas (YU, et al., 2013, ZHU, et al., 2014, ZHAO, et al., 2016, ZHONG; ZHANG; CHENG, 2018, NEUPANE, et al., 2018). Utilizando uma abordagem bioinformática é possível não só identificar sequências relacionadas com estas proteínas, mas também fazer anotações e outras análises. O aumento do número de genomas e transcriptomas públicos gerados com tecnologia de Sequenciamento de Nova Geração ou NGS (*Next-Generation Sequencing*) provém da informação que pode ser utilizada na procura de sequências relacionadas com proteínas de interesse biotecnológico, industrial ou de uso em melhoramento de espécies agrícolas (LI, et al, 2016).

Os primeiros projetos relacionados com a identificação de RGA's aplicaram uma abordagem nas que são utilizadas sequências de proteínas previamente relatadas para o desenho de *Primers*, que permitem amplificar, clonar e sequenciar os genes das proteínas de interesse (JOSHI e NAYAK, 2013. WAN, et al, 2012. MARONE, et al. 2013). Mas as novas ferramentas de bioinformática que permitem identificar, traduzir, comparar e prever as possíveis estruturas e funções de sequências de nucleotídeos e aminoácidos facilitam a identificação e caracterização deste tipo de sequências (LI, et al., 2016).

Os Modelos de Markov Ocultos ou HMM (*Hidden Markov Models*) têm sido muito utilizados para o desenvolvimento de softwares bioinformáticas, principalmente na procura de perfis probabilísticos, que permitem identificar membros de uma família de proteínas dentro de bases de dados públicas (DeFONZO; ALUFFI PENTINI; PARISI, 2007). Bases de dados como Pfam, InterProscan, Unitpro, e softwares como MEME, HMMER, entre outros utilizam HMM como algoritmo para a identificação de padrões em sequências de nucleotídeos ou aminoácidos que possam ser relacionados como domínios, motivos ou famílias de proteínas específicas. Utilizando estas ferramentas é possível identificar sequências, fazer alinhamentos, anotações e outros tipos de análises que permitam classificar e caracterizar sequências de interesse (KROGH, 1998). Os estudos relacionados com a identificação de sequências de RGA's utilizando ferramentas bioinformáticas são comuns, mas existem poucos relacionados com café sendo a maioria deles pouco específicos ou antigos.

1.6.1 PROCEDIMENTO PARA A IDENTIFICAÇÃO DE RGA'S

Sekhwal (2015) descreveu o procedimento comum utilizado para a identificação de RGA's de plantas com ferramentas bioinformáticas. O procedimento normalmente inclui quatro passos básicos: o primeiro que inclui a construção de uma base de dados com RGA's de referência, sendo as bases de dados mais utilizadas para esse propósito o GenBank e a PRGdb (*Planta R-Genes Database*) (SANSEVERINO, et al. 2010). O segundo passo inclui o uso de ferramentas de alinhamento de sequências para identificar sequências candidatas a RGA. Dentre os softwares mais utilizados, é possível mencionar o BLAST (*Basic Local Alignment Search Tool*), em que normalmente é utilizado um *e-valeu* de corte (entre $1e-5$ a -1) que permita filtrar somente aquelas sequências altamente relacionadas com os RGA curados (SEKHWAL, et al., 2015). O terceiro passo inclui a utilização de múltiplas ferramentas bioinformáticas capazes de identificar domínios, motivos e repetições de proteínas, tais como PFAM, InterProScan, ClustalW, entre outros. O último passo classifica as sequências candidatas a RGA baseadas nos resultados das análises anteriores (SEKHWAL, et al., 2015).

OBJETIVOS

OBJETIVO GERAL.

- Identificar proteínas de resistência do tipo NBS-LRR envolvidas na resistência efetiva para a ferrugem do café utilizando uma abordagem bioinformática.

OBJETIVOS ESPECÍFICOS

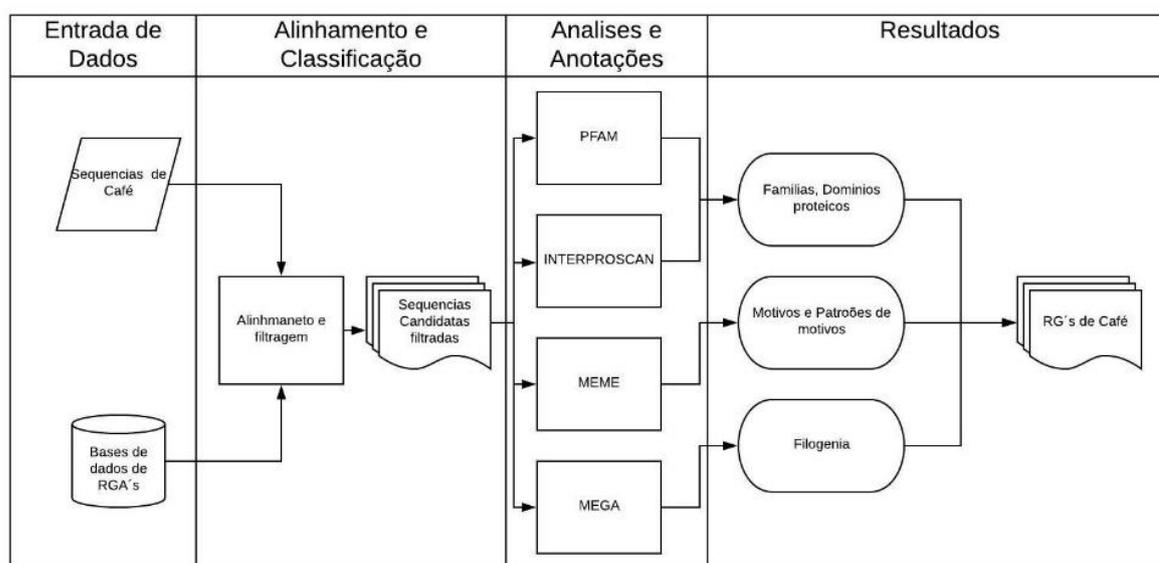
- Identificar e classificar sequências de nucleotídeos relacionados com proteínas de resistência em bases de dados públicas de Café utilizando ferramentas de bioinformática;
- Analisar as relações filogenéticas entre as sequências candidatas a gene de resistência (RG) de *C. arabica* e *C. canephora*;
- Identificar genes de resistência candidatos a serem utilizados em programas de melhoramento da resistência genética a ferrugem de variedades do café arábico.

3.MATERIAIS E MÉTODOS

3.1. DESENHO DO DIAGRAMA DE FLUXO DE TRABALHO

O método utilizado para identificar e classificar RG's de café foi baseado no método proposto por SEKHWAL et al. (2015) e LI, et al., (2016). Desenhamos um diagrama de fluxo de trabalho adaptado a nossa necessidade, incluindo os quatro passos dos procedimentos padrão para a identificação de RG's (Figura 8).

Figura 8 – Diagrama de fluxo para a identificação de RG's em seqüências públicas de Café



Fonte: Elaboração do autor

No primeiro passo são descritas as entradas dos dados biológicos para serem analisados, assim como a base de dados de RGA's utilizada para identificar as seqüências candidatas. No segundo passo, mostram-se os alinhamentos e a filtragem dos dados em categorias baseadas na classe de proteínas de resistência. No terceiro passo, são descritos os softwares utilizados nas análises de anotações e filogenia. Por último no quarto passo, são descritos os resultados das análises. As setas pretas apontam o fluxo do processo.

Depois procuramos a informação de seqüências de nucleotídeos disponíveis, gratuitos, e públicos de RGA's e Café com o objetivo de gerar as bases de dados utilizadas na primeira etapa do trabalho (arquivos *Input*). As bases de dados foram alinhadas entre elas e filtradas para selecionar somente aquelas seqüências que possuem alta similaridade com as seqüências RGA's utilizadas como isca. As seqüências filtradas foram submetidas a análises de busca de motivos, domínios proteicos, famílias etc. Finalmente as seqüências relacionadas com as classes de proteínas de interesse foram submetidas a análise filogenética.

Paralelamente, nas bases de dados públicas de café, que possuem software de análise *on line*, foram feitas busca por Palavras Chaves ou KS (*Keyword Search*) com o objetivo de identificar

sequências previamente reportadas e filtrá-las dos resultados. Esta metodologia também foi utilizada nas bases de dados de RGA's, mas com o objetivo de classificar as sequências regatadas e eliminar dados redundantes. Posteriormente, as sequências classificadas foram utilizadas em alinhamentos com as sequências de café. Foram feitas anotações, análises filogenéticas e, por fim, foram desenhados *primers* com o objetivo de identificar possível genes de resistência em amostras de DNA/RNA de café e genes candidatos ao uso em programas de melhoramento genético de variedades de café resistentes a doenças, especialmente aqueles relacionados com resistência à Ferrugem.

3.2. DATA MINING

Neste estudo, foram utilizados unicamente sequências de bases de dados públicas de café e de RGA's. Todas a sequências foram obtidas em formato *fasta* e foram armazenadas em arquivos classificados por espécie de café (no caso das bases de dados de café) e pelo tipo de RGA's (no caso das bases de dados de RGA's de referência).

3.2.1 BASES DE DADOS DE CAFÉ

Os dados do café foram obtidos de três bases de dados públicas as quais são descritas a seguir:

- A primeira base de dados utilizada foi o repositório do *Brazilian Coffee Genoma Project Consortium* (BCGP) o qual possui uma base de dados criada e administrada pela Universidade de Campinas (UNICAMP), pela Empresa Brasileira de Pesquisa Agropecuária (Embrapa) e a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), dentre outras. Este site fornece resultados do “*Brazilian ESTs sequencing*”, incluindo contigs, SNPs, anotações, domínios de proteínas, entre outros (MONDEGO et al., 2011). Desta base de dados foi recuperada sequências de nucleotídeos de três espécies de café: *C. arabica*, *C. canephora*, e *C. racimosa*, assim como sequências do Híbrido de Timor (TH = *C. arabica* x *C. canephora*).
- A segunda base de dados utilizada foi o *Coffee Genome Hub* (CGH), uma base de dados francesa feita e atualizada pelo IRD (*Institut de Recherche pour le Dévelpppement*) e o CIRAD (*Agricultural Research Center for International Development*). Os dados disponíveis no site são sequências completas do genoma de *C. canephora* juntamente com a estrutura do gene, pseudo-cromossomos, informação do produto gênico, metabolismo, classes de genes, transcriptômica (ESTs, RNA-Seq), marcadores

genéticos e mapas genéticos, dentre outros (DENOEUDE et al., 2014). Do CGH recuperaram-se os *contigs* do genoma de *C. canephora* com anotações do pseudo-cromossomos individuais e o genoma montado em conjunto.

- A terceira base de dados utilizada foi *Coffea arabica Genome* (CAG), uma base de dados feita pelo consórcio World Coffee Research. Esta contém dados dos *scaffolds* do sequenciamento do genoma da variedade Bourbon de *C. arabica*. É importante destacar que esta última base de dados foi lançada em 2018 e ainda não possui publicações relacionadas. No entanto, a montagem do genoma encontra-se liberada para pesquisa. Dessa base de dados foram recuperados todos os *scaffolds* do sequenciamento de *C. arabica* var. Bourbon.

3.2.2 BASES DE DADOS DE RGA's

As sequências de RGA's foram obtidas de duas bases de dados, utilizando os mesmos termos usados nas bases de dados de café como palavras-chaves. As bases de dados utilizadas foram:

O Phytozome v12.1, é uma base de dados comparativa de genomas de plantas sequenciados pelo Departamento da Energia dos Estados Unidos (*US. Energy Department*) e o Programa *JGI Plant Science*. O Phytozome é uma base de dados que contém os genomas anotados de mais de 90 espécies diferentes de plantas (GOODSTEIN et al., 2011). Desta base de dados foram resgatadas sequências anotadas de RGA's das plantas disponíveis.

- O PRG data-base é uma base de dados especializada no armazenamento, cura e discussão de RG de plantas. É administrada pelo CRG (*Center for Genomic Regulation*), a *Università Degli Studi de Napoli Federico II*, o CRAG (*Centre for Research In Agricultural Genomics*) e a Plataforma SEQUENTIA do Reino da Espanha (SANSEVERINO et al., 2010). Desta base de dados foram resgatadas exclusivamente as sequências curadas, anotadas e publicadas dos RGA's das espécies disponível nas bases de dados.

