The Pennsylvania State University

The Graduate School

EVOLUTIONARY AND FUNCTIONAL GENETICS OF DISEASE RESISTANCE IN *THEOBROMA CACAO* AND ITS WILD RELATIVES

A Dissertation in

Ecology

by

Noah P. Winters

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2022

The dissertation of Noah P. Winters was reviewed and approved by the following:

James Marden
Professor of Biology
Dissertation Co-Advisor
Co-Chair of Committee

Mark Guiltinan
J. Franklin Styer Professor of Horticultural Botany and Plant Molecular Biology
Dissertation Co-Advisor
Co-Chair of Committee

Claude dePamphilis
Huck Chair of Plant Biology and Evolutionary Genomics

Kevin Hockett
Assistant Professor of Microbial Ecology
Lloyd Huck Early Career Professor

Jason Kaye
Distinguished Professor of Soil Biogeochemistry
Chair of the Ecology Intercollegiate Graduate Degree Program

ABSTRACT

Plants have complex and dynamic immune systems that have evolved over millennia to help them resist pathogen invasion. Humans have worked to incorporate these evolved defenses into crops through breeding. However, many crop cultivars only harness a fraction of the overall genetic diversity available to a given species, or have such a long history of domestication that most diversity has been lost. Evaluating previously neglected germplasm for desirable traits, such as disease resistance, is therefore an essential step towards breeding crops that are adapted to both current and emerging threats. In this dissertation, we examine the evolution of defense response across populations of *Theobroma cacao* L. and wild species of *Theobroma*, with the goal of identifying genetic elements essential for protection against the cacao pathogen *Phytophthora palmivora*.

In Chapter 2, we combine data from RNA-sequencing, un-targeted metabolomics, and whole genome sequencing to discover genes and pathways associated with resistance. We found significant differences in transcriptional response across populations, indicating lineage-specific defenses. Among the processes shared across populations, however, was phenylpropanoid biosynthesis, a metabolic pathway with well-documented roles in plant defense. One of the genes in this pathway, caffeoyl-shikimate esterase (CSE), was up-regulated in response to pathogen challenge across all four populations, indicating its broad importance for cacao's defense response. Further experimental evidence suggested this gene synthesizes the antimicrobial compound caffeic acid, a known inhibitor of *Phytophthora* species. Together, our results indicate most of the expression variation associated with resistance is unique to populations. Moreover, they suggest using a small subset of clones to determine the basis of resistance to *P. palmivora,* as has been done in breeding programs for over five decades, provides limited power for discovering potentially useful genetic variation.

Natural variation in resistance to *P. palmivora* is well documented across cacao lineages, but little is known about resistance in its wild, non-cacao relatives. In Chapter 3, we use non-cacao *Theobroma* species to investigate the evolution of defense response across the genus. We discovered both lineage-specific and conserved aspects of defense response, including upregulation of the phenylpropanoid pathway. Of particular interest were *TcBBE8* and *TcWRKY29*, a pair of genes that were upregulated in response to pathogen challenge across five species of *Theobroma* and displayed evidence of positive selection. These results suggest some aspects of defense against *P. palmivora* are orthologous and, therefore, of fundamental important to defense across *Theobroma.*

Nucleotide-binding leucine rich repeats receptors (NLR) are essential components of plant immunity. NLR evolution is complex and dynamic, full of rapid expansions, contractions, and polymorphism. The hundreds of high-quality plant genomes generated over the last two decades have provided substantial insight into the evolutionary dynamics of NLR genes. Despite steadily decreasing sequencing costs, the difficulty of sequencing, assembling, and annotating high-quality genomes has resulted in comparatively little genome-wide information on intraspecies NLR diversity. In Chapter 4, we investigate the evolution of NLR genes across 11 high quality genomes of cacao. We found 3-fold variation in NLR copy number across genotypes, a pattern primarily driven by the expansion of NLR clusters by tandem and proximal duplication. Together, our results suggest local duplications can radically reshape gene families over short evolutionary time scales, creating a source of NLR diversity that could be utilized to enrich our understanding of both plant-pathogen interactions and resistance breeding.

This dissertation helps advance genomic research and resistance breeding in cacao in two significant ways. First, it identifies both conserved and lineage-specific aspects of *Theobroma*'s defense against *P. palmivora,* indicating the potential value of wild germplasm to breeding programs. In doing so, it also identifies several high priority candidate genes for further

functional characterization. Second, it classifies and analyzes immune receptor complement across a diverse set of cacao accessions, generating new knowledge about intraspecific evolution of large gene families and creating a resource for future NLR experimentation.

TABLE OF CONTENTS

LIST OF FIGURES ix

LIST OF TABLES xviii

PREFACE xix

ACKNOWLEDGEMENTS xxi

Chapter 1: Literature Review 1

1.1 *Theobroma cacao*: Taxonomy, cultivation, and biology 1
Taxonomy of Theobroma cacao and related species 1
Distribution and diversity of Theobroma cacao 1
History of cacao cultivation 3
Cacao genetics and genomics 4
Wild Theobroma species as a source of beneficial traits 7
1.2 Pathogens of cacao 8
Black pod rot – Phytophthora spp. 8
Frosty pod rot – Moniliophthora rorei 9
Witches' broom – Moniliophthora perniciosa 10
Ceratocystis wilt of cacao – Ceratocystis cacaofunesta 11
Cacao swollen shoot virus 12
Resistance breeding in cacao 13
1.3. Plant-pathogen interactions 14
Induced versus constitutive defenses 14
PAMP and DAMP triggered immunity 15
Effector molecules manipulate host physiology 16
NLR recognition of effectors 17
Induced phytohormone and chemical defenses 19
1.4. Evolution of plant genomes 20
1.5 Dissertation Overview 22
References 24

Chapter 2: A Combination of Conserved and Diverged Responses Underlie *Theobroma cacao*'s Defense Response to *Phytophthora palmivora* 48

Abstract 48
Introduction 49
Results 51
Cacao genotypes and populations sampled for this study 51
Different sets of genes are responsible for defense against P. palmivora across all four populations 53
Common functional groups underlie different sets of pathogen responsive genes 57
Functional analysis of a candidate gene for caffeic acid synthesis 61
Population branch statistics identify differentially expressed genes under selection 63

Discussion 66
Materials and Methods 70
Plant propagation 70
Genotype phylogeny 71
Transcriptome experimental design and treatment 71
Sample preparation and sequencing 73
Genome meta-assembly 74
Pseudochromosome construction 76
Assembly evaluation and validation 76
Repeat library construction 77
Generation of gene annotation evidence 77
Gene prediction and functional assignment 78
Expression quantification, differential expression, and gene ontology enrichment 80
TcCSE cloning and over-expression in Nicotiana benthamiana 83
Plant metabolite extraction from selected transcriptome tissue samples 85
Phytophthora palmivora growth inhibition and zoospore preparation 86
Genome scan for selection 87
References 89

Chapter 3: A Conserved Set of Orthologous Genes are Involved in Defense Against *Phytophthora palmivora* across *Theobroma* species 103

Abstract 103
Introduction 103
Materials and Methods 105
Plant phenotyping and sample selection 105
Transcriptome experimental design 106
RNA extraction, and sequencing 107
Transcriptome assembly, mapping, and expression quantification 108
Differential expression analysis and gene ontology enrichment 110
Orthogroup classification and resistance class assignment 111
Analysis of log_2 fold change across species 111
Measures of selection 112
Results 113
Theobroma spp. displayed variation in disease resistance to Phytophthora palmivora 113
Supertranscript statistics reveal contiguous and complete transcriptome assemblies 116
Theobroma spp. displayed overlapping functional response to P. palmivora 119
Differentially expressed orthogroups are shared across Theobroma 120
Core orthogroups display consistent expression responses across Theobroma 123
Conserved orthogroups show evidence of positive selection 128
Discussion 129
References 134

Chapter 4: Local Gene Duplications Drive NLR Copy Number Variation Across Multiple Genotypes of *Theobroma cacao* 143

Abstract 143
Introduction 143
Materials and Methods 146
Genome assembly and annotation 146
Genotype phylogeny 147
NLR classification and categorization 148
Gene duplication analysis 150
Genome synteny analysis 151
Pseudogene identification 151
Transposable element analysis 152
Statistical analyses 153
Results 153
Cacao genotypes displayed a high degree of copy number variation in NLR genes 153
NLR copy number variation was not distributed evenly throughout the genome 155
High and low copy number genotypes evolved independently multiple times 157
Annotation quality varied across genotypes but did not explain NLR copy number variation 158
Variation in transposable element content did not explain NLR copy number variation 160
Tandem and proximal duplications were primarily responsible for NLR copy number variation 163
Discussion 169
References 175

Chapter 5: Retrospective 189

Outro 189
Induced defense responses across populations of *T. cacao* 190
Induced defense responses across *Theobroma* 192
Molecular evolution of cacao's immune receptors 194
Conclusion 195
References 196

Appendix A: Supplementary Figures 245

Chapter 2 245
Chapter 3 253

LIST OF FIGURES

Figure 2-1: **Overview of cacao genotypes and populations sampled for this study**. (A) Maximum likelihood phylogeny of *T. cacao* genotypes based on 23,439 SNPs. White and gray boxes beside the phylogeny indicate whether genotypes were considered resistant (grey) or susceptible (black) to *P. palmivora* according to Fister et al. 2020. Numbers on the nodes indicate bootstrap support and colors at the tips indicate population membership: Guiana (blue), Iquitos (red), Marañón (green), and Nanay (orange). (B) Map displaying approximate center of origin for each of the four populations sampled for this study. (C) Pairwise F_{ST} estimates for each population. 53

Figure 2-2: **Different sets of genes are responsible for defense against *P. palmivora* across all four populations.** (A) Overlap of differentially expressed genes for *P. palmivora* treatment versus control (top) and between resistant versus susceptible genotypes (bottom). The blue, red, green, and orange bars represent genes that are only differentially expressed in Guiana Iquitos, Marañón, or Nanay, respectively. The pink bar indicates genes that are differentially expressed across all four populations. Numbers above the bars indicate the number of differentially expressed genes in that specific intersection. (B) Pairwise Spearman correlations of $\log_2$ fold changes for all genes investigated in this study, for *P. palmivora* treatment versus control (top) and between resistant versus susceptible genotypes (bottom). The bottom triangle is the Spearman correlation coefficient. The top triangle is the correlation coefficient depicted as an ellipse, the shape of which depends on the size of the coefficient. Stars indicate statistical significance ($p < 0.001$), tested using Spearman's rho. (C) Overlap of enriched GO terms (Fisher's exact test: FDR-adjusted p-value < 0.05) for *P. palmivora* treatment versus control (top) and resistant versus susceptible genotypes (bottom). The blue, red, green, and orange bars represent GO terms that are only enriched in Guiana Iquitos, Marañón, or Nanay, respectively. The pink bar indicates GO terms that are significantly enriched across all four populations. Numbers above the bars indicate the number of enriched GO terms in that specific intersection. 57

Figure 2-3: **Common functional groups underlie different sets of pathogen responsive genes.** (A) Enriched GO terms (Fisher's exact test: FDR-adjusted p-value < 0.05) and their median fold change for *P. palmivora* treatment versus control. Colored points indicate population membership: Guiana (blue), Iquitos (red), Marañón (green), or Nanay (orange). Point size is scaled to median fold change for the differentially expressed genes belonging to that term. (B) Enriched GO terms (Fisher's exact test: FDR-adjusted p-value < 0.05) and their median fold change for resistant versus susceptible genotypes. Colored points indicate population membership: Guiana (blue), Iquitos (red), Marañón (green), or Nanay (orange). Point size is scaled to median fold change for the differentially expressed genes belonging to that term. (C) The percentage of genes from each population that are unique, calculated for each GO term that is enriched in all four populations. Terms that are significantly enriched in *P. palmivora* treatment versus control are on top, and terms that are significantly enriched in resistant versus susceptible genotypes are on bottom. Each point represents the percentage of differentially expressed genes

belonging to a single GO term (indicated by color) that are unique to each population. For instance, Guiana has 22 differentially expressed genes in GO:0009834, of which 5 of them are not differentially expressed in any other population (4/22 = 22.7%). Means are shown as open triangles.....60

Figure 2-4: ***TcCSE* is involved in resistance to *P. palmivora.*** (A) Expression of *TcCSE* (SCA-6_Chr6v1_17513) across each population for control (blue) and treatment (yellow). Open diamonds indicate mean expression for susceptible genotypes and open circles indicate mean expression for resistant genotypes. (B) Relative abundance of caffeic acid in *N. benthamiana* plants transformed with 35s:TcCSE or an empty vector control, at both 48 and 96 hours post transformation. Means are shown as open triangles. Over-expression of *TcCSE* results in significantly higher caffeic acid accumulation relative to controls (t-test 48 hpi: p-value = 0.0164; t-test 96 hpi: p-value = 0.0174). (C) Mycelial area of *P. palmivora* cultures grown on plates of V8 media versus plates of V8 media amended with 2mM caffeic acid. Means are shown as open triangles. Plates amended with 2mM caffeic acid significantly inhibited mycelial growth (t-test: p-value < 0.001). (D) Relative abundance of caffeic acid for cacao leaves mock inoculated with water, challenged with *P. palmivora* zoospores, or zoospores only. Means are shown as open triangles. Cacao leaves challenged with zoospores accumulated significantly more caffeic acid than either mock inoculated or zoospore-only controls (t-tests: p-values < 0.001). Mock inoculated leaves had significantly more caffeic acid than zoospore-only controls (t-test: p-value = 0.022). € Relative abundance of caffeic acid in samples challenged with plugs of V8 media (blue) versus plugs of *P. palmivora* mycelia (yellow). There were no significant differences between treatment, phenotype, or the treatment*phenotype interaction (one-way ANOVA, Intensity ~ Treatment + Phenotype + Treatment*Phenotype: p-values > 0.05). 62

Figure 2-5: **Population branch statistics identify differentially expressed genes under selection.** (A) Population branch statistics can estimate lineage-specific selection leading to resistant genotypes. Branch lengths represent the magnitude of allele frequency change. For loci evolving neutrally in both resistant and susceptible genotypes, differences in allele frequency between resistant and susceptible individuals of the same population (S1, R1) will be *smaller* than allele frequency differences between susceptible individuals from two separate populations (S1, S2) (top). For loci under selection in resistant genotypes, differences in allele frequency between resistant and susceptible individuals of the same population (S1, R1) will be *greater* than allele frequency differences between susceptible individuals from two separate populations (S1, S2) (bottom). High PBS scores indicate genes that are under selection among resistant individuals from a given population. (B) Overlap of genic and non-genic regions designated as selection outliers (top 1% of their respective PBS distributions). PBS was estimated for 5 Kb upstream of each gene (top), the gene body (middle), and 5 Kb downstream of each gene (bottom). The blue, red, green, and orange bars represent genes that are only designated as selection outliers in Guiana, Iquitos, Marañón, or Nanay, respectively. Numbers above the bars indicate the number of selection outliers in that specific intersection. (C) Venn diagrams displaying the overlap between differentially expressed and genes under selection in resistant genotypes. Colors indicate population membership: blue (Guiana), red (Iquitos), green (Marañón), and orange (Nanay). 65

Figure 3-1: Variation in resistance to *P. palmivora* across *Theobroma.* Seven *Theobroma spp.* were assayed for resistance to *P. palmivora* strain C-14. Each dot represents the mean (n = 3) lesion area for an individual leaf. Letters indicate significant differences in mean lesion area (Welch's ANOVA, p-value < 0.001; Games-Howell post-hoc test, adjusted p-values < 0.05). Means are shown as blue dots. Species that share letters do not have significantly different means. 115

Figure 3-2: Split-plot design for RNA-seq experiment. Trees were sampled over three consecutive days. Each day, a single tree from each species was collected and processed. From each tree, two leaves were used for treatment with *P. palmivora* (purple) and two leaves were used for controls (blue). Leaves from the same tree and treatment combination were pooled before library preparation (L) and sequencing. 116

Figure 3-3: Transcriptome assembly quality and completeness metrics. (A) Coding sequence (CDS) length distributions for all four non-cacao *Theobroma spp.* and the reference transcriptome *T. cacao* (SPEC 54/1). Letters indicate significant differences in mean CDS length (Welch's ANOVA, p-value < 0.001; Games-Howell post-hoc test, adjusted p-values < 0.01). Means are shown as blue dots. (B) Proportion of complete, fragmented, and missing BUSCO genes for each non-cacao *Theobroma spp.* and the reference transcriptome *T. cacao* (SPEC 54/1). For all comparisons, there was a significant association between species (*T. cacao* SPEC 54/1 x *Theobroma spp.*) and BUSCO completeness (chi-square goodness-of-fit, p-values < 0.05). 118

Figure 3-4: Differentially expressed genes and enriched gene ontology terms. (A) Differentially expressed supertranscripts for each species (unadjusted p-value < 0.05). White bars indicate downregulated supertranscripts and black bars indicate upregulated superstranscripts. (B) Overlap of significantly enriched GO terms (FDR-adjusted p-values < 0.05). The green, orange, purple, and blue bars represent GO terms that are only enriched in *T. angustifolium, T. bicolor, T. grandiflorum,* and *T. mammosum,* respectively. The pink bar indicates GO terms that are significantly enriched across all four species. Numbers above the bars indicate the number of GO terms in each specific intersection. 119

Figure 3-5: GO terms enriched across all four *Theobroma spp*. Boxplots display the distribution of $\log_2$ fold changes for each GO term, for each species. Each colored point represents the $\log_2$ fold change for a single differentially expressed supertranscript (unadjusted p-value < 0.05). Means are shown as blue dots. 120

Figure 3-6: Differentially expressed orthogroups and their designated resistance classes. (A) Proportion of differentially expressed orthogroups in each resistance class. Orthogroups that are differentially expressed across all four species are CORE (blue). Those differentially expressed across two or three species are SHELL (pink). And orthogroups differentially expressed in only a single species are CLOUD (purple). Orthgroups containing one or more differentially expressed supertranscript are themselves considered differentially expressed. (B) Overlap of differentially expressed orthogroups. The green, orange, purple, and blue bars represent orthogroups that are only differentially expressed in *T. angustifolium, T. bicolor, T. grandiflorum,* and *T. mammosum,* respectively. The pink bar indicates orthogroups

that are differentially expressed across all four species. Numbers above the bars indicate the number of orthogroups in each specific intersection. (C) Average number of differentially expressed supertranscripts for each orthogroup and resistance class. Each point represents the mean number of differentially expressed supertranscripts per species for a given orthogroup. Numbers indicate the mean for each class. Letters indicate significant differences between class means (Welch's ANOVA, p-value< 0.001; Games-Howell post-hoc test, adjusted p-values < 0.05). 122

Figure 3-7: **Differentially expressed orthogroups in *T. cacao* and non-cacao *Theobroma spp***. (A) Overlap between orthogroups that are differentially expressed in *T. cacao* (grey) and orthogroups belonging to each resistance class: CORE (blue), SHELL (pink), and CLOUD (purple). (B) Mean $\log_2$ fold change correlations between orthogroups differentially expressed in both *T. cacao* (Chapter 2) and each non-cacao *Theobroma spp.* Each point represents the mean $\log_2$ fold change for a single orthogroup. Blue points are CORE orthogroups whose mean |LFC| < 1 in *T. cacao,* non-cacao *Theobroma spp.*, or both. Red points are CORE orthogroups whose mean |LFC| > 1 in both *T. cacao* and non-cacao *Theobroma spp.* Gray points are not in the CORE resistance class.....124

Figure 3-8: Gene family phylogeny for orthogroup 60, FAD-binding berberine bridge enzymes. Sequence IDs are colored according to their lineage: basal angiosperm (blue), basal eudicot (purple), monocot (orange), rosid (red), and asterid (green). All *Theobroma* species, including *T. cacao*, are shown in black. *T. cacao* sequences are from the SCA-6 genome. Node values indicate SH-like local supports calculated by FastTree. SH supports > 80 are not shown. Bars in the right panel indicate $\log_2$ fold changes. The box indicates the clade containing TcBBE8 (SCA-6_Chr6v1_16921) and its close orthologs across *Theobroma*. 126

Figure 3-9: Gene family phylogeny for orthogroup 361, WRKY transcription factors. Sequence IDs are colored according to their lineage: basal angiosperm (blue), basal eudicot (purple), monocot (orange), rosid (red), and asterid (green). All *Theobroma* species, including *T. cacao*, are shown in black. *T. cacao* sequences are from the SCA-6 genome. Node values indicate SH-like local supports calculated by FastTree. SH supports > 80 are not shown. Bars in the right panel indicate $\log_2$ fold changes. Boxes indicate the clades containing TcWRKY29 (SCA-6_Chr3v1_10161), TcWRKY22 (SCA-6_Chr1v1_03377), and TcWRKY69 (SCA-6_Chr6v1_18337), as well as their close orthologs across *Theobroma.* 127

Figure 3-10: **Orthogroups with signatures of positive selection.** (A) Proportion of orthogroups that have signatures of episodic, diversifying selection. CORE orthogroups with |LFC| > 1 in both *T. cacao* and non-cacao *Theobroma spp.* (top) were compared to an equal number (n = 48) of orthogroups drawn at random (bottom). Positive selection was significantly associated with orthogroup type (CORE vs random) (chi-sq. goodness-of-fit, p-value < 0.001). (B) Distribution of bootstrap replicates for both CORE orthogroups with |LFC| > 1 in both *T. cacao* and non-cacao *Theobroma spp.* (red), and orthogroups drawn at random (grey). The proportion of orthogroups displaying signatures of selection was significantly higher for CORE orthogroups with |LFC| > 1 in both *T. cacao* and non-cacao *Theobroma spp.* than for orthogroups drawn at random (t-test, p-value < 0.001). (C) The

proportion of each gene under selection. CORE orthogroups with |LFC| > 1 in both *T. cacao* and non-cacao *Theobroma spp.* (red) were compared to an equal number of orthogroups drawn at random (grey). Differences were not significant (t-test, p-value > 0.05). 129

Figure 4-1: NLR architecture and copy number across cacao genomes. (A) The four canonical NLR architectures. Proteins containing a TIR domain (PF01582 and PF13676) were categorized as TNL. Those containing an RPW8 domain (PF05659) were categorized as RNL. Proteins containing an Rx N-terminal domain (PF18052), or an NLR-parser CC annotation and a CC annotation from COILS were categorized as CNL. Proteins containing an NB-ARC domain and a leucine-rich repeat (LRR) domain, but no other domains, were categorized as NL. (B) NLR copy number across all classes and genotypes. NL, CNL, TNL, and RNLs are shown as blue, yellow, teal, and black, respectively. Each point represents the number of NLR copies for a particular genotype. Means are represented by diamonds. Lines represent 95% confidence intervals. High CNV genotypes had significantly more NLR genes than low CNV genotypes (mean difference = 225.11, Mann-Whitney test: p-value < 0.01). Other than the RNL class, differences in mean NLR number between Low CNV and High CNV genotypes were significant for all classes (negative binomial GLM: NLR # ~ CNV Group + NLR Class + CNV Group * NLR Class, adjusted p-values < 0.01). 155

Figure 4-2: Distribution of NLR genes across each genome. (A-B) Number of NLR genes or NLR pseudogenes on each chromosome. Orange depicts High CNV genotypes and purple depicts Low CNV genotypes. Each point represents the number of NLR genes or NLR pseudogenes for a particular genotype. NLR genes or NLR pseudogenes on Chr0 do not belong to one of the 10 chromosome-oriented scaffolds. Means are represented by diamonds. Lines represent 95% confidence intervals. 157

Figure 4-3: Phylogeny of cacao genotypes sampled for this study. Phylogenetic tree of the 11 cacao genotypes used in this study, constructed using 1,364 single copy genes. Four non-cacao species of *Theobroma* were additionally used as outgroups. Numbers on each node represent posterior support values calculated by ASTRAL. With the exception of CCN-51 and ICS-1, both of which are hybrids, tip colors indicate population membership. Non-cacao *Theobroma spp.* are shown in grey. CNV class (High, Low, or No Information) of each genotype is shown in orange, purple, of white, respectively. Disease phenotypes are shown for witches' broom disease (WBD), frosty pod rot (FPR), Ceratocystis wilt of cacao (CWC) and black pod rot (BPR). Blue indicates resistant, red indicates susceptible, and white indicates no information was available. 158

Figure 4-4: Genome annotation quality metrics. (A) BUSCO completeness for each genome used in this study, separated by Low CNV (left) and High CNV (right). The proportion of complete, fragmented, and missing BUSCOs are shown in green, orange, and beige, respectively. Differences in the mean proportion of complete, fragmented, and missing genes between Low CNV and High CNV genotypes were significant (one-way ANOVA: Proportion ~ CNV Group + BUSCO Class + CNV Group * BUSCO Class, p-value < 0.001; Tukey's HSD, adjusted p-value < 0.01).

(B) Distribution of AED scores for each genotype's classified NLR genes. Mean AED score was not significantly different between Low CNV and High CNV groups (mean difference = 0.018, t-test: p-value < 0.001). (C) The total number of genes annotated in each of the 11 genomes used in this study, separated by Low CNV (left) and High CNV (right). NLR genes are shown in black and non-NLR genes are shown in white. There was no significant difference in gene number between Low CNV and High CNV genotypes (mean difference = 1080.89 genes, Mann-Whitney test: p-value > 0.05). 160

Figure 4-5: Transposable element abundance for High and Low CNV genotypes. Abundance of the five most common transposable elements in cacao genomes. Orange depicts High CNV genotypes and purple depicts Low CNV genotypes. Each point represents the number of transposable elements for a particular genotype. Means are represented by diamonds. Lines represent 95% confidence intervals. Differences in mean TE abundance between Low and High CNV genotypes were not significant (negative binomial GLM: # TE ~ CNV + TE Class + CNV*TE Class, adjusted p-values > 0.05). 161

Figure 4-6: Distribution of transposable elements across each genome. Density of the five most common transposable elements across each cacao chromosome. Orange depicts High CNV genotypes and purple depicts Low CNV genotypes. Each point represents the number TEs for a particular genotype. Means are represented by diamonds. Lines represent 95% confidence intervals. Stars indicate significant differences in mean TE abundance between chromosomes (negative binomial GLM: # TE ~ Chrom + TE Class + Chrom*TE Class, adjusted p-values < 0.05). Differences in mean TE abundance between Low and High CNV genotypes on each chromosome were not significant (negative binomial GLM: # TE ~ CNV + Chrom + TE Class + Chrom*TE Class*CNV, adjusted p-values > 0.05)..... 163

Figure 4-7: Types of gene duplication. Duplicate genes were classified into one of five categories: singletons, dispersed, proximal, tandem, or WGD/segmental duplicates. All NLRs were first classified as singletons, i.e. genes with no history of recent duplication (A). If NLR genes contained significant BLAST hits elsewhere in the genome, they were reclassified as dispersed duplicates (B). Dispersed duplicates were then further categorized as proximal or tandem based on distance between hits. If < 20 genes separated the NLR duplicates, they were considered proximal (C). If NLR duplicates were immediately adjacent to one another, they were considered tandem (D). Lastly, NLR duplicates that were anchors of collinear blocks, as defined by MCScanX's algorithm, were classified as WGD/segmental duplicates (E). 164

Figure 4-8: Patterns of NLR duplication across each genome. (A) The proportion of NLR (black) and non-NLR (grey) genes in each duplication class. Each point represents the proportion of NLR or non-NLR genes for a particular genotype. Means are represented by diamonds. Lines represent 95% confidence intervals. All differences in mean proportion between NLR and non-NLR genes were significant (one-way ANOVA: Proportion ~ Gene Type + Duplicate Type + Gene Type * Duplicate Type, p-value < 0.001; Tukey's HSD, adjusted p-values < 0.001). (B) The number of NLR genes belonging to each duplication class, for both Low CNV (purple) and High CNV (orange) genotypes. Points represent the number of NLR

genes for a particular genotype. Means are represented by diamonds. Lines represent 95% confidence intervals. All differences in mean NLR number between Low CNV and High CNV groups were significant (negative binomial GLM: # NLR Duplicates ~ Duplicate Type + CNV Group + Duplicate Type * CNV Group, adjusted p-values < 0.001). 165

Figure 4-9: NLR duplications across domain architectures. NL, CNL, TNL, and RNLs are shown as blue, yellow, teal, and black, respectively. Each point represents the number of NLR copies for a particular genotype. Means are represented by diamonds. Lines represent 95% confidence intervals. 166

Figure 4-10: Number, size, and location of NLR clusters. (A) The genomic distribution of NLRs in each duplicate type, for Low CNV (purple) and High CNV (orange) genotypes. Each point represents the number of NLRs for a particular genotype. Boxes outline the four chromosomes with the highest NLR density. Means are represented by diamonds. Lines represent 95% confidence intervals. (B) Number and size of NLR clusters for each genotype, for Low CNV (purple) and High CNV (orange) genotypes. Each point represents a single NLR cluster. Mean cluster size for each genotype is represented by a diamond. Lines represent 95% confidence intervals. Boxed values indicate the number of NLR clusters for each genotype. Differences in both mean cluster number and mean cluster size between Low CNV and High CNV genotypes were significant (cluster number: mean difference = 6.32, Mann-Whitney test, p-value < 0.01; cluster size: mean difference = 11.48, Mann-Whitney test, p-value < 0.05). 168

Figure 4-11. Synteny of an NLR cluster expanded through local duplications. Tandem and proximal duplications drove the expansion of NLR copy number in this homologous region of chromosome 7. NLR genes and NLR pseudogenes are shown as orange and blue bars, respectively. The phylogeny on the left indicates evolutionary relationships between the 11 genotypes used for this study. Labels on the right side indicate whether a genotype is in the Low or High CNV group. 169

Supplemental Figure S2-1: Proportion of genes that are unique to each population for various sized subsamples, ranging from 200 to 2000 genes, for *P. palmivora* treatment (left) or R/S phenotype (right). Genes were either ranked by $|\log_2$ fold change| before subsampling (blue), or subsampled at random (orange). Each dot represents one of four populations sampled. Means are represented as crosses. For every sample size, the proportion of genes unique to each population was significantly higher when the genes were drawn at random (one-way ANOVA, Proportion Unique Genes ~ Sample Size + Subsample Method + Sample Size:Sample Method: p-values < 0.001). 245

Supplemental Figure S2-2: Overlap of differentially expressed paralogs (i.e. paralogous genes with ≥ 95% identity).The blue, red, green, and orange bars represent GO terms that are only enriched in Guiana Iquitos, Marañón, or Nanay, respectively. The pink bar indicates orthogroups that are significantly enriched across all four populations. Numbers above the bars indicate the number of orthogroups in that specific intersection. 246

Supplemental Figure S2-3: Overlap of differentially expressed orthogroups (i.e. orthogroups containing 1 or more differentially expressed genes).The blue, red, green, and orange bars represent GO terms that are only enriched in Guiana Iquitos, Marañón, or Nanay, respectively. The pink bar indicates orthogroups that are significantly enriched across all four populations. Numbers above the bars indicate the number of orthogroups in that specific intersection. 247

Supplemental Figure S2-4: Pairwise Spearman correlations of mean log_2 fold changes for all orthogroups included in this study. All genes were first classified into orthogroups, then mean log_2 fold change for each orthogroup and population were then calculated. The bottom triangle is the Spearman correlation coefficient. The top triangle is the correlation coefficient depicted as an ellipse, the shape of which depends on the size of the coefficient. Stars indicate statistical significance ($p < 0.001$), tested using Spearman's rho. 248

Supplemental Figure S2-5. Pairwise Spearman correlations of log_2 fold changes for 1:1 orthologs between *A. thaliana* and its close relatives (left) and between accessions of *A. thaliana* (right). The bottom triangle is the Spearman correlation coefficient. The top triangle is the correlation coefficient depicted as an ellipse, the shape of which depends on the size of the coefficient. Stars indicate statistical significance ($p < 0.001$), tested using Spearman's rho. Data are from Winkelmüller et al., 2021, *The Plant Cell.*.....249

Supplemental Figure S2-6. Expression of differentially expressed genes that are either unique to a single population (red) or shared across populations, for *P. palmivora* treatment (left) or R/S phenotype (right). Asterisks indicate statistical significance. For treatment, the genes unique to Guiana and Marañón had significantly higher expression than the genes shared among populations (one-way ANOVA, p -value < 2e-16; Tukey's HSD, FDR-adjusted p-value < 0.001). And for phenotype, the genes unique to Guiana, Marañón, and Nanay had significantly higher expression (one-way ANOVA, p -value < 2e-16; Tukey's HSD, FDR-adjusted p-value < 0.001). 250

Supplemental Figure S2-7. Environmental covariates included in the GLM used for differential expression. (Left) tray position for each plant in the greenhouse, corresponding to supplemental figure #. (Right) developmental stage of the leaves sampled for the transcriptome experiment. 251

Supplemental Figure S2-8. Two tables were aligned parallel to one another. On each bench, there were 3 trays, with approx. 30 plants on each tray. To minimize the effect of gradients in temperature, humidity, and light within the greenhouse, we kept the distance between tables to < 2 ft. We treated the plants in each tray with either pathogen or V8 control, such that parallel trays never experienced the same treatment. We randomized the placement of plants in each tray, with the caveat that the same genotype was in a mirrored position on both tables. Thus for each pair of plants within a genotype, one would receive pathogen treatment and one would receive control treatment. If there was an odd number of plants for a given genotype, or if a genotype only had one representative plant, the odd-numbered individual would be paired with an individual within the same population and resistance/susceptibility class. Lastly, if a genotype within the same population and

resistance/susceptibility class was unavailable, we used a genotype in the same resistance/susceptibility class from a different population. Color indicates population membership. 251

Supplemental Figure S2-9. Distribution of biological replicates for each genotype included in the transcriptome experiment. Color indicates population membership: Guiana (blue), Iquitos (red), Marañón (green), and Nanay (orange). (R) indicates resistant genotypes and (S) indicates susceptible genotypes. 252

LIST OF TABLES

Table 3-1: All species within the genus *Theobroma* and their assigned section (Cuatrecasas, 1964). Adapted from Zhang et al. 2011. 113

Table 3-2: Transcriptome assembly statistics and BUSCO scores for each *Theobroma spp.* sampled. 116

Table S3-1: Orthogroups that are differentially expressed in response to *P. palmivora* challenge in both wild *Theobroma spp.* and *Theobroma cacao.* 253

PREFACE

Chapter 1 was written by Noah Winters.

For Chapter 2, the RNA-seq experiment was designed by Noah Winters, Melanie Perryman, Dr. Naomi Altman, Dr. Peter Tiffin, Dr. Siela Maximova, Dr. Claude dePamphilis, Dr. Mark Guiltinan, Dr. Jim Marden, Paula Ralph, Dr. Eric Wafula, and Dr. Tuomas Hämälä. Noah Winters and Melanie Perryman performed the experiment. Genome assembly and annotation, including RNA and DNA extractions, were completed by Dr. Eric Wafula, Dr. Prakash Timilsina, Paula Ralph, and Dr. Sarah Prewitt. Analysis of RNA-seq and genomic data, including differential expression and genomic scans for selection, were performed by Noah Winters and Dr. Tuomas Hämälä, with guidance from Dr. Jim Marden and Dr. Naomi Altman. Candidate gene experimentation, including LC-MS/MS extractions, gene cloning, heterologous over-expression, and growth inhibition experiments, were performed by Noah Winters and Dr. Ben Knollenberg. Writing and editing were performed by Noah Winters, Dr. Eric Wafula, Dr. Prakash Timilsina, Dr. Sarah Prewitt, Dr. Ben Knollenberg, Dr. Tuomas Hämälä, Dr. Jim Marden, and Dr. Mark Guiltinan.

For Chapter 3, the RNA-seq experiment was designed by Noah Winters and Naomi Altman. Plant phenotyping experiments were completed by Noah Winters, with assistance from Lara Waldt, Nicholas Moreno, Allan Mata Quirós, and Dr. Mariela Leandro-Muñoz. The RNA-seq experiment, including tissue collection and RNA extraction, was performed by Noah Winters. All analyses were performed by Noah Winters, with guidance from Dr. Eric Wafula, Dr. Jim Marden, Dr. Mark Guiltinan, and Dr. Claude dePamphilis. All writing was completed by Noah Winters, incorporating revisions suggested by Dr. Mark Guiltinan and Dr. Jim Marden.

For Chapter 4, all data analysis, including NLR identification, gene duplicate classification, and pseudogene annotation, was completed by Noah Winters, with guidance from

Dr. Eric Wafula, Dr. Claude dePamphilis, Dr. Mark Guiltinan, and Dr. Jim Marden. Genome assembly and annotation were completed by Dr. Eric Wafula. Transposable element annotations were completed by Dr. Prakash Timilsina. All writing was completed by Noah Winters, incorporating revisions suggested by Dr. Mark Guiltinan and Dr. Jim Marden.

Chapter 5 was written by Noah Winters.

ACKNOWLEDGEMENTS

"I believe that the community - in the fullest sense: a place and all its creatures - is the smallest unit of health and that to speak of the health of an isolated individual is a contradiction in terms."

--Wendell Berry, Health is Membership

Rarely are you given the time and space to contemplate and give thanks to the communities that formed you. One such occasion was our wedding, where, for the first time, 25 years of mine and Haley's lives came together in a single place. We found this experience transformative. And, while a pale comparison, I would like to bring these communities together again here.

I grew up in community, which is to say I grew up surrounded by adults that loved me and were invested in my future. Much of this was in the context of church. For me and my siblings, the ACRC congregation provided a deep sense of comfort and security. The peace I found there is, to this day, the most durable aspect of my faith, the pearl I grasp most closely when I am overwhelmed by feelings of doubt and despair. Thank you for that gift.

To my committee members, particularly my advisors Jim and Mark, thank you for giving me this opportunity. The autonomy you gave me to pursue my curiosities, even in the face of grant requirements and dissertation deadlines, helped form me into the (slightly less inept) scientist I am today. Your measured guidance and consideration for my well-being will be sorely missed.

To my lab mates, thank you for accommodating my naïve, and at time reckless, ideas and behaviors. Your patience with me was, and is, a blessing. Lena, thank you for not letting me fall through the cracks. This process would have been impossible without you constantly advocating on my behalf.

To the educators that showed me kindness despite my ineptitude, particularly Dr. Edwards, Dr. Gillen, and Mrs. Hershberger, thank you.

To the friends I made in State College, you have made these last six years the most joyous of my life.

To chat, you are simultaneously the funniest and most infuriating people I know. Thank you for lightening my heart.

To my family, in every meaningful sense you have carried me through this process. As my greatest cheerleaders and deepest sources of support, you have sacrificed for me in ways I do not fully understand. Haley, thank you for sustaining me when I so desperately wanted to give up. Your enthusiasm for life brings profound joy to me and everyone else around you. Thank you for being my partner.

Just as Wendell Berry describes in Health is Membership, I am whole because I belong not only to myself, but to you, to these communities. To you all I dedicate this dissertation.

Chapter 1: Literature Review

1.1 *Theobroma cacao*: Taxonomy, cultivation, and biology

Taxonomy of Theobroma cacao and related species

Theobroma cacao L., the tree from which chocolate is derived, is a tropical understory plant native to the Amazon basin and southern Mexico (D. Zhang and Motilal 2016). There are 22 species in the genus *Theobroma,* cacao being the most economically important (D. Zhang et al. 2011; Cuatrecasas 1964). The only other widely cultivated species is *T. grandiflorum*, otherwise known as cupuaçu. Each species of *Theobroma* belongs to one of six sections primarily based on the structure of their pods and flowers, as well as vegetative characteristics such as germination, growth type, leaf hairs, and branching patterns (D. Zhang et al. 2011; Cuatrecasas 1964). The sections are *Andropetalum, Glossopetalum, Oreanthes, Rhytidocarpus, Telmatocarpus,* and *Theobroma* (Chapter 3). Cacao belongs to the section *Theobroma* and is considered to have the most apomorphic features, including glabrous leaves and partially woody pericarp (D. Zhang et al. 2011; Cuatrecasas 1964). *Theobroma* belongs to the mallow family (Malvaceae), a group of flowering plants containing 244 genera and 4,225 species. Many of these species are also of economic importance, including cotton (*Gossypium hirsutum*), okra (*Abelmoschus esculentus*), jute (*Corchorus olitorius*), and durian (*Durio zibethinus*) (Christenhusz and Byng 2016).

Distribution and diversity of Theobroma cacao

There are currently 11 recognized cacao genetic groups, hereafter populations, all of which are named after their geographic center of origin or the population's most traditional

cultivar (Juan C. Motamayor et al. 2008; D. Zhang et al. 2012). The populations are as follows: Amelonado, Guiana, Iquitos, Cacao Nacional Boliviano (CNB), Contamana, Criollo, Curaray, Marañon, Nacional, Nanay, and Purús.

Cacao populations are genetically differentiated, with pairwise F_{ST} estimates between 0.2 and 0.4 (Hartl and Clark 2007; Cornejo et al. 2018). Despite genetic and morphological differences between populations, all known genotypes can be crossed with one another (Bartley 2005). The high degree of differentiation across populations is likely due to a number of factors, including cacao's ecology and evolutionary history. Cacao is pollinated by Ceratopogonid midges (Arnold et al. 2019) and has large fruits that are difficult to disperse over long distances. Together these factors lead to limited gene flow between populations, facilitating local adaptation and further differentiation (J. C. Motamayor et al. 2002). Populations of cacao are hypothesized to have formed via several mechanisms. The first hypothesis contends that populations were widespread throughout the Western Amazon prior to the most recent glacial period 22,000-13,000 BP (Thomas et al. 2012). However, the dramatic changes in temperature and precipitation brought about by glaciation altered the distribution of many Amazonian plant species, including cacao, limiting them to geographically isolated glacial refugia (J. C. Motamayor et al. 2002). This, coupled with cacao's specific pollinators and limited dispersal, resulted in the formation of distinct populations via an isolation-by-distance like process (D. Zhang et al. 2006, 2012). The second hypothesis posits that human mediated movement led to long-range dispersal events. The bottlenecks that accompany dispersal, followed by potential domestication, led to the formation of genetically differentiated populations (J. C. Motamayor et al. 2002; Juan C. Motamayor et al. 2008). The greatest evidence for this is found in the Criollo group. In one study using RFLP and microsatellite loci, Criollo cacao from Central America was found to be more homozygous and less genetically diverse than individuals collected from the Upper Amazon, evidence of a genetic bottleneck that could have been caused by human-mediated dispersal and domestication (J. C.

Motamayor et al. 2002). Another study using whole genome sequencing and a more diverse set of samples, found strong signatures of domestication in Criollo dating back nearly 4,000 years (Cornejo et al. 2018). The third hypothesis is that ancient ridgelines, called paleoarches, created dispersal barriers that separated populations (Hubert et al. 2007; Juan C. Motamayor et al. 2008). The distribution of cacao populations appears to match these paleoarches, but there is considerable ambiguity regarding their location (Wesselingh and Salo 2006; Juan C. Motamayor et al. 2008).

History of cacao cultivation

While genomic evidence suggests cacao domestication began in Mesoamerica approximately 3,600 years ago, multiple lines of archaeological evidence support its use in the Upper Amazon as far back as 5,300 years ago (Zarrillo et al. 2018). Most often consumed as a drink, chocolate served an important ceremonial role in the courts of Mesoamerican empires like the Olmecs, Aztecs, and Mayans (Dillinger et al. 2000). Following exportation by Spanish explorers, chocolate became similarly important in post-Columbian Europe (D. Zhang and Motilal 2016). By the mid-17th century cacao was grown across Latin America, Southeast Asia, Africa, and the Caribbean (D. Zhang and Motilal 2016). It was also during this period that we begin seeing various diseases, called “blights” or “blasts”, that devastated cacao cultivation, continuing to this day. For example, sometime in the 1630s, a blight (“alhorra”) resulted in the destruction of more than 50% of Venezuela’s cacao trees (Ferry 2020). Between 1727 and 1732, more than 95% of Martinique’s cacao trees were lost to either a root rot fungus or a leaf-eating caterpillar (Kimber 2006). These disease outbreaks, as well as extreme weather events like hurricanes and droughts (W. H. Johnson 1912; Shephard 2018), were powerful determinants of where cacao was grown and what varieties were used (D. Zhang and Motilal 2016). This fact

remains true into the modern era. For instance, prior to 1990 Brazil was the 3rd largest producer of cacao in the world (Meinhardt et al. 2008). However, following an outbreak of the fungal disease *Moniliophthora perniciosa,* commonly referred to as witches' broom, it became a net importer of cacao (Meinhardt et al. 2008). Today, cacao production occurs in growing regions around the world, but is predominantly centered in West Africa (Ghana and Côte d'Ivoire) and Southeast Asia (Indonesia).

Cacao genetics and genomics

The first cacao genome, Criollo B97-61/B2 (hereafter Criollo), was sequenced in 2011 (Xavier Argout et al. 2011), with another, Matina 1-6 (hereafter Matina), sequenced shortly thereafter in 2013 (Juan C. Motamayor et al. 2013). Both studies estimated cacao's genome size to be approximately 450 Mbp, almost 4 times smaller than its close relative okra (*Abelmoschus esculentus,* 1,666 Mbp) and approximately 2.5 times larger than *Arabidopsis thaliana* (156 Mbp) (Pellicer and Leitch 2020). These early cacao genomes revealed 10 chromosomes ($2n = 20$), the origin of which was hypothesized to be 11 chromosome fusions. Moreover, both studies found variation in genes related to flavonoid and anthocyanin biosynthesis. The Criollo genome revealed the expansion of flavonoid-related biosynthetic genes relative to *Arabidopsis thaliana.* Likewise, the Matina genome revealed *TcMYB113* has the same function as homologous genes in Rosaceae and Brassicaceae, namely the control of fruit pigmentation through the regulation of anthocyanin biosynthesis (Gonzalez et al. 2008). Lastly, these genomes also revealed clusters of both nucleotide-binding leucine rich repeat receptors (NLR) and leucine rich receptor-like kinases (LRK), critical components of pathogen recognition and defense.

Both genomes have hastened the functional characterization of many cacao genes involved in a diverse set of processes, including anthocyanin biosynthesis, somatic

embryogenesis, defense response, and flowering time. For instance, anthocyanidin synthase, anthocyanidin reductase, and leucoanthocyanidin reductase, were all shown to be functional orthologs of their *A. thaliana* counterparts (Y. Liu et al. 2013). The cacao homolog of the gene *BABY BOOM,* an important regulator of plant totipotency, was shown to induce plant somatic embryogenesis (Jha and Kumar 2018; Florez et al. 2015). Homologs of defense-related transcription factors *NON-EXPRESSIOR OF PR GENES 1/3* (*TcNPR1* and *TcNPR3*) were shown to be important regulators of cacao defense. *TcNPR1* partially complemented an *A. thaliana npr1* mutant, decreasing susceptibility to *Pseudomonas syringae* pv. tomato DC3000 (Shi et al. 2010). Over-expression of this same gene in cacao also resulted in decreased susceptibility to *Phytophthora tropicalis* (Fister et al. 2015). Likewise, *TcNPR3* has been shown to negatively regulate *TcNPR1*, resulting in increased susceptibility. Transient knockdowns of *TcNPR3,* either through miRNA or CRISPR-Cas9 genome editing, have shown increased resistance to *P. tropicalis* (Shi et al. 2013; Fister et al. 2018). Lastly, cacao homologs of the flowering time regulator *FLOWERING LOCUS T* (*FT*) partially rescued *A. thaliana ft-10* mutants. Cacao tissue transformed with *AtFT* resulted in precocious flowering in tissue culture.

The Criollo and Matina genomes have also given us insight into the evolution of defense-related genes and their activation during pathogen challenge. For instance, the Matina genome, along with 70 other land plant genomes, helped identify a strong association between the duplication of NLR genes and the generation of novel plant miRNAs to regulate them (Y. Zhang et al. 2016). The Matina genome was also used to discover unique, non-canonical NLR domain architectures across multiple land plant families (Sarris et al. 2016). The Criollo genome helped resolve the history of whole genome duplication events in eudicot species (F. Li et al. 2015), and revealed cacao has not experienced any whole genome duplications since the gamma duplication prior to the separation of core and basal eudicots (Jiao et al. 2011). Many genomic and transcriptomic studies have discovered large scale differential expression of defense-associated

gene families in response to a variety of pathogens, including *Phyopthora palmivora, Phytophthora megakarya*, *Colletotrichum theobromicola,* and *Moniliophthora perniciosa* (Fister, Mejia, et al. 2016; Teixeira et al. 2014; Shahin S. Ali, Shao, Lary, Strem, et al. 2017; D. N. Pokou et al. 2019). These gene families include pathogenesis-related (PR) genes, NLR receptors, and WRKY transcription factors. Multiple studies have also highlighted the role of metabolism in cacao's defense response. Ali et al. found secondary metabolite pathways associated with the production of phenylpropanoid and terpenoid compounds were upregulated in cacao challenged by *Phytophthora palmivora* and *Phytophthora megakarya* (Shahin S. Ali, Shao, Lary, Strem, et al. 2017). Functional studies examining the role of phenylpropanoid compounds in cacao have found them to be inhibitory to pathogen growth (Knollenberg et al. 2020; Widmer and Laurent 2006).

Recently, a series of studies have extended our understanding of cacao intraspecies diversity and local adaptation using a combination of genomics and transcriptomics. In the first study, a collection of low coverage whole genome sequencing data from a diverse set of 200 cacao genomes revealed strong domestication ~3,600 years ago. Most of this domestication was detected in signatures of selection from metabolic genes involved in anthocyanin biosynthesis and disease resistance genes. This domestication, however, appeared to come at the cost of accumulating deleterious mutations that decrease individual fitness (Cornejo et al. 2018). In the second study, RNA-seq data collected from 31 genotypes spread across four populations (Guiana, Iquitos, Marañon, and Nanay) revealed signatures of polygenic adaptation in co-expressed gene clusters (Hämälä et al. 2020). These results suggest unique challenges present in each population's environment shape aspects of flowering time regulation, protein modification, and water transport important for the growth and survival. Lastly, local adaptation was also observed in a separate study examining structural variation across the same 31 genotypes. Using high quality genome assemblies, Hämälä et al. found structural variants that were both highly

associated with gene expression across populations, as well as displayed strong signatures of local adaptation (Hämälä et al. 2021). Together, these results highlight important biological differences across populations, further supporting the need for large and diverse comparative analyses that help identify loci important for breeding more resilient cacao.

Wild Theobroma species as a source of beneficial traits

Crop wild relatives represent a reservoir of potentially beneficial traits. For example, wild relatives of wheat and barley are extremely drought tolerant, and could help improve cereal productivity (Nevo and Chen 2010). The wild tomato species *Solanum chilense* shows variation in resistance to the devastating *Solanum* pathogen *Phytophthora infestans*, and could therefore be a valuable source of durable resistance genes (Stam, Scheikl, and Tellier 2017). One of the best illustrations of the value wild relatives represent is found in wine grapes (*Vitis vinifera*). In the late 19th century the Homopteran pathogen *Daktulosphaira vitifoliae,* commonly called phylloxera, devastated to European vineyards (Granett et al. 2001). It was not until scientists discovered American *Vitis* species were resistant to phylloxera, leading to the production of resistant rootstocks and hybrid varieties, that *Vitis vinifera* was saved from destruction. Even today, the majority of phylloxera resistant rootstocks were developed from wild relatives found in America (Granett et al. 2001).

Despite their potential importance as a source of beneficial traits, little research has been done on cacao's 21 wild relatives. This is likely because hybridization between *T. cacao* and wild *Theobroma spp.* often yields developmentally stunted progeny (Martinson 1966). Research on *T. bicolor* and *T. grandiflorum* suggests they display natural fruit abscission, a trait that would decrease labor costs, thereby increasing profitability of cacao cultivation. Moreover, work on *T. microcarpum* suggests it could be a beneficial source of traits controlling canopy architecture (D.

Zhang et al. 2011). Very little information is known about disease resistance across wild *Theobroma spp.* (Harry C. Evans 2016a; Rocha 1966)

1.2 Pathogens of cacao

Cacao is attacked by a variety of pests and pathogens, together which lead to an estimated 40% pre-harvest yield loss annually (Ploetz 2007; B. A. Bailey and Meinhardt 2018). These pathogens and pests belong to multiple kingdoms and phyla, from insects and viruses to fungi and oomycetes. While the geographic range for some of these diseases overlap, many of the worst have remained, at least for now, isolated to specific growing regions (B. A. Bailey and Meinhardt 2018). The following is a description of the fungal, virus, and oomycete pathogens of cacao.

Black pod rot – Phytophthora spp.

Cacao black pod rot is a disease characterized by necrosis of pod tissue and the seeds and pulp contained within. It is caused by seven species in the stramenopile oomycete genus *Phytophthora* (Shahin S. Ali, Shao, Lary, Strem, et al. 2017; Shahin S. Ali, Shao, Lary, Kronmiller, et al. 2017; Surujdeo-Maharaj et al. 2016). The species are as follows: *P. palmivora, P. megakarya, P. citrophthora, P capsici, P. megasperma, P. katsurae* (Guest 2007). Of these species, *P. palmivora* and *P. megakarya* are responsible for the greatest annual yield loss (Shahin S. Ali, Shao, Lary, Kronmiller, et al. 2017) and are particularly problematic in West Africa.

P. palmivora and *P. megakarya* survive in warm, wet climates like those found in cacao growing regions (B. A. Bailey and Meinhardt 2018). They infect plants through motile zoospores that effectively move through thin layers on water on the leaf surface, but both species can survive in soil for years as mycelia and/or chlamydospores (Bailey et al. 2016). *P. palmivora* and

P. megakarya are hemibiotrophic pathogens, meaning they colonize and feed on living tissue before becoming saprophytic. Both pathogens invade pod tissue through wounds, stomata, or by penetration of the pod epidermis using appressoria (Shahin S. Ali, Shao, Lary, Kronmiller, et al. 2017). *P. megakarya,* named for its large chromosomes (Morales-Cruz et al. 2020), penetrates the pod epidermis more readily than *P. palmivora,* and is therefore less reliant on wounding or open stomata (S. S. Ali et al. 2016). Upon entry into the cell, pathogen haustoria secrete effector proteins to subvert plant immunity or create a metabolic environment better suited to pathogen growth (Shahin S. Ali, Shao, Lary, Kronmiller, et al. 2017; Akrofi et al. 2015).

Evidence suggests *P. palmivora*'s center of origin is Southeast Asia, but it is currently distributed worldwide, including Indonesia, Malaysia, Guam, Trinidad, Jamaica, United States, Colombia, Argentina, Peru, and Côte d'Ivoire (Y. Guo et al. 2021). Consistent with its broad dispersal, *P. palmivora* is a generalist pathogen than can infect a wide variety of plant species, including cacao, papaya, rubber, and coconut (Jianan Wang et al. 2020; Erwin and Ribeiro 1996). *P. megakarya*, on the other hand, is believed to be specifically adapted to infect cacao, though plants native to West Africa likely act as reservoirs (Shahin S. Ali, Shao, Lary, Kronmiller, et al. 2017; Akrofi et al. 2015).

Frosty pod rot – Moniliophthora rorei

Frosty pod rot of cacao is characterized by pod necrosis and the formation of white secondary inoculum on the surface of the pod epidermis (Meinhardt et al. 2008). Frosty pod rot is caused by the basidiomycete fungal pathogen *Moniliophthora rorei,* a hemibiotroph closely related to the witches' broom-causing fungus *Moniliophthora perniciosa* (Harry C. Evans 2016a; Meinhardt et al. 2008). Infection of cacao pods occurs via basidiospores that germinate on pod tissue, invading either through open stomata or hyphal penetration of the epidermis using an

appressorium (Harry C. Evans 2016a). Juvenile cacao pods are the most susceptible to frost pod rot. Most *M. rorei* infections undergo a prolonged biotrophic phase (40-90 days), during which the infection is difficult to detect (B. A. Bailey et al. 2018). Upon switching to a saprophytic phase, characteristic white spores form on the outside of pods. *M. rorei* can infect young cacao leaves in the lab, but it is not common in the field and does not result in the production of secondary spores (Harry C. Evans 2016a).

Evidence suggests *M. rorei*'s center of origin is Colombia, but it is now widespread across South America (Phillips-Mora, Aime, and Wilkinson 2007). Management of frosty pod rot can be achieved through a combination of cultural practices, chemical agents, biological control, and resistance breeding. While breeding programs have successfully generated clones tolerant to the disease, there are very few tolerant genotypes and no known accessions with complete resistance (H. C. Evans 2002; Romero Navarro et al. 2017). Moreover, *M. rorei* has been shown to infect all non-cacao *Theobroma spp.* as well as all species in cacao's sister genus *Herrania* (Phillips-Mora and Wilkinson 2007).

Witches' broom – Moniliophthora perniciosa

Witches' broom of cacao is caused by the basidiomycete pathogen *Moniliophthora perniciosa* (Meinhardt et al. 2008). The first symptoms of a witches' broom infection are the proliferation of auxiliary shoots, forming bush-like growths on the tree referred to as green brooms (Teixeira et al. 2014). After 30-60 days, tissue experiencing biotrophic infection becomes necrotic (Meinhardt et al. 2008; Harry C. Evans 2016b). Any necrotic tissue can then give rise, following a series of wet and dry periods, to additional basidiospores capable of infecting new tissue. This necrotic phase results in the necrosis of pod tissue, rendering it unusable (Mondego et al. 2008). Cacao flower cushions can also be infected by *M. perniciosa*, resulting in

parthenocarpic fruits, i.e. fruits that lack seeds (Melnick et al. 2012). Moreover, repeated infections will result in plant mortality.

M. perniciosa is native to Latin America, similar to its sister species *M. rorei.* However, unlike its sister species, *M. perniciosa* is thought to have co-evolved with cacao, rather than a wild *Thebroma spp.* (Harry C. Evans 2016b, 2016a). Most species in *Theobroma* and *Herrania* are potential hosts of *M. perniciosa.* In fact, *M. perniciosa* is known to attack a wide range of species, including those in Solanales, Malphighiales, Lamiales, and Malvales (Harry C. Evans 2016b). Witches' broom disease is problematic across many growing regions, but the hardest hit are in Brazil (Meinhardt et al. 2008).

Ceratocystis wilt of cacao – Ceratocystis cacaofunesta

Ceratocystis wilt of cacao is caused by the ascomycete fungus *Ceratocystis cacaofunesta* (Cabrera et al. 2016). Cacao trees susceptible to *C. cacaofunesta* are rapidly colonized by the pathogen, resulting in systemic infection followed by wilted leaves and eventual death (Ambrosio et al. 2013). For this reason, *C. cacaofunesta* presents a substantial problem to not only farms but germplasm collections as well. *C. cacaofunesta* is a xylem pathogen, often invading passively through wounds created by insects, e.g. bark beetles (Mazón, Díaz, and Gaviria 2013). Once inside the stem, *C. cacaofunesta* colonizes the xylem parenchyma and eventually the secondary xylem, thus restricting the flow of water and nutrients causing the wilted leaves characteristic of this disease (Clérivet et al. 2000). *C. cacaofunesta* can also infect through the roots of plants, similar to other species in Ceratocystis, causing necrosis and plant death (J. A. Johnson, Harrington, and Engelbrecht 2005). *C. cacaofunesta* is distributed across Latin America and the Carribbean. As a genus, *Ceratocystis* is distributed worldwide, but diversity measurements using

microsatellite and RFLP markers suggest South America may be its center of origin (Engelbrecht et al. 2007; Cabrera et al. 2016).

Cacao swollen shoot virus

Cacao swollen shoot disease is caused by a virus of the same name (CSSV), and is the major viral pathogen of cacao (Muller 2016). CSSV infection results in anatomical abnormalities such as cell proliferation in the xylem, phloem, and cortex (Jacquot et al. 1999), resulting in the swollen shoots and roots characteristic of this disease. CSSV also infects leaves, with symptoms ranging from red discoloration of leaf veins in young flushes, to diffuse yellow flecking, blotches, and streaks in mature leaves (Posnette and Robertson 1950; Muller 2016). Some strains of CSSV can even result in spherical cacao fruits. As the CSSV infection becomes systemic in susceptible varieties, large scale defoliation and dieback occur, eventually leading to plant death (Muller 2016).

CSSV is vectored by at least 14 species of mealybugs, typical of many Badnaviruses, and does not spread quickly (Roivainen 1976). This slow, short-ranged dispersal has limited its spread. CSSV is believed to be native to West Africa, and is currently only found in the Eastern Hemisphere (Muller 2016). Several species in Malvales native to this region act as reservoirs for CSSV, including *Ceiba pentandra, Adansonia digitata,* and multiple *Cola spp.* (Posnette, Robertson, and Todd 1950). Management of CSSV involves eradication of infected limbs and whole trees, control of mealybug populations, and breeding resistant cultivars. Similar to other cacao pathogens, this last management strategy is made more difficult by the fact that no completely resistant genotypes have been discovered (Thresh and Owusu 1986).

Resistance breeding in cacao

For centuries, farmers and breeders have sought to generate cacao clones that are highly productive, fine flavor, and resistant to disease (Cornejo et al. 2018). Since the 1950s, much of this work has centered around a single collection of cacao plants collected in the 1930s and 1940s called the "Pound Collection" (D. Zhang and Motilal 2016). Responding to a witches' broom outbreak in Trinidad in the 1920s, F.J. Pound and colleagues prepared an expedition to the Upper Amazon to search for resistant germplasm (D. Zhang and Motilal 2016). Collection sites included the Amazon tributaries Rio Ucayali, Rio Marañón, and Rio Morona (Bartley 2005; D. Zhang et al. 2009). Individuals collected during this expedition include mother trees from the groups designated "Iquitos Mixed Calabacillo" (IMC), "Morona" (MO), "Nanay" (NA), "Parinari" (PA), "Scavina" (SCA), and "Refractario" (D. Zhang and Motilal 2016). The first five of these groups were collected during his Peruvian expeditions from 1938-1943, while the Refractario group was later collected in Western Ecuador. In 1945 Pound revisited sites from his previous expeditions, collecting new germplasm that he labeled "Pound" or "P" clones. Together, from 1938-1945 F.J. Pound collected more than 130 mother trees (Efombagn et al. 2008; N. D. Pokou et al. 2009; D. Zhang and Motilal 2016).

Of these seven groups, IMC and Scavina are the most widely utilized in breeding programs, followed by Nanay, Refractario, and Parinari (D. Zhang and Motilal 2016). From these groups, NA-32, NA-34, IMC-67, and IMC-47 have been used in every major breeding program since the 1950s (Bartley 2005). Moreover, on-farm genetic diversity in West Africa and Indonesia, where cacao is an introduced crop species, represents a similarly small fraction of genetic diversity derived from the Pound collection and Trinitario hybrids (Efombagn et al. 2008; N. D. Pokou et al. 2009; D. Zhang and Motilal 2016).

SCA-6 and Pound-7 have been particularly important clones for resistance breeding and identification of quantitative trait loci (QTL) (Faleiro et al. 2006; Risterucci et al. 2003; Brown et al. 2005, 2007; Lanaud et al. 2009). Together, they have helped identify dozens of QTLs for black pod rot, witches' broom disease, and frosty pod rot. Many of these QTLs, however, span large genomic regions, or dovetail together in such a way that they encompass hundreds or even thousands of genes (Lanaud et al. 2009). This lack of resolution is due to a number of factors, including small mapping populations, low marker density, and few generations (often only F1 or F2). Recent advances in breeding technology, however, are attempting to rectify this problem. For instance, a single nucleotide polymorphism (SNP) chip with 15,000 markers was recently developed, making high-density linkage maps feasible (Livingstone et al. 2017). Moreover, technological advances in breeding methodology, such as genomic selection, are now being used to help hasten breeding programs (Romero Navarro et al. 2017).

1.3. Plant-pathogen interactions

Induced versus constitutive defenses

Plant resistance to pathogen invasion and/or herbivory occurs through two broad categories: constitutive defenses that are pre-formed, and induced defense responses that are produced upon pathogens challenge (Van Zandt 2007). Examples of constitutive defenses include defensive structures and inhibitory metabolites. For example, thorns offer protection from herbivory by large mammals like kudu, impala, and goats (Cooper and Owen-Smith 1986). Likewise, nicotine and clovamide are constitutively expressed defense metabolites that inhibit invading herbivores and pathogens in tobacco and cacao, respectively (Knollenberg et al. 2020; Steppuhn et al. 2004). Examples of induced defense responses include recruitment of natural

enemies via the production of green leaf volatiles (Pare and Tumlinson 1999), and the accumulation of inhibitory polyphenolic compounds like medicarpin (Jizong Wang, Wang, et al. 2019; Jizong Wang, Hu, et al. 2019). Both constitutive defenses and herbivore challenge are beyond the scope of this dissertation, thus the remaining review will be focused on induced responses to microbial pathogen challenge.

Plant recognition of invading pathogens occurs through a two-tiered system comprised of extracellular and intracellular components (Dodds and Rathjen 2010; Jones and Dangl 2006). Extracellular detection of pathogens occurs through pattern recognition receptors (PRR) responding to conserved molecular patterns (Zipfel 2014; Chinchilla et al. 2007). Intracellular detection occurs through NLR recognition of pathogen secreted small molecules, called effectors (Jones and Dangl 2006; Johal and Briggs 1992; Lewis et al. 2013).

PAMP and DAMP triggered immunity

Extracellular recognition of plant pathogens occurs through two types of PRRs: receptor-like proteins (RLP) and receptor-like kinases (RLK) (Kourelis and van der Hoorn 2018). These two classes of PRR detect extracellular molecules either given off by the pathogen or the host (Zipfel 2014). Pathogen associated microbial patterns (PAMP) are conserved proteins or molecules emitted by pathogenic microbes (Zipfel 2014). The most well-known PAMP is a conserved 22 amino acid fragment on the N-terminus of bacterial flagellin, called flg22 (Chinchilla et al. 2007; Felix et al. 1999). Many plants recognize various epitopes of this protein fragment through the RLK *FLAGELLIN SENSISITIVE 2* (*FLS2*) (Chinchilla et al. 2007; Gómez-Gómez and Boller 2000). Detection of flg22 by *FLS2* orthologs has been demonstrated in *Arabidopsis thaliana* (Gómez-Gómez and Boller 2000), *Nicotiana benthamiana* (Fürst et al. 2020)*, Vitis vinifera* (Trdá et al. 2014), *Oryza sativa* (Takai et al. 2008), and many other species.

Binding of flg22 by *FLS2* results in the recruitment of the RLK *BRI1-ASSOCIATED RECEPTOR KINASE* (*BAK1*) (Chinchilla et al. 2007). Together, *FLS2* and *BAK1* initiate the plant immune response to pathogen challenge. This takes the form of ion fluxes at the cell surface, followed by an increase in the cytosolic concentration of Ca^{2+}, the production of reactive oxygen species, and mitogen-activated protein kinase (MAPK) cascades (Boller and Felix 2009). Changes in cytosolic Ca^{2+} and MAPK cascades transmit pathogen information to the nucleus, resulting in the transcriptional changes characteristic of pattern triggered immunity (PTI) (Zipfel 2014). Other PAMPs are also elicited by pathogens, including *NECROSIS AND ETHYLENE-INDUCING PEPTIDE 1* (*NEP1*) (Pemberton and Salmond 2004) and bacterial peptidoglycans (Willmann et al. 2011).

The other class of extracellular molecular signals detected by PRRs are damage associated molecular patterns (DAMP), which are derived from damaged host tissue (Zipfel 2014). The most well-known DAMPs are the oligogalacturonides (OG) (Hou et al. 2019). During pathogen invasion, microbes secrete pectate lyases to destroy the plant cell wall (Doares et al. 1995). Destruction of the cell wall results in the production of OGs, which are detected by PRRs in the wall-associated kinases family (Brutus et al. 2010; Choi and Klessig 2016). The resulting defense response is similar to that elicited by PAMPs, filled with Ca^{2+} flux, ROS production, and MAPK cascades (Dodds and Rathjen 2010).

