Genome-wide identification, characterization, and comparative analysis of NLR resistance genes in *Coffea spp.*

Mariana de Lima Santos [1*], Mário Lúcio Vilela de Resende [1], Gabriel Sérgio Costa Alves[2], Jose Carlos Huguet-Tapia[3], Márcio Fernando Ribeiro de Resende Júnior[4 †], Jeremy Todd Brawner[3* †]

[1]Laboratório de Fisiologia do Parasitismo, Faculdade de Ciências Agrárias, Departamento de Fitopatologia, Universidade Federal de Lavras, Lavras, Minas Gerais, Brazil

[2]Laboratório de Processos Biológicos e Produtos Biotecnológicos, Instituto de Ciências Biológicas, Departamento de Biologia Celular, Universidade de Brasília, Brasília, Distrito Federal, Brazil

[3]Institute of Food and Agricultural Sciences, Department of Plant Pathology, University of Florida, Gainesville, Florida, FL, United States

[4] Institute of Food and Agricultural Sciences, Horticultural Sciences Department, University of Florida, Gainesville, Florida, FL, United States

*** Correspondence:**
Corresponding Authors
santos-ml@outlook.com
jeremybrawner@ufl.edu

†These authors have contributed equally to this work and share last authorship

Keywords: resistance genes, Coffea, nucleotide-binding site leucine-rich repeat, Genome-wide, NLR-annotator.

Abstract

The largest family of disease resistance genes in plants are nucleotide-binding site leucine-rich repeat genes (NLRs). The products of these genes are responsible for recognizing avirulence proteins (Avr) of phytopathogens and triggering specific defense responses. Identifying NLRs in plant genomes with standard gene annotation software is challenging due to their multidomain nature, sequence diversity, and clustered genomic distribution. We present the results of a genome-wide scan and comparative analysis of NLR loci in three coffee species (*Coffea canephora*, *Coffea eugenioides* and their interspecific hybrid *Coffea arabica*). A total of 1311 non-redundant NLR loci were identified in *C. arabica*, 927 in *C. canephora*, and 1079 in *C. eugenioides*, of which 809, 562, and 695 are complete loci, respectively. The NLR-annotator tool used in this study showed extremely high sensitivities and specificities (over 99%) and increased the detection of putative NLRs in the reference coffee genomes. The NLRs loci in coffee are distributed among all chromosomes and are organized mostly in clusters. The *C. arabica* genome presented a smaller number of NLR loci when compared to the sum of the parental genomes (*C. canephora*, and *C. eugenioides*). There are orthologous NLRs (orthogroups) shared between coffee, tomato, potato, and reference NLRs and those that are shared only among coffee species, which provides clues about the functionality and evolutionary history of these orthogroups.

Phylogenetic analysis demonstrated orthologous NLRs shared between *C. arabica* and the parental genomes and those that were possibly lost. The NLR family members in coffee are subdivided into two main groups: TIR-NLR (TNL) and non-TNL. The non-TNLs seem to represent a repertoire of resistance genes that are important in coffee. These results will support functional studies and contribute to a more precise use of these genes for breeding disease-resistant coffee cultivars.

1 Introduction

Coffee is a globally important agricultural commodity that plays a significant economic role in producing and consuming countries (Krishnan, 2017). The genus *Coffea* consists of more than 100 botanical species (Davis et al., 2006), however, the most cultivated species are *Coffea canephora* and *Coffea arabica*. *C. canephora* is diploid (2n = 2x = 22 chromosomes) (Denoeud et al., 2014), while *C. arabica* is a allotetraploid (2n = 4x = 44 chromosomes) (Tran et al., 2018) originated from natural hybridization between *C. canephora* and *C. eugenidoides* (Lashermes et al., 1999; Bawin et al., 2020). Among the more than 50 coffee-producing countries, Brazil, Vietnam, Colombia, and Indonesia are major producers, with Brazil being the largest producer by volume. Currently, coffee diseases are the main factor affecting productivity (Cerda et al., 2017). Examples of diseases associated with coffee include cercosporiosis (*Cercospora coffeicola*), bacterial blight (*Pseudomonas syringae* pv. *Garcae*), anthracnose (*Colletotrichum coffeanum*), root-knot nematodes (*Meloidogyne spp*.), coffee berry disease – CBD (*Colletotrichum kahawae*), and coffee leaf rust – CLR (*Hemileia vastatrix*) (Cabral et al., 2016; Krishnan, 2017). CLR is one of the most devastating diseases found in coffee and is present in all regions of the world where coffee is grown (McCook and Vandermeer, 2015; Cabral et al., 2016). Currently, 95% of *C. arabica* varieties cultivated in Brazil are susceptible to CLR due to the emergence of variants of the pathogen (Cabral et al., 2016). Given the increasing problem of plant pathogens in coffee production, a greater understanding of the set of receptors regulating the plant immune system of coffee is needed.

Throughout evolution, plants have developed sophisticated systems to defend themselves from pathogens. The plant immune system involves both broad-spectrum and specific recognition of pathogens. Broad-spectrum recognition is related to the detection of pathogen-associated molecular patterns (PAMP), such as fungal chitin or bacterial flagella, by pattern recognition receptors (PRR) that are anchored to the plasma membrane and trigger the PAMP-triggered immunity (PTI) (Boutrot and Zipfel, 2017). Specific recognition, on the other hand, primarily involves receptors encoded by resistance genes (R genes) that detect the presence of pathogen effector proteins and trigger effector-triggered immunity (ETI) (Jones and Dangl, 2006). Both types of recognition occur dynamically and continuously, converging into signaling pathways that activate essential mechanisms for downstream responses to pathogen recognition (Lu and Tsuda, 2021; Yuan et al., 2021).

The R genes have been extensively studied in several crops to facilitate their greater use in plant breeding (Jupe et al., 2013; Wan et al., 2013; Lozano et al., 2015; Inturrisi et al., 2020; Steuernagel et al., 2020). The protein products of these genes recognize directly or indirectly effector proteins that are secreted by pathogens (Kourelis and Van Der Hoorn, 2018) and trigger a series of signaling steps that lead to the hypersensitive response (HR) that activates cell death and potentially leads to systemic acquired resistance (SAR) (Kachroo and Robin, 2013; Jones et

al., 2016). The largest and most diverse group of R genes found in plants belong to the nucleotide-binding site leucine-rich repeat family (NLR or NBS-LRR) (Jones et al., 2016). The proteins encoded by these genes are typically modular, and many have a variable N-terminal domain-containing Toll/interleucina-1 (TIR) or coiled-coil (CC) receptors. As well, the nucleotide-binding domain (NB-ARC or NBS) is a canonical feature of NRLs, that is shared with human apoptotic protease-activating factor-1 (*APAF-1*) and *Caenorhabditis elegans* death-4 (*CED-4*) proteins. A C-terminal region comprising a variable number of leucine-rich repeats (LRRs) is another common feature in NLR genes (Jones et al., 2016; Shao et al., 2019). The NB-ARC domain is highly conserved and is involved in the active and inactive state of the NLS protein (Bonardi et al., 2012; Jones et al., 2016). This domain presents motifs that are characteristic of the ATPase family, such as p-loop, kinase 2, and the RNBS (Resistance Nucleotide Binding Site) A, RNBS-C, and RNBS-D motifs (Van Ghelder et al., 2019). Mutations in specific residues within these motifs may cause the loss of protein function or self-activation and interfere with the regulation or activation of defense mechanisms. (Monteiro and Nishimura, 2018; Bezerra-Neto et al., 2020).

The domains described above provide a more refined classification of NLRs proteins, which may be characterized into two main groups: TNLs (TIR-NLRs) or non-TNL (which include CNLs - CC-NLRs). The truncation of a single domain, such as LRR (CN or TN), TIR or CC (NL), in both C and N terminal domains (N), may also be used to classify NLR genes (Monteiro and Nishimura, 2018). Additionally, atypical or non-canonical integrated domains (IDs) that act as decoys and play roles in oligomerization or downstream signaling may be present, demonstrating the structural diversity of this NLR family (Kroj et al., 2016; Wang et al., 2021). The number of NLRs in plant genomes varies greatly and is often organized in tandem, which facilitates duplication, contraction, and transposition and provides a reservoir of genetic variation that allows plant evolutionary dynamics to respond to changes phytopathogen populations (Barragan and Weigel, 2021). These genes are often under selection pressure, resulting in a large number of pseudogenes and variable loci content within the same species, among species, and across plant populations (Schatz et al., 2014; Steuernagel et al., 2015; Sun et al., 2020; Hufford et al., 2021).

The knowledge of how NLRs are distributed throughout the genome and their diversity is of great interest as it may reveal new sources of resistance that may be used to develop new cultivars (Monteiro and Nishimura, 2018). The growing number of sequenced plant genomes facilitates the search for novel NLR and has led to the genome-wide analysis of NLR genes (Denoeud et al., 2014; Song et al., 2015; Scott et al., 2020; Wang et al., 2021). However, its large number, frequently clustered genomic distribution, and low expression in uninfected tissues make cataloging NLR genes challenging and often underestimates the number of NLRs in genomes (Jupe et al., 2013; Steuernagel et al., 2015, 2020). To mitigate this problem, some specific gene/loci NLR annotation pipelines have been developed to augment standard gene annotation software and improve our ability to identify and locate genes belonging to this family. Some examples of these pipelines are NBSPred (Kushwaha et al., 2016), NLGenomeSweeper (Toda et al., 2020), and NLR-annotator, a new version of NLR-parser (Steuernagel et al., 2015, 2020). The NLR-Annotator is a tool used to annotate NLR loci that use 20 highly curated motifs present in NLR proteins and does not depend on the support of transcript data (Jupe et al., 2012; Steuernagel et al., 2020). Since it was published, this tool has been applied in several studies to prospect and annotate R genes in recently sequenced genomes (Muliyar et al., 2020; Read et al.,

2020; Scott et al., 2020), to check the completeness of previous annotations (Muliyar et al., 2020), and for studies of the resistance-related locus (Jost et al., 2020).

The genome of *C. canephora* was published in 2014, which allowed the first genome-wide NLR study in coffee (Denoeud et al., 2014). In 2018, the *C. arabica* and *C. eugenioides* genomes were deposited at the NCBI, providing an essential resource for studying the structure and evolution of NLRs in arabica coffee and the contribution of the genomes that gave rise to this species. For coffee production to continue advancing in producing regions worldwide, adequate disease management is of great importance. A range of strategies must be used to control the main phytosanitary problems associated with coffee production. Using these genomic resources is essential for informing breeding strategies aimed at developing resistance to disease in coffee. Given the above, this study aimed to: (i) identify NLR loci in *C. arabica*, *C. canephora*, and *C. eugenioides* genomes using the NLR-annotator tool and discuss the improvements in annotation derived from the use of a specific pipeline for NLR genes in coffee, (ii) catalog, classify and characterize the distribution of NLRs loci in the *coffee spp.* genomes, and (iii) understand the contribution of *C. canephora* and *C. eugenioides* to the NLR repertoire of *C. arabica*.

2 Materials and Methods

2.1 Coffee Genomic Resources

Three genomes were used in this study. The *C. arabica* (Caturra red - Cara_1.0, GenBank assembly accession: GCA_003713225.1) and *C. eugenioides* (Ceug_1.0, GenBank assembly accession: GCA_003713205.1) genomes are available from the NCBI (National Center for Biotechnology Information) database (https://www.ncbi.nlm.nih.gov/) and the *C. canephora* genome is available at Coffee genome hub (https://coffee-genome.org/) (Denoeud et al., 2014). For the three species, the genome files, sets of predicted proteins, predicted genes, and GFF (General Feature Format) were used.

2.2 Identification of loci NLR in *Coffea* spp. genomes

The identification of NLR loci in *Coffea spp.* was accomplished by the NLR-Annotator, (Steuernagel et al., 2020) using the default parameters. The tool uses combinations of short motifs of 15 to 50 amino acids to classify a genomic locus as an NLR. These motifs were defined using domains of known NLR proteins used as a training set in a study carried out by Jupe et al. 2012 (Supplementary Table 1).