Todas as sequências foram salvas em arquivos “.fasta” nomeadas utilizando um formato similar ao das bases de dado de café. Finalmente, as sequências das duas bases de dados foram misturadas num só arquivo “.fasta” baseado no tipo de RGA’s que eles contêm.

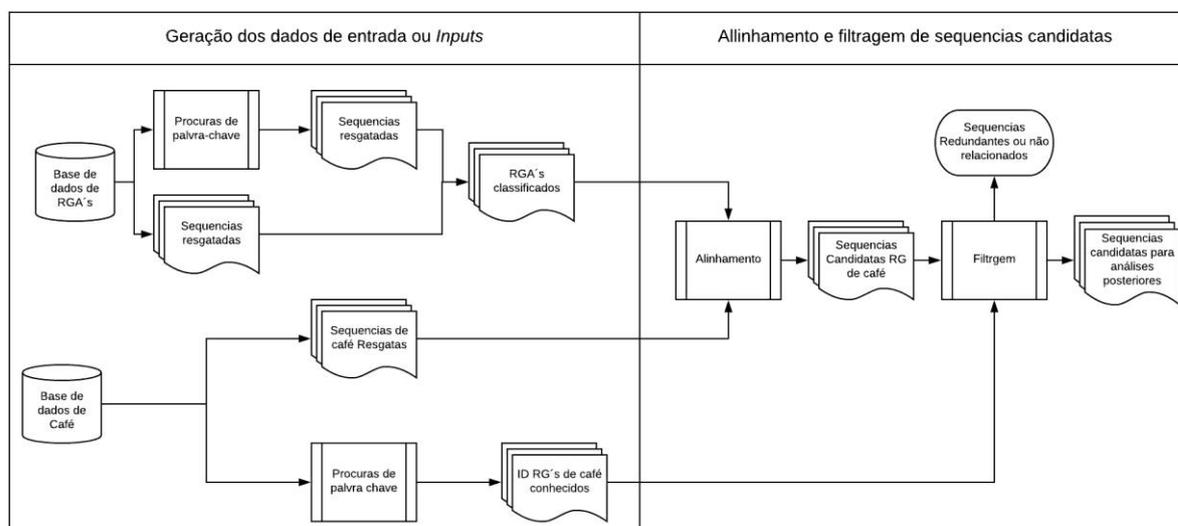
3.2.3 BUSCA POR PALAVRAS-CHAVES

O BCGP e CGH possuem ferramentas web para análise de sequência e, foram feitas KS nestas duas bases de dados de café utilizando aquelas “palavras” relacionadas com proteínas da classe NBS-LRR. As palavras utilizadas foram: “NBS”, “LRR”, “NBS-LRR”, “CNL”, “TNL” ou variantes como “CC-NBS-LRR”, “TIR-NBS-LRR” entre outras. Neste caso, armazenou-se somente os ID das sequências que posteriormente foram eliminadas dos resultados dos alinhamentos, utilizando um *script* na plataforma RStudio com o pacote *seqinr*. Também foram feitas KS no site do Phytozome utilizando os mesmos termos que nas bases de dados de café, mas com o objetivo de classificar as sequencias resgatadas.

3.3. ALINHAMENTOS E FILTRAGEM

Foram feitos alinhamentos com os dados resgatados das bases de dados de café e RGA’s, utilizando-se o software NCBI-BLAST v2.7.1 na versão livre para Linux. Utilizou-se os parâmetros padrão do *blastn* ($e\text{-value} < 1e\text{-}5$) e os resultados foram armazenados em arquivos “.csv”. Os arquivos dos resultados dos alinhamentos e os arquivos “.fasta” das bases de dados de café foram utilizados junto como *scripts* na plataforma RStudio com o pacote *seqinr* para resgatar as sequências candidatas a RG de café. Foram gerados arquivos individuais para cada classe de RG e, paralelamente, foram eliminadas todas a sequências redundantes dos arquivos gerados (Figura 9).

Figura 9: Sub-etapas do primer e segundo passo do diagrama de fluxo de trabalho para a identificação de RG's em café.



Fonte: Elaboração do autor

No primeiro passo, apresentam-se os arquivos e processos utilizados para a construção dos arquivos utilizados nos passos seguintes. No segundo passo, são apresentados os processos de alinhamento de sequências e filtragem de dados. As setas pretas demonstram o fluxo do processo.

3.4 ANÁLISES PARA ANOTAÇÕES

As sequências de café resultantes das análises anteriores foram submetidas a análises de anotações, desenho de primer degenerados e análises filogenéticas (Figura 10). No caso das anotações, o objetivo foi coletar toda a informação possível das sequências para poder classificá-las com base nas características funcionais e estruturais. Para cumprir este objetivo, as seguintes análises foram realizadas:

3.4.1 BUSCA DE PADRÕES DE MOTIVOS

A busca de padrões de motivos foram feitas utilizando o pacote de ferramentas MEME Suite v5.0.1 (*Motivo-Based Sequence Analysis Tools*) na sua versão web. O MEME Suite é um kit de softwares que permite os usuários fazerem diferentes tipos de análises ou procura de motivos, seja com motivos de referência ou buscas em bases de dados de proteínas. (TIMOTHY, et al., 2009). A busca de motivos é normalmente utilizada para procurar sítios ativos de enzimas, identificar regiões envolvidas em estrutura e estabilidade proteica, identificar características químicas particulares (termo estabilidade) e classificação de proteínas em famílias (TIMOTHY, 2008). Nós utilizamos o MEME Suite para procurar padrões de motivos e classificação de famílias de proteínas. Foram utilizados os parâmetros padrão do MEME modificando apenas o

número de motivos a serem procurados, que em nosso caso foi de 30 motivos, um padrão muito utilizado neste tipo de análises (ZHAO, et al., 2016). Os resultados do MEME foram armazenados em arquivos “.txt” e “.jpg”.

3.4.2 PROCURA DE DOMÍNIOS DE PROTEÍNAS

As sequências resultantes dos alinhamentos anteriores foram utilizadas para identificar domínios relacionados como RG's. Utilizou-se para esse fim o Pfam 32.0 e o InterProScan v70.0. O Pfam é uma base de dados de famílias de proteína, cada uma das quais foi definida por pelo menos dois alinhamentos de sequências e um perfil de HMM (FINN, et al., 2014). A versão web do Pfam permite fazer análises de procura de domínios, motivos e repetições em sequências tanto de DNA e RNA como de proteínas. (FINN, et al., 2015). Nós submetemos à análise pelo Pfam as sequências de nucleotídeos resultantes das análises anteriores. Todos os resultados foram armazenados num arquivo “.csv”. Por fim as sequências que não apresentaram resultados relacionadas com as famílias de interesse para este estudo foram eliminadas e não foram utilizadas para as análises posteriores.

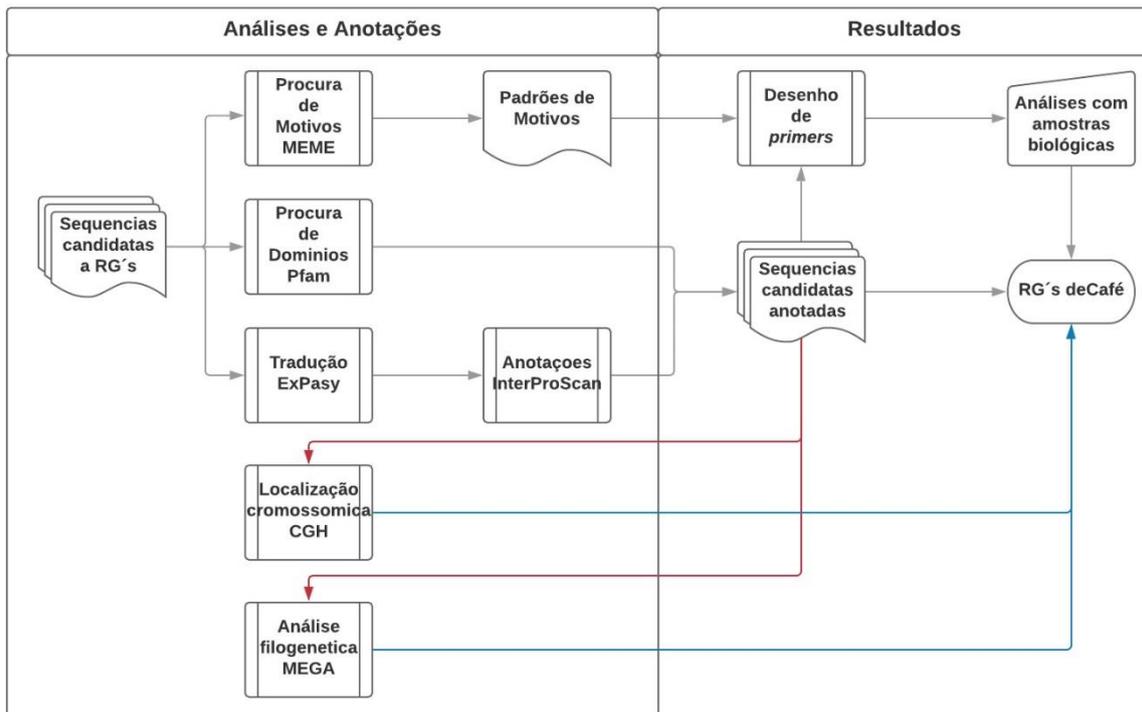
As sequências também foram analisadas no Interproscan na versão web, uma base de dados utilizadas para classificar proteínas dentro de famílias, prever a presença de domínios importantes e sítios ativos, assim como repetições e motivos, utilizando diferentes bases de dados de proteínas (FINN, et al. 2017). Nós traduzimos as sequências resultantes das análises anteriores e utilizamos apenas a região codificante das sequências na análise. As traduções foram feitas no software ExPASy *Translate* e, os resultados do ExPASy foram armazenados em arquivos “.fasta” utilizando o ID da sequência de nucleotídeos correspondente, como identificado nas sequências de aminoácidos resultado da tradução. Os resultados do Interproscan foram anexados aos resultados do Pfam num arquivo “.csv”.

3.5. DESENHO DE PRIMERS DEGENERADOS

Foram utilizados os resultados do MEME e as sequências filtradas da análise de anotações para o desenho de *primers*. Utilizamos o Software AliView v1.22 para desenhar os *primers* na interface gráfica do software. Primeiro as sequências foram alinhadas com algoritmo MUSCLE (*Multiple Sequence Comparison by Log-Expectation*), as sequências foram agrupadas de acordo com o nível de similaridade entre elas e depois, em cada grupo de sequências, foram identificados os motivos reportados pelo MEME. Finalmente, foram desenhados *primers* degenerados (*Degenerate Primer*) baseados no nível de degeneração de

cada motivo, e utilizamos somente aqueles motivos que tivessem um nível de degeneração menor de 1000. Confirmaram-se todos os parâmetros dos DP (GC%, *primer melting temperature*, *self-dimerization*, *hetero-dimerization*, *hairpin*) (BURPO, 2001) utilizando o Software OligoAnalyzer v3.1.

Figura 10 – Diagrama das sub-etapas do terceiro e quarto passo do diagrama de fluxo de trabalho para a identificação de RG's de café.



Fonte: Elaboração do autor

No terceiro passo são descritas as análises feitas nas sequências candidatas. No quarto passo são descritos os resultados das análises. As setas pretas mostram o fluxo do processo e, as setas vermelha e azul, o processo de *feedback* nas análises de localização cromossômica e filogenias, os quais precisaram de sequências previamente anotadas.