Effector molecules manipulate host physiology

In order to subvert PTI, pathogens have evolved effectors, small molecules, usually proteins, that are secreted into the apoplast or cell cytosol to disrupt host immunity and create a metabolic environment more suitable for pathogen growth (Zhou et al. 2011; Dodds and Rathjen 2010; Jones and Dangl 2006). Effector proteins interfere with various aspects of plant physiology

and immunity, including secondary metabolite production and gene expression. For example, the effector protein *HopZ1* from *Pseudomonas syringae* targets isoflavone biosynthesis in soybean, promoting infection (Zhou et al. 2011). Transcription activator-like effectors (TALE) in *Xanthomonas spp.* promote infection by entering the host nucleus and directly promoting the expression of susceptibility genes (Boch, Bonas, and Lahaye 2014). Lastly, *PSEUDOMONAS OUTER PROTEIN P2* (*PopP2*), an effector from *Ralstonia solanacearum,* acetylates lysine residues in WRKY proteins, disrupting their function and subsequent immune responsiveness (Sarris et al. 2015).

NLR recognition of effectors

Recognition of pathogen effectors takes place through intracellular NLR receptors (Lewis et al. 2013; Johal and Briggs 1992; Kourelis and van der Hoorn 2018). NLR detection of effectors can be distilled into three broad mechanisms: direct, indirect, or decoy/integrated decoy recognition (Van de Weyer et al. 2019a; Kourelis and van der Hoorn 2018). Direct recognition takes place when an NLR protein detects the presence of an effector by interacting with it. An example of direct recognition can be seen in the flax NLRs *L5, L6,* and *L7*, all of which interact with the *Melampsora lini* effector *AvrL567* (Dodds et al. 2006, 2004). Indirect recognition occurs when a pathogen effector disrupts any aspect of host immunity that is guarded by an NLR. Indirect recognition can be seen in the interaction between *Tobacco mosaic virus p50* and *N RECEPTOR-INTERACTING PROTEIN 1* (*NRIP1*). *NRIP1* is recruited to the cytoplasm by *p50*, where it is recognized by tobacco's *N* NLR thereby initiating a defense response (Caplan et al. 2008; Kourelis and van der Hoorn 2018). Lastly, decoy/integrated decoy recognition is similar to indirect recognition of guardees, except rather than disrupting an active component of plant immunity the effector disrupts a decoy. For instance, the effector *HopZ1a* from *Pseudomonas*

syringae belongs to the YopJ effector family known for disrupting kinases, thereby dampening plant immune responses (Mukherjee et al. 2006; Ma et al. 2006). In response to co-evolutionary pressure imposed by *HopZ1a*, A. *thaliana* has evolved a non-functional pseudokinase, *HOPZ-ETI-DEFICIENT 1* (*ZED1*), that acts as a decoy for *HopZ1a*-mediated acetylation (Lewis et al. 2013). Detecting *ZED1* acetylation by *HopZ1a*, the NLR *HOPZ-ACTIVATED-RESISTANCE 1* (*ZAR1*) launches a robust defense response (Lewis et al. 2013). Some NLRs form integrated decoys, where a decoy domain is directly fused to the NLR protein. When the decoy domain is perturbed a defense response is initiated. This can be seen in *A. thaliana RESISTANCE TO RALSTONIA SOLANACEARUM 1* (*RRS1*), which has an integrated WRKY transcription factor domain on the NLR's C-terminal end. Acetylation of this WRKY domain by *AvrRps4* from *Ralstonia solanacearum* results in a robust defense response (Sarris et al. 2015; Le Roux et al. 2015).

NLR receptors mediate a defense response called effector triggered immunity (ETI). ETI is very similar to PTI, involving ion fluxes, increased Ca^{2+} concentrations, ROS bursts, MAPK cascades, and defense gene induction (Dodds and Rathjen 2010). The differences between PTI and ETI, however, were outlined by Jones and Dangl (2006) in the so-called "zig-zag" model (Jones and Dangl 2006). The zig-zag model suggested that the principle difference between these two modes of defense is amplitude, where ETI is just an amplified version of PTI. However, ETI has the possibility to induce hypersensitive response (HR), a type of localized cell death at the point of pathogen invasion (Jones and Dangl 2006; Balint-Kurti 2019). This localized cell death often, but not necessarily, results in disease resistance (Balint-Kurti 2019).

Until recently, how NLRs manage to initiate HR was a mystery (Jeffery L. Dangl and Jones 2019). In a series of papers published in 2019, Wang et al. described the crystal structure of an NLR in both inactive and intermediate states (Jizong Wang, Wang, et al. 2019; Jizong Wang, Hu, et al. 2019). They revealed the NLR protein *ZAR1*, upon uridylation of its pseudokinase PBS-

LIKE PROTEINS 2 (*PBL2*) by the effector *AvrAC*, oligomerizes to form what the authors called a resistosome (Jizong Wang, Hu, et al. 2019). They named it as such because the pentangular structure appeared very similar to mammalian inflammasomes, like those formed by the NLR caspase activation and recruitment domains (CARD) in mice (Tenthorey et al. 2017). After oligomerization, the *ZAR1* inflammasome associates with the plasma membrane, creating a ion channel that leads to Ca^{2+} influx, ROS production and, subsequently, cell death (Bi et al. 2021).

While PTI and ETI were long considered separate biochemical reactions converging on a similar set of signaling pathways, the distinctions between the two processes have become increasingly less clear (Ngou, Jones, and Ding 2021; Lu and Tsuda 2021). For instance, we now recognize there is a greater degree of crosstalk between PTI and ETI than originally understood. Activation of PTI by flg22 has been shown to inhibit NLR-activated kinases (Hatsugai et al. 2017). Likewise, two recent studies demonstrated evidence that RBOHD-mediated ROS bursts were stronger when plants were treated with both an elicitor (flg22) and an effector (*AvrRpt2*), suggesting the ETI-PTI crosstalk can also be synergistic (Yuan et al. 2021; Ngou et al. 2021).

Induced phytohormone and chemical defenses

Beyond the ion fluxes, ROS bursts, and MAPK cascades associated with the acute defense responses of PTI and ETI, plants have evolved broader, longer lasting systemic responses as well (Fu and Dong 2013; Luna et al. 2012). Systemic acquired resistance (SAR) is activated by signaling pathways initiated during PTI and ETI, resulting in the production of a large set of phytohormones that regulate broad-spectrum disease resistance (Y. Chen et al. 2015). These phytohormones are auxin, salicylic acid (SA), jasmonic acid (JA), ethylene (ET), brassinosteroids (BRs), abscisic acid (ABA), gibberellic acid (GA), and cytokinin (CA). These eight phytohormones are all involved in systemic acquired resistance at some level, usually through

interaction with the SA or JA pathways (H. Guo and Ecker 2004; Divi, Rahman, and Krishna 2010; Anderson et al. 2004; De Bruyne, Höfte, and De Vleesschauwer 2014).

The two most well-studied regulators of SAR are SA and JA (Beckers and Spoel 2006; Vlot, Dempsey, and Klessig 2009). SA, considered the master regulator of SAR, is activated upon challenge by biotrophic and hemibiotrophic pathogens like *Puccinia graminis* and *Phytophthora spp.* (Vlot, Dempsey, and Klessig 2009). JA, on the other hand, is activated upon challenge from necrotrophic pathogens and herbivores like *Botrytis cinerea* and *Manduca sexta* (Browse 2009). The SA and JA pathways interact in both antagonistic and synergistic ways (Beckers and Spoel 2006), many of which are mediated by the SA receptors *NPR1*, *NPR3*, and *NPR4* (L. Liu et al. 2016; Chai et al. 2014). In cacao, both *TcNPR1* and *TcNPR3* have a demonstrated role in resistance to multiple *Phytophthora spp.* (Shi et al. 2010, 2013; Fister et al. 2018).

One hallmark of SAR is the activation of pathogenesis-related (PR) genes (van Loon, Rep, and Pieterse 2006). PR genes are involved in a variety immune-related functions, including degradation of pathogen cell walls, inhibition of proteinases, and maintenance of the redox environment (van Loon, Rep, and Pieterse 2006). In cacao, PR genes are upregulated in response to both fungal and oomycete pathogens (Fister, Mejia, et al. 2016; Snyder-Leiby and Furtek 1995). For instance, the a cacao class I endochitinase was differentially expressed in response to fungal elicitor treatment (Snyder-Leiby and Furtek 1995). Stable over-expression of this gene resulted in significant inhibition of the fungal pathogen *Colletotrichum gloeosporioides* (Siela N. Maximova et al. 2006).

1.4. Evolution of plant genomes

Understanding the size, structure, and complexity of plant genomes is essential for addressing current and emerging agricultural threats. Over the past 20 years, revolutions in

sequencing technology and computational methods have resulted in an abundance of high-quality genome assemblies (Michael and VanBuren 2020; Goodstein et al. 2012). These genome assemblies have transformed our understanding of gene and genome function, organization, and evolution For instance, phylogenomic analyses to identify gene duplication on a genomic scale revealed two ancestral whole genome duplications (WGD), one shared by seed plants and another shared by angiosperms (Jiao et al. 2011). Subsequent analyses have supported these paleopolyploidy events, as well as a host of others that, alongside the expansion of repeats and transposable elements, have contributed to the approximately 2,400-fold variation in genome size seen across land plants, from the 61 Mbp *Genlisea tuberosa* to the 149 Gbp *Paris japonica* (F. Li et al. 2015; D'Hont et al. 2012; Garcia-Mas et al. 2012; Pellicer and Leitch 2020; Michael 2014). Despite this variation in genome size, however, gene content has remained approximately the same, varying just 2-fold (Proost et al. 2011; Salse 2012).

Variation in gene content, while not as extreme as genome size, is caused by various types of gene duplication, including local, dispersed, and whole genome duplication events (Flagel and Wendel 2009). Tandem and proximal gene duplications occur when genes are copied locally, either through unequal crossing over (Leister 2004), slipped strand mispairing (Levinson and Gutman 1987), or TEs (Krasileva 2019; S. Kim et al. 2017). This results in clusters of tandemly arrayed genes (B. C. Meyers et al. 1998; Michelmore and Meyers 1998). Dispersed and segmental duplicates similarly form through unequal crossing over and TE-mediated movement (S. Kim et al. 2017; Y. Li et al. 2012), events in which regions of varying size are duplicated across the genome. Because duplicated genes are free of selective constraint, they can accumulate mutations (Ohno 1970), leading to gene duplicates with new (Yang et al. 2015) or modified functions (Force et al. 1999). Duplicate genes can also lose their function entirely, becoming nonfunctional (Blake C. Meyers et al. 2003; Van de Weyer et al. 2019a).

In the case of whole genome duplications, this process of duplicate gene loss is called fractionation, wherein duplicated regions of the genome are returned to a diploid state (Cheng et al. 2018). This pattern can be seen in the set of benchmarking universal single copy orthologs (BUSCO), a set of genes that are maintained as a single copy across all land plants, despite drastically different histories of genome duplication (Waterhouse et al. 2013; Simão et al. 2015). Other gene families, however, are preferentially retained following whole genome duplication (Lorin et al. 2018; Wu et al. 2008). For instance, starch synthesis genes in rice have many polyploidy-derived duplicates, indicative of their preferential retention after whole genome duplication (Wu et al. 2008). Together, these three processes lead to the expansion and contraction of gene families, subsequently helping drive variation in gene copy number across species.

1.5 Dissertation Overview

Plants have complex and dynamic immune systems that have evolved over millennia to help them resist pathogen invasion. The combined evolutionary forces of mutation, selection, and drift have worked together to shape variation in disease resistance across taxonomic scales. Humans have recognized the importance of natural variation in plant disease resistance for millennia, beginning with the Greek philosopher Theophrastus in 300 B.C.E. (Bockus et al. 2001; Theophrastus 1989). Harnessing variation for crop improvement followed shortly thereafter, and the effects of subsequent selection for yield, flavor, disease resistance and many other traits can be seen in a host of species (Meyer and Purugganan 2013). Since then, disease resistance has become an expansive and rapidly evolving field. Every subheading in this literature review, indeed many individual sentences, have been the subject of intensively researched review articles.

Despite this depth of knowledge, many aspects of plant disease resistance remain opaque for species that are difficult to grow, transform, and cross, such as tree crops like cacao. This dissertation is an effort to bridge that division. At the heart of this work are two fundamental questions: How did cacao's defense mechanisms evolve? And can we use this evolutionary information to identify genes important for disease resistance? At a superficial level these questions appear simple. Clarifying these gaps in understanding, however, involves untangling complex networks of action and reaction, both within and between species. Moreover, cacao populations are diverse, each shaped by their own evolutionary histories and unique environments, adding further complexity to our understanding of its defense response. To make this complexity tractable, each chapter focuses on a distinct aspect of cacao's defense response and how it evolved, across both populations and closely related species.

While extensive research has been done to identify variation in cacao's defense response, most of this work has been limited to a narrow set of genotypes. Moreover, much of the cacao germplasm currently used by farmers was generated from just a handful of clones. To gain a deeper understanding of cacao's natural variation in defense response, Chapter 2 investigates defense against *Phytophthora palmivora* across 31 genotypes representing four populations of cacao. Each population contains both genotypes that are resistant to *P. palmivora* as well as genotypes that are susceptible. Using a combination of transcriptomic, genomic, and metabolomic datasets we search for genes important for disease resistance across both populations and phenotype classes.

Extending this rationale one degree further, in Chapter 3 we again investigate defense response to *P. palmivora*, this time across four close relatives of *T. cacao*. The underlying assumption being that at least some portion of disease resistance is monophyletic, i.e. originated once and is shared across species. We compare resistance phenotypes and defense responses in these wild *Theobroma spp.* to those seen across populations of cacao, in an attempt to identify

genes or pathways that are indispensable to *Theobroma*'s disease resistance. In doing so, we identified several genes and gene families that were differentially expressed in response to *P. palmivora* challenge across *Theobroma,* and displayed signatures of positive selection indicative of long-term involvement in plant-pathogen interactions.

Many thousands of genes in each species are involved in defense response. The relative contribution of each gene to the overall disease phenotype is variable, but few are more integral to plant defense than NLR immune receptors. Investigating intraspecific NLR diversity is, therefore, an important aspect of understanding how a species interacts with its environment, and provides insight into the ways NLR variation, to the extent that it exists, can be harnessed for crop improvement. However, due to the limited number of high-quality genomes, we know little about NLR variation within a single species. In Chapter 4, we examine NLR diversity across 11 high-quality cacao genomes. Similar to previous studies, we observed complex evolutionary patterns that led to significant variation in NLR copy number. While NLR copy number was not associated with any documented resistance phenotypes, the identification and delineation of these genes is an important resource for scientists and breeders alike.

Finally, in Chapter 5, we describe both the promise and limitations of this work, outlining future experiments that could further elucidate the mechanisms underlying cacao's defense response. As advances in sequencing technology and analysis continue to decrease the cost of high-quality genome assemblies, the exploration of natural variation in a host of traits will be possible. The implications of this cost abatement for comparative genomics and cacao improvement is discussed.

References

Akrofi, Andrews Yaw, Ishmael Amoako-Atta, Michael Assuah, and Eric Kumi Asare. 2015. "Black Pod Disease on Cacao (Theobroma Cacao, L) in Ghana: Spread of Phytophthora Megakarya and Role of Economic Plants in the Disease Epidemiology." *Crop Protection (Guildford, Surrey)* 72 (June): 66–75.

Ali, S. S., I. Amoako-Attah, R. A. Bailey, M. D. Strem, M. Schmidt, A. Y. Akrofi, S. Surujdeo-Maharaj, et al. 2016. "PCR-Based Identification of Cacao Black Pod Causal Agents and Identification of Biological Factors Possibly Contributing ToPhytophthora Megakarya's Field Dominance in West Africa." *Plant Pathology* 65 (7): 1095–1108.

Ali, Shahin S., Jonathan Shao, David J. Lary, Brent Kronmiller, Danyu Shen, Mary D. Strem, Ishmael Amoako-Attah, et al. 2017. "Phytophthora Megakarya and P. Palmivora, Closely Related Causal Agents of Cacao Black Pod Rot, Underwent Increases in Genome Sizes and Gene Numbers by Different Mechanisms." *Genome Biology and Evolution* 9 (3): 536–57.

Ali, Shahin S., Jonathan Shao, David J. Lary, Mary D. Strem, Lyndel W. Meinhardt, and Bryan A. Bailey. 2017. "Phytophthora Megakarya and P. Palmivora, Causal Agents of Black Pod Rot, Induce Similar Plant Defense Responses Late during Infection of Susceptible Cacao Pods." *Frontiers in Plant Science* 8 (February): 169.

Ambrosio, Alinne Batista, Leandro Costa do Nascimento, Bruno V. Oliveira, Paulo José P. L. Teixeira, Ricardo A. Tiburcio, Daniela P. Toledo Thomazella, Adriana F. P. Leme, et al. 2013. "Global Analyses of Ceratocystis Cacaofunesta Mitochondria: From Genome to Proteome." *BMC Genomics* 14 (1): 91.

Anderson, Jonathan P., Ellet Badruzsaufari, Peer M. Schenk, John M. Manners, Olivia J. Desmond, Christina Ehlert, Donald J. Maclean, Paul R. Ebert, and Kemal Kazan. 2004. "Antagonistic Interaction between Abscisic Acid and Jasmonate-Ethylene Signaling

Pathways Modulates Defense Gene Expression and Disease Resistance in Arabidopsis." *The Plant Cell* 16 (12): 3460–79.

Argout, Xavier, Jerome Salse, Jean-Marc Aury, Mark J. Guiltinan, Gaetan Droc, Jerome Gouzy, Mathilde Allegre, et al. 2011. "The Genome of Theobroma Cacao." *Nature Genetics* 43 (2): 101–8.

Arnold, Sarah E. J., Samantha J. Forbes, David R. Hall, Dudley I. Farman, Puran Bridgemohan, Gustavo R. Spinelli, Daniel P. Bray, et al. 2019. "Floral Odors and the Interaction between Pollinating Ceratopogonid Midges and Cacao." *Journal of Chemical Ecology* 45 (10): 869–78.

Asselin, Jo Ann E., Jinshan Lin, Alvaro L. Perez-Quintero, Irene Gentzel, Doris Majerczak, Stephen O. Opiyo, Wanying Zhao, et al. 2015. "Perturbation of Maize Phenylpropanoid Metabolism by an AvrE Family Type III Effector from Pantoea Stewartii." *Plant Physiology* 167 (3): 1117–35.

Bailey, Bryan A., Shahin S. Ali, Andrews Y. Akrofi, and Lyndel W. Meinhardt. 2016. "Phytophthora Megakarya, a Causal Agent of Black Pod Rot in Africa." In *Cacao Diseases*, 267–303. Cham: Springer International Publishing.

Bailey, Bryan A., Harry C. Evans, Wilbert Phillips-Mora, Shahin S. Ali, and Lyndel W. Meinhardt. 2018. "Moniliophthora Roreri, Causal Agent of Cacao Frosty Pod Rot." *Molecular Plant Pathology* 19 (7): 1580–94.

Bailey, Bryan A., and Lyndel W. Meinhardt, eds. 2018. *Cacao Diseases*. Cham, Switzerland: Springer International Publishing.

Balint-Kurti, Peter. 2019. "The Plant Hypersensitive Response: Concepts, Control and Consequences." *Molecular Plant Pathology* 20 (8): 1163–78.

Bartley, B. G. D. 2005. "The Utilization of the Genetic Resources." In *The Genetic Diversity of Cacao and Its Utilization*, 309–22. Wallingford: CABI.

Beckers, G. J. M., and S. H. Spoel. 2006. “Fine-Tuning Plant Defence Signalling: Salicylate versus Jasmonate.” *Plant Biology (Stuttgart, Germany)* 8 (1): 1–10.

Bi, Guozhi, Min Su, Nan Li, Yu Liang, Song Dang, Jiachao Xu, Meijuan Hu, et al. 2021. “The ZAR1 Resistosome Is a Calcium-Permeable Channel Triggering Plant Immune Signaling.” *Cell* 184 (13): 3528-3541.e12.

Biffen, R. H. 1905. “Mendel’s Laws of Inheritance and Wheat Breeding.” *The Journal of Agricultural Science* 1 (1): 4–48.

Boch, Jens, Ulla Bonas, and Thomas Lahaye. 2014. “TAL Effectors--Pathogen Strategies and Plant Resistance Engineering.” *The New Phytologist* 204 (4): 823–32.

Bockus, William W., Jon A. Appel, Robert L. Bowden, Allan K. Fritz, Bikram S. Gill, T. Joe Martin, Rollin G. Sears, Dallas L. Seifers, Gina L. Brown-Guedira, and Merle G. Eversmeyer. 2001. “Success Stories: Breeding for Wheat Disease Resistance in Kansas.” *Plant Disease* 85 (5): 453–61.

Boller, Thomas, and Georg Felix. 2009. “A Renaissance of Elicitors: Perception of Microbe-Associated Molecular Patterns and Danger Signals by Pattern-Recognition Receptors.” *Annual Review of Plant Biology* 60 (1): 379–406.

Brown, J. Steven, Wilbert Phillips-Mora, Emilio J. Power, Cheryl Krol, Cuauhtemoc Cervantes-Martinez, Juan Carlos Motamayor, and Raymond J. Schnell. 2007. “Mapping QTLs for Resistance to Frosty Pod and Black Pod Diseases and Horticultural Traits InTheobroma CacaoL.” *Crop Science* 47 (5): 1851–58.

Brown, J. Steven, R. J. Schnell, J. C. Motamayor, Uilson Lopes, David N. Kuhn, and James W. Borrone. 2005. “Resistance Gene Mapping for Witches’ Broom Disease in Theobroma Cacao L. in an F2 Population Using SSR Markers and Candidate Genes.” *Journal of the American Society for Horticultural Science. American Society for Horticultural Science* 130 (3): 366–73.

Browse, John. 2009. “Jasmonate Passes Muster: A Receptor and Targets for the Defense Hormone.” *Annual Review of Plant Biology* 60 (1): 183–205.

Brutus, Alexandre, Francesca Sicilia, Alberto Macone, Felice Cervone, and Giulia De Lorenzo. 2010. “A Domain Swap Approach Reveals a Role of the Plant Wall-Associated Kinase 1 (WAK1) as a Receptor of Oligogalacturonides.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (20): 9452–57.

Cabrera, Odalys García, Eddy Patricia López Molano, Juliana José, Javier Correa Álvarez, and Gonçalo Amarante Guimarães Pereira. 2016. “Ceratocystis Wilt Pathogens: History and Biology—Highlighting C. Cacaofunesta, the Causal Agent of Wilt Disease of Cacao.” In *Cacao Diseases*, 383–428. Cham: Springer International Publishing.

Caplan, Jeffrey L., Padmavathi Mamillapalli, Tessa M. Burch-Smith, Kirk Czymmek, and S. P. Dinesh-Kumar. 2008. “Chloroplastic Protein NRIP1 Mediates Innate Immune Receptor Recognition of a Viral Effector.” *Cell* 132 (3): 449–62.

Chai, Jinyu, Jian Liu, Jun Zhou, and Da Xing. 2014. “Mitogen-Activated Protein Kinase 6 Regulates NPR1 Gene Expression and Activation during Leaf Senescence Induced by Salicylic Acid.” *Journal of Experimental Botany* 65 (22): 6513–28.

Chen, Li-Qing. 2014. “SWEET Sugar Transporters for Phloem Transport and Pathogen Nutrition.” *The New Phytologist* 201 (4): 1150–55.

Cheng, Feng, Jian Wu, Xu Cai, Jianli Liang, Michael Freeling, and Xiaowu Wang. 2018. “Gene Retention, Fractionation and Subgenome Differences in Polyploid Plants.” *Nature Plants* 4 (5): 258–68.

Chinchilla, Delphine, Cyril Zipfel, Silke Robatzek, Birgit Kemmerling, Thorsten Nürnberger, Jonathan D. G. Jones, Georg Felix, and Thomas Boller. 2007. “A Flagellin-Induced Complex of the Receptor FLS2 and BAK1 Initiates Plant Defence.” *Nature* 448 (7152): 497–500.

Choi, Hyong Woo, and Daniel F. Klessig. 2016. “DAMPs, MAMPs, and NAMPs in Plant Innate Immunity.” *BMC Plant Biology* 16 (1): 232.

Clérivet, Alain, Véronique Déon, Ibtissam Alami, Fredérique Lopez, Jean-Paul Geiger, and Michel Nicole. 2000. “Tyloses and Gels Associated with Cellulose Accumulation in Vessels Are Responses of Plane Tree Seedlings (Platanus × Acerifolia) to the Vascular Fungus Ceratocystis Fimbriata f. Sp Platani.” *Trees (Berlin, Germany: West)* 15 (1): 25–31.

Cooper, Susan M., and Norman Owen-Smith. 1986. “Effects of Plant Spinescence on Large Mammalian Herbivores.” *Oecologia* 68 (3): 446–55.

Cornejo, Omar E., Muh-Ching Yee, Victor Dominguez, Mary Andrews, Alexandra Sockell, Erika Strandberg, Donald Livingstone 3rd, et al. 2018. “Population Genomic Analyses of the Chocolate Tree, Theobroma Cacao L., Provide Insights into Its Domestication Process.” *Communications Biology* 1 (1): 167.

Cuatrecasas, José. 1964. *Cacao and Its Allies: A Taxonomic Revision of the Genus Theobroma*. Smithsonian Institution.

Dangl, Jeffery L., and Jonathan D. G. Jones. 2019. “A Pentangular Plant Inflammasome.” *Science (New York, N.Y.)* 364 (6435): 31–32.

De Bruyne, Lieselotte, Monica Höfte, and David De Vleesschauwer. 2014. “Connecting Growth and Defense: The Emerging Roles of Brassinosteroids and Gibberellins in Plant Innate Immunity.” *Molecular Plant* 7 (6): 943–59.

D’Hont, Angélique, France Denoeud, Jean-Marc Aury, Franc-Christophe Baurens, Françoise Carreel, Olivier Garsmeur, Benjamin Noel, et al. 2012. “The Banana (Musa Acuminata) Genome and the Evolution of Monocotyledonous Plants.” *Nature* 488 (7410): 213–17.

Dillinger, T. L., P. Barriga, S. Escárcega, M. Jimenez, D. Salazar Lowe, and L. E. Grivetti. 2000. "Food of the Gods: Cure for Humanity? A Cultural History of the Medicinal and Ritual Use of Chocolate." *The Journal of Nutrition* 130 (8S Suppl): 2057S-72S.

Divi, Uday K., Tawhidur Rahman, and Priti Krishna. 2010. "Brassinosteroid-Mediated Stress Tolerance in Arabidopsis Shows Interactions with Abscisic Acid, Ethylene and Salicylic Acid Pathways." *BMC Plant Biology* 10 (1): 151.

Doares, S. H., T. Syrovets, E. W. Weiler, and C. A. Ryan. 1995. "Oligogalacturonides and Chitosan Activate Plant Defensive Genes through the Octadecanoid Pathway." *Proceedings of the National Academy of Sciences of the United States of America* 92 (10): 4095–98.

Dodds, Peter N., Gregory J. Lawrence, Ann-Maree Catanzariti, Michael A. Ayliffe, and Jeffrey G. Ellis. 2004. "The Melampsora Lini AvrL567 Avirulence Genes Are Expressed in Haustoria and Their Products Are Recognized inside Plant Cells." *The Plant Cell* 16 (3): 755–68.

Dodds, Peter N., Gregory J. Lawrence, Ann-Maree Catanzariti, Trazel Teh, Ching-I A. Wang, Michael A. Ayliffe, Bostjan Kobe, and Jeffrey G. Ellis. 2006. "Direct Protein Interaction Underlies Gene-for-Gene Specificity and Coevolution of the Flax Resistance Genes and Flax Rust Avirulence Genes." *Proceedings of the National Academy of Sciences of the United States of America* 103 (23): 8888–93.

Dodds, Peter N., and John P. Rathjen. 2010. "Plant Immunity: Towards an Integrated View of Plant-Pathogen Interactions." *Nature Reviews. Genetics* 11 (8): 539–48.

Efombagn, Ives Bruno M., Juan C. Motamayor, Olivier Sounigo, Albertus B. Eskes, Salomon Nyassé, Christian Cilas, Ray Schnell, Maria J. Manzanares-Dauleux, and Maria Kolesnikova-Allen. 2008. "Genetic Diversity and Structure of Farm and GenBank

Accessions of Cacao (Theobroma Cacao L.) in Cameroon Revealed by Microsatellite Markers." *Tree Genetics & Genomes* 4 (4): 821–31.

Engelbrecht, C. J. B., T. C. Harrington, A. C. Alfenas, and C. Suarez. 2007. "Genetic Variation in Populations of the Cacao Wilt Pathogen, Ceratocystis Cacaofunesta." *Plant Pathology* 56 (6): 923–33.

Erwin, Donald C., and Olaf K. Ribeiro. 1996. *Phytophthora Diseases Worldwide*. St Paul: American Phytopathological Society.

Evans, H. C. 2002. "Invasive Neotropical Pathogens of Tree Crops." In *Tropical Mycology: Volume 2, Micromycetes*, 83–112. Wallingford: CABI.

Evans, Harry C. 2016a. "Frosty Pod Rot (Moniliophthora Roreri)." In *Cacao Diseases*, 63–96. Cham: Springer International Publishing.

———. 2016b. "Witches' Broom Disease (Moniliophthora Perniciosa): History and Biology." In *Cacao Diseases*, 137–77. Cham: Springer International Publishing.

Faleiro, F. G., V. T. Queiroz, U. V. Lopes, C. T. Guimaraes, J. L. Pires, M. M. Yamada, I. S. Araújo, et al. 2006. "Mapeamento Genético Molecular Do Cacaueiro (Theobroma Cacao L.) e QTLs Associados Aresistência Avassoura-de-Bruxa." *Euphytica/ Netherlands Journal of Plant Breeding* 149: 227–35.

Felix, G., J. D. Duran, S. Volko, and T. Boller. 1999. "Plants Have a Sensitive Perception System for the Most Conserved Domain of Bacterial Flagellin." *The Plant Journal: For Cell and Molecular Biology* 18 (3): 265–76.

Ferry, Robert J. 2020. *The Colonial Elite of Early Caracas*. Berkeley, CA: University of California Press.

Fister, Andrew S., Lena Landherr, Siela N. Maximova, and Mark J. Guiltinan. 2018. "Transient Expression of CRISPR/Cas9 Machinery Targeting TcNPR3 Enhances Defense Response in Theobroma Cacao." *Frontiers in Plant Science* 9 (March): 268.

Fister, Andrew S., Luis C. Mejia, Yufan Zhang, Edward Allen Herre, Siela N. Maximova, and Mark J. Guiltinan. 2016. “Theobroma Cacao L. Pathogenesis-Related Gene Tandem Array Members Show Diverse Expression Dynamics in Response to Pathogen Colonization.” *BMC Genomics* 17 (1). https://doi.org/10.1186/s12864-016-2693-3.

Fister, Andrew S., Shawn T. O’Neil, Zi Shi, Yufan Zhang, Brett M. Tyler, Mark J. Guiltinan, and Siela N. Maximova. 2015. “Two Theobroma Cacao Genotypes with Contrasting Pathogen Tolerance Show Aberrant Transcriptional and ROS Responses after Salicylic Acid Treatment.” *Journal of Experimental Botany* 66 (20): 6245–58.

Flagel, Lex E., and Jonathan F. Wendel. 2009. “Gene Duplication and Evolutionary Novelty in Plants.” *The New Phytologist* 183 (3): 557–64.

Florez, Sergio L., Rachel L. Erwin, Siela N. Maximova, Mark J. Guiltinan, and Wayne R. Curtis. 2015. “Enhanced Somatic Embryogenesis in Theobroma Cacao Using the Homologous BABY BOOM Transcription Factor.” *BMC Plant Biology* 15 (1): 121.

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. 1999. “Preservation of Duplicate Genes by Complementary, Degenerative Mutations.” *Genetics* 151 (4): 1531–45.

Fu, Zheng Qing, and Xinnian Dong. 2013. “Systemic Acquired Resistance: Turning Local Infection into Global Defense.” *Annual Review of Plant Biology* 64 (1): 839–63.

Fürst, Ursula, Yi Zeng, Markus Albert, Anna Kristina Witte, Judith Fliegmann, and Georg Felix. 2020. “Perception of Agrobacterium Tumefaciens Flagellin by FLS2XL Confers Resistance to Crown Gall Disease.” *Nature Plants* 6 (1): 22–27.

Galloway, L. F., and C. B. Fenster. 2000. “Population Differentiation in an Annual Legume: Local Adaptation.” *Evolution; International Journal of Organic Evolution* 54 (4): 1173–81.

Garcia-Mas, Jordi, Andrej Benjak, Walter Sanseverino, Michael Bourgeois, Gisela Mir, Víctor M. González, Elizabeth Hénaff, et al. 2012. “The Genome of Melon (Cucumis Melo L.).” *Proceedings of the National Academy of Sciences of the United States of America* 109 (29): 11872–77.

Gómez-Gómez, L., and T. Boller. 2000. “FLS2: An LRR Receptor-like Kinase Involved in the Perception of the Bacterial Elicitor Flagellin in Arabidopsis.” *Molecular Cell* 5 (6): 1003–11.

Gonzalez, Antonio, Mingzhe Zhao, John M. Leavitt, and Alan M. Lloyd. 2008. “Regulation of the Anthocyanin Biosynthetic Pathway by the TTG1/BHLH/Myb Transcriptional Complex in Arabidopsis Seedlings.” *The Plant Journal: For Cell and Molecular Biology* 53 (5): 814–27.

Goodstein, David M., Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D. Hayes, Joni Fazo, Therese Mitros, et al. 2012. “Phytozome: A Comparative Platform for Green Plant Genomics.” *Nucleic Acids Research* 40 (Database issue): D1178-86.

Granett, J., M. A. Walker, L. Kocsis, and A. D. Omer. 2001. “Biology and Management of Grape Phylloxera.” *Annual Review of Entomology* 46 (1): 387–412.

Guest, David. 2007. “Black Pod: Diverse Pathogens with a Global Impact on Cocoa Yield.” *Phytopathology* 97 (12): 1650–53.

Guo, Hongwei, and Joseph R. Ecker. 2004. “The Ethylene Signaling Pathway: New Insights.” *Current Opinion in Plant Biology* 7 (1): 40–49.

Guo, Yufang, Monique L. Sakalidis, Gabriel Andres Torres-Londono, and Mary K. Hausbeck. 2021. “Population Structure of a Worldwide Phytophthora Palmivora Collection Suggests Lack of Host Specificity and Reduced Genetic Diversity in South America and the Caribbean.” *Plant Disease*, no. PDIS-05-20-1055-RE (December): PDIS05201055RE.

Hämälä, Tuomas, Mark J. Guiltinan, James H. Marden, Siela N. Maximova, Claude W. dePamphilis, and Peter Tiffin. 2020. “Gene Expression Modularity Reveals Footprints of Polygenic Adaptation in Theobroma Cacao.” *Molecular Biology and Evolution* 37 (1): 110–23.

Hämälä, Tuomas, Eric K. Wafula, Mark J. Guiltinan, Paula E. Ralph, Claude W. dePamphilis, and Peter Tiffin. 2021. “Genomic Structural Variants Constrain and Facilitate Adaptation in Natural Populations of Theobroma Cacao, the Chocolate Tree.” *Proceedings of the National Academy of Sciences of the United States of America* 118 (35): e2102914118.

Hartl, Daniel L., and Andrew G. Clark. 2007. *Principles of Population Genetics*. 4th ed. New York, NY: Oxford University Press.

Hatsugai, Noriyuki, Daisuke Igarashi, Keisuke Mase, You Lu, Yayoi Tsuda, Suma Chakravarthy, Hai-Lei Wei, et al. 2017. “A Plant Effector-Triggered Immunity Signaling Sector Is Inhibited by Pattern-Triggered Immunity.” *The EMBO Journal* 36 (18): 2758–69.

Hou, Shuguo, Zunyong Liu, Hexi Shen, and Daoji Wu. 2019. “Damage-Associated Molecular Pattern-Triggered Immunity in Plants.” *Frontiers in Plant Science* 10 (May): 646.

Hubert, Nicolas, Fabrice Duponchelle, Jesus Nuñez, Carmen Garcia-Davila, Didier Paugy, and Jean-François Renno. 2007. “Phylogeography of the Piranha Genera Serrasalmus and Pygocentrus: Implications for the Diversification of the Neotropical Ichthyofauna.” *Molecular Ecology* 16 (10): 2115–36.

Jacquot, E., L. S. Hagen, P. Michler, O. Rohfritsch, C. Stussi-Garaud, M. Keller, M. Jacquemond, and P. Yot. 1999. “In Situ Localization of Cacao Swollen Shoot Virus in Agroinfected Theobroma Cacao.” *Archives of Virology* 144 (2): 259–71.

Jha, Priyanka, and Vijay Kumar. 2018. “BABY BOOM (BBM): A Candidate Transcription Factor Gene in Plant Biotechnology.” *Biotechnology Letters* 40 (11–12): 1467–75.

Jiao, Yuannian, Norman J. Wickett, Saravanaraj Ayyampalayam, André S. Chanderbali, Lena Landherr, Paula E. Ralph, Lynn P. Tomsho, et al. 2011. “Ancestral Polyploidy in Seed Plants and Angiosperms.” *Nature* 473 (7345): 97–100.

Johal, G. S., and S. P. Briggs. 1992. “Reductase Activity Encoded by the HM1 Disease Resistance Gene in Maize.” *Science (New York, N.Y.)* 258 (5084): 985–87.

Johnson, Jason A., Thomas C. Harrington, and C. J. B. Engelbrecht. 2005. “Phylogeny and Taxonomy of the North American Clade of the Ceratocystis Fimbriata Complex.” *Mycologia* 97 (5): 1067–92.

Johnson, William Henry. 1912. *Cocoa, Its Cultivation and Preparation*. London,: Murray,.

Jones, Jonathan D. G., and Jeffery L. Dangl. 2006. “The Plant Immune System.” *Nature* 444 (7117): 323–29.

Kim, Seungill, Jieun Park, Seon-In Yeom, Yong-Min Kim, Eunyoung Seo, Ki-Tae Kim, Myung-Shin Kim, et al. 2017. “New Reference Genome Sequences of Hot Pepper Reveal the Massive Evolution of Plant Disease-Resistance Genes by Retroduplication.” *Genome Biology* 18 (1): 210.

Kimber, C. 2006. *Martinique Revisited*. College Station, TX: Texas A & M University Press.

Knollenberg, Benjamin J., Guo-Xing Li, Joshua D. Lambert, Siela N. Maximova, and Mark J. Guiltinan. 2020. “Clovamide, a Hydroxycinnamic Acid Amide, Is a Resistance Factor Against Phytophthora Spp. in Theobroma Cacao.” *Frontiers in Plant Science* 11 (December): 617520.

Kourelis, Jiorgos, and Renier A. L. van der Hoorn. 2018. “Defended to the Nines: 25 Years of Resistance Gene Cloning Identifies Nine Mechanisms for R Protein Function.” *The Plant Cell* 30 (2): 285–99.

Krasileva, Ksenia V. 2019. “The Role of Transposable Elements and DNA Damage Repair Mechanisms in Gene Duplications and Gene Fusions in Plant Genomes.” *Current Opinion in Plant Biology* 48 (April): 18–25.

Lanaud, C., O. Fouet, D. Clément, M. Boccara, A. M. Risterucci, S. Surujdeo-Maharaj, T. Legavre, and X. Argout. 2009. “A Meta–QTL Analysis of Disease Resistance Traits of Theobroma Cacao L.” *Molecular Breeding: New Strategies in Plant Improvement* 24 (4): 361–74.

Le Roux, Clémentine, Gaëlle Huet, Alain Jauneau, Laurent Camborde, Dominique Trémousaygue, Alexandra Kraut, Binbin Zhou, et al. 2015. “A Receptor Pair with an Integrated Decoy Converts Pathogen Disabling of Transcription Factors to Immunity.” *Cell* 161 (5): 1074–88.

Leister, Dario. 2004. “Tandem and Segmental Gene Duplication and Recombination in the Evolution of Plant Disease Resistance Gene.” *Trends in Genetics: TIG* 20 (3): 116–22.

Levinson, G., and G. A. Gutman. 1987. “Slipped-Strand Mispairing: A Major Mechanism for DNA Sequence Evolution.” *Molecular Biology and Evolution* 4 (3): 203–21.

Lewis, Jennifer D., Amy Huei-Yi Lee, Jana A. Hassan, Janet Wan, Brenden Hurley, Jacquelyn R. Jhingree, Pauline W. Wang, et al. 2013. “The Arabidopsis ZED1 Pseudokinase Is Required for ZAR1-Mediated Immunity Induced by the Pseudomonas Syringae Type III Effector HopZ1a.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (46): 18722–27.

Li, Fuguang, Guangyi Fan, Cairui Lu, Guanghui Xiao, Changsong Zou, Russell J. Kohel, Zhiying Ma, et al. 2015. “Genome Sequence of Cultivated Upland Cotton (Gossypium Hirsutum TM-1) Provides Insights into Genome Evolution.” *Nature Biotechnology* 33 (5): 524–30.

Li, Yiyuan, Jianhui Xiao, Jiajie Wu, Jialei Duan, Yue Liu, Xingguo Ye, Xin Zhang, et al. 2012. "A Tandem Segmental Duplication (TSD) in Green Revolution Gene Rht-D1b Region Underlies Plant Height Variation." *The New Phytologist* 196 (1): 282–91.

Liu, Lijing, Fathi-Mohamed Sonbol, Bethany Huot, Yangnan Gu, John Withers, Musoki Mwimba, Jian Yao, Sheng Yang He, and Xinnian Dong. 2016. "Salicylic Acid Receptors Activate Jasmonic Acid Signalling through a Non-Canonical Pathway to Promote Effector-Triggered Immunity." *Nature Communications* 7 (October): 13099.

Liu, Yi, Zi Shi, Siela Maximova, Mark J. Payne, and Mark J. Guiltinan. 2013. "Proanthocyanidin Synthesis in Theobroma Cacao: Genes Encoding Anthocyanidin Synthase, Anthocyanidin Reductase, and Leucoanthocyanidin Reductase." *BMC Plant Biology* 13 (1): 202.

Livingstone, Donald, 3rd, Conrad Stack, Guiliana M. Mustiga, Dayana C. Rodezno, Carmen Suarez, Freddy Amores, Frank A. Feltus, Keithanne Mockaitis, Omar E. Cornejo, and Juan C. Motamayor. 2017. "A Larger Chocolate Chip-Development of a 15K Theobroma Cacao L. Snp Array to Create High-Density Linkage Maps." *Frontiers in Plant Science* 8 (December): 2008.

Loon, L. C. van, M. Rep, and C. M. J. Pieterse. 2006. "Significance of Inducible Defense-Related Proteins in Infected Plants." *Annual Review of Phytopathology* 44 (1): 135–62.

Lorin, Thibault, Frédéric G. Brunet, Vincent Laudet, and Jean-Nicolas Volff. 2018. "Teleost Fish-Specific Preferential Retention of Pigmentation Gene-Containing Families after Whole Genome Duplications in Vertebrates." *G3 (Bethesda, Md.)* 8 (5): 1795–1806.

Lu, You, and Kenichi Tsuda. 2021. "Intimate Association of PRR- and NLR-Mediated Signaling in Plant Immunity." *Molecular Plant-Microbe Interactions: MPMI* 34 (1): 3–14.

Luna, Estrella, Toby J. A. Bruce, Michael R. Roberts, Victor Flors, and Jurriaan Ton. 2012. "Next-Generation Systemic Acquired Resistance." *Plant Physiology* 158 (2): 844–53.

Ma, Wenbo, Frederick F. T. Dong, John Stavrinides, and David S. Guttman. 2006. “Type III Effector Diversification via Both Pathoadaptation and Horizontal Transfer in Response to a Coevolutionary Arms Race.” *PLoS Genetics* 2 (12): e209.

Martinson, Veronica A. 1966. “Hybridization of Cacao and Theobroma Grandiflora.” *The Journal of Heredity* 57 (4): 134–36.

Maximova, Siela N., Jean-Philippe Marelli, Ann Young, Sharon Pishak, Joseph A. Verica, and Mark J. Guiltinan. 2006. “Over-Expression of a Cacao Class I Chitinase Gene in Theobroma Cacao L. Enhances Resistance against the Pathogen, Colletotrichum Gloeosporioides.” *Planta* 224 (4): 740–49.

Mazón, Marina, Francisco Díaz, and Juan C. Gaviria. 2013. “Effectiveness of Different Trap Types for Control of Bark and Ambrosia Beetles (Scolytinae) in Criollo Cacao Farms of Mérida, Venezuela.” *International Journal of Pest Management* 59 (3): 189–96.

Meinhardt, Lyndel W., Johana Rincones, Bryan A. Bailey, M. Catherine Aime, Gareth W. Griffith, Dapeng Zhang, and Gonçalo A. G. Pereira. 2008. “Moniliophthora Perniciosa, the Causal Agent of Witches’ Broom Disease of Cacao: What’s New from This Old Foe?” *Molecular Plant Pathology* 9 (5): 577–88.

Melnick, Rachel L., Jean-Philippe Marelli, Richard C. Sicher, Mary D. Strem, and Bryan A. Bailey. 2012. “The Interaction of Theobroma Cacao and Moniliophthora Perniciosa, the Causal Agent of Witches’ Broom Disease, during Parthenocarpy.” *Tree Genetics & Genomes* 8 (6): 1261–79.

Meyers, B. C., D. B. Chin, K. A. Shen, S. Sivaramakrishnan, D. O. Lavelle, Z. Zhang, and R. W. Michelmore. 1998. “The Major Resistance Gene Cluster in Lettuce Is Highly Duplicated and Spans Several Megabases.” *The Plant Cell* 10 (11): 1817–32.

Meyers, Blake C., Alexander Kozik, Alyssa Griego, Hanhui Kuang, and Richard W. Michelmore. 2003. “Genome-Wide Analysis of NBS-LRR-Encoding Genes in Arabidopsis.” *The Plant Cell* 15 (4): 809–34.

Michael, Todd P. 2014. “Plant Genome Size Variation: Bloating and Purging DNA.” *Briefings in Functional Genomics* 13 (4): 308–17.

Michael, Todd P., and Robert VanBuren. 2020. “Building Near-Complete Plant Genomes.” *Current Opinion in Plant Biology* 54 (April): 26–33.

Michelmore, R. W., and B. C. Meyers. 1998. “Clusters of Resistance Genes in Plants Evolve by Divergent Selection and a Birth-and-Death Process.” *Genome Research* 8 (11): 1113–30.

Mondego, Jorge M. C., Marcelo F. Carazzolle, Gustavo G. L. Costa, Eduardo F. Formighieri, Lucas P. Parizzi, Johana Rincones, Carolina Cotomacci, et al. 2008. “A Genome Survey of Moniliophthora Perniciosa Gives New Insights into Witches’ Broom Disease of Cacao.” *BMC Genomics* 9 (1): 548.

Morales-Cruz, Abraham, Shahin S. Ali, Andrea Minio, Rosa Figueroa-Balderas, Jadran F. García, Takao Kasuga, Alina S. Puig, Jean-Philippe Marelli, Bryan A. Bailey, and Dario Cantu. 2020. “Independent Whole-Genome Duplications Define the Architecture of the Genomes of the Devastating West African Cacao Black Pod Pathogen Phytophthora Megakarya and Its Close Relative Phytophthora Palmivora.” *G3 (Bethesda, Md.)* 10 (7): 2241–55.

Motamayor, J. C., A. M. Risterucci, P. A. Lopez, C. F. Ortiz, A. Moreno, and C. Lanaud. 2002. “Cacao Domestication I: The Origin of the Cacao Cultivated by the Mayas.” *Heredity* 89 (5): 380–86.

Motamayor, Juan C., Philippe Lachenaud, Jay Wallace da Silva e Mota, Rey Loor, David N. Kuhn, J. Steven Brown, and Raymond J. Schnell. 2008. “Geographic and Genetic

Population Differentiation of the Amazonian Chocolate Tree (Theobroma Cacao L).” *PloS One* 3 (10): e3311.

Motamayor, Juan C., Keithanne Mockaitis, Jeremy Schmutz, Niina Haiminen, Donald Livingstone 3rd, Omar Cornejo, Seth D. Findley, et al. 2013. “The Genome Sequence of the Most Widely Cultivated Cacao Type and Its Use to Identify Candidate Genes Regulating Pod Color.” *Genome Biology* 14 (6): r53.

Mukherjee, Sohini, Gladys Keitany, Yan Li, Yong Wang, Haydn L. Ball, Elizabeth J. Goldsmith, and Kim Orth. 2006. “Yersinia YopJ Acetylates and Inhibits Kinase Activation by Blocking Phosphorylation.” *Science (New York, N.Y.)* 312 (5777): 1211–14.

Muller, Emmanuelle. 2016. “Cacao Swollen Shoot Virus (CSSV): History, Biology, and Genome.” In *Cacao Diseases*, 337–58. Cham: Springer International Publishing.

Nevo, Eviatar, and Guoxiong Chen. 2010. “Drought and Salt Tolerances in Wild Relatives for Wheat and Barley Improvement.” *Plant, Cell & Environment* 33 (4): 670–85.

Ngou, Bruno Pok Man, Hee-Kyung Ahn, Pingtao Ding, and Jonathan D. G. Jones. 2021. “Mutual Potentiation of Plant Immunity by Cell-Surface and Intracellular Receptors.” *Nature* 592 (7852): 110–15.

Ngou, Bruno Pok Man, Jonathan D. G. Jones, and Pingtao Ding. 2021. “Plant Immune Networks.” *Trends in Plant Science*, September. https://doi.org/10.1016/j.tplants.2021.08.012.

Pare, P. W., and J. H. Tumlinson. 1999. “Plant Volatiles as a Defense against Insect Herbivores.” *Plant Physiology* 121 (2): 325–32.

Pellicer, Jaume, and Ilia J. Leitch. 2020. “The Plant DNA C-Values Database (Release 7.1): An Updated Online Repository of Plant Genome Size Data for Comparative Studies.” *The New Phytologist* 226 (2): 301–5.

Pemberton, Clare L., and George P. C. Salmond. 2004. “The Nep1-like Proteins-a Growing Family of Microbial Elicitors of Plant Necrosis.” *Molecular Plant Pathology* 5 (4): 353–59.

Phillips-Mora, W., M. C. Aime, and M. J. Wilkinson. 2007. “Biodiversity and Biogeography of the Cacao (Theobroma Cacao) Pathogen Moniliophthora Roreri in Tropical America.” *Plant Pathology* 56 (6): 911–22.

Phillips-Mora, W., and M. J. Wilkinson. 2007. “Frosty Pod of Cacao: A Disease with a Limited Geographic Range but Unlimited Potential for Damage.” *Phytopathology* 97 (12): 1644–47.

Ploetz, Randy C. 2007. “Cacao Diseases: Important Threats to Chocolate Production Worldwide.” *Phytopathology* 97 (12): 1634–39.

Pokou, Désiré N., Andrew S. Fister, Noah Winters, Mathias Tahi, Coulibaly Klotioloma, Aswathy Sebastian, James H. Marden, Siela N. Maximova, and Mark J. Guiltinan. 2019. “Resistant and Susceptible Cacao Genotypes Exhibit Defense Gene Polymorphism and Unique Early Responses to Phytophthora Megakarya Inoculation.” *Plant Molecular Biology* 99 (4–5): 499–516.

Pokou, N. D., J. A. K. N’Goran, Ph Lachenaud, A. B. Eskes, J. C. Montamayor, R. Schnell, M. Kolesnikova-Allen, D. Clément, and A. Sangaré. 2009. “Recurrent Selection of Cocoa Populations in Côte d’Ivoire: Comparative Genetic Diversity between the First and Second Cycles.” *Plant Breeding = Zeitschrift Fur Pflanzenzuchtung* 128 (5): 514–20.

Posnette, A. F., and N. F. Robertson. 1950. “Virus Diseases of Cacao in West Africa.” *The Annals of Applied Biology* 37 (3): 363–77.

Posnette, A. F., N. F. Robertson, and J. Mca Todd. 1950. “Virus Diseases of Cacao in West Africa.” *The Annals of Applied Biology* 37 (2): 229–40.

Proost, Sebastian, Pedro Pattyn, Tom Gerats, and Yves Van de Peer. 2011. “Journey through the Past: 150 Million Years of Plant Genome Evolution.” *The Plant Journal: For Cell and Molecular Biology* 66 (1): 58–65.

Risterucci, A. M., D. Paulin, M. Ducamp, J. A. K. N’Goran, and C. Lanaud. 2003. “Identification of QTLs Related to Cocoa Resistance to Three Species of Phytophthora.” *Theoretical and Applied Genetics* 108 (1): 168–74.

Rocha, Hermínio M. 1966. “La Importancia de Las Sustancias Polifenólicas En El Mecanismo Fisiológico de La Resistencia de Cacao (Theobroma Cacao L.) a Phytophthora Palmivora (Butl.) Butl.” IICA, Turrialba (Costa Rica).

Roivainen, Osmo. 1976. “Transmission of Cocoa Viruses by Mealybugs (Homoptera: Pseudococcidae).” *Agricultural and Food Science* 48 (3): 203–304.

Romero Navarro, J. Alberto, Wilbert Phillips-Mora, Adriana Arciniegas-Leal, Allan Mata-Quirós, Niina Haiminen, Guiliana Mustiga, Donald Livingstone Iii, et al. 2017. “Application of Genome Wide Association and Genomic Prediction for Improvement of Cacao Productivity and Resistance to Black and Frosty Pod Diseases.” *Frontiers in Plant Science* 8 (November): 1905.

Salse, Jérôme. 2012. “In Silico Archeogenomics Unveils Modern Plant Genome Organisation, Regulation and Evolution.” *Current Opinion in Plant Biology* 15 (2): 122–30.

Sarris, Panagiotis F., Volkan Cevik, Gulay Dagdas, Jonathan D. G. Jones, and Ksenia V. Krasileva. 2016. “Comparative Analysis of Plant Immune Receptor Architectures Uncovers Host Proteins Likely Targeted by Pathogens.” *BMC Biology* 14 (1): 8.

Sarris, Panagiotis F., Zane Duxbury, Sung Un Huh, Yan Ma, Cécile Segonzac, Jan Sklenar, Paul Derbyshire, et al. 2015. “A Plant Immune Receptor Detects Pathogen Effectors That Target WRKY Transcription Factors.” *Cell* 161 (5): 1089–1100.

Saunders, James, and Nichole O'neill. 2004. "The Characterization of Defense Responses to Fungal Infection in Alfalfa." *BioControl (Dordrecht, Netherlands)* 49 (6): 715–28.

Shephard, Charles. 2018. *An Historical Account of the Island of Saint Vincent*. Franklin Classics Trade Press.

Shi, Zi, Siela N. Maximova, Yi Liu, Joseph Verica, and Mark J. Guiltinan. 2010. "Functional Analysis of the Theobroma Cacao NPR1 Gene in Arabidopsis." *BMC Plant Biology* 10 (1): 248.

Shi, Zi, Yufan Zhang, Siela N. Maximova, and Mark J. Guiltinan. 2013. "TcNPR3 from Theobroma Cacao Functions as a Repressor of the Pathogen Defense Response." *BMC Plant Biology* 13 (1): 204.

Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. "BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs." *Bioinformatics (Oxford, England)* 31 (19): 3210–12.

Snyder-Leiby, T. E., and D. B. Furtek. 1995. "A Genomic Clone (Accession No. U30324) from Theobroma Cacao L. with High Similarity to Plant Class I Endochitinase Sequences." *Plant Physiology* 109: 338.

Stam, Remco, Daniela Scheikl, and Aurélien Tellier. 2017. "The Wild Tomato Species Solanum Chilense Shows Variation in Pathogen Resistance between Geographically Distinct Populations." *PeerJ* 5 (e2910): e2910.

Steppuhn, Anke, Klaus Gase, Bernd Krock, Rayko Halitschke, and Ian T. Baldwin. 2004. "Nicotine's Defensive Function in Nature." *PLoS Biology* 2 (8): E217.

Surujdeo-Maharaj, S., T. N. Sreenivasan, L. A. Motilal, and P. Umaharan. 2016. "Black Pod and Other Phytophthora Induced Diseases of Cacao: History, Biology, and Control." In *Cacao Diseases*, 213–66. Cham: Springer International Publishing.

Takai, Ryota, Akira Isogai, Seiji Takayama, and Fang-Sik Che. 2008. “Analysis of Flagellin Perception Mediated by Flg22 Receptor OsFLS2 in Rice.” *Molecular Plant-Microbe Interactions: MPMI* 21 (12): 1635–42.

Teixeira, Paulo José Pereira Lima, Daniela Paula de Toledo Thomazella, Osvaldo Reis, Paula Favoretti Vital do Prado, Maria Carolina Scatolin do Rio, Gabriel Lorencini Fiorin, Juliana José, et al. 2014. “High-Resolution Transcript Profiling of the Atypical Biotrophic Interaction between Theobroma Cacao and the Fungal Pathogen Moniliophthora Perniciosa.” *The Plant Cell* 26 (11): 4245–69.

Tenthorey, Jeannette L., Nicole Haloupek, José Ramón López-Blanco, Patricia Grob, Elise Adamson, Ella Hartenian, Nicholas A. Lind, et al. 2017. “The Structural Basis of Flagellin Detection by NAIP5: A Strategy to Limit Pathogen Immune Evasion.” *Science (New York, N.Y.)* 358 (6365): 888–93.

Theophrastus. 1989. *Enquiry into Plants: Bks. I-V v. 1*. Translated by A. F. Hort. Loeb Classical Library, No. 7. London, England: LOEB.

Thomas, Evert, Maarten van Zonneveld, Judy Loo, Toby Hodgkin, Gea Galluzzi, and Jacob van Etten. 2012. “Present Spatial Diversity Patterns of Theobroma Cacao L. in the Neotropics Reflect Genetic Differentiation in Pleistocene Refugia Followed by Human-Influenced Dispersal.” *PloS One* 7 (10): e47676.

Thresh, J. M., and G. K. Owusu. 1986. “The Control of Cocoa Swollen Shoot Disease in Ghana: An Evaluation of Eradication Procedures.” *Crop Protection (Guildford, Surrey)* 5 (1): 41–52.

Tigano, Anna, and Vicki L. Friesen. 2016. “Genomics of Local Adaptation with Gene Flow.” *Molecular Ecology* 25 (10): 2144–64.

Trdá, Lucie, Olivier Fernandez, Freddy Boutrot, Marie-Claire Héloir, Jani Kelloniemi, Xavier Daire, Marielle Adrian, et al. 2014. “The Grapevine Flagellin Receptor VvFLS2

Differentially Recognizes Flagellin-Derived Epitopes from the Endophytic Growth-Promoting Bacterium Burkholderia Phytofirmans and Plant Pathogenic Bacteria." *The New Phytologist* 201 (4): 1371–84.

Van de Weyer, Anna-Lena, Freddy Monteiro, Oliver J. Furzer, Marc T. Nishimura, Volkan Cevik, Kamil Witek, Jonathan D. G. Jones, Jeffery L. Dangl, Detlef Weigel, and Felix Bemm. 2019. "A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis Thaliana." *Cell* 178 (5): 1260-1272.e14.

Van Zandt, Peter A. 2007. "Plant Defense, Growth, and Habitat: A Comparative Assessment of Constitutive and Induced Resistance." *Ecology* 88 (8): 1984–93.

Vlot, A. Corina, D'maris Amick Dempsey, and Daniel F. Klessig. 2009. "Salicylic Acid, a Multifaceted Hormone to Combat Disease." *Annual Review of Phytopathology* 47 (1): 177–206.

Wang, Jianan, Michael D. Coffey, Nicola De Maio, and Erica M. Goss. 2020. "Repeated Global Migrations on Different Plant Hosts by the Tropical PathogenPhytophthora Palmivora." *BioRxiv*. bioRxiv. https://doi.org/10.1101/2020.05.13.093211.

Wang, Jizong, Meijuan Hu, Jia Wang, Jinfeng Qi, Zhifu Han, Guoxun Wang, Yijun Qi, Hong-Wei Wang, Jian-Min Zhou, and Jijie Chai. 2019. "Reconstitution and Structure of a Plant NLR Resistosome Conferring Immunity." *Science (New York, N.Y.)* 364 (6435): eaav5870.

Wang, Jizong, Jia Wang, Meijuan Hu, Shan Wu, Jinfeng Qi, Guoxun Wang, Zhifu Han, et al. 2019. "Ligand-Triggered Allosteric ADP Release Primes a Plant NLR Complex." *Science (New York, N.Y.)* 364 (6435): eaav5868.

Waterhouse, Robert M., Fredrik Tegenfeldt, Jia Li, Evgeny M. Zdobnov, and Evgenia V. Kriventseva. 2013. "OrthoDB: A Hierarchical Catalog of Animal, Fungal and Bacterial Orthologs." *Nucleic Acids Research* 41 (Database issue): D358-65.

Wesselingh, F. P., and J. A. Salo. 2006. “A Miocene Perspective on the Evolution of the Amazonian Biota.” *Scripta Geologica* 133: 439–58.

Widmer, Timothy L., and Nathalie Laurent. 2006. “Plant Extracts Containing Caffeic Acid and Rosmarinic Acid Inhibit Zoospore Germination of Phytophthora Spp. Pathogenic to Theobroma Cacao.” *European Journal of Plant Pathology* 115 (4): 377–88.

Willmann, Roland, Heini M. Lajunen, Gitte Erbs, Mari-Anne Newman, Dagmar Kolb, Kenichi Tsuda, Fumiaki Katagiri, et al. 2011. “Arabidopsis Lysin-Motif Proteins LYM1 LYM3 CERK1 Mediate Bacterial Peptidoglycan Sensing and Immunity to Bacterial Infection.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (49): 19824–29.

Wu, Yufeng, Zhengge Zhu, Ligeng Ma, and Mingsheng Chen. 2008. “The Preferential Retention of Starch Synthesis Genes Reveals the Impact of Whole-Genome Duplication on Grass Evolution.” *Molecular Biology and Evolution* 25 (6): 1003–6.

Yang, Zhenzhen, Eric K. Wafula, Loren A. Honaas, Huiting Zhang, Malay Das, Monica Fernandez-Aparicio, Kan Huang, et al. 2015. “Comparative Transcriptome Analyses Reveal Core Parasitism Genes and Suggest Gene Duplication and Repurposing as Sources of Structural Novelty.” *Molecular Biology and Evolution* 32 (3): 767–90.

Yuan, Minhang, Zeyu Jiang, Guozhi Bi, Kinya Nomura, Menghui Liu, Yiping Wang, Boying Cai, Jian-Min Zhou, Sheng Yang He, and Xiu-Fang Xin. 2021. “Pattern-Recognition Receptors Are Required for NLR-Mediated Plant Immunity.” *Nature* 592 (7852): 105–9.

Zarrillo, Sonia, Nilesh Gaikwad, Claire Lanaud, Terry Powis, Christopher Viot, Isabelle Lesur, Olivier Fouet, et al. 2018. “The Use and Domestication of Theobroma Cacao during the Mid-Holocene in the Upper Amazon.” *Nature Ecology & Evolution* 2 (12): 1879–88.

Zhang, Dapeng, Enrique Arevalo-Gardini, Sue Mischke, Luis Zúñiga-Cernades, Alejandro Barreto-Chavez, and Jorge Adriazola Del Aguila. 2006. “Genetic Diversity and Structure

of Managed and Semi-Natural Populations of Cocoa (Theobroma Cacao) in the Huallaga and Ucayali Valleys of Peru." *Annals of Botany* 98 (3): 647–55.

Zhang, Dapeng, Michel Boccara, Lambert Motilal, Sue Mischke, Elizabeth S. Johnson, David R. Butler, Bryan Bailey, and Lyndel Meinhardt. 2009. "Molecular Characterization of an Earliest Cacao (Theobroma Cacao L.) Collection from Upper Amazon Using Microsatellite DNA Markers." *Tree Genetics & Genomes* 5 (4): 595–607.

Zhang, Dapeng, Antonio Figueira, Lambert Motilal, Philippe Lachenaud, and Lyndel W. Meinhardt. 2011. "Theobroma." In *Wild Crop Relatives: Genomic and Breeding Resources*, 277–96. Berlin, Heidelberg: Springer Berlin Heidelberg.

Zhang, Dapeng, Windson July Martínez, Elizabeth S. Johnson, Eduardo Somarriba, Wilberth Phillips-Mora, Carlos Astorga, Sue Mischke, and Lyndel W. Meinhardt. 2012. "Genetic Diversity and Spatial Structure in a New Distinct Theobroma Cacao L. Population in Bolivia." *Genetic Resources and Crop Evolution* 59 (2): 239–52.

Zhang, Dapeng, and Lambert Motilal. 2016. "Origin, Dispersal, and Current Global Distribution of Cacao Genetic Diversity." In *Cacao Diseases*, 3–31. Cham: Springer International Publishing.

Zhang, Yu, Rui Xia, Hanhui Kuang, and Blake C. Meyers. 2016. "The Diversification of Plant NBS-LRR Defense Genes Directs the Evolution of MicroRNAs That Target Them." *Molecular Biology and Evolution* 33 (10): 2692–2705.

Zhou, Huanbin, Jian Lin, Aimee Johnson, Robyn L. Morgan, Wenwan Zhong, and Wenbo Ma. 2011. "Pseudomonas Syringae Type III Effector HopZ1 Targets a Host Enzyme to Suppress Isoflavone Biosynthesis and Promote Infection in Soybean." *Cell Host & Microbe* 9 (3): 177–86.

Zipfel, Cyril. 2014. "Plant Pattern-Recognition Receptors." *Trends in Immunology* 35 (7): 345–51.

Chapter 2: A Combination of Conserved and Diverged Responses Underlie *Theobroma cacao*'s Defense Response to *Phytophthora palmivora*

Abstract

Plants have complex and dynamic immune systems that have evolved over millennia to help them resist pathogen invasion. Humans have worked to incorporate these evolved defenses into crops through breeding. However, many crop cultivars only harness a fraction of the overall genetic diversity available to a given species, or have such a long history of domestication that most diversity has been lost. Evaluating previously neglected germplasm for desirable traits, such as disease resistance, is therefore an essential step towards breeding crops that are adapted to both current and emerging threats. Here, we examine the evolution of defense response across four populations of *Theobroma cacao* L., with the goal of identifying genetic elements essential for protection against the oomycete pathogen *Phytophthora palmivora*. By combining data from RNA-sequencing, un-targeted metabolomics, and whole genome sequencing we have discovered several genes and pathways associated with resistance, primarily within rather than between populations. Among these processes is phenylpropanoid biosynthesis, a metabolic pathway with well-documented roles in plant defense. One of the genes in this pathway, caffeoyl-shikimate esterase (CSE), was up-regulated across all four populations, indicating its broad importance for cacao's defense response. Further experimental evidence suggests this gene synthesizes the antimicrobial compound caffeic acid, a known inhibitor of *Phytophthora spp.* Together, our results indicate most expression variation associated with resistance is unique to populations. Moreover, they suggest using a small subset of clones to determine the basis of resistance to *P. palmivora,* as has been done in breeding programs for over five decades, provides limited power for discovering potentially useful genetic variation.

Introduction

For thousands of years humans have worked to incorporate a wide variety of desirable traits into crops through breeding. This process has led to an erosion of genetic diversity through artificial selection that can be detrimental to further crop improvement (Shahin S. Ali, Shao, Lary, Kronmiller, et al. 2017; Morales-Cruz et al. 2020; S. S. Ali et al. 2016), and raises the strong possibility that standing genetic variation in wild ancestors could be a rich source of new alleles (Kremling et al., 2018; Troyer, 1990; Zhao et al., 2018).

Harnessing genetic diversity from wild populations is a particularly attractive possibility for genetic variation affecting pathogen resistance. This is because large population size and balancing (diversifying) selection often maintains genetic variation at loci that are co-evolving with locally abundant pathogens (Stam, Silva-Arias, and Tellier 2019; Stahl et al. 1999; Koenig et al. 2019; Bellis et al. 2020). When populations are spread across broad geographic areas, this co-evolution creates a rich tapestry of alleles conferring resistance to a diverse set of microbes. Evaluating previously neglected germplasm from wild populations for desirable traits, such as disease resistance, is therefore an essential step toward breeding crops that are adapted to both current and emerging threats. In this study, we examine whether genotypes from wild populations of the tree crop *Theobroma cacao* L. can be used to efficiently identify genes conferring resistance to the oomycete pathogen *Phytophthora palmivora*.