In summary, the pipeline is divided into three steps: 1) dissection of genomic input sequence into 20-kb fragments overlapping by 5 kb; 2) translating each fragment into all six reading frames and searching for the motifs associated with NLR by NLR-Parser to create an xml-based report file. The NLR-Parser searches for combinations of doublets or triplets of motifs and records their nucleotide positions, disregarding motifs that occur randomly. Finally in step 3, the NLR-Annotator uses the xml file as input, integrates data from all fragments, evaluates positions and combinations of motifs. In this step, the NB-ARC motifs are used as the principal seed to annotate NLR loci, generate output files (.gff, .bed - Browser Extensible Data, .txt and file of the NB-ARC motifs as multiple alignments to complete loci) based on coordinates and orientation the initial input genomic sequence (Steuernagel et al., 2020).

Each section of the genomic sequence associated with a single NLR is called an 'NLR locus' and this refers to an NB-ARC domain (or associated motif) followed or not by one or more leucine-rich repeats (LRRs). From the sets of motifs that are identified, these loci are classified as complete (containing the P-loop, at least three consecutive NB-ARC motifs, and at least one LRR), complete (pseudogenes), partial and partial (pseudogenes) (Supplementary Table 1). Therefore, the NLR-Annotator identifies the NLR loci that are either active genes or pseudogenes. The number of NLR loci and their classification is described in the output file.txt (Steuernagel et al., 2020 and https://github.com/steuernb/NLR-Annotator).

2.3 Validation of the NLR-annotator sensitivity and specificity in coffee genomes

To validate the sensitivity and specificity of the NLR-annotator in the coffee genomes, we initially classified the protein sequences of the three genomes using PfamScan (https://pfam.xfam.org/) version 1.5 with an e-value of less than 1E- 5 and models from Pfam-A. Subsequently, proteins that had the NB-ARC domain (PF00931) were filtered, and from this process, we obtained the ID of the genes corresponding to each protein. With the list of gene model IDs of the NLR family, it was then possible to filter the GFF files and obtain the positions of the genes that had already been annotated in each genome.

We identified overlapping intervals to compare the NLR loci detected by NLR-Annotator and the NLR genes that had already been annotated in the genomes. We used the information from .gff files from both annotations for an overlay analysis using bedtools intersect (version 2.29.2). An overlap was only considered if both, the locus, and gene, were on the same strand. This analysis made it possible to distinguish the loci identified by NLR-annotator that were or were not overlapping with the gene models from the reference genomes.

For NLR genes already annotated in the genomes and not identified by NLR-annotator, a search for motifs by NLR-Parser was performed to obtain the xml and txt output (options -c and -o) as well the detection of conserved domains using the NCBI Conserved Domain Database (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) for nucleotides sequences. Standard parameters were used for the conserved domains analysis, except for the threshold (E-value), which was set to 1E-5. The Graphical summary was set to provide a concise view of the results. To characterize the NLR loci only found by NLR-annotator and to make sure they were homologous with NLRs already annotated in plants, we aligned the sequences for these loci with NCBI's non-redundant protein database (nr) (https://blast.ncbi.nlm.nih.gov) using BLASTx (BLAST - version 2.10.1 with the max_target_seqs option set to 5). For loci that did not have homology with NLRs proteins, a conserved domain analysis was also performed as previously described.

The sensitivity of the pipeline was calculated as the ratio of the number of loci identified by NLR-annotator (including motifs detected by NLR-Parser in the second step of the pipeline) to the number of NLRs genes already annotated in the genomes. The specificity was calculated as the ratio of the number of loci identified by NLR-annotator that are related to NLRs genes or have characteristic domains of that family to the total number of loci identified. Characteristic domains were defined as domains overlapping with the annotations already described in the studied genomes, homology with NLR proteins by BLASTx or NB-ARC domains identified with conserved domains analysis.

2.4 Distribution of NLR loci in Coffee's chromosomes

In order to visualize the distribution of NLR loci on chromosomes of the three analyzed coffee species, the annotation files from NLR annotator (.txt) were used to extract the genomic position and classifications of the loci. The chromosome size information in Mb was obtained from the NCBI (for *C. arabica* and *C. eugenioides*) and Coffee genome hub (for *C. canephora*) and the visualization was created using the R software with the chromoMap package (Anand and Lopez, 2020). ChromoMap, divides the chromosomes as a continuous composition of loci. Each locus, consist of a specific genomic range determined algorithmically based on chromosome length and then the annotations are inserted. The detailed annotation information on each locus NLRs (complete, complete pseudogene, partial and partial pseudogene) is displayed in an HTML file.

2.5 Prediction of genes in the complete loci found only by the NLR-annotator

Gene prediction was performed using the AUGUSTUS program version 3.3.3 (http://bioinf.uni-greifswald.de/augustus/) (Stanke et al., 2006) using gene models from *Solanum lycopersicum* and allowing for the prediction of only complete genes.

2.6 Orthologous groups and Phylogenetic Analyses

The complete loci identified in the coffee genomes by the NLR-annotator, being those loci that overlap with gene models of the reference genomes and loci that were annotated by AUGUSTUS as putative genes were the focus of ortholog and phylogenetic analysis. In order to make a comparison with the set of coffee NLRs, 326 NLR loci identified in tomato (*Solanum licopersicum* - Heinz 1706) by Andolfo et al., 2014, 755 loci identified in potato (*Solanum tuberosum*) by Jupe et al., 2013, 67 NLR reference genes (functionally characterized protein) obtained from The Plant Resistance Genes database - PRGDB (http://prgdb.org/prgdb/, S2 table) (Osuna-Cruz et al., 2018) and the *CED-4* gene from *Caenorhabditis elegans* (outgroup) were also added. All these sequences were classified according to the rules of motifs established by the NLR-annotator and only those considered as complete NLR were used for these analyses. This criterion was used to standardize the methodology for classifying loci as complete or not.

The amino acid sequences of the NB-ARC domain were extracted from the set of complete NLR loci for the 5 species (*C. arabica*, *C. canephora*, *C. eugenioides*, *S. licopersicum* and *S. tuberosum*) along with the reference genes by NLR -annotator (parameter -a). All NB-ARC domain of these complete loci (hereafter called NLRs) were compared with each other using BLASTP, all-by-all (E value <1e-10). The markov clustering algorithm was performed with inflation value of 1.5 and then NLRs in the same cluster were classified as orthologous subgroups by OrthoMCL version 1.4 (standard parameters) (Li et al., 2003). In order to analyze and visualize the number of orthogroups shared between the species and the ones that are unique to a single species, we used the UpSetR package in R (Conway et al., 2017).

The NLRs clustered into single-copy orthogroups (orthogroups that have one copy of each NLR present once in each of the 5 genomes or reference NLRs) by OrthoMCL were used to construct a phylogenetic tree. The sequences were aligned using MAFFT version 6.903 (Katoh et al., 2002), with the --auto parameter to select the best alignment strategy. The tree was inferred using RAxML version 8.2.10 (Stamatakis, 2014) with the PROTGAMMAAUTO model (the JTT model was selected as having the highest likelihood) with 100 bootstrap replicates. A second

phylogenetic tree classifying the coffee NLRs was also constructed with the above-mentioned parameters using the entire set of complete NLRs. Coffee NLRs were classified in the tree using the previous classification describing 67 reference NLRs (Supplementary Table 2) and the tomato and potato NLRs (Jupe et al., 2012, 2013; Andolfo et al., 2014). The trees were visualized and edited using the Interactive Tree of Life (iTOL) tool (Letunic and Bork, 2021).

3 Results

3.1 NLRs Identification, validation of the sensitivity and specificity of NLR-annotator in coffee genomes

A total of 1318 loci were identified for *C. arabica*, 932 for *C. canephora*, and 1081 for *C. eugenioides* (Supplementary Table 3). In each species, we identified some loci that are in the same position but were separated by the NLR-annotator as two distinct NLRs. We found 7, 5, and 2 repeated loci in the aforementioned species, respectively (Supplementary Table 3 highlighted in blue). Considering these repeated loci when counting the number and distribution of the NLRs loci on the chromosomes, the most complete classification was considered. After the identification of these regions, it was found that there were 1311 non-redundant loci for *C. arabica* (627 from the *C. canephora* subgenome - CaC, 650 from the *C. eugenioides* subgenome - CaE and 34 Unassigned - Un), 927 for *C. canephora* (559 mapped on chromosomes and 367 Un) and 1079 for *C. eugenioides* (944 mapped on chromosomes and 135 Un). The number of complete loci found in each species was 809 (*C. arabica*), 562 (*C. canehora*), and 695 (*C.eugenioides*).

To examine whether there was a consensus between the gene models for NLRs that have previously been annotated in the genomes and loci identified by NLR-annotator, an overlap analysis was performed. PfamScan analyses were conducted on the set of predicted proteins and the subsequent selection of NLR proteins annotated in each genome showed that 1015, 709, and 869 genes encoded proteins (including isoforms) containing the NB-ARC domain in the *C. arabica*, *C. canephora*, and *C. eugenioides* genomes, respectively (Figure 1, Supplementary Table 4). The overlap between the genomic positions of these genes and the positions of loci from NLR-annotator showed that of 1311, 927, and 1079 loci identified by NLR-annotator for *C. arabica*, *C. canephora*, and *C.eugenioides* respectively, 1013 (99,80%), 687 (96,90%), and 857 (98,62%) overlap with the genes already annotated in the reference genomes. A total of 298, 240 and 222 non overlapping NLRs were found, respectively (Figure 1, Table 1). We also noticed that there are cases in which more than one NLR loci overlapped with a single NLR gene, and the opposite was also found in all three genomes. A Venn diagram representing these data is shown in figure 1 as the intersection and data are highlighted in Supplementary Table 5.

The overlap analysis also made it possible to identify genes annotated in the reference genomes that did not overlap with any locus from NLR-annotator. To examine these genes, an NLR-parser analysis, with options -c (file.xml) and -o (file.txt), was performed on this set. Among the genes not identified by NLR-annotator for *C. arabica* (18), *C. canephora* (25), and *C. eugenioides* (24), 7, 3, and 4, respectively, were below the standard threshold (1E-5) for the MAST based motif search used by NLR-Parser. Additionally, 9, 17, and 16 genes present motifs that were detectable using the standard threshold but did not contain at least three consecutive motifs belonging to the NB-ARC domain or presented as motifs in random order. These loci were therefore not

annotated in the third step of the NLR-annotator (Supplementary Table 6). After this analysis, we also identified and confirmed genes that were not found by NLR-annotator. Two genes were not found in *C. arabica* (LOC113737176 and LOC113735982), five genes in *C. canephora*, (Cc02_g12220, Cc03_g10360, Cc07_g18800, Cc00_g21910 and Cc00_g35420) and four genes in *C. eugenioides* (LOC113766771, LOC113766774, LOC113766615 and LOC113777141). Supplementary text 1, supplementary table 6 and supplementary figure 1 present additional details from this analysis. After these analyses, it was possible to verify that the NLR annotator showed a sensitivity of 99.8%, 99.4%, and 99.7% for *C. arabica*, *C. canephora* and *C. eugenioides*, respectively.

As stated above, the overlap analysis also made it possible to detect that the NLR-annotator identified loci that were complete, complete (pseudogenes), partial and partial (pseudogenes) that did not overlap with genes already annotated in reference genomes (Figure 1, Table 1). To further investigate these loci and ensure that they were indeed related to genes encoding NLR proteins, a BLASTx analysis was performed. This analysis showed that of the 298, 240, and 222 loci in *C. arabica*, *C canephora*, and *C. eugenioides*, only 7, 4, and 6 did not show homology with resistance proteins being found among the five best hits, respectively. (Supplementary Table 7, highlighted in orange).

To describe the sequences that did not show homology to NLRs proteins by BLASTx, a conserved domains analysis was performed (Supplementary Figure 2). Many of these loci do not show homology with NLRs proteins because most of the sequence contains domains related to the family of proteins involved in the activity of transposable elements such as ribonuclease H (RNase H) and reverse transcriptases (RTs). However, it was also possible to identify characteristic domains of NLR proteins such as NB-ARC, Toll / interleukin-1 receptor (TIR), RX-CC_like, and Rx_N, suggesting that these loci cannot be considered false positives. Only three loci did not present characteristic domains, Chr11c_nlr_73_Ca, chr0_nlr_300_Cc and Chr8_nlr_67_Ce and all these loci were partial (pseudogenes). These loci were removed from further analysis. From these results, it was possible to verify that the specificity of the NLR-annotator was 99.9% in all three genomes.