3.7 ANÁLISES FILOGENÉTICAS

Foram utilizadas para as análises filogenéticas as sequências agrupadas na análise anterior. De cada grupo foi gerada uma sequência chamada de “sequências representativas”, que além de serem as mais conservadas, foram as que tiveram a maior cobertura das sequências do mesmo grupo. As sequências representativas foram editadas e posteriormente, submetidas a análise no software MEGA X (*Molecular Evolutionary Genetics Analysis*) na sua versão para Linux. Às sequências candidatas para RG de café, foram anexadas sequências de CNL de referência de outras espécies de plantas de acordo com a classe de RG analisada. As Árvores

foram construídas utilizando o método de inferência de Máxima Verossimilhança ou ML (*Maximum Likelihood*), o test filogenético de *Bootstrap* com 1000 replicatas, o modelo de substituição de nucleotídeos utilizado foi o *General Time Reversible Gamma* distribuído com sítios invariantes. (*Gamma Distributed with Invariant Sites* ou GTR+G+I) e um *cutoff* de 90%.

3.8 LOCALIZAÇÃO CROMOSSÔMICA

Finalmente, as sequências candidatas de *C. arabica* resultantes das análises anteriores foram submetidas à localização cromossômica na base de dados CGH, base de dados que oferece uma montagem completa do genoma de *C. canephora* e a estrutura dos 11 pseudo-cromossomes. Ela tem sido utilizada nos projetos de sequenciamento de *C. arabica* como genoma de referência pela relação evolutiva que as duas espécies compartilham e também porque oferece ferramentas para realizar análises de sequências. Utilizamos o Blastx do CGH para identificar a localização de loci relacionado com as sequências candidatas a RG de café. Os resultados desta análise foram armazenados em arquivos “.csv” e “.jpg”.

4 RESULTADOS E DISCUSSÃO

4.1. DATA MINING

4.1.1 BASES DE DADOS DO CAFÉ

Foram resgatadas um total de 127.491 sequências de café (Tabela 1). Os dados resgatados foram armazenados e classificados por tipo de base de dado e espécie ou variedade de café. Um total de seis arquivos com sequências de nucleotídeos foram gerados com os dados resgatados das bases de dados de café.

Tabela 1: Resultados do processo de *Data Mining*.

Base de dado/Espécie de café	Número de Sequências
Cc-CGH (<i>C. canephora</i>)	25.574
Ca-BCGP (<i>C. arabica</i>)	4.147
Cc-BCGP (<i>C. canephora</i>)	7.684
Cr-BCGP (<i>C. racimosa</i>)	3.113
TH-BCGP (Híbrido de Timor)	8.662
Ca-WCR (<i>C. arabica</i>)	78.311

Fonte: Elaboração do autor

*Apresentam-se os nomes dos arquivos contendo a sequências resgatadas e o número de sequências contido em cada arquivo.

4.1.2 BASES DE DADOS DE RGA's

Resgatou-se um total de 43.252 sequências relacionadas com RG's das duas bases de dados. Delas, apenas cento e doze sequências de referência foram resgatadas da PRGdb. O resto de sequências de RGA's foram resgatas por meio dos KS no Phytozome.

4.1.3 PROCURA DE PALAVRA-CHAVE (KS)

Da base de dados CGH foram resgatados 1.751 ID de sequências no total, da base de dados BCGP foram regatados 214 ID e por último, do Phytozome foram resgatadas 43.140 sequências relacionadas com as “palavras-chaves” utilizadas (Tabela 2). É importante destacar que as sequências identificadas com a “palavra” LRR foram sempre as mais numerosas. Os LRR operam numa ampla variedade de interações moleculares no processo imune das plantas e estão presentes em quase todas as classes de proteínas de resistência, o que poderia explicar a grande quantidade de sequências identificadas (AYLWIN; RAMNIK, 2011). Os resultados para a “palavra” NBS apresentu menor número de sequências que no caso anterior, mas sempre

apresentou uma distribuição similar em todas as bases de dados. O domínio NBS está presente em diferentes famílias de proteínas, não só envolvida na imunidade e resistências dos organismos senão também e outra funções metabólicas diferentes (PARCA, et al., 2012). Observou-se que a maioria das sequências identificadas com essa “palavra” não apresentavam relação com proteínas de resistências.

Tabela 2: Número de sequências recuperadas por busca de palavras-chaves nas bases de dados do café e de RG

Nomes/ Keywords	NBS	LRR	CNL	TNL
Cc-CGH	25	1075	586	65
Ca-BCGP	66	68	3	-
Cc-BCGP	28	34	2	-
Cr-BCGP	6	7	-	-
RG-Phytozome	2.438	33.981	708	798

Fonte: Elaboração do autor.

* (“-“ = Não foram encontrados resultados). Apresentam-se os nomes das bases de dados e o número total de sequências resgatadas nos alinhamentos com sequências de referência de RGA’s.

Finalmente, no caso das “palavras” CNL e TNL os resultados contrastam com o reportado para *Arabidopsis sp*, em que o número de sequências relacionadas como TNL é maior que as relacionadas como CNL (MEYERS, et al., 2003). Nós identificamos uma maior proporção de sequências relacionadas como CNL nas bases de dados utilizadas. Esta característica poderia indicar que os CNL de café provêm de um ancestral mais antigo que os TNL, ou que os CNL de sofreram uma quantidade maior de eventos de duplicação e recombinação no processo de co-evolução planta-patógeno.

4.2 ALINHAMENTOS E FILTRAGEM

Ao todo, 259 sequências de nucleotídeos das duas espécies de café estudadas apresentaram alta similaridade com sequências de referências relacionadas com proteínas CNL e 95 com TNL uma relação de quase 3 sequências CNL para uma TNL. As bases de dados que apresentaram maior número de sequências alinhadas foram: CGH (*C. canephora*) e a WCR (*C. arabica*), que representam 81% das sequências identificadas (Tabela 3). A base de dados BCGP, infelizmente, não apresentou sequências novas para nenhuma das classes de proteínas de resistência de interesse.

Tabela 3: Número de sequências candidatas a RG's resultado dos alinhamentos sem redundâncias.

Bases de dados/RG	CNL	TNL
Cc-CGH	40	26
Cc-BCGP	17	10
Ca-BCGP	8	6
Cr-BCGP	2	1
TH-BCGP	9	10
Ca-WCR	183	42

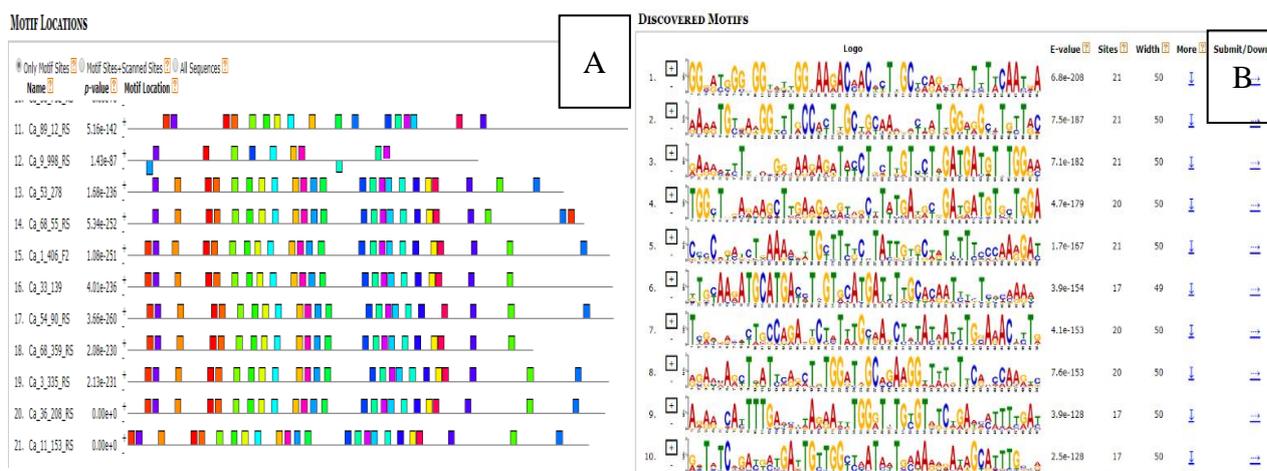
Fonte: Elaboração do autor.

Os resultados obtidos da base de dados CGH também tinham sido reportados anteriormente, mas algumas das anotações desta base de dados não estão de acordo com as anotações feitas neste estudo. Algumas sequências da CGH não apresentam os domínios típicos das proteínas com as quais foram relacionadas originalmente, ou não possuem todos os domínios relacionados com elas. As sequências dos alinhamentos e KS da CGH foram filtradas para eliminar dados redundantes utilizados nas análises seguintes com o objetivo de melhorar as anotações feitas previamente e aumentar a informação disponível destas sequências. Finalmente, todas as sequências identificadas na base de dados WCR foram submetidas às análises posteriores em vista que, esta base de dados ainda não possui publicações relacionadas com as sequências identificadas.

4.3 BUSCA DE MOTIVOS

Por meio da análise de motivos, foi possível agrupar as sequências baseado no padrão de motivos apresentado pelo MEME na sua interface gráfica. Os gráficos de linhas e caixas apresentam os motivos identificados com maior frequência nas sequências, assim como a sua localização (Figura 11). O gráfico de “Motivos Descobertos” (*Discovered Motifs*) apresenta as sequências dos motivos identificados na análise. Utilizou-se esta informação para identificar os melhores candidatos para o desenho de *primers*. No gráfico “Localização de Motivos” (*Motif Locations*) é possível observar o padrão de motivos que possuem as sequências. Estes padrões ajudaram a classificar as sequências, eliminando aquelas que possuem pouca similaridade e, sequências pequenas, além de auxiliar na identificação de sequências que poderiam precisar maior edição. Os padrões de motivos dos nossos resultados foram muito similares aos reportados por Zhao et al. (2016).

Figura 11 – Exemplo dos resultados da interface gráfica do software MEME Suite.



Fonte: Elaboração do autor.

A: apresenta o gráfico de Localização de Motivos, as caixas coloridas representam os motivos identificados e a localização nas sequências submetidas. B: apresenta o gráfico de Descoberta de Motivo, apresentam-se a sequência de nucleotídeos ou aminoácidos que compõem os motivos identificados pelo software.

Um resultado importante obtido nesta análise foi a identificação dos motivos *Ploop* e *GLPL* nas sequências de café arábico e robusta (Figura 12). Estes têm sido relacionados com o domínio NBS das proteínas de resistências das classes CNL e TNL de plantas, e têm sido utilizados como marcadores para a identificação de sequências relacionadas com resistência a doenças em outros estudos (NOIR, et al. 2004). Outro resultado interessante foi a identificação de quatro motivos conservados relacionados com o domínio NBS e presentes exclusivamente nas sequências de Ca-WCR, e quatro mais nas sequências de Cc-CGH 2 relacionadas com o domínio NBS e dois mais relacionadas com o domínio LRR. Também foi possível reconhecer padrões de motivos que permitem diferenciar variantes do domínio CC no N-terminal que, no caso do café arábico foi de três diferentes padrões e no caso de café robusta foi de quatro (ANEXO 4). Os resultados do MEME foram utilizados junto aos resultados das anotações para identificar motivos que possam ser utilizados para o desenho de *primers*.

Figura 12 – Localização do motivo *Ploop* em seqüências de café robusta.



Fonte: Elaboração do autor.

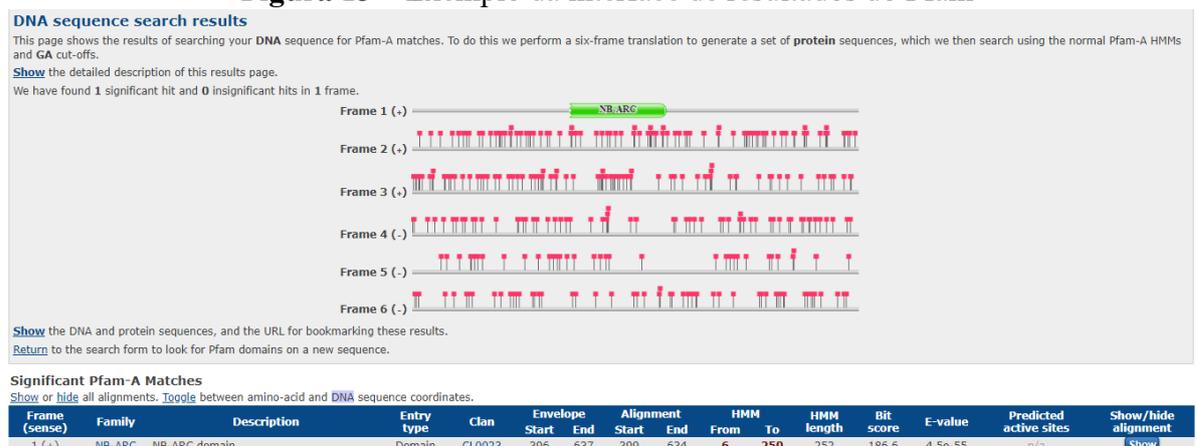
O quadro vermelho amostra a localização do motivo *Ploop* num alinhamento múltiplo feito com o software AliView.