Theobroma cacao L., the tree from which chocolate is derived, is a tropical understory plant native to the Amazon basin (Harry C. Evans 2016a; B. A. Bailey and Meinhardt 2018). Cocoa and cocoa butter, the products created by fermenting, drying, and processing cacao seeds, form the basis of a chocolate and confectionary market worth approximately $100 billion (B. A. Bailey and Meinhardt 2018; Ploetz 2007). Cacao is distributed across ten strongly-differentiated populations that are hypothesized to have evolved via ancient ridgelines, glacial refugia, and/or

human management (Juan C. Motamayor et al. 2008; Cornejo et al. 2018; B. A. Bailey and Meinhardt 2018). While there is some evidence for domestication and human-induced genetic bottlenecks, most cacao germplasm is thought to exist in its ancestral state (Cornejo et al. 2018; Juan C. Motamayor et al. 2008).

Annual yield loss in cacao is caused by a variety of pests and pathogens, the worst of which is black pod rot (Evans 2016a). Black pod rot is caused by four *Phytophthora* species and alone accounts for over 10% of pre-harvest yield loss (Ploetz 2007). The two most damaging members of this quartet are *Phytophthora megakarya* and its sister species *Phytophthora palmivora* (Shahin S. Ali, Shao, Lary, Kronmiller, et al. 2017; Morales-Cruz et al. 2020; S. S. Ali et al. 2016). Native to southeast Asia, *P. palmivora* is a generalist pathogen that causes extensive yield loss to a range of hosts, including cacao, oil palm, and papaya (Gumtow, Wu, Uchida, & Tian, 2018; Mchau & Coffey, 1994; Torres et al., 2016). Breeding programs for tree crops like cacao are extremely difficult and time consuming, taking one to two decades or more to produce commercially viable clones. Moreover, small mapping populations and, until recently, low marker density make identification of quantitative trait loci (QTL) difficult, identifying large genomic regions home to hundreds or even thousands of genes (Gutiérrez et al. 2021; Livingstone et al. 2017; Lanaud et al. 2009).

Despite these difficulties, several breeding programs have successfully generated high yielding clones with partial resistance to black pod rot (Boza et al., 2014; Gutiérrez et al., 2021). These programs, while successful, have been centered around a small number of resistant genotypes collected in the 1930s. Most alleles conferring resistance to black pod rot are, therefore, derived from a very small set of parents. Such limited diversity leaves clones predisposed to breakthrough infections by rapidly evolving pathogens (Badet & Croll, 2020). Thus, producing clones durably resistant to pathogen challenge requires consideration of the genetic diversity that exists in cacao's many wild populations (B. A. Bailey and Meinhardt 2018).

Here, we test the hypothesis that wild populations represent diverse and potentially valuable sources of genetic variation by examining defense responses across four populations of *Theobroma cacao* L. Through the use of genomic, transcriptomic, and metabolomic data, collected in a unified experimental design, we identify both conserved and divergent aspects of cacao's defense response. Our results indicate that wild populations of crop species offer much more genetic diversity than any single individual or narrowly selected set of genotypes and can thus provide a diverse array of novel alleles for crop improvement.

Results

Cacao genotypes and populations sampled for this study

We selected 31 cacao genotypes based on previously described levels of resistance to the black pod rot pathogen *Phytophthora palmivora* (Fister et al. 2020). Each sampled genotype belongs to one of four populations named for their original geographic location (Juan C. Motamayor et al. 2008): Guiana, Iquitos, Marañón, or Nanay (Figure 2-1A). From each population, we chose the four most resistant and four most susceptible individuals for further experimentation, with the exception of Nanay, from which we sampled only 3 resistant genotypes. Cacao is native to the Amazon basin and it is there that most of its genetic diversity, including the populations used in this study, can be found (B. A. Bailey and Meinhardt 2018; Juan C. Motamayor et al. 2008). The range of these four populations extends from Peru and Ecuador in the west, to Brazil and Guiana in the east (Figure 1B) (Cornejo et al. 2018). Cacao is an understory tree that produces small, delicate flowers primarily pollinated by Ceratopogonid midges (Arnold et al. 2019). The fruits that form after successful pollination events are large, oblong, and dispersed locally by mammals (Hockings, Yamakoshi, & Matsuzawa, 2017). The

combined effect of this limited pollen and fruit dispersal is the formation of local populations with limited gene flow between them. Previous work on these genotypes show that this limited gene flow is partially reflected in highly divergent (Hartl and Clark 2007) F_{ST} estimates (Figure 1C) (Hämälä et al. 2020), and may have led to the formation of locally adaptive genetic variation (Hämälä et al. 2021).

To investigate how this divergence among populations affects the evolution of cacao's defense response, and to discover potentially novel mechanisms underlying tolerance to *P. palmivora*, we performed an RNA-seq experiment. The previously mentioned 31 genotypes were first imported as grafted plants from the ex situ germplasm collection at the Tropical Agricultural Research and Higher Education Center (CATIE). Approximately 300 rooted cuttings were taken from these grafted plants, of which 141 individual plants representing 27 genotypes survived. To minimize the effects of greenhouse gradients in temperature, humidity and other abiotic factors, 6 week old plants were distributed across the greenhouse in a pseudo-randomized block design as described in the Materials and Methods (*Transcriptome experimental design and treatment*). We challenged individual leaves on each plant with multiple plugs of *P. palmivora* mycelia or mock inoculant and collected samples 8 hours post inoculation (hpi). We chose 8 hours because it provided an estimation of early defense response but, according to preliminary evidence, was still late enough to observe transcriptional changes in specific defense-associated genes.

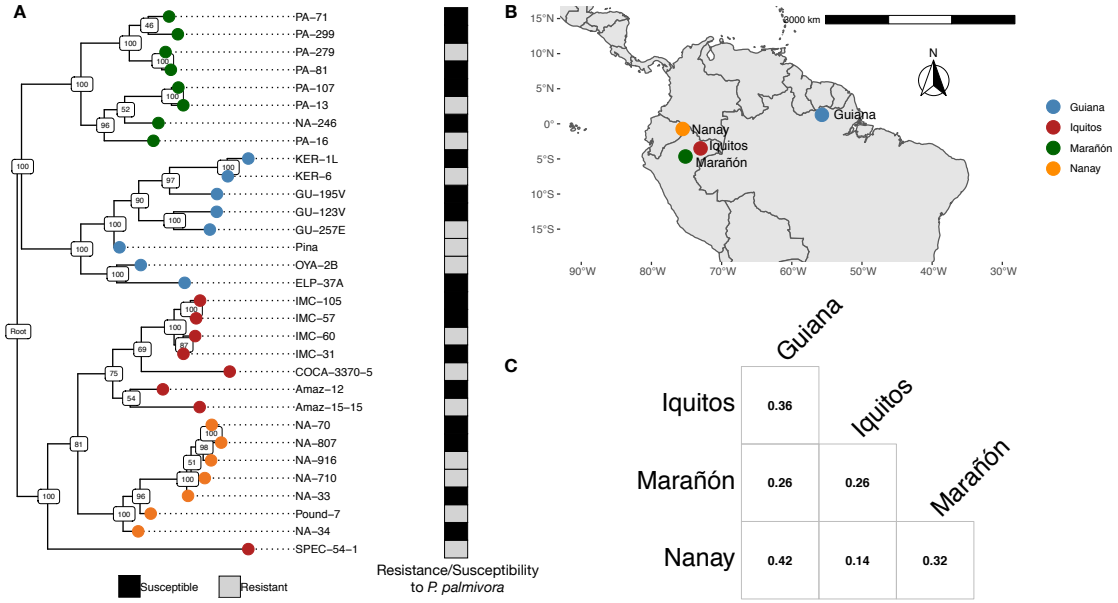


Figure 2-1: **Overview of cacao genotypes and populations sampled for this study**. (A) Maximum likelihood phylogeny of *T. cacao* genotypes based on 23,439 SNPs. White and gray boxes beside the phylogeny indicate whether genotypes were considered resistant (grey) or susceptible (black) to *P. palmivora* according to Fister et al. 2020. Numbers on the nodes indicate bootstrap support and colors at the tips indicate population membership: Guiana (blue), Iquitos (red), Marañón (green), and Nanay (orange). (B) Map displaying approximate center of origin for each of the four populations sampled for this study. (C) Pairwise F_{ST} estimates for each population.

Different sets of genes are responsible for defense against P. palmivora across all four populations

The 141 samples we sequenced had an average of 8 million QuantSeq reads per library, of which approximately 80% mapped to SCA-6. Because 3' tagging methods like QuantSeq produce a single read per transcript, less than a third of the reads normally necessary for traditional RNA-seq are required (Corley, Troy, Bosco, & Wilkins, 2019). Thus, even low coverage QuantSeq libraries can capture moderately expressed genes (Hua et al., 2021). After testing for differential expression using DESeq2, we chose the top 1000 genes ranked by absolute

log_2 fold change (LFC) to analyze further. We examined two types of transcriptional response, hereafter referred to as our main effects: response to pathogen treatment and differences between resistant/susceptible (R/S) phenotypes. We also examined the combined response of both treatment and phenotype (additive effects). Treatment X phenotype interaction effects were weak and rare across all populations (total N = 37 at FDR-adjusted p-value < 0.05) and were therefore omitted from the final analysis. For each of our main effects, we started by examining the proportion of differentially expressed genes that were shared across populations. Of the 1000 genes chosen from each population, over 40% were unique, i.e. not shared with any other populations (Treatment: $Mean_{\%\ unique} = 41.9$, $SEM_{\%\ unique} = 0.7$; R/S Phenotype: $Mean_{\%\ unique} = 43.7$, $SEM_{\%\ unique} = 0.8$; Figure 2-2A). Moreover, not only were many of the genes from each population unique, LFC correlations among all genes examined in this study (approximately 17k) revealed inconsistent responses (Figure 2-2B). This reveals that genes across all four populations responded differently to both pathogen challenge and R/S phenotype.

We chose an arbitrary LFC cutoff, rather than one based on p-values after multiple test correction, because limitations in sample size and intra-population variation resulted in a loss of statistical power at the group level. To further verify that our LFC cutoff did not bias interpretation of the results, we performed the same analysis on two different subsets of our data. First, we examined the effect of using a traditional, adjusted p-value cutoff (Benjamini and Hochberg 1995). We observed a larger proportion of genes that were unique to each population, for both pathogen treatment ($Mean_{\%\ unique} = 55.3$, $SEM_{\%\ unique} = 14.9$) and R/S phenotype ($Mean_{\%\ unique} = 68.6$, $SEM_{\%\ unique} = 7.3$). Second, we examined the effect of using different sized gene set cutoffs, ranging from 200 to 2000 genes. For each sample size, the proportion of genes that were unique to each population was between 30-40%, and was significantly lower than if the genes were selected at random (Supplemental Figure S2-1; one-way ANOVA, Proportion Unique Genes ~ Sample Size + Subsample Method + Sample Size x Sample Method: p-values < 0.001).

Hence, the high degree of population uniqueness was not due to size of the subset or chance detection of lowly expressed genes. Therefore, we hereafter refer to the top 1000 genes from each population as differentially expressed.

Recent gene duplications can result in highly similar, and likely redundant, copies of the same gene. If each population is using different, yet closely related and functionally redundant genes to respond to *P. palmivora,* our observation that population responses were largely non-overlapping may be inflated. To test whether closely related genes were behaving similarly across populations, we clustered paralogs using a 95% identity cutoff. We then calculated the proportion of paralogous clusters that were unique to a given population or shared across populations. For both the pathogen treatment and R/S phenotype main effects, the majority of differentially expressed genes in each population had no close paralogs. This resulted in patterns very similar to those in Figure 1 (Treatment: $Mean_{\%\ unique} = 40.1$, $SEM_{\%\ unique} = 0.6$; Phenotype: $Mean_{\%\ unique} = 41.9$, $SEM_{\%\ unique} = 1.0$; Supplemental Figure S2-2). Therefore, differences in differentially expressed genes among populations for both the *P. palmivora* treatment and R/S phenotype did not seem to be inflated by the differential expression of closely related paralogs.

To investigate the potential for functional redundancy in less closely related paralogs, we classified genes into orthogroups, i.e. narrowly defined protein families inferred to have a single ancestral gene among the species we were comparing (Wall et al. 2008; Emms and Kelly 2019). We then calculated the proportion of differentially expressed orthogroups that were unique to each population versus shared across populations (Supplemental Figure S2-3). To be considered differentially expressed, orthogroups only needed to contain a single differentially expressed gene from one of our four populations. We found a greater degree of shared orthogroups than we did individual genes ($Mean_{\%\ unique} = 28.9$, $SEM_{\%\ unique} = 1.2$; t-test, p-value < 0.001) and R/S phenotype ($Mean_{\%\ unique} = 32.3$, $SEM_{\%\ unique} = 2.6$; t-test, p-value < 0.05) main effects. Average LFC among orthogroups, however, was again weakly correlated across populations

(Supplemental Figure S2-4). Thus, each population used different, but often evolutionarily related genes to respond to *P. palmivora.*

There have been very few studies that examine stress response across many genotypes from multiple populations, making it difficult to compare cacao's divergent gene expression response with other plant species. However, our result is in stark contrast with at least one recent study in *Arabidopsis*, wherein the evolution of immunity-related gene expression was tested by treating *A. thaliana* and its close relatives with the microbial elicitor flg22. Of the genes differentially expressed in response to flg22, the proportion of 1:1 orthologs unique to each species was approximately 20 – 31% (Winkelmüller et al., 2021). When their focus was limited to solely *A. thaliana* genotypes, the proportion of genes private to each genotype decreased even further, falling to approximately 3.5 – 12.5%. Moreover, average LFC correlations between differentially expressed 1:1 orthologs, both between species and within species, were considerably higher than we observed among cacao populations (between species, $Mean_{correlation}$ = 72.8, $SEM_{correlation}$= 0.9; within species, $Mean_{correlation}$ = 87.7, $SEM_{correlation}$ = 3.9; Supplemental Figure S2-5). These results suggest that differences in effective population size, generation time, and population connectivity can give rise to taxa with diverse evolutionary histories and function, and that *A. thaliana*, while an excellent model for many questions, does not necessarily predict responses across species.

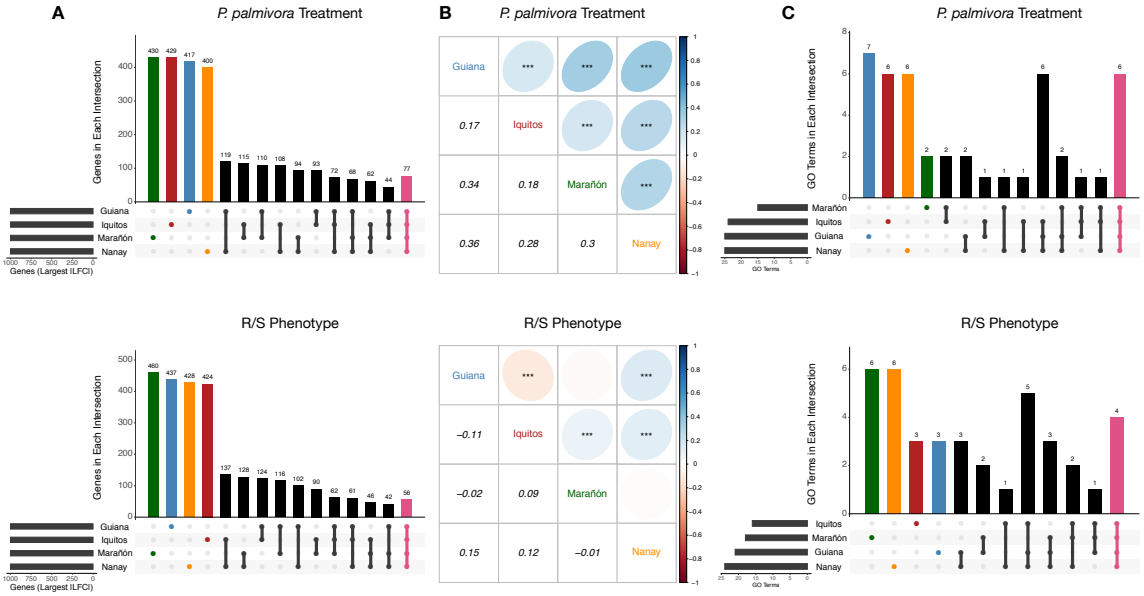


Figure 2-2: **Different sets of genes are responsible for defense against *P. palmivora* across all four populations.** (A) Overlap of differentially expressed genes for *P. palmivora* treatment versus control (top) and between resistant versus susceptible genotypes (bottom). The blue, red, green, and orange bars represent genes that are only differentially expressed in Guiana Iquitos, Marañón, or Nanay, respectively. The pink bar indicates genes that are differentially expressed across all four populations. Numbers above the bars indicate the number of differentially expressed genes in that specific intersection. (B) Pairwise Spearman correlations of $\log_2$ fold changes for all genes investigated in this study, for *P. palmivora* treatment versus control (top) and between resistant versus susceptible genotypes (bottom). The bottom triangle is the Spearman correlation coefficient. The top triangle is the correlation coefficient depicted as an ellipse, the shape of which depends on the size of the coefficient. Stars indicate statistical significance ($p < 0.001$), tested using Spearman's rho. (C) Overlap of enriched GO terms (Fisher's exact test: FDR-adjusted p-value < 0.05) for *P. palmivora* treatment versus control (top) and resistant versus susceptible genotypes (bottom). The blue, red, green, and orange bars represent GO terms that are only enriched in Guiana Iquitos, Marañón, or Nanay, respectively. The pink bar indicates GO terms that are significantly enriched across all four populations. Numbers above the bars indicate the number of enriched GO terms in that specific intersection.

Common functional groups underlie different sets of pathogen responsive genes

Because genes often function in redundant and overlapping ways within networks, a large number of genes unique to each population does not preclude overlapping functional responses.

We compared functional similarity among our differentially expressed genes, either in response to pathogen challenge or between R/S phenotype, using gene ontology (GO) terms (Figure 2-3A and B). There were more shared GO terms than individual genes (Treatment: $Mean_{\%\ unique} = 22.6$, $SEM_{\%\ unique} = 3.2$; Phenotype: $Mean_{\%\ unique} = 22.8$, $SEM_{\%\ unique} = 4.1$; Figure 2-2C), suggesting many of the defense-related genes in each population, while often unique, underly shared functional responses. Even the GO terms that were unique to each population, however, often shared similarity, e.g. “response to auxin” and “auxin homeostasis”. While we tried to reduce redundant GO terms by exploiting the parent-child structure of GO directed acyclic graphs, some partially overlapping terms remained. Thus, the proportion of functional categories that were private to each population was likely lower than estimated above.

The list of GO terms significantly enriched across all populations presumably represents a ‘core’ defense response that contains well-known defense-related processes. For the pathogen treatment main effect, these included “response to molecule of fungal origin” (GO:0002238), “induced systemic resistance” (GO:0009682), “response to gibberellin” (GO:0009739), “lignin biosynthetic process” (GO:0009809), “plant-type secondary cell wall biogenesis” (GO:0009834), and “response to cadmium ion” (GO:0046686). For R/S phenotype main effect, we saw “response to molecule of fungal origin” (GO:0002238), “response to insect” (GO:0009625), “response to jasmonic acid” (GO0009753), and “cinnamic acid biosynthetic process” (GO:0009800). Even within these core GO terms, however, 30-40% of the genes responding in each population were unique (Figure 2-3C). This mirrors the pattern observed when examining all differentially expressed genes (Figure 2-2A). Thus, even within this small, conserved subset of cacao’s defense response, there were many genes within each population that were responding uniquely.

The genes shared across cacao populations included a cast of well-known defense-mediators. Those responding to pathogen treatment across all four populations included multiple WRKY transcription factors (Bhattarai, Atamian, Kaloshian, & Eulgem, 2010; Mukhtar,

Deslandes, Auriac, Marco, & Somssich, 2008), as well as chitinase and endochitinase proteins (Y. J. Zhu et al. 2003; Siela N. Maximova et al. 2006). Less well-known, but strongly upregulated, defense mediators included Gretchen Hagen3 (GH3) and multiple berberine bridge enzymes (BBE) (X. Zou et al. 2019; Benedetti et al. 2018; Rodrigues Oblessuc, Vaz Bisneta, and Melotto 2019; Locci et al. 2019). Likewise, there were also several well-known defense regulators among the genes responding to R/S phenotype across all four populations. These included a serine-threonine protein kinase (putative LRK10), a nucleotide-binding leucine rich repeat protein (NLR), and several lipoxygenase enzymes, all of which represent protein families with well-known roles in pathogen detection, signal transduction, and subsequent defense (Feuillet, Schachermayr, and Keller 1997; Kourelis and van der Hoorn 2018; Bell, Creelman, and Mullet 1995). Lastly, we also observed many genes involved in the formation of metabolites derived from the phenylpropanoid pathway, such as flavonoids, lignins, and hydroxycinnamic acids. Among these metabolic genes were flavin-dependent monooxygenases, caffeic acid 3-O-methyltransferases, hydroxycinnamoyltransferses, and caffeoyl shikimate esterase (*TcCSE*) (Chezem, Memon, Li, Weng, & Clay, 2017; Návarová, Bernsdorff, Döring, & Zeier, 2012; G.-F. Wang et al., 2015; M. Wang et al., 2018).

This set of differentially expressed metabolic genes suggests a diverse array of potential secondary metabolites, some of which are likely anti-microbial. *TcCSE* (SCA-6_Chr6v1_17513), however, stood out as a particularly attractive experimental candidate for several reasons. First, *TcCSE* was consistently upregulated in response to pathogen challenge across all four populations (Figure 4A). Second, *TcCSE* is a member of the phenylpropanoid pathway and is responsible for hydrolyzing caffeoyl shikimate into the hydroxycinnamic acid (HCAA) caffeate (caffeic acid) (Vanholme et al., 2013). HCAAs are well-known anti-microbial secondary metabolites involved in a diverse array of plant-pathogen interactions (Muroi et al. 2009; Fitzgerald et al. 2004; Knollenberg et al. 2020; Khan et al. 2021; Widmer and Laurent 2006). Together, these results

indicate *TcCSE* could be a potentially important, and yet uncharacterized, gene involved in cacao's defense response. Accordingly, we performed a series of functional experiments involving *TcCSE*, both to verify our candidate gene identification approach and to evaluate this particular gene as a potential breeding marker.

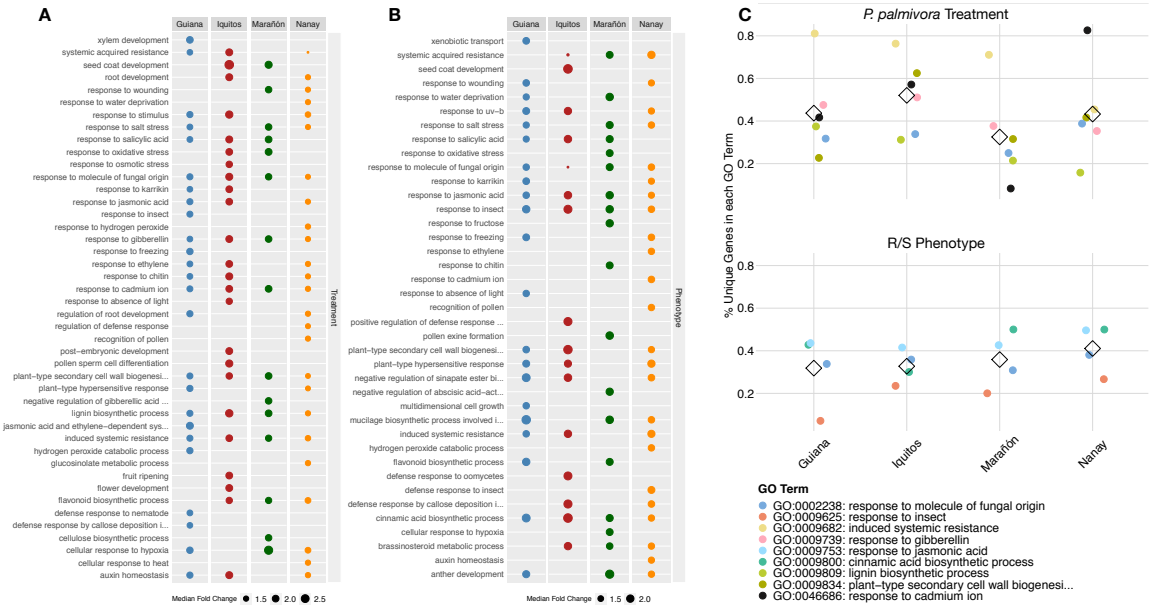


Figure 2-3: **Common functional groups underlie different sets of pathogen responsive genes.** (A) Enriched GO terms (Fisher's exact test: FDR-adjusted p-value < 0.05) and their median fold change for *P. palmivora* treatment versus control. Colored points indicate population membership: Guiana (blue), Iquitos (red), Marañón (green), or Nanay (orange). Point size is scaled to median fold change for the differentially expressed genes belonging to that term. (B) Enriched GO terms (Fisher's exact test: FDR-adjusted p-value < 0.05) and their median fold change for resistant versus susceptible genotypes. Colored points indicate population membership: Guiana (blue), Iquitos (red), Marañón (green), or Nanay (orange). Point size is scaled to median fold change for the differentially expressed genes belonging to that term. (C) The percentage of genes from each population that are unique, calculated for each GO term that is enriched in all four populations. Terms that are significantly enriched in *P. palmivora* treatment versus control are on top, and terms that are significantly enriched in resistant versus susceptible genotypes are on bottom. Each point represents the percentage of differentially expressed genes belonging to a single GO term (indicated by color) that are unique to each population. For instance, Guiana has 22 differentially expressed genes in GO:0009834, of which 5 of them are not differentially expressed in any other population (4/22 = 22.7%). Means are shown as open triangles.

Functional analysis of a candidate gene for caffeic acid synthesis

To begin characterizing *TcCSE*'s role in cacao's defense response, we first verified its function through transient, heterologous over-expression in *N. benthamiana*. We began by cloning *TcCSE* from the Criollo B97-61/B2 variety of cacao driven by a E12-Ω CaMV-35S constitutive promoter. We then transiently transformed *N. benthamiana* plants with our 35S:TcCSE vector or an empty vector control. Consistent with its documented function, transient over-expression of *TcCSE* resulted in significant caffeic acid accumulation relative to our empty vector control (mean intensity difference = 13157.05, t-test 48hpi: p-value = 0.0164; mean intensity difference = 12993.19, t-test 96hpi: p-value = 0.0174; Figure 2-4B). This pattern was true for samples collected at both 48 and 96 hours post transformation.

While caffeic acid has been shown to directly inhibit *P. palmivora* zoospores (Widmer and Laurent 2006), its inhibitory effects towards mycelia have not been tested. Moreover, it remains unclear whether caffeic acid is directly inhibitory *in planta*. To address these points we performed two experiments. First, we grew *P. palmivora* on plates with or without 2 mM caffeic acid. As expected, plates containing 2 mM caffeic acid significantly inhibited mycelial growth (mean colony area difference = 8.9 cm^2, t-test: p-value < 0.001; Figure 2-4C). Second, to determine whether caffeic acid is mobilized to the site of infection *in planta*, which is necessary for direct contact and subsequent inhibition, we placed *P. palmivora* zoospore droplets on detached cacao leaves. After 24 hours, we collected the droplets and measured caffeic acid concentration using liquid chromatography mass spectrometry (LC-MS/MS). Caffeic acid concentration was significantly higher in the zoospore droplets on the surface of leaves than either mock inoculated (mean intensity difference = 677.03) or zoospore only (mean difference = 875.59) controls (t-tests: p-values < 0.001; Figure 2-4D). Mock inoculated leaves had significantly more caffeic acid than zoospore-only controls (mean difference = 198.57, t-test: p-

value = 0.022). These results, combined with previously published evidence (Widmer and Laurent 2006), are consistent with direct inhibition and suggest that caffeic acid is an important part of cacao's early defense response towards *P. palmivora.*

We next used an LC-MS/MS untargeted metabolomics approach to test the hypothesis that cacao plants with higher *TcCSE* expression had higher levels of caffeic acid 8 hours after challenge with *P. palmivora* mycelia (Figure 2-4E). There were no significant differences between treatment, phenotype, or the treatment X phenotype interaction (one-way ANOVA, Intensity ~ Treatment + Phenotype + Treatment X Phenotype: p-values > 0.05). This result did not support our initial hypothesis, but as we elaborate in the discussion, leaves may respond differently to zoospores and mycelia, and/or sampling one metabolite at one time point may not have been sufficient to characterize the relevant phenotype.

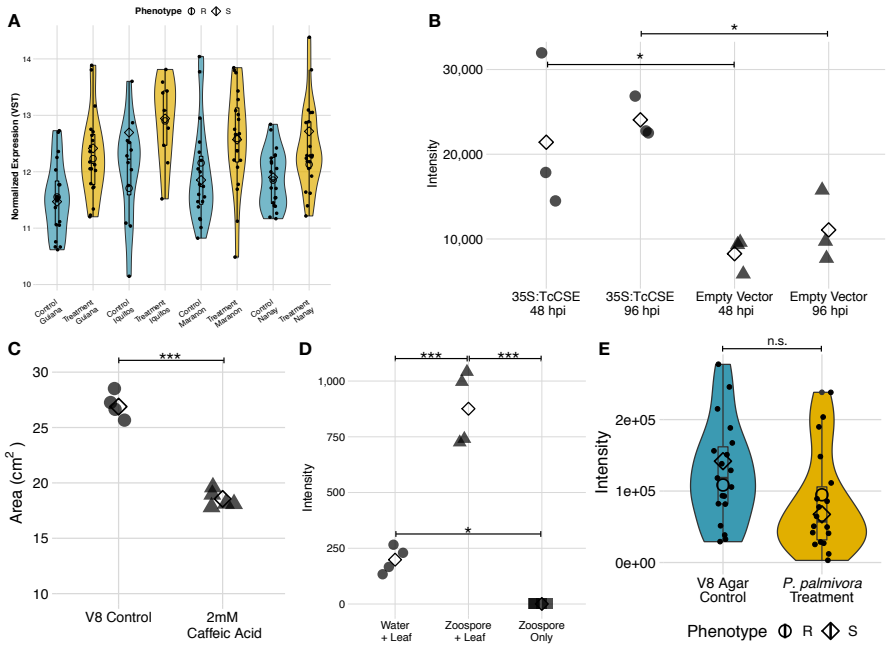


Figure 2-4: ***TcCSE* is involved in resistance to *P. palmivora.*** (A) Expression of *TcCSE* (SCA-6_Chr6v1_17513) across each population for control (blue) and treatment (yellow). Open diamonds indicate mean expression for susceptible genotypes and open circles indicate mean expression for resistant genotypes. (B) Relative abundance of caffeic acid in *N. benthamiana* plants transformed with 35s:TcCSE or an empty vector control, at both 48 and 96 hours post transformation. Means are shown as open triangles. Over-expression of *TcCSE* results in

significantly higher caffeic acid accumulation relative to controls (t-test 48 hpi: p-value = 0.0164; t-test 96 hpi: p-value = 0.0174). (C) Mycelial area of *P. palmivora* cultures grown on plates of V8 media versus plates of V8 media amended with 2mM caffeic acid. Means are shown as open triangles. Plates amended with 2mM caffeic acid significantly inhibited mycelial growth (t-test: p-value < 0.001). (D) Relative abundance of caffeic acid for cacao leaves mock inoculated with water, challenged with *P. palmivora* zoospores, or zoospores only. Means are shown as open triangles. Cacao leaves challenged with zoospores accumulated significantly more caffeic acid than either mock inoculated or zoospore-only controls (t-tests: p-values < 0.001). Mock inoculated leaves had significantly more caffeic acid than zoospore-only controls (t-test: p-value = 0.022). € Relative abundance of caffeic acid in samples challenged with plugs of V8 media (blue) versus plugs of *P. palmivora* mycelia (yellow). There were no significant differences between treatment, phenotype, or the treatment*phenotype interaction (one-way ANOVA, Intensity ~ Treatment + Phenotype + Treatment*Phenotype: p-values > 0.05).

Population branch statistics identify differentially expressed genes under selection

Many of the differentially expressed genes detected in our transcriptome experiment, both in response to pathogen challenge and between R/S phenotypes, were unique to each population (Figure 2-1A). This suggests that at least some aspect of each population's defense response against *P. palmivora* is lineage-specific, that resistance versus susceptibility may be mediated by different genes depending on the population, and that wild populations of crop plants can be a rich source of novel alleles for plant breeding. To determine the extent to which natural selection has shaped resistance and susceptibility in each population, we used population branch statistics (PBS) to estimate the lineage-specific genetic differentiation associated with resistant genotypes in each population (Figure 2-5A). Specifically, this approach detects alleles displaying divergence between resistant and susceptible genotypes, revealing loci that may be under selection. We estimated PBS for the coding region of each gene, as well as 5 Kb on both the 5' (hereafter upstream) and the 3' ends (hereafter downstream). Thus each gene has three PBS values. Genic and non-genic regions in the top 1% of their respective PBS distributions were considered selection outliers. Estimating PBS allows us to relate expression differences to

selection outliers, providing another powerful method for detecting genes important for disease resistance.

Across all four populations, we detected 1,016 selection outliers in the 5Kb upstream region, as well as 915 and 1,003 in the gene body and 5Kb downstream regions, respectively (Figure 2-5B). The vast majority of these selection outliers are unique to each population. This pattern is similar to that observed among the differentially expressed genes, which again suggests each population has evolved lineage-specific aspects of their defense response. Among these selection outliers, 158 were also differentially expressed in response to pathogen challenge, R/S phenotype, or both (Figure 2-5C). This indicates that at least some of the differentially expressed genes have been evolving differently among resistant genotypes. Many of these genes can be found within the 'core' GO terms mentioned previously, including the cinnamic acid biosynthetic process, induced systemic resistance, response to gibberellin, response to jasmonic acid, and response to molecule of fungal origin. Three of the genes defined as selection outliers are differentially expressed across all four genetic groups: WRKY transcription factor 29 (*TcWRKY29;* SCA-6_Chr3v1_10161, pathogen treatment), berberine bridge enzyme 8 (*TcBBE8*; SCA-6_Chr6v1_16921, pathogen treatment), and flavin containing dimethylaniline monoxygenase 1 (*TcFMO1*; SCA-6_Chr9v1_23321, pathogen treatment and R/S phenotype). The fact these three genes are differentially expressed, present in the small number of GO terms enriched across all four populations, and show signatures of divergence among resistant genotypes makes them highly attractive candidates for future experimentation.

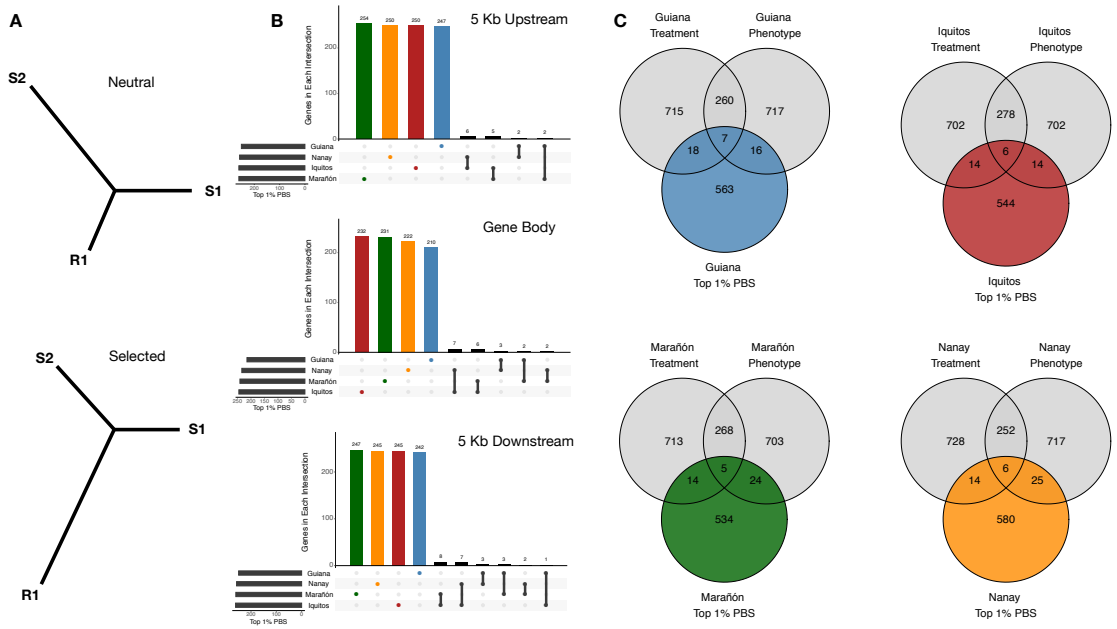


Figure 2-5: **Population branch statistics identify differentially expressed genes under selection.** (A) Population branch statistics can estimate lineage-specific selection leading to resistant genotypes. Branch lengths represent the magnitude of allele frequency change. For loci evolving neutrally in both resistant and susceptible genotypes, differences in allele frequency between resistant and susceptible individuals of the same population (S1, R1) will be *smaller* than allele frequency differences between susceptible individuals from two separate populations (S1, S2) (top). For loci under selection in resistant genotypes, differences in allele frequency between resistant and susceptible individuals of the same population (S1, R1) will be *greater* than allele frequency differences between susceptible individuals from two separate populations (S1, S2) (bottom). High PBS scores indicate genes that are under selection among resistant individuals from a given population. (B) Overlap of genic and non-genic regions designated as selection outliers (top 1% of their respective PBS distributions). PBS was estimated for 5 Kb upstream of each gene (top), the gene body (middle), and 5 Kb downstream of each gene (bottom). The blue, red, green, and orange bars represent genes that are only designated as selection outliers in Guiana, Iquitos, Marañón, or Nanay, respectively. Numbers above the bars indicate the number of selection outliers in that specific intersection. (C) Venn diagrams displaying the overlap between differentially expressed and genes under selection in resistant genotypes. Colors indicate population membership: blue (Guiana), red (Iquitos), green (Marañón), and orange (Nanay).

Discussion

Plant pathogens are responsible for extensive annual yield loss in crop species, a problem that is likely to become worse due to climate change. Through breeding, humans have sought to mitigate the damage these pathogens cause by harnessing natural variation in resistance/susceptibility. However, hybrids created in plant breeding programs often represent only a small proportion of the overall genetic diversity available to a species. Wild populations of crop species are therefore important reservoirs of genetic diversity. Here, we used genomic, transcriptomic, and metabolomic data to investigate the evolution of defense response across four populations of cacao, with the goal of identifying resistance alleles that could potentially be incorporated into breeding programs.

Differential expression analysis revealed a rich set of defense-associated genes that change their expression level either in response to pathogen challenge or between resistant/susceptible individuals. Many of these differentially expressed genes (30-40%) are unique to each population (Figure 2-2A). That is, ~40% of genes that are differentially expressed in one population will not be differentially expressed in any of the other three. However, despite this high degree of lineage-specificity in transcriptional response, there are also a large number of differentially expressed genes that appear to underly a common set of biological processes (Figure 2-2C). These processes represent functional annotation categories that are enriched across all four populations, indicating a fundamental importance for particular defense responses across the species. These include both broad and specific categories, like induced systemic resistance and cinnamic acid biosynthetic process, respectively. Furthermore, although 30-40% of the genes belonging to these shared GO terms were lineage-specific (Figure 2-3C), many of them have a high potential for functional redundancy. For instance, within the cinnamic acid biosynthetic pathway we observed lineage-specific expression and/or evolutionary rate differences in four

separate genes encoding putative caffeic acid 3-O methyltransferases (*TcCOMT*), as well as two separate genes for both shikimate O-hyroxycinnamoyltransferase (*TcHST*) and laccase-14 (*TcLAC14*). Likewise, for the lignin biosynthetic pathway we observed four putative *TcHST* genes and seven separate laccase genes. Thus, while each of our populations likely possess unique solutions to pathogen challenge, at least a portion of their defense responses seem to converge upon common pathways producing potentially analogous functions. Some of the variation may represent lineage-specific differences in the timing of defense gene regulation. It may also arise due to lineage-specific co-evolution with pathogen effectors, which could drive high evolutionary rates and divergence among genetically isolated host lineages.

Of the nine processes that were enriched across all four populations, either in response to pathogen challenge or R/S phenotype, lignin biosynthetic process and cinnamic acid biosynthesis stand out for several reasons. First, as part of the phenylpropanoid pathway, both processes are well-known contributors to plant defense against a wide range of pathogens. For instance, lignin and monolignols play a role in hypersensitive response and penetration defense against fungi and oomycetes (Bhuiyan, Selvaraj, Wei, & King, 2009; Menden, Kohlhoff, & Moerschbacher, 2007). Genes involved in lignin biosynthesis interact with nucleotide binding leucine-rich repeat proteins to modulate plant defense (G.-F. Wang et al., 2015). Hydroxycinnamic acids such as p-coumaroyagmatine, feruloylagmatine, p-coumaroylputrescine, and feruloylputrescine confer defense to the fungal pathogen *Alternaria brassicicola* in *Arabidopsis thaliana* (Muroi et al. 2009). The phenolic aldehyde vanillin, a derivative of ferulic acid, hinders the growth of multiple bacterial species by dissipating ion gradients and thereby inhibiting respiration (Fitzgerald et al. 2004). The hydroxycinnamic acid amide clovamide indirectly inhibits three species of *Phytophthora*, including *P. palmivora* (Knollenberg et al. 2020). And lastly, caffeic acid and its derivatives both directly and indirectly inhibit many pathogens, among them *P. palmivora* and *P. megakarya* (Widmer and Laurent 2006; Khan et al. 2021).

The last of these compounds, caffeic acid, is particularly interesting because the gene responsible for catalyzing the reaction from caffeoyl shikimate to caffeic acid, *TcCSE*, displays consistent upregulation across all four populations (Figure 2-4A). To test whether caffeic acid and *TcCSE* were involved in defense response against *P. palmivora* we performed a series of experiments. We first verified the function of *TcCSE* through heterologous over-expression in *N. benthamiana*, confirming the accumulation of caffeic acid both 48 and 96 hours post transformation (Figure 2-4B). Caffeic acid was both inhibitory to *P. palmivora* mycelia and mobilized to the point of infection at the leaf surface (Figure 2-4C-D). Despite these results, however, genotypes displaying upregulated *TcCSE* in our transcriptome experiment did not display increased caffeic acid accumulation (Figure 2-4E). This result could have arisen from multiple factors. First, *TcCSE* expression could precede caffeic acid accumulation. This possibility is supported by the fact that both the *TcCSE* over-expression experiment (Figure 2-4B) and the caffeic acid mobilization experiment (Figure 2-4D) were completed at > 24 hours, whereas tissue for our transcriptome experiment was collected 8 hours post inoculation. Second, it could be the case that caffeic acid was converted into lignin via sinapic acid (Yamauchi, Yasuda, and Fukushima 2002), which would not be detected using our metabolite extraction protocol. And lastly, caffeic acid could have been converted into one of many possible caffeic acid derivatives that are difficult to predict and quantify (Khan et al. 2021). Together, our results indicate that *TcCSE* and caffeic acid are potentially important components of cacao's defense, though we so far lack a complete understanding of expression time course and fate of resulting metabolites.

We found that major aspects of cacao defense responses against *P. palmivora* were lineage-specific, and, therefore, resistance versus susceptibility appears to be mediated by different genes depending on the population. To further test this possibility, we estimated lineage-specific selection associated with each population's resistant genotypes. Similar to our differential

expression results, there was no consistent set of rapidly evolving resistance-associated genes across all four populations. That is, different sets of genes displayed evidence of selection in each population's resistant genotypes (Figure 2-5B). Among the genes displaying evidence of selection are a small number that are also differentially expressed across all four populations: *TcWRKY29* (SCA-6_Chr3v1_10161), *TcBBE8* (SCA-6_Chr6v1_16921), and *TcFMO1* (SCA-6_Chr9v1_23321) (Figure 2-5C). Despite multiple lines of evidence supporting the importance of these genes, none of them appear to be present in resistance QTLs (Gutiérrez et al. 2021; Lanaud et al. 2009). This is likely because the QTLs were predicted based on progeny from only a handful of parent clones that represent a small fraction of cacao's overall genetic diversity. Thus, even genes that are conserved across wild populations are not always detected and may therefore present novel opportunities for breeding.

Producing cacao varieties that are durably resistant to pathogens requires the development of crop improvement methods that harness underutilized germplasm and rapidly identify disease associated alleles. With the advent of new sequencing technologies and readily available analytical tools, we are now in an era where the benefits of cacao's genetic diversity can be fully realized. In this study, we investigated the evolution of defense response against *P. palmivora* across four divergent populations of cacao. Consistent with the high genetic differentiation among these populations, we observed both lineage-specific transcriptional differences and historical responses to selection. This suggests these populations have experienced genetic drift and/or adapted to their local microbial communities in ways that affect their defenses against *P. palmivora*. Genes and pathways that responded consistently across all four populations include *TcCSE*, *TcFMO1, TcWRKY29,* and *TcBBE8*, as well as pathways involved in the biosynthesis of phenylpropanoids. Together, our results indicate cacao's defenses against *P. palmivora* are mediated by a network of both conserved and divergent responses, and

suggests wild cacao populations are a source of genetic diversity that could help improve the health of both tree and farmer.

Materials and Methods

Plant propagation

We selected 31 cacao genotypes for experimentation based on their resistance/susceptibility to the black-pod rot causing pathogen *Phytophthora palmivora* (Fister et al. 2020). To begin propagation of plants required for this study, we first imported grafted trees from the Tropical Agricultural Research and Higher Education Center (CATIE). From those grafted plants we created rooted cuttings according to a previously described method (S. N. Maximova et al. 2005). We took cuttings from plant material that had an approx. 0.5 cm stem and was beginning to become woody. Cuts were made approx. 3 mm above the node, attempting to capture a single leaf. Each leaf was cut in half. We submerged the woody portion of each cutting in rooting hormone (1:1 IBA potassium salt and NAA, 0.2 g total (0.1 g each) in 50 mL 50% EtOH) and placed them in wet sand (Quikrete, medium grade), so that the leaf petiole is just above the surface. Finally, we placed the cuttings into a misting chamber (every 10 minutes for 6 seconds) surrounded by shade cloth and supplemented natural light using LED lights (16 hr photoperiod, 6am - 10pm).

Once cuttings developed roots and were starting to put on new growth (approx. 4 weeks after cutting), we re-potted the plants into D40H D-pots from Stuewe (Tangent, OR). Peat mix was used to plug the bottom of the pots before filling them the rest of the way with a wetted mixture of 4:2:1 Perlite:Sand:Turface. We placed rooted plants on automatic drip irrigation lines (1 dripper/plant) and watered 3 times per day: 8am for 10 minutes, 12pm for 6 minutes, and 6pm

for 6 minutes. Finally, plants were left in the misting chamber for 2 weeks to allow for them to recover before being transferred out of the mist chamber into a temperature and humidity-controlled greenhouse. For the duration of the experiment, the greenhouse was kept at 80-90% relative humidity, 76 °C at night, and 83 °C during the day. Of the approximately 300 rooting cuttings that were taken, only 141 survived and were healthy enough for experimentation. The number of replicates per genotype, population, and resistance/susceptibility class varied (Supplemental Figure S2-9).

Genotype phylogeny

Evolutionary relationships among the 27 genotypes used in this study were assessed using a phylogeny inferred from 23,439 single nucleotide polymorphisms (SNP). The tree was constructed using the SNPhylo pipeline (Lee et al. 2014). Briefly, SNPs were filtered out if they had a minor allele frequency < 0.1 (-m 0.1) and were missing in more than 10% of the genotypes (-M 0.1). Moreover, since SNPs in linkage disequilibrium provide redundant information, only a single SNP per linkage disequilibrium block was used for phylogenetic reconstruction (-l 0.6). SNPs were concatenated and aligned using MUSCLE (Edgar 2004). Lastly, a maximum likelihood phylogeny was constructed using DNAML in PHYLIP (Felsenstein 1993). Phenotype information was mapped next to the phylogeny based on previously described levels of resistance or susceptibility to *P. palmivora* (Fister et al. 2020).

Transcriptome experimental design and treatment

In a greenhouse, 2 tables were aligned parallel to one another (Supplemental Figure S2-8). On each bench, we placed 3 trays, with approximately 30 plants on each tray. To minimize the

effect of gradients in temperature, humidity, and light within the greenhouse, we kept the distance between tables to < 2 ft. We treated the plants in each tray with either pathogen or V8 control, such that parallel trays never experienced the same treatment. In order to minimize edge versus interior effects, environmental factors being confounded with population or resistance/susceptibility etc., we randomized the placement of plants in each tray, with the caveat that the same genotype was in a mirrored position on both tables. Thus for each pair of plants within a genotype, one would receive pathogen treatment and one would receive control treatment. If there was an odd number of plants for a given genotype, or if a genotype only had one representative plant, the odd-numbered individual would be paired with an individual within the same population *and* resistance/susceptibility class (Fister et al. 2020). And if a genotype within the same population and resistance/susceptibility class was unavailable, we used a genotype in the same resistance/susceptibility class from a different population. The plants were moved to their respective positions one week before the experiment in order to allow them to acclimate.

To create pathogen cultures for infection, we created cultures of *P. palmivora* strain Gh-ER1349 on V8 media as previously described (Fister, Shi, et al. 2016). Briefly, plugs of pathogen were taken out of liquid nitrogen 3 weeks before the experiment, dried, and placed on V8 agar. Plates were placed in the dark at 27 °C. Then, 1.5 weeks before the experiment, pathogen cultures were sub-cultured onto new V8 plates. Finally, two days before the experiment, *P. palmivora* plates were once again sub-cultured to create 120 thin (10 mL) V8 agar plates. Plates were then left to grow in the dark at 27 °C until the experiment.

Prior to inoculation, we evaluated each plant. We selected 2 leaves from each plant for inoculation based on size and health. All leaves were graded as stage D, D/E (transitioning from D to E) or E. Inoculation was done on the abaxial side of the selected leaves using either 1.5 cm mycelia plugs taken from the growing edge of the culture, or 1.5 cm plugs of the V8 control.

Inoculations were done an hour after sunset and green headlamps were worn to limit the effect of light on the plants. We placed 6 agar plugs of either pathogen mycelia or V8 control on each of the selected leaves, avoiding veins or damaged portions of the leaf as much as possible. After all 6 plugs were placed, we sprayed each leaf with a fine mist of water to limit desiccation of the agar plug. After 8 hours, leaves were collected following the same order as inoculation. Both leaves were carefully removed from the plant, making sure agar plugs remained attached. The leaves were then placed on a cutting board and a 1.75 cm cork borer was used to take a leaf disc with the center of the agar plug as the center of the disc. In this way, the entire agar plug plus a small amount of surrounding tissue was cut from each leaf. We then removed the agar plugs and pooled the 12 leaf discs (6 from each of 2 leaves) into a single 2 mL tube. Tubes were immediately flash frozen in liquid nitrogen before finally being stored at -80 °C.

Sample preparation and sequencing

Tissue was ground using pre-chilled (frozen at -80 °C) stainless steel beads (2 x 2.3 mm, and 1 x 3.2 mm) in a Qiagen (Hilden, Germany) TissueLyzer for 3 rounds of 40 seconds. Tubes were re-frozen after each round to prevent thawing. Once tissue was ground into a fine powder, samples were once again stored at -80 °C.

We extracted RNA from 100 mg of ground tissue. RNA extraction protocol was adapted from Thermo Fisher Scientific's small scale RNA isolation protocol (Publication No. MAN0000243) for PureLink™ Plant RNA Reagent (Life Technologies, Carlsbad, CA, USA). The following modifications were made: Homemade extraction buffer (1 mL) from US Patent US6875757B2 was substituted for 0.5 mL PureLink™ Plant RNA Reagent, samples were vortexed until homogenized in buffer, all spins were done at 16,000 x g at 4 °C, 200 ul of NaCl was used, 600 ul of chloroform was used for first organic extraction, then chloroform extraction

was repeated 1 time using an equal volume of chloroform to aqueous layer (typically 1 mL) , 3 x 1mL ethanol washes were performed to improve sample purity, and nucleic acid pellets were allowed to dry for 10 minutes before resuspension in 20 uL VWR molecular grade water.

Once RNA extractions were completed, we assessed the purity and concentration using a NanoDrop 2000/2000c (Thermo Fisher Scientific, Waltham, MA, USA). RNA integrity was determined by running 1 uL of the purified RNA on a 1.5% agarose gel, making sure both the 28S and 18S rRNA bands were intact. DNA contamination was removed from RNA using Thermo Fisher DNAse1 (RNAse-free, catalog #EN0521) and the manufacturer's protocol (Publication No. MAN0012000). After DNAse treatment, we further purified the RNA using a Zymo RNA Clean and Concentrator kit (Catalog #R1013; Irvine, CA) following the recommended protocol in the manufacturer's manual. RNA was eluted from columns in 15 uL. Prior to sequencing, we determined final RNA concentration and integrity using an Agilent 4200 Tapestation System. Samples with less than 44 ng/uL and/or a RIN less than 5.0 were re-extracted.

Transcriptome sequencing was performed using the Pennsylvania State University Genomics Core Facility. Lexogen QuantSeq libraries were created using the manufacturer's protocol. Samples were then run in 5 batches, 32 samples per batch, on an Illumina NextSeq 550 in High Output mode with 75 bp reads, producing approx. 12 million reads per library.

Genome meta-assembly

DNA was extracted and sequenced according to previously outlined methods (Hämälä et al. 2021). The linked read data for the *Theobroma cacao* genotype SCA-6 were assembled with Supernove v2.1 (Weisenfeld et al. 2017) at five different raw read coverage depths of approximately 56x, 62x, 68x, 75x, and 85x based on the estimated genome sizes. We translated

the Supernova assembly graph to create two parallel pseudohaplotype FASTA representations of the genome (pseudohap2 style) and utilized one pseudohaplotype from each of the five assemblies for subsequent post-processing. Among these five pseudohaplotype assemblies, we designated one of them as the optimum primary Supernova assembly using a combination of assembly metrics. Utilized assembly metrics include: completeness of annotated conserved land plant (embryophyta) single-copy BUSCO genes (Simão et al., 2015, Waterhouse et al., 2018), contig and scaffold contiguity (L50), and an assembly size closer to the estimated haploid genome size (Supplementary Tables S10-S12). Quickmerge (Chakraborty, Baldwin-Brown, Long, & Emerson, 2016) was then used to incrementally improve the back-bone assembly by bridging gaps and joining contigs using the remaining four primary pseudohaplotype assemblies in decreasing order of assembly quality. After each merging step, the resulting meta-assembly was assessed for contiguity, completeness, and assembly size, only being retained if all three displayed improvement. Assembly errors introduced during *de novo* assembly and merging were corrected using the Tigmint (Jackman et al. 2018) and ARCS (Yeo et al. 2018) algorithms. Tigmint aligns linked reads to an assembly to identify potential errors, then breaks assembled sequences at the boundaries of these errors. The assembly is then re-scaffolded into highly contiguous sequences with ARCS utilizing the long-distance information contained in the linked reads. Gapfiller v1.10 (Boetzer and Pirovano 2012) was used to iteratively fill gaps between contigs using paired-end reads from both the short insert Illumina libraries and the 10x Chromium libraries. Finally, those same reads were used by Pilon v1.23 to correct base errors and local mis-assemblies.

Pseudochromosome construction

Chloroplast, mitochondrial, and contaminant sequences present in the meta-assembly were removed prior to construction of the nuclear pseudochromosomes. To identify these extraneous DNA sequences, the meta-assembly was searched against the NCBI nucleotide collection database (*nt*) using Megablast (Chen, Ye, Zhang, & Xu, 2015). Meta-assembly sequences with hits in the *nt* database were then queried against the NCBI taxonomy database to determine their taxonomic attribution. Meta-assembly sequences with best hits to non-embryophytes (land plants) were considered contaminants and discarded. We performed a second iteration of Megablast searches of the remaining meta-assembly sequences (embryophyte-only) against the NCBI RefSeq plant organelles database to identify chloroplast and mitochondrial sequences. Meta-assembly sequences with high similarity (> 80% identity and > 50% coverage) to sequences in the plant organelles database were also discarded. Finally, the remaining nuclear contigs and scaffolds were ordered and oriented into pseudomolecules with RaGOO (Alonge et al. 2019) using the *Theobroma cacao* L. cultivar Matina 1-6 v1.1 (Juan C. Motamayor et al. 2013) reference chromosomes.

Assembly evaluation and validation

We assessed the SCA-6 meta-assembly for contiguity, completeness, and structural accuracy by comparing it to the two published *Theobroma cacao* chromosome level reference assemblies of Matina 1-6 v2.1 and Criollo B97-61/B2 v2.0. Both the contig and scaffold assembly metrics were evaluated in addition to completeness of universally conserved single copy genes using the BUSCO land plants (embryophyta) benchmark gene set. Whole genome

synteny comparison between Matina 1-6 v2.1 and the five genotypes in this study were plotted with DGenies using whole genome DNA alignments performed with minimap2.

Repeat library construction

Prior to annotation, repetitive and TE-rich regions of the genome must be masked, lest they be annotated as protein-coding genes. We did so according to the MAKER-P repeat masking protocol (Campbell et al. 2014). MITE-Hunter (Han & Wessler, 2010) and LTRharvest/LTRdigest (Ellinghaus, Kurtz, & Willhoeft, 2008; Steinbiss, Willhoeft, Gremme, & Kurtz, 2009) were used to collect consensus miniature inverted-repeat transposable elements (MITEs) and long terminal repeat retrotransposons (LTRs) from the meta-assembly, respectively. LTRs were first filtered to remove false positives and elements with nested insertions, then combined with the MITEs to mask the genomes. The unmasked regions of the genomes were then annotated for *de novo* repetitive sequences using RepeatModeler1 (http://www.repeatmasker.org/RepeatModeler). Finally, all collected repetitive sequences were compared to a BLAST database of plant proteins from SwissProt and RefSeq, where proteins from transposable elements are excluded. Sequences with significant hits to the protein database were excluded from the repeat masking library.

Generation of gene annotation evidence

In order to capture robust transcript data to support genome annotation, we sequenced pooled RNA from a diverse array of cacao tissue samples available in the Guiltnan-Maximova lab. All harvested tissues were flash frozen in liquid nitrogen immediately on collection, homogenized to fine powder, and stored in liquid nitrogen or at -80 °C for RNA extraction. Total

RNA was isolated from cacao tissue samples using Purelink Plant RNA Reagent following the same protocol outlined above. Extracted samples were cleaned by ethanol precipitation (Zumbo, 1932) before sample pooling. RNA extracted from tissue following salicylic acid treatment were collected and processed as previously described (Fister et al. 2015). Individual and pool RNA integrity was assessed on an Agilent 2100 Bioanalyzer System. Illumina TruSeq libraries (150 nt) were prepared using RNA pools at Pennsylvania State University, The Hucks Institutes Genomics Core Facility. Libraries were then sequenced on an Illumina NextSeq 550 in high output mode at the same facility.

Raw RNA-Seq reads were trimmed to remove low-quality bases as well as embedded adaptor sequences and filtered to discard short read fragments using Trimmomatic v0.33 (Bolger, Lohse, and Usadel 2014) . We then used FastQC v0.10.1 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to assess the overall sequence quality before and after trimming. Cleaned reads from each tissue sample were *de novo* assembled using Trinity (Haas et al. 2013) with the default parameters. The resulting transcriptome assemblies were post-processed with the PlantTribes 2 AssemblyPostProcessor (https://github.com/dePamphilis/PlantTribes) to select contigs with potential coding regions to use as evidence for gene annotation.

Gene prediction and functional assignment

Protein-coding gene annotations from the reference *Theobroma cacao* genomes of Matina 1-6 v2.1 and Criollo B97-61/B2 v2.0 were separately transferred to pseudomolecules of the SCA-6 meta-assembly using the FLO (https://github.com/wurmlab/flo) pipeline, which is based on the UCSC Genome Browser Kent-Toolkit (Kuhn, Haussler, & Kent, 2013). We then utilized the MAKER annotation pipeline (release 3.01.02) (Holt and Yandell 2011) to update

transferred annotations with evidence data and to predict gene models with *ab initio* gene finders. Repetitive and low complexity regions of the pseudomolecules were first masked with RepeatMasker in MAKER using the previously described cacao-specific repeat library. The annotation evidence provided to MAKER includes previously described tissue- and stress-specific transcriptome assemblies. Additionally, proteomes of nine representative Malvid genomes, including *Gossypium raimodii*, *Gossypium hirsutum*, *Arabidopsis thaliana*, *Carica papaya, Citrus sinensis, Citrus clementina, Eucalyptus grandis, Panica granatum*, and *Populus trichocarpa* were provided as cross-species homology evidence. In the initial run of MAKER, transferred annotations were updated with evidence data and additional annotations were predicted with Augustus using a cacao training set. A second iteration of MAKER was performed using both Augustus and SNAP *ab initio* gene finders to further improve the quality of gene models (Korf, 2004; Stanke et al., 2006). We selected approximately 5,000 high confidence gene models from the initial MAKER run to train SNAP Hidden Markov models used to predict gene structure. MAKER only replaced a previously predicted gene model if annotation evidence suggested that a model from the second run was better. Complete functional annotation of gene sets was performed using the Blast2GO (Conesa et al., 2005) functional annotation module. The best functional descriptors for gene products were assigned following BLASTp searches against the UniProt/SwissProt databases. Additionally, gene models were assigned to KEGG (http://www.kegg.jp/) pathways and annotated with protein family domains as detected by InterProScan (Quevillon et al. 2005). Identified domains were directly translated into gene ontology terms.

Expression quantification, differential expression, and gene ontology enrichment

Illumina 75 bp reads were trimmed to remove adapters using trimmomatic (Bolger, Lohse, and Usadel 2014). Reads were aligned to the SCA-6 meta-assembly using STAR (Dobin et al., 2013) and quantified using featureCounts (Liao, Smyth, & Shi, 2014)**.** Differential expression analysis was performed using DESeq2 (Love, Huber, & Anders, 2014). Due to variation in temperature, humidity, and leaf developmental stage across the experiment, we included both tray and leaf developmental stage as covariates in the model (Supplemental Figure S2-7). Moreover, because the experiment was unbalanced, i.e. containing inconsistent sample sizes both within and between phenotype classes and populations, we provided custom contrast matrices to DESeq2 for the differential expression calculations (Supplemental Figure S2-9). The contrast matrices attempt to add weights that help mitigate the bias introduced by differences in sample size and were calculated as follows:

Treatment contrast:

$$\frac{\left(\frac{1}{\sum \text{Treatment}_{\text{Resistant}}} + \frac{1}{\sum \text{Treatment}_{\text{Susceptible}}}\right) - \left(\frac{1}{\sum \text{Control}_{\text{Resistant}}} + \frac{1}{\sum \text{Control}_{\text{Susceptible}}}\right)}{2}$$

Phenotype contrast:

$$\frac{\left(\frac{1}{\sum \text{Treatment}_{\text{Resistant}}} + \frac{1}{\sum \text{Control}_{Resistant}}\right) - \left(\frac{1}{\sum \text{Treatment}_{Susceptible}} + \frac{1}{\sum \text{Control}_{\text{Susceptible}}}\right)}{2}$$

Interaction contrast:

$$\frac{\left(\frac{1}{\sum \text{Treatment}_{\text{Resistant}}} - \frac{1}{\sum Treatment_{\text{Susceptible}}}\right) - \left(\frac{1}{\sum \text{Control}_{Resistant}} - \frac{1}{\sum \text{Control}_{\text{Susceptible}}}\right)}{2}$$

Here, treatment effects describe those genes that respond to pathogen treatment, but do not show differences between resistant and susceptible genotypes. Phenotype effects describe those genes that display differences between resistant and susceptible genotypes, but do not differ between treatment and control samples. Finally, additive effects are just those genes that respond to both treatment and phenotype, thereby representing the overlap between the two main effects. Very few interactive effects were observed in our study, so we chose to omit them. After running differential expression, we chose the top 1000 genes ranked by absolute $\log_2$ fold change (LFC) to run gene ontology enrichment. We chose an arbitrary LFC cutoff, rather than one based on p-values after multiple test correction, because limitations in sample size and inter-genotype variation resulted in a loss of statistical power at the group level. To verify that our LFC cutoff did not cause spurious results (Figure 2-2A), we performed the same analysis on two different subsets of our data. First, we analyzed only those genes that were significantly differentially expressed (FDR-adjusted p-value < 0.05). Second, to verify that the large proportion of genes private to each population was not due to random chance, we compared the overlap of two types of subsamples. In the first type of subsample, we ranked the genes in each population by LFC before taking samples of size N, where N = 200 – 2000 genes. This is the exact same protocol we used to choose our top 1000 differentially expressed genes. In the second type of subsample, we randomly sampled gene sets of size N, where N = 200 – 2000 genes. For both types of subsample we calculated the proportion of unique genes in each population, for each sized sample. We calculated whether differences in subsamples (LFC-ranked versus random subset) were significant using a one-way ANOVA followed by Tukey's honest significant different (Tukey HSD).

Lastly, we verified that the genes unique to each population did not display significantly lower expression than the genes shared between populations (Supplemental Figure S2-6). For both the treatment and phenotype main effects, the genes unique to specific populations were not

systematically biased towards lower expression. In fact, for treatment, the genes unique to Guiana and Marañón had significantly higher expression than the genes shared among populations (one-way ANOVA, p -value < 2e-16; Tukey's HSD, adjusted p-value < 0.001). And for phenotype, the genes unique to Guiana, Marañón, and Nanay had significantly higher expression (one-way ANOVA, p -value < 2e-16; Tukey's HSD, adjusted p-value < 0.01).

We used the top 1000 genes, ranked by |LFC|, from each population for further analysis. We performed gene ontology (GO) enrichment analysis using topGO v2.38.1 (algorithm = "classic", statistic = "fisher"), which produced a large list of enriched GO terms (FDR-adjusted p-value < 0.05). Because gene ontologies are organized as directed acyclic graphs (DAG), leading to parent-child relationships between specific terms, GO enrichment methods often produce large, unwieldy lists that contain redundant information that complicates further analysis. Therefore, we exploited the structure of the DAGs to prioritize GO terms that lie close to the tips of the graphs using GOxploreR (Manjang, Tripathi, Yli-Harja, Dehmer, & Emmert-Streib, 2020). In this way, terms providing the most specific information were carried forward for further analysis. We then grouped similar GO terms using Lin's measure of semantic similarity as implemented in REVIGO (Supek, Bošnjak, Škunca, & Šmuc, 2011).

In order to determine whether each population was using different yet evolutionarily related genes to defend themselves against P. palmivora, we classified all predicted proteins in the SCA-6 genome into orthologous gene families. This was done using PlantTribes (Wall et al. 2008), which employs a combination of BLAST (Altschul, 1990) and hidden Markov models (Eddy 2011) to infer groups of genes that share a single common ancestor among a diverse set of 37 high quality plant genomes (https://github.com/dePamphilis/PlantTribes).

TcCSE cloning and over-expression in Nicotiana benthamiana

Cacao cDNA was prepared with DNaseI-treated RNA from stage A/B leaf tissue (cacao genotype SCA-6) using M-MuLV Reverse Transcriptase (NEB M0253S; New England Biolabs, Ipswich, MA, USA). *TcCSE* (CDS: 960 bp) was cloned from cDNA using Phusion DNA Polymerase (NEB 0530S) and the primers TcCSE_for and TcCSE_rev (Supplemental Table S2.1). The primers introduced BsaI sites with overhangs 1 and 4 on the 5' and 3' end of the amplicon, respectively, for later subcloning into pGK19.0923 by Golden Gate assembly (see below). The amplicon was cloned into pMiniT 2.0 using the NEB PCR Cloning Kit (NEB E1202S) and verified by Sanger sequencing.