3.2 Distribution of NLR loci in the Coffea spp. Genome

Considering all detected loci, in *C. canephora*, chromosomes 3 and 11 have the most significant number of identified loci, including complete, complete (pseudogene), partial and partial (pseudogene). For *C. eugenioides*, chromosomes 3 and 11 also contain many NLR and large numbers of loci were also found on chromosomes 5 and 8. The total number of loci found on chromosome 8 is almost the same as the number identified on 11, however, the number of complete NLR on chromosome 8 is higher, with 16 more loci identified. For *C. arabica*, chromosomes 3 and 11 from the *C. canephora* and *C. eugenioides* subgenomes, respectively, also have the largest number of NLR. This what was also found on chromosome 8 from the *C. eugenioides* subgenoma. For *C. arabica*, the *C. eugenoides* subgenome generally has a slightly higher number of NLRs loci as reported above. The number of loci of this subgenome on chromosomes 8 and 11 stand out in comparison to *C. canephora* subgenome, with 34 and 30 more loci, respectively. The chromosomes with the fewest loci for the three species are 9 and 10, and chromosome 4 specifically for *C. eugenioides*. The number of loci in unmapped sequences

(Unassigned) for the *C. canephora* reference genome represent 39.7%, which was much higher than the number of loci found in *C. eugenioides* (12.5%) and *C. arabica* (2,6%) (Figure 2A).

The chromosomal location of these loci in the three species demonstrated that most loci are organized in clusters and are unevenly distributed across the entire chromosome. In addition, there are clusters that have the four different types of loci or at least two types, presenting a stretch of complete, complete (pseudogene), partial and/or partial (pseudogene). Not all loci were clustered, we also found loci of the four types that were physically isolated in chromosomes (Figure 2B).

3.3 Gene prediction for complete loci found only by NLR-annotator

Since the NLR-annotator is not a gene predictor but is a tool to annotate loci associated with NLRs, the gene-finding program AUGUSTUS was used to characterize the loci found only by NLR-annotator and that were classified as complete (S3 Table highlighted in orange). This analysis aimed to verify whether these complete loci could be considered potential gene models. This analysis showed that of the 70 and 67 complete loci for *C. arabica* and *C. canephora*, 64 and 66, are potential gene models, respectively. For *C. eugenioides*, all 71 loci were predicted as potential genes. The loci that were not identified as potential genes are in red in supplementary table 3.

3.4 Ortholog Groups and Phylogenetic Analysis

From the ortholog group analysis conducted using OrthoMCL, 803 complete loci of *C. arabica*, 561 of *C. canephora* and 695 of *C. eugenioides* were used. Six and 1 loci of *C. arabica* and *C. canephora*, respectively, were removed from analysis because they are complete loci that are not overlapping gene models or were not identified as putative genes by AUGUSTUS analysis. Additionally, 151 tomato loci (out of 326) and 403 potato loci (out of 755) that were classified as complete loci by NLR-annotator as well as 67 reference NLRs and *CED-4* were used. Out of a total of 2681 NLRs, 2038 (76%) were grouped into 593 orthologous groups, hereinafter referred to as orthogroups (Figure 3, Supplementary Table 8). The number of coffee NLRs present within these orthogroups were 591, 427, 584, which represents 73.6%, 76.1% and 84% of the total NLRs found for *C. arabica*, *C. canephora* and *C. eugenioides*, respectively. Two hundred and seventy-two orthogroups were in single-copies, containing 647 NLRs, of which only 7 are reference NLRs.

There were 10 orthogroups shared by all species and reference NLRs and 11 were shared only among species and not the reference NLRs. As expected, the greatest number of orthogroups were shared among coffee NLRs, 200 orthogroups containing 783 NLRs were shared only among *C. arabica* (296: 163 CaE, 130 CaC e 3 un), *C. canephora* (215) and *C. eugenioides* (272), respectively. The comparison between *C. arabica* NLRs with only one of the diploid species showed that *C. eugenioides* shares a slightly higher number of orthogroups (78) than *C. canephora* (71) and also of NLRs within these orthogroups (orthogroup Ca/Ce = 87/96 NLRs, orthogroup Ca/Cc = 86/74 NLRs). When the comparison was only between the two diploid species, it was observed that 34 orthogroups are shared only between them. The number of orthogroups shared between NLRs of the same coffee species was 31 in *C. eugenioides*, 24 in *C. arabica* and 9 in *C. canephora*.

The number of NLR orthogroups shared only between the three coffee species and one solanum specie was higher among potato (9) than tomato (3) NLRs, however it should be noted that the number of potato NLR in the analysis was almost 3 times larger than the number of tomato NLR. Forty-six orthogroups contain NLRs from at least one coffee species and one solanum species. Fifteen orthogroups were shared between reference NLRs, and at least one coffee and a solanum species, and grouped 21 reference NLRs (Figure 3, Supplementary Table 8, highlighted in light blue). Of these, *C. canephora* and/or *C. eugenioides* are present in three orthogroups with reference genes in which *C. arabica* is absent (ORTHOMCL16: Cc, Soly, Soltu e *Hero*; ORTHOMCL17: Ce, Soly, Soltu e *Rpiblb1*; ORTHOMCL24: Cc, Ce, Soly, Soltu e *VAT*), indicating these orthogroups are not present in the hybrid. Four orthogroups were clustered in only the three coffee species and reference NLRs (ORTHOMCL1, ORTHOMCL19, ORTHOMCL119 and ORTHOMCL199, Figure 3 and Supplementary Table 8, highlighted in dark blue), which contained *Lr10*, *MLA1*, *MLA10*, *MLA13*, *Mla12*, *Mla6*, *Pi36*, *Pikm2TS*, *FOM-2*, *Rdg2a* e *Pm3*. The percentage of orphans (i.e., NLRs not assigned to any ortholog group by OrthoMCL) among coffee NLRs was highest in *C. arabica* (26.4% -212) followed by *C. canephora* (23.9% - 134) and *C. eugenioides* (16% -111) (Supplementary Table 9).

The phylogenetic tree of single-copy orthologous NLRs showed that most clades are shared only among coffee species (Figure 4), but it was also possible to observe clades that clustered NLRs from solanum, coffee, and reference. Among the clades that clustered coffee NLRs, 71 presented groupings of orthologs between *C. arabica*, *C. canephora* and *C. eugenioides*, and most of these are located within the same chromosome. One of these clades, in addition to grouping NLRs of the three coffee species, includes the reference NLR *RPS2* (RESISTANCE to P. SYRINGAE 2) (Figure 4). This clade was supported by a high bootstrap value (100%) and was grouped in the ORTHOMCL45 (Supplementary Table 8). All loci in this cluster were found on chromosome 6 for the three coffee species. Clades that contained NLRs of *C. arabica* and *C. canephora*, *C. arabica* and *C. eugenioides* and a few *C. canephora* and *C. eugenioides* were also observed. These are within the same chromosome or on different chromosomes.

Phylogenetic analysis for coffee NLRs classification revealed that members of the NLR superfamily are grouped into 2 main groups: TIR-NLR (including TNL and NLs) and non-TNLs (including CNLs and NLs) (Figure 5). NLRs belonging to the non-TNL group outnumbered those in the TNL group in coffee genomes. We also found that the non-TNLs group is divided into 13 subgroups and that all subgroups had NLRs from all studied coffee genomes. The same pattern occurred in the TNL group. Within non-TNLs subgroups it was possible to observe clades with a greater number of NLRs from *C. arabica* that are shared with *C. eugenioides* (bands on the outer ring of the tree with a predominance of green and blue colors). There were exclusive coffee clades as well as clades that contained NLRs that were shared with potato, tomato, and reference NLRs.

4 Discussion

4.1 NLRs Identification and use of NLR-annotator in coffee genomes

In this study, we investigated loci related to genes of the NLR family in three coffee genomes and compared these loci with NLR from other plants. The annotation of genes in this family is a high priority in plant genome sequencing and annotation projects because losses from pathogens

are among the main problems for sustainable agriculture (Steuernagel et al., 2015). A better understanding of disease resistance in crops will provide plant breeders with tools that may be used to produce long-term solutions for dealing with future environmental change. A catalog of NLRs loci, whether complete genes or pseudogenes, within and between species, provides a toolbox for exploring NLRs that has not previously been described (Jones et al., 2016). Given the importance of coffee and the availability of the recent *C. arabica* and *C. eugenioides* genomes, the study of NLRs loci in these species represents an essential source of information for the development of new disease-resistant cultivars.

As NLR-annotator has already been validated in *C. canephora* (Steuernagel et al., 2020), we initially used this genome to ensure the reproducibility of the tool in our study and subsequently applied it with the *C. arabica* and *C. eugenioides* genomes. The tool identified 932 loci NLR for *C. canephora*, which is in accordance with Steuernagel et al., 2020, and also predicted of two distinct NLRs loci within the same genomic position that was previously reported. This repeatable annotation is the result of a complete NB-ARC domain preceded by a truncated NB-ARC domain, which makes the tool use the two NB-ARC domains as distinct seeds to identify two NLRs for the same locus. This has also been reported when this tool was used on the wheat genome (Steuernagel et al., 2020). The sensitivities and specificities of this tool in coffee genomes were extremely high (above 99%). In the *Arabidopsis thaliana* genome, which represents a well-annotated model plant genome, this tool achieved 95% sensitivity, and all loci that were found were validated to be associated with NLRs (specificity of 100%) (Steuernagel et al., 2020). In the Nipponbare reference genome of rice, the detection success rate was 99.2% (Read et al., 2020).

As NLR genes have repeated and clustered genomic distributions in plants, their representation in genomes using standard gene callers can be underestimated (Jupe et al., 2013; Steuernagel et al., 2015, 2020). In addition to the high rate of specificity and sensitivity, the NLR-annotator allowed for the identification of complete loci for coffee in regions distinct from the gene models already annotated in the reference genomes. This study, therefore, increased the number of putative NLR genes detected in the reference genomes of coffee species. The complete loci identified by NLR-annotator that did not overlap the reference genome annotation have also been reported in rice (Read et al., 2020). It is also relevant to highlight that as this tool is not limited to searching for functional genes, the complete (pseudogenes) loci that did not overlap with annotations of the reference genome were also identified for *C. arabica* (90), *C. canephora* (56), and *C eugenioides* (65). The location of these loci also represent an important resource, since non-functional alleles identified in sequenced accessions may represent functional NLRs in other individuals of the same species (Steuernagel et al., 2020). The caturra cultivar (*C. arabica*) sequenced and used in this study, for example, is used as a susceptible control to differentiate *Hemileia vastatrix* races among differential clones (Zambolim and Caixeta, 2021). Pseudogene regions in this genome may indicate functional genes present in other coffee cultivars.

Our data showed that 18 of the 20 loci found only by NLR-annotator, that did not present homology to NLRs proteins by BLASTx analysis, have protein domains involved in the activity of transposable elements (TE). It is known that TE are abundant in plant genomes and that they play an important role in adaptive evolution and contribute to the evolutionary dynamics of plant-pathogen interactions (Malacarne et al., 2012; Zhang et al., 2014; Kim et al., 2017). Many R genes are flanked by TE, which in addition to being sources of genetic variability, are involved

in suppressing or increasing the expression of these genes (Seidl and Thomma, 2017). The Ty3-gypsy-like TE performance has been reported in a region around the S_H3 locus associated with CLR resistance. This TE has been described in *C. arabica* subgenomes, replacing the orthologous counterpart of *C. canephora* with that of *C. eugenioides* (homoeologous non-reciprocal transposition - HNRT) (Cenci et al., 2012). Moreover, there is evidence of functional R genes that have evolved through TE-mediated duplications (Seidl and Thomma, 2017), which demonstrates their importance in the evolutionary changes and expansion of NLR receptors and justifies the presence of domains related to TE in the studied loci (Zhang et al., 2014; Kim et al., 2017).