4.4. PROCURA DE DOMÍNIOS E ANOTAÇÕES

Todas as seqüências de CGH e WCR, resultado da filtragem, foram submetidas a análises de anotações nos softwares Pfam e InterProScan (Figura 13). Foram selecionadas somente aquelas seqüências nas quais foi possível identificar domínios relacionados com as classes CNL e TNL.

Para *C. canephora* foram identificadas 260 seqüências entre KS e alinhamentos. Nelas foi possível identificar pelo menos dois domínios (CC-NBS) relacionados com proteínas do tipo CNL, sendo estes os mais numerosos, uma característica similar ao relatado em outros estudos (PLOCIK et al. 2004, WAN et al. 2012, LOZANO et al. 2015). Para TNL foram identificadas seis seqüências relacionadas com essa classe de proteínas, o que difere com o relatado para *Arabidopsis sp.*, em que o maior número de seqüências relacionadas como proteínas de resistência codificam proteínas de tipo TNL (MEYERS, et al., 2003). Este resultado poderia indicar que os genes de resistência de café do tipo TNL poderiam ter se desenvolvido mais recentemente que os CNL no processo de especiação, como têm sido observados em outras famílias de planta (SEO, et al., 2016).

Figura 13 – Exemplo da interface de resultados do Pfam



Fonte: Elaboração do autor.

A caixa verde representa o domínio NBS de uma sequência submetida a análise, os pontos vermelhos representam os *stops* códons identificados para cada marco de leitura.

Para *C. arabica* foram identificadas 97 sequências candidatas a CNL por meio de alinhamentos de sequências com a base de dados de RGA's, mas foi possível identificar domínios relacionados como esta classe em apenas 58 sequências. Igual ao caso de robusta, os CNLs foram a classe que apresentou maior número de sequências identificadas, mas o número de CNL ficou menor que em robusta. Esta característica poderia explicar a maior susceptibilidade do café arábico a doenças, devido ao menor número de fatores que podem ser objeto de recombinação no desenvolvimento de novos RG's (JOSHI; NAYAK, 2013). É conhecido que os loci, onde são localizados este tipo de RG, são susceptíveis à recombinação, gerando novos RG's por pressão biótica, o que resulta em resistência específica a doenças (SEO, et al., 2016). O menor número de CNLs identificados no café arábico limita este processo e poderia dificultar os desenvolvimentos de novos fatores de resistência na planta.

É importante falar que a variedade Bourbon, a qual foi utilizada pelo WCR para o sequenciamento do genoma de café, é altamente susceptível à ferrugem, o que poderia sustentar os resultados obtidos nos alinhamentos, um menor número de fatores disponível no genoma da planta pode limitar a sua resposta imunitária frente a patógenos altamente adaptados como é o caso da ferrugem (SEO, et al., 2016). No caso dos TNL foram identificadas 29 sequências candidatas resultados dos alinhamentos com a base de dados de RGA's, mas não foi possível identificar nenhuma que tivesse pelo menos dois domínios típicos da classe. Este resultado poderia indicar mesmo que no caso de robusta que os TNL são de recente surgimento em cafés, ou por outro lado, que os TNL provêm de mais de um gene.

4.5. DESENHO DE PRIMERS DEGENERADOS

Para o desenho de *primers* foram utilizados os resultados das análises de motivos e as sequências anotadas nas análises anteriores. As sequências foram agrupadas com base na similaridade entre elas. Foram utilizados os motivos identificados pelo MEME para seleccionar as regiões mais conservadas nelas (Figura 14). Calculou-se o nível de degeneração dos motivos, assim como todos os parâmetros de qualidade (%GC, TM, *Hairpin*, *Self-dimerization*, *hetero-dimerization*). O nível de degeneração foi calculado com a seguinte formula.

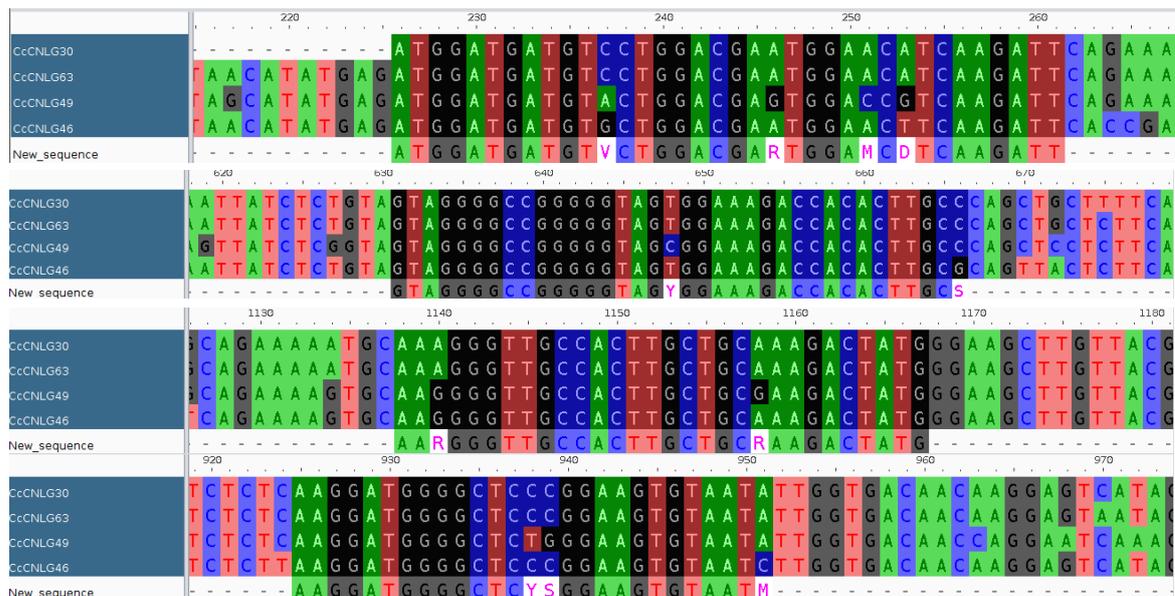
$$ND = npcAA1 * npcAA2 * npcAA_n \dots$$

Onde ND é o Nível de degeneração

$npcAA_n$ é o número de possíveis códon para o Aminoácido n

Foram desenhados 18 pares de *primers* com as sequências de robusta (ANEXO 1), os níveis de degeneração ficaram entre 192 para os *primers* menos degenerados e 1223 para os mais degenerados, níveis menores aos relatados em outros estudos (NOIR, et al. 2004). As porcentagens de CG ficaram entre 42 e 61%, e as temperaturas de *melting* entre 50 a 60 °C. Finalmente, todos os *primers* apresentaram baixa probabilidade de *Hairpin*, *self-dimerization* e *hetero-dimerization*. Para café arábico, foram desenhados também 19 pares de *primers* com características similares às de robusta (ANEXO 2).

Figura 14 – Motivos conservados utilizados para o desenho de *primers*.



Fonte: Elaboração do autor:

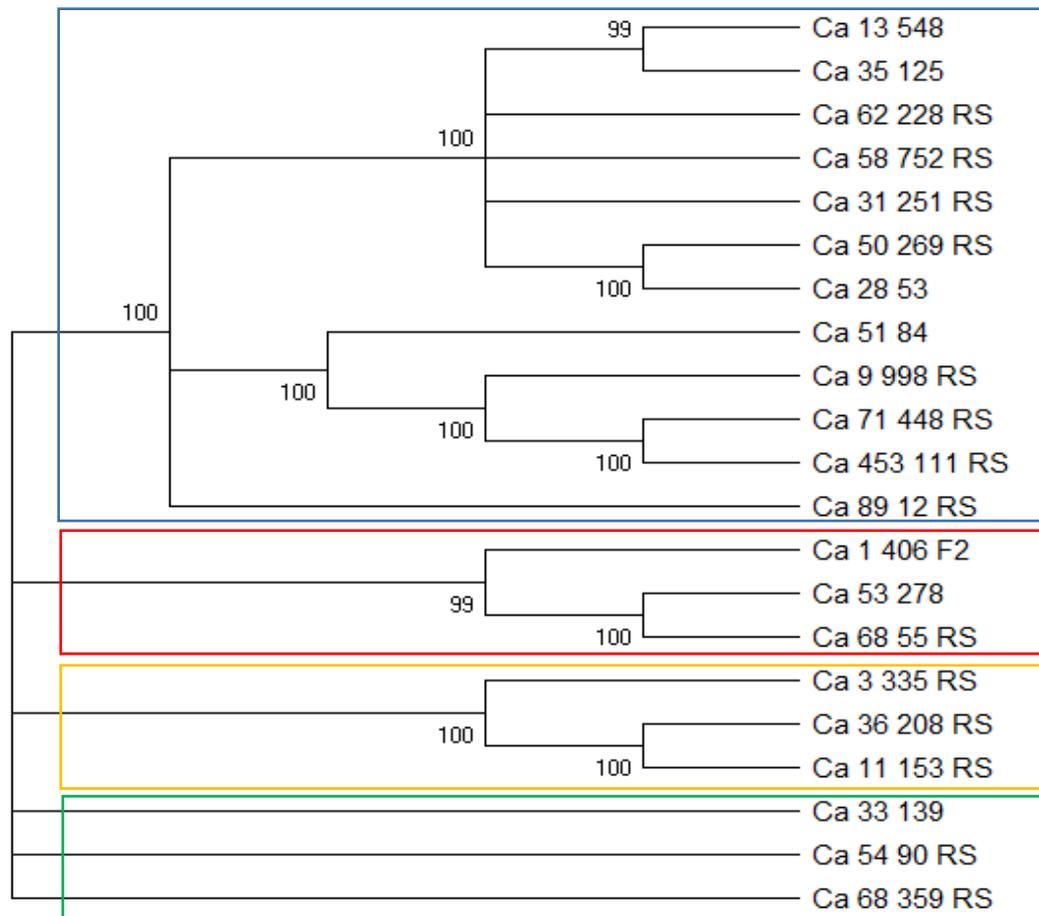
São amostrados em destaque exemplo de motivos utilizados no desenho de primer degenerados.

4.6 ANÁLISES FILOGENÉTICAS

Todas as sequências resultantes das análises de anotações foram editadas no AliView, retiradas as suas regiões não codificantes. Depois foram agrupadas com base na sua similaridade. Finalmente, de cada grupo foi gerada uma sequência chamada “Sequências Representativas” para serem utilizadas nas análises e reduzir o número de sequências redundantes. As análises filogenéticas amostraram resultados similares àqueles obtidos em outros estudos (LOZANO, et al., 2015). Uma das principais características identificadas nas sequências submetidas a análise foi a presença de quatro *clusters* principais de CNL de *C.arabica* – WCR, uma característica típica desta família e que se correlaciona com os padrões de motivos observados nos resultados do MEME Suite, (Figura 15). Outro resultado interessante foi o caso da sequência Ca_89_12 a qual é agrupada no *cluster* 1 (*Cluster azul*), mas aparece isolado dentro do cluster, mesmo apresentando os três domínios típicos dos CNL, a relação com a sequência CcCNLG20 e a sequência de referência PRGDB00061440(PI8) de girassol (*Helianthus annuus*), (ANEXO 5). Uma proteína relacionada com a resistência efetiva de *Plasmospora halstedii*, fungo patógeno responsável do nanismo do girassol, a principal doença da planta (RADWAN, 2004).