To facilitate rapid subcloning of *TcCSE* and other coding sequences into an overexpression vector, the binary vector pGZ12.0501 (GenBank KF871320.1) was converted into a GoldenGate assembly compatible vector (Lebedenko, Birikh, Plutalov, & Berlin YuA, 1991; Valla & Lale, 2016). To achieve this, the PDK intron from pHANNIBAL (GenBank: AJ311872.1) was amplified by PCR (Phusion polymerase) with the primers PDK_BsaI_for and PDK_BsaI_rev. PDK_BsaI_for introduced one SpeI and two BsaI restriction sites on the 5' end of the amplicon and PDK_BsaI_rev introduced two BsaI and one HpaI restriction sites on the 3' end of the amplicon (Supplemental Table S2.1), resulting in the following amplicon with BsaI restriction sites with unique overhangs (in parentheses): (TGCC)/BsaI recognition site 1 (reversed) – BsaI recognition site 2/(GCAA) – PDK intron – (ACTA)/BsaI recognition site 3 (reversed) – BsaI recognition site 4/(TTAC). The amplicon was digested with SpeI and HpaI restriction enzymes and ligated into pGZ12.0501 between SpeI and HpaI sites using T4 DNA Ligase 4 (NEB M0202S). This resulting vector is referred to as pGK19.0923.

For Golden Gate assembly, pMiniT 2.0 plasmid harboring the *TcCSE* candidate coding sequences with BsaI adapters (sites 1 and 4) (~150 ng) was mixed with pGK19.0923 plasmid

(~50 ng) in 1x T4 DNA Ligase buffer (NEB B0202S), with T4 DNA Ligase (NEB M0202S, 200 U) and BsaI-HF-v2 (NEB R3733S, 10U) in a total reaction volume of 10 µl. The reaction mixture was incubated at 37°C for 30 minutes, followed by 30 cycles of 37°C (5 minutes)/16°C (5 minutes), and a final heat denaturation at 60°C (5 minutes). The product was transformed into *E. coli* (10-beta) for selection on LB-kanamycin plates. The resulting vector will be referred to as 35S:TcCSE and places the *TcCSE* coding sequence after the E12-Ω CaMV-35S constitutive promoter (Mitsuhara et al., 1996).

35S:TcCSE and the empty vector control pGH00.0126 (GenBank KF018690.1) (S. Maximova et al. 2003) were transformed into *Agrobacterium tumefaciens* strain AGL1 (Lazo, Stein, & Ludwig, 1991) by electroporation. The *A. tumefaciens* cultures were grown overnight in liquid 523 media to an optical density (OD_{600nm}) of ~1 as previously described (Fister, Shi, et al. 2016). Cells were pelleted by centrifugation (15 minutes at 5,000 x *g*) and the cell pellet was re-suspended in sterile water to an optical density ($OD_{600\ nm}$) of 0.4±0.02 for *Nicotiana benthamiana* infiltration and transient expression.

Four volumes of *A. tumefaciens* culture harboring either the empty vector or 35S:TcCSE constructs were mixed with one volume of p19 culture (*A. tumefaciens* with pDGB3alpha2_35S:P19:Tnos, Addgene #GB1203; Addgene, Watertown, MA, USA) (Sarrion-Perdigones et al., 2013) for co-infiltration.

N. benthamiana plants were grown to 4-5 weeks from seed. Stage 2 and 3 leaves, according to Ma et al. 2012 (Ma, Lukasik, Gawehns, & Takken, 2012), were infiltrated with *A. tumefaciens* cultures on the abaxial side using a needle-less syringe as previously described (Bach et al., 2014).

At 48 and 96 hours after infiltration, 1.5 cm (I.D.) holes were punched out using a cork borer from *N. benthamiana* leaf tissue expressing the GFP marker gene included in both pGH00.0126 and pGK19.0923 backbones. Two leaf discs from the same plant were placed in a 2

ml screw cap tube containing 1 ml of 80/20/0.1 methanol/water/formic acid (v/v/v) and constituted one sample. Samples were heated at 80°C for 30 minutes. The supernatant was dried in a SpeedVac and the resulting pellet was dissolved in an equal volume of 90/10/0.1 water/methanol/formic acid (v/v/v), filtered (0.2 μm, nylon), and loaded into HPLC vials for LC-MS/MS analysis.

Samples were run in negative ion mode on an AB SCIEX 5600 Triple TOF with a Shimadzu Prominence UFLC at Pennsylvania State University's Metabolomics Core Facility at the Huck Insitutes of the Life Science. We followed the instrument specifications previously outlined in Knollenberg et al. 2020.

We analyzed spectral and separation data coming from the LC-MS/MS instrument using the XCMS v3.8.2 package in R v3.6.3. Feature detection was performed using the following parameters: ppm = 15, minimum peak width = 5, maximum peak width = 20, signal/noise threshold = 6, m/z diff = 0.01, integration method = 1, prefilter peaks = 3, prefilter intensity = 100, noise filter = 0. Peaks were then grouped according to the following parameters: bw = 5, minimum fraction = 0.4, m/z width = 0.015, minimum number samples = 1, maximum features = 100. We subtracted the mass of a single proton (1.007276 Da) from the monoisotopic mass of caffeic acid (180.04225873 Da) to identify putative caffeic acid metabolites. We discovered a putative caffeic acid metabolite at a median m/z of 179.0354 and a median retention time of 489.2016 seconds. Our putative caffeic acid metabolite was then confirmed using MS-DIAL v4.0 (Tsugawa et al., 2015) to extract and confirm the MS/MS spectrum.

Plant metabolite extraction from selected transcriptome tissue samples

We extracted metabolites from leaf discs collected during the RNA-seq experiment (*Transcriptome experimental design and treatment*) according to previously described methods

(De Vos et al. 2007; Knollenberg et al. 2020). We flash froze leaf discs in liquid nitrogen and ground them in a mortar and pestle. Special care was taken to prevent the tissue from thawing. A 3:1 solvent to tissue ratio (µl:mg) was used to extract the metabolites, where the solvent was a solution of LC-MS/MS grade 80% methanol and 0.1% formic acid in water (v/v). Genistein was spiked into each sample to serve as an internal control (Calderón, Wright, Hurst, & van Breemen, 2009). Finally, we filtered residual particulates from the extract using spin columns (0.2 µm; Norgen Biotek Corp. Cat. #40000) before quantifying metabolites via LC-MS/MS. LC-MS/MS samples were again run using the specifications outlined in the previous section (*TcCSE Cloning and Over-expression in Nicotiana benthamiana*).

Phytophthora palmivora growth inhibition and zoospore preparation

We performed growth inhibition assays to assess whether caffeic acid was capable of directly inhibiting *Phytophthora palmivora* strain Gh-ER1349 mycelial growth. First, pathogen cultures were taken out of storage in liquid nitrogen and grown on 20% V8 media (Fister, Shi, et al. 2016) for two days. After two days, we sub-cultured the leading edge of the culture onto new plates with or without 2 mM caffeic acid. Plates were stored upside-down in the dark at 27 °C for two days, after which we determined mycelial growth inhibition using ImageJ (Schneider, Rasband, & Eliceiri, 2012). We amended the plates with 2 mM caffeic acid because this concentration is on the low end of what has previously been considered physiologically relevant (Widmer and Laurent 2006). We prepared *P. palmivora* zoospores for the metabolite mobilization assay according to the following protocol. We created 125 mL Erlenmeyer flasks containing 25 mL V8 media. We placed two mycelial plugs in each flask and sealed them with foil and parafilm. In order to make sure pathogen cultures were kept in darkness, flasks were placed in a cardboard box in the incubator (27 °C) for 7 days. After 7 days, flasks were place in 24 hour light

for 4 days, again at 27 °C. After this 11 day period, we induced zoospores by first flooding each flask with 25mL sterile, ice cold water. Flooded flasks were then placed in the refrigerator (4 °C) for 45 minutes before placing them back in the incubator (27 °C) for 30 min. We calculated the concentration of newly created zoospores using a hemocytometer. Finally, we resuspended zoospores in 50 mL Falcon tubes and immediately used them for experimentation.

Genome scan for selection

We searched for signals of selection at the genome level by using previously published short-read sequence data from the 31 genotypes (Hämälä et al. 2020). After removing low-quality reads and sequencing adapters with Trimmomatic (Bolger, Lohse, and Usadel 2014), we aligned the surviving reads to the SCA-6 meta-assembly using BWA-MEM (Li, 2013). We removed duplicated reads with SAMtools (Li et al., 2009) and called variable sites using BCFtools (Li, 2011). We only used reads with mapping- and base-quality ≥ 20 in the variant calling. The variant calls were then filtered to only keep biallelic SNPs with the following requirements: site- and genotype-quality ≥ 20, read coverage ≥ 6, < 20% missing data, and minor allele frequency > 0.05.

We used population branch statistics (PBS) (Yi et al., 2010) to estimate the genetic differentiation of lineages leading into the resistant genotypes of each population. Standard differentiation measures, such as F_{ST} or d_{XY}, can detect signals of differential selection, but they generally cannot distinguish which of the populations has been the target of selection. To detect lineage-specific selection, PBS uses an outgroup to polarize differentiation measures between two closely related populations. Assuming a closely related population pair 1 and 2, and an outgroup 3, PBS for population 1 is estimated as:

$$\mathrm{PBS}_1 = \frac{\mathrm{T}_{12} + \mathrm{T}_{13} - \mathrm{T}_{23}}{2},$$

where T_i is a relative divergence time: $T = -\ln(1 - F_{ST})$. Here, using the F_{ST} estimator by Hudson 1992 (Hudson, Slatkin, & Maddison, 1992), we first quantified differentiation between the resistant and susceptible genotypes of each population. Then, to find selection specifically acting on the resistant class, we combined the susceptible genotypes from the three remaining populations to act as an outgroup. The reasoning behind this approach is that alleles responding to pathogen-mediated selection in the resistant genotypes should be either neutral or deleterious in the susceptible genotypes, revealing longer-than-expected branch lengths leading into the resistant lineages. To better associate the selection signals with results from the transcriptome experiment, we estimated PBS specifically for each gene, including the surrounding regulatory regions. Consistent with previously published methods (Choudhury et al., 2014; Hsieh et al., 2017; Schweizer et al., 2019), we categorized the top 1% of PBS scores as selection outliers.

Acknowledgments

Thank you to Lena Sheaffer for her assistance in project and laboratory management. Thank you to Francisco Menendez Burns, Zach Dashner, and Akiva Shalit-Kaneh for the tissue they contributed for genome annotation. Thank you to Craig Praul and the Huck Institutes of Life Sciences Genomics Core Facility. This work was supported by The Pennsylvania State University College of Agricultural Sciences, the Huck Institutes of the Life Sciences, the Penn State Endowed Program in Molecular Biology of Cacao, NSF Plant Genome Research Award 1546863 and by the Agriculture and Food Research Initiative (grant number 2018-07789 and accession number 1019277) from the USDA National Institute of Food and Agriculture.

References

Ali, S. S., Amoako-Attah, I., Bailey, R. A., Strem, M. D., Schmidt, M., Akrofi, A. Y., … Bailey, B. A. (2016). PCR-based identification of cacao black pod causal agents and identification of biological factors possibly contributing toPhytophthora megakarya's field dominance in West Africa. *Plant Pathology*, *65*(7), 1095–1108.

Ali, Shahin S., Shao, J., Lary, D. J., Kronmiller, B., Shen, D., Strem, M. D., … Bailey, B. A. (2017). Phytophthora megakarya and P. palmivora, closely related causal agents of cacao black pod rot, underwent increases in genome sizes and gene numbers by different mechanisms. *Genome Biology and Evolution*, *9*(3), 536–557.

Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F. J., … Schatz, M. C. (2019). RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biology*, *20*(1), 224.

Altschul, S. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.

Arnold, S. E. J., Forbes, S. J., Hall, D. R., Farman, D. I., Bridgemohan, P., Spinelli, G. R., … Stevenson, P. C. (2019). Floral odors and the interaction between pollinating Ceratopogonid midges and cacao. *Journal of Chemical Ecology*, *45*(10), 869–878.

Bach, S. S., Bassard, J.-É., Andersen-Ranberg, J., Møldrup, M. E., Simonsen, H. T., & Hamberger, B. (2014). High-throughput testing of terpenoid biosynthesis candidate genes using transient expression in Nicotiana benthamiana. *Methods in Molecular Biology (Clifton, N.J.)*, *1153*, 245–255.

Badet, T., & Croll, D. (2020). The rise and fall of genes: origins and functions of plant pathogen pangenomes. *Current Opinion in Plant Biology*, *56*, 65–73.

Bailey, B. A., & Meinhardt, L. W. (Eds.). (2018). *Cacao diseases*. Cham, Switzerland: Springer International Publishing.

Bell, E., Creelman, R. A., & Mullet, J. E. (1995). A chloroplast lipoxygenase is required for wound-induced jasmonic acid accumulation in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(19), 8675–8679.

Bellis, E. S., Kelly, E. A., Lorts, C. M., Gao, H., DeLeo, V. L., Rouhan, G., … Lasky, J. R. (2020). Genomics of sorghum local adaptation to a parasitic plant. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(8), 4243–4251.

Benedetti, M., Verrascina, I., Pontiggia, D., Locci, F., Mattei, B., De Lorenzo, G., & Cervone, F. (2018). Four Arabidopsis berberine bridge enzyme-like proteins are specific oxidases that inactivate the elicitor-active oligogalacturonides. *The Plant Journal: For Cell and Molecular Biology*, *94*(2), 260–273.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300.

Bhattarai, K. K., Atamian, H. S., Kaloshian, I., & Eulgem, T. (2010). WRKY72-type transcription factors contribute to basal immunity in tomato and Arabidopsis as well as gene-for-gene resistance mediated by the tomato R gene Mi-1. *The Plant Journal: For Cell and Molecular Biology*, *63*(2), 229–240.

Bhuiyan, N. H., Selvaraj, G., Wei, Y., & King, J. (2009). Gene expression profiling and silencing reveal that monolignol biosynthesis plays a critical role in penetration defence in wheat against powdery mildew invasion. *Journal of Experimental Botany*, *60*(2), 509–521.

Boetzer, M., & Pirovano, W. (2012). Toward almost closed genomes with GapFiller. *Genome Biology*, *13*(6), R56.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114–2120.

Boza, E. J., Motamayor, J. C., Amores, F. M., Cedeño-Amador, S., Tondo, C. L., Livingstone, D. S., … Gutiérrez, O. A. (2014). Genetic characterization of the cacao cultivar CCN 51: Its impact and significance on global cacao improvement and production. *Journal of the American Society for Horticultural Science. American Society for Horticultural Science*, *139*(2), 219–229.

Calderón, A. I., Wright, B. J., Hurst, W. J., & van Breemen, R. B. (2009). Screening antioxidants using LC-MS: case study with cocoa. *Journal of Agricultural and Food Chemistry*, *57*(13), 5693–5699.

Campbell, M. S., Law, M., Holt, C., Stein, J. C., Moghe, G. D., Hufnagel, D. E., … Yandell, M. (2014). MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiology*, *164*(2), 513–524.

Chakraborty, M., Baldwin-Brown, J. G., Long, A. D., & Emerson, J. J. (2016). Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Research*, *44*(19), e147.

Chen, Y., Ye, W., Zhang, Y., & Xu, Y. (2015). High speed BLASTN: an accelerated MegaBLAST search tool. *Nucleic Acids Research*, *43*(16), 7762–7768.

Chezem, W. R., Memon, A., Li, F.-S., Weng, J.-K., & Clay, N. K. (2017). SG2-type R2R3-MYB transcription factor MYB15 controls defense-induced lignification and basal immunity in Arabidopsis. *The Plant Cell*, *29*(8), 1907–1926.

Choudhury, A., Hazelhurst, S., Meintjes, A., Achinike-Oduaran, O., Aron, S., Gamieldien, J., … Ramsay, M. (2014). Population-specific common SNPs reflect demographic histories and highlight regions of genomic plasticity with functional relevance. *BMC Genomics*, *15*(1), 437.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics (Oxford, England)*, *21*(18), 3674–3676.

Corley, S. M., Troy, N. M., Bosco, A., & Wilkins, M. R. (2019). QuantSeq. 3' Sequencing combined with Salmon provides a fast, reliable approach for high throughput RNA expression analysis. *Scientific Reports*, *9*(1), 18895.

Cornejo, O. E., Yee, M.-C., Dominguez, V., Andrews, M., Sockell, A., Strandberg, E., … Motamayor, J. C. (2018). Population genomic analyses of the chocolate tree, Theobroma cacao L., provide insights into its domestication process. *Communications Biology*, *1*(1), 167.

De Vos, R. C. H., Moco, S., Lommen, A., Keurentjes, J. J. B., Bino, R. J., & Hall, R. D. (2007). Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nature Protocols*, *2*(4), 778–791.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, *7*(10), e1002195.

Ellinghaus, D., Kurtz, S., & Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, *9*(1), 18.

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238.

Evans, H. C. (2016). Frosty Pod Rot (Moniliophthora roreri). In *Cacao Diseases* (pp. 63–96). Cham: Springer International Publishing.

Feuillet, C., Schachermayr, G., & Keller, B. (1997). Molecular cloning of a new receptor-like kinase gene encoded at the Lr10 disease resistance locus of wheat. *The Plant Journal: For Cell and Molecular Biology*, *11*(1), 45–52.

Fister, A. S., Leandro-Muñoz, M. E., Zhang, D., Marden, J. H., Tiffin, P., dePamphilis, C., … Guiltinan, M. J. (2020). Widely distributed variation in tolerance to Phytophthora palmivora in four genetic groups of cacao. *Tree Genetics & Genomes*, *16*(1). doi:10.1007/s11295-019-1396-8

Fister, A. S., O'Neil, S. T., Shi, Z., Zhang, Y., Tyler, B. M., Guiltinan, M. J., & Maximova, S. N. (2015). Two Theobroma cacao genotypes with contrasting pathogen tolerance show aberrant transcriptional and ROS responses after salicylic acid treatment. *Journal of Experimental Botany*, *66*(20), 6245–6258.

Fister, A. S., Shi, Z., Zhang, Y., Helliwell, E. E., Maximova, S. N., & Guiltinan, M. J. (2016). Protocol: transient expression system for functional genomics in the tropical tree Theobroma cacao L. *Plant Methods*, *12*(1), 19.

Fitzgerald, D. J., Stratford, M., Gasson, M. J., Ueckert, J., Bos, A., & Narbad, A. (2004). Mode of antimicrobial action of vanillin against Escherichia coli, Lactobacillus plantarum and Listeria innocua. *Journal of Applied Microbiology*, *97*(1), 104–113.

Gumtow, R., Wu, D., Uchida, J., & Tian, M. (2018). A Phytophthora palmivora extracellular cystatin-like protease inhibitor targets papain to contribute to virulence on papaya. *Molecular Plant-Microbe Interactions: MPMI*, *31*(3), 363–373.

Gutiérrez, O. A., Puig, A. S., Phillips-Mora, W., Bailey, B. A., Ali, S. S., Mockaitis, K., … Motamayor, J. C. (2021). SNP markers associated with resistance to frosty pod and black pod rot diseases in an F1 population of Theobroma cacao L. *Tree Genetics & Genomes*, *17*(3). doi:10.1007/s11295-021-01507-w

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512.

Hämälä, T., Guiltinan, M. J., Marden, J. H., Maximova, S. N., dePamphilis, C. W., & Tiffin, P. (2020). Gene expression modularity reveals footprints of polygenic adaptation in Theobroma cacao. *Molecular Biology and Evolution*, *37*(1), 110–123.

Hämälä, T., Wafula, E. K., Guiltinan, M. J., Ralph, P. E., dePamphilis, C. W., & Tiffin, P. (2021). Genomic structural variants constrain and facilitate adaptation in natural populations of Theobroma cacao, the chocolate tree. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(35), e2102914118.

Han, Y., & Wessler, S. R. (2010). MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Research*, *38*(22), e199.

Hartl, D. L., & Clark, A. G. (2007). *Principles of population genetics* (4th ed.). New York, NY: Oxford University Press.

Hockings, K. J., Yamakoshi, G., & Matsuzawa, T. (2017). Dispersal of a Human-Cultivated Crop by Wild Chimpanzees (Pan troglodytes verus) in a Forest–Farm Matrix. *International Journal of Primatology*, *38*(2), 172–193.

Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, *12*(1), 491.

Hsieh, P., Hallmark, B., Watkins, J., Karafet, T. M., Osipova, L. P., Gutenkunst, R. N., & Hammer, M. F. (2017). Exome sequencing provides evidence of polygenic adaptation to a fat-rich animal diet in indigenous Siberian populations. *Molecular Biology and Evolution*, *34*(11), 2913–2926.

Hua, L., Stevenson, S. R., Reyna-Llorens, I., Xiong, H., Kopriva, S., & Hibberd, J. M. (2021). The bundle sheath of rice is conditioned to play an active role in water transport as well as sulfur assimilation and jasmonic acid synthesis. *The Plant Journal: For Cell and Molecular Biology*, *107*(1), 268–286.

Hudson, R. R., Slatkin, M., & Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics*, *132*(2), 583–589.

Jackman, S. D., Coombe, L., Chu, J., Warren, R. L., Vandervalk, B. P., Yeo, S., … Birol, I. (2018). Tigmint: Correcting assembly errors using linked reads from large molecules. doi:10.1101/304253

Khan, F., Bamunuarachchi, N. I., Tabassum, N., & Kim, Y.-M. (2021). Caffeic acid and its derivatives: Antimicrobial drugs toward microbial pathogens. *Journal of Agricultural and Food Chemistry*, *69*(10), 2979–3004.

Knollenberg, B. J., Li, G.-X., Lambert, J. D., Maximova, S. N., & Guiltinan, M. J. (2020). Clovamide, a Hydroxycinnamic Acid Amide, Is a Resistance Factor Against Phytophthora spp. in Theobroma cacao. *Frontiers in Plant Science*, *11*, 617520.

Koenig, D., Hagmann, J., Li, R., Bemm, F., Slotte, T., Neuffer, B., … Weigel, D. (2019). Long-term balancing selection drives evolution of immunity genes in Capsella. *ELife*, *8*. doi:10.7554/eLife.43606

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*, 59.

Kourelis, J., & van der Hoorn, R. A. L. (2018). Defended to the nines: 25 years of resistance gene cloning identifies nine mechanisms for R protein function. *The Plant Cell*, *30*(2), 285–299.

Kremling, K. A. G., Chen, S.-Y., Su, M.-H., Lepak, N. K., Romay, M. C., Swarts, K. L., … Buckler, E. S. (2018). Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature*, *555*(7697), 520–523.

Kuhn, R. M., Haussler, D., & Kent, W. J. (2013). The UCSC genome browser and associated tools. *Briefings in Bioinformatics*, *14*(2), 144–161.

Lanaud, C., Fouet, O., Clément, D., Boccara, M., Risterucci, A. M., Surujdeo-Maharaj, S., … Argout, X. (2009). A meta–QTL analysis of disease resistance traits of Theobroma cacao L. *Molecular Breeding: New Strategies in Plant Improvement*, *24*(4), 361–374.

Lazo, G. R., Stein, P. A., & Ludwig, R. A. (1991). A DNA transformation-competent Arabidopsis genomic library in Agrobacterium. *Bio/Technology*, *9*(10), 963–967.

Lebedenko, E. N., Birikh, K. R., Plutalov, O. V., & Berlin YuA. (1991). Method of artificial DNA splicing by directed ligation (SDL). *Nucleic Acids Research*, *19*(24), 6757–6761.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)*, *27*(21), 2987–2993.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Retrieved from http://arxiv.org/abs/1303.3997

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079.

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, *30*(7), 923–930.

Livingstone, D., 3rd, Stack, C., Mustiga, G. M., Rodezno, D. C., Suarez, C., Amores, F., … Motamayor, J. C. (2017). A larger chocolate chip-development of a 15K Theobroma cacao L. snp array to create high-density linkage maps. *Frontiers in Plant Science*, *8*, 2008.

Locci, F., Benedetti, M., Pontiggia, D., Citterico, M., Caprari, C., Mattei, B., … De Lorenzo, G. (2019). An Arabidopsis berberine bridge enzyme-like protein specifically oxidizes cellulose oligomers and plays a role in immunity. *The Plant Journal: For Cell and Molecular Biology*, *98*(3), 540–554.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550.

Ma, L., Lukasik, E., Gawehns, F., & Takken, F. L. W. (2012). The use of agroinfiltration for transient expression of plant resistance and fungal effector proteins in Nicotiana benthamiana leaves. *Methods in Molecular Biology (Clifton, N.J.)*, *835*, 61–74.

Mangelsdorf, P. C. (1986). The origin of corn. *Scientific American*, *255*(2), 80–86.

Manjang, K., Tripathi, S., Yli-Harja, O., Dehmer, M., & Emmert-Streib, F. (2020). Graph-based exploitation of gene ontology using GOxploreR for scrutinizing biological significance. *Scientific Reports*, *10*(1), 16672.

Maximova, S., Miller, C., Antúnez de Mayolo, G., Pishak, S., Young, A., & Guiltinan, M. J. (2003). Stable transformation of Theobroma cacao L. and influence of matrix attachment regions on GFP expression. *Plant Cell Reports*, *21*(9), 872–883.

Maximova, S. N., Young, A., Pishak, S., Miller, C., Traore, A., & Guiltinan, M. J. (2005). Integrated System for Propagation of Theobroma cacao L. In *Protocol for Somatic Embryogenesis in Woody Plants* (pp. 209–227). Berlin/Heidelberg: Springer-Verlag.

Maximova, Siela N., Marelli, J.-P., Young, A., Pishak, S., Verica, J. A., & Guiltinan, M. J. (2006). Over-expression of a cacao class I chitinase gene in Theobroma cacao L. enhances resistance against the pathogen, Colletotrichum gloeosporioides. *Planta*, *224*(4), 740–749.

Mchau, G. R. A., & Coffey, M. D. (1994). Isozyme diversity in Phytophthora palmivora: evidence for a southeast Asian centre of origin. *Mycological Research*, *98*(9), 1035–1043.

Menden, B., Kohlhoff, M., & Moerschbacher, B. M. (2007). Wheat cells accumulate a syringyl-rich lignin during the hypersensitive resistance response. *Phytochemistry*, *68*(4), 513–520.

Mitsuhara, I., Ugaki, M., Hirochika, H., Ohshima, M., Murakami, T., Gotoh, Y., … Ohashi, Y. (1996). Efficient promoter cassettes for enhanced expression of foreign genes in dicotyledonous and monocotyledonous plants. *Plant & Cell Physiology*, *37*(1), 49–59.

Morales-Cruz, A., Ali, S. S., Minio, A., Figueroa-Balderas, R., García, J. F., Kasuga, T., … Cantu, D. (2020). Independent whole-genome duplications define the architecture of the genomes of the devastating West African cacao black pod pathogen Phytophthora megakarya and its close relative Phytophthora palmivora. *G3 (Bethesda, Md.)*, *10*(7), 2241–2255.

Motamayor, J. C., Lachenaud, P., da Silva e Mota, J. W., Loor, R., Kuhn, D. N., Brown, J. S., & Schnell, R. J. (2008). Geographic and genetic population differentiation of the amazonian chocolate tree (Theobroma cacao L). *PloS One*, *3*(10), e3311.

Motamayor, J. C., Mockaitis, K., Schmutz, J., Haiminen, N., Livingstone, D., 3rd, Cornejo, O., … Kuhn, D. N. (2013). The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biology*, *14*(6), r53.

Mukhtar, M. S., Deslandes, L., Auriac, M.-C., Marco, Y., & Somssich, I. E. (2008). The Arabidopsis transcription factor WRKY27 influences wilt disease symptom development caused by Ralstonia solanacearum. *The Plant Journal: For Cell and Molecular Biology*, *56*(6), 935–947.

Muroi, A., Ishihara, A., Tanaka, C., Ishizuka, A., Takabayashi, J., Miyoshi, H., & Nishioka, T. (2009). Accumulation of hydroxycinnamic acid amides induced by pathogen infection and identification of agmatine coumaroyltransferase in Arabidopsis thaliana. *Planta*, *230*(3), 517–527.

Návarová, H., Bernsdorff, F., Döring, A.-C., & Zeier, J. (2012). Pipecolic acid, an endogenous mediator of defense amplification and priming, is a critical regulator of inducible plant immunity. *The Plant Cell*, *24*(12), 5123–5141.

Ploetz, R. C. (2007). Cacao diseases: important threats to chocolate production worldwide. *Phytopathology*, *97*(12), 1634–1639.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., & Lopez, R. (2005). InterProScan: protein domains identifier. *Nucleic Acids Research*, *33*(Web Server issue), W116-20.

Rodrigues Oblessuc, P., Vaz Bisneta, M., & Melotto, M. (2019). Common and unique Arabidopsis proteins involved in stomatal susceptibility to Salmonella enterica and Pseudomonas syringae. *FEMS Microbiology Letters*, *366*(16). doi:10.1093/femsle/fnz197

Sarrion-Perdigones, A., Vazquez-Vilar, M., Palací, J., Castelijns, B., Forment, J., Ziarsolo, P., … Orzaez, D. (2013). GoldenBraid 2.0: a comprehensive DNA assembly framework for plant synthetic biology. *Plant Physiology*, *162*(3), 1618–1631.

Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, *9*(7), 671–675.

Schweizer, R. M., Velotta, J. P., Ivy, C. M., Jones, M. R., Muir, S. M., Bradburd, G. S., … Cheviron, Z. A. (2019). Physiological and genomic evidence that selection on the transcription factor Epas1 has altered cardiovascular function in high-altitude deer mice. *PLoS Genetics*, *15*(11), e1008420.

Stahl, E. A., Dwyer, G., Mauricio, R., Kreitman, M., & Bergelson, J. (1999). Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. *Nature*, *400*(6745), 667–671.

Stam, R., Silva-Arias, G. A., & Tellier, A. (2019). Subsets of NLR genes show differential signatures of adaptation during colonization of new habitats. *The New Phytologist*, *224*(1), 367–379.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research*, *34*(Web Server issue), W435-9.

Steinbiss, S., Willhoeft, U., Gremme, G., & Kurtz, S. (2009). Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Research*, *37*(21), 7002–7013.

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS One*, *6*(7), e21800.

Torres, G. A., Sarria, G. A., Martinez, G., Varon, F., Drenth, A., & Guest, D. I. (2016). Bud rot caused by Phytophthora palmivora: A destructive emerging disease of oil palm. *Phytopathology*, *106*(4), 320–329.

Troyer, A. F. (1990). A retrospective view of corn genetic resources. *The Journal of Heredity*, *81*(1), 17–24.

Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., … Arita, M. (2015). MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature Methods*, *12*(6), 523–526.

Valla, S., & Lale, R. (Eds.). (2016). *DNA cloning and assembly methods*. New York, NY: Humana Press.

Vanholme, R., Cesarino, I., Rataj, K., Xiao, Y., Sundin, L., Goeminne, G., … Boerjan, W. (2013). Caffeoyl shikimate esterase (CSE) is an enzyme in the lignin biosynthetic pathway in Arabidopsis. *Science (New York, N.Y.)*, *341*(6150), 1103–1106.

Wall, P. K., Leebens-Mack, J., Müller, K. F., Field, D., Altman, N. S., & dePamphilis, C. W. (2008). PlantTribes: a gene and gene family resource for comparative genomics in plants. *Nucleic Acids Research*, *36*(Database issue), D970-6.

Wang, G.-F., He, Y., Strauch, R., Olukolu, B. A., Nielsen, D., Li, X., & Balint-Kurti, P. J. (2015). Maize homologs of hydroxycinnamoyltransferase, a key enzyme in lignin biosynthesis, bind the nucleotide binding leucine-rich repeat Rp1 proteins to modulate the defense response. *Plant Physiology*, *169*(3), 2230–2243.

Wang, M., Zhu, X., Wang, K., Lu, C., Luo, M., Shan, T., & Zhang, Z. (2018). A wheat caffeic acid 3-O-methyltransferase TaCOMT-3D positively contributes to both resistance to sharp eyespot disease and stem mechanical strength. *Scientific Reports*, *8*(1). doi:10.1038/s41598-018-24884-0

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., & Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome Research*, *27*(5), 757–767.

Widmer, T. L., & Laurent, N. (2006). Plant extracts containing caffeic acid and rosmarinic acid inhibit zoospore germination of Phytophthora spp. pathogenic to Theobroma cacao. *European Journal of Plant Pathology*, *115*(4), 377–388.

Winkelmüller, T. M., Entila, F., Anver, S., Piasecka, A., Song, B., Dahms, E., … Tsuda, K. (2021). Gene expression evolution in pattern-triggered immunity within Arabidopsis thaliana and across Brassicaceae species. *The Plant Cell*, *33*(6), 1863–1887.

Yamauchi, K., Yasuda, S., & Fukushima, K. (2002). Evidence for the biosynthetic pathway from sinapic acid to syringyl lignin using labeled sinapic acid with stable isotope at both

methoxy groups in Robinia pseudoacacia and Nerium indicum. *Journal of Agricultural and Food Chemistry*, *50*(11), 3222–3227.

Yeo, S., Coombe, L., Warren, R. L., Chu, J., & Birol, I. (2018). ARCS: scaffolding genome drafts with linked reads. *Bioinformatics*, *34*(5), 725–731.

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., … Wang, J. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. *Science (New York, N.Y.)*, *329*(5987), 75–78.

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., … Huang, X. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature Genetics*, *50*(2), 278–284.

Zhu, Y. J., Qiu, X., Moore, P. H., Borth, W., Hu, J., Ferreira, S., & Albert, H. H. (2003). Systemic acquired resistance induced by BTH in papaya. *Physiological and Molecular Plant Pathology*, *63*(5), 237–248.

Zou, X., Long, J., Zhao, K., Peng, A., Chen, M., Long, Q., … Chen, S. (2019). Overexpressing GH3.1 and GH3.1L reduces susceptibility to Xanthomonas citri subsp. citri by repressing auxin signaling in citrus (Citrus sinensis Osbeck). *PloS One*, *14*(12), e0220017.

Zumbo, P. (1932). Ethanol precipitation. *Weill Cornell Medical College*, 1–12.

Chapter 3: A Conserved Set of Orthologous Genes are Involved in Defense Against *Phytophthora palmivora* across *Theobroma* species

Abstract

The oomycete pathogen *Phytophthora palmivora* is responsible for extensive annual yield loss in a variety of crop species. Among those species is *Theobroma cacao*, the tree from which chocolate is derived. Natural variation in resistance to *P. palmivora* is well documented across cacao lineages, but little is known about resistance in its wild, non-cacao relatives. In this study, we used non-cacao Theobroma species to investigate the evolution of defense response across *Theobroma.* We discovered both lineage-specific and conserved aspects of defense response, including upregulation of the phenylpropanoid pathway. Of particular interest was *TcBBE8* and *TcWRKY29*, a pair of genes that were upregulated across five species of *Theobroma* and displayed evidence of positive selection. Together, our results suggest some aspects of defense against *P. palmivora* are orthologous and are, therefore, fundamentally important to defense across *Theobroma.*

Introduction

Theobroma cacao L., a tropical understory plant native to the Amazon basin (Harry C. Evans 2016a; D. Zhang and Motilal 2016) and the tree from which chocolate is derived, forms the basis of a market worth approximately $100 billion per year (Ploetz 2007; Bailey and Meinhardt 2016). Nearly 70% of the world's cacao is grown by small-holding farmers in just three countries: Ghana, Côte d'Ivoire, and Indonesia (ICCO). However, nearly 40% of pre-harvest yield is lost annually due to a variety of pests and pathogens, causing economic hardship for millions of farmers (Wood and Lass 2001). One such devastating pathogen is the hemibiotrophic oomycete

Phytophthora palmivora (E.J. Butler). Native to Southeast Asia (Mchau and Coffey 1994; Jianan Wang et al. 2020), *P. palmivora* is one of four main *Phytophthora* species that cause black pod rot, a disease characterized by necrosis of pod tissue and the seeds contained inside (Acebo-Guerrero, Hernández-Rodríguez, Heydrich-Pérez, El Jaziri, & Hernández-Lauzardo, 2012). Developing approaches to mitigate *P. palmivora*-mediated yield loss is therefore critical for both farmers and chocolate companies alike.

One such approach involves breeding superior cacao varieties that are high yielding, fine flavor, and resistant to *P. palmivora.* While disease resistant clones exist, most contemporary breeding programs have focused on a handful clones from the Pound collection (Boza et al. 2014; Gutiérrez et al. 2021; D. Zhang and Motilal 2016). This has resulted in underutilization of cacao's broader genetic diversity, most of which exists in its ancestral state due to limited human intervention (Cornejo et al. 2018). This wild germplasm exists in two forms. The first comprises ten geographically and genetically isolated populations of *T. cacao* spread across South America (Juan C. Motamayor et al. 2008; D. Zhang and Motilal 2016). These populations are, for the most part, strongly differentiated (Cornejo et al. 2018; Hämälä et al. 2020), suggesting they may have evolved lineage-specific disease resistance mechanisms that would be valuable to breeders (Chapter 2).

The second form of wild germplasm exists as 21 non-cacao *Theobroma* species*,* all of which are also native to the Amazon basin and Southern Mexico (Cuatrecasas 1964; Richardson et al. 2015; D. Zhang et al. 2011). Despite their potential economic importance, little research has been conducted on these 21 *Theobroma* species (D. Zhang et al. 2011; B. A. Bailey and Meinhardt 2018). This is likely because interspecific barriers to hybridization prevent wild relatives from being incorporated into *T. cacao* breeding programs (Silva, Venturieri, & Figueira, 2004). Focusing solely on hybridization barriers, however, ignores their potential evolutionary utility. That is, the application of knowledge gained from examining the evolution of defense

responses across *Theobroma* species could be used to improve *T. cacao* through breeding or genetic modification. In this study, we test whether response to *P. palmivora* attack is mediated by orthologous genes, or has evolved independently in multiple lineages of *Theobroma.* The underlying assumption is that at least some portion of resistance to *P. palmivora* is monophyletic, i.e. arose in a species ancestral to *Theobroma*, despite the fact that *T. cacao* and *P. palmivora* did not co-evolve.

Through RNA sequencing and molecular evolutionary analyses, we have identified gene families that respond consistently to pathogen challenge across four *Theobroma* species with contrasting levels of resistance to *P. palmivora*: *T. angustifolium, T. bicolor, T. grandiflorum,* and *T. mammosum*. Together, our results support wild *Theobroma spp.* as a potentially valuable source of underutilized germplasm, particularly when used to identify core defense mediators in the agriculturally important crop *T. cacao.*

Materials and Methods

Plant phenotyping and sample selection

We examined resistance to the pathogenic oomycete *P. palmivora* across non-cacao *Theobroma spp.* present in the *ex situ* germplasm collection at the Tropical Agricultural Research and Higher Education Center (CATIE) in Turrialba, Costa Rica (Table 3.1). We tested seven species for resistance to *P. palmivora*: *T. mammosum* (Andropetalum)*, T. angustifolium* (Glossopetalum)*, T. grandiflorum* (Glossopetalum)*, T. simiarum* (Glossopetalum)*, T. speciosum* (Oreanthes)*, T. bicolor* (Rhytidocarpus)*,* and *T. microcarpum* (Telmatocarpus)*.* Leaves were sampled over the course of two weeks from June to July 2019. Disease assays were conducted as previously described (Fister et al. 2020). Briefly, leaves were collected from adult trees between

0800 and 1000 each morning using a metal poleaxe. Care was taken to select leaves that did not display any visible signs of damage or pathogen infection. Because leaf developmental timeline is less well-characterized for non-cacao *Theobroma spp.* than it is for *T. cacao,* we had to use fully mature leaves for disease assays. Once collected, leaves were placed in Ziplock bags with wet paper towel and brought back to the lab. Leaves were washed with tap water and tissue from the leaf apex and petiole were cut so leaf sections would fit inside a petri dish. To prevent desiccation of the leaf sections, wounds were sealed with molten parafilm. Likewise, petri dishes were filled with sterile paper towels and wetted with sterile water. Leaves were placed in the petri dishes adaxial-side down. Three plugs from the leading edge of 2-day old *Phytophthora palmivora* (strain C-14) mycelia grown on V8 media (Fister, Shi, et al. 2016) were placed on the right side of the leaf's midvein. As controls, three plugs of V8 media (no mycelia) were placed on the left side of the midvein as well. A total of 204 inoculations were performed across the seven species we examined. Since we are primarily interested in transcriptional differences that lead to resistance or susceptibility, we attempted to minimize the effect of differences in cuticle thickness by scoring the abaxial side of the leaves with a razor blade prior to pathogen challenge. Once challenged with pathogen, leaves were wetted with sterile water, petri dishes were sealed with parafilm, and placed in an incubator at 27 °C. After 48 hours, petri dishes were removed from the incubator and photographed. Lesion area was measured using ImageJ. The two most resistance and two most susceptible species from each clade were chosen for further experimentation (Figure 3-1).

Transcriptome experimental design

The two most resistant and two most susceptible species based on the phenotyping experiments (Figure 3-1) were carried forward for further transcriptome analysis. The

transcriptome experiment followed a split-plot design, where tree, the more difficult factor to randomize, was treated as the blocking factor (Figure 3-2). Over three consecutive days, we sampled leaves from a single tree for each species. From each tree, we took four leaves, two for *P. palmivora* treatment and two for controls. Leaves were sampled, processed, and challenged with pathogen similar to the disease phenotyping assay outline above. The only exception being that leaves were treated with either *P. palmivora* plugs or control plugs, not both. Within each tree, we randomized the order in which we processed each species. A cork borer was used to punch out leaf discs surrounding the necrotic lesion area 48 hours post inoculation. Leaf discs were then put into 2 mL cryovial tubes and flash frozen using liquid nitrogen.

RNA extraction, and sequencing

RNA extraction followed the protocol outlined in Chapter 2. Briefly, frozen tissue was ground in pre-chilled mortar and pestles and the fine powder was stored at -80 °C until use. We extracted RNA from 100 mg of frozen tissue and followed the protocol outlined in Thermo Fisher Scientific's small scale RNA isolation, with several modifications (Publication No. MAN0000243). First, 1 mL homemade RNA extraction buffer (US Patent US6875757B2) was substituted for 0.5 mL PureLink™ Plant RNA Reagent (Life Technologies, Carlsbad, CA, USA). Tissue was vortexed until it was completely homogenized in buffer. Second, 200 uL of NaCl was used rather than 100 uL. Third, 600 uL of chloroform was used rather than 300 uL for the first organic extraction. An additional chloroform extraction was performed using 1:1 chloroform:aqueous layer. Lastly, we performed 3 ethanol washes before drying the nucleic acid pellets for 10 min and resuspending them in VWR molecular grade water (VWR, Radnor, PA, USA). All spins were performed at 16,000 x g at 4 °C. After RNA extraction, DNA contamination was removed through treatment with DNase I (Publication No. MAN0012000).

Enzyme and buffers were removed after DNase I treatment using Zymo RNA Clean and Concentrator kits (Catalog #R1013; Zymo Research, Irvine, CA, USA). We determined RNA integrity and concentration using an Agilent 4200 Tapestation System. For both pathogen treatment and controls, we pooled RNA extracted from two separate leaves belonging to the same tree prior to sequencing (Figure 3-2).

All library construction and sequencing was done at the Pennsylvania State University Genomics Core Facility. Stranded, single end, 150 nt libraries were sequenced on two high output runs of an Illumina NextSeq 550. This generated approximately 30 million reads per sample and approximately 200 million reads per species (Table 3.2).

Transcriptome assembly, mapping, and expression quantification

Adapters and low quality bases were trimmed from the reads using Trimmomatic v0.38 (SE -phred33 ILLUMINACLIP:TruSeq3-SE:2:30:10 LEADING:3 SLIDINGWINDOW:4:15 MINLEN:50) (Bolger, Lohse, and Usadel 2014). Reads were then assembled into transcripts using Trinity v2.11.0 (--seqtype fq --single --SS_lib_type R --no_normalize_reads --no_cleanup --bflyHeapSpaceMax 20) (Haas et al. 2013). Transcripts were post-processed into putative coding sequences and their corresponding amino acids using TransDecoder (https://github.com/TransDecoder/TransDecoder) as implemented in the PlantTribes v2.0 AssemblyPostProcessor pipeline (Wall et al. 2008). Non-embyrophyte contaminants were then cleaned from predicted coding sequences using a BLAST-based procedure. First, predicted coding sequences were searched against the NCBI nonredundant (nr) database. The BLAST hits were then queried against NCBI's taxonomy database to assign taxonomic class. Finally, assembled sequences with top hits outside embryophyta (land plants) were discarded using a custom set of Bash and Perl scripts.

For a variety of reasons, including RNA degradation, genome heterozygosity, alternative splicing etc., transcriptome assemblies are often highly fragmented (Honaas et al. 2016). This can lead to multiple assembled transcripts originating from the same gene, which can cause redundant read mapping and inappropriate expression quantification. To address this problem, we created 'supertranscripts' by generating consensus sequences from transcripts belonging to the same Trinity cluster (Honaas et al. 2016). We first separated predicted coding sequences and amino acids by clusters, i.e. transcripts possessing identical IDs other than the isoform suffix. We then aligned each cluster of amino acid sequences with MAFFT v7.20 (L-INS-i) (Katoh et al. 2005). The coding sequences were then forced onto these amino acid alignments to create codon alignments. From each cluster, a set of hidden Markov models (HMM) were created from the amino acid and coding sequence alignments using HMMER v3.1b1 (Eddy 2011). The majority-rule consensus (>50%) sequence was then called from each HMM using hmmemit (-c -o). This consensus sequence represents a cluster's putative supertranscript. Finally, to remove premature stop codons and other potential artifacts that may have been introduced during supertranscript construction, putative supertranscripts were cleaned using the PlantTribes v2.0 PostAssemblyProcessor.

We assessed the quality of each assembly in two ways (Table 3.2 and Figure 3-3). First, we examined assembly summary statistics relative to a leaf transcriptome (*T. cacao* SPEC 54/1) that was assembled for genome annotation in Chapter 2. *T. cacao* SPEC 54/1 was chosen for comparison because: (1) it was assembled and cleaned according to the methods outlined above, and (2) SPEC 54/1 is known to be homozygous, an important factor in transcriptome assembly (Kajitani et al. 2014). Assembly summary statistics allow for valuable technical comparisons, but do not necessarily indicate how well a transcriptome's gene content has been captured. To estimate how completely we captured each species' gene space, we quantified the completeness

of basic universal single copy orthologs (BUSCO) for each species (Simão et al. 2015), again comparing our results to the *T. cacao* SPEC 54/1 transcriptome.

Differential expression analysis and gene ontology enrichment

Supertranscript abundance was quantified using Kallisto (-i -o -b 100 –single -l 200 -s 20 -t 5) (Bray, Pimentel, Melsted, & Pachter, 2016) and plugged directly into limma voom (Law, Chen, Shi, & Smyth, 2014) for differential expression analysis. Our experiment was implemented as a split-plot design with tree as a blocking factor. An unadjusted p-value cutoff of 0.05 was used to define supertranscripts as differentially expressed. We used unadjusted p-values rather than p-values corrected for multiple testing for two reasons. First, we were primarily interested in using the differential expression results to identify groups of orthologous genes that were responding consistently across species, rather than identify specific genes that may be important for disease resistance. When looking at sets of aggregated genes (e.g. GO terms, orthogroups) we are less worried about multiple test correction, since it is unlikely that we would observe an enriched GO term or shared orthogroup due to false positives alone. Moreover, if out differential expression results contained only false positives, we would expect between 850-1,000 differentially expressed supertranscripts for each species (assuming an $\alpha = 0.05$). Instead, we observed approximately 1,500-3,500 supertranscripts for each species, indicating the presence of true positives. Second, our small sample size and large standard error made FDR-adjusted p-values > 0.05 for most supertranscripts (Benjamini and Hochberg 1995).

Differentially expressed supertranscripts were assigned to GO terms according to their best hit in the SCA-6 genome, before being queried for functional enrichment according to the methods outlined in Chapter 2. Briefly, we began by performing gene ontology enrichment using Fisher's exact tests implemented in topGO v2.38.1 (algorithm = "classic", statistic = "fisher")

(Alexa & Rahnenführer, 2009). This resulted in a large list of significantly enriched (FDR-adjusted p-value < 0.05), but redundant, GO terms for each species. We limited this redundancy by exploiting the structure of each ontology's directed acyclic graph, prioritizing GO terms near the tips using GOxploreR (Manjang, Tripathi, Yli-Harja, Dehmer, & Emmert-Streib, 2020). Lastly, REVIGO's implementation of Lin's semantic similarity was used to collapse similar GO terms (Xiao et al. 2008; Z. Zhu et al. 2016). This resulted in a list of non-redundant, significantly enriched GO terms (Figure 3-4).

Orthogroup classification and resistance class assignment

The PlantTribes v2.0 GeneFamilyClassifier pipeline (--scaffold 37Gv1.0 --method OrthoFinder --classifier both) was used to classify supertranscripts into orthogroups, i.e. sets of genes inferred to have a single common ancestor among the species we were comparing. Orthogroups with at least one differentially expressed supertranscript were considered differentially expressed orthogroups. These differentially expressed orthogroups were then classified into 'core', 'shell, or 'cloud' resistance classes according to how broadly they were shared (Figure 3-6), nomenclature borrowed from Van de Weyer et al. 2019. Those that were shared across all four species were considered core. Those shared across two or three species were considered shell. And orthogroups differentially expressed within a single species were considered cloud.

Analysis of log_2 fold change across species

To gain a better understanding of how defense response evolved in *Theobroma*, and to better predict groups of genes that may be important for resistance specifically in *T. cacao*, we

compared orthogroup expression from the *T. cacao* transcriptome results presented in Chapter 2 to our non-cacao *Theobroma spp*. Mean orthogroup $\log_2$ fold change (LFC) for each *Theobroma spp.* was compared to mean LFC across all populations of *T. cacao*. Differentially expressed orthogroups that were strongly responsive (|LFC| > 1), shared across all four non-cacao *Theobroma spp.*, i.e. core, *and* also differentially expressed in at least one population from Chapter 2, were labeled as 'core & |LFC| > 1' and carried forward for further analysis (Table B.1 and Figure 3-7).

Measures of selection

We tested whether differentially expressed orthogroups (N = 48) shared across *Theobroma* were evolving under diversifying selection using HyPhy's branch-site unrestricted statistical test for episodic diversification (BUSTED) (--alignment --tree --branches --output). BUSTED is a branch-site method that, given a set of foreground and background branches, tests whether a subset of codons in a gene have undergone positive selection. It does so by first fitting two codon models to foreground and background branches. Each codon model contains three dN/dS (ω) classes, where dN/dS is the ratio of nonsynonymous substitutions per nonsynonymous sites (dN) to synonymous substitutions per synonymous sites (dS). In the first model, called the unconstrained model, positive selection is allowed (dN/dS > 1). In the second model, referred to as the null or constrained model, positive selection is not allowed (dN/dS ≤ 1). The unconstrained model fit is then compared to the constrained model fit using a likelihood ratio test. A significant result indicates that at least one codon on at least one of the foreground branches has experience positive selection (Figure 3-8).

We began by classifying all supertranscripts predicted during transcriptome assembly into orthogroups, as described above. From each orthogroup, we extracted sequences for all

Theobroma spp., as well as a subset of the species used for classification: *Elaeis guineensis* (Arecaceae), *Oryza sativa* (Poaceae), *Lactuca sativa* (Asteraceae), *Solanum lycopersicum* (Solanaceae), *Arabidopsis thaliana* (Brassicaceae), *Theobroma cacao* (Malvaceae), *Medicago truncatula* (Fabaceae), *Vitis vinifera* (Vitaceae), *Aquilegia coerulea* (Ranunculaceae), *Amborella trichopoda* (Amborellaceae). We then aligned each orthogroup at the amino acid level using the MAFFT v7.205 L-INS-I algorithm, unless a gene family was > 1000 sequences, in which case --auto was used (Katoh et al. 2005). The coding sequences were then forced onto the amino acids to create a codon alignment using a custom Perl script. To improve codon alignments, we trimmed columns that were primarily composed of gaps using TrimAl (-gappyout) (Capella-Gutiérrez, Silla-Martínez, & Gabaldón, 2009), and completely removed sequences that were composed of >70% gaps. Trees were built from each orthogroup alignment using FastTree v2.1.10 (-nt -gtr) (Price, Dehal, and Arkin 2010). Finally, BUSTED models were implemented using HyPhy (Pond et al., 2005). All *Theobroma spp.*, including *T. cacao*, were used as the foreground while all other species were used as background.

Results

Theobroma spp. displayed variation in disease resistance to Phytophthora palmivora

Seven species of *Theobroma* were tested for their resistance to *Phytophthora palmivora* (strain C-14) using detached leaf assays. These species spanned five of six sections of the genus *Theobroma,* excluding *T. cacao* (section Theobroma), as it has been extensively assayed in previous studies (Table 3.1) (Fister et al. 2020).

Table 3-1: All species within the genus *Theobroma* and their assigned section (Cuatrecasas, 1964). Adapted from Zhang et al. 2011.

Species	Section	Sampled in this study	Number of trees sampled	Number of leaves sampled
T. mammosum	Andropetalum	X	1	11
T. angustifolium	Glossopetalum	X	1	11
T. canumanense	Glossopetalum			
T. chocoense	Glossopetalum			
T. cirmolinae	Glossopetalum			
T. grandiflorum	Glossopetalum	X	3	15
T. hylaeum	Glossopetalum			
T. nemorale	Glossopetalum			
T. obovatum	Glossopetalum			
T. simiarum	Glossopetalum	X	2	9
T. sinuosum	Glossopetalum			
T. stipulatum	Glossopetalum			
T. subincanum	Glossopetalum			
T. bernouillii	Oreanthes			
T. glaucum	Oreanthes			
T. speciosum	Oreanthes	X	1	4
T. sylvestre	Oreanthes			
T. velutinum	Oreanthes			
T. bicolor	Rhytidocarpus	X	2	10
T. gileri	Telmatocarpus			
T. microcarpum	Telmatocarpus	X	1	8

T. cacao	Theobroma			

We performed a total of 204 inoculations, testing an average of 9.7 leaves per species. *T. angustifolium* displayed the greatest susceptibility to *P. palmivora* while *T. mammosum* displayed the greatest resistance, although mean lesion area was not significantly different between *T. mammosum, T. grandiflorum, T. microcarpum,* and *T. speciosum* (Welch's ANOVA: p-value < 0.001; Games-Howell post-hoc test: adjusted p-values < 0.05; Figure 3-1).

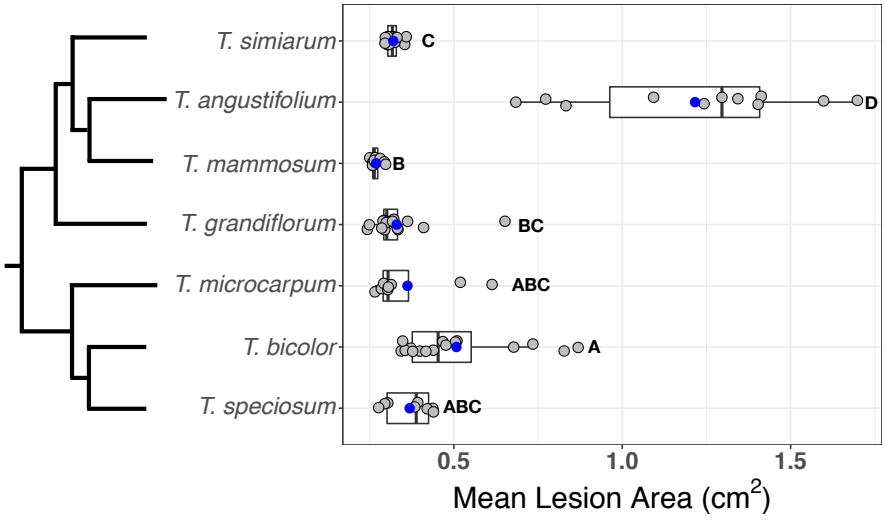


Figure 3-1: Variation in resistance to *P. palmivora* across *Theobroma.* Seven *Theobroma spp.* were assayed for resistance to *P. palmivora* strain C-14. Each dot represents the mean (n = 3) lesion area for an individual leaf. Letters indicate significant differences in mean lesion area (Welch's ANOVA, p-value < 0.001; Games-Howell post-hoc test, adjusted p-values < 0.05). Means are shown as blue dots. Species that share letters do not have significantly different means.

To test for the presence of conserved, resistance-related transcripts and orthogroups across *Theobroma*, we chose the two most resistant and the two most susceptible species (Figure 3-1) for further experimentation: *T. mammosum* (Andropetalum; Resistant), *T. grandiflorum* (Glossopetalum; Resistant), *T. angustifolium* (Glossopetalum; Susceptible), and *T. bicolor*

(Rhytidocarpus; Susceptible). We organized our experiment as a split-plot design with each species possessing three separate biological replicates (trees) (Figure 3-2).

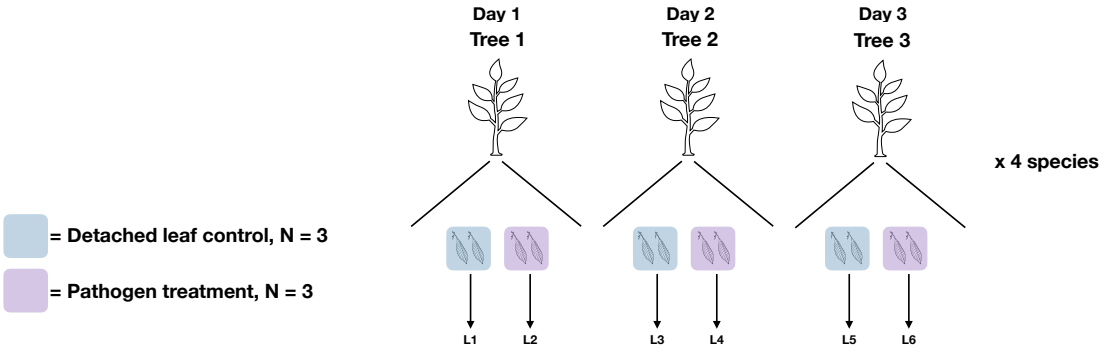


Figure 3-2: Split-plot design for RNA-seq experiment. Trees were sampled over three consecutive days. Each day, a single tree from each species was collected and processed. From each tree, two leaves were used for treatment with *P. palmivora* (purple) and two leaves were used for controls (blue). Leaves from the same tree and treatment combination were pooled before library preparation (L) and sequencing.

Supertranscript statistics reveal contiguous and complete transcriptome assemblies

After contaminant removal and supertranscript construction, each species had approximately 21,000 to 25,000 putative coding sequences (Table 3.2), close to the approximately 28,000 genes estimated from *T. cacao* genomic data.

Table 3-2: Transcriptome assembly statistics and BUSCO scores for each *Theobroma spp.* sampled.

	T. angustifolium	***T. bicolor***	***T. grandiflorum***	***T. mammosum***	***T. cacao* (SPEC 54/1)**
Assembly Statistics					
Total reads (millions)	184.6	172.6	182.3	173.6	—
Total plant transcripts	120,037	113,607	112,944	109,608	—

Mean transcript CDS length (bp)	689.07	640.73	666.25	731.22	—
Median transcript CDS length (bp)	510	492	492	531	—
Transcript N50	810	720	780	897	—
Total supergenes	24,691	24,709	21,090	23,805	25,335
Mean supergene CDS length (bp)	769.07	718.33	797.92	829.71	832.56
Median supergene CDS length (bp)	555	537	603	597	576
Supergene N50	975	870	1005	1080	1098
BUSCO (%)					
Complete	63.2	52.2	62.0	68.3	58.1
Single Copy	61.7	50.8	60.7	66.8	56.6
Duplicated	1.5	1.4	1.3	1.5	1.5
Fragmented	18.9	26.2	19.6	16.8	23.3
Missing	17.9	21.6	18.4	14.9	18.6

We assessed the quality of our assemblies relative to a *T. cacao* SPEC 54/1 leaf transcriptome assembled using the same methodology. We began by examining differences in the distributions of coding sequence lengths across our *Theobroma* species, an important metric for assessing transcriptome assembly contiguity. The *T. cacao* SPEC 54/1 transcriptome was significantly more contiguous than all other *Theobroma* assemblies, with the exception of *T. mammosum* (Welch's ANOVA: p-value< 0.001; Games-Howell post-hoc test: adjusted p-values < 0.01; Figure 3-3A). Some of this difference, however, was driven by the extremely large number of transcripts belonging to each species, which provides power to achieve statistical significance for even small differences in mean length. While technical measurements like mean coding sequence length and N50 are important indicators of assembly quality, they fail to assess completeness, i.e. the degree to which an assembly has adequately captured a species' gene content. To evaluate assembly completeness, we searched each transcriptome assembly for the presence of

basic universal single copy orthologs (BUSCO). BUSCO completeness was higher among non-cacao *Theobroma spp*. relative to *T. cacao* SPEC 54/1, the single exception being *T. bicolor.* For all comparisons, there was a significant association between species (*T. cacao* SPEC 54/1 x *Theobroma spp.*) and BUSCO completeness (chi-square goodness-of-fit: p-value < 0.05; Figure 3-3B). Thus, despite the fact our *Theobroma spp.* assemblies are slightly less contiguous than the *T. cacao* SPEC 54/1 assembly, our BUSCO scores suggest we were able to adequately assemble *Theobroma* gene space. Together, these metrics indicate we have assembled transcriptomes suitable for downstream analyses.

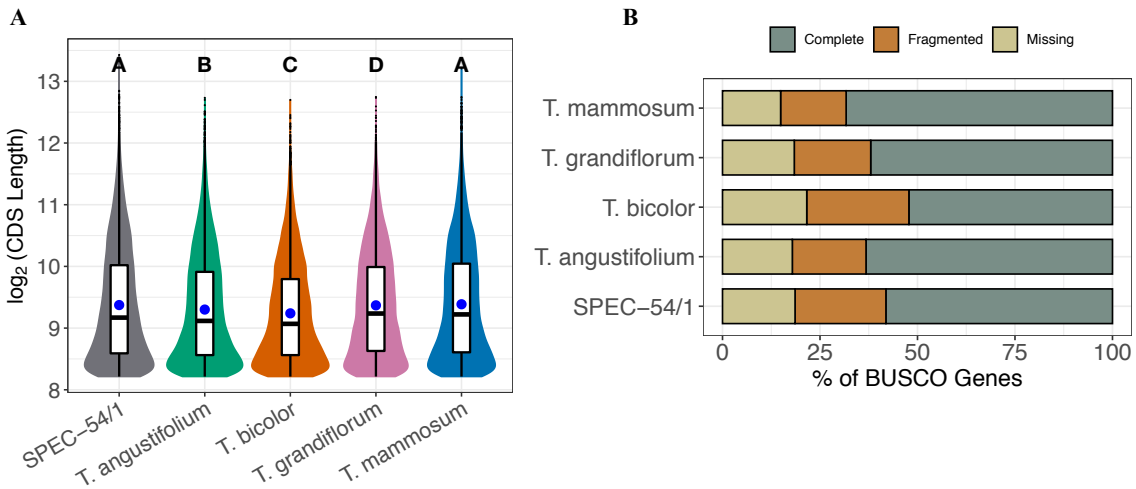


Figure 3-3: Transcriptome assembly quality and completeness metrics. (A) Coding sequence (CDS) length distributions for all four non-cacao *Theobroma spp.* and the reference transcriptome *T. cacao* (SPEC 54/1). Letters indicate significant differences in mean CDS length (Welch's ANOVA, p-value < 0.001; Games-Howell post-hoc test, adjusted p-values < 0.01). Means are shown as blue dots. (B) Proportion of complete, fragmented, and missing BUSCO genes for each non-cacao *Theobroma spp.* and the reference transcriptome *T. cacao* (SPEC 54/1). For all comparisons, there was a significant association between species (*T. cacao* SPEC 54/1 x *Theobroma spp.*) and BUSCO completeness (chi-square goodness-of-fit, p-values < 0.05).

Theobroma spp. displayed overlapping functional response to P. palmivora

Differential expression analysis revealed thousands of supertranscripts that were significantly responding to *P. palmivora* in each species (Figure 3-4A). *Theobroma mammosum* had the largest number of differentially expressed supertranscripts (3,324) while *T. angustifolium* had the lowest (1,613). An approximately equal number of upregulated and downregulated supertranscripts was observed for each species. There was a total of 93 significantly enriched gene ontology (GO) terms (Figure 3-4B). While there were many GO terms unique to each species, six GO terms were enriched across all four: 'response to salicylic acid' (GO:0009751), 'response to jasmonic acid' (GO:0009753), 'coumarin biosynthetic process' (GO:0009805), 'lignin biosynthetic process' (GO:0009809), 'oxidation-reduction process' (GO:0055114), and 'response to karrikin' (GO:0080167) (Figure 3-5).

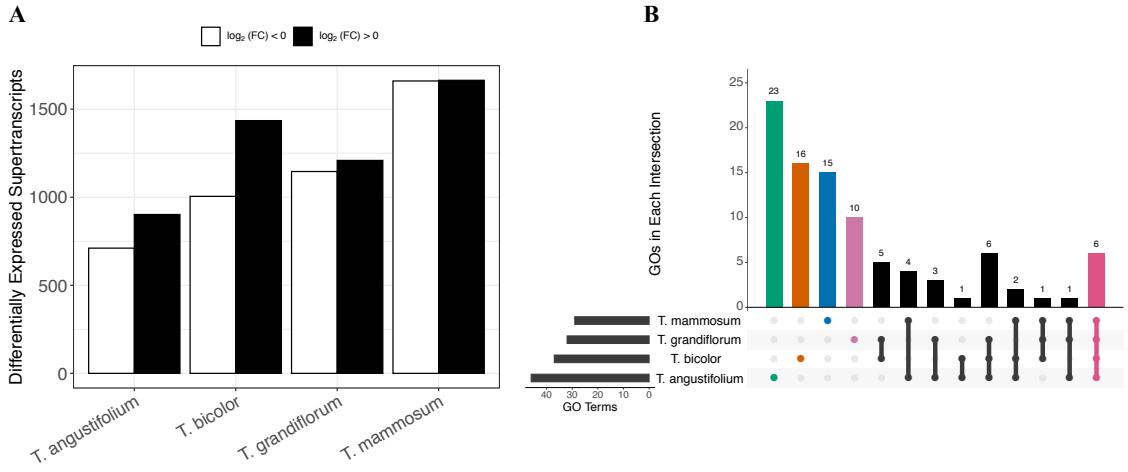


Figure 3-4: Differentially expressed genes and enriched gene ontology terms. (A) Differentially expressed supertranscripts for each species (unadjusted p-value < 0.05). White bars indicate downregulated supertranscripts and black bars indicate upregulated superstranscripts. (B) Overlap of significantly enriched GO terms (FDR-adjusted p-values < 0.05). The green, orange, purple, and blue bars represent GO terms that are only enriched in *T. angustifolium, T. bicolor, T. grandiflorum,* and *T. mammosum,* respectively. The pink bar indicates GO terms that are significantly enriched across all four species. Numbers above the bars indicate the number of GO terms in each specific intersection.

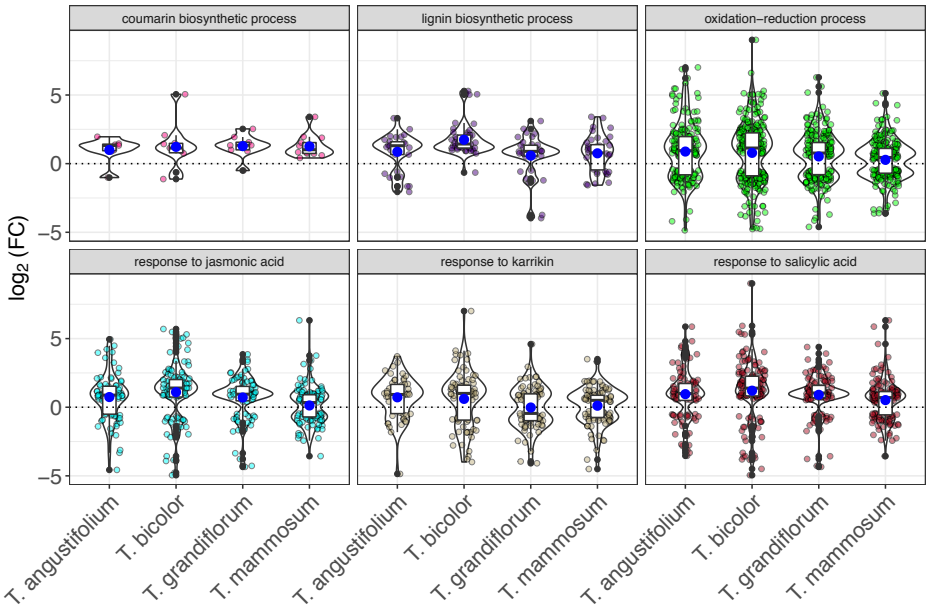


Figure 3-5: GO terms enriched across all four *Theobroma spp*. Boxplots display the distribution of $\log_2$ fold changes for each GO term, for each species. Each colored point represents the $\log_2$ fold change for a single differentially expressed supertranscript (unadjusted p-value < 0.05). Means are shown as blue dots.