4.2 Distribution of NLR loci in the *Coffea spp.* Genome

Although *C. arabica* results from a natural interspecific hybridization event between *C. canephora* and *C. eugenioides*, the number of loci found was not proportional to the sum of the two subgenomes, showing that the hybrid has a relatively smaller number of NLRs loci. The *C. canephora* genome size is about 690 Mbp, and the *C. eugenioides* is 665 Mbp (Noirot et al., 2003; Clarindo and Carvalho, 2011; Hamon et al., 2015). The *C. arabica* genome, on the other hand, is slightly smaller than the sum of its two combined parental genomes (about 1276 Mbp) (Hamon et al., 2015). This may explain the smaller number of NLRs in this species. Genome contraction is common in amphidiploids, which may be related to chromosomal rearrangements, including duplication, insertions, and deletions after initial hybridization (Hamon et al., 2015). An example of the number of NLRs being smaller than the sum of the corresponding parents was reported in *Brassica juncea* (Indian mustard), a species formed by hybridization between the diploid Brassica species of *B. rapa*, and *B. nigra* (Inturrisi et al., 2020). Moreover, differences in the genome assembly quality may also have interfered with the identification of NLR loci.

Among the three species analyzed, the only one with a genome-wide NLR study already reported is *C. canephora* (Denoeud et al., 2014). The NLR gene data from the previous study agrees with much of our findings. A large number of NLR loci in unanchored scaffolds for this species has also been described. Here 210 complete NLR loci were identified in unanchored scaffolds for *C. canephora*, while in the first description of the manually curated genes, 213 were not mapped. The number of mapped NLR genes was 348, while in our study, there were 352 complete loci. In *C. canephora*, it has also been reported that NLRs genes are located on all chromosomes, but with an increased number found on chromosomes 1, 3, 5, 8, 11, which together represented 70.1% of the mapped NLR genes. Although we have highlighted chromosomes 3 and 11 as having a greater number of NLR, chromosomes 1, 5, and 8 also contain large numbers of NLR loci in the three species studied here. Together, all these chromosomes represent 68.2, 71.4 and 70.0% of the total of NLR loci mapped for *C. arabica*, *C. canephora*, and *C. eugenioides*, respectively. Moreover, the low number of NLR genes on chromosomes 9 and 10 had been previously reported was confirmed in this study (Denoeud et al., 2014). These comparisons show that the three species display a conserved pattern with regards to the chromosomal distribution of NLR loci.

The NLR loci found in the three studied coffee species are arranged in clusters that group complete loci, pseudogenes and partial. These genes tend to group together and provide birth-and-death events for functional NLRs (Ling et al., 2021). In these clusters it is possible to find tandem gene duplications, recombination hotspots or active transposon elements functioning as a

reservoir of genetic variation to generate specificity for new pathogen variants (Michelmore and Meyers, 1998; Zhang et al., 2014). Within plant genomes many R genes have been found to reside in clusters (Jupe et al., 2012; Andolfo et al., 2014, 2021; Seo et al., 2016; Zheng et al., 2016; Read et al., 2020). The S_H3 locus in coffee, for example, corresponds to a complex cluster of multiple genes, including CNL-like NLR genes (Ribas et al., 2011; Cenci et al., 2012). The number of complete or functional loci in plants represents a fraction of the total number of loci found (Jupe et al., 2012; Seo et al., 2016). This happens precisely because the evolutionary dynamics within these clusters favor the coexistence of functional genes, pseudogenes, and partial genes, which differ between plants in consequence of evolutionary routes for certain pathosystems.

Recent discoveries show that NLRs can be multi domain receptors, that is they present domains integrated to the canonical form NLR or TNL/CNL (Bailey et al., 2018; Wang et al., 2021). Knowing regions of the genome that have this canonical form can facilitate the description of non-canonical integrated domains that are upstream or downstream from the more conserved region (Monteiro and Nishimura, 2018). Other relevant information is that activation of NLRs often happens in complexes and there is evidence that truncated NLRs can form heterocomplexes with complete NLRs, or may act as the main receptors in defense activation (Monteiro and Nishimura, 2018). NLRs truncated as *CbCN* (*Capsicum baccatum* – CC-NB-ARC), and *TN2* (TIR-NB-ARC) act in resistance to pathogens (Zhao et al., 2015; Son et al., 2021). This evidence reinforces the importance of knowing loci related to R genes, whether complete, pseudogenes or partial.

4.3 Ortholog groups and Phylogenetic Analyses

The genus *Coffea* belongs to the Rubiaceae family, which is in the asterid clade that also contains the Solanaceae family. Many studies have used species of the genus *Solanum* to obtain insights into the genomic and evolutionary architecture of coffee (Lin et al., 2005; Lefebvre-Pautigny et al., 2010; Denoeud et al., 2014). Species of the genus *Solanum* have also been used as models for understanding the molecular processes related to plant-pathogen interaction. This supports their use for comparative approaches to lead to discoveries of NLR loci or functionally important gene families in coffee (Andolfo et al., 2021). Our results showed that of the 17 reference NLRs that have been cloned and characterized in species of the genus *Solanum* and that were also used in this study, 8 of them are present in shared orthogroups with coffee (Supplementary Table 2 and Supplementary Table 8), being 1 TNL (*Gro1-4)* and 7 CNL (*Hero*, *Prf*, *Rpi-blb1*, *Rx2*, *Sw-5*, *Tm-2a*, *Tm-2*). These genes have been found to be involved in resistance to a diverse group of pathogens including viruses, oomycetes, bacteria and nematodes (Bendahmane et al., 2000; Van Der Vossen et al., 2003; Paal et al., 2004; Andolfo et al., 2021). In total, 46 orthogroups were shared between coffee and solanum indicating that these orthologs were probably present before the speciation of these two groups. Reference NLRs characterized in species such as *Hordeum vulgare*, *Oryza sativa*, *Triticum aestivum*, *Glycine max*, *Arabidopsis thaliana* e *Cucumis melo* also share othogroups with coffee NLRs and all of these NLR belong to the CNL class (Supplementary Table 2 and Supplementary Table 8). These orthogroups are important as they indicate roles that may be inferred and further investigated in coffee. An interesting orthogroup that obtained high support in the phylogenetic tree was the one that clustered the *RPS2* reference NLR as well as NLRs present in all three coffee species. *RPS2 i*s a

resistance gene of *Arabidopsis thaliana* that confers resistance to *Pseudomonas syringae* bacteria that express the *avrRpt2* avirulence gene (Bent et al., 1994; Mindrinos et al., 1994).

In general, plant species exhibit differences in the number of NLR genes contained within their genomes. Amplification of certain groups has been detected (Seo et al., 2016). This diversity has not been associated with genome size or phylogenetic relationships, but is related to the specialization of each particular host (Wan et al., 2013; Lozano et al., 2015; Seo et al., 2016; Zheng et al., 2016; Steuernagel et al., 2020). In all three coffee species, a set of orphan genes and orthogroups that share NLRs within the same species were detected. In tomato, 45 of ~320 NLRs sequences are more similar to each other than to any other sequences compared (Andolfo et al., 2021) and orthogroups that share NLRs within the same species of *Solanum* are attributed to duplication events that generate different gene repertoires and result in species-specific subfamilies (Seo et al., 2016).

The single-copy ortrogroups provide more reliable results for interpreting evolutionary processes between groups of evaluated genes by allowing for the identification of true orthologs between different groups of plants (Zimmer et al., 2007; Duarte et al., 2010). The results from the single-copy ortrogroups tree showed that some orthologous NLRs seem to have a common ancestor only among coffee species. The S_H3 locus, for example, was described as being shared only among coffee species suggesting that the ancestral copy S_H3-CNL was inserted into the S_H3 locus after the divergence of the *Solanum* and *Coffea* lineages (Ribas et al., 2011). The clades that clustered orthologous NLRs from *C. arabica*, *C. canephora* and *C. eugenioides* probably represent NLR present in both ancestral diploids genomes, which were passed to *C. arabica* genome. On the other hand, clades that clustered only *C. canephora* and *C. eugenioides* loci may represent ancestral NLRs that were lost in *C. arabica* or that underwent so many modifications in this species that makes it difficult to find homology between these NLRs. These NLR may provide valuable resistance mechanisms that are not present in the *C. arabica* hybrid. Nucleotide level changes, such as deletions, insertions and rearrangements have been observed in coffee RGA (Resistance gene analogues) (Noir et al., 2001; Hendre et al., 2011). It is also known that it is very likely that the sequenced genotypes of *C. canephora* and *C. eugenioides* present significant differences from the ancestral donors of *C. arabica* subgenomes, which may explain the lack of homology in certain NLR groups (Cenci et al., 2012).

The NB-ARC is the most conserved domain in the NLR gene family. Despite the conservation of this domain, it is possible to distinguish the TIR (TNL) and non-TIR classes based on different residues inside the motifs present in this region (Jones et al., 2016; Shao et al., 2019; Van Ghelder et al., 2019). Therefore, this domain has been used to describe the phylogenetic relationships between the sequences of this group and classify them (Andolfo et al., 2014). The classification of NLRs in coffee revealed that the non-TNL class were present in greater numbers than those of the TNL group in each of the three analyzed coffee genomes. It is known that non-TNL genes that include many CNL are widely distributed in monocots and dicots, while TNL are mainly found in dicots (McHale et al., 2006; Zheng et al., 2016). The low frequency of TNLs in coffee agrees with results found for species of the solanum group, such as pepper, tomato and potato (Andolfo et al., 2014; Seo et al., 2016). Our results are also consistent with the low frequency of TNLs found in previous studies of coffee (Hendre et al., 2011; Denoeud et al., 2014). Thus, it is possible to suggest that, as in solanum, non-TNLs represent an important repertoire of resistance genes in coffee. Additionally, the TNL group and the non-TNLs

subgroups contained NLRs from *C.arabica*, *C. canephora* and *C. eugenioides*, indicating conservation of the NLR classes across coffee genomes and suggesting that all subgroups were present in a common ancestor, similar to what has been described for comparisons of species within the solanum group (Seo et al., 2016).

In the two phylogenetic trees analyzed, the clades group coffee NLRs that are mostly present in the same chromosomes but groupings of NLRs present on different chromosomes were also detected. Genes located on the same chromosome tend to group into subclades in the phylogenetic tree. However, rearrangement events of the chromosomes can affect NLR loci and modify their genomic order or location (Cenci et al., 2012; Zheng et al., 2016; Ling et al., 2021).

Considering the relevance of coffee, few studies have been conducted addressing the identification of NLR in genomes of this crop. RGA studies using degenerate primers for NB-ARC region have already been performed (Noir et al., 2001; Hendre et al., 2011; Kumar, 2012), in addition to studies in S_H3 locus (Cenci et al., 2010, 2012; Lashermes et al., 2010), but very little is known about the NLR family in cultivated (*C.arabica* e *C. canephora*) and uncultivated coffee species (such as *C. eugenioides*). This is the first study focused on genome wide identification of NLRs in the *C. arabica* genome, and also adds information to the existing report for the *C. canephora* genome (Denoeud et al., 2014). The Genome-wide identification of coffee NLRs allows for more in-depth molecular studies provides further information for identifying candidate genes for cloning and subsequent functional validation of NLR genes while also expanding the range of NLR that are available for breeding of this crop (Seo et al., 2016).

5 Conclusion

Our analysis showed that the use of a specific pipeline for resistance genes was efficient in detecting NLR loci in the studied coffee genomes and increased the information available for the location of these loci in *C. arabica*, *C. canephora* and *C. eugenioides*. The NLR loci in the three coffee species are unevenly distributed across all chromosomes and are mostly arranged in clusters. The number of loci in *C. arabica* is less than the sum of the NLR loci from the parents of this hybrid. Single-copy NLR orthogroups investigated in a phylogenetic tree identified orthologous NLRs that are shared between *C. arabica* and the parental genomes as well as NLR that were possibly lost. Coffee NLRs and some functionally characterized NLRs share common orthogroups, which provides clues to the functionality of coffee NLRs, and paves the way for further investigation. Loci from the NLR family are subdivided into two main groups in coffee: TIR-NLR (TNL) and non-TNL, but non-TNLs are present in greater number and seem to represent an important repertoire of resistance genes in coffee.