As sequências Ca_33_139, Ca_54_90 e Ca_68_359 aparecem isoladas na árvore filogenética (*Cluster verde*). As primeiras duas apresentam relações filogenéticas com a sequência CcCNLG75 do café robusta. A sequência Ca_68_359 foi relacionada exclusivamente com a sequência CcCNLG66, também de robusta. Estes resultados poderiam indicar que estas proteínas têm sua origem no sub-genoma de robusta, porém nenhuma delas apresentou relações filogenéticas com as proteínas de referência do PRGdb o que poderia indicar que elas são exclusivas do gênero *Coffea*.

Figura 15 – Árvores filogenéticas das sequências candidatas a CNL de *C. arabica*-WCR.

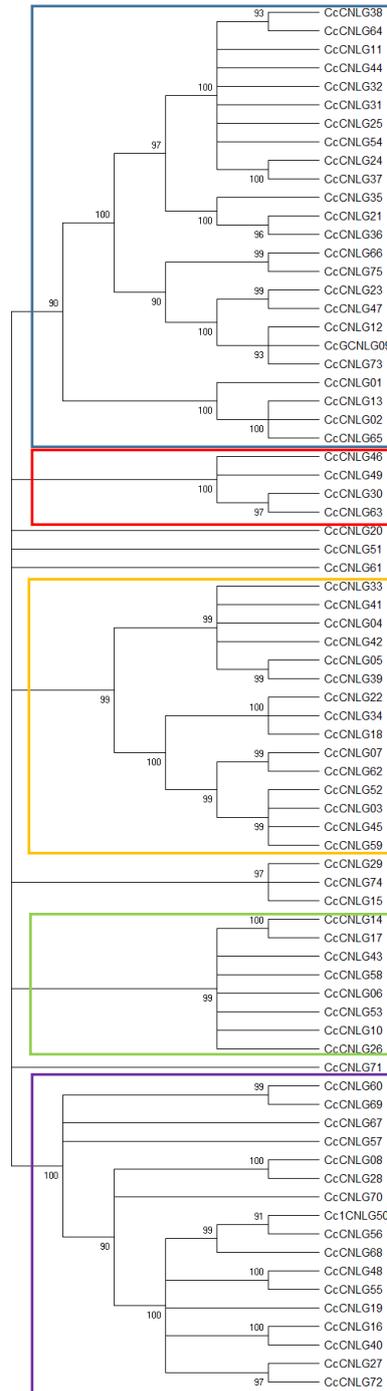


Fonte:Elaboração do autor

Método de inferência: ML, Test filogenético: *Bootstrap* com 1000 replicatas, Modelo de substituição: GTR+G+I, *Cutoff*: 90%.

A análise filogenética de café robusta apresentou um maior número de *clusters* que no caso que arábica (Figura 16), o que poderia confirmar a presença de uma maior variedade de fatores de resistência diferentes, fenômeno observado em outros resultados deste estudo; assim como maior especificidade na atividade destes fatores. Esta característica poderia ter sido desenvolvida após a especiação do café arábica, considerando que a origem desta última têm como um dos pais o café robusta. Outra possível explicação é a perda destes fatores no processo evolutivo do café arábico. Se a diversificação destes genes tivesse sido antes do processo de especiação de café arábica, este último possuiria um maior número de fatores. No entanto, a recombinação como regiões homólogas de *Coffea eugenioides* durante o processo de especiação poderiam explicar este fenômeno.

Figura 16: Árvore filogenética das sequências candidatas a CNL de *C. canephora*–CGH.



Fonte: Elaboração do autor

Método de inferência: ML, Test filogenético: *Boostrap* com 1000 replicatas, Modelo de substituição: GTR+G+I, *Cutoff*: 90%.

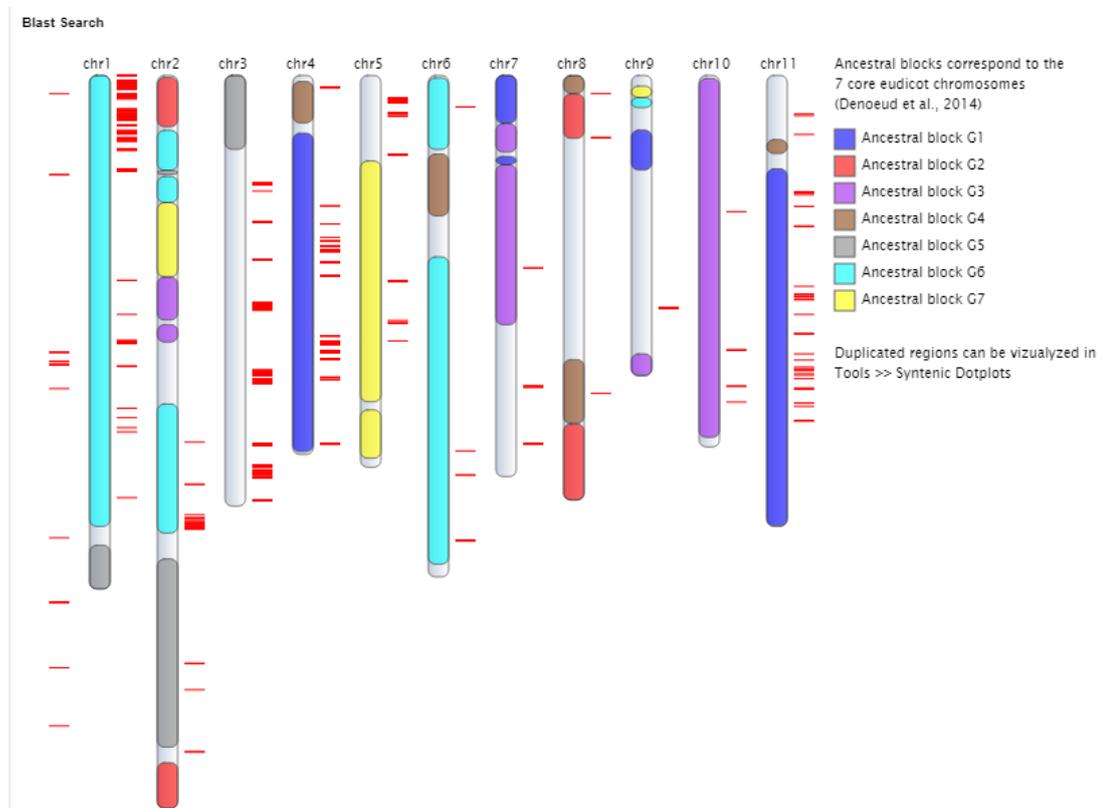
No caso do café robusta as sequências CcCNLG71, CcCNLG20, CcCNLG51 e CcCNLG61 não apresentou relações filogenéticas com outras sequências de *C. canephora*, mas sim com sequencias de café arábico (ANEXO 5). O que poderia indicar que estas proteínas são também exclusivas do gênero *Coffea*. A sequência CcCNLG71 não apresentou relaciones filogenéticas

com sequências de café arábico, e apresentou relações filogenéticas exclusivamente com a sequência de referência PRGDB00000045(Prf), uma proteína de “tomate passa” (*Solanum pimpinellifolium*), relacionada com o controle efetivo de *Pseudomonas syringae*. (CHANG, et al., 2002).

4.7 LOCALIZAÇÃO CROMOSSÔMICA

Finalmente, as sequências candidatas para CNL de café arábico foram submetidas à identificação de localização cromossômica, sendo 2239 loci relacionados com as sequências candidatas a CNL identificadas (Figura 17). Todos os cromossomos do café apresentaram regiões relacionadas com as sequências, candidatos e submetidos a análise, mas foram os blocos ancestrais 1 e 5 dos eudicotiledones que apresentaram o maior número de loci, o que poderia indicar a origem evolutiva destes genes. A identificação das regiões de cromossomos com alto número de loci relacionados como proteínas de resistência poderia ajudar nos programas de melhoramento genético que utilizam técnicas citogenéticas para o desenvolvimento de variedades de café arábica.

Figura 17 - Localização cromossômica das sequências candidatas para CNL de *C. arabica* – WCR.



Fonte: Elaboração do autor

As linhas vermelhas representam a localização dos loci relacionados com as sequências submetidas, os quadros coloridos representam os blocos ancestrais das plantas dicotiledôneas.

5. CONSIDERAÇÕES FINAIS E CONCLUSÕES

Foi possível identificar, analisar e classificar sequências de proteínas de café altamente relacionados com proteínas de resistência do tipo CNL, assim como fazer anotações que permitiram aumentar a informação disponível sobre fatores de resistência a doenças em café.

Identificamos motivos conservados exclusivos nas sequências de café arábico e robusta que, além de poderem ser utilizados para a identificação de novos fatores de resistência para uso em programa de melhoramento genético, também podem ser utilizados como marcadores moleculares para diferenciação varietal.

Identificamos sequências exclusivas tanto do café arábico como do café robusta, assim como sequências altamente relacionadas com proteínas de resistências de outras plantas relatas com antecedência.

Finalmente, é de vital importância continuar identificando possíveis genes de resistência de café de diferentes classes que permitam aumentar os fatores moleculares candidatos a programas de melhoramento genético desta planta, assim como o sequenciamento do transcriptoma do sistema patogênico *H. vastatrix* – *C. arabica*. Com isso permite-se desenvolver novas estratégias de controle de doenças em café.

6. REFERENCIAS

- ABIC (Associação Brasileira da Industria do Café). A expansão do café no Brasil
- AFZAL A.J., WOOD A.J., LIGHTFOOT D.A. Plant Receptor-Like Serine Threonine Kinase: Roles in signaling and Plant Defense. **MPMI**, v. 21, n. 5, p. 507-517. 2007.
- AVELINO, J., WILLOCQUET, L., SAVARY, S. Effect of crop management patterns on coffee rust epidemics. **Plant Pathology**, v. 53, p. 541-547. 2004.
- AVELINO, J., RIVAS, G. La roya anaranjada del cafeto. **HAL**. v. 47, 2013.
- AYLWIN, N., RAMNIK, J. X. Leucine-rich repeat (LRR) proteins Integrators of pattern recognition and signaling in immunity. **Autophagy**, v. 7, n. 9, p. 1082-1084. 2011.
- Banco Central de Honduras – BCH. Honduras em Cifras 2015-2017. Tegucigalpa. 2017.
- BELKHADIR, Y., et al. Arabidopsis RIN4 Negatively Regulates Disease Resistance Mediated by RPS2 and RMP1 Downstream or Independent of the NDR1 Signal Modulator and Is Not Required for the Virulence Functions of Bacteria Type III Effectors AvrRpt2 or AvrRpm1. **The Plant Cell**, v. 16, p. 2822-2835. 2004.
- BIGEARD, J., COLCOMBET, J., HIRT, H. Signaling Mechanisms in Pattern-Triggered Immunity (PTI). **Molecular Plant**, v. 8, p. 521-539. 2015.
- BOLLER, T., FELIX, G. A Renaissance of Elicitors: Perception of Microbe-Associated Molecular Pattern and Danger Signals by Pattern-Recognition Receptor. **Annu. Rev. Plant Biol.** v. 60, p. 379-406. 2009.
- BURPO F.J. (2001). A critical review of PCR primers design algorithms and cross-hybridation case study. Department of Chemical Engineering, Stanford University.
- CHANG, J.H. et al. Functional analyses of the Pto resistance gene family in tomato and the identification of a minor resistance determinant in a susceptible haplotype. **Mol. Plant Microbe Interact**, v. 15, n. 3, p. 281-291.
- CLARKE, R.J., MACRAE, R. (Ed.) Coffee: chemistry. 2. Ed. New York: Elsevier, p. 305, 1989.
- CONRATH, A. Systemic Acquired Resistance. **Plant Signal Behav**, v. 1, n. 4, p. 179-184. 2006
- CRAWFORD J. History of coffee. **Journal of the Statistical Society of London**, v. 15, n. 1, 1852. 50-58.
- CRISTANCHO, M.A. et al. Outbreak of coffee leaf rust (*Hemileia vastatrix*) in Colombia. **New Disease Reports**, v. 25, p. 19. 2012.
- DANGL, J.L., JONES, J.D.G. Plant Pathogens and integrated defence responses to infection. **Nature**, v. 411, p. 826-833. 2001

DAVIS, A.P., et al. An annotated taxonomic conspectus of the genus *Coffea* (Rubiaceae). **Botanical Journal of the Linnean Society**, v. 152, p. 465-512. 2006.