Differentially expressed orthogroups are shared across Theobroma

The presence of shared functional response to *P. palmivora* across *Theobroma spp.* suggests, but does not necessarily guarantee, that some aspects of defense are mediated by orthologous genes. To estimate the proportion of defense that is orthologous, we classified supertranscripts into orthogroups using PlantTribes v2.0. We designated orthogroups as differentially expressed if they contained at least one differentially expressed supertranscript. Each differentially expressed orthogroup was then assigned to a resistance class based on its degree of overlap with the other three species. Orthogroups that were differentially expressed across all four species were called ‘core’. Those differentially expressed across 2-3 species were

called ‘shell’. And orthogroups that were differentially expressed in only a single species were called ‘cloud’. Most differentially expressed orthogroups were found in shell and cloud, with core having the smallest proportion (Figure 3-6A-B). With the exception of *T. mammosum*, there were a greater number of differentially expressed orthogroups that were shared among species than unique to a single species.

We set our criterion for defining differentially expressed orthogroups at one supertranscript, an admittedly liberal cutoff. This has the benefit of capturing expression conservation in small gene families, with potentially only a single supertranscript differentially expressed in each species. However, using a one supertranscript cutoff also increases the potential for false positives, especially when using unadjusted p-values to define differential expression. This problem is likely greatest for the cloud resistance class, which only requires a single differentially expressed supertranscript for inclusion. To test whether the pattern of orthogroup differential expression we observe is driven by false positives, we calculated the average number of supertranscripts that are differentially expressed in each orthogroup and resistance class (Figure 3-6C). Averages at or only slightly greater than one suggest most of the orthogroups in that resistance class were designated as differentially expressed based on evidence from only a single supertranscript. And, while differential expression of a single supertranscript could be authentic, the risk of false positive identification is high. For the cloud and shell resistance classes, the average number of differentially expressed supertranscripts in each orthogroup was very close to one (M_{CLOUD} = 1.09, SEM_{CLOUD} = 0.007; M_{SHELL} = 1.24, SEM_{SHELL} = 0.01; Figure 3-6C). This is consistent with the idea that most of the orthogroups assigned to these resistance classes were driven by a single supertranscript and were therefore likely false positives. The average number of differentially expressed supertranscripts was significantly higher for the core resistance class (M_{CORE} = 2.55, SEM_{CORE} = 0.20; Welch’s ANOVA: p-value< 0.001; Games-Howell post-hoc test: adjusted p-values < 0.05). Moreover, it is improbable that a single orthogroup would, by chance,

contain differentially expressed supertranscripts from four separate species. Thus while false positives likely present a problem for all resistance classes, the problem seemed to be most pronounced in the cloud and shell classes. Therefore, we focused the remainder of our analyses on the 317 core orthogroups, i.e. orthogroups that were differentially expressed across all four species.

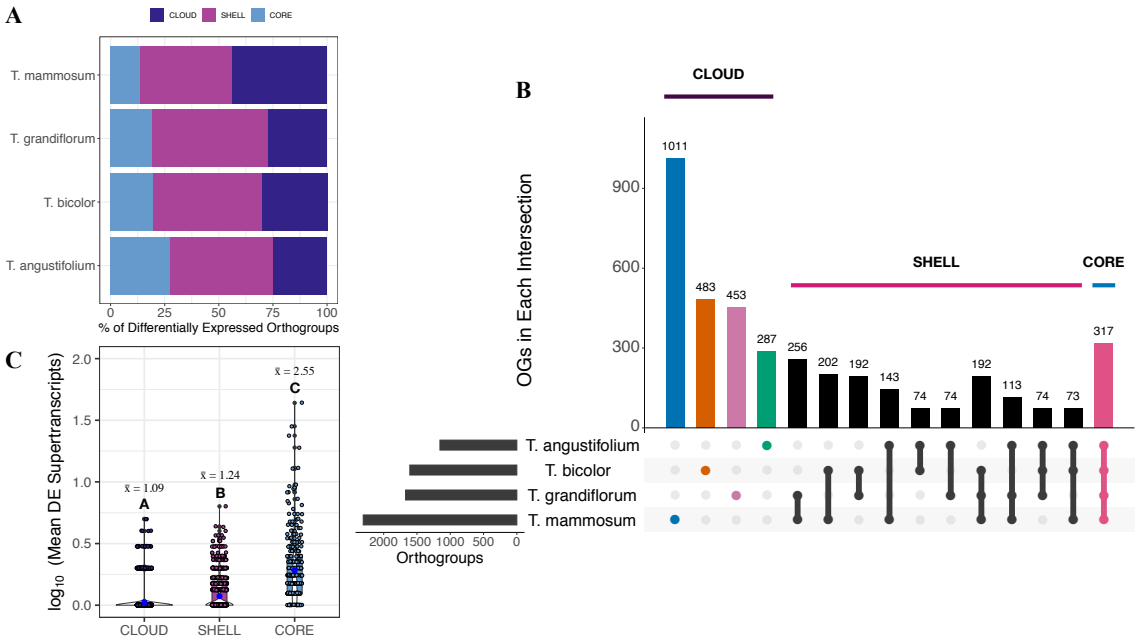


Figure 3-6: Differentially expressed orthogroups and their designated resistance classes. (A) Proportion of differentially expressed orthogroups in each resistance class. Orthogroups that are differentially expressed across all four species are CORE (blue). Those differentially expressed across two or three species are SHELL (pink). And orthogroups differentially expressed in only a single species are CLOUD (purple). Orthgroups containing one or more differentially expressed supertranscript are themselves considered differentially expressed. (B) Overlap of differentially expressed orthogroups. The green, orange, purple, and blue bars represent orthogroups that are only differentially expressed in *T. angustifolium, T. bicolor, T. grandiflorum,* and *T. mammosum,* respectively. The pink bar indicates orthogroups that are differentially expressed across all four species. Numbers above the bars indicate the number of orthogroups in each specific intersection. (C) Average number of differentially expressed supertranscripts for each orthogroup and resistance class. Each point represents the mean number of differentially expressed supertranscripts per species for a given orthogroup. Numbers indicate the mean for each class. Letters indicate significant differences between class means (Welch's ANOVA, p-value< 0.001; Games-Howell post-hoc test, adjusted p-values < 0.05).

Core orthogroups display consistent expression responses across Theobroma

To gain a better understanding of how defense response evolved across *Theobroma*, and to assess how consistent our observed defense responses were across experimental treatments, time scales, and species, we incorporated additional expression evidence from Chapter 2 into our analyses. We began by classifying the *T. cacao* SCA-6 genome into orthogroups using the pipeline described above. We then used the differentially expressed genes, ranked by |LFC|, from Chapter 2 (see Chapter 2 Materials & Methods) to define differentially expressed orthogroups. An orthogroup only needed a single differentially expressed gene from a single population to be considered a differentially expressed orthogroup. In total, we observed 1,133 differentially expressed orthogroups across all four populations examined in Chapter 2. Of those, 733 were also differentially expressed in the non-cacao species (Figure 3-7A). For most orthogroups, mean LFC was weakly, but significantly, correlated between *Theobroma spp.* and *T. cacao* (Figure 3-7B). Several of these core orthogroups, however, had strong responses (|LFC| > 1) across both datasets (Table B.1). Thus, while LFC may not be strongly correlated in a broad sense, some othogroups seemed to be consistently important for *Theobroma*'s defense response.

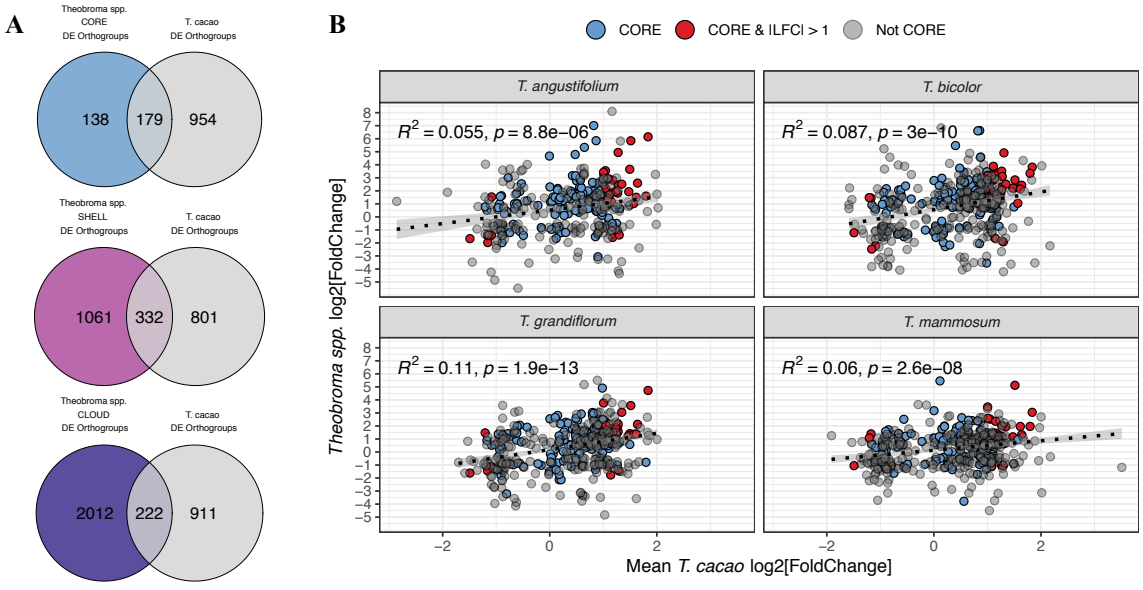


Figure 3-7: **Differentially expressed orthogroups in *T. cacao* and non-cacao *Theobroma spp*.** (A) Overlap between orthogroups that are differentially expressed in *T. cacao* (grey) and orthogroups belonging to each resistance class: CORE (blue), SHELL (pink), and CLOUD (purple). (B) Mean $\log_2$ fold change correlations between orthogroups differentially expressed in both *T. cacao* (Chapter 2) and each non-cacao *Theobroma spp.* Each point represents the mean $\log_2$ fold change for a single orthogroup. Blue points are CORE orthogroups whose mean $|LFC| < 1$ in *T. cacao,* non-cacao *Theobroma spp.*, or both. Red points are CORE orthogroups whose mean $|LFC| > 1$ in both *T. cacao* and non-cacao *Theobroma spp.* Gray points are not in the CORE resistance class.

These orthogroups included a diverse array of gene families with both well-known and potentially novel roles in defense (Table B.1). For instance, we observed both the chitinase and endochitinase gene families, proteins known to be antipathogenic in many species, including cacao (S. Maximova et al. 2003; Y. J. Zhu et al. 2003). Four gene families involved in the biosynthesis and modification of hydroxycinnamic acids were also observed, many of which are anti-microbial (Khan et al. 2021; Widmer and Laurent 2006; Fitzgerald et al. 2004; Knollenberg et al. 2020). Isoeugenol synthases, a family of proteins responsible for biosynthesis of the broad-spectrum antimicrobial phenolic isoeugenol, were upregulated 3 – 32 fold in each species (Table B.1) (Ferreira et al., 2018; Hyldgaard, Mygind, Piotrowska, Foss, & Meyer, 2015).

Perhaps the two most interesting orthogroups, however, were OG60 and OG361, which contain berberine-bridge and WRKY transcription factor proteins, respectively. In Chapter 2, proteins in these two families, *TcBBE8* (SCA-6_Chr6v1_16921) and *TcWRKY29* (SCA-6_Chr3v1_10161), were shown to be both differentially expressed upon pathogen challenge and under selection in resistant varieties. Phylogenies for OG60 and OG361 revealed closely-related orthologs that were responding consistently across species (Figures 3.8 and 3.9). Supertranscripts belonging to the same clade as *TcBBE8* were 2-24 fold upregulated in response to pathogen challenge (Figure 3-8). While other genes from non-cacao *Theobroma* species displayed upregulation following pathogen challenge, few had *T. cacao* orthologs displaying similar expression patterns. Similar to *TcBBE8*, genes in the same clade as *TcWRKY29* were 2-3 fold

upregulated (Figure 3-9). Moreover, we also observed consistent upregulation of two other defense-associated WRKY transcription factors, *TcWRKY22* (SCA-6_Chr1v1_03377), and *TcWRKY69* (SCA-6_Chr6v1_18337). Such consistent responses across different species, time points, experimental designs, and pathogen strains, suggests these two genes are likely key components of cacao's defense response.

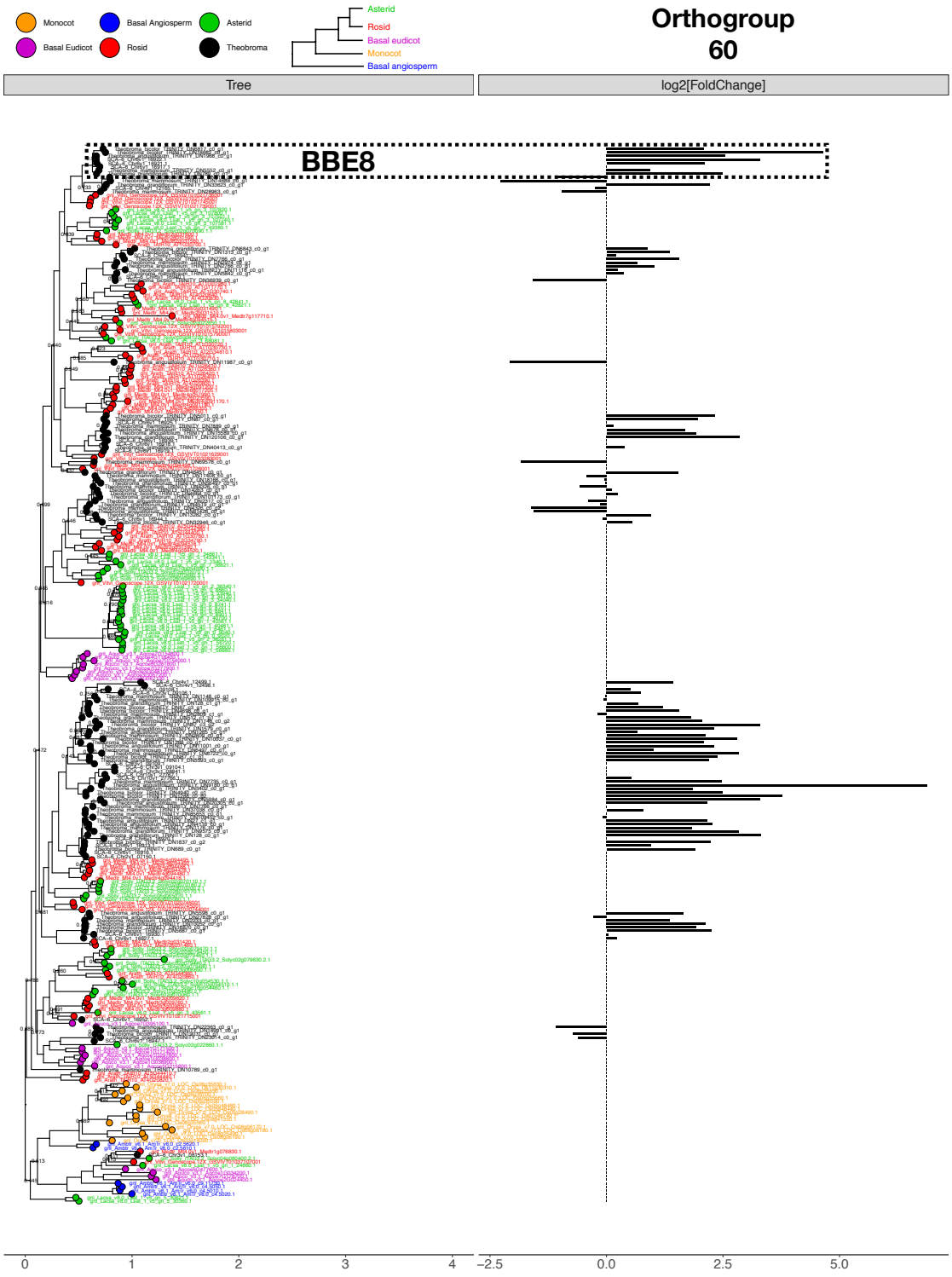


Figure 3-8: Gene family phylogeny for orthogroup 60, FAD-binding berberine bridge enzymes. Sequence IDs are colored according to their lineage: basal angiosperm (blue), basal eudicot (purple), monocot (orange), rosid (red), and asterid (green). All *Theobroma* species, including *T. cacao*, are shown in black. *T. cacao* sequences are from the SCA-6 genome. Node

values indicate SH-like local supports calculated by FastTree. SH supports > 80 are not shown. Bars in the right panel indicate $\log_2$ fold changes. The box indicates the clade containing TcBBE8 (SCA-6_Chr6v1_16921) and its close orthologs across *Theobroma*.

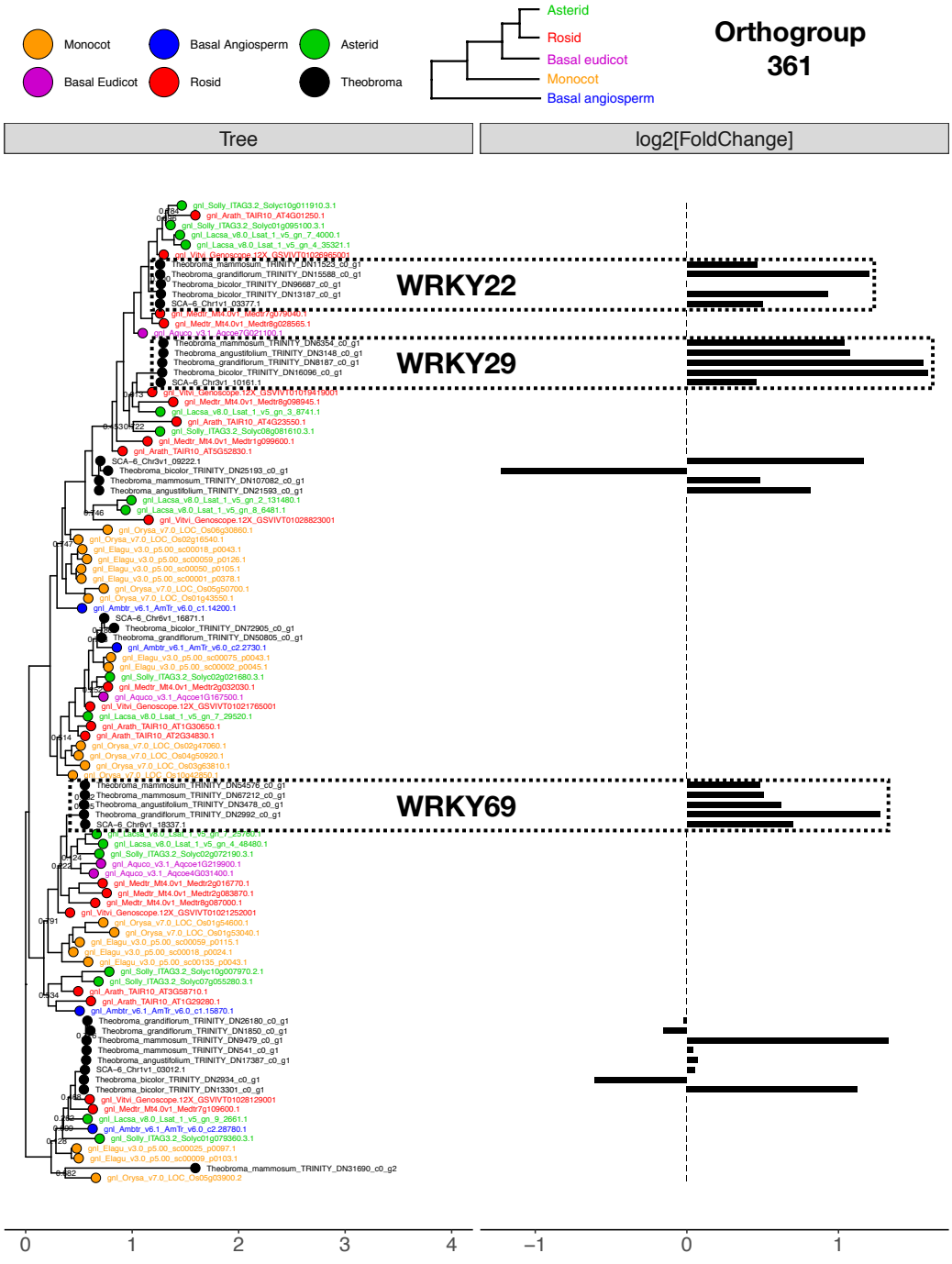


Figure 3-9: Gene family phylogeny for orthogroup 361, WRKY transcription factors. Sequence IDs are colored according to their lineage: basal angiosperm (blue), basal eudicot

(purple), monocot (orange), rosid (red), and asterid (green). All *Theobroma* species, including *T. cacao*, are shown in black. *T. cacao* sequences are from the SCA-6 genome. Node values indicate SH-like local supports calculated by FastTree. SH supports > 80 are not shown. Bars in the right panel indicate $\log_2$ fold changes. Boxes indicate the clades containing TcWRKY29 (SCA-6_Chr3v1_10161), TcWRKY22 (SCA-6_Chr1v1_03377), and TcWRKY69 (SCA-6_Chr6v1_18337), as well as their close orthologs across *Theobroma.*

Conserved orthogroups show evidence of positive selection

To examine how selection shaped the conserved aspects of *Theobroma'*s defense outlined above (Table B.1), we performed branch-site tests using BUSTED to look for evidence of episodic diversifying selection among all *Theobroma* genes in each orthogroup. There was a significant association between the number of orthogroups with evidence of positive selection and orthogroup type (CORE & |LFC| > 1 versus random) (chi-sq goodness-of-fit, p-value < 0.001; Figure 3-10A). Of the 48 orthogroups that were differentially expressed in all five *Theobroma* species and had mean |LFC| > 1, 46 of them displayed significant (FDR-adjust p-value < 0.05) signatures of positive selection. This is higher than the 48 orthogroups we selected at random, of which only 31 showed significant signatures of positive selection. Moreover, bootstrap replication (n = 1000) further supported the idea that the proportion of orthogroups under selection was significantly higher for core orthogroups with |LFC| > 1 than for randomly selected orthogroups (t-test, p-value < 0.05; Figure 3-10B). For both sets, and equal proportion of each gene was under selection (t-test, p-value > 0.05; Figure 3-10C).

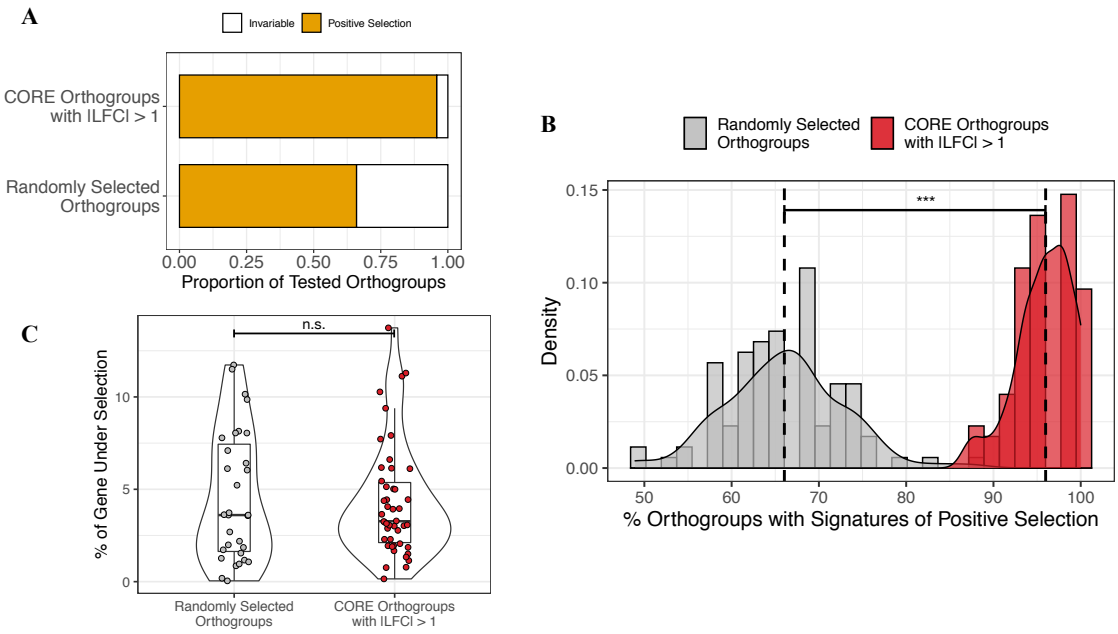


Figure 3-10: **Orthogroups with signatures of positive selection.** (A) Proportion of orthogroups that have signatures of episodic, diversifying selection. CORE orthogroups with |LFC| > 1 in both *T. cacao* and non-cacao *Theobroma spp.* (top) were compared to an equal number (n = 48) of orthogroups drawn at random (bottom). Positive selection was significantly associated with orthogroup type (CORE vs random) (chi-sq. goodness-of-fit, p-value < 0.001). (B) Distribution of bootstrap replicates for both CORE orthogroups with |LFC| > 1 in both *T. cacao* and non-cacao *Theobroma spp.* (red), and orthogroups drawn at random (grey). The proportion of orthogroups displaying signatures of selection was significantly higher for CORE orthogroups with |LFC| > 1 in both *T. cacao* and non-cacao *Theobroma spp.* than for orthogroups drawn at random (t-test, p-value < 0.001). (C) The proportion of each gene under selection. CORE orthogroups with |LFC| > 1 in both *T. cacao* and non-cacao *Theobroma spp.* (red) were compared to an equal number of orthogroups drawn at random (grey). Differences were not significant (t-test, p-value > 0.05).

Discussion

The oomycete pathogen *Phytophthora palmivora* is responsible for extensive annual yield loss in *Theobroma cacao.* In this study, we used non-cacao *Theobroma spp.* to investigate the evolution of defense response across *Theobroma,* with the goal of identifying conserved defense mechanisms that can be incorporated into breeding programs.

Disease resistance assays revealed variation in tolerance/susceptibility to *P. palmivora* across seven species of *Theobroma* (Figure 3-1). Our results are, to our knowledge, one of only two datasets documenting disease resistance phenotypes in these cacao wild relatives. The other was a thesis written by H.M. Rocha in 1966 and found results consistent with our own, namely that *T. bicolor* was significantly more susceptible to *P. palmivora* than T. grandiflorum (Rocha 1966). Our primary interest was to uncover aspects of disease resistance shared across *Theobroma,* to better identify candidate genes important in *T. cacao.* To that end, we chose the two most resistant and two most susceptible non-cacao *Theobroma spp.* for RNA sequencing: *T. angustifolium, T. bicolor, T. grandiflorum,* and *T. mammosum*.

All four species displayed a strong transcriptional response to pathogen challenge, each with thousands of differentially expressed genes (unadjusted p-value < 0.05) (Figure 3-4A). Many of these genes have inferred functions associated with familiar gene ontology terms, some of which were shared across all four species (Figures 3.4B and 3.5). Among these highly shared gene ontology terms, the coumarin (GO:0009805) and lignin biosynthetic (GO:0009809) processes particularly stand out. Both lignin and coumarins are created via the phenylpropanoid pathway, which is responsible for making a diverse array of polyphenolic compounds. Some of these polyphenolics are known to be important for defense against pathogenic microbes. For instance, various hydroxycinnamic acids are active antimicrobial agents (Fitzgerald et al. 2004; Khan et al. 2021; Muroi et al. 2009; Widmer and Laurent 2006), two of which are involved in the *T. cacao – P. palmivora* interaction: clovamide (Knollenberg et al. 2020) and caffeic acid (Chapter 2). Likewise, coumarins are also involved in a diverse array of plant-microbe interactions, including some involving *P. palmivora*. Scopoletin, a particular type of coumarin, has been shown to accumulate in cell cultures of *Hevea brasiliensis* (Malvaceae) during *P. palmivora*-derived elicitor treatment (Dutsadee & Nunta, 2008). And certain *H. brasiliensis* accessions resistant to *P. palmivora* accumulate coumarins faster than susceptible accessions

(Churngchow and Rattarasarn 2001). Coumarin accumulation has also been observed in *Corchorus olitorius* (Malvaceae) when challenged with the fungal pathogen *Helminthosporium turcicum* (Zeid 2002).

Shared gene ontology terms suggest aspects of *Theobroma*'s defense against *P. palmivora* arose in a common ancestor of these four species, and potentially even predates the formation of *Theobroma* as a genus. It is still possible, however, that non-orthologous genes are mediating similar functions across the genus. To explicitly test which aspects of defense response were mediated by orthologous genes, we sorted supertranscripts into orthogroups. We discovered over 300 orthogroups that were differentially expressed across all four species, i.e. core orthogroups (Figure 3-6A-B). Long-term expression conservation suggests these orthogroups are of fundamental importance to disease resistance. Moreover, they also act as important indicators of where to search for genetic variation in cacao. That is, resistance and/or susceptibility to *P. palmivora* in cacao may be caused by genetic variation in one or more of these core orthogroups.

To determine how consistent these results were across experiments, and to make them more translatable to *T. cacao,* we incorporated expression data from Chapter 2 into our analyses (Figure 3-7). We began by classifying *T. cacao* genes into orthogroups. We then used the criteria outlined above to designate orthogroups as differentially expressed. Interestingly, a large number of core orthogroups display strong fold change ($|LFC| > 1$) in response to *P. palmivora* across both non-cacao *Theobroma spp.* and *T. cacao* (Table S3.1). Many of these orthogroups are known to be important for defense, such as the chitinase and endochitinase gene families (S. Maximova et al. 2003; Y. J. Zhu et al. 2003). Two of them, however, stand out as particularly interesting: OG60 and OG361. This is because proteins belonging to these two families, *TcBBE8* (SCA-6_Chr6v1_16921) and *TcWRKY29* (SCA-6_Chr3v1_10161), were repeatedly shown to be interesting candidates in Chapter 2. Both genes were consistently upregulated across all four cacao populations in response to pathogen challenge and displayed signatures of selection among

resistant genotypes. Phylogenies for both OG60 and OG361 revealed a number of closely related orthologs in non-cacao *Theobroma spp.* that are similarly upregulated (Figures 3.8 and 3.9). Supertranscripts in the same clade as *TcBBE8* (OG60) were between 2 and 24 fold upregulated in response to *P. palmivora* (Figure 3-8). Likewise, supertranscripts in the same clade as *TcWRKY29* were between 2 and 3 fold upregulated (Figure 3-9).

These results suggest that *TcBBE8* and *TcWRKY29*, and their corresponding orthologs, could be important components of resistance to *P. palmivora.* Indeed, these results make some sense given each gene's function in *Arabidopsis thaliana. TcBBE8* (AT1G3700) belongs to class of FAD-oxidases called berberine bridge enzymes, some of which are important mediators of resistance to pathogens. For instance, *bbe8* knockouts in *A. thaliana* display reduced stomatal aperture following inoculation with the bacterial pathogen *Pseudomona syringae* (*Pst*) DC3000 (Rodrigues Oblessuc, Vaz Bisneta, and Melotto 2019). Control of stomatal aperture is a key characteristic of innate immune response that helps limit disease progression (Melotto, Underwood, Koczan, Nomura, & He, 2006), but it is manipulated by bacterial phytotoxins during pathogen invasion. Therefore, reduced stomatal aperture upon *Pst* inoculation indicates *A. thaliana BBE8* is involved in coronatine-induced re-opening of stomata (Melotto et al. 2017), which *bbe8* knockouts help mitigate. Another berberine bridge enzyme in *A. thaliana*, *AtBBE22*, oxidizes cellulose oligomers following attack by *Botrytis cinerea*, thereby preventing cellodextrins from becoming a source of carbon for the pathogen (Locci et al. 2019). *TcWRKY29* (AT4G23550) belongs to a class of ancient transcription factors that form an essential component of the plant immune response (Gkizi et al., 2016). *WRKY29* is activated by bacterial flagellin and is often used as a marker for pattern triggered immunity (Fuechtbauer et al., 2018; Göhre, Jones, Sklenář, Robatzek, & Weber, 2012).

As mentioned above, the fact we observe differentially expressed orthogroups across five species of *Theobroma* suggests some aspects of defense response in this genus have been present

for a long time, evolving alongside, and because of, continuous pathogen challenge. This plant-pathogen co-evolution has left fingerprints of positive selection throughout each species' genome. To examine how selection has operated on orthogroups whose response to pathogen is conserved across both *T. cacao* and non-cacao *Theobroma spp.,* we searched for evidence of positive selection using branch-site tests. Of the 48 orthogroups investigated (CORE & |LFC| > 1; Table S3.1), we found 95.8% displayed significant evidence of selection (Figure 3-10A). Perhaps unsurprisingly, OG60 and OG361 were both among these. Thus, two separate methods (population branch statistic and branch-site tests) have indicated selection is operating on genes in these two orthogroups, further highlighting their importance for *Theobroma*'s defense. Moreover, nearly 96% of these conserved orthogroups display evidence of positive selection, a significantly higher proportion than randomly selected orthogroups, suggesting they are not only important presently, but have likely been important aspects of *Theobroma*'s defense for thousands or even millions of years.

Future experimentation on these core orthogroups that display signatures of long-term diversifying selection should be prioritized. Of particular interest are those genes that also display evidence of selection among resistant and susceptible varieties of cacao, such as *TcBBE8* and *TcWRKY29*. Consistent with their fold induction, both genes should be functionally characterized using transient over-expression experiments followed by pathogen bioassays. Furthermore, *in silico* examination of the gene body and regulatory regions for these two genes should be performed across a diverse set of cacao accessions. This would help reveal SNPs and/or SNVs segregating between resistant and susceptible varieties of cacao, further supporting the connection between genotype and phenotype.

Phythophthora palmivora presents one of the greatest threats to cacao production worldwide. Breeding resistant varieties is one approach to mitigate *P. palmivora'*s worst effects, but most contemporary breeding programs have focused on only a handful of clones.

Understanding how *T. cacao'*s wild germplasm can be utilized is therefore an essential step towards breeding clones adapted to both current and emerging threats. Together, our results suggest *Theobroma*'s response *P. palmivora* is a diverse network of both lineage-specific and conserved defenses. Moreover, they provide phenotypic and evolutionary evidence from wild relatives identifying genes and gene families beneficial for *T. cacao*'s future viability as a cultivated crop for producing chocolate.

Acknowledgments

Thank you to Lena Sheaffer for her assistance in project and laboratory management. Thank you to Lara Waldt, Nicholas Moreno, Allan Mata Quirós, and Dr. Mariela Leandro-Muñoz for their help with tissue collection and phenotyping. Thank you to Craig Praul and the Huck Institutes of Life Sciences Genomics Core Facility. This work was supported by The Pennsylvania State University College of Agricultural Sciences, the Huck Institutes of the Life Sciences, the Penn State Endowed Program in Molecular Biology of Cacao, NSF Plant Genome Research Award 1546863 and by the Agriculture and Food Research Initiative (grant number 2018-07789 and accession number 1019277) from the USDA National Institute of Food and Agriculture.

References

Acebo-Guerrero, Yanelis, Annia Hernández-Rodríguez, Mayra Heydrich-Pérez, Mondher El Jaziri, and Ana N. Hernández-Lauzardo. 2012. "Management of Black Pod Rot in Cacao (Theobroma CacaoL.): A Review." *Fruits* 67 (1): 41–48.

Alexa, Adrian, and Jörg Rahnenführer. 2009. “Gene Set Enrichment Analysis with TopGO.” *Bioconductor Improv* 27: 1–26.

Bailey, Bryan A., and Lyndel W. Meinhardt, eds. 2018. *Cacao Diseases*. Cham, Switzerland: Springer International Publishing.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society* 57 (1): 289–300.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics (Oxford, England)* 30 (15): 2114–20.

Boza, Edward J., Juan Carlos Motamayor, Freddy M. Amores, Sergio Cedeño-Amador, Cecile L. Tondo, Donald S. Livingstone, Raymond J. Schnell, and Osman A. Gutiérrez. 2014. “Genetic Characterization of the Cacao Cultivar CCN 51: Its Impact and Significance on Global Cacao Improvement and Production.” *Journal of the American Society for Horticultural Science. American Society for Horticultural Science* 139 (2): 219–29.

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. “Near-Optimal Probabilistic RNA-Seq Quantification.” *Nature Biotechnology* 34 (5): 525–27.

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. “TrimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses.” *Bioinformatics (Oxford, England)* 25 (15): 1972–73.

Churngchow, Nunta, and Matinee Rattarasarn. 2001. “Biosynthesis of Scopoletin in Hevea Brasiliensis Leaves Inoculated with Phytophthora Palmivora.” *Journal of Plant Physiology* 158 (7): 875–82.

Cornejo, Omar E., Muh-Ching Yee, Victor Dominguez, Mary Andrews, Alexandra Sockell, Erika Strandberg, Donald Livingstone 3rd, et al. 2018. “Population Genomic Analyses of the

Chocolate Tree, Theobroma Cacao L., Provide Insights into Its Domestication Process." *Communications Biology* 1 (1): 167.

Cuatrecasas, José. 1964. *Cacao and Its Allies: A Taxonomic Revision of the Genus Theobroma*. Smithsonian Institution.

Dutsadee, Chinnapun, and Churngchow Nunta. 2008. "Induction of Peroxidase, Scopoletin, Phenolic Compounds and Resistance in Hevea Brasiliensis by Elicitin and a Novel Protein Elicitor Purified from Phytophthora Palmivora." *Physiological and Molecular Plant Pathology* 72 (4–6): 179–87.

Eddy, Sean R. 2011. "Accelerated Profile HMM Searches." *PLoS Computational Biology* 7 (10): e1002195.

Evans, Harry C. 2016. "Frosty Pod Rot (Moniliophthora Roreri)." In *Cacao Diseases*, 63–96. Cham: Springer International Publishing.

Ferreira, Sávio Benvindo, Tassiana Barbosa Dantas, Daniele de Figuerêdo Silva, Paula Benvindo Ferreira, Thamara Rodrigues de Melo, and Edeltrudes de Oliveira Lima. 2018. "In Silico and in Vitro Investigation of the Antifungal Activity of Isoeugenol against Penicillium Citrinum." *Current Topics in Medicinal Chemistry* 18 (25): 2186–96.

Fister, Andrew S., Mariela E. Leandro-Muñoz, Dapeng Zhang, James H. Marden, Peter Tiffin, Claude dePamphilis, Siela Maximova, and Mark J. Guiltinan. 2020. "Widely Distributed Variation in Tolerance to Phytophthora Palmivora in Four Genetic Groups of Cacao." *Tree Genetics & Genomes* 16 (1). https://doi.org/10.1007/s11295-019-1396-8.

Fister, Andrew S., Zi Shi, Yufan Zhang, Emily E. Helliwell, Siela N. Maximova, and Mark J. Guiltinan. 2016. "Protocol: Transient Expression System for Functional Genomics in the Tropical Tree Theobroma Cacao L." *Plant Methods* 12 (1): 19.

Fitzgerald, D. J., M. Stratford, M. J. Gasson, J. Ueckert, A. Bos, and A. Narbad. 2004. “Mode of Antimicrobial Action of Vanillin against Escherichia Coli, Lactobacillus Plantarum and Listeria Innocua.” *Journal of Applied Microbiology* 97 (1): 104–13.

Fuechtbauer, Winnie, Temur Yunusov, Zoltán Bozsóki, Aleksandr Gavrin, Euan K. James, Jens Stougaard, Sebastian Schornack, and Simona Radutoiu. 2018. “LYS12 LysM Receptor Decelerates Phytophthora Palmivora Disease Progression in Lotus Japonicus.” *The Plant Journal: For Cell and Molecular Biology* 93 (2): 297–310.

Gkizi, Danai, Silke Lehmann, Floriane L’Haridon, Mario Serrano, Epaminondas J. Paplomatas, Jean-Pierre Métraux, and Sotirios E. Tjamos. 2016. “The Innate Immune Signaling System as a Regulator of Disease Resistance and Induced Systemic Resistance Activity against Verticillium Dahliae.” *Molecular Plant-Microbe Interactions: MPMI* 29 (4): 313–23.

Göhre, Vera, Alexandra M. E. Jones, Jan Sklenář, Silke Robatzek, and Andreas P. M. Weber. 2012. “Molecular Crosstalk between PAMP-Triggered Immunity and Photosynthesis.” *Molecular Plant-Microbe Interactions: MPMI* 25 (8): 1083–92.

Gutiérrez, Osman A., Alina S. Puig, Wilbert Phillips-Mora, Bryan A. Bailey, Shahin S. Ali, Keithanne Mockaitis, Raymond J. Schnell, et al. 2021. “SNP Markers Associated with Resistance to Frosty Pod and Black Pod Rot Diseases in an F1 Population of Theobroma Cacao L.” *Tree Genetics & Genomes* 17 (3). https://doi.org/10.1007/s11295-021-01507-w.

Haas, Brian J., Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D. Blood, Joshua Bowden, Matthew Brian Couger, et al. 2013. “De Novo Transcript Sequence Reconstruction from RNA-Seq Using the Trinity Platform for Reference Generation and Analysis.” *Nature Protocols* 8 (8): 1494–1512.

Hämälä, Tuomas, Mark J. Guiltinan, James H. Marden, Siela N. Maximova, Claude W. dePamphilis, and Peter Tiffin. 2020. “Gene Expression Modularity Reveals Footprints of Polygenic Adaptation in Theobroma Cacao.” *Molecular Biology and Evolution* 37 (1): 110–23.

Hyldgaard, Morten, Tina Mygind, Roxana Piotrowska, Morten Foss, and Rikke L. Meyer. 2015. “Isoeugenol Has a Non-Disruptive Detergent-like Mechanism of Action.” *Frontiers in Microbiology* 6 (July): 754.

Katoh, Kazutaka, Kei-Ichi Kuma, Hiroyuki Toh, and Takashi Miyata. 2005. “MAFFT Version 5: Improvement in Accuracy of Multiple Sequence Alignment.” *Nucleic Acids Research* 33 (2): 511–18.

Khan, Fazlurrahman, Nilushi Indika Bamunuarachchi, Nazia Tabassum, and Young-Mog Kim. 2021. “Caffeic Acid and Its Derivatives: Antimicrobial Drugs toward Microbial Pathogens.” *Journal of Agricultural and Food Chemistry* 69 (10): 2979–3004.

Knollenberg, Benjamin J., Guo-Xing Li, Joshua D. Lambert, Siela N. Maximova, and Mark J. Guiltinan. 2020. “Clovamide, a Hydroxycinnamic Acid Amide, Is a Resistance Factor Against Phytophthora Spp. in Theobroma Cacao.” *Frontiers in Plant Science* 11 (December): 617520.

Law, Charity W., Yunshun Chen, Wei Shi, and Gordon K. Smyth. 2014. “Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts.” *Genome Biology* 15 (2): R29.

Locci, Federica, Manuel Benedetti, Daniela Pontiggia, Matteo Citterico, Claudio Caprari, Benedetta Mattei, Felice Cervone, and Giulia De Lorenzo. 2019. “An Arabidopsis Berberine Bridge Enzyme-like Protein Specifically Oxidizes Cellulose Oligomers and Plays a Role in Immunity.” *The Plant Journal: For Cell and Molecular Biology* 98 (3): 540–54.

Manjang, Kalifa, Shailesh Tripathi, Olli Yli-Harja, Matthias Dehmer, and Frank Emmert-Streib. 2020. “Graph-Based Exploitation of Gene Ontology Using GOxploreR for Scrutinizing Biological Significance.” *Scientific Reports* 10 (1): 16672.

Maximova, S., C. Miller, G. Antúnez de Mayolo, S. Pishak, A. Young, and M. J. Guiltinan. 2003. “Stable Transformation of Theobroma Cacao L. and Influence of Matrix Attachment Regions on GFP Expression.” *Plant Cell Reports* 21 (9): 872–83.

Mchau, Godwin R. A., and Michael D. Coffey. 1994. “Isozyme Diversity in Phytophthora Palmivora: Evidence for a Southeast Asian Centre of Origin.” *Mycological Research* 98 (9): 1035–43.

Melotto, Maeli, William Underwood, Jessica Koczan, Kinya Nomura, and Sheng Yang He. 2006. “Plant Stomata Function in Innate Immunity against Bacterial Invasion.” *Cell* 126 (5): 969–80.

Melotto, Maeli, Li Zhang, Paula R. Oblessuc, and Sheng Yang He. 2017. “Stomatal Defense a Decade Later.” *Plant Physiology* 174 (2): 561–71.

Motamayor, Juan C., Philippe Lachenaud, Jay Wallace da Silva e Mota, Rey Loor, David N. Kuhn, J. Steven Brown, and Raymond J. Schnell. 2008. “Geographic and Genetic Population Differentiation of the Amazonian Chocolate Tree (Theobroma Cacao L).” *PloS One* 3 (10): e3311.

Muroi, Atsushi, Atsushi Ishihara, Chihiro Tanaka, Akihiro Ishizuka, Junji Takabayashi, Hideto Miyoshi, and Takaaki Nishioka. 2009. “Accumulation of Hydroxycinnamic Acid Amides Induced by Pathogen Infection and Identification of Agmatine Coumaroyltransferase in Arabidopsis Thaliana.” *Planta* 230 (3): 517–27.

Pond, Sergei L. Kosakovsky, Simon D. W. Frost, and Spencer V. Muse. 2005. “HyPhy: Hypothesis Testing Using Phylogenies.” *Bioinformatics (Oxford, England)* 21 (5): 676–79.

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2010. “FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments.” *PloS One* 5 (3): e9490.

Richardson, James E., Barbara A. Whitlock, Alan W. Meerow, and Santiago Madriñán. 2015. “The Age of Chocolate: A Diversification History of Theobroma and Malvaceae.” *Frontiers in Ecology and Evolution* 3 (November). https://doi.org/10.3389/fevo.2015.00120.

Rocha, Hermínio M. 1966. “La Importancia de Las Sustancias Polifenólicas En El Mecanismo Fisiológico de La Resistencia de Cacao (Theobroma Cacao L.) a Phytophthora Palmivora (Butl.) Butl.” IICA, Turrialba (Costa Rica).

Rodrigues Oblessuc, Paula, Mariana Vaz Bisneta, and Maeli Melotto. 2019. “Common and Unique Arabidopsis Proteins Involved in Stomatal Susceptibility to Salmonella Enterica and Pseudomonas Syringae.” *FEMS Microbiology Letters* 366 (16). https://doi.org/10.1093/femsle/fnz197.

Silva, Carlos Rogério Sousa, Giorgini Augusto Venturieri, and Antonio Figueira. 2004. “Description of Amazonian Theobroma L. Collections, Species Identification, and Characterization of Interspecific Hybrids.” *Acta Botanica Brasilica* 18 (2): 333–41.

Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. “BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs.” *Bioinformatics (Oxford, England)* 31 (19): 3210–12.

Supek, Fran, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. 2011. “REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms.” *PloS One* 6 (7): e21800.

Wall, P. Kerr, Jim Leebens-Mack, Kai F. Müller, Dawn Field, Naomi S. Altman, and Claude W. dePamphilis. 2008. “PlantTribes: A Gene and Gene Family Resource for Comparative Genomics in Plants.” *Nucleic Acids Research* 36 (Database issue): D970-6.

Wang, Jianan, Michael D. Coffey, Nicola De Maio, and Erica M. Goss. 2020. “Repeated Global Migrations on Different Plant Hosts by the Tropical PathogenPhytophthora Palmivora.” *BioRxiv*. bioRxiv. https://doi.org/10.1101/2020.05.13.093211.

Widmer, Timothy L., and Nathalie Laurent. 2006. “Plant Extracts Containing Caffeic Acid and Rosmarinic Acid Inhibit Zoospore Germination of Phytophthora Spp. Pathogenic to Theobroma Cacao.” *European Journal of Plant Pathology* 115 (4): 377–88.

Wood, Gar, and R. A. Lass. 2001. *Cocoa*. PDF. Edited by G. A. R. Wood and R. A. Lass. 4th ed. Philadelphia, PA: Blackwell Science.

Zeid, Aisha Hussein Saleh Abou. 2002. “Stress Metabolites from Corchorus Olitorius L. Leaves in Response to Certain Stress Agents.” *Food Chemistry* 76 (2): 187–95.

Zhang, Dapeng, Antonio Figueira, Lambert Motilal, Philippe Lachenaud, and Lyndel W. Meinhardt. 2011. “Theobroma.” In *Wild Crop Relatives: Genomic and Breeding Resources*, 277–96. Berlin, Heidelberg: Springer Berlin Heidelberg.

Zhang, Dapeng, and Lambert Motilal. 2016. “Origin, Dispersal, and Current Global Distribution of Cacao Genetic Diversity.” In *Cacao Diseases*, 3–31. Cham: Springer International Publishing.

Zhu, Yun J., Xiaohui Qiu, Paul H. Moore, Wayne Borth, John Hu, Stephen Ferreira, and Henrik H. Albert. 2003. “Systemic Acquired Resistance Induced by BTH in Papaya.” *Physiological and Molecular Plant Pathology* 63 (5): 237–48.

Chapter 4: Local Gene Duplications Drive NLR Copy Number Variation Across Multiple Genotypes of *Theobroma cacao*

Abstract

Nucleotide-binding leucine rich repeats receptors (NLR) are an essential component plant immunity. NLR evolution is complex and dynamic, full of rapid expansions, contractions, and polymorphism. The hundreds of high-quality plant genomes that have been generated over the last two decades have provided substantial insight into the evolutionary dynamics of NLR genes. Despite steadily decreasing sequencing costs, the difficulty of sequencing, assembling, and annotating high-quality genomes has resulted in comparatively little genome-wide information on intraspecies NLR diversity. In this study, we investigated the evolution of NLR genes across 11 high quality genomes of the chocolate tree, *Theobroma cacao.* We found 3-fold variation in NLR copy number across genotypes, a pattern primarily driven by the expansion of NLR clusters by tandem and proximal duplication. Together, our results suggest local duplications can radically reshape gene families over short evolutionary time scales, creating a source of NLR diversity that could be utilized to enrich our understanding of both plant-pathogen interactions and resistance breeding.

Introduction

Plant immunity is principally composed of two layers. The first layer is formed by extracellular receptors that sense and respond to conserved molecular patterns, called pattern recognition receptors (PRR) (Zipfel 2014; Gómez-Gómez and Boller 2000; Chinchilla et al. 2007). In an attempt to subvert detection by extracellular immune receptors, many pathogens have evolved effector proteins that are secreted directly into the plant cell to either dampen plant

immune responsiveness (Zhou et al. 2011) or produce metabolic environments favorable to pathogen growth (L.-Q. Chen 2014). In response to this effector secretion, plants have evolved a second, intracellular layer of pathogen recognition. This intracellular recognition is mediated by immune receptors called nucleotide-binding leucine-rich repeat proteins (NLR) (Flor 1971; Johal and Briggs 1992; Białas et al. 2021). Following pathogen perception, PRRs and NLRs mediate defense responses using partially overlapping mechanisms that result in the production of reactive oxygen species, regulation of phytohormones like salicylic and jasmonic acid, and increased expression of defense-related weapon proteins, along with many other alterations (S. Maximova et al. 2003; Kobayashi et al. 2012; Klessig et al. 2000; Ngou, Jones, and Ding 2021).

NLR genes have received particular attention because of their ability to cause hypersensitive response, a type of qualitative resistance typified by programmed cell death and subsequent cessation of disease progression (Jones and Dangl 2006; J. L. Dangl and Jones 2001). NLR-mediated defense was originally discovered by Henry Harold Flor in the 1940s while breeding flax cultivars resistant to the rust pathogen *Melampsora lini* (Flor 1971). Flor described this phenomenon as the gene-for-gene concept of plant-pathogen interaction, wherein a specific NLR gene recognizes a specific pathogen effector, leading to resistance. Since then, much work has been done to characterize NLR-effector interactions. We now know of at least nine unique molecular mechanisms for NLR-mediated recognition of effector proteins (Kourelis and van der Hoorn 2018).

Most NLR genes contain just three domains. The center of the protein is composed of a nucleotide-binding (NB-ARC) domain that is homologous to human Apaf-1 and CED-4 NB-ARC domains (van der Biezen and Jones 1998). The N-terminal ends of NLRs are variable, containing either a Toll/interleukin-1 domain (TIR) or a coiled-coil motif (CC). And lastly, the C-terminal ends contain a set of leucine-rich repeats (LRR) that vary in length. Many NLRs maintain this canonical structure, but variation in gene architectures is widespread (Van de Weyer et al. 2019a).

Due to their essential role in pathogen recognition and subsequent defense response, NLR genes and pathogen effectors are in a constant arms race. This tight co-evolutionary relationship has resulted in the expansion of NLR and effector repertoires (W. Wang et al. 2021; B. C. Meyers et al. 1998; Haas et al. 2009). For instance, it is not uncommon for NLRs to constitute 1-3% of a species' total gene space (Y. Zhang et al. 2016). At the same time, however, NLR copy number across plant genomes is also highly variable, from as few as 55 in watermelon (Lin et al. 2013) to as many as 2,151 in wheat (Andersen et al. 2020). Likewise, limited evidence suggests NLR copy number can vary 1-2 fold within a species (Van de Weyer et al. 2019a; M.-S. Kim et al. 2021). Examining NLR complement across multiple individuals of the same species helps us understand how populations interact with their environments and provides insight into the ways NLR variation, to the extent that it exists, can be harnessed through breeding.

By virtue of being highly dynamic and repetitive, NLR genes are particularly difficult to assemble, annotate, and analyze. Genome resequencing efforts, moreover, are inherently reliant on reference genomes, limiting their utility for understanding natural variation in gene content, novel domain architectures, and genomic organization. Thus, genome-wide analysis of NLR variation across a species requires high quality *de novo* genome assemblies or the use of enrichment methods to perform high throughput sequencing of target genes (Jupe et al. 2013; Stam, Scheikl, and Tellier 2016). Here, we explored the NLR content of 11 high quality genome assemblies from the chocolate tree, *Theobroma cacao.* We found that NLR copy number was highly variable across genotypes, a phenomenon largely driven by tandem and segmental duplications. Our results provide additional insight into the evolution of NLRs across individuals of the same species and suggests local duplications can drastically alter gene content over short evolutionary time scales.

Materials and Methods

Genome assembly and annotation

We analyzed 11 highly contiguous genomes for this study, including both cacao reference genomes, Criollo B97-61/B2 v2.0 and Matina 1-6 v2.1 (Juan C. Motamayor et al. 2013; X. Argout et al. 2017), as well as nine other genomes that we sequenced, assembled, and annotated: CCN-51, GU-257E, ICS-1, NA-246, IMC-105, NA-807, Pound-7, SCA-6, SPEC 54/1. All assemblies were completed using Illumina 10X linked reads, collected and sequenced as described in Chapter 2 and (Hämälä et al. 2021). Briefly, linked-reads were assembled using Supernova v2.1 (Weisenfeld et al. 2017) at five different raw read coverage depths: 56x, 62x, 68x, 75x, and 85x. The number of reads included in each coverage depth was determined according to estimated genome size. Each of the five coverage depths had two pseudohaplotype assemblies, one of which was chosen for post-processing. From the resulting five pseudohaplotype assemblies, a single representative was chosen as the meta-assembly backbone using a combination of metrics. These metrics included: completeness of benchmarking universal single copy orthologs (BUSCO) (Simão et al. 2015), contig and scaffold L50, and an assembly size that was consistent with the estimated haploid genome size. The remaining pseudohaplotype assemblies were then used to bridge gaps and join contigs, iteratively improving the meta-assembly backbone for each genotype. Assembly errors were corrected with TigMint (Jackman et al. 2018) before being re-scaffolded by ARCS (Yeo et al. 2018). Gaps were filled using GapFiller v1.10 (Boetzer and Pirovano 2012) and the resulting assembly for each genotype was called its meta-assembly. Chloroplast, mitochondria, and non-embryophyte contaminant sequences were removed using the BLAST-based procedure outlined in Chapters 2 and 3. Finally, each meta-

assembly was ordered and oriented onto pseudomolecules (chromosomes) using RaGOO (Alonge et al. 2019) and the *T. cacao* Matina 1-6 v1.1 genome assembly (Juan C. Motamayor et al. 2013).

Before beginning annotation of the meta-assemblies, regions containing a high density of repeats or transposable elements were identified and masked using the MAKER-P repeat masking protocol (Campbell et al. 2014). A diverse set of tissues were sampled to generate transcripts that could be used as genome annotation evidence (Chapter 2 *Materials and Methods*). Transcripts were created using the *de novo* assembly protocol outlined in both Chapters 2 and 3. Meta-assembly annotation took place in two steps. First, annotations from the Criollo B97-61/B2 v2.0 and Matina 1-6 v2.1 reference genomes were transferred to the meta-assemblies using the FLO pipeline (https://github.com/wurmlab/flo). Second, the assembled annotation evidence, as well as evidence from nine other species in Malvaceae, was used to create *de novo* annotations via the MAKER pipeline (Holt and Yandell 2011). These steps resulted in highly contiguous and complete genome assemblies that were suitable for analysis of complex, repeat rich genomic regions.

Genotype phylogeny

We used singly copy orthologs to create a maximum likelihood phylogeny for the 11 genotypes analyzed in this study, as well as four non-cacao *Theobroma spp.* (Simão et al. 2015). First, complete BUSCOs were extracted from each genotype's predicted proteome. Sequence for wild *Theobroma spp.* were extracted from the transcriptome assemblies in Chapter 3. Only complete BUSCOs present in > 3 species were used for phylogenetic analysis. Each set of BUSCOs was then aligned using MAFFT (L-INS-i) (Katoh et al. 2005) and gene tress were constructed using FastTree v2.0 (Price, Dehal, and Arkin 2010). Lastly, a species tree was created

from all 1,364 gene trees using the coalescent-based species tree estimation program ASTRAL (Mirarab et al. 2014).

We collected phenotype information from the International Cacao Germplasm Database (http://www.icgd.rdg.ac.uk/) for three problematic cacao diseases: Ceratocystis wilt of cacao (CWC), frosty pod rot (FPR), and witches' broom disease (WBD). These diseases were chosen because they had the largest collection of phenotype information in the database. Data on black pod rot phenotypes (*P. palmivora*) were taken from a previously published study (Fister et al. 2020). We filtered phenotype information using several criteria. First, we removed any information that was based on disease incidence in the field, since this is highly dependent on environmental conditions. Second, we removed sources of information that did not contain discernable phenotypes, e.g. "intermediate". Lastly, since most clones have multiple phenotype estimates, we classified phenotype classes as numeric values and took the average, i.e. susceptible = 1, moderately susceptible = 2, moderately resistant = 3, and tolerant = resistant = 4. Phenotype averages that were ≤ 2 were considered susceptible and ≥ 3 were considered resistant.

NLR classification and categorization

Nucleotide-binding leucine-rich repeat receptors (NLR) were detected using a combination of custom, domain-specific hidden Markov models (HMM) and the NLR-identification tool NLR-parser (Steuernagel et al. 2015). We created a set of cacao-specific HMMs that were designed to detect homology with three canonical NLR domains: nucleotide-binding domain (NB-ARC), resistance to powdery mildew domain (RPW8), and the Toll/interleukin-1 receptor domain (TIR). In plants, each of these domains are diagnostic for this gene family (Van de Weyer et al. 2019a). HMMs were created by first identifying domains in each genome's predicted proteome using Interproscan v5.32-71.0 (-appl Pfam) (Quevillon et al.

2005). Proteins containing high confidence hits to NB-ARC (PF00931), TIR (PF01582 and PF13676), or RPW8 (PF05659) Pfam domains (e-values ≤ 1e-60 for NB-ARC or 1e-40 for TIR and RPW8) were used for further HMM construction (Mistry et al. 2021). This identified a large number of proteins containing NB-ARC and/or TIR domains. Domain sequences were then extracted from their respective proteins and aligned using MAFFT (--auto) (Katoh et al. 2005). After alignment, phylogenies were constructed from each domain alignment using RAxML (-m PROTCATJTT -p 1234). In order to limit the bias highly similar, and therefore redundant, NB-ARC and TIR domain sequences could introduce during HMM classification, we used a clustering approach to identify unique sequences. This approach used usearch v11 (-cluster_tree -id 0.98) (Edgar 2010) to first cluster domain sequences that were ≥ 98% identical. A single representative was then selected from each cluster. This reduced the number of NB-ARC sequences by > 93% (152/2137) and the number of TIR sequences by nearly 50% (87/189). There were not enough RPW8 domains for redundancy to be an issue in HMM construction and classification, so no clustering and filtering was performed. Finally, representative NB-ARC, TIR, and RPW8 sequences were once again aligned using MAFFT (L-INS-i) and HMMs were built using HMMER v3.3 (hmmbuild) (Eddy 2011). This resulted in one HMM classifier for each of the three canonical domains.

The three HMM classifiers were then used to detect the presence of NB-ARC, TIR, or RPW8 domains from the predicted proteomes of all 11 genotypes. Proteins containing at least one high confidence (e-value ≤ 1e-4) hit were classified as NLRs and carried forward for further analysis. Because the HMM classifiers were constructed using domain sequences, they were not able to identify proteins containing coiled-coil (CC) motifs. To address this problem, we used the NLR identification tool NLR-parser. NLR-parser uses the MEME v4.9.1 suite (T. L. Bailey et al. 2009) to detect small stretches of sequence that are highly similar to a pre-defined set of NLR-specific motifs (Jupe et al. 2012), including CC motifs. Therefore, proteins identified as having

CC motifs by NLR-parser were incorporated into our list of putative NLRs. Domain architectures for all putative NLRs were then assessed using Interproscan v5.32-71.0 (-appl Pfam, COILS) and used to categorize NLRs into TNL, RNL, CNL, or NL using previously outlined criteria (Van de Weyer et al. 2019a). Proteins containing a TIR domain were categorized as TNL and those containing an RPW8 were categorized as RNL. Proteins containing an Rx N-terminal domain (PF18052), or an NLR-parser CC annotation *and* a CC annotation from COILS were categorized as CNL. And lastly, proteins containing an NB-ARC domain *and* a leucine-rich repeat (LRR) domain, but no other domains, were categorized as NL (Figure 1A).

Gene duplication analysis

Gene duplication histories were characterized using MCScanX's duplicate gene classifier (Y. Wang et al. 2012). First, all-by-all BLASTp searches were performed using each genotype's proteome with an e-value cutoff of 1e-10. These BLAST hits were then used as input to categorize duplication types. To do so, MCScanX first ordered genes according to their chromosomal location and categorized them as singletons. BLAST hits were then used to identify genes containing hits elsewhere in the genome. Any gene containing a hit elsewhere was called a dispersed duplicate. Dispersed duplicates were then further categorized as proximal duplicates if they were no more than 20 genes away from a BLAST hit, and tandem duplicates if they were one gene away from a BLAST hit. Lastly, genes classified as anchors by MCScsanX were categorized as segmental or whole genome duplicates (WGD).

NLR clustering was performed using a custom set of R and Bash scripts. First, we calculated the number of NLR genes in non-overlapping genomic windows of 1 Mbp. More than 50% of an NLR needed to overlap a window for it to be counted. Adjacent windows that both contained at least one NLR were then merged. This process was repeated until there were no

remaining windows that could be merged. Sets of merged windows were considered NLR clusters if they contained ≥ 3 NLR genes.

Genome synteny analysis

Pairwise comparisons of genome collinearity were performed using MCScanX (match_score = 50, gap_penalty = -1, match_size = 5, max_gaps = 20, repCut = 300, repDiv = 30). First, putative orthologs were identified with all-vs-all BLASTp (e-value 1e-10) and used to define collinear blocks according to the MCScanX algorithm. To do so, MCScanX first sorted BLAST hits according to their chromosomal positions. Long chains of collinear genes were then identified and collinear blocks longer than five genes were reported. Lastly, adjacent, collinear gene pairs were then used as anchors to align collinear blocks, identifying syntenic regions between genomes.

Pseudogene identification

Pseudogenized NLR genes were identified according to the MAKER-P pseudogene identification pipeline (Campbell et al. 2014; C. Zou et al. 2009). First, we searched for genomic regions containing high sequence similarity to NLR genes using tBLASTn (e-value ≤ 1e-20). These regions of sequence similarity were then used as input into the pseudogene pipeline. Non-genic regions were sorted to remove hits ≤ 30 amino acids in length and ≤ 40% identity. These filtered regions were considered putative pseudoexons. Putative pseudoexons that significantly matched (e-value < 1e-5) repetitive sequences, as defined in RepBase v.12 (Jurka et al. 2005), were removed. The remaining pseudoexons were linked together to form contigs based on two criteria: (1) the best BLAST hit for both pseudoexons was the same parent NLR, and (2) the

sequence space between the matching pseudoexons was inside the 99th percentile of the intron length distribution. These contigs represented putative pseudogenes. NLR integrated domains, i.e. non-canonical NLR domains that are fused to NLR genes, presented a challenge because they would result in the identification on non-NLR pseudoexons and subsequent pseudogenes. Therefore, putative pseudogenes were translated in all six frames and their domains were identified using Interproscan v5.32-71.0 (-appl Pfam). Pseudogenes that did not contain any common NLR domains were removed. We considered the following NLR domains as common: LRR (PF00560, PF07725, PF12799, PF13855), NB-ARC (PF00931), TIR (PF01582, PF13676), RPW8 (PF05659), and CC (PF18052). This filtered set of pseudogenes represented putatively non-functional NLR genes.

Transposable element analysis

Transposable elements (TE) from each genome were annotated according to the procedure outlined in Chapter 2 and followed the MAKER-P repeat masking protocol. First, miniature inverted-repeat transposable elements (MITE) were identified using MITE-Hunter (Han and Wessler 2010). Likewise, long terminal repeat retrotransposons (LTR) were identified using LTRharvest/LTRdigest (Ellinghaus, Kurtz, and Willhoeft 2008; Steinbiss et al. 2009). *De novo* repetitive sequences were predicted using RepeatModeler1 (http://www.repeatmasker.org/RepeatModeler). Predicted TEs were then searched against a SwissProt and RefSeq protein database. Any TEs containing significant hits to the database were excluded from further analysis. We chose to analyze five TE classes: DNA transposons, LINE, SINE, and LTR retrotransposons, and rolling circle Helitrons.

Statistical analyses

All statistical analyses were performed using R v3.6 (Team R.C. 2013). Negative binomial regressions were performed using the MASS v7.3-53.1 package (Venables and Ripley 2002) and pairwise significance was calculated using the emmeans v1.5.4 package (Lenth et al. 2018). Model assumptions were checked using performance v0.8.0 (Lüdecke et al. 2021). All confidence intervals were calculated using Hmisc v4.4-2 (Harrell and Dupont 2006). Plots were created using ggplot2 v3.3.5 (Wickham 2009) and genoPlotR v0.8.11 (Guy, Kultima, and Andersson 2010).