6 Conflict of Interest

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

7 Author Contributions

MLi, GS, and MF: conceptualization. MLi, MLu, GS, JC, MF, and JT: methodology. MLi, JC: software, formal analysis, investigation. MLi writing – original draft preparation. MLi, MLu, GS, JC, MF, and JT: visualization. MLi, MLu, GS, JC, MF, and JT: validation. MLu, MF, and JT: resources, supervision, project administration, and funding acquisition. MLi, MLu, GS, JC, MF, and JT: data curation; writing – review and editing. All authors have read and agreed to the published version of the manuscript.

8 Funding

This research was funded by Coordination for the Improvement of Higher Education Personnel (CAPES) and University of Florida to MLS, and the USDA McIntire-Stennis project FLA-PLP-005931 “Developing and scaling up the next generation of healthy forests" to JTB.

9 Acknowledgments

We are grateful to the Universidade Federal de Lavras, the Universidade de Brasília, the University of Florida, U.S. Department of Agriculture’s Agriculture and Food Research Initiative grant FLA-PLP006039, University of Florida’s Institute for Food and Agricultural Science, and Corn Breeding and Genomics program for supporting this collaborative research project.

References

Anand, L., and Lopez, C. M. R. (2020). chromoMap: An R package for Interactive Visualization and Annotation of Chromosomes.

Andolfo, G., D’agostino, N., Frusciante, L., and Ercolano, M. R. (2021). The tomato interspecific NB-LRR gene arsenal and its impact on breeding strategies. *Genes.* 12, 1–12. doi:10.3390/genes12020184.

Andolfo, G., Jupe, F., and Witek, K. (2014). Defining the full tomato NB-LRR resistance gene repertoire using genomic and cDNA RenSeq. *BMC plant* 14, 120, 2014. doi:10.1186/1471-2229-14-120.

Bailey, P. C., Schudoma, C., Jackson, W., Baggs, E., Dagdas, G., Haerty, W., et al. (2018). Dominant integration locus drives continuous diversification of plant immune receptors with exogenous domain fusions. *Genome Biol.* 19, 1–18. doi:10.1186/s13059-018-1392-6.

Barragan, A. C., and Weigel, D. (2021). Plant NLR Diversity: The Known Unknowns of Pan-NLRomes. *Plant Cell* 33, 814–831. doi:/10.1093/plcell/koaa002.

Bawin, Y., Ruttink, T., Staelens, A., Haegeman, A., Stoffelen, P., Mwanga Mwanga, J. C. I., et al. (2020). Phylogenomic analysis clarifies the evolutionary origin of *Coffea arabica*. *J.*

Syst. Evol. doi:10.1111/jse.12694.

Bendahmane, A., Querci, M., Kanyuka, K., and Baulcombe, D. C. (2000). Agrobacterium transient expression system as a tool for the isolation of disease resistance genes: Application to the *Rx2* locus in potato. *Plant J.* 21, 73–81. doi:/10.1046/j.1365-313x.2000.00654.x.

Bent, A. F., Kunkel, B. N., Dahlbeck, D., Brown, K. L., Schmidt, R., Giraudat, J., et al. (1994). RPS2 of Arabidopsis thaliana: A leucine-rich repeat class of plant disease resistance genes. *Science* 265, 1856–1860. doi:10.1126/science.8091210.

Bezerra-Neto, J. P., Araújo, F. C., Ferreira-Neto, J. R. C., Silva, R. L. O., Borges, A. N. C., Matos, M. K. S., et al. (2020). "NBS-LRR genes-Plant health sentinels: Structure, roles, evolution and biotechnological applications," in *Applied Plant Biotechnology for Improving Resistance to Biotic Stress* (Elsevier), 63–120. doi:10.1016/B978-0-12-816030-5.00004-5.

Bonardi, V., Cherkis, K., Nishimura, M. T., and Dangl, J. L. (2012). A new eye on NLR proteins: Focused on clarity or diffused by complexity? *Curr. Opin. Immunol.* 24, 41–50. doi:10.1016/j.coi.2011.12.006.

Boutrot, F., and Zipfel, C. (2017). Function, discovery, and exploitation of plant pattern recognition receptors for broad-spectrum disease resistance. *Annu. Rev. Phytopathol.* 55, 257–286. doi:10.1146/annurev-phyto-080614-120106.

Cabral, P. G. C., Maciel-Zambolim, E., Oliveira, S. A. S., Caixeta, E. T., and Zambolim, L. (2016). Genetic diversity and structure of *Hemileia vastatrix* populations on *Coffea spp.* *Plant Pathol.* 65, 196–204. doi:10.1111/ppa.12411.

Cenci, A., Combes, M. C., and Lashermes, P. (2010). Comparative sequence analyses indicate that Coffea (Asterids) and Vitis (Rosids) derive from the same paleo-hexaploid ancestral genome. *Mol. Genet. Genomics* 283, 493–501. doi:10.1007/s00438-010-0534-7.

Cenci, A., Combes, M. C., and Lashermes, P. (2012). Genome evolution in diploid and tetraploid Coffea species as revealed by comparative analysis of orthologous genome segments. *Plant Mol. Biol.* 78, 135–145. doi:10.1007/s11103-011-9852-3.

Cerda, R., Avelino, J., Gary, C., Tixier, P., Lechevallier, E., and Allinne, C. (2017). Primary and secondary yield losses caused by pests and diseases: Assessment and modeling in coffee. *PLoS One* 12, 1–17. doi:10.1371/journal.pone.0169133.

Clarindo, W. R., and Carvalho, C. R. (2011). Flow cytometric analysis using SYBR Green I for genome size estimation in coffee. *Acta Histochem.* 113, 221–225. doi:10.1016/j.acthis.2009.10.005.

Conway, J. R., Lex, A., and Gehlenborg, N. (2017). UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics* 33, 2938–2940. doi:10.1093/bioinformatics/btx364.

Davis, A. P., Govaerts, R., Bridson, D. M., and Stoffelen, P. (2006). An annotatated taxonomic of the genus coffea (Rubiaceae). *Bot. J. Linn. Soc.* 152, 465–512. doi:10.1111/j.1095-8339.2006.00584.x.

Denoeud, F., Carretero-Paulet, L., Dereeper, A., Droc, G., Guyot, R., Pietrella, M., et al. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* 345, 1181–1184. doi:10.1126/science.1255274.

Duarte, J. M., Wall, P. K., Edger, P. P., Landherr, L. L., Ma, H., Pires, J. C., et al. (2010). Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels. *BMC Evol. Biol.* 10, 1–18. doi:10.1186/1471-2148-10-61.

Hamon, P., Hamon, S., Razafinarivo, N. J., Guyot, R., Siljak-Yakovlev, S., Couturon, E., et al. (2015). "Coffea Genome Organization and Evolution," in *Coffee in Health and Disease Prevention*, 29–37. doi:10.1016/B978-0-12-409517-5.00004-8.

Hendre, P. S., Bhat, P. R., Krishnakumar, V., Aggarwal, R. K., and Donini, P. (2011). Isolation and characterization of resistance gene analogues from Psilanthus species that represent wild relatives of cultivated coffee endemic to India. *Genome* 54, 377–390. doi:10.1139/g11-004.

Hufford, M. B., Seetharam, A. S., Woodhouse, M. R., Chougule, K. M., Coletta, D., Tittes, S., et al. (2021). De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science*. doi:10.1126/science.abg5289.

Inturrisi, F., Bayer, P. E., Yang, H., Tirnaz, S., Edwards, D., and Batley, J. (2020). Genome-wide identification and comparative analysis of resistance genes in *Brassica juncea*. *Mol. Breed.* 40, 1–14. doi:10.1101/2020.12.15.422814.

Jones, J. D. G., and Dangl, J. L. (2006). The plant immune system. *Nature* 444, 323–329. doi:10.1038/nature05286.

Jones, J. D. G., Vance, R. E., and Dangl, J. L. (2016). Intracellular innate immune surveillance devices in plants and animals. *Science* 354. doi:10.1126/science.aaf6395.

Jost, M., Singh, D., Lagudah, E., Park, R. F., and Dracatos, P. (2020). Fine mapping of leaf rust resistance gene Rph13 from wild barley. *Theor. Appl. Genet.* 133, 1887–1895. doi:10.1007/s00122-020-03564-6.

Jupe, F., Pritchard, L., Etherington, G. J., MacKenzie, K., Cock, P. J. A., Wright, F., et al. (2012). Identification and localisation of the NB-LRR gene family within the potato genome. *BMC Genomics* 13, 1–14. doi:10.1186/1471-2164-13-75.

Jupe, F., Witek, K., Verweij, W., Śliwka, J., Pritchard, L., Etherington, G. J., et al. (2013). Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. *Plant J.* 76, 530–544. doi:10.1111/tpj.12307.

Kachroo, A., and Robin, G. P. (2013). Systemic signaling during plant defense. *Curr. Opin. Plant Biol.* 16, 527–533. doi:10.1016/j.pbi.2013.06.019.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–66. doi:/10.1093/nar/gkf436.

Kim, S., Park, J., Yeom, S. I., Kim, Y. M., Seo, E., Kim, K. T., et al. (2017). New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. *Genome Biol.* 18, 1–11. doi:10.1186/s13059-017-1341-9.

Kourelis, J., and Van Der Hoorn, R. A. (2018). Defended to the Nines: 25 years of Resistance Gene Cloning Identifies Nine Mechanisms for R Protein Function. *Plant Cell* 30, 285–299. doi:10.1105/tpc.17.00579.

Krishnan, S. (2017). Sustainable Coffee Production. *Oxford Res. Encycl. Environ. Sci.* doi:10.1093/acrefore/9780199389414.013.224.

Kroj, T., Chanclud, E., Michel-Romiti, C., Grand, X., and Morel, J. B. (2016). Integration of decoy domains derived from protein targets of pathogen effectors into plant immune receptors is widespread. *New Phytol.* 210, 618–626. doi:10.1111/nph.13869.

Kumar, D. (2012). Isolation of Nucleotide Binding Site (NBS)-Leucine Rich Repeat (LRR) Resistant Gene Analogs (Rgas) In Arabica Coffee (*Coffea arabica* L. Cv S.288). *J. Biotechnol. Biomater.* 02. doi:10.4172/2155-952x.1000146.

Kushwaha, S. K., Chauhan, P., Hedlund, K., and Ahren, D. (2016). NBSPred: A support vector machine-based high-throughput pipeline for plant resistance protein NBSLRR prediction. *Bioinformatics* 32, 1223–1225. doi:10.1093/bioinformatics/btv714.

Lashermes, P., Combes, M. C., Ribas, A., Cenci, A., Mahé, L., and Etienne, H. (2010). Genetic and physical mapping of the SH3 region that confers resistance to leaf rust in coffee tree (*Coffea arabica* L.). *Tree Genet. Genomes* 6, 973–980. doi:10.1007/s11295-010-0306-x.

Lashermes, P., Combes, M. C., Robert, J., Trouslot, P., D'Hont, A., Anthony, F., et al. (1999). Molecular characterization and origin of the Coffea arabica L. genome. *Mol. Gen. Genet.* 261, 259–266. doi:10.1007/s004380050965.

Lefebvre-Pautigny, F., Wu, F., Philippot, M., Rigoreau, M., Priyono, Zouine, M., et al. (2010). High resolution synteny maps allowing direct comparisons between the coffee and tomato genomes. *Tree Genet. Genomes* 6, 565–577. doi:10.1007/s11295-010-0272-3.

Letunic, I., and Bork, P. (2021). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* 49, W293–W296. doi:10.1093/nar/gkw290.

Li, L., Stoeckert, C. J. J., and Roos, D. S. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res.* 13, 2178–2189. doi:10.1101/gr.1224503.candidates.

Lin, C., Mueller, L. A., Carthy, J. M., Crouzillat, D., Pétiard, V., and Tanksley, S. D. (2005). Coffee and tomato share common gene repertoires as revealed by deep sequencing of seed and cherry transcripts. *Theor. Appl. Genet.* 112, 114–130. doi:10.1007/s00122-005-0112-2.

Ling, J., Xie, X., Gu, X., Zhao, J., Ping, X., Li, Y., et al. (2021). High-quality chromosome-level genomes of *Cucumis metuliferus* and *Cucumis melo* provide insight into Cucumis genome evolution. *Plant J.* 107, 136–148. doi:10.1111/tpj.15279.