DeFONZO, V., ALUFFI-PENTINO, F., PARISI, V. Hidden Markov Models in Bioinformatics. **Current Bioinformatics**, v. 2, p. 49-61. 2007.

DE KOCHKO, A., et al. Advances in *Coffea* genomics. **Botanical Research**, v. 53, 2010. 23.

DENOEUDE F., et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science*, **Plant Genomics**, v. 345, n. 6201, p. 1181-1184. 2014.

DESLANDES L., et al. Resistance to *Ralstonia solanacearum* in *Arabidopsis thaliana* is conferred by the recessive RRS-1 gene, a member of a novel family of resistance genes. **PANS**, v. 99, n. 4, p. 2404-2409. 2002.

ELMORE, J.M., COAKER, G. The Role of the Plasma Membrane H⁺-ATPase in Plant - Microbe Interactions. **Molecular Plant**, v. 4, n. 3, p. 416-427. 2011.

FAO. Statistic Yearbook 2013. Chart 89: Exports of coffee, tea, cocoa and spices, constant 2005 US\$ (2000-2010). FAO Statistics Division.

FINN, R.D., et al. Pfam: the protein families' databases. **Nucleic Acids Research**. v. 42, p. D222-D230. 2014.

FINN R.D., et al. The Pfam protein families' databases: towards a more sustainable future. **Nuclei Acids Research**, v. 44, p. D279-D285. 2015.

FINN, R.D., et al. InterPro in 2017 – beyond protein family and domain annotations. **Nucleic Acids Research**, v. 45, p. D190-D199. 2017.

FLOR, H.H. Current Status of the Gen-for-Gen Concept. **Annu. Rev. Phytopathol**, v. 9, p. 275-296. 1971.

GONZALES, N. **Historia del Café en Marcala**: 1.ed. Tegucigalpa: Guaymuras, 2012.

GOODSTEIN D.M., et al. Phytozome: a comparative platform for green plant genomics. **Nuclei Acid Research**, v. 40, p. D1178-D1186. 2011.

GROß, F., DURNER, J., GAUPELS, F. Nitric Oxide, antioxidante and prooxidants in plant defense responses. **Frontiers in Plant Science**, v. 4. n. 419. 2013

GRUPTA, K.J. et al. On the Origin of nitric oxide. **Trends in Planta Science**, v. 16, n. 3, p. 160-168. 2011.

GURURANI M.A., et al. Plant disease resistance genes: current status and future directions. **Physiological and Molecular Plant Pathology**, v. 78, p. 51-56. 2012.

HARLAN, J.R. Ethiopia: A center of diversity. **Economic Botany**, v. 23, n. 4, 309-314. 1969.

HE, P., SHAN, L., SHEEN, J. Elicitation and suppression of microbe-associated molecular pattern-triggered immunity in plant-microbe interactions. **Cellular Microorganismology**, v. 9, p. 1385-1396. 2007.

HEALEY, A. et al. Protocol: a simple method for extracting next generation sequencing quality genomic DNA from recalcitrant plant species. **Plant Methods**, v. 10, n. 12. 2014.

HEESE, A. et al. The receptor-like kinase SERK3/BAK1 is a central regulator of innate immunity in plants. **Plant Biology**, v.104, p. 12217-12222. 2007.

INSTITUTO HONDUREÑO DEL CAFÉ – IHCAFE. Memoria cosecha 2016-2017. Tegucigalpa: 2017.

INTERNATIONAL COFFEE ORGANIZATION – ICO. Total Production by all exporting countries. London: 2018.

INTERNATIONAL COFFEE ORGANIZATION - ICO. Imports of coffee by selected importing countries. London: 2018.

JEWORUTSKI E. et al. Early signalling through the Arabidopsis pattern recognition receptor FLS2 and EFR involves Ca-associated opening of plant membrane anion channels. **Plant Journal**, v. 62, p. 367-378. 2010.

JIMENEZ, E. Café I. (*G. coffea*). **Reduca (Biologia). Serie Botánica**, v. 7, n. 2, 2014. 113-132.

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

JOSHI, R.K., NAYAK, S. Perspectives of genomic diversification and molecular recombination towards R-genes evolution in plants. **Physiol Mol Biol Plants**, v. 19, n. 1, p. 1-9. 2013.

KROGH, A. An Introduction to Hidden Markov Models for Biological Sequences. Elsevier, p. 45-63. 1998.

LASHERMES, P. et al. Molecular Characterization and origin of the *Coffea arabica* L. genoma. **Mol Gen Genet**. v. 261, p. 259-266. 1999.

LI, L., et al. The FLS2-associated kinase BIK1 directly phosphorylates the NADPH oxidase RbohD to control plant immunity. **Cell Host Microbe**, v. 15, p. 329-338. 2014.

LI, P., et al. TGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGA) in plants. **BMC Genomics**, v. 17, n. 852, p. 1-10. 2016

LOZANO R., et al. Identification and distribution of the NBS-LRR gene family in the cassava genome. **BMC Genomics**, v. 16, n. 360. 2015.

MA, W. et al. Innate Immunity Signaling: Cytosolic Ca²⁺ Elevation Is Linked to Downstream Nitric Oxide Generation through the Action of Calmodulin or a Calmodulin-Like Protein. **Plant Physiology**, v. 148, p. 818-828. 2008

- MARONE, D., et al. Genetics basis of qualitative and quantitative resistance to powdery mildew in wheat: from consensus region to candidate genes. **BMC Genomics**, v. 14, n. 562, p. 1-17. 2013.
- MAURIN, O. et al. Towards a Phylogeny for *Coffea* (Rubiaceae): Identifying Well-supported Lineages Based on Nuclear and Plastid DNA Sequences. **Annals of Botany**. Oxford. V 100, p. 1565-1583, 2007.
- MCHALE L., TAN X., KOHEL P., MICHELMORE R. Plant NBS-LRR proteins: a adaptable guards. **Genome Biology**, 7:212. 2006.
- MEJURIA, T., NEUHOFF, D., KOPKE U. The Status Of Coffee Production And The Potential For Organic Conversion In Ethiopia. **Conference on International Agriculture Research for Development**. Berlin, 5-7 de October de 2004.
- MEYERS, B.C., et al. Genome-Wide Analysis of NBS-LRR-Encoding Genes in Arabidopsis. **The Plant Cell**, v. 15, p. 809-934. 2003
- MOFFETT P., FARNHAM G., PEART J., BAULCOMBE D.C. Interaction between domains of a plant NBS-LRR protein in disease resistance-related cell death. **The EMBO journal**, v. 21, n. 17, p. 4511-4519. 2002.
- MONDEGO J.M., et al. An EST-based analysis identifies new genes and reveals distinctive gene expression features of Coffee arabica and *Coffea canephora*. **BMC Plant Biology**, v. 11, n. 30, p. 1-20. 2011.
- MUKHTAR S.H., et al. Independently Evolved Virulence Effectors Converge onto Hubs in a Plant Immune System Network. **Science**, v. 333, n. 6042, p. 596-601. 2011.
- NEUPANE, S., et al. Genome-wide identification of NBS-Encoding Resistance Genes in Sunflower (*Helianthus annuus* L.). **Genes**, v. 9, n. 384, p. 1-21. 2018.
- NOIR, S. et al. Construction and characterization of a BAC library for genome analysis of the allotetraploid coffee species (*Coffea arabica* L.). **Theor Appl Genet**. v. 109, p. 225-230. 2004
- NUHSE. T.S., et al. Quantitative phosphor proteomic analysis of plasma membrane proteins reveals regulatory mechanisms of planta immune responses. **Plant Journal**, v. 51, p. 931-940. 2007
- NÜRNNERGER, T., KEMMERLING, B. PAMP-Triggeted Immunity in Plants. **Advances in Botanical Research**, v. 51. 2009.
- OJITO-RAMOS, K., PORTAL, O. Inducción del sistema immune de plantas. **Biotechnologia vegetal**, v. 10, n. 1, p. 3-19. 2010
- OKAZAKI, Y., SAITO, K. Roles of lipids as signaling molecules and mitigators during stress response in plants. **The Plant Journal**, v. 79, p. 584-596. 2014.

PARCA, L. et al. Identification of Nucleotide-Binding Site in protein Structure: A Novel Approach Based on Nucleotide Modularity. **Plos one**, v. 7, n. 11, p. 1-12. 2012.

PLOCIK, A., LAYDEN, J., KESSELI, R. Comparative analysis of NBS domain sequences of NBS-LRR disease resistance genes from sunflower, lettuce, and chicory. **Molecular Phylogenetic and Evolution**, v. 31, p. 153-163.2004.

PRAKASH, N. S. et al. Introgression molecular analysis of a leaf rust resistance gene from *Coffea liberica* into *C. arabica* L. **Theor Appl Genet**, v. 109, p. 1311-1317. 2004.

RADWAN, O. Development of PCR Markers for the PI5/PI8 locus for resistance to *Plasmopara halstedii* in sunflower, *Helianthus annuus* L. from complete CC-NBS-LRR sequences. **Theor Appl Genet**, v. 109, n. 1, p. 176-185. 2004

RIBAS A.F., et al. Organization and molecular evolution of a disease-resistance gene cluster in coffee trees. **BMC Genomics**, v. 12, p. 240. 2011.

ROJO-JIMENEZ, E. Café I (G. *Coffea*). **Reduca**, v. 7, n. 2, p. 113-132. 2014.

ROVA, J.H., et al. A trnL-F cpDNA sequence study of the Condamineae-Rondeletiae-Sipaneae complex with implications on the phylogeny of the Rubiaceae. 2002.

RYAN, C.A., HUFFAKER, A., YAMAGUCHI, Y. New insights into innate immunity in *Arabidopsis*, **Cellular Microorfanismology**, v. 9, p. 1902-1908. 2007.

SANSEVERINO, W., et al. PRGdb: a bioinformatics platform for plant resistance gene analysis. **Nucleic Acids Research**, v. 38, p. D814-D821. 2010.

SCHWESSINGER, B., ZIPFEL, C. News from the frontline: recent insights into PAMP-triggered immunity in plants. **Current Opinion in Plant Biology**, v. 11, p. 389-395. 2008.

SEHNKE, P.C., DELILLE, J.M., FERL, R.J. Consummating Signal Transduction: The Role of 14-3-3 Proteins in the Completion of Signal-Induced Transitions in Proteins Activity. **The Plant Cell**, p. S339-354. 2002.

SEKHWAL, M.K., et al. Disease Resistance Gene Analogs (RGA's) in Plants. **Int. J. Mol. Sci.** v. 16, p. 19248-19290. 2015.

SEO, E., et al. Genome-Wide Comparative Analyses Reveal the Dynamic Evolution of Nucleotide-Binding Leucine Rich Repeat Gene Family among Solanaceae Plants. **Frontiers in Plant Science**, v. 7, n.1205. 2016.

SERA, G. H. et al. Resistance to Leaf Rust in Coffee Carrying SH3 Gene and others SH Genes. **Brazilian archives of biology and technology**, v. 50, n. 5, p. 753-755. 2007.

SHARMA T.R., et al. Resistance Genes Analogues as a Tool for Rapid Identification and Cloning of Disease Resistance Genes in Plants – A review. **J. Plant Biochemistry and Biotechnology**, v. 18, n. 1, p. 01-11. 2009.