Results

Cacao genotypes displayed a high degree of copy number variation in NLR genes

We used 11 genotypes for this study, each of which belongs to one of seven previously described (Juan C. Motamayor et al. 2008) genetic groups: Criollo B97-61/B2 (Criollo), Matina 1-6 v2.1 (Amelonado), CCN-51 (Hybrid), GU-257E (Guiana), ICS-1 (Hybrid), NA-246 (Marañón), IMC-105 (Iquitos), NA-807 (Nanay), Pound-7 (Nanay), SCA-6 (Contamana), and SPEC 54/1 (Iquitos). The exceptions to this were CCN-51 and ICS-1, which are both hybrids. For each genome, we identified NLR genes using a combination of custom HMM classifiers and previously developed tools (Steuernagel et al. 2015). In total, we identified 2,563 NLR genes across 11 genomes. We further categorized these NLRs into groups based on their domain architecture (Figure 4-1A). Typical NLR genes are tripartite, possessing a variable N-terminal domain, a conserved NB-ARC domain, and a C-terminal end containing a set of leucine-rich repeats of varying length. Based on the presence and/or absence of these domains and their organization, we categorized NLR genes into four classes: NL, CNL, TNL, and RNL (Figure 4-

1A). RNL copy number ($Mean_{RNL}$ = 3.09, SEM_{RNL} = 0.25) appeared to be conserved across genotypes, consistent with their role as an ancient clade of helper NLRs (Figure 4-1B) (Lapin et al. 2019; Jubic et al. 2019). Copy number in NL ($Mean_{NL}$ = 118.27, SEM_{NL} = 21.05), CNL ($Mean_{CNL}$ = 89.64, SEM_{CNL} = 14.28), and TNL classes ($Mean_{TNL}$ = 21.09, SEM_{TNL} = 2.86), however, was highly variable, with particular divergence seen in NL and CNL. Total NLR number varied across genotypes and seemed to fall into two discrete classes: genotypes with high NLR copy number ($Mean_{HighCNV}$ = 314.86, $SEM_{HighCNV}$ = 13.72) and genotypes with low copy number ($Mean_{LowCNV}$ = 89.75, SEM_{LowCNV} = 2.98), hereafter referred to as High CNV and Low CNV, respectively (Figure 4-1B). There was 3-fold difference in NLR copy number between these two groups (mean difference = 225.11, Mann-Whitney test: p-value < 0.01), most of which was driven by expansion and/or contraction of the NL and CNL classes (negative binomial GLM: NLR # ~ CNV Group + NLR Class + CNV Group * NLR Class, adjusted p-values < 0.01).

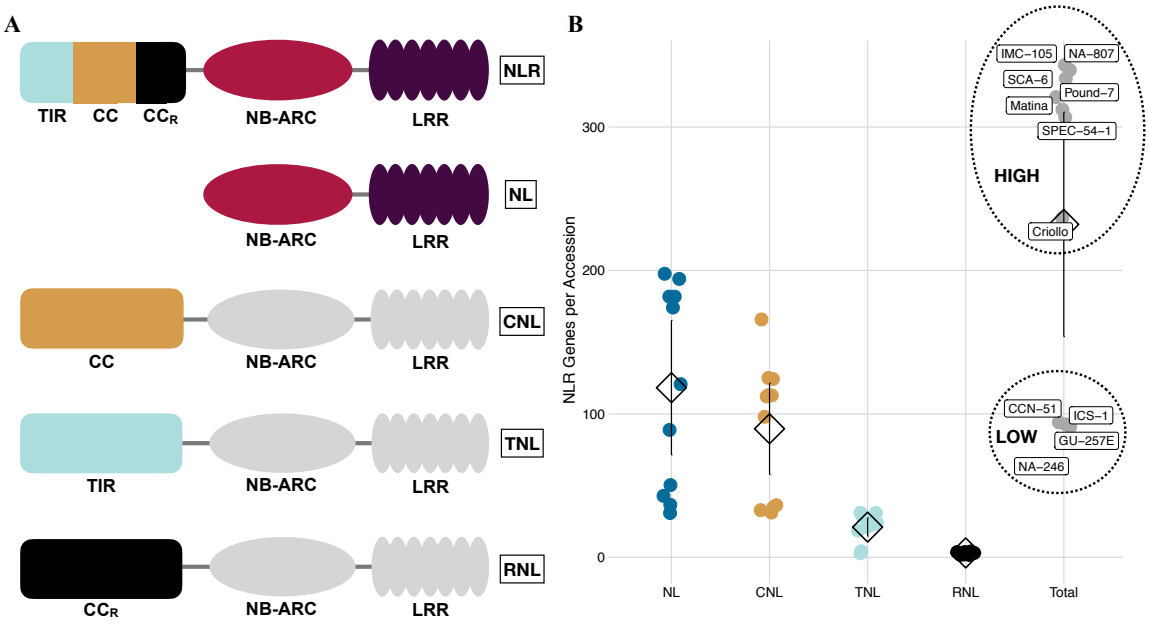


Figure 4-1: NLR architecture and copy number across cacao genomes. (A) The four canonical NLR architectures. Proteins containing a TIR domain (PF01582 and PF13676) were categorized as TNL. Those containing an RPW8 domain (PF05659) were categorized as RNL. Proteins containing an Rx N-terminal domain (PF18052), or an NLR-parser CC annotation and a CC annotation from COILS were categorized as CNL. Proteins containing an NB-ARC domain and a leucine-rich repeat (LRR) domain, but no other domains, were categorized as NL. (B) NLR copy number across all classes and genotypes. NL, CNL, TNL, and RNLs are shown as blue, yellow, teal, and black, respectively. Each point represents the number of NLR copies for a particular genotype. Means are represented by diamonds. Lines represent 95% confidence intervals. High CNV genotypes had significantly more NLR genes than low CNV genotypes (mean difference = 225.11, Mann-Whitney test: p-value < 0.01). Other than the RNL class, differences in mean NLR number between Low CNV and High CNV genotypes were significant for all classes (negative binomial GLM: NLR # ~ CNV Group + NLR Class + CNV Group * NLR Class, adjusted p-values < 0.01).

NLR copy number variation was not distributed evenly throughout the genome

NLR genes were distributed across all 10 cacao chromosomes, but most (45-84%) were localized to just four: Chr5, Chr6, Chr7, and Chr10 (Figure 4-2A). This is true for both Low CNV and High CNV genotypes. Consistent with a birth-and-death model, these four chromosomes also

contained the greatest concentration of pseudogenes in the High CNV genotypes (Figure 4-2B). Low CNV genotypes, however, were more variable. Three of the four Low CNV genotypes, CCN-51, GU-257E, and NA-246, had zero NLR genes, while ICS-1 had 2x the number of NLR pseudogenes as it did NLRs (Figure 4-2B). ICS-1 pseudogenes were distributed across chromosomes 5, 6, 7, and 10, similar to the High CNV genotypes, but patterns of gene duplication varied. While the high CNV genotypes had approximately 1.6 pseudogenes for every 1 parent NLR (Mean = 1.63, SEM = 0.09), ICS-1 had 4.6. That is, ICS-1 had a large number of pseudogenes (n = 198) coming from a narrow number of parents (n = 43). These results indicate NLR genes, at least in the High CNV genotypes, expanded on a small number of chromosomes, helping drive the observed differences in copy number across genotypes. This NLR expansion was then followed by pseudogenization according to a birth-and-death model. This process, however, likely occurred differently in ICS-1, for reasons we elaborate in the discussion.

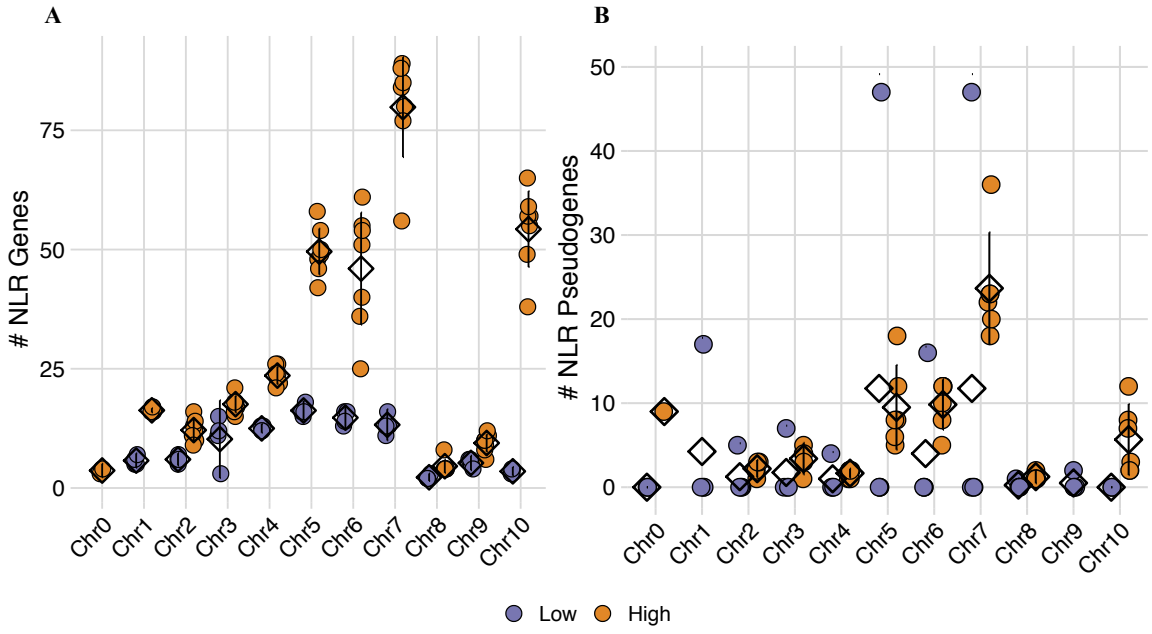


Figure 4-2: Distribution of NLR genes across each genome. (A-B) Number of NLR genes or NLR pseudogenes on each chromosome. Orange depicts High CNV genotypes and purple depicts Low CNV genotypes. Each point represents the number of NLR genes or NLR pseudogenes for a particular genotype. NLR genes or NLR pseudogenes on Chr0 do not belong to one of the 10 chromosome-oriented scaffolds. Means are represented by diamonds. Lines represent 95% confidence intervals.

High and low copy number genotypes evolved independently multiple times

To test whether low versus high copy number was segregating across the cacao phylogeny, we created a species tree from 1,364 single copy ortholog trees (Figure 4-3). Population-level relationships were recovered for Nanay but not for Iquitos. This is consistent with other phylogenetic trees (Chapter 2), and likely occurred because SPEC 54/1 is highly differentiated from other Iquitos genotypes. While low versus high copy number was consistent within each clade, e.g. both NA-807 and Pound-7 belong to High CNV, there was not consistency within populations, e.g. SPEC 54/1 versus IMC-105. Likewise, more broadly defined clades are also inconsistent with respect to their CNV group, e.g. GU-257E, NA-246, and SCA-6. These results indicate High CNV, as a trait, independently evolved multiple times. It is likely, however, that, given a larger sample size, copy number variation among genotypes would become less discrete, forming a more continuous distribution that fills in the gap between High and Low CNV genotypes. Thus, it may be necessary to assess NLR copy number variation across the phylogeny as quantitative rather than qualitative. Lastly, CNV group was not associated with resistance to Ceratocystis, black pod rot, frosty pod rot, or witch's broom disease (Figure 4-3).

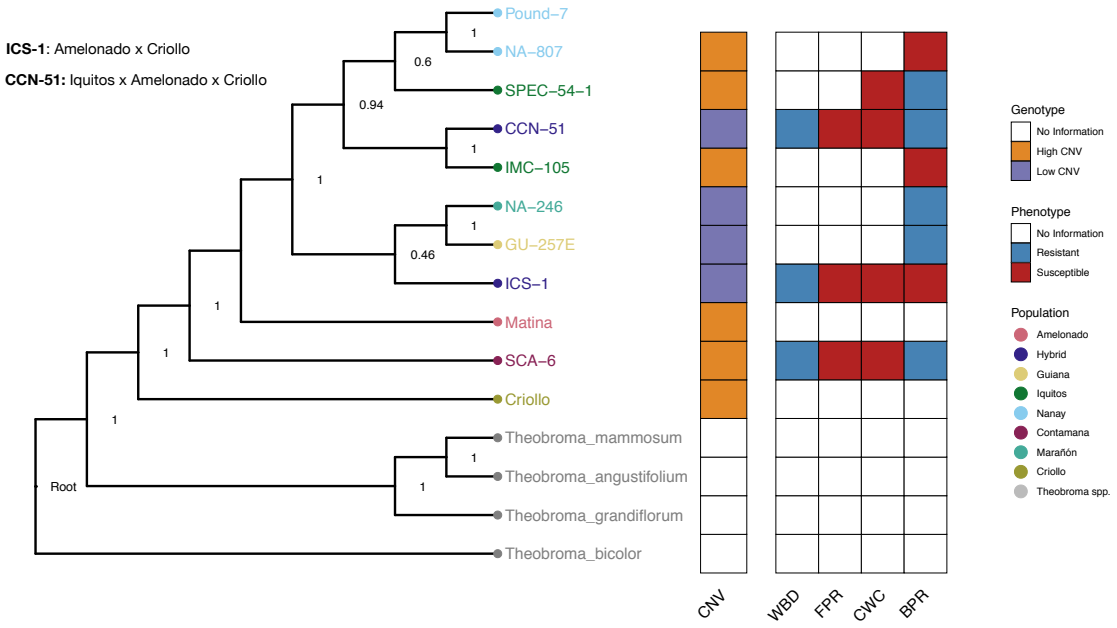


Figure 4-3: Phylogeny of cacao genotypes sampled for this study. Phylogenetic tree of the 11 cacao genotypes used in this study, constructed using 1,364 single copy genes. Four non-cacao species of *Theobroma* were additionally used as outgroups. Numbers on each node represent posterior support values calculated by ASTRAL. With the exception of CCN-51 and ICS-1, both of which are hybrids, tip colors indicate population membership. Non-cacao *Theobroma spp.* are shown in grey. CNV class (High, Low, or No Information) of each genotype is shown in orange, purple, of white, respectively. Disease phenotypes are shown for witches' broom disease (WBD), frosty pod rot (FPR), Ceratocystis wilt of cacao (CWC) and black pod rot (BPR). Blue indicates resistant, red indicates susceptible, and white indicates no information was available.

Annotation quality varied across genotypes but did not explain NLR copy number variation

To confirm differences in annotation quality across genomes did not drive variation in NLR copy number, we analyzed three separate metrics of annotation quality for each genome (Figure 4-4). We began by comparing BUSCO completeness across all 11 genotypes. The Low CNV group had a significantly lower proportion of both complete and fragmented BUSCOs relative to the High CNV group (one-way ANOVA: Proportion ~ CNV Group + BUSCO Class + CNV Group * BUSCO Class, p-value < 0.001), indicative of less complete genome annotations (Figure 4-4A). The mean difference for complete and fragmented BUSCOs was 7.38% and

4.75%, respectively. The average proportion of complete BUSCOs for both CNV groups, however, was ≥ 90%, which is generally considered highly complete (Simão et al. 2015). Next, we used MAKER's annotation edit distance (AED) to assess the quality of annotated NLR genes. AED is a measure of how well annotations match aligned transcripts and protein data and has scores ranging from 0 to 1. An AED of 0 indicates a perfect match between an annotation and its evidence, while an AED of 1 indicates complete discordance. Annotations with AED scores < 0.2 are considered extremely high quality (Holt and Yandell 2011). Mean AED was significantly different between Low CNV genotypes and High CNV genotypes (t-test: p-value < 0.001; Figure 4-4B). However, both CNV groups had AED distributions with means centered near 0.2 (0.19 for Low CNV and 0.20 for High CNV, mean difference = 0.018). This indicates NLR annotations for both Low CNV and High CNV genotypes are not likely to be spurious. Lastly, we also assessed annotation quality by investigating differences in total annotated gene space among Low CNV and High CNV genotypes (Figure 4-4C). While the Low CNV genotypes possessed fewer annotated genes ($Mean_{Gene\ \#}$ = 25,064.25 genes) than High CNV genotypes ($Mean_{Gene\ \#}$ = 26,145.14 genes), the difference was not significant (mean difference = 1080.89 genes, Mann-Whitney test: p-value > 0.05). Moreover, this approximately 4% variation in total gene content between High and Low CNV genotypes is much lower than the approximately 300% variation in NLR copy number. Thus, the only way differences in annotated gene content could have driven the observed patterns of NLR copy number variation is if annotation was systematically biased against NLRs. Together, these results suggest differences in NLR copy number are not the result of technical differences in annotation quality but are due to authentic differences in duplication histories across genotypes.

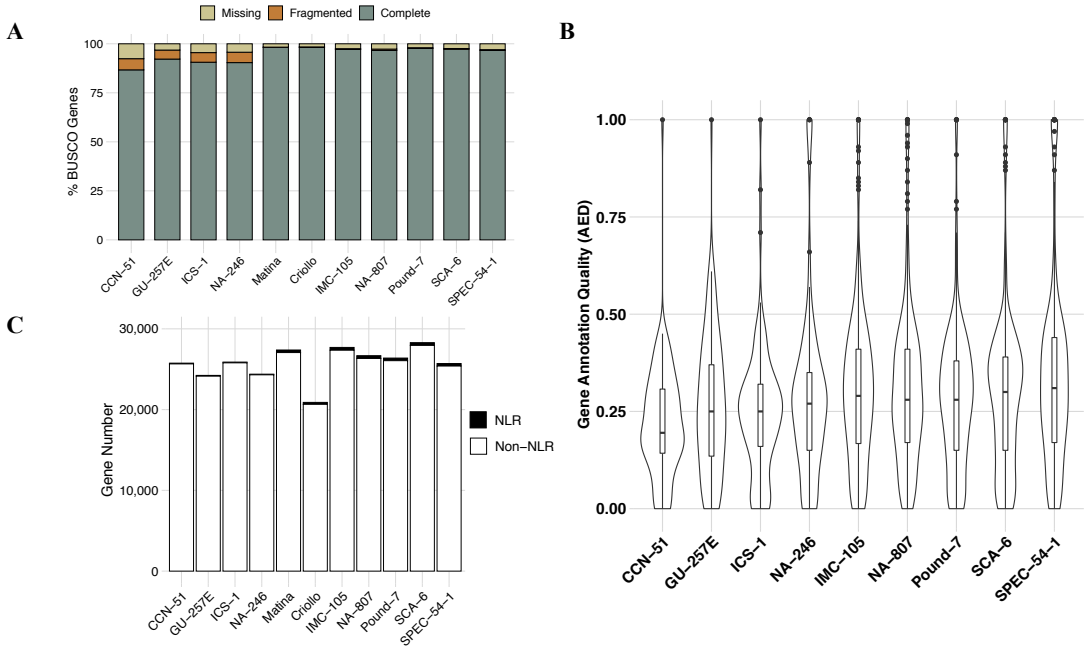


Figure 4-4: Genome annotation quality metrics. (A) BUSCO completeness for each genome used in this study, separated by Low CNV (left) and High CNV (right). The proportion of complete, fragmented, and missing BUSCOs are shown in green, orange, and beige, respectively. Differences in the mean proportion of complete, fragmented, and missing genes between Low CNV and High CNV genotypes were significant (one-way ANOVA: Proportion ~ CNV Group + BUSCO Class + CNV Group * BUSCO Class, p-value < 0.001; Tukey's HSD, adjusted p-value < 0.01). (B) Distribution of AED scores for each genotype's classified NLR genes. Mean AED score was not significantly different between Low CNV and High CNV groups (mean difference = 0.018, t-test: p-value < 0.001). (C) The total number of genes annotated in each of the 11 genomes used in this study, separated by Low CNV (left) and High CNV (right). NLR genes are shown in black and non-NLR genes are shown in white. There was no significant difference in gene number between Low CNV and High CNV genotypes (mean difference = 1080.89 genes, Mann-Whitney test: p-value > 0.05).

Variation in transposable element content did not explain NLR copy number variation

Gene duplication is a common feature of plant genomes. Duplications can take place through many mechanisms, such unequal crossing over during meiosis (S. Kim et al. 2017), non-homologous end joining during DNA damage repair (S. Kim et al. 2017), slipped-strand mispairing (Xu et al. 2021), and transposable element-mediated operations (S. Kim et al. 2017). The last of these possibilities was recently shown to play a role in NLR duplication and

diversification in two species in the genus *Capsicum*: *C. baccatum* and *C. chinense* (S. Kim et al. 2017). To investigate whether TEs were responsible for the patterns of NLR copy number variation across genotypes, we annotated both well-known and uncharacterized TEs from 10 of the 11 cacao genotypes in this study. We examined the abundance of five known TE classes: DNA transposons (Vladimir V. Kapitonov and Jurka 2006), SINE, LINE, and LTR retrotransposons (Kramerov and Vassetzky 2011; Xiao et al. 2008), and rolling circle Helitrons (V. V. Kapitonov and Jurka 2001). LTR elements were by far the most abundant, followed by DNA, LINE, SINE, and Helitron elements. There were no significant differences in TE abundance between High CNV and Low CNV genotypes (negative binomial GLM: # TE ~ CNV + TE Class + CNV*TE Class, adjusted p-values > 0.05; Figure 4-5).

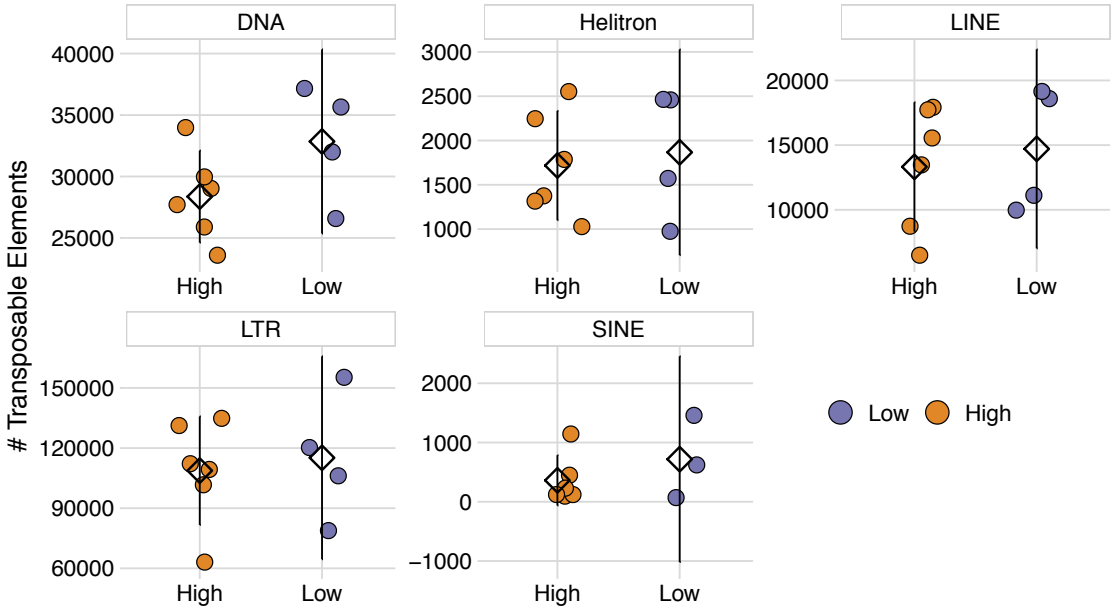


Figure 4-5: Transposable element abundance for High and Low CNV genotypes. Abundance of the five most common transposable elements in cacao genomes. Orange depicts High CNV genotypes and purple depicts Low CNV genotypes. Each point represents the number of transposable elements for a particular genotype. Means are represented by diamonds. Lines represent 95% confidence intervals. Differences in mean TE abundance between Low and High CNV genotypes were not significant (negative binomial GLM: # TE ~ CNV + TE Class + CNV*TE Class, adjusted p-values > 0.05).

Aggregate patterns of TE abundance across genomes, however, may mask differences in local TE density that drive gene duplications. Therefore, we also investigated the abundance of TE classes on each chromosome for both Low and High CNV genotypes. Patterns of TE density across chromosomes did not match those seen for NLRs (Figure 4-6). That is, very few TE classes had significantly higher abundance among NLR-dense chromosomes, i.e. Chr5, Chr6, Chr7, and Chr10 (negative binomial GLM: # TE ~ Chrom + TE Class + Chrom*TE Class). There were significantly more Helitron TEs on Chr5, Chr6, and Chr10 relative to Chr8 (adjusted p-values < 0.05), and significantly more LTR TEs on Chr5 relative to Chr8 (adjusted p-value < 0.01). This, however, is primarily because Chr8 had the lowest TE abundance across all classes, and likely has nothing to do with NLR accumulation. Moreover, none of the chromosomes displayed significant differences in TE content between Low and High CNV genotypes (negative binomial GLM: # TE ~ CNV + Chrom + TE Class + Chrom*TE Class*CNV, adjusted p-values > 0.05). These results suggest variation in NLR content, both within and between genomes, was not explained by variation in TE content and potential TE-mediated duplications.

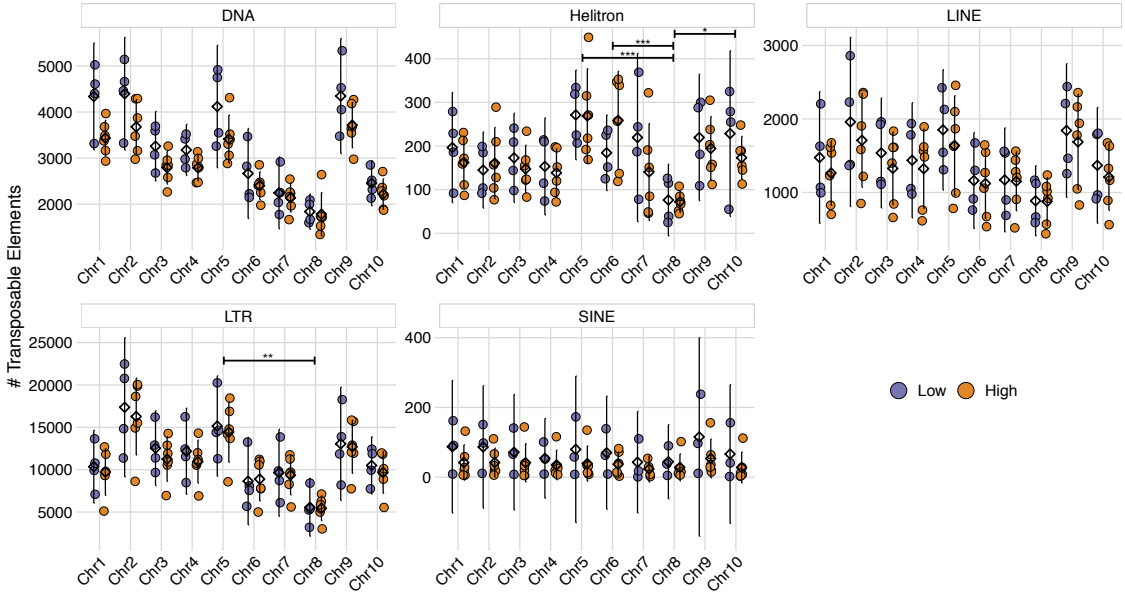


Figure 4-6: Distribution of transposable elements across each genome. Density of the five most common transposable elements across each cacao chromosome. Orange depicts High CNV genotypes and purple depicts Low CNV genotypes. Each point represents the number TEs for a particular genotype. Means are represented by diamonds. Lines represent 95% confidence intervals. Stars indicate significant differences in mean TE abundance between chromosomes (negative binomial GLM: # TE ~ Chrom + TE Class + Chrom*TE Class, adjusted p-values < 0.05). Differences in mean TE abundance between Low and High CNV genotypes on each chromosome were not significant (negative binomial GLM: # TE ~ CNV + Chrom + TE Class + Chrom*TE Class*CNV, adjusted p-values > 0.05).

Tandem and proximal duplications were primarily responsible for NLR copy number variation

Confident that differences in NLR copy number across genotypes was not due to technical discrepancies in annotation quality or mediated by transposable elements, we began investigating gene duplication histories across all 11 genotypes. To do this, we used MCScanX's duplicate gene classifier to categorize NLR genes as singletons, dispersed, proximal, tandem, or WGD/segmental duplicates (Figure 4-7). All NLRs were first classified as singletons, i.e. genes with no history of recent duplication (Figure 4-7A). If NLR genes contained significant BLAST hits elsewhere in the genome, they were reclassified as dispersed duplicates (Figure 4-7B). Dispersed duplicates were then further categorized as proximal or tandem based on distance between hits. If < 20 genes separated the NLR duplicates, they were considered proximal (Figure 4-7C). If NLR duplicates were immediately adjacent to one another, they were considered tandem (Figure 4-7D). Lastly, NLR duplicates that were anchors of collinear blocks, as defined by MCScanX's algorithm (Y. Wang et al. 2012), were classified as WGD/segmental duplicates (Figure 4-7E).

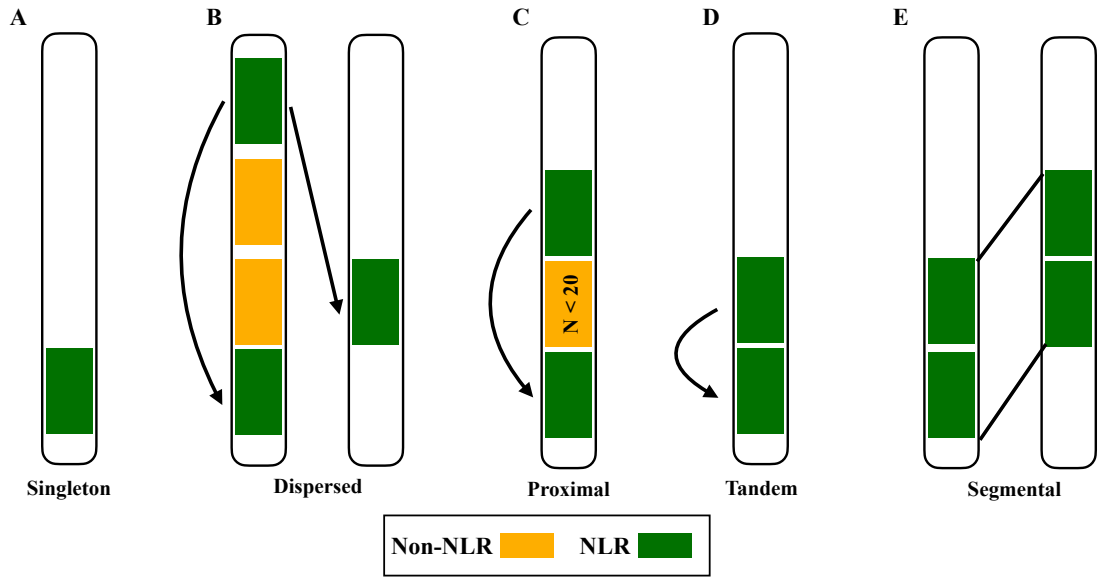


Figure 4-7: Types of gene duplication. Duplicate genes were classified into one of five categories: singletons, dispersed, proximal, tandem, or WGD/segmental duplicates. All NLRs were first classified as singletons, i.e. genes with no history of recent duplication (A). If NLR genes contained significant BLAST hits elsewhere in the genome, they were reclassified as dispersed duplicates (B). Dispersed duplicates were then further categorized as proximal or tandem based on distance between hits. If < 20 genes separated the NLR duplicates, they were considered proximal (C). If NLR duplicates were immediately adjacent to one another, they were considered tandem (D). Lastly, NLR duplicates that were anchors of collinear blocks, as defined by MCScanX's algorithm, were classified as WGD/segmental duplicates (E).

NLR genes had disproportionately higher tandem (mean difference = 23.9%) and proximal (mean difference = 24.3%) duplication rates relative to non-NLR genes (one-way ANOVA: Proportion ~ Gene Type + Duplicate Type + Gene Type * Duplicate Type, p-value < 0.001; Tukey's HSD, adjusted p-values < 0.001; Figure 4-8A). Likewise, NLR genes had a significantly lower proportion of both dispersed (mean difference = 7.1%) and segmental (mean difference = 17.5%) duplications relative to non-NLR genes (one-way ANOVA: Proportion ~ Gene Type + Duplicate Type + Gene Type * Duplicate Type, p-value < 0.001; Tukey's HSD, adjusted p-values < 0.01; Figure 4-8A). These results are consistent with previous findings that NLR evolution is primarily driven by local duplication events (Blake C. Meyers et al. 2003; B. C.

Meyers et al. 1998). There were very few singleton NLRs relative to non-NLR genes. Therefore, we focused the remaining analyses on tandem, proximal, dispersed, and segmental duplicates.

While tandem and proximal duplications drove NLR evolution broadly, it was unclear how differences in NLR copy number across genotypes arose. To test this, we examined the NLR duplication history of both Low CNV and High CNV groups (Figure 4-8B). High CNV genotypes had significantly more NLR duplicates across all types (negative binomial GLM: # NLR Duplicates ~ Duplicate Type + CNV Group + Duplicate Type * CNV Group, adjusted p-values < 0.001). Again, however, the largest difference was seen for tandem and proximal duplicates, which together accounted for 73.7% of the total difference in NLR number between High CNV and Low CNV genotypes (165.9/225.1). Thus, it appears that tandem and proximal duplication events not only drive differences between NLR and non-NLR genes, but also drive variation in NLR content across genotypes.

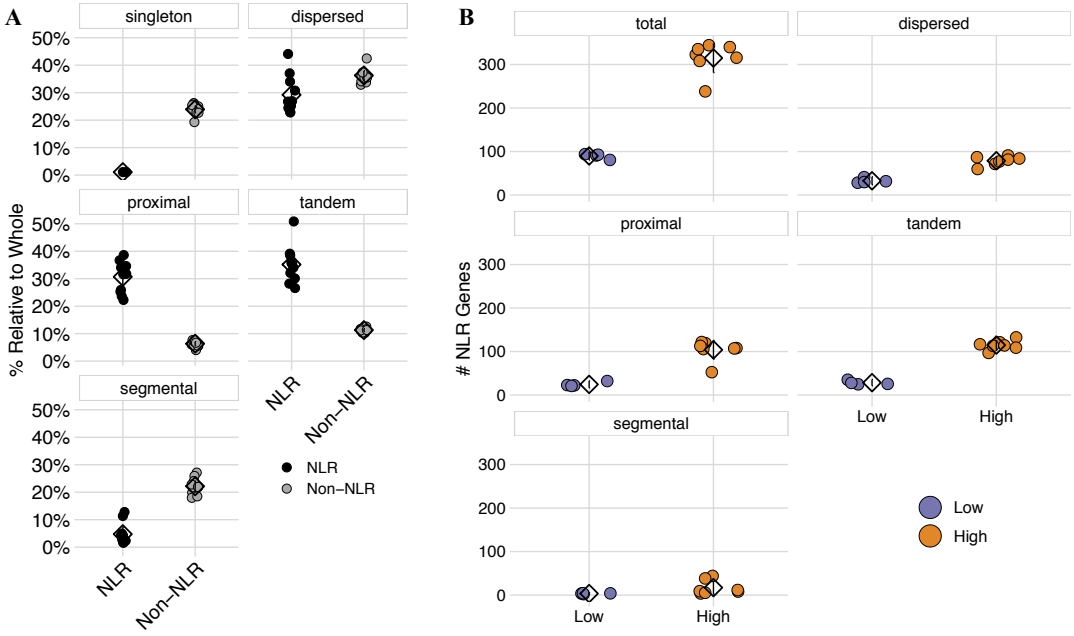


Figure 4-8: Patterns of NLR duplication across each genome. (A) The proportion of NLR (black) and non-NLR (grey) genes in each duplication class. Each point represents the proportion of NLR or non-NLR genes for a particular genotype. Means are represented by diamonds. Lines represent 95% confidence intervals. All differences in mean proportion between NLR and non-

NLR genes were significant (one-way ANOVA: Proportion ~ Gene Type + Duplicate Type + Gene Type * Duplicate Type, p-value < 0.001; Tukey's HSD, adjusted p-values < 0.001). (B) The number of NLR genes belonging to each duplication class, for both Low CNV (purple) and High CNV (orange) genotypes. Points represent the number of NLR genes for a particular genotype. Means are represented by diamonds. Lines represent 95% confidence intervals. All differences in mean NLR number between Low CNV and High CNV groups were significant (negative binomial GLM: # NLR Duplicates ~ Duplicate Type + CNV Group + Duplicate Type * CNV Group, adjusted p-values < 0.001).

Difference in tandem and proximal duplication rates appeared to drive differences in NLR copy number among genotypes, but it was unclear whether duplication type was biased towards specific gene architectures. To test this, we calculated the number of NL, CNL, TNL, and RNL genes belonging to each duplication type. Across all four types of duplication, the proportion of NLRs in each architecture class was constant (Figure 4-9). There was, however, a large difference in NLR copy number within both NL and CNL. Thus, duplications, as a whole, were biased towards NL and CNL architectures, and this bias was consistent across duplication types.

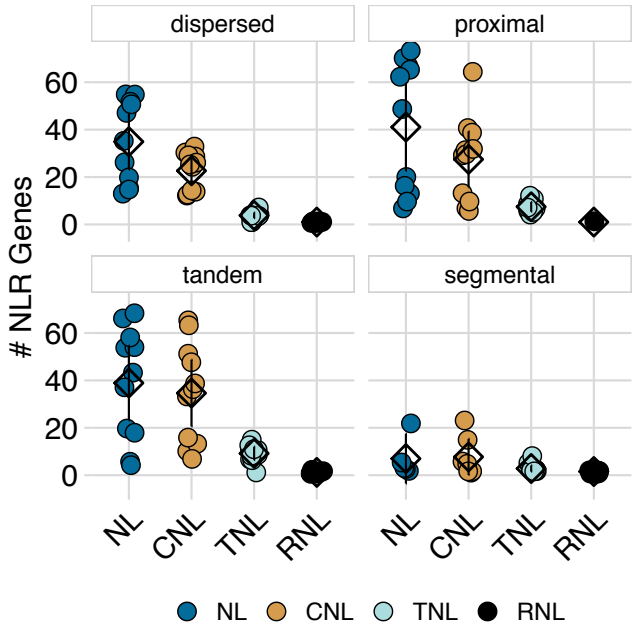


Figure 4-9: NLR duplications across domain architectures. NL, CNL, TNL, and RNLs are shown as blue, yellow, teal, and black, respectively. Each point represents the number of NLR

copies for a particular genotype. Means are represented by diamonds. Lines represent 95% confidence intervals.

Most NLR copy number variation occurred on just four chromosomes: Chr5, Chr6, Chr7, and Chr10. To test whether tandem and proximal duplications were responsible for the formation of these NLR hotspots, we investigated the distribution of NLRs across both chromosomes and duplication types (Figure 4-10). While Chr7 harbors a higher number of dispersed duplicates than other chromosomes, NLR density on Chr5, Chr6, Chr7, and Chr10 was primarily driven by tandem and proximal duplications (Figure 4-10A). Likewise, differences in copy number between Low CNV and High CNV genotypes was associated with tandem and proximal duplications on these chromsomes. Consistent with these results, there was a significant difference in both the number and size of NLR clusters between Low CNV and High CNV genotypes (Figure 4-10B), a difference predominantly caused by proximal and tandem duplications (cluster number: mean difference = 6.32, Mann-Whitney test, p-value < 0.01; cluster size: mean difference = 11.48, Mann-Whitney test, p-value < 0.05). Thus, it appears that most NLR copy number variation is occurring on just four chromosomes, and these differences are principally due to proximal and tandem duplications driving expansion of NLR clusters. The expansion of one such NLR cluster can be seen in Figure 4-11. While the Low CNV genotypes have NLRs in this 2 Mbp region, cluster expansion via tandem and proximal duplications in High CNV genotypes is striking.

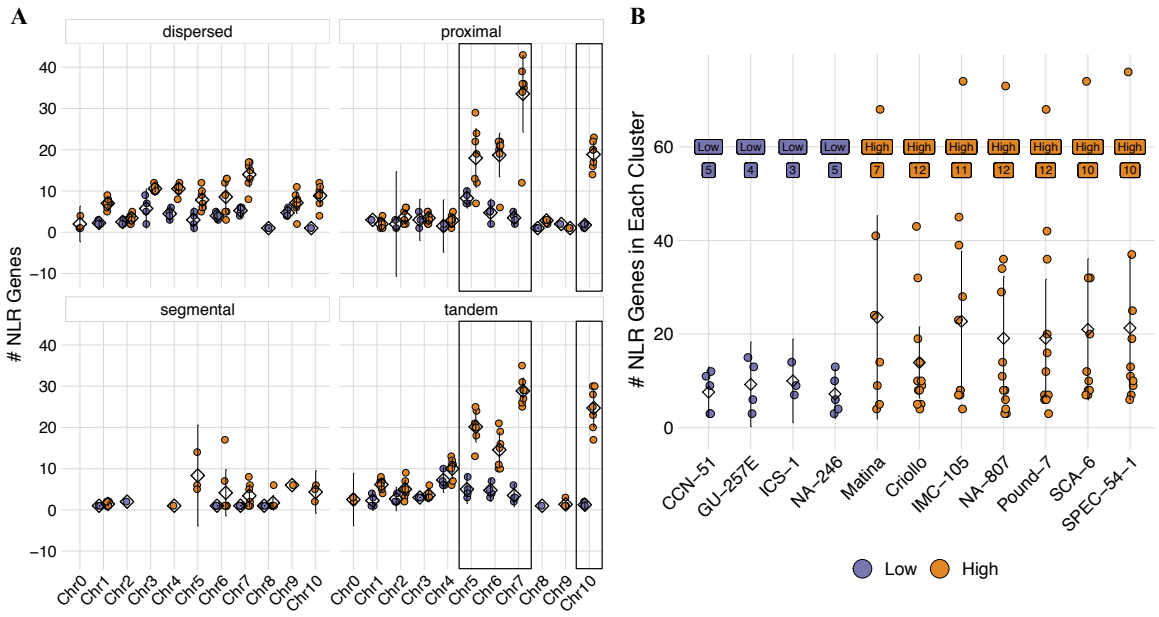


Figure 4-10: Number, size, and location of NLR clusters. (A) The genomic distribution of NLRs in each duplicate type, for Low CNV (purple) and High CNV (orange) genotypes. Each point represents the number of NLRs for a particular genotype. Boxes outline the four chromosomes with the highest NLR density. Means are represented by diamonds. Lines represent 95% confidence intervals. (B) Number and size of NLR clusters for each genotype, for Low CNV (purple) and High CNV (orange) genotypes. Each point represents a single NLR cluster. Mean cluster size for each genotype is represented by a diamond. Lines represent 95% confidence intervals. Boxed values indicate the number of NLR clusters for each genotype. Differences in both mean cluster number and mean cluster size between Low CNV and High CNV genotypes were significant (cluster number: mean difference = 6.32, Mann-Whitney test, p-value < 0.01; cluster size: mean difference = 11.48, Mann-Whitney test, p-value < 0.05).

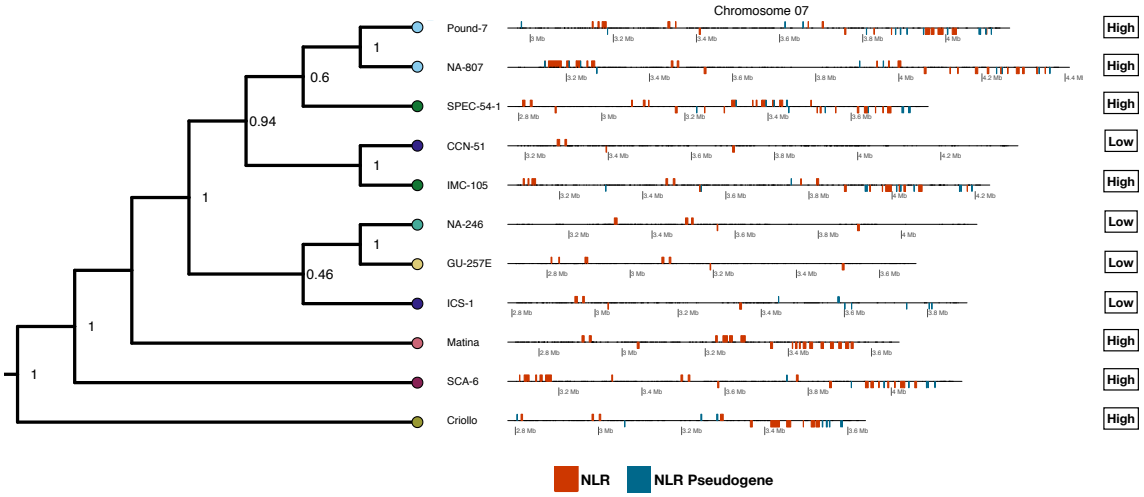


Figure 4-11. Synteny of an NLR cluster expanded through local duplications. Tandem and proximal duplications drove the expansion of NLR copy number in this homologous region of chromosome 7. NLR genes and NLR pseudogenes are shown as orange and blue bars, respectively. The phylogeny on the left indicates evolutionary relationships between the 11 genotypes used for this study. Labels on the right side indicate whether a genotype is in the Low or High CNV group.

Discussion

Recognition of pathogen challenge is the first step in plant defense response and subsequent resistance or susceptibility. This recognition takes place either extracellularly, by PRRs, or intracellularly, by NLRs (S. Kim et al. 2017). Co-evolution between pathogen effectors and plant NLR genes has resulted in an arms race typified by large and diverse repertoires of both effector and NLR gene families (Haas et al. 2009; W. Wang et al. 2021; B. C. Meyers et al. 1998). The large number of high-quality genomes sequenced over the last two decades have revealed huge variation in NLR copy number between even closely related species (Y. Zhang et al. 2016). However, due to the high cost of sequencing, assembling, and annotating high-quality genomes, we still have a very limited understanding of genome-wide NLR diversity within a

single species. In this study, we investigated the evolution of NLR content across 11 high-quality genome assemblies of the chocolate tree, *Theobroma cacao,* with the goal of gaining further insight into both NLR evolution and cacao's interactions with its microbial environment.

NLR copy number was divided into two discrete groups, between which there was a 3-fold difference in gene content (Figure 4-1B). It is hard to know how this compares to other species, but the limited evidence we have suggests this is a high degree of variation. For instance, across 64 genotypes of *A. thaliana* Van de Weyer et al. found an approximately 1.5-fold variation in NLR copy number (Van de Weyer et al. 2019a). Likewise, NLR copy number varied < 1-fold across five genotypes of *C. annuum* (S. Kim et al. 2017). This variation is likely possible for many reasons, but the two most probable factors are limited gene flow among populations and a long divergence time. Limited gene flow between populations of cacao (Chapter 2) means variation is not homogenized through mating, allowing for greater diversification (Slatkin 1993). Long divergence times, i.e. time since speciation, means there is greater opportunity for variants to arise and fix in populations. Cacao diverged from its most recent common ancestor 9.9 million years ago (Richardson et al. 2015), making it an old lineage relative to the other two species for which we have estimates of NLR copy number variation. *A. thaliana* diverged from other *Arabidopsis* species approximately 6 million years ago (Xiao et al. 2008; Z. Zhu et al. 2016), and *C. annuum* diverged from other *Capsicum* species sometime in the last 3.4 million years (Särkinen et al. 2013; Carrizo García et al. 2016). While speciation time is certainly not the only factor controlling genetic variability, it is still important for population-level differentiation (Haag et al. 2005). Together, long divergence times and stratified populations may have helped facilitate large differences in NLR gene content across genotypes. The selective forces that drive the maintenance of high NLR copy number, however, are far less clear. On one hand, NLR genes are energetically costly to produce in the absence of their cognate pathogen (Tian et al. 2003), and their mis-regulation can lead to autoimmunity (Lolle et al. 2017). At the same time, within-

species NLR diversity is important for the colonization of new habitats in wild tomato populations (Stam, Silva-Arias, and Tellier 2019), and decreased NLR polymorphism is associated with greater susceptibility to species-specific pathogens (Marden et al. 2017). Thus, depending on the scenario, copy number variation could be either beneficial or detrimental.

Differences in NLR copy number were not associated with resistance or susceptibility to a range of cacao diseases (Figure 4-3). This is largely unsurprising, since most cacao pathogens did not co-evolve with cacao and are therefore unlikely to be strong drives of copy number variation (B. A. Bailey and Meinhardt 2018). The exceptions to this being witches' broom and, to some extent, frost pod rot, but even they were not associated with NLR copy number (B. A. Bailey et al. 2018; Meinhardt et al. 2008). It is still possible, however, that copy number variation was driven by co-evolution with one or multiple unknown pathogens.

Most variation in NLR content was localized to Chr5, Chr6, Chr7, and Chr10 (Figure 4-2A), together which accounted for 45-85% of all NLRs, depending on genotype. It is on these four chromosomes that we also observe the greatest concentration of pseudogenes (Figure 4-2B). These results are consistent with previous findings, that expansion of NLR clusters occurs via a birth-and-death process leading to the formation of NLR hotspots (Blake C. Meyers et al. 2003; Mizuno et al. 2020). Developed by Ota and Nei as a counterpoint to models of concerted evolution, the birth-and-death model was originally conceived as an explanation for patterns of evolution observed in the animal major histocompatibility complex (MHC) and immunoglobulin (Ig) gene families (Ota and Nei 1994; Nei, Gu, and Sitnikova 1997). They found that members of these gene families were more closely related to orthologous genes in other species than they were genes from the same species, inconsistent with expectations under a model of concerted evolution. MHC and Ig genes instead followed a pattern of repeated gene duplication followed by either retention, sometimes for long evolutionary time (Klein 1987), or pseudogenization.

Birth-and-death evolution has now been used to explain the expansion or contraction of many gene families, including the ubiquitin family (Nei, Rogozin, and Piontkivska 2000), the fatty acyl-CoA reductase family (Finet et al. 2019), and the ATP-binding cassette family (Annilo et al. 2006). Clusters of NLR genes were first shown to evolve through a birth-and-death like process in lettuce (B. C. Meyers et al. 1998), but this pattern that has been repeatedly confirmed in numerous other species (Stam et al. 2019; Jeong et al. 2001; Mizuno et al. 2020; Blake C. Meyers et al. 2003). This same pattern of birth-and-death evolution was present across our cacao genomes, evidenced by the fact High CNV genotypes had a much higher number of NLR pseudogenes than Low CNV genotypes, and that pseudogene density exactly mirrored NLR density (Figure 2A-B).

The singular exception to the birth-and-death model was ICS-1, which possessed 2x the number of pseudogenes as it did NLR genes, and 3-4x the number of pseudogenes as the High CNV genotypes (Figure 4-2B). ICS-1 pseudogenes were localized to the four NLR-dense chromosomes, similar to the High CNV genotypes, but patterns of pseudogenization were different between the two sets. Most of ICS-1's pseudogenes originate from a narrow set of parents, rather than the 1:1 or 2:1 pseudogene:parent NLR relationship we see in the High CNV genotypes. ICS-1's unique pattern of NLR abundance could be the result of two possible scenarios. First, ICS-1 may have undergone rapid expansion of NLRs in a manner similar to the high CNV genotypes, but somewhere along the line went through a large scale pseudogenization event, likely mediated by retrotransposition (Esnault, Maestre, and Heidmann 2000; Ding et al. 2006). This option, however, seems unlikely given TE abundance was similar across all genotypes (Figure 4-5 and 4-6). The second option is that pseudogene expansion may have occurred through the same mechanisms that produced High and Low CNV genotypes, but rather than duplicating functional NLRs, pseudogenes were duplicated (S. Kim et al. 2017). This scenario is supported by the pattern of pseudogene parentage outlined above. That is, repeated

duplication of a pseudogene would result in the a skewed pseudogene:parent NLR ratio, as we observe in ICS-1 relative to the High CNV genotypes. A combination of these two scenarios is also possible, e.g. pseudogenization via retrotransposition followed by pseudogene duplication by unequal crossing over. Ascertaining the respective likelihoods of these scenarios, however, requires further comparative analyses. For instance, if ICS-1 underwent a large scale pseudogenization event, most of its NLR pseudogenes should be syntenic to functional NLRs in the High CNV genotypes. Likewise, processed pseudogenes formed through retrotransposition should lack introns and have a 3' poly-A tail (Ding et al. 2006). If ICS-1 pseudogene expansion occurred through duplication of existing pseudogenes, however, intron-exon architecture should be more similar across duplicates than they are to their respective parent NLRs. Interestingly, ICS-1 is susceptible to several cacao pathogens (Figure 4-3) (Fister et al. 2020; Phillips-Mora et al. 2005) but is also known for its high yield (Araújo et al. 2009), indicative of a potential trade-off between growth and defense. However, further work is required to test this hypothesis.

Differences in NLR copy number appeared to be caused by tandem and proximal duplications that resulted in the formation of NLR clusters. High CNV genotypes had significantly more and larger NLR clusters than Low CNV genotypes (Figure 4-10B), nearly all of which were localized to the same four chromosomes mentioned above (Figure 4-10A). This rapid expansion of NLRs could have been caused by two mechanisms. The first is TE-mediated gene duplication, which has a demonstrated role in NLR expansion (S. Kim et al. 2017). Gene duplication by TEs is most often accomplished by class I transposable elements like LTR and LINE elements (Xiao et al. 2008; Zhu et al. 2016). However, rolling circle Helitrons and DNA Pack-MULEs are also capable of gene duplication (Jiang et al. 2011; Lai et al. 2005). To this end, we investigated the abundance and density of five TE classes across 10 of our genomes. We found no association between TE abundance and NLR abundance (Figure 4-5 and 4-6), both when viewing the data in aggregate and when separating TEs by chromosome. Thus, it appears

cacao's rapid expansion of NLRs was not TE-mediated. The other mechanism most likely to generate tandem and proximal duplications is unequal crossing over (Leister 2004; Michelmore and Meyers 1998), which occurs when homologous sequences are incorrectly paired during meiosis. Once tandem and proximal duplicates are formed, the risk of further homologous mismatches increases, resulting in elongation of duplicate arrays and the subsequent expansion of gene families. Indeed, many well-known gene clusters are the result of unequal crossing over, including the human CYP2D6 cluster (Heim and Meyer 1992), the fruit fly glutamate tRNA cluster (Hosbach, Silberklang, and McCarthy 1980), and the flax M locus (Anderson et al. 1997). And, while we did not explicitly test whether unequal crossing over caused the observed patterns of NLR expansion, our results are consistent with this mechanism.

NLR genes are one of the first layers of pathogen defense in plants. Investigating NLR diversity across multiple populations of a single species is therefore necessary to understand how organisms interact with the environment, shaping their ecology, and for harnessing their diversity to breed more resilient crops. However, due to their complex evolutionary histories, investigating intraspecies NLR evolution is challenging. Here, we examined the evolution of NLR genes across 11 genotypes of *Theobroma cacao*. Together, our results suggest local duplications can radically reshape gene families over short evolutionary time scales, creating a source of NLR diversity that could be utilized to enrich our understanding of both plant-pathogen interactions and resistance breeding.

Acknowledgments

Thank you to Lena Sheaffer for her assistance in project and laboratory management. Thank you to Paula Ralph for her work extracting genomic DNA. Thank you to Craig Praul and the Huck Institutes of Life Sciences Genomics Core Facility. This work was supported by The

Pennsylvania State University College of Agricultural Sciences, the Huck Institutes of the Life Sciences, the Penn State Endowed Program in Molecular Biology of Cacao, NSF Plant Genome Research Award 1546863 and by the Agriculture and Food Research Initiative (grant number 2018-07789 and accession number 1019277) from the USDA National Institute of Food and Agriculture.

References

Alonge, Michael, Sebastian Soyk, Srividya Ramakrishnan, Xingang Wang, Sara Goodwin, Fritz J. Sedlazeck, Zachary B. Lippman, and Michael C. Schatz. 2019. "RaGOO: Fast and Accurate Reference-Guided Scaffolding of Draft Genomes." *Genome Biology* 20 (1): 224.

Andersen, Ethan J., Madhav P. Nepal, Jordan M. Purintun, Dillon Nelson, Glykeria Mermigka, and Panagiotis F. Sarris. 2020. "Wheat Disease Resistance Genes and Their Diversification through Integrated Domain Fusions." *Frontiers in Genetics* 11 (August): 898.

Anderson, Peter A., Gregory J. Lawrence, Bronwyn C. Morrish, Michael A. Ayliffe, E. Jean Finnegan, and Jeffrey G. Ellis. 1997. "Inactivation of the Flax Rust Resistance Gene M Associated with Loss of a Repeated Unit within the Leucine-Rich Repeat Coding Region." *The Plant Cell* 9 (4): 641.

Annilo, Tarmo, Zhang-Qun Chen, Sergey Shulenin, Julie Costantino, Lauren Thomas, Hong Lou, Stefan Stefanov, and Michael Dean. 2006. "Evolution of the Vertebrate ABC Gene Family: Analysis of Gene Birth and Death." *Genomics* 88 (1): 1–11.

Araújo, Ioná S., Gonçalo A. de Souza Filho, Messias G. Pereira, Fábio G. Faleiro, Vagner T. de Queiroz, Cláudia T. Guimarães, Maurílio A. Moreira, et al. 2009. "Mapping of

Quantitative Trait Loci for Butter Content and Hardness in Cocoa Beans (Theobroma Cacao L.).” *Plant Molecular Biology Reporter* 27 (2): 177–83.

Argout, X., G. Martin, G. Droc, O. Fouet, K. Labadie, E. Rivals, J. M. Aury, and C. Lanaud. 2017. “The Cacao Criollo Genome v2.0: An Improved Version of the Genome for Genetic and Functional Genomic Studies.” *BMC Genomics* 18 (1). https://doi.org/10.1186/s12864-017-4120-9.

Bailey, Bryan A., Harry C. Evans, Wilbert Phillips-Mora, Shahin S. Ali, and Lyndel W. Meinhardt. 2018. “Moniliophthora Roreri, Causal Agent of Cacao Frosty Pod Rot.” *Molecular Plant Pathology* 19 (7): 1580–94.

Bailey, Bryan A., and Lyndel W. Meinhardt, eds. 2018. *Cacao Diseases*. Cham, Switzerland: Springer International Publishing.

Bailey, Timothy L., Mikael Boden, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, and William S. Noble. 2009. “MEME SUITE: Tools for Motif Discovery and Searching.” *Nucleic Acids Research* 37 (Web Server issue): W202-8.

Białas, Aleksandra, Thorsten Langner, Adeline Harant, Mauricio P. Contreras, Clare Em Stevenson, David M. Lawson, Jan Sklenar, et al. 2021. “Two NLR Immune Receptors Acquired High-Affinity Binding to a Fungal Effector through Convergent Evolution of Their Integrated Domain.” *ELife* 10 (July). https://doi.org/10.7554/eLife.66961.

Biezen, E. A. van der, and J. D. Jones. 1998. “The NB-ARC Domain: A Novel Signalling Motif Shared by Plant Resistance Gene Products and Regulators of Cell Death in Animals.” *Current Biology: CB* 8 (7): R226-7.

Boetzer, Marten, and Walter Pirovano. 2012. “Toward Almost Closed Genomes with GapFiller.” *Genome Biology* 13 (6): R56.

Campbell, Michael S., Meiyee Law, Carson Holt, Joshua C. Stein, Gaurav D. Moghe, David E. Hufnagel, Jikai Lei, et al. 2014. “MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations.” *Plant Physiology* 164 (2): 513–24.

Carrizo García, Carolina, Michael H. J. Barfuss, Eva M. Sehr, Gloria E. Barboza, Rosabelle Samuel, Eduardo A. Moscone, and Friedrich Ehrendorfer. 2016. “Phylogenetic Relationships, Diversification and Expansion of Chili Peppers (Capsicum, Solanaceae).” *Annals of Botany* 118 (1): 35–51.

Chen, Li-Qing. 2014. “SWEET Sugar Transporters for Phloem Transport and Pathogen Nutrition.” *The New Phytologist* 201 (4): 1150–55.

Chinchilla, Delphine, Cyril Zipfel, Silke Robatzek, Birgit Kemmerling, Thorsten Nürnberger, Jonathan D. G. Jones, Georg Felix, and Thomas Boller. 2007. “A Flagellin-Induced Complex of the Receptor FLS2 and BAK1 Initiates Plant Defence.” *Nature* 448 (7152): 497–500.

Dangl, J. L., and J. D. Jones. 2001. “Plant Pathogens and Integrated Defence Responses to Infection.” *Nature* 411 (6839): 826–33.

Ding, Wenyong, Lin Lin, Bing Chen, and Jianwu Dai. 2006. “L1 Elements, Processed Pseudogenes and Retrogenes in Mammalian Genomes.” *IUBMB Life* 58 (12): 677–85.

Eddy, Sean R. 2011. “Accelerated Profile HMM Searches.” *PLoS Computational Biology* 7 (10): e1002195.

Edgar, Robert C. 2010. “Search and Clustering Orders of Magnitude Faster than BLAST.” *Bioinformatics (Oxford, England)* 26 (19): 2460–61.

Ellinghaus, David, Stefan Kurtz, and Ute Willhoeft. 2008. “LTRharvest, an Efficient and Flexible Software for de Novo Detection of LTR Retrotransposons.” *BMC Bioinformatics* 9 (1): 18.

Esnault, C., J. Maestre, and T. Heidmann. 2000. “Human LINE Retrotransposons Generate Processed Pseudogenes.” *Nature Genetics* 24 (4): 363–67.

Finet, Cédric, Kailey Slavik, Jian Pu, Sean B. Carroll, and Henry Chung. 2019. “Birth-and-Death Evolution of the Fatty Acyl-CoA Reductase (FAR) Gene Family and Diversification of Cuticular Hydrocarbon Synthesis in Drosophila.” *Genome Biology and Evolution* 11 (6): 1541–51.

Fister, Andrew S., Mariela E. Leandro-Muñoz, Dapeng Zhang, James H. Marden, Peter Tiffin, Claude dePamphilis, Siela Maximova, and Mark J. Guiltinan. 2020. “Widely Distributed Variation in Tolerance to Phytophthora Palmivora in Four Genetic Groups of Cacao.” *Tree Genetics & Genomes* 16 (1). https://doi.org/10.1007/s11295-019-1396-8.

Flor, H. H. 1971. “Current Status of the Gene-for-Gene Concept.” *Annual Review of Phytopathology* 9 (1): 275–96.

Gómez-Gómez, L., and T. Boller. 2000. “FLS2: An LRR Receptor-like Kinase Involved in the Perception of the Bacterial Elicitor Flagellin in Arabidopsis.” *Molecular Cell* 5 (6): 1003–11.

Guy, Lionel, Jens Roat Kultima, and Siv G. E. Andersson. 2010. “GenoPlotR: Comparative Gene and Genome Visualization in R.” *Bioinformatics (Oxford, England)* 26 (18): 2334–35.

Haag, Christoph R., Myriam Riek, Jürgen W. Hottinger, V. Ilmari Pajunen, and Dieter Ebert. 2005. “Genetic Diversity and Genetic Differentiation in Daphnia Metapopulations with Subpopulations of Known Age.” *Genetics* 170 (4): 1809–20.

Haas, Brian J., Sophien Kamoun, Michael C. Zody, Rays H. Y. Jiang, Robert E. Handsaker, Liliana M. Cano, Manfred Grabherr, et al. 2009. “Genome Sequence and Analysis of the Irish Potato Famine Pathogen Phytophthora Infestans.” *Nature* 461 (7262): 393–98.

Hämälä, Tuomas, Eric K. Wafula, Mark J. Guiltinan, Paula E. Ralph, Claude W. dePamphilis, and Peter Tiffin. 2021. “Genomic Structural Variants Constrain and Facilitate Adaptation

in Natural Populations of Theobroma Cacao, the Chocolate Tree." *Proceedings of the National Academy of Sciences of the United States of America* 118 (35): e2102914118.

Han, Yujun, and Susan R. Wessler. 2010. "MITE-Hunter: A Program for Discovering Miniature Inverted-Repeat Transposable Elements from Genomic Sequences." *Nucleic Acids Research* 38 (22): e199.

Heim, M. H., and U. A. Meyer. 1992. "Evolution of a Highly Polymorphic Human Cytochrome P450 Gene Cluster: CYP2D6." *Genomics* 14 (1): 49–58.

Holt, Carson, and Mark Yandell. 2011. "MAKER2: An Annotation Pipeline and Genome-Database Management Tool for Second-Generation Genome Projects." *BMC Bioinformatics* 12 (1): 491.

Hosbach, H. A., M. Silberklang, and B. J. McCarthy. 1980. "Evolution of a D. Melanogaster Glutamate TRNA Gene Cluster." *Cell* 21 (1): 169–78.

Jackman, Shaun D., Lauren Coombe, Justin Chu, Rene L. Warren, Benjamin P. Vandervalk, Sarah Yeo, Zhuyi Xue, et al. 2018. "Tigmint: Correcting Assembly Errors Using Linked Reads from Large Molecules." *BioRxiv*. bioRxiv. https://doi.org/10.1101/304253.

Jelesko, J. G., R. Harper, M. Furuya, and W. Gruissem. 1999. "Rare Germinal Unequal Crossing-over Leading to Recombinant Gene Formation and Gene Duplication in Arabidopsis Thaliana." *Proceedings of the National Academy of Sciences of the United States of America* 96 (18): 10302–7.

Jeong, S. C., A. J. Hayes, R. M. Biyashev, and M. A. Saghai Maroof. 2001. "Diversity and Evolution of a Non-TIR-NBS Sequence Family That Clusters to a Chromosomal ''hotspot" for Disease Resistance Genes in Soybean." *Theoretical and Applied Genetics* 103 (2–3): 406–14.

Jiang, Ning, Ann A. Ferguson, R. Keith Slotkin, and Damon Lisch. 2011. "Pack-Mutator-like Transposable Elements (Pack-MULEs) Induce Directional Modification of Genes

through Biased Insertion and DNA Acquisition." *Proceedings of the National Academy of Sciences of the United States of America* 108 (4): 1537–42.

Johal, G. S., and S. P. Briggs. 1992. "Reductase Activity Encoded by the HM1 Disease Resistance Gene in Maize." *Science (New York, N.Y.)* 258 (5084): 985–87.

Jones, Jonathan D. G., and Jeffery L. Dangl. 2006. "The Plant Immune System." *Nature* 444 (7117): 323–29.

Jubic, Lance M., Svenja Saile, Oliver J. Furzer, Farid El Kasmi, and Jeffery L. Dangl. 2019. "Help Wanted: Helper NLRs and Plant Immune Responses." *Current Opinion in Plant Biology* 50 (August): 82–94.

Jupe, Florian, Leighton Pritchard, Graham J. Etherington, Katrin Mackenzie, Peter J. A. Cock, Frank Wright, Sanjeev Kumar Sharma, et al. 2012. "Identification and Localisation of the NB-LRR Gene Family within the Potato Genome." *BMC Genomics* 13 (1): 75.

Jupe, Florian, Kamil Witek, Walter Verweij, Jadwiga Sliwka, Leighton Pritchard, Graham J. Etherington, Dan Maclean, 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." *The Plant Journal: For Cell and Molecular Biology* 76 (3): 530–44.

Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. 2005. "Repbase Update, a Database of Eukaryotic Repetitive Elements." *Cytogenetic and Genome Research* 110 (1–4): 462–67.

Kapitonov, V. V., and J. Jurka. 2001. "Rolling-Circle Transposons in Eukaryotes." *Proceedings of the National Academy of Sciences of the United States of America* 98 (15): 8714–19.

Kapitonov, Vladimir V., and Jerzy Jurka. 2006. "Self-Synthesizing DNA Transposons in Eukaryotes." *Proceedings of the National Academy of Sciences of the United States of America* 103 (12): 4540–45.

Katoh, Kazutaka, Kei-Ichi Kuma, Hiroyuki Toh, and Takashi Miyata. 2005. “MAFFT Version 5: Improvement in Accuracy of Multiple Sequence Alignment.” *Nucleic Acids Research* 33 (2): 511–18.

Kim, Myung-Shin, Geun Young Chae, Soohyun Oh, Jihyun Kim, Hyunggon Mang, Seungill Kim, and Doil Choi. 2021. “Comparative Analysis of de Novo Genomes Reveals Dynamic Intra-Species Divergence of NLRs in Pepper.” *BMC Plant Biology* 21 (1): 247.

Kim, Seungill, Jieun Park, Seon-In Yeom, Yong-Min Kim, Eunyoung Seo, Ki-Tae Kim, Myung-Shin Kim, et al. 2017. “New Reference Genome Sequences of Hot Pepper Reveal the Massive Evolution of Plant Disease-Resistance Genes by Retroduplication.” *Genome Biology* 18 (1): 210.

Klein, J. 1987. “Origin of Major Histocompatibility Complex Polymorphism: The Trans-Species Hypothesis.” *Human Immunology* 19 (3): 155–62.

Klessig, D. F., J. Durner, R. Noad, D. A. Navarre, D. Wendehenne, D. Kumar, J. M. Zhou, et al. 2000. “Nitric Oxide and Salicylic Acid Signaling in Plant Defense.” *Proceedings of the National Academy of Sciences of the United States of America* 97 (16): 8849–55.