Lozano, R., Hamblin, M. T., Prochnik, S., and Jannink, J. L. (2015). Identification and distribution of the NBS-LRR gene family in the Cassava genome. *BMC Genomics* 16, 1–14. doi:10.1186/s12864-015-1554-9.

Lu, Y., and Tsuda, K. (2021). Intimate Association of PRR- and NLR-Mediated Signaling in Plant Immunity. *Mol. Plant-Microbe Interact.* 34, 3–14. doi:10.1094/mpmi-08-20-0239-ia.

Malacarne, G., Perazzolli, M., Cestaro, A., Sterck, L., Fontana, P., van de Peer, Y., et al. (2012). Deconstruction of the (paleo)polyploid grapevine genome based on the analysis of transposition events involving NBS resistance genes. *PLoS One* 7. doi:10.1371/journal.pone.0029762.

McCook, S., and Vandermeer, J. (2015). The Big Rust and the Red Queen: Long-Term Perspectives on Coffee Rust Research. *Phytopathology* 105, 1164–1173. doi:10.1094/PHYTO-04-15-0085-RVW.

McHale, L., Tan, X., Koehl, P., and Michelmore, R. W. (2006). Plant NBS-LRR proteins: Adaptable guards. *Genome Biol.* 7. doi:10.1186/gb-2006-7-4-212.

Michelmore, R. W., and Meyers, B. C. (1998). Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res.* 8, 1113–1130. doi:10.1101/gr.8.11.1113.

Mindrinos, M., Katagiri, F., Yu, G. L., and Ausubel, F. M. (1994). The A. thaliana disease resistance gene RPS2 encodes a protein containing a nucleotide-binding site and leucine-rich repeats. *Cell* 78, 1089–1099. doi:10.1016/0092-8674(94)90282-8.

Monteiro, F., and Nishimura, M. T. (2018). Structural, functional, and genomic diversity of plant NLR proteins: An evolved resource for rational engineering of plant immunity. *Annu. Rev. Phytopathol.* 56, 243–267. doi:10.1146/annurev-phyto-080417-045817.

Muliyar, R. K., Chowdappa, P., Behera, S. K., Kasaragod, S., Gangaraj, K. P., Kotimoole, C. N., et al. (2020). Assembly and Annotation of the Nuclear and Organellar Genomes of a Dwarf Coconut (Chowghat Green Dwarf) Possessing Enhanced Disease Resistance. *Omi. A J. Integr. Biol.* 24, 726–742. doi:10.1089/omi.2020.0147.

Noir, S., Combes, M.-C., Anthony, F., and Lashermes, P. (2001). Origin, diversity and evolution of NBS-type disease-resistance gene homologues in coffee trees (Coffea L.). *Mol. Genet. Genomics* 265, 654–662. doi:10.1007/s004380100459.

Noirot, M., Poncet, V., Barre, P., Hamon, P., Hamon, S., and De Kochko, A. (2003). Genome size variations in diploid African Coffea species. *Ann. Bot.* 92, 709–714. doi:10.1093/aob/mcg183.

Osuna-Cruz, C. M., Paytuvi-Gallart, A., Di Donato, A., Sundesha, V., Andolfo, G., Cigliano, R. A., et al. (2018). PRGdb 3.0: A comprehensive platform for prediction and analysis of plant disease resistance genes. *Nucleic Acids Res.* 46, D1197–D1201. doi:10.1093/nar/gkx1119.

Paal, J., Henselewski, H., Muth, J., Meksem, K., Menéndez, C. M., Salamini, F., et al. (2004). Molecular cloning of the potato Gro1-4 gene conferring resistance to pathotype Ro1 of the root cyst nematode Globodera rostochiensis, based on a candidate gene approach. *Plant J.* 38, 285–297. doi:10.1111/j.1365-313X.2004.02047.x.

Read, A. C., Moscou, M. J., Zimin, A. V., Pertea, G., Meyer, R. S., Purugganan, M. D., et al. (2020). Genome assembly and characterization of a complex zfBED-NLR gene-containing disease resistance locus in Carolina Gold Select rice with Nanopore sequencing. *PLoS Genet.* 16, 1–24. doi:10.1371/journal.pgen.1008571.

Ribas, A. F., Cenci, A., Combes, M. C., Etienne, H., and Lashermes, P. (2011). Organization and molecular evolution of a disease-resistance gene cluster in coffee trees. *BMC Genomics* 12. doi:10.1186/1471-2164-12-240.

Schatz, M. C., Maron, L. G., Stein, J. C., Hernandez Wences, A., Gurtowski, J., Biggers, E., et al. (2014). Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. *Genome Biol.* 15, 506. doi:10.1186/s13059-014-0506-z.

Scott, A. D., Zimin, A. V, Puiu, D., Workman, R., Britton, M., Zaman, S., et al. (2020). A Reference Genome Sequence for Giant Sequoia. *G3: Genes, Genomes, Genetics*. 10, 3907–3919. doi: 10.1534/g3.120.401612.

Seidl, M. F., and Thomma, B. P. H. J. (2017). Transposable Elements Direct The Coevolution between Plants and Microbes. *Trends Genet.* 33, 842–851. doi:10.1016/j.tig.2017.07.003.

Seo, E., Kim, S., Yeom, S. I., and Choi, D. (2016). Genome-wide comparative analyses reveal the dynamic evolution of nucleotide-binding leucine-rich repeat gene family among solanaceae plants. *Front. Plant Sci.* 7, 1–13. doi:10.3389/fpls.2016.01205.

Shao, Z. Q., Xue, J. Y., Wang, Q., Wang, B., and Chen, J. Q. (2019). Revisiting the Origin of Plant NBS-LRR Genes. *Trends Plant Sci.* 24, 9–12. doi:10.1016/j.tplants.2018.10.015.

Son, S., Kim, S., Lee, K., Oh, J., Choi, I., Do, J., et al. (2021). The *Capsicum baccatum*-Specific Truncated NLR Protein CbCN Enhances the Innate Immunity against Colletotrichum acutatum. *Int. J. Mol. Sci.* 22. doi:10.3390/ijms22147672.

Song, W., Wang, B., Li, X., Wei, J., Chen, L., Zhang, D., et al. (2015). Identification of Immune Related LRR-Containing Genes in Maize (*Zea mays L.*) by Genome-Wide Sequence Analysis. *Int. J. Genomics* 2015. doi:10.1155/2015/231358.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi:10.1093/bioinformatics/btu033.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* 34, W435–W439. doi:10.1093/nar/gkl200.

Steuernagel, B., Jupe, F., Witek, K., Jones, J. D. G., and Wulff, B. B. H. (2015). NLR-parser: Rapid annotation of plant NLR complements. *Bioinformatics* 31, 1665–1667. doi:10.1093/bioinformatics/btv005.

Steuernagel, B., Witek, K., Krattinger, S. G., Ramirez-Gonzalez, R. H., Schoonbeek, H. J., Yu, G., et al. (2020). The NLR-annotator tool enables annotation of the intracellular immune receptor repertoire. *Plant Physiol.* 183, 468–482. doi:10.1104/pp.19.01273.

Sun, X., Jiao, C., Schwaninger, H., Chao, C. T., Ma, Y., Duan, N., et al. (2020). Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nat. Genet.* 52, 1423–1432. doi:10.1038/s41588-020-00723-9.

Toda, N., Rustenholz, C., Baud, A., Le Paslier, M. C., Amselem, J., Merdinoglu, D., et al. (2020). NLGenomeSweeper: A tool for genome-wide NBS-LRR resistance gene identification. *Genes.* 11, 333. doi:10.3390/genes11030333.

Tran, H. T. M., Ramaraj, T., Furtado, A., Lee, L. S., and Henry, R. J. (2018). Use of a draft genome of coffee (*Coffea arabica*) to identify SNPs associated with caffeine content. *Plant Biotechnol. J.* 9, 1756–1766. doi:10.1111/pbi.12912.

Van Der Vossen, E., Sikkema, A., Te Lintel Hekkert, B., Gros, J., Stevens, P., Muskens, M., et al. (2003). An ancient R gene from the wild potato species *Solanum bulbocastanum* confers broad-spectrum resistance to *Phytophthora infestans* in cultivated potato and tomato. *Plant J.* 36, 867–882. doi:10.1046/j.1365-313X.2003.01934.x.

Van Ghelder, C., Parent, G. J., Rigault, P., Prunier, J., Giguère, I., Caron, S., et al. (2019). The large repertoire of conifer NLR resistance genes includes drought responsive and highly diversified RNLs. *Sci. Rep.* 9, 1–13. doi:10.1038/s41598-019-47950-7.

Wan, H., Yuan, W., Bo, K., Shen, J., Pang, X., and Chen, J. (2013). Genome-wide analysis of NBS-encoding disease resistance genes in *Cucumis sativus* and phylogenetic study of NBS-encoding genes in Cucurbitaceae crops. *BMC Genomics* 14, 109. doi:10.1186/1471-2164-14-109.

Wang, Z., Ren, H., Xu, F., Lu, G., Cheng, W., Que, Y., et al. (2021). Genome-Wide Characterization of NLRs in *Saccharum spontaneum* L. and Their Responses to Leaf Blight in Saccharum. *Agronomy* 11, 153. doi:10.3390/plants10020322.

Yuan, M., Jiang, Z., Bi, G., Nomura, K., Liu, M., He, S. Y., et al. (2021). Pattern-recognition receptors are required for NLR-mediated plant immunity. *Nature*, 1–5. doi:10.1101/2020.04.10.031294.

Zambolim, L., and Caixeta, E. T. (2021). An overview of physiological specialization ofcoffee leaf rust – new designation of Pathotypes. *Int. J. Curr. Res.* 13, 15564–15575.

Zhang, R., Murat, F., Pont, C., Langin, T., and Salse, J. (2014). Paleo-evolutionary plasticity of plant disease resistance genes. *BMC Genomics* 15. doi:10.1186/1471-2164-15-187.

Zhao, T., Rui, L., Li, J., Nishimura, M. T., Vogel, J. P., Liu, N., et al. (2015). A Truncated NLR Protein, TIR-NBS2, Is Required for Activated Defense Responses in the exo70B1 Mutant. *PLoS Genet.* 11, 1–28. doi:10.1371/journal.pgen.1004945.

Zheng, F., Wu, H., Zhang, R., Li, S., He, W., Wong, F. L., et al. (2016). Molecular phylogeny and dynamic evolution of disease resistance genes in the legume family. *BMC Genomics* 17, 1–13. doi:10.1186/s12864-016-2736-9.

Zimmer, A., Lang, D., Richardt, S., Frank, W., Reski, R., and Rensing, S. A. (2007). Dating the early evolution of plants: Detection and molecular clock analyses of orthologs. *Mol. Genet. Genomics* 278, 393–402. doi:10.1007/s00438-007-0257-6.

Figure 1. Venn diagrams representing the overlap between the loci from NLR-annotator and NLR genes annotated in the *C. arabica*, *C. canephora,* and *C. eugenioides* reference genomes. The colors represent the origin of the annotation, with blue indicating those annotated by NLR-annotator and green indicating those found in the reference genome. The intersection refers to the overlaps that occurred once or more than once.

Figure 2. Number and chromosomal distribution of NLR loci in *C. arabica, C. canephora* e *C. eugenioides*. (A) The chromosomes with the highest number of NLR loci are highlighted in dark blue, and those with the smallest number of NRL are highlighted in light blue. CPL identifies the completeness of NLR as: C= complete, Cps= complete (pseudogene), P= partial, Pps= partial pseudogene and Un= unassigned. **(B)** The chromosomes represented in *C. arabica* refer to the two subgenomes with the first originating from *C. canephopr*a and the second originating from *C. eugenioides*. The chromosomal distribution represented in this figure does not show all loci for each region but identifies all regions that contain NLRs loci. A more detailed view of these chromosomes with the detail of all regions may be found at: https://1drv.ms/u/s!As084N7WlXAIhMZfxevH93zPgU-YhQ?e=0KQ8Iy. Access the link and download the HTML file.