- TEKETAY, D. History, botany and ecological requirements of coffee. **Walia**, v. 20, p. 28-50. 1999.
- TESTERINK, C., et al. Phosphatidic acid binds to and inhibits the activity of Arabidopsis CTR1. **Journal of Experimental Botany**, v. 58, n. 14, p. 3905-3914. 2007.
- TIMOTHY L., et al. Noble, "MEME SUITE: tools for motifs discovery and searching", **Nucleic Acids Research**, v. 37, p. W202-W208. 2009.
- TIMOYHY L. Bioinformatics: Chapter 12, Discovering Sequence Motifs. Totowa: Human Press. 2008.
- VILLEGAS-GARCÍA, C., BAEZA-ARAGÓN, C.A. Diseminación de *Hemileia vastatrix* Berk. y Br. A nivel del árbol, en un foco natural. **Cenicafe**, v. 41, n. 2, p. 39-49. 1990.
- WALLER, J.M., BIGGER, M., HILLOCKS, R.J. Coffee Pests, Diseases and their management. CABI, Cambridge: 2007.
- WAN H., et al. Analysis of TIR- and non-TIR-NBS-LRR disease resistance gene analogous in pepper: characterization, genetic variation, functional divergence and expression patterns. **BMC Genomics**, v. 13, n. 502. 2012.
- WANG, C., et al. Involvement of Phospholipase D in Wound-Induced Accumulation of Jasmonic Acid in Arabidopsis. **The Plant Cell**, v. 12, p. 2237-2246. 2000.
- YAMAGUCHI, Y., PEARCE, G., RYAN C.A. The cell surface leucine-rich-repeat receptor for AtPep1, an endogenous peptide elicitor in Arabidopsis, is functional in transgenic tobacco cells. **Plant Biology**, v. 103, n. 26, p. 10104-10109. 2006.
- YINGQIANG W.W., et al. Specific Targeting of the Arabidopsis Resistance Protein RPW8.2 to the Interfacial Membrane Encasing the Fungal Haustorium Renders Broad-Spectrum Resistance to Powdery Mildew. **The Plant Cell**, v. 21, p. 2998-2913. 2009.
- YU J., et al. Genome-wide comparative analysis of NBS-encoding genes between Brassia species and Arabidopsis thaliana. **BMC Genomics**, v. 15, n. 3, p. 1-18. 2014.
- ZHAO Y., et al. Bioinformatics Analysis of NBS-LRR Encoding Resistance Genes in *Setaria italica*. **Biochem Genet**, v. 54, n. 3. p. 232-248. 2016.
- ZHONG, Y., ZHANG, X., CHENG, Z.M. Linage-specific duplications of NBS-LRR genes occurring before the divergence of six *Fragaria* species. **BMC Genomics**, v. 19, n. 128, p. 1-15. 2018.
- ZHU, Q.H., et al. miR482 Regulation of NBS-LRR Defense genes During Fungal Pathogen Infection in Cotton. **Plos One**, v. 8, n. 12, p 1-11. 2013.
- ZIPFEL, C. Pattern-recognition receptor in plant innate immunity. **Current Opinion in Immunology**, v. 20, p. 10-16. 2008.

ZHANG, Y., et al. Phospholipase D α 1 and Phosphatidic Acid Regulate NADPH Oxidase Activity and Production of Reactive Oxygen Species in ABA-Mediated Stoma Closure in Arabidopsis. **The Plant Cell**, v. 21, p. 2357-2377. 2009.

7. ANEXOS

ANEXO 1: *Primers degenerados de C. arabica*

ID	Nseq	F-Motiv	FND	FID	FTM	F%GC	SEQ-FP	R-Motiv	RND	RID	RTM	R%GC	SEQ-RP	TA	L
G1	8	MAETVL	768	F1	53.1	50	ATGGCNGARACNGTNYTN	GLCFKA	768	R1	51	47.2	NGCYTTRAARCANARNCC	50.05	2655
G2	3	FAGYRH	768	F2	54.6	47.2	TTYGCNGGNTAYRGNCA	PSGIQH	1152	R2	52.2	51.9	RTGYTG DATNCCNRWNGG	51.4	2438
G3	4	KQVREA	768	F3	54.3	52.8	AARCARGTNRGRGARGCN	YQGEGLC	1536	R3	57.4	52.4	RCANARNCCYTCNCCYTGR	53.38	2477
G4	7	CFLKDAD	768	F4	53.8	42.9	TGYTTYTNAARGARGCNGAY	DPMPVLE	1536	R4	57.4	54.8	YTCNARNACNGGCATNGGRTC	53.6	2262
G5	5	MADPVI	384	F5	53.8	51.9	ATGGCNGAYCCNGTNATH	GPDSQW	384	R5	55.8	61.1	CCAYTGNRWRTCNGGNCC	52.8	3142
G6	5	MADPVI	384	F6	50.5	46.3	ATGACNGAYCCNGTNATH	GPDSQW	384	R6	55.8	61.1	CCAYTGNRWRTCNGGNCC	51.15	2138
G7	8	RWRMAS	864	F7	56.2	58.3	RGNTGGRGNATGGCNWYN	MGRLTC	1152	R7	53.2	50	RCANGTNAANCYCNCCAT	52.7	1702
G8	5	WLKQLEE	575	F8	53.6	42.9	TGGYTNAARCARYTNGARGAR	WIWLSYQ	432	R8	51.2	39.7	YTGR TANRWNARCCADATCCA	50.4	3623
G9	4	MIRGFLA	864	F9	55.5	46.8	ATGATHRGNGGNTTYTNGCN	CHGTPQ	512	R9	58.8	61.1	YTGNGGNGTNCRTGRCA	55.15	4567
G10	6	AEGFLH	768	F10	53.1	52.8	GCNGARGGNTTYTNCAY	PDWFGK	128	R10	53	52.8	YTTNCCRAACCARTCNGG	51.05	2237
G11	4	PILGMG	1152	F11	53	54.6	CCNATHYTNNGGNATGGGN	NEFDWSG	128	R11	54.9	45.2	NCCNRWCCARTCRAAYCRTT	51.95	2156
G12	7	PNEEEM	256	F12	55.2	52.4	CCNAAYGARGARGGGNATG	LQRAANS	768	R12	51.5	47.2	NGTNGGNARYTCYTRCA	51.35	3329
G13	4	GHYVVG	1024	F13	53.6	55.6	NCCNACNACRTARTGNCC	MTPKGFG	1024	R13	58.3	52.4	NCCRAANCCYTTNGGNGTCAT	53.95	2144
G14	4	MAEAAL	768	F14	59.4	61.1	ATGGCNGARGCNGCNYTN	GAEWPK	256	R14	58.9	63.9	YTTNGGCCAYTCNGCNCC	57.15	3408
G15	4	AFRWRMA	2651	F15	60.8	59.5	GCNTTYRGNTGGRGNATGGCN	WFGNFAA	2651	R15	59.6	52.4	NGCNGCRAARTTNCRAACCA	58.2	1152
G16	3	FQSHAW	192	F16	55.7	50	CAYTTYCAYWYNCAYG CNTGG	MKGRHC	192	R16	53.3	50	RCARTGNCYNCYTTTCAT	52.5	1802
G17	3	IVGMGG	768	F17	56.2	57.4	ATHGTNGGNATGGGNGGN	PEWMGS	192	R17	54.9	58.3	NRWNCCATCCAYTCNGG	55.25	2948
G18	4	E[DE]EVFKA	256	F18	53	45.2	GARGAYGARGTNTTYAARGCN	EGLOPH	768	R18	55.5	58.3	RTGNGGYTGNARNCCYTC	52.3	2033
G19	4	VAMIQAV	1536	F19	56.5	51.6	GTNGCNATGATHCARGCNGTN	GDQFPRW	768	R19	59.2	59.5	CCANCYNGGRAAYTGRTCNC	56	2168

ID: Identificadores do Grupo de *primers*, Nseq: Numero de Sequências utilizadas para os desenhos de *primers*, F-Motiv: Motivo utilizado para o desenho do *primer forward*, FND: Nivel de degeneração do *primer forward*, FID: Identificador do *primer forward*, FTM: Têmpertura de *melting* do *primer forward*, F%GC: Contudo de GC do *primer forward*, SEQ-FP: Sequência do *primer forward* em sentido

5'-3, R, R-Motiv: Motivo utilizado par ao desenho do *primer reverse*, RND: Nivel de degeneração do *primer reverse*, RID: Identificador do *primer reverse*, RTM: Temperatura de *melting* do *primer reverse*, R%GC: Contudo de GC do *primer reverse*, SEQ-RP: Sequência do *primer reverse* em sentido 5'-3, TA: Temperatura de *annealing* para o grupo de primers, L: Tamanho esperado do fragmento.

ANEXO 2: Primers degenerados de *C. arabica*

ID	Nseq	F-Motiv	FND	FID	FTM	F%GC	SEQ-FP	R-Motiv	RND	RID	RTM	R%GC	SEQ-RP	TA	L
G1	8	MAETVL	768	F1	53.1	50.0	ATGGCNGARACNGTNYTN	GLCFKA	768	R1	51.0	47.2	NGCYTTTRAARCANARNCC	50.05	2565
G2	3	GHMKAF	128	F2	50.7	47.2	GGNCAYATGAARGCNTTY	HIPEIV	576	R2	46.9	42.6	NACDATYTCNGGDATRTG	46.8	2766
G3	4	KQVREA	768	F3	54.3	52.8	AARCARGTNRGRGARGCN	YQGEGLC	1536	R3	57.4	52.4	RCANARNCCYTCNCCYTGRTA	53.85	2477
G4	7	CFLKAD	768	F4	53.8	42.9	TGYTTYTNAARGARGCNGAY	DPMPVLE	1536	R4	57.4	54.8	YTCNARNACNGGCATNGGRTC	53.6	2262
G5	5	MADPVI	384	F5	53.8	51.9	ATGGCNGAYCCNGTNATH	GPDSQW	384	R5	55.8	61.1	CCAYTGNRWRTCNGGNCC	52.8	3142
G6	5	MADPVI	384	F6	50.5	46.3	ATGACNGAYCCNGTNATH	GPDSQW	384	R6	55.8	61.1	CAYTGNRWRTCNGGNCC	51.15	3138
G7	8	MADPVI	384	F7	50.5	46.3	ATGACNGAYCCNGTNATH	YNEFDWS	192	R7	49.7	35.7	NRWCCARTCRAAYTCRTRTA	48.1	2793
G8	5	WLKQLEE	576	F8	53.6	42.9	TGGYTNAARCARYTNGARGAR	WIWLSYQ	432	R8	51.2	39.7	YTGRTANRWNARCCADATCCA	50.4	3623
G9	6	MIRGFLA	864	F9	55.5	46.8	ATGATHRGNNGNTTYTNGCN	CHGTPQ	512	R9	58.8	61.1	YTGNGGNGTNCRTGRCA	55.15	4567
G10	6	ASEQIG	1152	F10	52.7	51.9	GCNWYNGARCARATHGGN	NLCPSN	1152	R9	47.8	38.	NRWRAARTTNRWNGGRCA	48.25	3566
G11	4	ANERIG	1152	F11	52.4	51.9	GCNAAYGARRGNATHGGN	QIGSET	1152	R11	49.3	46.3	NGTYTCNRWNCCDATYTG	48.85	2909
G12	7	PNEEEGM	256	F12	55.2	52.4	CCNAAYGARGARGARGGNATG	CKELPT	768	R12	51.5	47.2	NGTNGGNARYTCYTRCA	51.35	3329
G13	4	ANERIG	1152	F13	52.4	51.9	GCNAAYGARRGNATHGGN	QIGSET	1152	R13	49.3	46.3	NGTYTCNRWNCCDATYTG	48.85	2921
G14	4	MAEAAL	768	F14	59.4	61.1	ATGGCNGARGCNGCNYTN	GAEWPK	256	R14	58.9	63.9	YTTNGGCCAYTCNGCNCC	57.15	3408
G15	4	VGFKKDV	1024	F15	51.2	40.5	GTNGGNTTYAARAARGAY	QWSKVS	576	R15	52.2	40.5	NRWNACYTTNRWCCAYTGNRW	49.7	3077
G16	4	KQVREV	768	F16	51.4	47.2	AARCARGTNRGNGARGTN	IIRQFL	1296	R16	43.7	31.5	ATHATHRGNCAARTTYTN	45.55	2534
G17	4	DKILPL	1728	F17	45.0	38.0	NARNGGNARDATYTRTC	PDQIQH	192	R17	51.2	49.1	RTGYTGATYTYGTCNGG	46.1	3464
G18	4	E[DE]EVFKA	256	F18	53.0	45.2	GARGANGARGTNTTYAARGCN	EGLQPH	768	R18	55.5	58.3	RTGNGGYTGNARNCCYTC	52.3	2033
G19	4	MIQAVL	768	F19	49.7	43.5	ATGATHCARGCNGTNYTN	CPKLKT	768	R19	49.4	41.7	NGTYTTNARYTTNGGRCA	47.55	2477

ID: Identificadores do Grupo de *primers*, Nseq: Número de Sequências utilizadas para os desenhos de *primers*, F-Motiv: Motivo utilizado para o desenho do *primer forward*, FND: Nível de degeneração do *primer forward*, FID: Identificador do *primer forward*, FTM: Temperatura de *melting* do *primer forward*, F%GC: Contudo de GC do *primer forward*, SEQ-FP: Sequência do *primer forward* em sentido 5'-3', R, R-Motiv: Motivo utilizado para o desenho do *primer reverse*, RND: Nível de degeneração do *primer reverse*, RID: Identificador do *primer reverse*, RTM: Temperatura de *melting* do *primer reverse*, R%GC: Contudo de GC do *primer reverse*, SEQ-RP: Sequência do *primer reverse* em sentido 5'-3', TA: Temperatura de *annealing* para o grupo de primers, L: Tamanho esperado do fragmento.