Kobayashi, Michie, Miki Yoshioka, Shuta Asai, Hironari Nomura, Kazuo Kuchimura, Hitoshi Mori, Noriyuki Doke, and Hirofumi Yoshioka. 2012. “StCDPK5 Confers Resistance to Late Blight Pathogen but Increases Susceptibility to Early Blight Pathogen in Potato via Reactive Oxygen Species Burst.” *The New Phytologist* 196 (1): 223–37.

Kourelis, Jiorgos, and Renier A. L. van der Hoorn. 2018. “Defended to the Nines: 25 Years of Resistance Gene Cloning Identifies Nine Mechanisms for R Protein Function.” *The Plant Cell* 30 (2): 285–99.

Kramerov, D. A., and N. S. Vassetzky. 2011. “Origin and Evolution of SINEs in Eukaryotic Genomes.” *Heredity* 107 (6): 487–95.

Lai, Jinsheng, Yubin Li, Joachim Messing, and Hugo K. Dooner. 2005. “Gene Movement by Helitron Transposons Contributes to the Haplotype Variability of Maize.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (25): 9068–73.

Lapin, Dmitry, Viera Kovacova, Xinhua Sun, Joram A. Dongus, Deepak Bhandari, Patrick von Born, Jaqueline Bautor, et al. 2019. “A Coevolved EDS1-SAG101-NRG1 Module Mediates Cell Death Signaling by TIR-Domain Immune Receptors.” *The Plant Cell* 31 (10): 2430–55.

Leister, Dario. 2004. “Tandem and Segmental Gene Duplication and Recombination in the Evolution of Plant Disease Resistance Gene.” *Trends in Genetics: TIG* 20 (3): 116–22.

Lin, Xiao, Yu Zhang, Hanhui Kuang, and Jiongjiong Chen. 2013. “Frequent Loss of Lineages and Deficient Duplications Accounted for Low Copy Number of Disease Resistance Genes in Cucurbitaceae.” *BMC Genomics* 14 (1): 335.

Lolle, Signe, Christiaan Greeff, Klaus Petersen, Milena Roux, Michael Krogh Jensen, Simon Bressendorff, Eleazar Rodriguez, Kenneth Sømark, John Mundy, and Morten Petersen. 2017. “Matching NLR Immune Receptors to Autoimmunity in Camta3 Mutants Using Antimorphic NLR Alleles.” *Cell Host & Microbe* 21 (4): 518-529.e4.

Lüdecke, Daniel, Mattan Ben-Shachar, Indrajeet Patil, Philip Waggoner, and Dominique Makowski. 2021. “Performance: An R Package for Assessment, Comparison and Testing of Statistical Models.” *Journal of Open Source Software* 6 (60): 3139.

Marden, J. H., S. A. Mangan, M. P. Peterson, E. Wafula, H. W. Fescemyer, J. P. Der, C. W. dePamphilis, and L. S. Comita. 2017. “Ecological Genomics of Tropical Trees: How Local Population Size and Allelic Diversity of Resistance Genes Relate to Immune Responses, Cosusceptibility to Pathogens, and Negative Density Dependence.” *Molecular Ecology* 26 (9): 2498–2513.

Maximova, S., C. Miller, G. Antúnez de Mayolo, S. Pishak, A. Young, and M. J. Guiltinan. 2003. "Stable Transformation of Theobroma Cacao L. and Influence of Matrix Attachment Regions on GFP Expression." *Plant Cell Reports* 21 (9): 872–83.

Meinhardt, Lyndel W., Johana Rincones, Bryan A. Bailey, M. Catherine Aime, Gareth W. Griffith, Dapeng Zhang, and Gonçalo A. G. Pereira. 2008. "Moniliophthora Perniciosa, the Causal Agent of Witches' Broom Disease of Cacao: What's New from This Old Foe?" *Molecular Plant Pathology* 9 (5): 577–88.

Meyers, B. C., D. B. Chin, K. A. Shen, S. Sivaramakrishnan, D. O. Lavelle, Z. Zhang, and R. W. Michelmore. 1998. "The Major Resistance Gene Cluster in Lettuce Is Highly Duplicated and Spans Several Megabases." *The Plant Cell* 10 (11): 1817–32.

Meyers, Blake C., Alexander Kozik, Alyssa Griego, Hanhui Kuang, and Richard W. Michelmore. 2003. "Genome-Wide Analysis of NBS-LRR-Encoding Genes in Arabidopsis." *The Plant Cell* 15 (4): 809–34.

Michelmore, R. W., and B. C. Meyers. 1998. "Clusters of Resistance Genes in Plants Evolve by Divergent Selection and a Birth-and-Death Process." *Genome Research* 8 (11): 1113–30.

Mighell, A. J., N. R. Smith, P. A. Robinson, and A. F. Markham. 2000. "Vertebrate Pseudogenes." *FEBS Letters* 468 (2–3): 109–14.

Mirarab, S., R. Reaz, Md S. Bayzid, T. Zimmermann, M. S. Swenson, and T. Warnow. 2014. "ASTRAL: Genome-Scale Coalescent-Based Species Tree Estimation." *Bioinformatics (Oxford, England)* 30 (17): i541-8.

Mistry, Jaina, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A. Salazar, Erik L. L. Sonnhammer, Silvio C. E. Tosatto, et al. 2021. "Pfam: The Protein Families Database in 2021." *Nucleic Acids Research* 49 (D1): D412–19.

Mizuno, Hiroshi, Satoshi Katagiri, Hiroyuki Kanamori, Yoshiyuki Mukai, Takuji Sasaki, Takashi Matsumoto, and Jianzhong Wu. 2020. “Evolutionary Dynamics and Impacts of Chromosome Regions Carrying R-Gene Clusters in Rice.” *Scientific Reports* 10 (1): 872.

Motamayor, Juan C., Philippe Lachenaud, Jay Wallace da Silva e Mota, Rey Loor, David N. Kuhn, J. Steven Brown, and Raymond J. Schnell. 2008. “Geographic and Genetic Population Differentiation of the Amazonian Chocolate Tree (Theobroma Cacao L).” *PloS One* 3 (10): e3311.

Motamayor, Juan C., Keithanne Mockaitis, Jeremy Schmutz, Niina Haiminen, Donald Livingstone 3rd, Omar Cornejo, Seth D. Findley, et al. 2013. “The Genome Sequence of the Most Widely Cultivated Cacao Type and Its Use to Identify Candidate Genes Regulating Pod Color.” *Genome Biology* 14 (6): r53.

Nei, M., X. Gu, and T. Sitnikova. 1997. “Evolution by the Birth-and-Death Process in Multigene Families of the Vertebrate Immune System.” *Proceedings of the National Academy of Sciences of the United States of America* 94 (15): 7799–7806.

Nei, M., I. B. Rogozin, and H. Piontkivska. 2000. “Purifying Selection and Birth-and-Death Evolution in the Ubiquitin Gene Family.” *Proceedings of the National Academy of Sciences of the United States of America* 97 (20): 10866–71.

Ngou, Bruno Pok Man, Jonathan D. G. Jones, and Pingtao Ding. 2021. “Plant Immune Networks.” *Trends in Plant Science*, September. https://doi.org/10.1016/j.tplants.2021.08.012.

Novikova, Polina Yu, Nora Hohmann, Viktoria Nizhynska, Takashi Tsuchimatsu, Jamshaid Ali, Graham Muir, Alessia Guggisberg, et al. 2016. “Sequencing of the Genus Arabidopsis Identifies a Complex History of Nonbifurcating Speciation and Abundant Trans-Specific Polymorphism.” *Nature Genetics* 48 (9): 1077–82.

Ota, T., and M. Nei. 1994. “Divergent Evolution and Evolution by the Birth-and-Death Process in the Immunoglobulin VH Gene Family.” *Molecular Biology and Evolution* 11 (3): 469–82.

Phillips-Mora, W., J. Castillo, U. Krauss, E. Rodriguez, and M. J. Wilkinson. 2005. “Evaluation of Cacao (Theobroma Cacao) Clones against Seven Colombian Isolates of Moniliophthora Roreri from Four Pathogen Genetic Groups.” *Plant Pathology* 54 (4): 483–90.

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2010. “FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments.” *PloS One* 5 (3): e9490.

Quevillon, E., V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez. 2005. “InterProScan: Protein Domains Identifier.” *Nucleic Acids Research* 33 (Web Server issue): W116-20.

Richardson, James E., Barbara A. Whitlock, Alan W. Meerow, and Santiago Madriñán. 2015. “The Age of Chocolate: A Diversification History of Theobroma and Malvaceae.” *Frontiers in Ecology and Evolution* 3 (November). https://doi.org/10.3389/fevo.2015.00120.

Särkinen, Tiina, Lynn Bohs, Richard G. Olmstead, and Sandra Knapp. 2013. “A Phylogenetic Framework for Evolutionary Study of the Nightshades (Solanaceae): A Dated 1000-Tip Tree.” *BMC Evolutionary Biology* 13 (September): 214.

Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. “BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs.” *Bioinformatics (Oxford, England)* 31 (19): 3210–12.

Slatkin, Montgomery. 1993. “Isolation by Distance in Equilibrium and Non-Equilibrium Populations.” *Evolution; International Journal of Organic Evolution* 47 (1): 264–79.

Stam, Remco, Tetyana Nosenko, Anja C. Hörger, Wolfgang Stephan, Michael Seidel, José M. M. Kuhn, Georg Haberer, and Aurelien Tellier. 2019. “The de Novo Reference Genome and Transcriptome Assemblies of the Wild Tomato Species Solanum Chilense Highlights Birth and Death of NLR Genes between Tomato Species.” *G3 (Bethesda, Md.)* 9 (12): 3933–41.

Stam, Remco, Daniela Scheikl, and Aurélien Tellier. 2016. “Pooled Enrichment Sequencing Identifies Diversity and Evolutionary Pressures at NLR Resistance Genes within a Wild Tomato Population.” *Genome Biology and Evolution* 8 (5): 1501–15.

Stam, Remco, Gustavo A. Silva-Arias, and Aurelien Tellier. 2019. “Subsets of NLR Genes Show Differential Signatures of Adaptation during Colonization of New Habitats.” *The New Phytologist* 224 (1): 367–79.

Steinbiss, Sascha, Ute Willhoeft, Gordon Gremme, and Stefan Kurtz. 2009. “Fine-Grained Annotation and Classification of de Novo Predicted LTR Retrotransposons.” *Nucleic Acids Research* 37 (21): 7002–13.

Steuernagel, Burkhard, Florian Jupe, Kamil Witek, Jonathan D. G. Jones, and Brande B. H. Wulff. 2015. “NLR-Parser: Rapid Annotation of Plant NLR Complements.” *Bioinformatics (Oxford, England)* 31 (10): 1665–67.

Tian, D., M. B. Traw, J. Q. Chen, M. Kreitman, and J. Bergelson. 2003. “Fitness Costs of R-Gene-Mediated Resistance in Arabidopsis Thaliana.” *Nature* 423 (6935): 74–77.

Van de Weyer, Anna-Lena, Freddy Monteiro, Oliver J. Furzer, Marc T. Nishimura, Volkan Cevik, Kamil Witek, Jonathan D. G. Jones, Jeffery L. Dangl, Detlef Weigel, and Felix Bemm. 2019. “A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis Thaliana.” *Cell* 178 (5): 1260-1272.e14.

Vaughn, Justin N., and Jeffrey L. Bennetzen. 2014. “Natural Insertions in Rice Commonly Form Tandem Duplications Indicative of Patch-Mediated Double-Strand Break Induction and

Repair." *Proceedings of the National Academy of Sciences of the United States of America* 111 (18): 6684–89.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. 4th ed. Statistics and Computing. New York, NY: Springer.

Wang, Weidong, Liyang Chen, Kevin Fengler, Joy Bolar, Victor Llaca, Xutong Wang, Chancelor B. Clark, et al. 2021. "A Giant NLR Gene Confers Broad-Spectrum Resistance to Phytophthora Sojae in Soybean." *Nature Communications* 12 (1): 6263.

Wang, Yupeng, Haibao Tang, Jeremy D. Debarry, Xu Tan, Jingping Li, Xiyin Wang, Tae-Ho Lee, et al. 2012. "MCScanX: A Toolkit for Detection and Evolutionary Analysis of Gene Synteny and Collinearity." *Nucleic Acids Research* 40 (7): e49.

Weisenfeld, Neil I., Vijay Kumar, Preyas Shah, Deanna M. Church, and David B. Jaffe. 2017. "Direct Determination of Diploid Genome Sequences." *Genome Research* 27 (5): 757–67.

Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer.

Xiao, Han, Ning Jiang, Erin Schaffner, Eric J. Stockinger, and Esther van der Knaap. 2008. "A Retrotransposon-Mediated Gene Duplication Underlies Morphological Variation of Tomato Fruit." *Science (New York, N.Y.)* 319 (5869): 1527–30.

Yeo, Sarah, Lauren Coombe, René L. Warren, Justin Chu, and Inanç Birol. 2018. "ARCS: Scaffolding Genome Drafts with Linked Reads." *Bioinformatics* 34 (5): 725–31.

Zhang, Yu, Rui Xia, Hanhui Kuang, and Blake C. Meyers. 2016. "The Diversification of Plant NBS-LRR Defense Genes Directs the Evolution of MicroRNAs That Target Them." *Molecular Biology and Evolution* 33 (10): 2692–2705.

Zhou, Huanbin, Jian Lin, Aimee Johnson, Robyn L. Morgan, Wenwan Zhong, and Wenbo Ma. 2011. "Pseudomonas Syringae Type III Effector HopZ1 Targets a Host Enzyme to

Suppress Isoflavone Biosynthesis and Promote Infection in Soybean." *Cell Host & Microbe* 9 (3): 177–86.

Zhu, Zhenglin, Shengjun Tan, Yaqiong Zhang, and Yong E. Zhang. 2016. "LINE-1-like Retrotransposons Contribute to RNA-Based Gene Duplication in Dicots." *Scientific Reports* 6 (1): 24755.

Zipfel, Cyril. 2014. "Plant Pattern-Recognition Receptors." *Trends in Immunology* 35 (7): 345–51.

Zou, Cheng, Melissa D. Lehti-Shiu, Françoise Thibaud-Nissen, Tanmay Prakash, C. Robin Buell, and Shin-Han Shiu. 2009. "Evolutionary and Expression Signatures of Pseudogenes in Arabidopsis and Rice." *Plant Physiology* 151 (1): 3–15.

Chapter 5: Retrospective

Outro

Natural variation in flavor, yield, disease resistance, and a host of other agronomic traits has been observed for millennia. This has led to their selection during breeding and subsequent crop domestication (Meyer and Purugganan 2013; Theophrastus 1989). Understanding the genotypes that underlie natural variation in phenotypes, however, has only been possible more recently (Benfey and Mitchell-Olds 2008). Moreover, despite the fact genotype-to-phenotype relationships have been extensively documented in crops like corn (Cook et al. 2012), wheat (Li et al. 2021; Driever et al. 2014), and rice (W. Chen et al. 2014; J.-Y. Chen et al. 2018), knowledge in less experimentally tractable species, like cacao, is still limited. The goal of this work is to clarify gaps in our understanding of how defense response and immunity function within cacao, with the hope of generating knowledge that can be used for its improvement.

In Chapter 1, we outlined the two fundamental questions at the heart of this dissertation: How did cacao's defense mechanisms evolve? And can we use this evolutionary information to identify genes important for disease resistance? The subsequent chapters seek to answer pieces of these two questions, dividing them into smaller portions more amenable to experimentation. In Chapter 2, we examined defense response to *P. palmivora* across four populations of cacao. We observed a large set of genes in each population that were responding uniquely, i.e. not shared across the other three groups. These results suggest wild, underutilized cacao populations could act as a rich source of genetic diversity for future crop improvement. Chapter 3 provides a broader look at immunity across the genus, investigating the evolution of defense response to *P. palmivora* across four non-cacao species of *Theobroma*. Similar to our observations in Chapter 2, we found wild relatives of *T. cacao* had both lineage-specific and conserved aspects of their

defense response. We observed several of these conserved responses in cacao as well, suggesting they are fundamentally important for defense across *Theobroma.* These analyses further support the use of evolutionary and comparative methods to identify loci that are important for disease resistance. Finally, in Chapter 4, we examined the evolution of NLR receptors across 11 cacao genotypes. We found 3-fold variation in NLR copy number across genotypes, a pattern predominantly driven by tandem and proximal duplications. These results provide an unprecedented look at intraspecific NLR diversity in cacao and give us new understanding of how cacao interacts with its natural environment. A more detailed discussion of these results is included below. We pay specific attention to the limitations of this work and future directions that could help further elucidate evolutionary and functional studies of cacao's defense response.

Induced defense responses across populations of *T. cacao*

We examined defense responses against *P. palmivora* across four populations of cacao, identifying thousands of differentially expressed genes. Approximately 40% of the genes differentially expressed in a given population were unique, i.e. not differentially expressed in any of the other three. This was true both when examining genes responding to treatment with *P. palmivora* and between resistant and susceptible varieties of cacao.

The lack of a coherent defense response suggests that some proportion of each population's immunity is lineage-specific, and that resistance/susceptibility may be determined by different genes in each population. To test this, we employed population branch statistics to estimate genetic divergence associated with resistant genotypes in each population. This revealed similar patterns as the gene expression results, albeit more pronounced, with over 75% of the selection outliers unique to each group.

Despite this uniqueness, there was considerable overlap among GO terms, suggesting each population is using somewhat different sets of genes to perform a similar set of functions. Among the terms shared across populations, both in response to *P. palmivora* treatment and resistance/susceptibility phenotype, were two that encompassed phenylpropanoid biosynthesis. Several genes involved in the core phenylpropanoid pathway were upregulated, including one called caffeoyl-shikimate esterase (*TcCSE*). *TcCSE* leads to the production of caffeic acid in many plants, an intriguing molecule owing to its well-documented antimicrobial activity (Widmer and Laurent 2006; Khan et al. 2021). Consistent with *TcCSE*'s putative function, heterologous over-expression in *N. benthamiana* resulted in the accumulation of caffeic acid. Moreover, *P. palmivora* growing on agar plates amended with caffeic acid were significantly inhibited. While the mechanism for caffeic acid's inhibition of *P. palmivora* remains unclear, these data suggest it could be cytotoxic (Khan et al. 2021).

Contrary to expectations, caffeic acid did not accumulate after pathogen challenge. This lack of caffeic acid accumulation could be due to a number of factors, including inappropriate time point selection, conversion to lignin, or derivatization into other antimicrobial compounds (Yamauchi, Yasuda, and Fukushima 2002; Khan et al. 2021). Without this piece of information, it is hard to determine the importance of caffeic acid to cacao's defense response. Future studies should focus on examining caffeic acid accumulation in pathogen-challenged tissue at time points spanning 0-48 hpi. When coupled with *TcCSE* expression measurements over the same period, this would provide greater understanding of both the importance of caffeic acid and the connection between gene expression and metabolite accumulation. Together, these results support the idea that each population has adapted to a unique array of environmental challenges, thereby generating a diverse set of potential defense responses that could be utilized by breeders to improve cacao disease resistance.

Induced defense responses across *Theobroma*

To better understand the evolution of defense response across the genus, we extended the rationale from Chapter 2 to include four non-cacao species of *Theobroma.* We challenged *T. angustifolium, T. bicolor, T. grandiflorum,* and *T. mammosum* with *P. palmivora* and assessed differential expression 48 hpi. We found, similar to Chapter 2, thousands of differentially expressed genes across all four species. Likewise, there was a large degree of overlap among GO categories, several of which were related to phenylpropanoid biosynthesis, again signaling the importance of this pathway.

We examined the degree to which defense response was orthologous, i.e. arose in a common ancestor of cacao, by investigating differentially expressed orthogroups. We found 317 orthogroups that were differentially expressed across all four species, some of which were also differentially expressed in cacao. Furthermore, branch-site tests suggested many of these orthogroups displayed signatures of diversifying selection. This set conserved defense responses likely arose in the common ancestor of these five species, and potentially even predates the formation of *Theobroma* as a genus, indicating their fundamental role in disease resistance.

The annotations for many of these orthogroups indicate well-established roles in defense response, such as the chitinase and endochitinase gene families (Siela N. Maximova et al. 2006; Y. J. Zhu et al. 2003). Two of them, OG60 and OG361, stand out as particularly interesting. This is because proteins belonging to these two families, *TcBBE8* (SCA-6_Chr6v1_16921) and *TcWRKY29* (SCA-6_Chr3v1_10161), were repeatedly shown to be interesting candidates in Chapter 2. Both genes were upregulated across all four cacao populations in response to pathogen challenge and displayed signatures of divergence among resistant genotypes in one or more populations. Phylogenies for both OG60 and OG361 revealed a number of closely related orthologs in non-cacao *Theobroma spp.* that were similarly upregulated. Thus, in two separate

experiments, taken at different time points, and with different *P. palmivora* strains, these two genes were both differentially expressed. Moreover, two separate methods, branch-site tests and population branch statistics, indicate these genes are evolving under selection.

While our current data strongly suggest the importance of these genes, their relevance to cacao defense response will need to be verified experimentally. Consistent with their fold induction, future experimentation should include over-expression followed by pathogen bioassays to assess their effect on lesion growth. Furthermore, alignment and visualization of these genomic regions should be performed across a diverse set of cacao clones, helping identify single base pairs or structural variants associated with disease resistance.

Secondary metabolites, particularly hydroxycinnamic acids (Knollenberg et al. 2020; Muroi et al. 2009), coumarins (Zeid 2002; Churngchow and Rattarasarn 2001), and purine alkaloids (Aneja and Gianfagna 2001), have demonstrated roles in resistance to pathogens across a wide range of species, including cacao. Genes with inferred functions associated with secondary metabolite biosynthesis were consistently upregulated in both the experiments from Chapter 2 and Chapter 3. These studies are not alone in identifying the differential expression of metabolite pathways during pathogen challenge (Shahin S. Ali, Shao, Lary, Strem, et al. 2017). To date, however, there has been little research on cacao's induced defense-metabolites (Neves Dos Santos et al. 2021). Investigating this diversity is therefore critical to understanding how resistant varieties defend themselves against pathogen invasion. To accomplish this, large scale pathogen challenge experiments performed across a diverse set of genotypes and/or *Theobroma spp.* followed by untargeted metabolomics (e.g. LC-MS/MS) would yield a rich set of defense-associated metabolites. At least one cacao dataset amenable to this type of study (Chapter 2, *Functional analysis of candidate gene for caffeic acid synthesis*) is currently available, but it was collected at an early infection time point that may not be representative of induced metabolite diversity. While bioinformatic tools to analyze metabolomic data have lagged behind those for

sequence data, approaches to rapidly identify metabolites from untargeted MS/MS spectra have recently been developed (M. Wang et al. 2020, 2016). Working backwards from metabolite accumulation to gene expression and/or metabolic QTL identification, similar to previously developed methods (Knollenberg et al. 2020; Bilbrey et al. 2021), could help identify candidate loci necessary for metabolite accumulation. Further functional genetics could clarify this relationship, helping identify candidate resistance markers for breeding.

Molecular evolution of cacao's immune receptors

As we saw in Chapters 2 and 3, hundreds or even thousands of genes are differentially regulated during pathogen invasion. Few of these genes, however, are as integral to plant defense as NLR receptors. Studying the intraspecific diversity of these receptors is therefore essential to understanding how a given species interacts with its microbial environment. Moreover, NLRs have a long history in breeding programs, helping create rust resistant wheat (D. Arora, Gross, and Brueggeman 2013), blast resistant rice (G.-L. Wang and Valent 2017), and blight resistant potato (Park et al. 2005).

In Chapter 4, we examined the diversity and evolution of NLR receptors across 11 genotypes of cacao. There was a 3-fold difference in NLR complement across the genotypes, twice as much variation as other species (Van de Weyer et al. 2019a; M.-S. Kim et al. 2021). Much of this variation appeared to be controlled by local tandem and proximal duplications, rather than dispersed or segmental duplications, and occured on just a few chromosomes. NLR density did not co-locate with TE density, consistent with the idea that unequal crossing over likely generated large clusters of NLRs.

For these results to be translatable to breeding programs, however, their specific roles in disease resistance will need to be clarified. Identification of NLR-effector interactions is difficult,

most often involving QTL (Díaz-Tatis et al. 2021) or association mapping (S. Arora et al. 2019), both of which are challenging in cacao. To make this process more tractable, NLRs involved in direct effector recognition could be predicted computationally. For instance, Shannon's entropy was recently used to identify hypervariable positions in orthologous NLRs (Prigozhin and Krasileva 2021). Many of the NLRs containing hypervariable regions were known to directly interact with their cognate effectors. Therefore, Shannon's entropy could be used to identify NLRs containing hypervariable regions. Putative interaction with cognate effectors could then be determined based on currently available machine learning approaches predicting protein structure and protein-protein interactions (Zeng et al. 2020; Jumper et al. 2021). This would result in a set of NLRs and their putative effectors that could be validated experimentally using protein-protein interaction assays like yeast two-hybrids.

Conclusion

Over the past 20 years, advances in genomics, metabolomics, and computation have fundamentally altered our understanding of many biological processes and have accelerated the acquisition of new knowledge. We are now able to search for the genetic underpinnings of natural variation in a host of traits, at a fraction of the previous cost. These technologies are, and will remain, a boon to agricultural research. This is especially true for both historically marginalized crop species and those that are difficult to study, such as cacao. In this dissertation, we have searched for genomic loci conferring resistance to *P. palmivora* in four cacao populations, examined the evolution of defense response across five species of *Theobroma,* and investigated the intraspecific diversity of NLR receptors in 11 genotypes of cacao. In doing so we have sequenced, extracted, assembled, and analyzed hundreds of transcriptome libraries, dozens of metabolite data sets, and tens of whole genomes. Despite these advances, disease resistance is an

incredibly complex trait, and more work needs to be done identifying genomic regions associated with resistance and susceptibility. It is an exciting time to be doing both basic and applied biological research, and I look forward to continuing this work going forward.

References

Acebo-Guerrero, Yanelis, Annia Hernández-Rodríguez, Mayra Heydrich-Pérez, Mondher El Jaziri, and Ana N. Hernández-Lauzardo. 2012. "Management of Black Pod Rot in Cacao (Theobroma CacaoL.): A Review." *Fruits* 67 (1): 41–48.

Akrofi, Andrews Yaw, Ishmael Amoako-Atta, Michael Assuah, and Eric Kumi Asare. 2015. "Black Pod Disease on Cacao (Theobroma Cacao, L) in Ghana: Spread of Phytophthora Megakarya and Role of Economic Plants in the Disease Epidemiology." *Crop Protection (Guildford, Surrey)* 72 (June): 66–75.

Alexa, Adrian, and Jörg Rahnenführer. 2009. "Gene Set Enrichment Analysis with TopGO." *Bioconductor Improv* 27: 1–26.

Ali, S. S., I. Amoako-Attah, R. A. Bailey, M. D. Strem, M. Schmidt, A. Y. Akrofi, S. Surujdeo-Maharaj, et al. 2016. "PCR-Based Identification of Cacao Black Pod Causal Agents and Identification of Biological Factors Possibly Contributing ToPhytophthora Megakarya's Field Dominance in West Africa." *Plant Pathology* 65 (7): 1095–1108.

Ali, Shahin S., Jonathan Shao, David J. Lary, Brent Kronmiller, Danyu Shen, Mary D. Strem, Ishmael Amoako-Attah, et al. 2017. "Phytophthora Megakarya and P. Palmivora, Closely Related Causal Agents of Cacao Black Pod Rot, Underwent Increases in Genome Sizes and Gene Numbers by Different Mechanisms." *Genome Biology and Evolution* 9 (3): 536–57.

Ali, Shahin S., Jonathan Shao, David J. Lary, Mary D. Strem, Lyndel W. Meinhardt, and Bryan A. Bailey. 2017. “Phytophthora Megakarya and P. Palmivora, Causal Agents of Black Pod Rot, Induce Similar Plant Defense Responses Late during Infection of Susceptible Cacao Pods.” *Frontiers in Plant Science* 8 (February): 169.

Alonge, Michael, Sebastian Soyk, Srividya Ramakrishnan, Xingang Wang, Sara Goodwin, Fritz J. Sedlazeck, Zachary B. Lippman, and Michael C. Schatz. 2019. “RaGOO: Fast and Accurate Reference-Guided Scaffolding of Draft Genomes.” *Genome Biology* 20 (1): 224.

Altschul, S. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10.

Ambrosio, Alinne Batista, Leandro Costa do Nascimento, Bruno V. Oliveira, Paulo José P. L. Teixeira, Ricardo A. Tiburcio, Daniela P. Toledo Thomazella, Adriana F. P. Leme, et al. 2013. “Global Analyses of Ceratocystis Cacaofunesta Mitochondria: From Genome to Proteome.” *BMC Genomics* 14 (1): 91.

Andersen, Ethan J., Madhav P. Nepal, Jordan M. Purintun, Dillon Nelson, Glykeria Mermigka, and Panagiotis F. Sarris. 2020. “Wheat Disease Resistance Genes and Their Diversification through Integrated Domain Fusions.” *Frontiers in Genetics* 11 (August): 898.

Anderson, Jonathan P., Ellet Badruzsaufari, Peer M. Schenk, John M. Manners, Olivia J. Desmond, Christina Ehlert, Donald J. Maclean, Paul R. Ebert, and Kemal Kazan. 2004. “Antagonistic Interaction between Abscisic Acid and Jasmonate-Ethylene Signaling Pathways Modulates Defense Gene Expression and Disease Resistance in Arabidopsis.” *The Plant Cell* 16 (12): 3460–79.

Anderson, Peter A., Gregory J. Lawrence, Bronwyn C. Morrish, Michael A. Ayliffe, E. Jean Finnegan, and Jeffrey G. Ellis. 1997. “Inactivation of the Flax Rust Resistance Gene M

Associated with Loss of a Repeated Unit within the Leucine-Rich Repeat Coding Region." *The Plant Cell* 9 (4): 641.

Aneja, Madhu, and Thomas Gianfagna. 2001. "Induction and Accumulation of Caffeine in Young, Actively Growing Leaves of Cocoa (Theobroma Cacao L.) by Wounding or Infection with Crinipellis Perniciosa." *Physiological and Molecular Plant Pathology* 59 (1): 13–16.

Annilo, Tarmo, Zhang-Qun Chen, Sergey Shulenin, Julie Costantino, Lauren Thomas, Hong Lou, Stefan Stefanov, and Michael Dean. 2006. "Evolution of the Vertebrate ABC Gene Family: Analysis of Gene Birth and Death." *Genomics* 88 (1): 1–11.

Araújo, Ioná S., Gonçalo A. de Souza Filho, Messias G. Pereira, Fábio G. Faleiro, Vagner T. de Queiroz, Cláudia T. Guimarães, Maurílio A. Moreira, et al. 2009. "Mapping of Quantitative Trait Loci for Butter Content and Hardness in Cocoa Beans (Theobroma Cacao L.)." *Plant Molecular Biology Reporter* 27 (2): 177–83.

Argout, X., G. Martin, G. Droc, O. Fouet, K. Labadie, E. Rivals, J. M. Aury, and C. Lanaud. 2017. "The Cacao Criollo Genome v2.0: An Improved Version of the Genome for Genetic and Functional Genomic Studies." *BMC Genomics* 18 (1). https://doi.org/10.1186/s12864-017-4120-9.

Argout, Xavier, Jerome Salse, Jean-Marc Aury, Mark J. Guiltinan, Gaetan Droc, Jerome Gouzy, Mathilde Allegre, et al. 2011. "The Genome of Theobroma Cacao." *Nature Genetics* 43 (2): 101–8.

Arnold, Sarah E. J., Samantha J. Forbes, David R. Hall, Dudley I. Farman, Puran Bridgemohan, Gustavo R. Spinelli, Daniel P. Bray, et al. 2019. "Floral Odors and the Interaction between Pollinating Ceratopogonid Midges and Cacao." *Journal of Chemical Ecology* 45 (10): 869–78.

Arora, D., T. Gross, and R. Brueggeman. 2013. “Allele Characterization of Genes Required for Rpg4-Mediated Wheat Stem Rust Resistance Identifies Rpg5 as the R Gene.” *Phytopathology* 103 (11): 1153–61.

Arora, Sanu, Burkhard Steuernagel, Kumar Gaurav, Sutha Chandramohan, Yunming Long, Oadi Matny, Ryan Johnson, et al. 2019. “Resistance Gene Cloning from a Wild Crop Relative by Sequence Capture and Association Genetics.” *Nature Biotechnology* 37 (2): 139–43.

Asselin, Jo Ann E., Jinshan Lin, Alvaro L. Perez-Quintero, Irene Gentzel, Doris Majerczak, Stephen O. Opiyo, Wanying Zhao, et al. 2015. “Perturbation of Maize Phenylpropanoid Metabolism by an AvrE Family Type III Effector from Pantoea Stewartii.” *Plant Physiology* 167 (3): 1117–35.

Bach, Søren Spanner, Jean-Étienne Bassard, Johan Andersen-Ranberg, Morten Emil Møldrup, Henrik Toft Simonsen, and Björn Hamberger. 2014. “High-Throughput Testing of Terpenoid Biosynthesis Candidate Genes Using Transient Expression in Nicotiana Benthamiana.” *Methods in Molecular Biology (Clifton, N.J.)* 1153: 245–55.

Badet, Thomas, and Daniel Croll. 2020. “The Rise and Fall of Genes: Origins and Functions of Plant Pathogen Pangenomes.” *Current Opinion in Plant Biology* 56 (August): 65–73.

Bailey, Bryan A., Shahin S. Ali, Andrews Y. Akrofi, and Lyndel W. Meinhardt. 2016. “Phytophthora Megakarya, a Causal Agent of Black Pod Rot in Africa.” In *Cacao Diseases*, 267–303. Cham: Springer International Publishing.

Bailey, Bryan A., Harry C. Evans, Wilbert Phillips-Mora, Shahin S. Ali, and Lyndel W. Meinhardt. 2018. “Moniliophthora Roreri, Causal Agent of Cacao Frosty Pod Rot.” *Molecular Plant Pathology* 19 (7): 1580–94.

Bailey, Bryan A., and Lyndel W. Meinhardt, eds. 2018. *Cacao Diseases*. Cham, Switzerland: Springer International Publishing.

Bailey, Timothy L., Mikael Boden, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, and William S. Noble. 2009. “MEME SUITE: Tools for Motif Discovery and Searching.” *Nucleic Acids Research* 37 (Web Server issue): W202-8.

Balint-Kurti, Peter. 2019. “The Plant Hypersensitive Response: Concepts, Control and Consequences.” *Molecular Plant Pathology* 20 (8): 1163–78.

Bartley, B. G. D. 2005. “The Utilization of the Genetic Resources.” In *The Genetic Diversity of Cacao and Its Utilization*, 309–22. Wallingford: CABI.

Beckers, G. J. M., and S. H. Spoel. 2006. “Fine-Tuning Plant Defence Signalling: Salicylate versus Jasmonate.” *Plant Biology (Stuttgart, Germany)* 8 (1): 1–10.

Bell, E., R. A. Creelman, and J. E. Mullet. 1995. “A Chloroplast Lipoxygenase Is Required for Wound-Induced Jasmonic Acid Accumulation in Arabidopsis.” *Proceedings of the National Academy of Sciences of the United States of America* 92 (19): 8675–79.

Bellis, Emily S., Elizabeth A. Kelly, Claire M. Lorts, Huirong Gao, Victoria L. DeLeo, Germinal Rouhan, Andrew Budden, et al. 2020. “Genomics of Sorghum Local Adaptation to a Parasitic Plant.” *Proceedings of the National Academy of Sciences of the United States of America* 117 (8): 4243–51.

Benedetti, Manuel, Ilaria Verrascina, Daniela Pontiggia, Federica Locci, Benedetta Mattei, Giulia De Lorenzo, and Felice Cervone. 2018. “Four Arabidopsis Berberine Bridge Enzyme-like Proteins Are Specific Oxidases That Inactivate the Elicitor-Active Oligogalacturonides.” *The Plant Journal: For Cell and Molecular Biology* 94 (2): 260–73.

Benfey, Philip N., and Thomas Mitchell-Olds. 2008. “From Genotype to Phenotype: Systems Biology Meets Natural Variation.” *Science (New York, N.Y.)* 320 (5875): 495–97.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society* 57 (1): 289–300.

Bhattarai, Kishor K., Hagop S. Atamian, Isgouhi Kaloshian, and Thomas Eulgem. 2010. “WRKY72-Type Transcription Factors Contribute to Basal Immunity in Tomato and Arabidopsis as Well as Gene-for-Gene Resistance Mediated by the Tomato R Gene Mi-1.” *The Plant Journal: For Cell and Molecular Biology* 63 (2): 229–40.

Bhuiyan, Nazmul H., Gopalan Selvaraj, Yangdou Wei, and John King. 2009. “Gene Expression Profiling and Silencing Reveal That Monolignol Biosynthesis Plays a Critical Role in Penetration Defence in Wheat against Powdery Mildew Invasion.” *Journal of Experimental Botany* 60 (2): 509–21.

Bi, Guozhi, Min Su, Nan Li, Yu Liang, Song Dang, Jiachao Xu, Meijuan Hu, et al. 2021. “The ZAR1 Resistosome Is a Calcium-Permeable Channel Triggering Plant Immune Signaling.” *Cell* 184 (13): 3528-3541.e12.

Białas, Aleksandra, Thorsten Langner, Adeline Harant, Mauricio P. Contreras, Clare Em Stevenson, David M. Lawson, Jan Sklenar, et al. 2021. “Two NLR Immune Receptors Acquired High-Affinity Binding to a Fungal Effector through Convergent Evolution of Their Integrated Domain.” *ELife* 10 (July). https://doi.org/10.7554/eLife.66961.

Biezen, E. A. van der, and J. D. Jones. 1998. “The NB-ARC Domain: A Novel Signalling Motif Shared by Plant Resistance Gene Products and Regulators of Cell Death in Animals.” *Current Biology: CB* 8 (7): R226-7.

Bilbrey, Emma A., Kathryn Williamson, Emmanuel Hatzakis, Diane Doud Miller, Jonathan Fresnedo-Ramírez, and Jessica L. Cooperstone. 2021. “Integrating Genomics and Multiplatform Metabolomics Enables Metabolite Quantitative Trait Loci Detection in Breeding-Relevant Apple Germplasm.” *The New Phytologist* 232 (5): 1944–58.

Boch, Jens, Ulla Bonas, and Thomas Lahaye. 2014. “TAL Effectors--Pathogen Strategies and Plant Resistance Engineering.” *The New Phytologist* 204 (4): 823–32.

Bockus, William W., Jon A. Appel, Robert L. Bowden, Allan K. Fritz, Bikram S. Gill, T. Joe Martin, Rollin G. Sears, Dallas L. Seifers, Gina L. Brown-Guedira, and Merle G. Eversmeyer. 2001. “Success Stories: Breeding for Wheat Disease Resistance in Kansas.” *Plant Disease* 85 (5): 453–61.

Boetzer, Marten, and Walter Pirovano. 2012. “Toward Almost Closed Genomes with GapFiller.” *Genome Biology* 13 (6): R56.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics (Oxford, England)* 30 (15): 2114–20.

Boller, Thomas, and Georg Felix. 2009. “A Renaissance of Elicitors: Perception of Microbe-Associated Molecular Patterns and Danger Signals by Pattern-Recognition Receptors.” *Annual Review of Plant Biology* 60 (1): 379–406.

Boza, Edward J., Juan Carlos Motamayor, Freddy M. Amores, Sergio Cedeño-Amador, Cecile L. Tondo, Donald S. Livingstone, Raymond J. Schnell, and Osman A. Gutiérrez. 2014. “Genetic Characterization of the Cacao Cultivar CCN 51: Its Impact and Significance on Global Cacao Improvement and Production.” *Journal of the American Society for Horticultural Science. American Society for Horticultural Science* 139 (2): 219–29.

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. “Near-Optimal Probabilistic RNA-Seq Quantification.” *Nature Biotechnology* 34 (5): 525–27.

Brown, J. Steven, Wilbert Phillips-Mora, Emilio J. Power, Cheryl Krol, Cuauhtemoc Cervantes-Martinez, Juan Carlos Motamayor, and Raymond J. Schnell. 2007. “Mapping QTLs for Resistance to Frosty Pod and Black Pod Diseases and Horticultural Traits InTheobroma CacaoL.” *Crop Science* 47 (5): 1851–58.

Brown, J. Steven, R. J. Schnell, J. C. Motamayor, Uilson Lopes, David N. Kuhn, and James W. Borrone. 2005. “Resistance Gene Mapping for Witches’ Broom Disease in Theobroma Cacao L. in an F2 Population Using SSR Markers and Candidate Genes.” *Journal of the American Society for Horticultural Science. American Society for Horticultural Science* 130 (3): 366–73.

Browse, John. 2009. “Jasmonate Passes Muster: A Receptor and Targets for the Defense Hormone.” *Annual Review of Plant Biology* 60 (1): 183–205.

Brutus, Alexandre, Francesca Sicilia, Alberto Macone, Felice Cervone, and Giulia De Lorenzo. 2010. “A Domain Swap Approach Reveals a Role of the Plant Wall-Associated Kinase 1 (WAK1) as a Receptor of Oligogalacturonides.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (20): 9452–57.

Cabrera, Odalys García, Eddy Patricia López Molano, Juliana José, Javier Correa Álvarez, and Gonçalo Amarante Guimarães Pereira. 2016. “Ceratocystis Wilt Pathogens: History and Biology—Highlighting C. Cacaofunesta, the Causal Agent of Wilt Disease of Cacao.” In *Cacao Diseases*, 383–428. Cham: Springer International Publishing.

Calderón, Angela I., Brian J. Wright, W. Jeffrey Hurst, and Richard B. van Breemen. 2009. “Screening Antioxidants Using LC-MS: Case Study with Cocoa.” *Journal of Agricultural and Food Chemistry* 57 (13): 5693–99.

Campbell, Michael S., Meiyee Law, Carson Holt, Joshua C. Stein, Gaurav D. Moghe, David E. Hufnagel, Jikai Lei, et al. 2014. “MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations.” *Plant Physiology* 164 (2): 513–24.

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. “TrimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses.” *Bioinformatics (Oxford, England)* 25 (15): 1972–73.

Caplan, Jeffrey L., Padmavathi Mamillapalli, Tessa M. Burch-Smith, Kirk Czymmek, and S. P. Dinesh-Kumar. 2008. “Chloroplastic Protein NRIP1 Mediates Innate Immune Receptor Recognition of a Viral Effector.” *Cell* 132 (3): 449–62.

Carrizo García, Carolina, Michael H. J. Barfuss, Eva M. Sehr, Gloria E. Barboza, Rosabelle Samuel, Eduardo A. Moscone, and Friedrich Ehrendorfer. 2016. “Phylogenetic Relationships, Diversification and Expansion of Chili Peppers (Capsicum, Solanaceae).” *Annals of Botany* 118 (1): 35–51.

Chai, Jinyu, Jian Liu, Jun Zhou, and Da Xing. 2014. “Mitogen-Activated Protein Kinase 6 Regulates NPR1 Gene Expression and Activation during Leaf Senescence Induced by Salicylic Acid.” *Journal of Experimental Botany* 65 (22): 6513–28.

Chakraborty, Mahul, James G. Baldwin-Brown, Anthony D. Long, and J. J. Emerson. 2016. “Contiguous and Accurate de Novo Assembly of Metazoan Genomes with Modest Long Read Coverage.” *Nucleic Acids Research* 44 (19): e147.

Chen, Jun-Yu, Hong-Wei Zhang, Hua-Li Zhang, Jie-Zheng Ying, Liang-Yong Ma, and Jie-Yun Zhuang. 2018. “Natural Variation at QHd1 Affects Heading Date Acceleration at High Temperatures with Pleiotropism for Yield Traits in Rice.” *BMC Plant Biology* 18 (1). https://doi.org/10.1186/s12870-018-1330-5.

Chen, Li-Qing. 2014. “SWEET Sugar Transporters for Phloem Transport and Pathogen Nutrition.” *The New Phytologist* 201 (4): 1150–55.

Chen, Wei, Yanqiang Gao, Weibo Xie, Liang Gong, Kai Lu, Wensheng Wang, Yang Li, et al. 2014. “Genome-Wide Association Analyses Provide Genetic and Biochemical Insights into Natural Variation in Rice Metabolism.” *Nature Genetics* 46 (7): 714–21.

Chen, Ying, Weicai Ye, Yongdong Zhang, and Yuesheng Xu. 2015. “High Speed BLASTN: An Accelerated MegaBLAST Search Tool.” *Nucleic Acids Research* 43 (16): 7762–68.

Cheng, Feng, Jian Wu, Xu Cai, Jianli Liang, Michael Freeling, and Xiaowu Wang. 2018. “Gene Retention, Fractionation and Subgenome Differences in Polyploid Plants.” *Nature Plants* 4 (5): 258–68.

Chezem, William R., Altamash Memon, Fu-Shuang Li, Jing-Ke Weng, and Nicole K. Clay. 2017. “SG2-Type R2R3-MYB Transcription Factor MYB15 Controls Defense-Induced Lignification and Basal Immunity in Arabidopsis.” *The Plant Cell* 29 (8): 1907–26.

Chinchilla, Delphine, Cyril Zipfel, Silke Robatzek, Birgit Kemmerling, Thorsten Nürnberger, Jonathan D. G. Jones, Georg Felix, and Thomas Boller. 2007. “A Flagellin-Induced Complex of the Receptor FLS2 and BAK1 Initiates Plant Defence.” *Nature* 448 (7152): 497–500.

Choi, Hyong Woo, and Daniel F. Klessig. 2016. “DAMPs, MAMPs, and NAMPs in Plant Innate Immunity.” *BMC Plant Biology* 16 (1): 232.

Choudhury, Ananyo, Scott Hazelhurst, Ayton Meintjes, Ovokeraye Achinike-Oduaran, Shaun Aron, Junaid Gamieldien, Mahjoubeh Jalali Sefid Dashti, Nicola Mulder, Nicki Tiffin, and Michèle Ramsay. 2014. “Population-Specific Common SNPs Reflect Demographic Histories and Highlight Regions of Genomic Plasticity with Functional Relevance.” *BMC Genomics* 15 (1): 437.

Christenhusz, Maarten J. M., and James W. Byng. 2016. “The Number of Known Plants Species in the World and Its Annual Increase.” *Phytotaxa* 261 (3): 201–17.

Churngchow, Nunta, and Matinee Rattarasarn. 2001. “Biosynthesis of Scopoletin in Hevea Brasiliensis Leaves Inoculated with Phytophthora Palmivora.” *Journal of Plant Physiology* 158 (7): 875–82.

Clérivet, Alain, Véronique Déon, Ibtissam Alami, Fredérique Lopez, Jean-Paul Geiger, and Michel Nicole. 2000. “Tyloses and Gels Associated with Cellulose Accumulation in Vessels Are Responses of Plane Tree Seedlings (Platanus × Acerifolia) to the Vascular

Fungus Ceratocystis Fimbriata f. Sp Platani." *Trees (Berlin, Germany: West)* 15 (1): 25–31.

Conesa, Ana, Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Manuel Talón, and Montserrat Robles. 2005. "Blast2GO: A Universal Tool for Annotation, Visualization and Analysis in Functional Genomics Research." *Bioinformatics (Oxford, England)* 21 (18): 3674–76.

Cook, Jason P., Michael D. McMullen, James B. Holland, Feng Tian, Peter Bradbury, Jeffrey Ross-Ibarra, Edward S. Buckler, and Sherry A. Flint-Garcia. 2012. "Genetic Architecture of Maize Kernel Composition in the Nested Association Mapping and Inbred Association Panels." *Plant Physiology* 158 (2): 824–34.

Cooper, Susan M., and Norman Owen-Smith. 1986. "Effects of Plant Spinescence on Large Mammalian Herbivores." *Oecologia* 68 (3): 446–55.

Corley, Susan M., Niamh M. Troy, Anthony Bosco, and Marc R. Wilkins. 2019. "QuantSeq. 3' Sequencing Combined with Salmon Provides a Fast, Reliable Approach for High Throughput RNA Expression Analysis." *Scientific Reports* 9 (1): 18895.

Cornejo, Omar E., Muh-Ching Yee, Victor Dominguez, Mary Andrews, Alexandra Sockell, Erika Strandberg, Donald Livingstone 3rd, et al. 2018. "Population Genomic Analyses of the Chocolate Tree, Theobroma Cacao L., Provide Insights into Its Domestication Process." *Communications Biology* 1 (1): 167.

Cuatrecasas, José. 1964. *Cacao and Its Allies: A Taxonomic Revision of the Genus Theobroma*. Smithsonian Institution.

Dangl, J. L., and J. D. Jones. 2001. "Plant Pathogens and Integrated Defence Responses to Infection." *Nature* 411 (6839): 826–33.

Dangl, Jeffery L., and Jonathan D. G. Jones. 2019. "A Pentangular Plant Inflammasome." *Science (New York, N.Y.)* 364 (6435): 31–32.

De Bruyne, Lieselotte, Monica Höfte, and David De Vleesschauwer. 2014. “Connecting Growth and Defense: The Emerging Roles of Brassinosteroids and Gibberellins in Plant Innate Immunity.” *Molecular Plant* 7 (6): 943–59.

De Vos, Ric C. H., Sofia Moco, Arjen Lommen, Joost J. B. Keurentjes, Raoul J. Bino, and Robert D. Hall. 2007. “Untargeted Large-Scale Plant Metabolomics Using Liquid Chromatography Coupled to Mass Spectrometry.” *Nature Protocols* 2 (4): 778–91.

D’Hont, Angélique, France Denoeud, Jean-Marc Aury, Franc-Christophe Baurens, Françoise Carreel, Olivier Garsmeur, Benjamin Noel, et al. 2012. “The Banana (Musa Acuminata) Genome and the Evolution of Monocotyledonous Plants.” *Nature* 488 (7410): 213–17.

Díaz-Tatis, Paula A., Juan C. Ochoa, Edgar M. Rico, Catalina Rodríguez, Adriana Medina, Boris Szurek, Paul Chavarriaga, and Camilo E. López. 2021. “RXam2, a NLR from Cassava (Manihot Esculenta) Contributes Partially to the Quantitative Resistance to Xanthomonas Phaseoli Pv. Manihotis.” *Plant Molecular Biology*, November. https://doi.org/10.1007/s11103-021-01211-2.

Dillinger, T. L., P. Barriga, S. Escárcega, M. Jimenez, D. Salazar Lowe, and L. E. Grivetti. 2000. “Food of the Gods: Cure for Humanity? A Cultural History of the Medicinal and Ritual Use of Chocolate.” *The Journal of Nutrition* 130 (8S Suppl): 2057S-72S.

Ding, Wenyong, Lin Lin, Bing Chen, and Jianwu Dai. 2006. “L1 Elements, Processed Pseudogenes and Retrogenes in Mammalian Genomes.” *IUBMB Life* 58 (12): 677–85.

Divi, Uday K., Tawhidur Rahman, and Priti Krishna. 2010. “Brassinosteroid-Mediated Stress Tolerance in Arabidopsis Shows Interactions with Abscisic Acid, Ethylene and Salicylic Acid Pathways.” *BMC Plant Biology* 10 (1): 151.

Doares, S. H., T. Syrovets, E. W. Weiler, and C. A. Ryan. 1995. “Oligogalacturonides and Chitosan Activate Plant Defensive Genes through the Octadecanoid Pathway.”

Proceedings of the National Academy of Sciences of the United States of America 92 (10): 4095–98.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics (Oxford, England)* 29 (1): 15–21.

Dodds, Peter N., Gregory J. Lawrence, Ann-Maree Catanzariti, Michael A. Ayliffe, and Jeffrey G. Ellis. 2004. "The Melampsora Lini AvrL567 Avirulence Genes Are Expressed in Haustoria and Their Products Are Recognized inside Plant Cells." *The Plant Cell* 16 (3): 755–68.

Dodds, Peter N., Gregory J. Lawrence, Ann-Maree Catanzariti, Trazel Teh, Ching-I A. Wang, Michael A. Ayliffe, Bostjan Kobe, and Jeffrey G. Ellis. 2006. "Direct Protein Interaction Underlies Gene-for-Gene Specificity and Coevolution of the Flax Resistance Genes and Flax Rust Avirulence Genes." *Proceedings of the National Academy of Sciences of the United States of America* 103 (23): 8888–93.

Dodds, Peter N., and John P. Rathjen. 2010. "Plant Immunity: Towards an Integrated View of Plant-Pathogen Interactions." *Nature Reviews. Genetics* 11 (8): 539–48.

Driever, S. M., T. Lawson, P. J. Andralojc, C. A. Raines, and M. A. J. Parry. 2014. "Natural Variation in Photosynthetic Capacity, Growth, and Yield in 64 Field-Grown Wheat Genotypes." *Journal of Experimental Botany* 65 (17): 4959–73.

Dutsadee, Chinnapun, and Churngchow Nunta. 2008. "Induction of Peroxidase, Scopoletin, Phenolic Compounds and Resistance in Hevea Brasiliensis by Elicitin and a Novel Protein Elicitor Purified from Phytophthora Palmivora." *Physiological and Molecular Plant Pathology* 72 (4–6): 179–87.

Eddy, Sean R. 2011. "Accelerated Profile HMM Searches." *PLoS Computational Biology* 7 (10): e1002195.

Edgar, Robert C. 2004. “MUSCLE: A Multiple Sequence Alignment Method with Reduced Time and Space Complexity.” *BMC Bioinformatics* 5 (August): 113.

———. 2010. “Search and Clustering Orders of Magnitude Faster than BLAST.” *Bioinformatics (Oxford, England)* 26 (19): 2460–61.

Efombagn, Ives Bruno M., Juan C. Motamayor, Olivier Sounigo, Albertus B. Eskes, Salomon Nyassé, Christian Cilas, Ray Schnell, Maria J. Manzanares-Dauleux, and Maria Kolesnikova-Allen. 2008. “Genetic Diversity and Structure of Farm and GenBank Accessions of Cacao (Theobroma Cacao L.) in Cameroon Revealed by Microsatellite Markers.” *Tree Genetics & Genomes* 4 (4): 821–31.

Ellinghaus, David, Stefan Kurtz, and Ute Willhoeft. 2008. “LTRharvest, an Efficient and Flexible Software for de Novo Detection of LTR Retrotransposons.” *BMC Bioinformatics* 9 (1): 18.

Emms, David M., and Steven Kelly. 2019. “OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics.” *Genome Biology* 20 (1): 238.

Engelbrecht, C. J. B., T. C. Harrington, A. C. Alfenas, and C. Suarez. 2007. “Genetic Variation in Populations of the Cacao Wilt Pathogen, Ceratocystis Cacaofunesta.” *Plant Pathology* 56 (6): 923–33.

Erwin, Donald C., and Olaf K. Ribeiro. 1996. *Phytophthora Diseases Worldwide*. St Paul: American Phytopathological Society.

Esnault, C., J. Maestre, and T. Heidmann. 2000. “Human LINE Retrotransposons Generate Processed Pseudogenes.” *Nature Genetics* 24 (4): 363–67.

Evans, H. C. 2002. “Invasive Neotropical Pathogens of Tree Crops.” In *Tropical Mycology: Volume 2, Micromycetes*, 83–112. Wallingford: CABI.

Evans, Harry C. 2016a. “Frosty Pod Rot (Moniliophthora Roreri).” In *Cacao Diseases*, 63–96. Cham: Springer International Publishing.

———. 2016b. “Witches’ Broom Disease (Moniliophthora Perniciosa): History and Biology.” In *Cacao Diseases*, 137–77. Cham: Springer International Publishing.

Faleiro, F. G., V. T. Queiroz, U. V. Lopes, C. T. Guimaraes, J. L. Pires, M. M. Yamada, I. S. Araújo, et al. 2006. “Mapeamento Genético Molecular Do Cacaueiro (Theobroma Cacao L.) e QTLs Associados Aresistência Avassoura-de-Bruxa.” *Euphytica/ Netherlands Journal of Plant Breeding* 149: 227–35.

Felix, G., J. D. Duran, S. Volko, and T. Boller. 1999. “Plants Have a Sensitive Perception System for the Most Conserved Domain of Bacterial Flagellin.” *The Plant Journal: For Cell and Molecular Biology* 18 (3): 265–76.

Felsenstein, Joseph. 1993. *PHYLIP (Phylogeny Inference Package), Version 3.5 c*. Joseph Felsenstein.

Ferreira, Sávio Benvindo, Tassiana Barbosa Dantas, Daniele de Figuerêdo Silva, Paula Benvindo Ferreira, Thamara Rodrigues de Melo, and Edeltrudes de Oliveira Lima. 2018. “In Silico and in Vitro Investigation of the Antifungal Activity of Isoeugenol against Penicillium Citrinum.” *Current Topics in Medicinal Chemistry* 18 (25): 2186–96.

Ferry, Robert J. 2020. *The Colonial Elite of Early Caracas*. Berkeley, CA: University of California Press.

Feuillet, C., G. Schachermayr, and B. Keller. 1997. “Molecular Cloning of a New Receptor-like Kinase Gene Encoded at the Lr10 Disease Resistance Locus of Wheat.” *The Plant Journal: For Cell and Molecular Biology* 11 (1): 45–52.

Finet, Cédric, Kailey Slavik, Jian Pu, Sean B. Carroll, and Henry Chung. 2019. “Birth-and-Death Evolution of the Fatty Acyl-CoA Reductase (FAR) Gene Family and Diversification of Cuticular Hydrocarbon Synthesis in Drosophila.” *Genome Biology and Evolution* 11 (6): 1541–51.

Fister, Andrew S., Lena Landherr, Siela N. Maximova, and Mark J. Guiltinan. 2018. “Transient Expression of CRISPR/Cas9 Machinery Targeting TcNPR3 Enhances Defense Response in Theobroma Cacao.” *Frontiers in Plant Science* 9 (March): 268.

Fister, Andrew S., Mariela E. Leandro-Muñoz, Dapeng Zhang, James H. Marden, Peter Tiffin, Claude dePamphilis, Siela Maximova, and Mark J. Guiltinan. 2020. “Widely Distributed Variation in Tolerance to Phytophthora Palmivora in Four Genetic Groups of Cacao.” *Tree Genetics & Genomes* 16 (1). https://doi.org/10.1007/s11295-019-1396-8.

Fister, Andrew S., Luis C. Mejia, Yufan Zhang, Edward Allen Herre, Siela N. Maximova, and Mark J. Guiltinan. 2016. “Theobroma Cacao L. Pathogenesis-Related Gene Tandem Array Members Show Diverse Expression Dynamics in Response to Pathogen Colonization.” *BMC Genomics* 17 (1). https://doi.org/10.1186/s12864-016-2693-3.

Fister, Andrew S., Shawn T. O’Neil, Zi Shi, Yufan Zhang, Brett M. Tyler, Mark J. Guiltinan, and Siela N. Maximova. 2015. “Two Theobroma Cacao Genotypes with Contrasting Pathogen Tolerance Show Aberrant Transcriptional and ROS Responses after Salicylic Acid Treatment.” *Journal of Experimental Botany* 66 (20): 6245–58.

Fister, Andrew S., Zi Shi, Yufan Zhang, Emily E. Helliwell, Siela N. Maximova, and Mark J. Guiltinan. 2016. “Protocol: Transient Expression System for Functional Genomics in the Tropical Tree Theobroma Cacao L.” *Plant Methods* 12 (1): 19.

Fitzgerald, D. J., M. Stratford, M. J. Gasson, J. Ueckert, A. Bos, and A. Narbad. 2004. “Mode of Antimicrobial Action of Vanillin against Escherichia Coli, Lactobacillus Plantarum and Listeria Innocua.” *Journal of Applied Microbiology* 97 (1): 104–13.

Flagel, Lex E., and Jonathan F. Wendel. 2009. “Gene Duplication and Evolutionary Novelty in Plants.” *The New Phytologist* 183 (3): 557–64.

Flor, H. H. 1971. “Current Status of the Gene-for-Gene Concept.” *Annual Review of Phytopathology* 9 (1): 275–96.

Florez, Sergio L., Rachel L. Erwin, Siela N. Maximova, Mark J. Guiltinan, and Wayne R. Curtis. 2015. “Enhanced Somatic Embryogenesis in Theobroma Cacao Using the Homologous BABY BOOM Transcription Factor.” *BMC Plant Biology* 15 (1): 121.

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. 1999. “Preservation of Duplicate Genes by Complementary, Degenerative Mutations.” *Genetics* 151 (4): 1531–45.

Fu, Zheng Qing, and Xinnian Dong. 2013. “Systemic Acquired Resistance: Turning Local Infection into Global Defense.” *Annual Review of Plant Biology* 64 (1): 839–63.

Fuechtbauer, Winnie, Temur Yunusov, Zoltán Bozsóki, Aleksandr Gavrin, Euan K. James, Jens Stougaard, Sebastian Schornack, and Simona Radutoiu. 2018. “LYS12 LysM Receptor Decelerates Phytophthora Palmivora Disease Progression in Lotus Japonicus.” *The Plant Journal: For Cell and Molecular Biology* 93 (2): 297–310.

Fürst, Ursula, Yi Zeng, Markus Albert, Anna Kristina Witte, Judith Fliegmann, and Georg Felix. 2020. “Perception of Agrobacterium Tumefaciens Flagellin by FLS2XL Confers Resistance to Crown Gall Disease.” *Nature Plants* 6 (1): 22–27.

Galloway, L. F., and C. B. Fenster. 2000. “Population Differentiation in an Annual Legume: Local Adaptation.” *Evolution; International Journal of Organic Evolution* 54 (4): 1173–81.

Garcia-Mas, Jordi, Andrej Benjak, Walter Sanseverino, Michael Bourgeois, Gisela Mir, Víctor M. González, Elizabeth Hénaff, et al. 2012. “The Genome of Melon (Cucumis Melo L.).” *Proceedings of the National Academy of Sciences of the United States of America* 109 (29): 11872–77.

Gkizi, Danai, Silke Lehmann, Floriane L’Haridon, Mario Serrano, Epaminondas J. Paplomatas, Jean-Pierre Métraux, and Sotirios E. Tjamos. 2016. “The Innate Immune Signaling System as a Regulator of Disease Resistance and Induced Systemic Resistance Activity

against Verticillium Dahliae." *Molecular Plant-Microbe Interactions: MPMI* 29 (4): 313–23.

Göhre, Vera, Alexandra M. E. Jones, Jan Sklenář, Silke Robatzek, and Andreas P. M. Weber. 2012. "Molecular Crosstalk between PAMP-Triggered Immunity and Photosynthesis." *Molecular Plant-Microbe Interactions: MPMI* 25 (8): 1083–92.

Gómez-Gómez, L., and T. Boller. 2000. "FLS2: An LRR Receptor-like Kinase Involved in the Perception of the Bacterial Elicitor Flagellin in Arabidopsis." *Molecular Cell* 5 (6): 1003–11.

Gonzalez, Antonio, Mingzhe Zhao, John M. Leavitt, and Alan M. Lloyd. 2008. "Regulation of the Anthocyanin Biosynthetic Pathway by the TTG1/BHLH/Myb Transcriptional Complex in Arabidopsis Seedlings." *The Plant Journal: For Cell and Molecular Biology* 53 (5): 814–27.

Goodstein, David M., Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D. Hayes, Joni Fazo, Therese Mitros, et al. 2012. "Phytozome: A Comparative Platform for Green Plant Genomics." *Nucleic Acids Research* 40 (Database issue): D1178-86.

Granett, J., M. A. Walker, L. Kocsis, and A. D. Omer. 2001. "Biology and Management of Grape Phylloxera." *Annual Review of Entomology* 46 (1): 387–412.

Guest, David. 2007. "Black Pod: Diverse Pathogens with a Global Impact on Cocoa Yield." *Phytopathology* 97 (12): 1650–53.

Gumtow, Rebecca, Dongliang Wu, Janice Uchida, and Miaoying Tian. 2018. "A Phytophthora Palmivora Extracellular Cystatin-like Protease Inhibitor Targets Papain to Contribute to Virulence on Papaya." *Molecular Plant-Microbe Interactions: MPMI* 31 (3): 363–73.

Guo, Hongwei, and Joseph R. Ecker. 2004. "The Ethylene Signaling Pathway: New Insights." *Current Opinion in Plant Biology* 7 (1): 40–49.

Guo, Yufang, Monique L. Sakalidis, Gabriel Andres Torres-Londono, and Mary K. Hausbeck. 2021. “Population Structure of a Worldwide Phytophthora Palmivora Collection Suggests Lack of Host Specificity and Reduced Genetic Diversity in South America and the Caribbean.” *Plant Disease*, no. PDIS-05-20-1055-RE (December): PDIS05201055RE.

Gutiérrez, Osman A., Alina S. Puig, Wilbert Phillips-Mora, Bryan A. Bailey, Shahin S. Ali, Keithanne Mockaitis, Raymond J. Schnell, et al. 2021. “SNP Markers Associated with Resistance to Frosty Pod and Black Pod Rot Diseases in an F1 Population of Theobroma Cacao L.” *Tree Genetics & Genomes* 17 (3). https://doi.org/10.1007/s11295-021-01507-w.

Guy, Lionel, Jens Roat Kultima, and Siv G. E. Andersson. 2010. “GenoPlotR: Comparative Gene and Genome Visualization in R.” *Bioinformatics (Oxford, England)* 26 (18): 2334–35.

Haag, Christoph R., Myriam Riek, Jürgen W. Hottinger, V. Ilmari Pajunen, and Dieter Ebert. 2005. “Genetic Diversity and Genetic Differentiation in Daphnia Metapopulations with Subpopulations of Known Age.” *Genetics* 170 (4): 1809–20.

Haas, Brian J., Sophien Kamoun, Michael C. Zody, Rays H. Y. Jiang, Robert E. Handsaker, Liliana M. Cano, Manfred Grabherr, et al. 2009. “Genome Sequence and Analysis of the Irish Potato Famine Pathogen Phytophthora Infestans.” *Nature* 461 (7262): 393–98.

Haas, Brian J., Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D. Blood, Joshua Bowden, Matthew Brian Couger, et al. 2013. “De Novo Transcript Sequence Reconstruction from RNA-Seq Using the Trinity Platform for Reference Generation and Analysis.” *Nature Protocols* 8 (8): 1494–1512.

Hämälä, Tuomas, Mark J. Guiltinan, James H. Marden, Siela N. Maximova, Claude W. dePamphilis, and Peter Tiffin. 2020. “Gene Expression Modularity Reveals Footprints of Polygenic Adaptation in Theobroma Cacao.” *Molecular Biology and Evolution* 37 (1): 110–23.

Hämälä, Tuomas, Eric K. Wafula, Mark J. Guiltinan, Paula E. Ralph, Claude W. dePamphilis, and Peter Tiffin. 2021. “Genomic Structural Variants Constrain and Facilitate Adaptation in Natural Populations of Theobroma Cacao, the Chocolate Tree.” *Proceedings of the National Academy of Sciences of the United States of America* 118 (35): e2102914118.

Han, Yujun, and Susan R. Wessler. 2010. “MITE-Hunter: A Program for Discovering Miniature Inverted-Repeat Transposable Elements from Genomic Sequences.” *Nucleic Acids Research* 38 (22): e199.

Hartl, Daniel L., and Andrew G. Clark. 2007. *Principles of Population Genetics*. 4th ed. New York, NY: Oxford University Press.

Hatsugai, Noriyuki, Daisuke Igarashi, Keisuke Mase, You Lu, Yayoi Tsuda, Suma Chakravarthy, Hai-Lei Wei, et al. 2017. “A Plant Effector-Triggered Immunity Signaling Sector Is Inhibited by Pattern-Triggered Immunity.” *The EMBO Journal* 36 (18): 2758–69.

Heim, M. H., and U. A. Meyer. 1992. “Evolution of a Highly Polymorphic Human Cytochrome P450 Gene Cluster: CYP2D6.” *Genomics* 14 (1): 49–58.

Hockings, Kimberley J., Gen Yamakoshi, and Tetsuro Matsuzawa. 2017. “Dispersal of a Human-Cultivated Crop by Wild Chimpanzees (Pan Troglodytes Verus) in a Forest–Farm Matrix.” *International Journal of Primatology* 38 (2): 172–93.