Figure 3. **Upset plot of orthologous NLR groups (orthogroups) among five species, *C. arabica* (Ca) *C. canephora* (Cc), *C. eugenioides* (Ce), *S. tuberosum* (Soltu) and *S. lycopersicum* (Soly) and NLRs de referência (Ref).** The orthogroups that cluster combinations of species/Ref NLRs is shown by the interconnected blue dots on the bottom panel. Unconnected blue dots show orthogroups that are present within the same species. The 'Set size' represents the total number of orthogroups per species/Ref. The 'intersection size' shows the number of orthogroups shared between species/Ref or within the same species/Ref. The orthogroups were clustered using OrthoMCL.

Figure 4. **Phylogenetic tree of single-copy NLR orthogroups.** The phylogenetic tree was constructed using RAxML and was based on 647 NLRs (domain NB-ARC) that were single-copy orthologs. The colored ring indicates coffee NLR clades, the green color represents *C. arabica* (Ca), red represents *C. canephora* (Cc) and blue represents *C. eugenioides* (Ce). Labels in black are coffee NLRs, green is used for *S. lycopersicum* (Soly), blue is used for *S. tuberosum* (Soltu) and pink indicates reference proteins (Ref). Bootstrap values above 70% are indicated on each branch with a brown circle. The pink background identifies clades that group orthologs of Ca, Cc and Ce. The clade highlighted in purple shows the coffee NLRs and RPS2 grouping.

Figure 5. **Phylogenetic tree for coffee NLRs classification**. NB-ARC domains from 2681 NLRs clustering *C. arabica* (*Ca*), *C. canephora (Cc), C. eugenioides* (Ce), *S. lycopersicum*, *S. tuberosum*, reference NLRs (pink balls) and CED-4 (root nematode outlier) were used to construct the tree. The tree was constructed using RAxML. The classifications of the reference NLRs and some *S. lycopersicum*, *S. tuberosum* NLRs were used to classify the coffee NLRs into TNLs and Non TNLs groups (inner ring - TNL = yellow, CNL = gray and NL = Purple). Subgroups in Non TNLs are indicated from I to XIII and alternating colors (green and purple). Gray and yellow background highlight coffee NLRs and the outer ring separates the NLRs for Ca, Cc and Ce in green, red and blue respectively.

Table 1. Loci identified using NLR-annotator that did not overlap with annotations of NLR genes from coffee reference genomes.

Species	Complete	Complete (pseudogene)	Partial	Partial (pseudogene)	Total
C. arabica	70	90	65	73	298
C. canephora	67	56	73	44	240
C. eugenioides	71	65	37	49	222

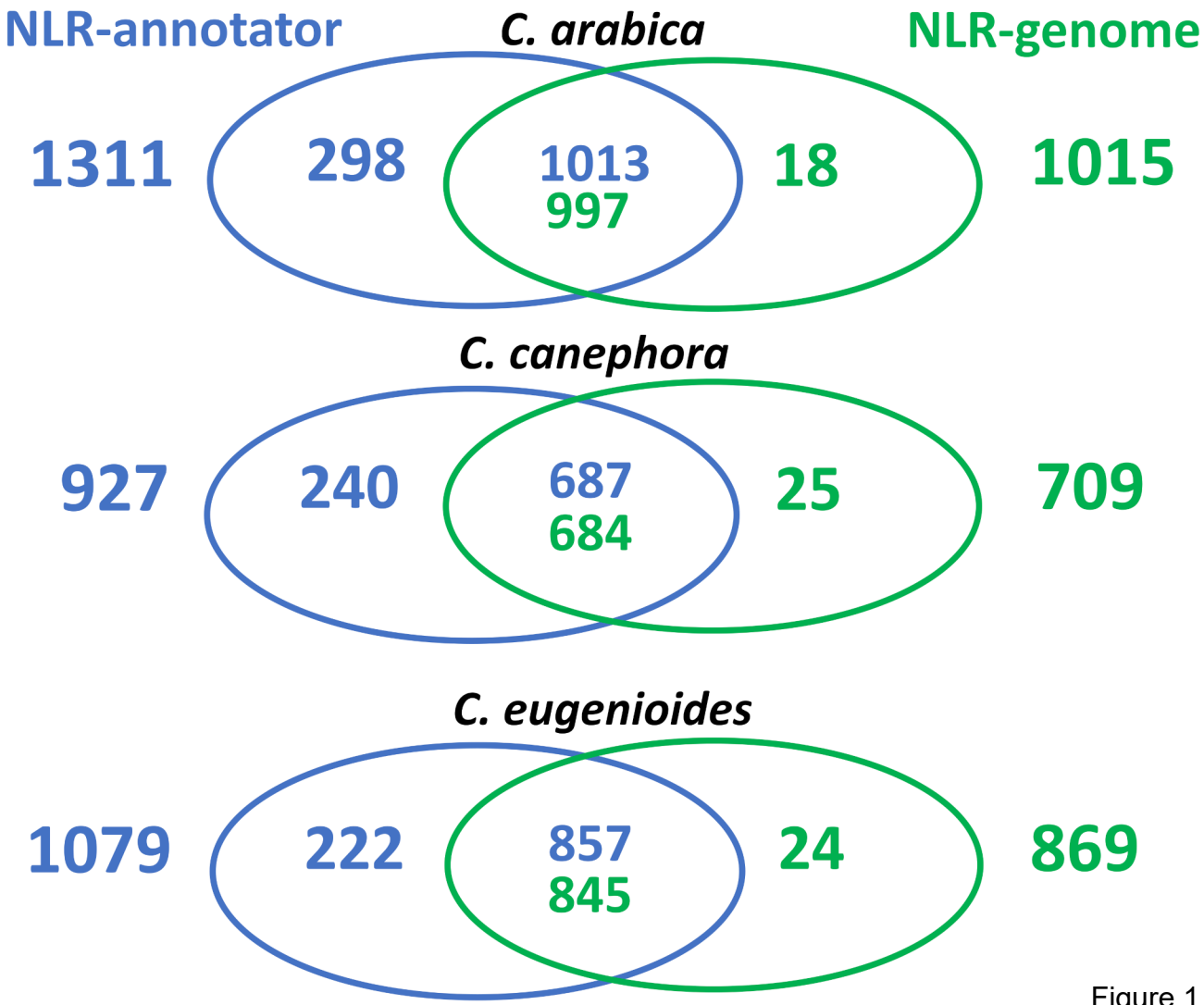

NLR-annotator
C. arabica
NLR-genome
1311
298
1013
997
18
1015
C. canephora
927
240
687
684
25
709
C. eugenioides
1079
222
857
845
24
869

Figure 1

A

Coffea arabica

CPL	Chr1c	Chr1e	Chr2c	Chr2e	Chr3c	Chr3e	Chr4c	Chr4e	Chr5c	Chr5e	Chr6c	Chr6e	Chr7c	Chr7e	Chr8c	Chr8e	Chr9c	Chr9e	Chr10c	Chr10e	Chr11c	Chr11e	Un	Total
C	42	48	17	22	71	65	25	21	42	49	23	15	44	43	35	54	11	7	16	9	54	73	23	809
Cps	21	13	7	8	38	24	5	6	28	19	3	6	14	15	12	23	1	4	1	1	12	23	5	289
P	5	4	4	6	13	12	1	4	4	5	10	2	5	6	4	3	1	0	5	8	10	7	6	125
Pps	5	2	2	3	8	4	2	1	1	2	2	4	2	4	4	9	2	0	4	3	10	13	0	88
Total	**73**	**67**	**30**	**39**	**130**	**105**	**33**	**32**	**75**	**75**	**38**	**27**	**65**	**68**	**55**	**89**	**15**	**11**	**26**	**21**	**86**	**116**	**34**	**1311**

Coffea canephora

CPL	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	Chr7	Chr8	Chr9	Chr10	Chr11	Un	Total
C	49	19	70	23	42	16	29	37	1	11	55	210	562
Cps	14	9	25	7	14	4	7	12	1	1	14	66	174
P	3	7	12	3	5	2	5	2	3	3	8	68	121
Pps	6	2	13	1	2	2	2	3	1	1	13	23	70
Total	**72**	**37**	**120**	**34**	**63**	**24**	**43**	**54**	**6**	**16**	**90**	**367**	**927**

Coffea eugenioides

CPL	Chr1	Chr2	Chr3	Chr4	Chr5	Chr6	Chr7	Chr8	Chr9	Chr10	Chr11	Un	Total
C	56	30	98	17	86	40	56	106	18	10	90	88	695
Cps	14	17	33	7	34	16	18	27	5	3	29	36	239
P	3	10	13	2	9	4	8	9	1	3	16	5	83
Pps	2	3	11	1	5	3	3	6	3	4	14	6	62
Total	**75**	**60**	**155**	**27**	**134**	**63**	**85**	**148**	**27**	**20**	**149**	**135**	**1079**