ANEXO 3: Principais proteínas de café relacionadas como proteínas de resistência de outras plantas relacionadas com antecedência

N°	ID	FRAME	FAMILY	DESCRIPTION	ENTRY TYPE	CLAN	ENVELOPE		ALIGNMENT		BIT SCORE	E-VALUE
							START	END	START	END		
1	>Ca_13_548	2 (+)	PMD	Plant mobile domain	Domain	n/a	2352	2613	2373	2611	55.7	4.2e-15
		3 (+)	Rx_N	Rx N-terminal domain	Domain	n/a	181	309	188	270	61.7	6.2e-17
		3 (+)	NB-ARC	NB-ARC domain	Domain	CL0023	354	605	358	573	173.5	4.3e-51
2	>Ca_15_189	6 (-)	NB-ARC	NB-ARC domain	Domain	CL0023	2690	2880	2692	2867	72.7	2.6e-20
		4 (-)	Rx_N	Rx N-terminal domain	Domain	n/a	2486	2591	2488	2574	63.1	2.2e-17
		1 (+)	NB-LRR	Late blight resistance protein R1	Family	n/a	3510	3705	3520	3700	50.3	2.1e-13
3	>Ca_15_193	3 (+)	NB-ARC	NB-ARC domain	Domain	CL0023	2936	3027	2950	3026	42.6	3.8e-11
		6 (-)	F-box	F-box domain	Domain	CL0271	608	653	610	644	22.6	6.9e-05
		6 (-)	LRR_2	Leucine Rich Repeat	Repeat	CL0022	771	796	771	796	18.5	0.0023
		5 (-)	FBD	FBD	Family	n/a	1049	1090	1053	1090	39.9	2.9e-10
4	>Ca_31_251	2 (+)	GTP_EFTU	Elongation factor Tu GTP binding domain	Domain	CL0023	5952	6004	5953	6002	31.1	1.5e-07
		5 (-)	Rx_N	Rx N-terminal domain	Domain	n/a	4398	4518	4404	4486	68.2	5.5e-19
		5 (-)	NB-ARC	NB-ARC domain	Domain	CL0023	4572	4807	4589	4793	164.7	2.1e-48
5	>Ca_33_139-1	1 (+)	NB-LRR	Late blight resistance protein R1	Family	n/a	4176	4357	4177	4329	59.0	4.7e-16
		1 (+)	NB-ARC	NB-ARC domain	Domain	CL0023	4595	4655	4597	4643	26.8	2.6e-06
6	>Cc07_g17810	1 (+)	NB-LRR	Late blight resistance protein R1	Family	n/a	138	377	144	354	55.8	4.4e-15
		1 (+)	NB-ARC	NB-ARC domain	Domain	CL0023	538	776	540	760	176.9	3.9e-52
7	>Cc00_g05490	1 (+)	NB-LRR	Late blight resistance protein R1	Family	n/a	3	220	5	200	52.7	3.8e-14
		1 (+)	NB-ARC	NB-ARC domain	Domain	CL0023	377	616	379	599	170.3	4.2e-50

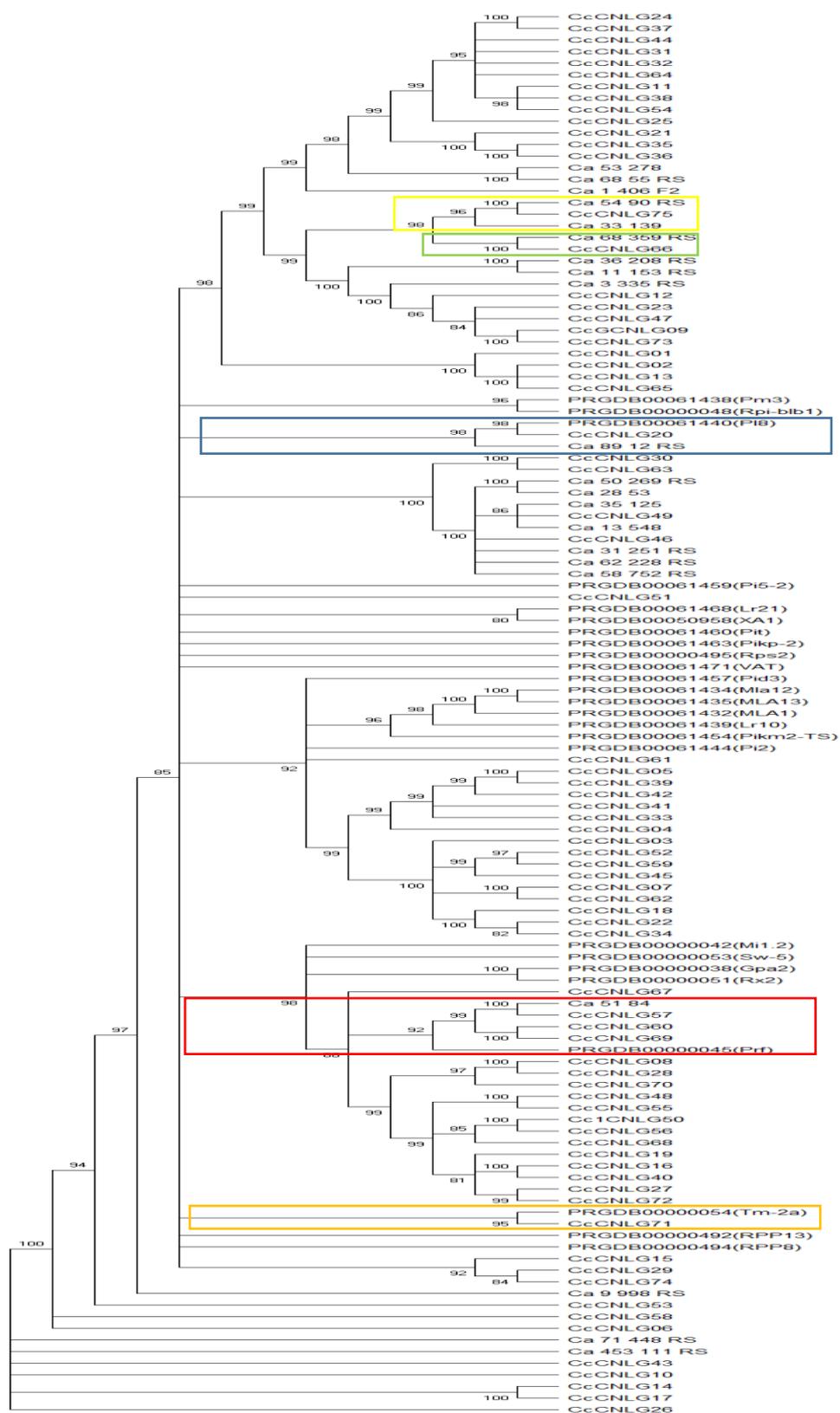
8	>Cc02_g33990	1 (+)	NB-LRR	Late blight resistance protein R1	Family	n/a	133	376	143	357	66.3	2.7e-18
		1 (+)	NB-ARC	NB-ARC domain	Domain	CL0023	537	777	539	759	164.7	2.1e-48
9	>Cc00_g10160	1 (+)	NB-LRR	Late blight resistance protein R1	Family	n/a	44	295	58	276	77.9	8.4e-22
		1 (+)	NB-ARC	NB-ARC domain	Domain	CL0023	456	693	457	678	166.5	5.9e-49
10	>Cc00_g05420	1 (+)	NB-LRR	Late blight resistance protein R1	Family	n/a	134	380	149	361	56.9	2E-15
11	>Cc00_g05470	1 (+)	NB-ARC	NB-ARC domain	Domain	CL0023	542	779	543	777	161.5	2E-47
		1 (+)	NB-LRR	Late blight resistance protein R1	Family	n/a	134	392	156	368	62.3	4.6e-17
12	>Cc04_g12020	1 (+)	NB-ARC	NB-ARC domain	Domain	CL0023	551	791	553	775	179.7	5.7e-53
		1 (+)	NB-LRR	Late blight resistance protein R1	Family	n/a	118	373	139	350	49.5	3.6e-13
13	>Cc00_g10170	1 (+)	NB-ARC	NB-ARC domain	Domain	CL0023	536	777	538	762	182.9	6.1e-54
		1 (+)	NB-LRR	Late blight resistance protein R1	Family	n/a	128	390	151	363	61.4	8.8e-17
		1 (+)	NB-ARC	NB-ARC domain	Domain	CL0023	549	789	551	774	183.2	4.9e-54
14	>Cc00_g19160	1 (+)	NB-LRR	Late blight resistance protein R1	Family	n/a	122	385	144	359	59.4	3.6e-16
		1 (+)	NB-ARC	NB-ARC domain	Domain	CL0023	545	786	547	771	176.5	5.3e-52
15	>Cc00_g03010	1 (+)	NB-LRR	Late blight resistance protein R1	Family	n/a	154	409	171	388	71.7	6.2e-20
		1 (+)	NB-ARC	NB-ARC domain	Domain	CL0023	569	817	570	800	187.5	2.4e-55
16	>Cc07_g20590	1 (+)	NB-LRR	Late blight resistance protein R1	Family	n/a	183	417	210	393	42.2	6.3e-11
		1 (+)	NB-ARC	NB-ARC domain	Domain	CL0023	574	814	576	799	180.5	3.2e-53

17	>Cc07_g21460	1 (+)	NB-LRR	Late blight resistance protein R1	Family	n/a	138	388	153	364	50.8	1.5e-13
		1 (+)	NB-ARC	NB-ARC domain	Domain	CL0023	545	785	547	780	183.3	4.6e-54
18	>Cc11_g03840	1 (+)	NB-LRR	Late blight resistance protein R1	Family	n/a	1	159	3	135	49.6	3.3e-13
		1 (+)	NB-ARC	NB-ARC domain	Domain	CL0023	316	555	322	554	163.2	6.2e-48
19	>Cc07_g17040	1 (+)	NB-LRR	Late blight resistance protein R1	Family	n/a	134	411	157	386	87.1	1.3e-24
		1 (+)	NB-ARC	NB-ARC domain	Domain	CL0023	564	813	566	789	183.9	3E-54
20	>Cc07_g20740	2 (+)	NB-LRR	Late blight resistance protein R1	Family	n/a	575	826	580	820	77.3	1.2e-21
		2 (+)	NB-ARC	NB-ARC domain	Domain	CL0023	985	1233	987	1219	180.2	4E-53

ANEXO 4: Motivos identificados exclusivamente em sequencias de *Coffea arabica* e *Coffea canephora*

Espécie	Motivo	e-value	Posição
<i>C. arabica</i>		2.4e-083	NBS
<i>C. arabica</i>		1.2e-079	NBS
<i>C. arabica</i>		3.2e-075	NBS
<i>C. arabica</i>		6.0e-076	NBS
<i>C. canephora</i>		1.5e-115	NBS
<i>C. canephora</i>		7.9e-113	LRR
<i>C. canephora</i>		5.3e-105	LRR
<i>C. canephora</i>		2.3e-071	NBS

ANEXO 5: Árvore filogenética das sequências candidatas a CNL de *C. canephora*-CGH, *C. arabica*-WCR e sequências de referência PRGdb.



Método de inferência: ML, Test filogenético: *Boostrap* com 1000 replicatas, Modelo de substituição: GTR+G+I, *Cutoff*: 90%.