B

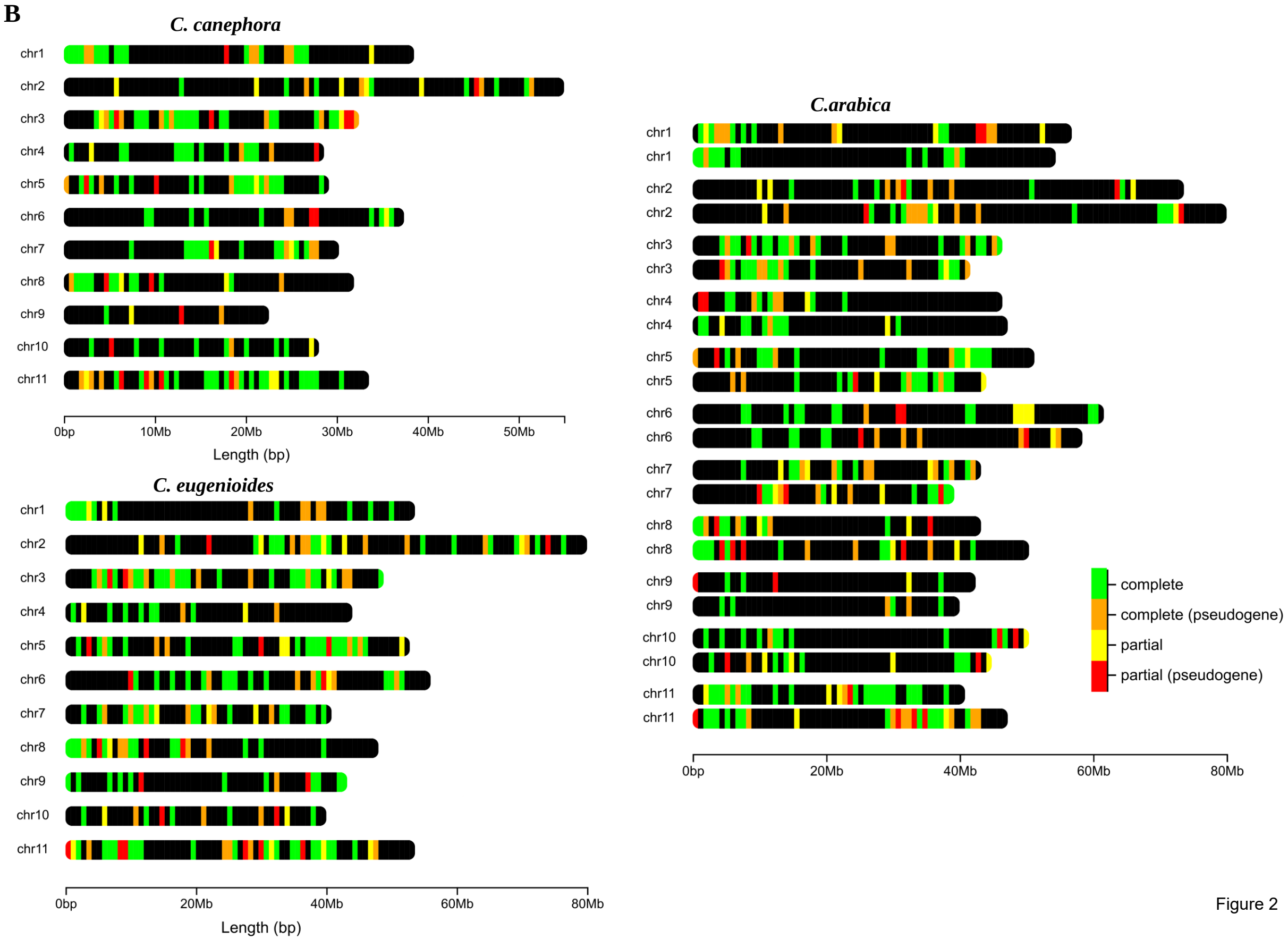


Figure 2

Figure 3

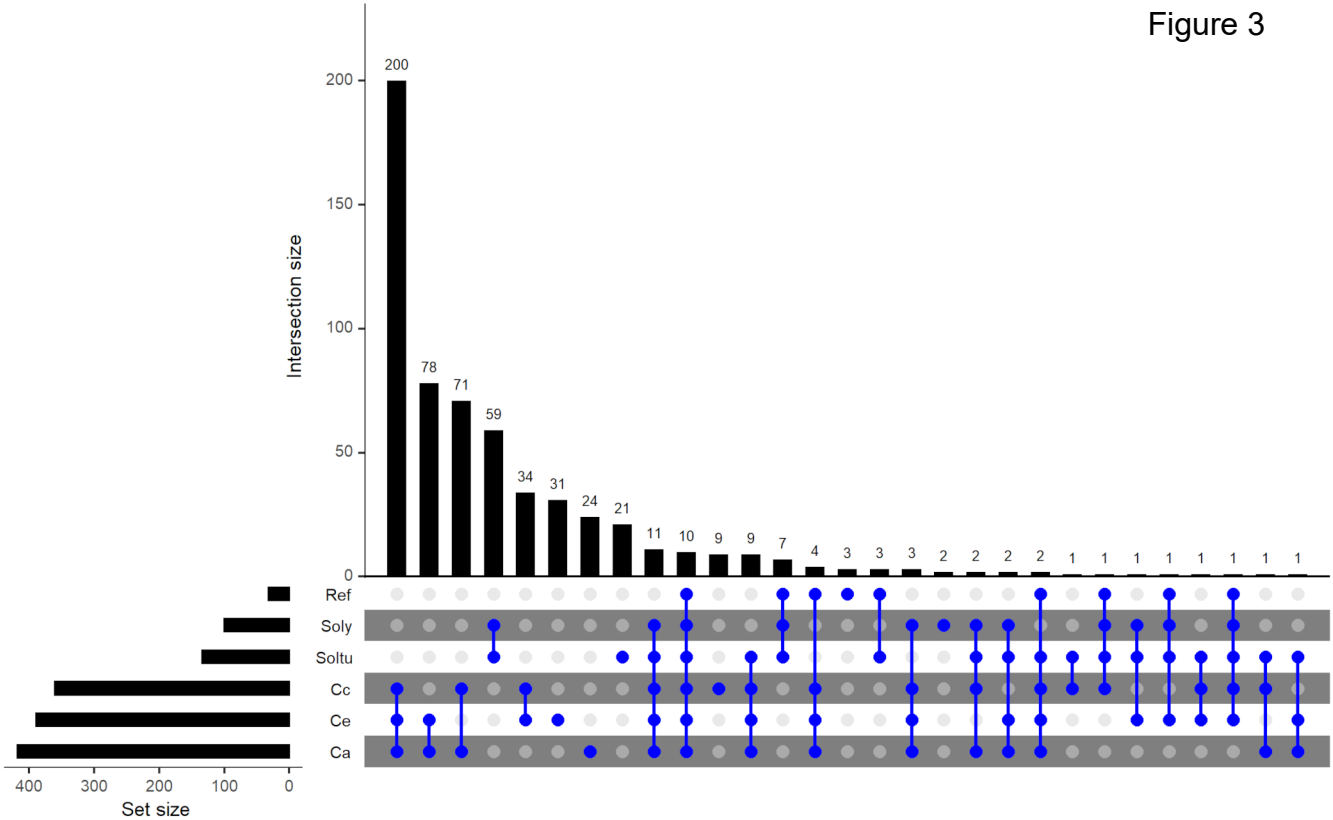

Intersection size
200
150
100
50
0
200
78
71
59
34
31
24
21
11
10
9
9
7
4
3
3
3
2
2
2
2
1
1
1
1
1
1
1
1
Ref
Soly
Soltu
Cc
Ce
Ca
400
300
200
100
0
Set size

Figure 4

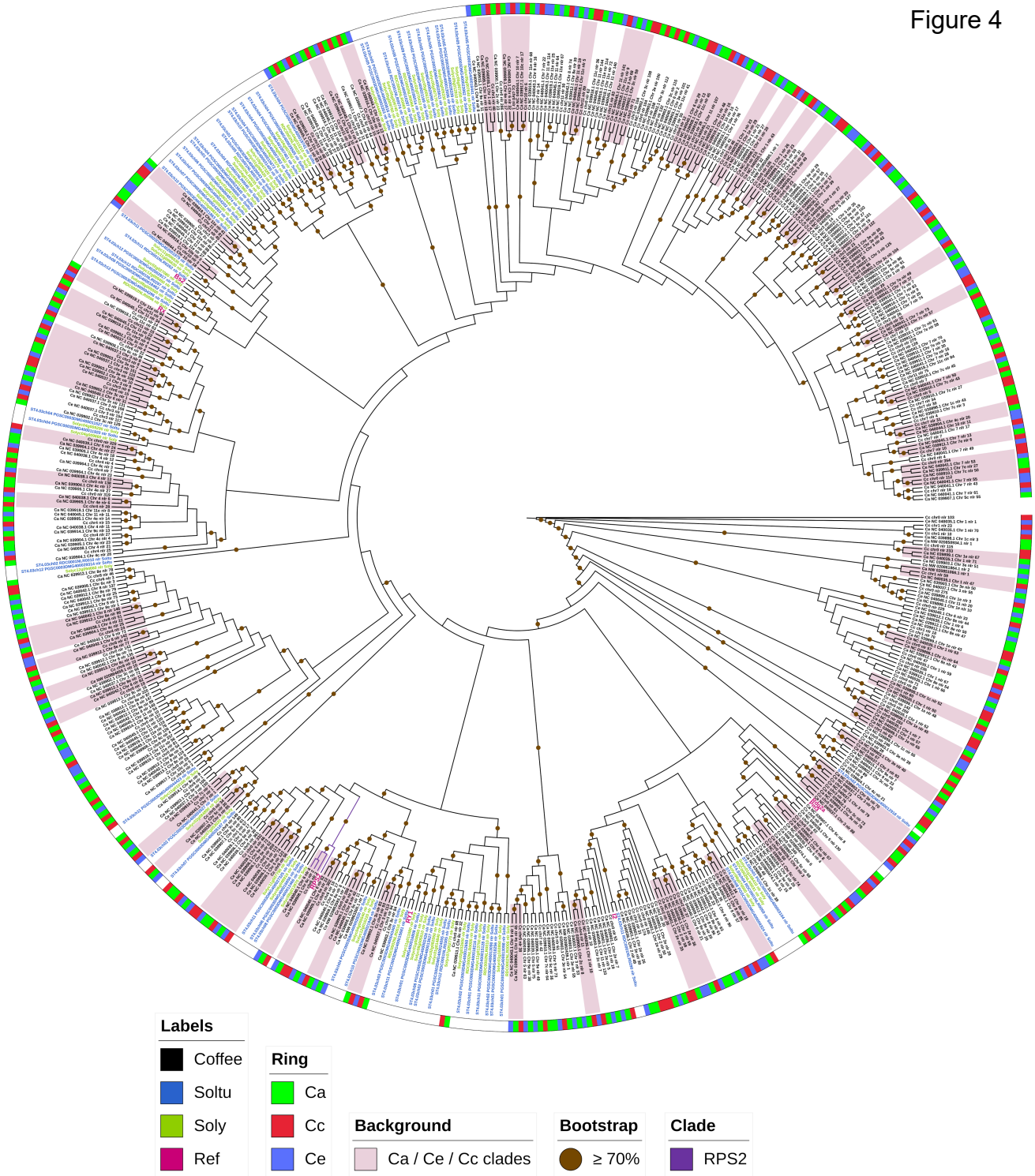
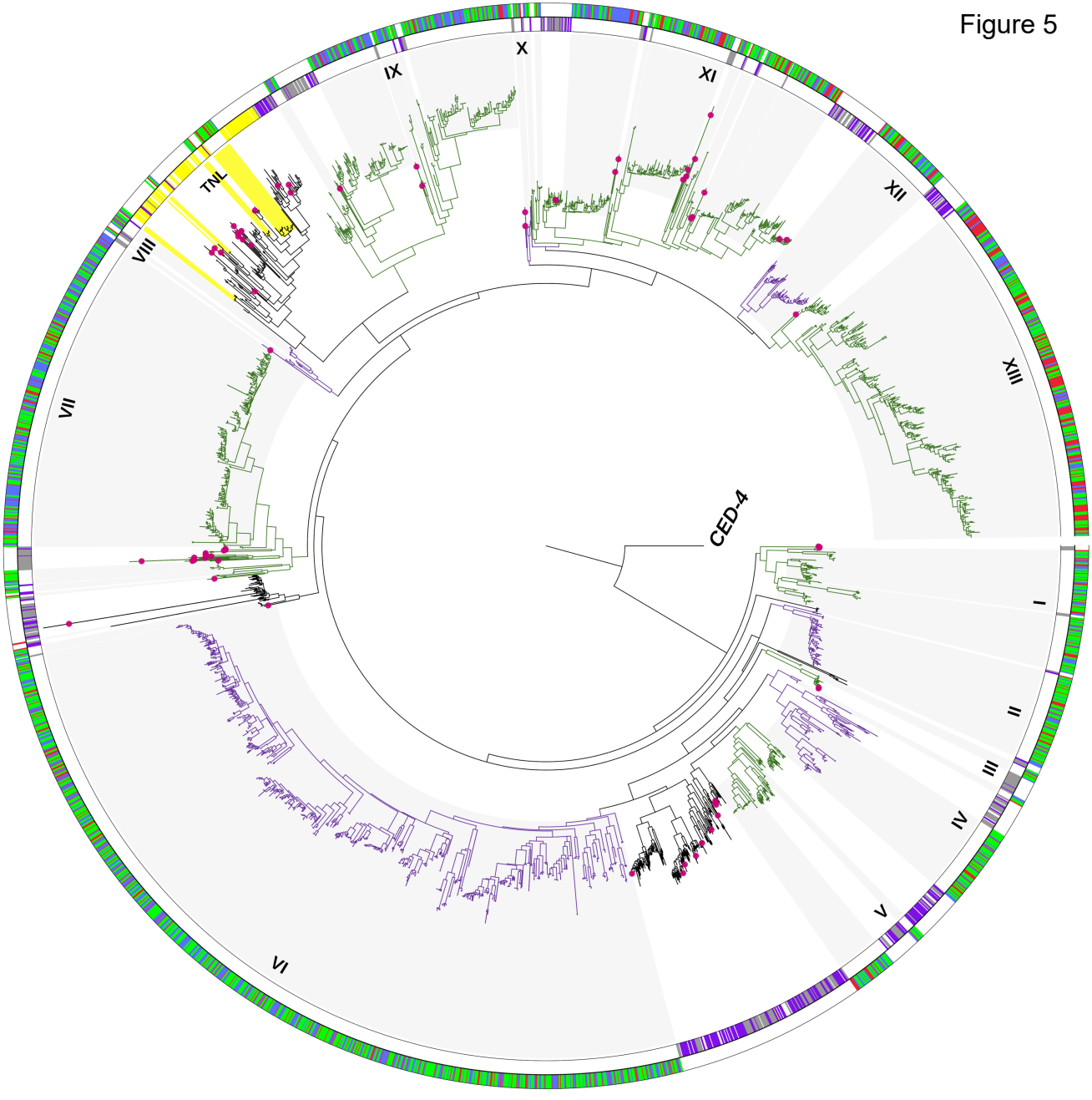

Labels
Coffee
Soltu
Soly
Ref
Ring
Ca
Cc
Ce
Background
Ca / Ce / Cc clades
Bootstrap
≥ 70%
Clade
RPS2

Figure 5



Inner ring	Outer ring	Backgrounds
TNL	Ca	Coffee NLRs - TNLs
CNL	Cc	Coffee NLRs - Non-TNLs
NL	Ce	

Reference NLRs