Holt, Carson, and Mark Yandell. 2011. “MAKER2: An Annotation Pipeline and Genome-Database Management Tool for Second-Generation Genome Projects.” *BMC Bioinformatics* 12 (1): 491.

Honaas, Loren A., Eric K. Wafula, Norman J. Wickett, Joshua P. Der, Yeting Zhang, Patrick P. Edger, Naomi S. Altman, J. Chris Pires, James H. Leebens-Mack, and Claude W. dePamphilis. 2016. “Selecting Superior DE Novo Transcriptome Assemblies: Lessons Learned by Leveraging the Best Plant Genome.” *PloS One* 11 (1): e0146062.

Hosbach, H. A., M. Silberklang, and B. J. McCarthy. 1980. “Evolution of a D. Melanogaster Glutamate TRNA Gene Cluster.” *Cell* 21 (1): 169–78.

Hou, Shuguo, Zunyong Liu, Hexi Shen, and Daoji Wu. 2019. “Damage-Associated Molecular Pattern-Triggered Immunity in Plants.” *Frontiers in Plant Science* 10 (May): 646.

Hsieh, Pinghsun, Brian Hallmark, Joseph Watkins, Tatiana M. Karafet, Ludmila P. Osipova, Ryan N. Gutenkunst, and Michael F. Hammer. 2017. “Exome Sequencing Provides Evidence of Polygenic Adaptation to a Fat-Rich Animal Diet in Indigenous Siberian Populations.” *Molecular Biology and Evolution* 34 (11): 2913–26.

Hua, Lei, Sean R. Stevenson, Ivan Reyna-Llorens, Haiyan Xiong, Stanislav Kopriva, and Julian M. Hibberd. 2021. “The Bundle Sheath of Rice Is Conditioned to Play an Active Role in Water Transport as Well as Sulfur Assimilation and Jasmonic Acid Synthesis.” *The Plant Journal: For Cell and Molecular Biology* 107 (1): 268–86.

Hubert, Nicolas, Fabrice Duponchelle, Jesus Nuñez, Carmen Garcia-Davila, Didier Paugy, and Jean-François Renno. 2007. “Phylogeography of the Piranha Genera Serrasalmus and Pygocentrus: Implications for the Diversification of the Neotropical Ichthyofauna.” *Molecular Ecology* 16 (10): 2115–36.

Hudson, R. R., M. Slatkin, and W. P. Maddison. 1992. “Estimation of Levels of Gene Flow from DNA Sequence Data.” *Genetics* 132 (2): 583–89.

Hyldgaard, Morten, Tina Mygind, Roxana Piotrowska, Morten Foss, and Rikke L. Meyer. 2015. “Isoeugenol Has a Non-Disruptive Detergent-like Mechanism of Action.” *Frontiers in Microbiology* 6 (July): 754.

Jackman, Shaun D., Lauren Coombe, Justin Chu, Rene L. Warren, Benjamin P. Vandervalk, Sarah Yeo, Zhuyi Xue, et al. 2018. “Tigmint: Correcting Assembly Errors Using Linked Reads from Large Molecules.” *BioRxiv*. bioRxiv. https://doi.org/10.1101/304253.

Jacquot, E., L. S. Hagen, P. Michler, O. Rohfritsch, C. Stussi-Garaud, M. Keller, M. Jacquemond, and P. Yot. 1999. “In Situ Localization of Cacao Swollen Shoot Virus in Agroinfected Theobroma Cacao.” *Archives of Virology* 144 (2): 259–71.

Jelesko, J. G., R. Harper, M. Furuya, and W. Gruissem. 1999. “Rare Germinal Unequal Crossing-over Leading to Recombinant Gene Formation and Gene Duplication in Arabidopsis Thaliana.” *Proceedings of the National Academy of Sciences of the United States of America* 96 (18): 10302–7.

Jeong, S. C., A. J. Hayes, R. M. Biyashev, and M. A. Saghai Maroof. 2001. “Diversity and Evolution of a Non-TIR-NBS Sequence Family That Clusters to a Chromosomal ’’hotspot” for Disease Resistance Genes in Soybean.” *Theoretical and Applied Genetics* 103 (2–3): 406–14.

Jha, Priyanka, and Vijay Kumar. 2018. “BABY BOOM (BBM): A Candidate Transcription Factor Gene in Plant Biotechnology.” *Biotechnology Letters* 40 (11–12): 1467–75.

Jiang, Ning, Ann A. Ferguson, R. Keith Slotkin, and Damon Lisch. 2011. “Pack-Mutator-like Transposable Elements (Pack-MULEs) Induce Directional Modification of Genes through Biased Insertion and DNA Acquisition.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (4): 1537–42.

Jiao, Yuannian, Norman J. Wickett, Saravanaraj Ayyampalayam, André S. Chanderbali, Lena Landherr, Paula E. Ralph, Lynn P. Tomsho, et al. 2011. “Ancestral Polyploidy in Seed Plants and Angiosperms.” *Nature* 473 (7345): 97–100.

Johal, G. S., and S. P. Briggs. 1992. “Reductase Activity Encoded by the HM1 Disease Resistance Gene in Maize.” *Science (New York, N.Y.)* 258 (5084): 985–87.

Johnson, Jason A., Thomas C. Harrington, and C. J. B. Engelbrecht. 2005. “Phylogeny and Taxonomy of the North American Clade of the Ceratocystis Fimbriata Complex.” *Mycologia* 97 (5): 1067–92.

Johnson, William Henry. 1912. *Cocoa, Its Cultivation and Preparation*. London,: Murray,.

Jones, Jonathan D. G., and Jeffery L. Dangl. 2006. "The Plant Immune System." *Nature* 444 (7117): 323–29.

Jubic, Lance M., Svenja Saile, Oliver J. Furzer, Farid El Kasmi, and Jeffery L. Dangl. 2019. "Help Wanted: Helper NLRs and Plant Immune Responses." *Current Opinion in Plant Biology* 50 (August): 82–94.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature* 596 (7873): 583–89.

Jupe, Florian, Leighton Pritchard, Graham J. Etherington, Katrin Mackenzie, Peter J. A. Cock, Frank Wright, Sanjeev Kumar Sharma, et al. 2012. "Identification and Localisation of the NB-LRR Gene Family within the Potato Genome." *BMC Genomics* 13 (1): 75.

Jupe, Florian, Kamil Witek, Walter Verweij, Jadwiga Sliwka, Leighton Pritchard, Graham J. Etherington, Dan Maclean, 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." *The Plant Journal: For Cell and Molecular Biology* 76 (3): 530–44.

Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. 2005. "Repbase Update, a Database of Eukaryotic Repetitive Elements." *Cytogenetic and Genome Research* 110 (1–4): 462–67.

Kajitani, Rei, Kouta Toshimoto, Hideki Noguchi, Atsushi Toyoda, Yoshitoshi Ogura, Miki Okuno, Mitsuru Yabana, et al. 2014. "Efficient de Novo Assembly of Highly Heterozygous Genomes from Whole-Genome Shotgun Short Reads." *Genome Research* 24 (8): 1384–95.

Kapitonov, V. V., and J. Jurka. 2001. “Rolling-Circle Transposons in Eukaryotes.” *Proceedings of the National Academy of Sciences of the United States of America* 98 (15): 8714–19.

Kapitonov, Vladimir V., and Jerzy Jurka. 2006. “Self-Synthesizing DNA Transposons in Eukaryotes.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (12): 4540–45.

Katoh, Kazutaka, Kei-Ichi Kuma, Hiroyuki Toh, and Takashi Miyata. 2005. “MAFFT Version 5: Improvement in Accuracy of Multiple Sequence Alignment.” *Nucleic Acids Research* 33 (2): 511–18.

Khan, Fazlurrahman, Nilushi Indika Bamunuarachchi, Nazia Tabassum, and Young-Mog Kim. 2021. “Caffeic Acid and Its Derivatives: Antimicrobial Drugs toward Microbial Pathogens.” *Journal of Agricultural and Food Chemistry* 69 (10): 2979–3004.

Kim, Myung-Shin, Geun Young Chae, Soohyun Oh, Jihyun Kim, Hyunggon Mang, Seungill Kim, and Doil Choi. 2021. “Comparative Analysis of de Novo Genomes Reveals Dynamic Intra-Species Divergence of NLRs in Pepper.” *BMC Plant Biology* 21 (1): 247.

Kim, Seungill, Jieun Park, Seon-In Yeom, Yong-Min Kim, Eunyoung Seo, Ki-Tae Kim, Myung-Shin Kim, et al. 2017. “New Reference Genome Sequences of Hot Pepper Reveal the Massive Evolution of Plant Disease-Resistance Genes by Retroduplication.” *Genome Biology* 18 (1): 210.

Kimber, C. 2006. *Martinique Revisited*. College Station, TX: Texas A & M University Press.

Klein, J. 1987. “Origin of Major Histocompatibility Complex Polymorphism: The Trans-Species Hypothesis.” *Human Immunology* 19 (3): 155–62.

Klessig, D. F., J. Durner, R. Noad, D. A. Navarre, D. Wendehenne, D. Kumar, J. M. Zhou, et al. 2000. “Nitric Oxide and Salicylic Acid Signaling in Plant Defense.” *Proceedings of the National Academy of Sciences of the United States of America* 97 (16): 8849–55.

Knollenberg, Benjamin J., Guo-Xing Li, Joshua D. Lambert, Siela N. Maximova, and Mark J. Guiltinan. 2020. “Clovamide, a Hydroxycinnamic Acid Amide, Is a Resistance Factor Against Phytophthora Spp. in Theobroma Cacao.” *Frontiers in Plant Science* 11 (December): 617520.

Kobayashi, Michie, Miki Yoshioka, Shuta Asai, Hironari Nomura, Kazuo Kuchimura, Hitoshi Mori, Noriyuki Doke, and Hirofumi Yoshioka. 2012. “StCDPK5 Confers Resistance to Late Blight Pathogen but Increases Susceptibility to Early Blight Pathogen in Potato via Reactive Oxygen Species Burst.” *The New Phytologist* 196 (1): 223–37.

Koenig, Daniel, Jörg Hagmann, Rachel Li, Felix Bemm, Tanja Slotte, Barbara Neuffer, Stephen I. Wright, and Detlef Weigel. 2019. “Long-Term Balancing Selection Drives Evolution of Immunity Genes in Capsella.” *ELife* 8 (February). https://doi.org/10.7554/eLife.43606.

Korf, Ian. 2004. “Gene Finding in Novel Genomes.” *BMC Bioinformatics* 5 (May): 59.

Kourelis, Jiorgos, and Renier A. L. van der Hoorn. 2018. “Defended to the Nines: 25 Years of Resistance Gene Cloning Identifies Nine Mechanisms for R Protein Function.” *The Plant Cell* 30 (2): 285–99.

Kramerov, D. A., and N. S. Vassetzky. 2011. “Origin and Evolution of SINEs in Eukaryotic Genomes.” *Heredity* 107 (6): 487–95.

Krasileva, Ksenia V. 2019. “The Role of Transposable Elements and DNA Damage Repair Mechanisms in Gene Duplications and Gene Fusions in Plant Genomes.” *Current Opinion in Plant Biology* 48 (April): 18–25.

Kremling, Karl A. G., Shu-Yun Chen, Mei-Hsiu Su, Nicholas K. Lepak, M. Cinta Romay, Kelly L. Swarts, Fei Lu, Anne Lorant, Peter J. Bradbury, and Edward S. Buckler. 2018. “Dysregulation of Expression Correlates with Rare-Allele Burden and Fitness Loss in Maize.” *Nature* 555 (7697): 520–23.

Kuhn, Robert M., David Haussler, and W. James Kent. 2013. “The UCSC Genome Browser and Associated Tools.” *Briefings in Bioinformatics* 14 (2): 144–61.

Lai, Jinsheng, Yubin Li, Joachim Messing, and Hugo K. Dooner. 2005. “Gene Movement by Helitron Transposons Contributes to the Haplotype Variability of Maize.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (25): 9068–73.

Lanaud, C., O. Fouet, D. Clément, M. Boccara, A. M. Risterucci, S. Surujdeo-Maharaj, T. Legavre, and X. Argout. 2009. “A Meta–QTL Analysis of Disease Resistance Traits of Theobroma Cacao L.” *Molecular Breeding: New Strategies in Plant Improvement* 24 (4): 361–74.

Lapin, Dmitry, Viera Kovacova, Xinhua Sun, Joram A. Dongus, Deepak Bhandari, Patrick von Born, Jaqueline Bautor, et al. 2019. “A Coevolved EDS1-SAG101-NRG1 Module Mediates Cell Death Signaling by TIR-Domain Immune Receptors.” *The Plant Cell* 31 (10): 2430–55.

Law, Charity W., Yunshun Chen, Wei Shi, and Gordon K. Smyth. 2014. “Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts.” *Genome Biology* 15 (2): R29.

Lazo, G. R., P. A. Stein, and R. A. Ludwig. 1991. “A DNA Transformation-Competent Arabidopsis Genomic Library in Agrobacterium.” *Bio/Technology* 9 (10): 963–67.

Le Roux, Clémentine, Gaëlle Huet, Alain Jauneau, Laurent Camborde, Dominique Trémousaygue, Alexandra Kraut, Binbin Zhou, et al. 2015. “A Receptor Pair with an Integrated Decoy Converts Pathogen Disabling of Transcription Factors to Immunity.” *Cell* 161 (5): 1074–88.

Lebedenko, E. N., K. R. Birikh, O. V. Plutalov, and Berlin YuA. 1991. “Method of Artificial DNA Splicing by Directed Ligation (SDL).” *Nucleic Acids Research* 19 (24): 6757–61.

Lee, Tae-Ho, Hui Guo, Xiyin Wang, Changsoo Kim, and Andrew H. Paterson. 2014. “SNPhylo: A Pipeline to Construct a Phylogenetic Tree from Huge SNP Data.” *BMC Genomics* 15 (1): 162.

Leister, Dario. 2004. “Tandem and Segmental Gene Duplication and Recombination in the Evolution of Plant Disease Resistance Gene.” *Trends in Genetics: TIG* 20 (3): 116–22.

Levinson, G., and G. A. Gutman. 1987. “Slipped-Strand Mispairing: A Major Mechanism for DNA Sequence Evolution.” *Molecular Biology and Evolution* 4 (3): 203–21.

Lewis, Jennifer D., Amy Huei-Yi Lee, Jana A. Hassan, Janet Wan, Brenden Hurley, Jacquelyn R. Jhingree, Pauline W. Wang, et al. 2013. “The Arabidopsis ZED1 Pseudokinase Is Required for ZAR1-Mediated Immunity Induced by the Pseudomonas Syringae Type III Effector HopZ1a.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (46): 18722–27.

Li, Fuguang, Guangyi Fan, Cairui Lu, Guanghui Xiao, Changsong Zou, Russell J. Kohel, Zhiying Ma, et al. 2015. “Genome Sequence of Cultivated Upland Cotton (Gossypium Hirsutum TM-1) Provides Insights into Genome Evolution.” *Nature Biotechnology* 33 (5): 524–30.

Li, Heng. 2011. “A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data.” *Bioinformatics (Oxford, England)* 27 (21): 2987–93.

———. 2013. “Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM.” *ArXiv [q-Bio.GN]*. arXiv. http://arxiv.org/abs/1303.3997.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics (Oxford, England)* 25 (16): 2078–79.

Li, Long, Zhi Peng, Xinguo Mao, Jingyi Wang, Chaonan Li, Xiaoping Chang, and Ruilian Jing. 2021. “Genetic Insights into Natural Variation Underlying Salt Tolerance in Wheat.” *Journal of Experimental Botany* 72 (4): 1135–50.

Li, Yiyuan, Jianhui Xiao, Jiajie Wu, Jialei Duan, Yue Liu, Xingguo Ye, Xin Zhang, et al. 2012. “A Tandem Segmental Duplication (TSD) in Green Revolution Gene Rht-D1b Region Underlies Plant Height Variation.” *The New Phytologist* 196 (1): 282–91.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. “FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics (Oxford, England)* 30 (7): 923–30.

Lin, Xiao, Yu Zhang, Hanhui Kuang, and Jiongjiong Chen. 2013. “Frequent Loss of Lineages and Deficient Duplications Accounted for Low Copy Number of Disease Resistance Genes in Cucurbitaceae.” *BMC Genomics* 14 (1): 335.

Liu, Lijing, Fathi-Mohamed Sonbol, Bethany Huot, Yangnan Gu, John Withers, Musoki Mwimba, Jian Yao, Sheng Yang He, and Xinnian Dong. 2016. “Salicylic Acid Receptors Activate Jasmonic Acid Signalling through a Non-Canonical Pathway to Promote Effector-Triggered Immunity.” *Nature Communications* 7 (October): 13099.

Liu, Yi, Zi Shi, Siela Maximova, Mark J. Payne, and Mark J. Guiltinan. 2013. “Proanthocyanidin Synthesis in Theobroma Cacao: Genes Encoding Anthocyanidin Synthase, Anthocyanidin Reductase, and Leucoanthocyanidin Reductase.” *BMC Plant Biology* 13 (1): 202.

Livingstone, Donald, 3rd, Conrad Stack, Guiliana M. Mustiga, Dayana C. Rodezno, Carmen Suarez, Freddy Amores, Frank A. Feltus, Keithanne Mockaitis, Omar E. Cornejo, and Juan C. Motamayor. 2017. “A Larger Chocolate Chip-Development of a 15K Theobroma Cacao L. Snp Array to Create High-Density Linkage Maps.” *Frontiers in Plant Science* 8 (December): 2008.

Locci, Federica, Manuel Benedetti, Daniela Pontiggia, Matteo Citterico, Claudio Caprari, Benedetta Mattei, Felice Cervone, and Giulia De Lorenzo. 2019. “An Arabidopsis Berberine Bridge Enzyme-like Protein Specifically Oxidizes Cellulose Oligomers and Plays a Role in Immunity.” *The Plant Journal: For Cell and Molecular Biology* 98 (3): 540–54.

Lolle, Signe, Christiaan Greeff, Klaus Petersen, Milena Roux, Michael Krogh Jensen, Simon Bressendorff, Eleazar Rodriguez, Kenneth Sømark, John Mundy, and Morten Petersen. 2017. “Matching NLR Immune Receptors to Autoimmunity in Camta3 Mutants Using Antimorphic NLR Alleles.” *Cell Host & Microbe* 21 (4): 518-529.e4.

Loon, L. C. van, M. Rep, and C. M. J. Pieterse. 2006. “Significance of Inducible Defense-Related Proteins in Infected Plants.” *Annual Review of Phytopathology* 44 (1): 135–62.

Lorin, Thibault, Frédéric G. Brunet, Vincent Laudet, and Jean-Nicolas Volff. 2018. “Teleost Fish-Specific Preferential Retention of Pigmentation Gene-Containing Families after Whole Genome Duplications in Vertebrates.” *G3 (Bethesda, Md.)* 8 (5): 1795–1806.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550.

Lu, You, and Kenichi Tsuda. 2021. “Intimate Association of PRR- and NLR-Mediated Signaling in Plant Immunity.” *Molecular Plant-Microbe Interactions: MPMI* 34 (1): 3–14.

Lüdecke, Daniel, Mattan Ben-Shachar, Indrajeet Patil, Philip Waggoner, and Dominique Makowski. 2021. “Performance: An R Package for Assessment, Comparison and Testing of Statistical Models.” *Journal of Open Source Software* 6 (60): 3139.

Luna, Estrella, Toby J. A. Bruce, Michael R. Roberts, Victor Flors, and Jurriaan Ton. 2012. “Next-Generation Systemic Acquired Resistance.” *Plant Physiology* 158 (2): 844–53.

Ma, Lisong, Ewa Lukasik, Fleur Gawehns, and Frank L. W. Takken. 2012. “The Use of Agroinfiltration for Transient Expression of Plant Resistance and Fungal Effector

Proteins in Nicotiana Benthamiana Leaves." *Methods in Molecular Biology (Clifton, N.J.)* 835: 61–74.

Ma, Wenbo, Frederick F. T. Dong, John Stavrinides, and David S. Guttman. 2006. "Type III Effector Diversification via Both Pathoadaptation and Horizontal Transfer in Response to a Coevolutionary Arms Race." *PLoS Genetics* 2 (12): e209.

Mangelsdorf, Paul C. 1986. "The Origin of Corn." *Scientific American* 255 (2): 80–86.

Manjang, Kalifa, Shailesh Tripathi, Olli Yli-Harja, Matthias Dehmer, and Frank Emmert-Streib. 2020. "Graph-Based Exploitation of Gene Ontology Using GOxploreR for Scrutinizing Biological Significance." *Scientific Reports* 10 (1): 16672.

Marden, J. H., S. A. Mangan, M. P. Peterson, E. Wafula, H. W. Fescemyer, J. P. Der, C. W. dePamphilis, and L. S. Comita. 2017. "Ecological Genomics of Tropical Trees: How Local Population Size and Allelic Diversity of Resistance Genes Relate to Immune Responses, Cosusceptibility to Pathogens, and Negative Density Dependence." *Molecular Ecology* 26 (9): 2498–2513.

Martinson, Veronica A. 1966. "Hybridization of Cacao and Theobroma Grandiflora." *The Journal of Heredity* 57 (4): 134–36.

Maximova, S., C. Miller, G. Antúnez de Mayolo, S. Pishak, A. Young, and M. J. Guiltinan. 2003. "Stable Transformation of Theobroma Cacao L. and Influence of Matrix Attachment Regions on GFP Expression." *Plant Cell Reports* 21 (9): 872–83.

Maximova, S. N., A. Young, S. Pishak, C. Miller, A. Traore, and M. J. Guiltinan. 2005. "Integrated System for Propagation of Theobroma Cacao L." In *Protocol for Somatic Embryogenesis in Woody Plants*, 209–27. Berlin/Heidelberg: Springer-Verlag.

Maximova, Siela N., Jean-Philippe Marelli, Ann Young, Sharon Pishak, Joseph A. Verica, and Mark J. Guiltinan. 2006. "Over-Expression of a Cacao Class I Chitinase Gene in

Theobroma Cacao L. Enhances Resistance against the Pathogen, Colletotrichum Gloeosporioides." *Planta* 224 (4): 740–49.

Mazón, Marina, Francisco Díaz, and Juan C. Gaviria. 2013. "Effectiveness of Different Trap Types for Control of Bark and Ambrosia Beetles (Scolytinae) in Criollo Cacao Farms of Mérida, Venezuela." *International Journal of Pest Management* 59 (3): 189–96.

Mchau, Godwin R. A., and Michael D. Coffey. 1994. "Isozyme Diversity in Phytophthora Palmivora: Evidence for a Southeast Asian Centre of Origin." *Mycological Research* 98 (9): 1035–43.

Meinhardt, Lyndel W., Johana Rincones, Bryan A. Bailey, M. Catherine Aime, Gareth W. Griffith, Dapeng Zhang, and Gonçalo A. G. Pereira. 2008. "Moniliophthora Perniciosa, the Causal Agent of Witches' Broom Disease of Cacao: What's New from This Old Foe?" *Molecular Plant Pathology* 9 (5): 577–88.

Melnick, Rachel L., Jean-Philippe Marelli, Richard C. Sicher, Mary D. Strem, and Bryan A. Bailey. 2012. "The Interaction of Theobroma Cacao and Moniliophthora Perniciosa, the Causal Agent of Witches' Broom Disease, during Parthenocarpy." *Tree Genetics & Genomes* 8 (6): 1261–79.

Melotto, Maeli, William Underwood, Jessica Koczan, Kinya Nomura, and Sheng Yang He. 2006. "Plant Stomata Function in Innate Immunity against Bacterial Invasion." *Cell* 126 (5): 969–80.

Melotto, Maeli, Li Zhang, Paula R. Oblessuc, and Sheng Yang He. 2017. "Stomatal Defense a Decade Later." *Plant Physiology* 174 (2): 561–71.

Menden, Barbara, Markus Kohlhoff, and Bruno M. Moerschbacher. 2007. "Wheat Cells Accumulate a Syringyl-Rich Lignin during the Hypersensitive Resistance Response." *Phytochemistry* 68 (4): 513–20.

Meyer, Rachel S., and Michael D. Purugganan. 2013. “Evolution of Crop Species: Genetics of Domestication and Diversification.” *Nature Reviews. Genetics* 14 (12): 840–52.

Meyers, B. C., D. B. Chin, K. A. Shen, S. Sivaramakrishnan, D. O. Lavelle, Z. Zhang, and R. W. Michelmore. 1998. “The Major Resistance Gene Cluster in Lettuce Is Highly Duplicated and Spans Several Megabases.” *The Plant Cell* 10 (11): 1817–32.

Meyers, Blake C., Alexander Kozik, Alyssa Griego, Hanhui Kuang, and Richard W. Michelmore. 2003. “Genome-Wide Analysis of NBS-LRR-Encoding Genes in Arabidopsis.” *The Plant Cell* 15 (4): 809–34.

Michael, Todd P. 2014. “Plant Genome Size Variation: Bloating and Purging DNA.” *Briefings in Functional Genomics* 13 (4): 308–17.

Michael, Todd P., and Robert VanBuren. 2020. “Building Near-Complete Plant Genomes.” *Current Opinion in Plant Biology* 54 (April): 26–33.

Michelmore, R. W., and B. C. Meyers. 1998. “Clusters of Resistance Genes in Plants Evolve by Divergent Selection and a Birth-and-Death Process.” *Genome Research* 8 (11): 1113–30.

Mighell, A. J., N. R. Smith, P. A. Robinson, and A. F. Markham. 2000. “Vertebrate Pseudogenes.” *FEBS Letters* 468 (2–3): 109–14.

Mirarab, S., R. Reaz, Md S. Bayzid, T. Zimmermann, M. S. Swenson, and T. Warnow. 2014. “ASTRAL: Genome-Scale Coalescent-Based Species Tree Estimation.” *Bioinformatics (Oxford, England)* 30 (17): i541-8.

Mistry, Jaina, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A. Salazar, Erik L. L. Sonnhammer, Silvio C. E. Tosatto, et al. 2021. “Pfam: The Protein Families Database in 2021.” *Nucleic Acids Research* 49 (D1): D412–19.

Mitsuhara, I., M. Ugaki, H. Hirochika, M. Ohshima, T. Murakami, Y. Gotoh, Y. Katayose, et al. 1996. “Efficient Promoter Cassettes for Enhanced Expression of Foreign Genes in Dicotyledonous and Monocotyledonous Plants.” *Plant & Cell Physiology* 37 (1): 49–59.

Mizuno, Hiroshi, Satoshi Katagiri, Hiroyuki Kanamori, Yoshiyuki Mukai, Takuji Sasaki, Takashi Matsumoto, and Jianzhong Wu. 2020. “Evolutionary Dynamics and Impacts of Chromosome Regions Carrying R-Gene Clusters in Rice.” *Scientific Reports* 10 (1): 872.

Mondego, Jorge M. C., Marcelo F. Carazzolle, Gustavo G. L. Costa, Eduardo F. Formighieri, Lucas P. Parizzi, Johana Rincones, Carolina Cotomacci, et al. 2008. “A Genome Survey of Moniliophthora Perniciosa Gives New Insights into Witches’ Broom Disease of Cacao.” *BMC Genomics* 9 (1): 548.

Morales-Cruz, Abraham, Shahin S. Ali, Andrea Minio, Rosa Figueroa-Balderas, Jadran F. García, Takao Kasuga, Alina S. Puig, Jean-Philippe Marelli, Bryan A. Bailey, and Dario Cantu. 2020. “Independent Whole-Genome Duplications Define the Architecture of the Genomes of the Devastating West African Cacao Black Pod Pathogen Phytophthora Megakarya and Its Close Relative Phytophthora Palmivora.” *G3 (Bethesda, Md.)* 10 (7): 2241–55.

Motamayor, J. C., A. M. Risterucci, P. A. Lopez, C. F. Ortiz, A. Moreno, and C. Lanaud. 2002. “Cacao Domestication I: The Origin of the Cacao Cultivated by the Mayas.” *Heredity* 89 (5): 380–86.

Motamayor, Juan C., Philippe Lachenaud, Jay Wallace da Silva e Mota, Rey Loor, David N. Kuhn, J. Steven Brown, and Raymond J. Schnell. 2008. “Geographic and Genetic Population Differentiation of the Amazonian Chocolate Tree (Theobroma Cacao L).” *PloS One* 3 (10): e3311.

Motamayor, Juan C., Keithanne Mockaitis, Jeremy Schmutz, Niina Haiminen, Donald Livingstone 3rd, Omar Cornejo, Seth D. Findley, et al. 2013. “The Genome Sequence of the Most Widely Cultivated Cacao Type and Its Use to Identify Candidate Genes Regulating Pod Color.” *Genome Biology* 14 (6): r53.

Mukherjee, Sohini, Gladys Keitany, Yan Li, Yong Wang, Haydn L. Ball, Elizabeth J. Goldsmith, and Kim Orth. 2006. “Yersinia YopJ Acetylates and Inhibits Kinase Activation by Blocking Phosphorylation.” *Science (New York, N.Y.)* 312 (5777): 1211–14.

Mukhtar, M. Shahid, Laurent Deslandes, Marie-Christine Auriac, Yves Marco, and Imre E. Somssich. 2008. “The Arabidopsis Transcription Factor WRKY27 Influences Wilt Disease Symptom Development Caused by Ralstonia Solanacearum.” *The Plant Journal: For Cell and Molecular Biology* 56 (6): 935–47.

Muller, Emmanuelle. 2016. “Cacao Swollen Shoot Virus (CSSV): History, Biology, and Genome.” In *Cacao Diseases*, 337–58. Cham: Springer International Publishing.

Muroi, Atsushi, Atsushi Ishihara, Chihiro Tanaka, Akihiro Ishizuka, Junji Takabayashi, Hideto Miyoshi, and Takaaki Nishioka. 2009. “Accumulation of Hydroxycinnamic Acid Amides Induced by Pathogen Infection and Identification of Agmatine Coumaroyltransferase in Arabidopsis Thaliana.” *Planta* 230 (3): 517–27.

Návarová, Hana, Friederike Bernsdorff, Anne-Christin Döring, and Jürgen Zeier. 2012. “Pipecolic Acid, an Endogenous Mediator of Defense Amplification and Priming, Is a Critical Regulator of Inducible Plant Immunity.” *The Plant Cell* 24 (12): 5123–41.

Nei, M., X. Gu, and T. Sitnikova. 1997. “Evolution by the Birth-and-Death Process in Multigene Families of the Vertebrate Immune System.” *Proceedings of the National Academy of Sciences of the United States of America* 94 (15): 7799–7806.

Nei, M., I. B. Rogozin, and H. Piontkivska. 2000. “Purifying Selection and Birth-and-Death Evolution in the Ubiquitin Gene Family.” *Proceedings of the National Academy of Sciences of the United States of America* 97 (20): 10866–71.

Neves Dos Santos, Fábio, Dilze Maria Argôlo Magalhães, Edna Dora Martins Newman Luz, Marcos Nogueira Eberlin, and Ana Valéria Colnaghi Simionato. 2021. “Metabolite Mass

Spectrometry Profiling of Cacao Genotypes Reveals Contrasting Resistances to Ceratocystis Cacaofunesta Phytopathogen." *Electrophoresis* 42 (23): 2519–27.

Nevo, Eviatar, and Guoxiong Chen. 2010. "Drought and Salt Tolerances in Wild Relatives for Wheat and Barley Improvement." *Plant, Cell & Environment* 33 (4): 670–85.

Ngou, Bruno Pok Man, Hee-Kyung Ahn, Pingtao Ding, and Jonathan D. G. Jones. 2021. "Mutual Potentiation of Plant Immunity by Cell-Surface and Intracellular Receptors." *Nature* 592 (7852): 110–15.

Ngou, Bruno Pok Man, Jonathan D. G. Jones, and Pingtao Ding. 2021. "Plant Immune Networks." *Trends in Plant Science*, September. https://doi.org/10.1016/j.tplants.2021.08.012.

Novikova, Polina Yu, Nora Hohmann, Viktoria Nizhynska, Takashi Tsuchimatsu, Jamshaid Ali, Graham Muir, Alessia Guggisberg, et al. 2016. "Sequencing of the Genus Arabidopsis Identifies a Complex History of Nonbifurcating Speciation and Abundant Trans-Specific Polymorphism." *Nature Genetics* 48 (9): 1077–82.

Ota, T., and M. Nei. 1994. "Divergent Evolution and Evolution by the Birth-and-Death Process in the Immunoglobulin VH Gene Family." *Molecular Biology and Evolution* 11 (3): 469–82.

Pare, P. W., and J. H. Tumlinson. 1999. "Plant Volatiles as a Defense against Insect Herbivores." *Plant Physiology* 121 (2): 325–32.

Park, Tae-Ho, Jack Gros, Anne Sikkema, Vivianne G. A. A. Vleeshouwers, Marielle Muskens, Sjefke Allefs, Evert Jacobsen, Richard G. F. Visser, and Edwin A. G. van der Vossen. 2005. "The Late Blight Resistance Locus Rpi-Blb3 from Solanum Bulbocastanum Belongs to a Major Late Blight R Gene Cluster on Chromosome 4 of Potato." *Molecular Plant-Microbe Interactions: MPMI* 18 (7): 722–29.

Pellicer, Jaume, and Ilia J. Leitch. 2020. “The Plant DNA C-Values Database (Release 7.1): An Updated Online Repository of Plant Genome Size Data for Comparative Studies.” *The New Phytologist* 226 (2): 301–5.

Pemberton, Clare L., and George P. C. Salmond. 2004. “The Nep1-like Proteins-a Growing Family of Microbial Elicitors of Plant Necrosis.” *Molecular Plant Pathology* 5 (4): 353–59.

Phillips-Mora, W., M. C. Aime, and M. J. Wilkinson. 2007. “Biodiversity and Biogeography of the Cacao (Theobroma Cacao) Pathogen Moniliophthora Roreri in Tropical America.” *Plant Pathology* 56 (6): 911–22.

Phillips-Mora, W., J. Castillo, U. Krauss, E. Rodriguez, and M. J. Wilkinson. 2005. “Evaluation of Cacao (Theobroma Cacao) Clones against Seven Colombian Isolates of Moniliophthora Roreri from Four Pathogen Genetic Groups.” *Plant Pathology* 54 (4): 483–90.

Phillips-Mora, W., and M. J. Wilkinson. 2007. “Frosty Pod of Cacao: A Disease with a Limited Geographic Range but Unlimited Potential for Damage.” *Phytopathology* 97 (12): 1644–47.

Ploetz, Randy C. 2007. “Cacao Diseases: Important Threats to Chocolate Production Worldwide.” *Phytopathology* 97 (12): 1634–39.

Pokou, Désiré N., Andrew S. Fister, Noah Winters, Mathias Tahi, Coulibaly Klotioloma, Aswathy Sebastian, James H. Marden, Siela N. Maximova, and Mark J. Guiltinan. 2019. “Resistant and Susceptible Cacao Genotypes Exhibit Defense Gene Polymorphism and Unique Early Responses to Phytophthora Megakarya Inoculation.” *Plant Molecular Biology* 99 (4–5): 499–516.

Pokou, N. D., J. A. K. N’Goran, Ph Lachenaud, A. B. Eskes, J. C. Montamayor, R. Schnell, M. Kolesnikova-Allen, D. Clément, and A. Sangaré. 2009. “Recurrent Selection of Cocoa

Populations in Côte d'Ivoire: Comparative Genetic Diversity between the First and Second Cycles." *Plant Breeding = Zeitschrift Fur Pflanzenzuchtung* 128 (5): 514–20.

Pond, Sergei L. Kosakovsky, Simon D. W. Frost, and Spencer V. Muse. 2005. "HyPhy: Hypothesis Testing Using Phylogenies." *Bioinformatics (Oxford, England)* 21 (5): 676–79.

Posnette, A. F., and N. F. Robertson. 1950. "Virus Diseases of Cacao in West Africa." *The Annals of Applied Biology* 37 (3): 363–77.

Posnette, A. F., N. F. Robertson, and J. Mca Todd. 1950. "Virus Diseases of Cacao in West Africa." *The Annals of Applied Biology* 37 (2): 229–40.

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2010. "FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments." *PloS One* 5 (3): e9490.

Prigozhin, Daniil M., and Ksenia V. Krasileva. 2021. "Analysis of Intraspecies Diversity Reveals a Subset of Highly Variable Plant Immune Receptors and Predicts Their Binding Sites." *The Plant Cell* 33 (4): 998–1015.

Proost, Sebastian, Pedro Pattyn, Tom Gerats, and Yves Van de Peer. 2011. "Journey through the Past: 150 Million Years of Plant Genome Evolution." *The Plant Journal: For Cell and Molecular Biology* 66 (1): 58–65.

Quevillon, E., V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez. 2005. "InterProScan: Protein Domains Identifier." *Nucleic Acids Research* 33 (Web Server issue): W116-20.

Richardson, James E., Barbara A. Whitlock, Alan W. Meerow, and Santiago Madriñán. 2015. "The Age of Chocolate: A Diversification History of Theobroma and Malvaceae." *Frontiers in Ecology and Evolution* 3 (November). https://doi.org/10.3389/fevo.2015.00120.

Risterucci, A. M., D. Paulin, M. Ducamp, J. A. K. N'Goran, and C. Lanaud. 2003. "Identification of QTLs Related to Cocoa Resistance to Three Species of Phytophthora." *Theoretical and Applied Genetics* 108 (1): 168–74.

Rocha, Hermínio M. 1966. "La Importancia de Las Sustancias Polifenólicas En El Mecanismo Fisiológico de La Resistencia de Cacao (Theobroma Cacao L.) a Phytophthora Palmivora (Butl.) Butl." IICA, Turrialba (Costa Rica).

Rodrigues Oblessuc, Paula, Mariana Vaz Bisneta, and Maeli Melotto. 2019. "Common and Unique Arabidopsis Proteins Involved in Stomatal Susceptibility to Salmonella Enterica and Pseudomonas Syringae." *FEMS Microbiology Letters* 366 (16). https://doi.org/10.1093/femsle/fnz197.

Roivainen, Osmo. 1976. "Transmission of Cocoa Viruses by Mealybugs (Homoptera: Pseudococcidae)." *Agricultural and Food Science* 48 (3): 203–304.

Romero Navarro, J. Alberto, Wilbert Phillips-Mora, Adriana Arciniegas-Leal, Allan Mata-Quirós, Niina Haiminen, Guiliana Mustiga, Donald Livingstone Iii, et al. 2017. "Application of Genome Wide Association and Genomic Prediction for Improvement of Cacao Productivity and Resistance to Black and Frosty Pod Diseases." *Frontiers in Plant Science* 8 (November): 1905.

Salse, Jérôme. 2012. "In Silico Archeogenomics Unveils Modern Plant Genome Organisation, Regulation and Evolution." *Current Opinion in Plant Biology* 15 (2): 122–30.

Särkinen, Tiina, Lynn Bohs, Richard G. Olmstead, and Sandra Knapp. 2013. "A Phylogenetic Framework for Evolutionary Study of the Nightshades (Solanaceae): A Dated 1000-Tip Tree." *BMC Evolutionary Biology* 13 (September): 214.

Sarrion-Perdigones, Alejandro, Marta Vazquez-Vilar, Jorge Palací, Bas Castelijns, Javier Forment, Peio Ziarsolo, José Blanca, Antonio Granell, and Diego Orzaez. 2013.

"GoldenBraid 2.0: A Comprehensive DNA Assembly Framework for Plant Synthetic Biology." *Plant Physiology* 162 (3): 1618–31.

Sarris, Panagiotis F., Volkan Cevik, Gulay Dagdas, Jonathan D. G. Jones, and Ksenia V. Krasileva. 2016. "Comparative Analysis of Plant Immune Receptor Architectures Uncovers Host Proteins Likely Targeted by Pathogens." *BMC Biology* 14 (1): 8.

Sarris, Panagiotis F., Zane Duxbury, Sung Un Huh, Yan Ma, Cécile Segonzac, Jan Sklenar, Paul Derbyshire, et al. 2015. "A Plant Immune Receptor Detects Pathogen Effectors That Target WRKY Transcription Factors." *Cell* 161 (5): 1089–1100.

Saunders, James, and Nichole O'neill. 2004. "The Characterization of Defense Responses to Fungal Infection in Alfalfa." *BioControl (Dordrecht, Netherlands)* 49 (6): 715–28.

Schneider, Caroline A., Wayne S. Rasband, and Kevin W. Eliceiri. 2012. "NIH Image to ImageJ: 25 Years of Image Analysis." *Nature Methods* 9 (7): 671–75.

Schweizer, Rena M., Jonathan P. Velotta, Catherine M. Ivy, Matthew R. Jones, Sarah M. Muir, Gideon S. Bradburd, Jay F. Storz, Graham R. Scott, and Zachary A. Cheviron. 2019. "Physiological and Genomic Evidence That Selection on the Transcription Factor Epas1 Has Altered Cardiovascular Function in High-Altitude Deer Mice." *PLoS Genetics* 15 (11): e1008420.

Shephard, Charles. 2018. *An Historical Account of the Island of Saint Vincent*. Franklin Classics Trade Press.

Shi, Zi, Siela N. Maximova, Yi Liu, Joseph Verica, and Mark J. Guiltinan. 2010. "Functional Analysis of the Theobroma Cacao NPR1 Gene in Arabidopsis." *BMC Plant Biology* 10 (1): 248.

Shi, Zi, Yufan Zhang, Siela N. Maximova, and Mark J. Guiltinan. 2013. "TcNPR3 from Theobroma Cacao Functions as a Repressor of the Pathogen Defense Response." *BMC Plant Biology* 13 (1): 204.

Silva, Carlos Rogério Sousa, Giorgini Augusto Venturieri, and Antonio Figueira. 2004. "Description of Amazonian Theobroma L. Collections, Species Identification, and Characterization of Interspecific Hybrids." *Acta Botanica Brasilica* 18 (2): 333–41.

Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. "BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs." *Bioinformatics (Oxford, England)* 31 (19): 3210–12.

Slatkin, Montgomery. 1993. "Isolation by Distance in Equilibrium and Non-Equilibrium Populations." *Evolution; International Journal of Organic Evolution* 47 (1): 264–79.

Snyder-Leiby, T. E., and D. B. Furtek. 1995. "A Genomic Clone (Accession No. U30324) from Theobroma Cacao L. with High Similarity to Plant Class I Endochitinase Sequences." *Plant Physiology* 109: 338.

Stahl, E. A., G. Dwyer, R. Mauricio, M. Kreitman, and J. Bergelson. 1999. "Dynamics of Disease Resistance Polymorphism at the Rpm1 Locus of Arabidopsis." *Nature* 400 (6745): 667–71.

Stam, Remco, Tetyana Nosenko, Anja C. Hörger, Wolfgang Stephan, Michael Seidel, José M. M. Kuhn, Georg Haberer, and Aurelien Tellier. 2019. "The de Novo Reference Genome and Transcriptome Assemblies of the Wild Tomato Species Solanum Chilense Highlights Birth and Death of NLR Genes between Tomato Species." *G3 (Bethesda, Md.)* 9 (12): 3933–41.

Stam, Remco, Daniela Scheikl, and Aurélien Tellier. 2016. "Pooled Enrichment Sequencing Identifies Diversity and Evolutionary Pressures at NLR Resistance Genes within a Wild Tomato Population." *Genome Biology and Evolution* 8 (5): 1501–15.

———. 2017. "The Wild Tomato Species Solanum Chilense Shows Variation in Pathogen Resistance between Geographically Distinct Populations." *PeerJ* 5 (e2910): e2910.

Stam, Remco, Gustavo A. Silva-Arias, and Aurelien Tellier. 2019. “Subsets of NLR Genes Show Differential Signatures of Adaptation during Colonization of New Habitats.” *The New Phytologist* 224 (1): 367–79.

Stanke, Mario, Oliver Keller, Irfan Gunduz, Alec Hayes, Stephan Waack, and Burkhard Morgenstern. 2006. “AUGUSTUS: Ab Initio Prediction of Alternative Transcripts.” *Nucleic Acids Research* 34 (Web Server issue): W435-9.

Steinbiss, Sascha, Ute Willhoeft, Gordon Gremme, and Stefan Kurtz. 2009. “Fine-Grained Annotation and Classification of de Novo Predicted LTR Retrotransposons.” *Nucleic Acids Research* 37 (21): 7002–13.

Steppuhn, Anke, Klaus Gase, Bernd Krock, Rayko Halitschke, and Ian T. Baldwin. 2004. “Nicotine’s Defensive Function in Nature.” *PLoS Biology* 2 (8): E217.

Steuernagel, Burkhard, Florian Jupe, Kamil Witek, Jonathan D. G. Jones, and Brande B. H. Wulff. 2015. “NLR-Parser: Rapid Annotation of Plant NLR Complements.” *Bioinformatics (Oxford, England)* 31 (10): 1665–67.

Supek, Fran, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. 2011. “REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms.” *PloS One* 6 (7): e21800.

Surujdeo-Maharaj, S., T. N. Sreenivasan, L. A. Motilal, and P. Umaharan. 2016. “Black Pod and Other Phytophthora Induced Diseases of Cacao: History, Biology, and Control.” In *Cacao Diseases*, 213–66. Cham: Springer International Publishing.

Takai, Ryota, Akira Isogai, Seiji Takayama, and Fang-Sik Che. 2008. “Analysis of Flagellin Perception Mediated by Flg22 Receptor OsFLS2 in Rice.” *Molecular Plant-Microbe Interactions: MPMI* 21 (12): 1635–42.

Teixeira, Paulo José Pereira Lima, Daniela Paula de Toledo Thomazella, Osvaldo Reis, Paula Favoretti Vital do Prado, Maria Carolina Scatolin do Rio, Gabriel Lorencini Fiorin, Juliana José, et al. 2014. “High-Resolution Transcript Profiling of the Atypical

Biotrophic Interaction between Theobroma Cacao and the Fungal Pathogen Moniliophthora Perniciosa." *The Plant Cell* 26 (11): 4245–69.

Tenthorey, Jeannette L., Nicole Haloupek, José Ramón López-Blanco, Patricia Grob, Elise Adamson, Ella Hartenian, Nicholas A. Lind, et al. 2017. "The Structural Basis of Flagellin Detection by NAIP5: A Strategy to Limit Pathogen Immune Evasion." *Science (New York, N.Y.)* 358 (6365): 888–93.

Theophrastus. 1989. *Enquiry into Plants: Bks. I-V v. 1*. Translated by A. F. Hort. Loeb Classical Library, No. 7. London, England: LOEB.

Thomas, Evert, Maarten van Zonneveld, Judy Loo, Toby Hodgkin, Gea Galluzzi, and Jacob van Etten. 2012. "Present Spatial Diversity Patterns of Theobroma Cacao L. in the Neotropics Reflect Genetic Differentiation in Pleistocene Refugia Followed by Human-Influenced Dispersal." *PloS One* 7 (10): e47676.

Thresh, J. M., and G. K. Owusu. 1986. "The Control of Cocoa Swollen Shoot Disease in Ghana: An Evaluation of Eradication Procedures." *Crop Protection (Guildford, Surrey)* 5 (1): 41–52.

Tian, D., M. B. Traw, J. Q. Chen, M. Kreitman, and J. Bergelson. 2003. "Fitness Costs of R-Gene-Mediated Resistance in Arabidopsis Thaliana." *Nature* 423 (6935): 74–77.

Tigano, Anna, and Vicki L. Friesen. 2016. "Genomics of Local Adaptation with Gene Flow." *Molecular Ecology* 25 (10): 2144–64.

Torres, G. A., G. A. Sarria, G. Martinez, F. Varon, A. Drenth, and D. I. Guest. 2016. "Bud Rot Caused by Phytophthora Palmivora: A Destructive Emerging Disease of Oil Palm." *Phytopathology* 106 (4): 320–29.

Trdá, Lucie, Olivier Fernandez, Freddy Boutrot, Marie-Claire Héloir, Jani Kelloniemi, Xavier Daire, Marielle Adrian, et al. 2014. "The Grapevine Flagellin Receptor VvFLS2 Differentially Recognizes Flagellin-Derived Epitopes from the Endophytic Growth-

Promoting Bacterium Burkholderia Phytofirmans and Plant Pathogenic Bacteria." *The New Phytologist* 201 (4): 1371–84.

Troyer, A. F. 1990. "A Retrospective View of Corn Genetic Resources." *The Journal of Heredity* 81 (1): 17–24.

Tsugawa, Hiroshi, Tomas Cajka, Tobias Kind, Yan Ma, Brendan Higgins, Kazutaka Ikeda, Mitsuhiro Kanazawa, Jean VanderGheynst, Oliver Fiehn, and Masanori Arita. 2015. "MS-DIAL: Data-Independent MS/MS Deconvolution for Comprehensive Metabolome Analysis." *Nature Methods* 12 (6): 523–26.

Valla, Svein, and Rahmi Lale, eds. 2016. *DNA Cloning and Assembly Methods*. Methods in Molecular Biology 1116. New York, NY: Humana Press.

Van de Weyer, Anna-Lena, Freddy Monteiro, Oliver J. Furzer, Marc T. Nishimura, Volkan Cevik, Kamil Witek, Jonathan D. G. Jones, Jeffery L. Dangl, Detlef Weigel, and Felix Bemm. 2019a. "A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis Thaliana." *Cell* 178 (5): 1260-1272.e14.

———. 2019b. "A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis Thaliana." *Cell* 178 (5): 1260-1272.e14.

Van Zandt, Peter A. 2007. "Plant Defense, Growth, and Habitat: A Comparative Assessment of Constitutive and Induced Resistance." *Ecology* 88 (8): 1984–93.

Vanholme, Ruben, Igor Cesarino, Katarzyna Rataj, Yuguo Xiao, Lisa Sundin, Geert Goeminne, Hoon Kim, et al. 2013. "Caffeoyl Shikimate Esterase (CSE) Is an Enzyme in the Lignin Biosynthetic Pathway in Arabidopsis." *Science (New York, N.Y.)* 341 (6150): 1103–6.

Vaughn, Justin N., and Jeffrey L. Bennetzen. 2014. "Natural Insertions in Rice Commonly Form Tandem Duplications Indicative of Patch-Mediated Double-Strand Break Induction and Repair." *Proceedings of the National Academy of Sciences of the United States of America* 111 (18): 6684–89.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. 4th ed. Statistics and Computing. New York, NY: Springer.

Vlot, A. Corina, D'maris Amick Dempsey, and Daniel F. Klessig. 2009. "Salicylic Acid, a Multifaceted Hormone to Combat Disease." *Annual Review of Phytopathology* 47 (1): 177–206.

Wall, P. Kerr, Jim Leebens-Mack, Kai F. Müller, Dawn Field, Naomi S. Altman, and Claude W. dePamphilis. 2008. "PlantTribes: A Gene and Gene Family Resource for Comparative Genomics in Plants." *Nucleic Acids Research* 36 (Database issue): D970-6.

Wang, Guan-Feng, Yijian He, Renee Strauch, Bode A. Olukolu, Dahlia Nielsen, Xu Li, and Peter J. Balint-Kurti. 2015. "Maize Homologs of Hydroxycinnamoyltransferase, a Key Enzyme in Lignin Biosynthesis, Bind the Nucleotide Binding Leucine-Rich Repeat Rp1 Proteins to Modulate the Defense Response." *Plant Physiology* 169 (3): 2230–43.

Wang, Guo-Liang, and Barbara Valent. 2017. "Durable Resistance to Rice Blast." *Science (New York, N.Y.)*. American Association for the Advancement of Science (AAAS).

Wang, Jianan, Michael D. Coffey, Nicola De Maio, and Erica M. Goss. 2020. "Repeated Global Migrations on Different Plant Hosts by the Tropical PathogenPhytophthora Palmivora." *BioRxiv*. bioRxiv. https://doi.org/10.1101/2020.05.13.093211.

Wang, Jizong, Meijuan Hu, Jia Wang, Jinfeng Qi, Zhifu Han, Guoxun Wang, Yijun Qi, Hong-Wei Wang, Jian-Min Zhou, and Jijie Chai. 2019. "Reconstitution and Structure of a Plant NLR Resistosome Conferring Immunity." *Science (New York, N.Y.)* 364 (6435): eaav5870.

Wang, Jizong, Jia Wang, Meijuan Hu, Shan Wu, Jinfeng Qi, Guoxun Wang, Zhifu Han, et al. 2019. "Ligand-Triggered Allosteric ADP Release Primes a Plant NLR Complex." *Science (New York, N.Y.)* 364 (6435): eaav5868.

Wang, Mingxun, Jeremy J. Carver, Vanessa V. Phelan, Laura M. Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, et al. 2016. “Sharing and Community Curation of Mass Spectrometry Data with Global Natural Products Social Molecular Networking.” *Nature Biotechnology* 34 (8): 828–37.

Wang, Mingxun, Alan K. Jarmusch, Fernando Vargas, Alexander A. Aksenov, Julia M. Gauglitz, Kelly Weldon, Daniel Petras, et al. 2020. “Mass Spectrometry Searches Using MASST.” *Nature Biotechnology* 38 (1): 23–26.

Wang, Minxia, Xiuliang Zhu, Ke Wang, Chungui Lu, Meiying Luo, Tianlei Shan, and Zengyan Zhang. 2018. “A Wheat Caffeic Acid 3-O-Methyltransferase TaCOMT-3D Positively Contributes to Both Resistance to Sharp Eyespot Disease and Stem Mechanical Strength.” *Scientific Reports* 8 (1). https://doi.org/10.1038/s41598-018-24884-0.

Wang, Weidong, Liyang Chen, Kevin Fengler, Joy Bolar, Victor Llaca, Xutong Wang, Chancelor B. Clark, et al. 2021. “A Giant NLR Gene Confers Broad-Spectrum Resistance to Phytophthora Sojae in Soybean.” *Nature Communications* 12 (1): 6263.

Wang, Yupeng, Haibao Tang, Jeremy D. Debarry, Xu Tan, Jingping Li, Xiyin Wang, Tae-Ho Lee, et al. 2012. “MCScanX: A Toolkit for Detection and Evolutionary Analysis of Gene Synteny and Collinearity.” *Nucleic Acids Research* 40 (7): e49.

Waterhouse, Robert M., Fredrik Tegenfeldt, Jia Li, Evgeny M. Zdobnov, and Evgenia V. Kriventseva. 2013. “OrthoDB: A Hierarchical Catalog of Animal, Fungal and Bacterial Orthologs.” *Nucleic Acids Research* 41 (Database issue): D358-65.

Weisenfeld, Neil I., Vijay Kumar, Preyas Shah, Deanna M. Church, and David B. Jaffe. 2017. “Direct Determination of Diploid Genome Sequences.” *Genome Research* 27 (5): 757–67.

Wesselingh, F. P., and J. A. Salo. 2006. “A Miocene Perspective on the Evolution of the Amazonian Biota.” *Scripta Geologica* 133: 439–58.

Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer.

Widmer, Timothy L., and Nathalie Laurent. 2006. "Plant Extracts Containing Caffeic Acid and Rosmarinic Acid Inhibit Zoospore Germination of Phytophthora Spp. Pathogenic to Theobroma Cacao." *European Journal of Plant Pathology* 115 (4): 377–88.

Willmann, Roland, Heini M. Lajunen, Gitte Erbs, Mari-Anne Newman, Dagmar Kolb, Kenichi Tsuda, Fumiaki Katagiri, et al. 2011. "Arabidopsis Lysin-Motif Proteins LYM1 LYM3 CERK1 Mediate Bacterial Peptidoglycan Sensing and Immunity to Bacterial Infection." *Proceedings of the National Academy of Sciences of the United States of America* 108 (49): 19824–29.

Winkelmüller, Thomas M., Frederickson Entila, Shajahan Anver, Anna Piasecka, Baoxing Song, Eik Dahms, Hitoshi Sakakibara, et al. 2021. "Gene Expression Evolution in Pattern-Triggered Immunity within Arabidopsis Thaliana and across Brassicaceae Species." *The Plant Cell* 33 (6): 1863–87.

Wood, Gar, and R. A. Lass. 2001. *Cocoa*. PDF. Edited by G. A. R. Wood and R. A. Lass. 4th ed. Philadelphia, PA: Blackwell Science.

Wu, Yufeng, Zhengge Zhu, Ligeng Ma, and Mingsheng Chen. 2008. "The Preferential Retention of Starch Synthesis Genes Reveals the Impact of Whole-Genome Duplication on Grass Evolution." *Molecular Biology and Evolution* 25 (6): 1003–6.

Xiao, Han, Ning Jiang, Erin Schaffner, Eric J. Stockinger, and Esther van der Knaap. 2008. "A Retrotransposon-Mediated Gene Duplication Underlies Morphological Variation of Tomato Fruit." *Science (New York, N.Y.)* 319 (5869): 1527–30.

Xu, Xiao-Dong, Jia-Yin Guan, Zi-Yi Zhang, Yu-Rou Cao, Kenneth B. Storey, Dan-Na Yu, and Jia-Yong Zhang. 2021. "Novel TRNA Gene Rearrangements in the Mitochondrial Genomes of Praying Mantises (Mantodea: Mantidae): Translocation, Duplication and

Pseudogenization." *International Journal of Biological Macromolecules* 185 (August): 403–11.

Yamauchi, Kazuchika, Seiichi Yasuda, and Kazuhiko Fukushima. 2002. "Evidence for the Biosynthetic Pathway from Sinapic Acid to Syringyl Lignin Using Labeled Sinapic Acid with Stable Isotope at Both Methoxy Groups in Robinia Pseudoacacia and Nerium Indicum." *Journal of Agricultural and Food Chemistry* 50 (11): 3222–27.

Yang, Zhenzhen, Eric K. Wafula, Loren A. Honaas, Huiting Zhang, Malay Das, Monica Fernandez-Aparicio, Kan Huang, et al. 2015. "Comparative Transcriptome Analyses Reveal Core Parasitism Genes and Suggest Gene Duplication and Repurposing as Sources of Structural Novelty." *Molecular Biology and Evolution* 32 (3): 767–90.

Yeo, Sarah, Lauren Coombe, René L. Warren, Justin Chu, and Inanç Birol. 2018. "ARCS: Scaffolding Genome Drafts with Linked Reads." *Bioinformatics* 34 (5): 725–31.

Yi, Xin, Yu Liang, Emilia Huerta-Sanchez, Xin Jin, Zha Xi Ping Cuo, John E. Pool, Xun Xu, et al. 2010. "Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude." *Science (New York, N.Y.)* 329 (5987): 75–78.

Yuan, Minhang, Zeyu Jiang, Guozhi Bi, Kinya Nomura, Menghui Liu, Yiping Wang, Boying Cai, Jian-Min Zhou, Sheng Yang He, and Xiu-Fang Xin. 2021. "Pattern-Recognition Receptors Are Required for NLR-Mediated Plant Immunity." *Nature* 592 (7852): 105–9.

Zarrillo, Sonia, Nilesh Gaikwad, Claire Lanaud, Terry Powis, Christopher Viot, Isabelle Lesur, Olivier Fouet, et al. 2018. "The Use and Domestication of Theobroma Cacao during the Mid-Holocene in the Upper Amazon." *Nature Ecology & Evolution* 2 (12): 1879–88.

Zeid, Aisha Hussein Saleh Abou. 2002. "Stress Metabolites from Corchorus Olitorius L. Leaves in Response to Certain Stress Agents." *Food Chemistry* 76 (2): 187–95.

Zeng, Min, Fuhao Zhang, Fang-Xiang Wu, Yaohang Li, Jianxin Wang, and Min Li. 2020. "Protein-Protein Interaction Site Prediction through Combining Local and Global

Features with Deep Neural Networks." *Bioinformatics (Oxford, England)* 36 (4): 1114–20.

Zhang, Dapeng, Enrique Arevalo-Gardini, Sue Mischke, Luis Zúñiga-Cernades, Alejandro Barreto-Chavez, and Jorge Adriazola Del Aguila. 2006. "Genetic Diversity and Structure of Managed and Semi-Natural Populations of Cocoa (Theobroma Cacao) in the Huallaga and Ucayali Valleys of Peru." *Annals of Botany* 98 (3): 647–55.

Zhang, Dapeng, Michel Boccara, Lambert Motilal, Sue Mischke, Elizabeth S. Johnson, David R. Butler, Bryan Bailey, and Lyndel Meinhardt. 2009. "Molecular Characterization of an Earliest Cacao (Theobroma Cacao L.) Collection from Upper Amazon Using Microsatellite DNA Markers." *Tree Genetics & Genomes* 5 (4): 595–607.

Zhang, Dapeng, Antonio Figueira, Lambert Motilal, Philippe Lachenaud, and Lyndel W. Meinhardt. 2011. "Theobroma." In *Wild Crop Relatives: Genomic and Breeding Resources*, 277–96. Berlin, Heidelberg: Springer Berlin Heidelberg.

Zhang, Dapeng, Windson July Martínez, Elizabeth S. Johnson, Eduardo Somarriba, Wilberth Phillips-Mora, Carlos Astorga, Sue Mischke, and Lyndel W. Meinhardt. 2012. "Genetic Diversity and Spatial Structure in a New Distinct Theobroma Cacao L. Population in Bolivia." *Genetic Resources and Crop Evolution* 59 (2): 239–52.

Zhang, Dapeng, and Lambert Motilal. 2016. "Origin, Dispersal, and Current Global Distribution of Cacao Genetic Diversity." In *Cacao Diseases*, 3–31. Cham: Springer International Publishing.

Zhang, Yu, Rui Xia, Hanhui Kuang, and Blake C. Meyers. 2016. "The Diversification of Plant NBS-LRR Defense Genes Directs the Evolution of MicroRNAs That Target Them." *Molecular Biology and Evolution* 33 (10): 2692–2705.

Zhao, Qiang, Qi Feng, Hengyun Lu, Yan Li, Ahong Wang, Qilin Tian, Qilin Zhan, et al. 2018. "Pan-Genome Analysis Highlights the Extent of Genomic Variation in Cultivated and Wild Rice." *Nature Genetics* 50 (2): 278–84.

Zhou, Huanbin, Jian Lin, Aimee Johnson, Robyn L. Morgan, Wenwan Zhong, and Wenbo Ma. 2011. "Pseudomonas Syringae Type III Effector HopZ1 Targets a Host Enzyme to Suppress Isoflavone Biosynthesis and Promote Infection in Soybean." *Cell Host & Microbe* 9 (3): 177–86.

Zhu, Yun J., Xiaohui Qiu, Paul H. Moore, Wayne Borth, John Hu, Stephen Ferreira, and Henrik H. Albert. 2003. "Systemic Acquired Resistance Induced by BTH in Papaya." *Physiological and Molecular Plant Pathology* 63 (5): 237–48.

Zhu, Zhenglin, Shengjun Tan, Yaqiong Zhang, and Yong E. Zhang. 2016. "LINE-1-like Retrotransposons Contribute to RNA-Based Gene Duplication in Dicots." *Scientific Reports* 6 (1): 24755.

Zipfel, Cyril. 2014. "Plant Pattern-Recognition Receptors." *Trends in Immunology* 35 (7): 345–51.

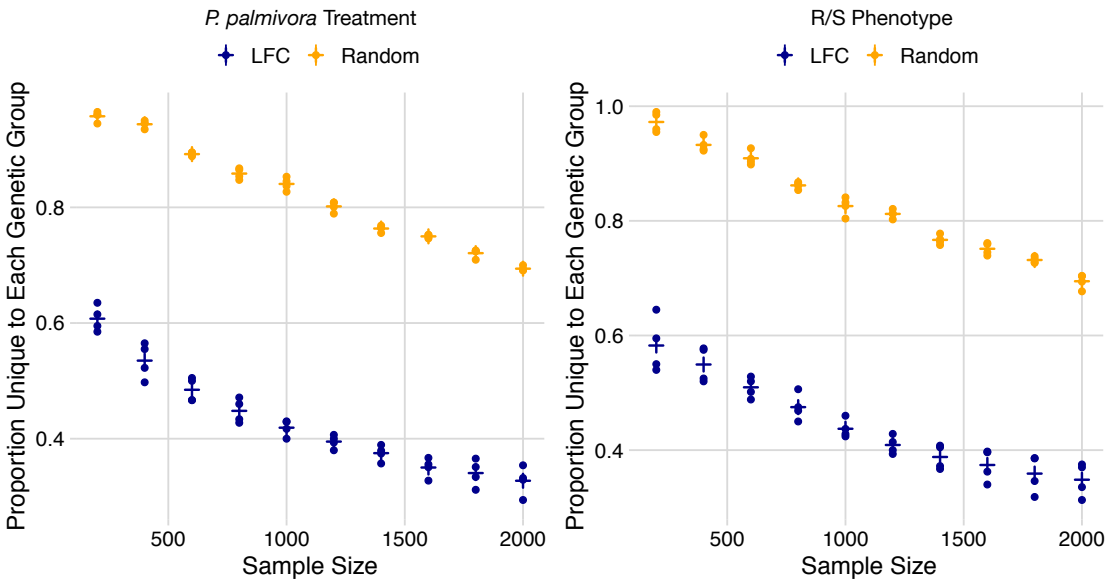
Zou, Cheng, Melissa D. Lehti-Shiu, Françoise Thibaud-Nissen, Tanmay Prakash, C. Robin Buell, and Shin-Han Shiu. 2009. "Evolutionary and Expression Signatures of Pseudogenes in Arabidopsis and Rice." *Plant Physiology* 151 (1): 3–15.

Zou, Xiuping, Junhong Long, Ke Zhao, Aihong Peng, Min Chen, Qin Long, Yongrui He, and Shanchun Chen. 2019. "Overexpressing GH3.1 and GH3.1L Reduces Susceptibility to Xanthomonas Citri Subsp. Citri by Repressing Auxin Signaling in Citrus (Citrus Sinensis Osbeck)." *PloS One* 14 (12): e0220017.

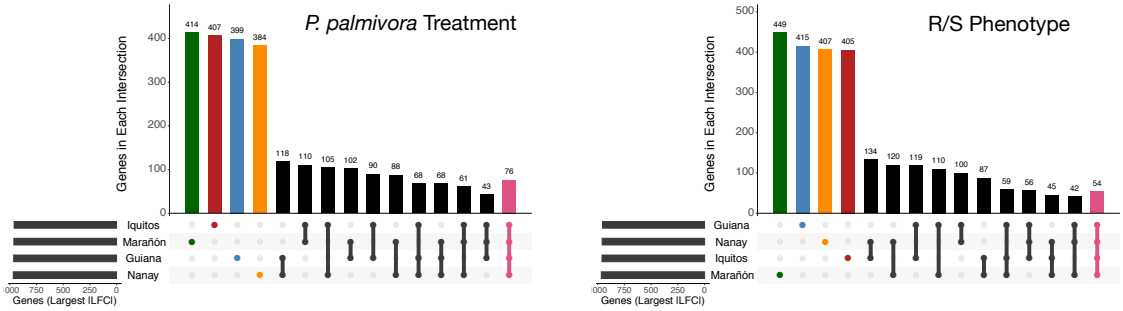
Zumbo, Paul. 1932. "Ethanol Precipitation." *Weill Cornell Medical College*, 1–12.

Appendix A: Supplementary Figures

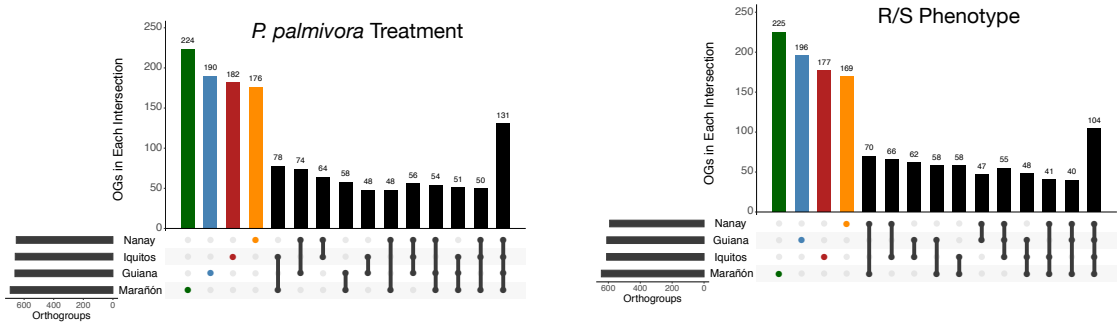
Chapter 2



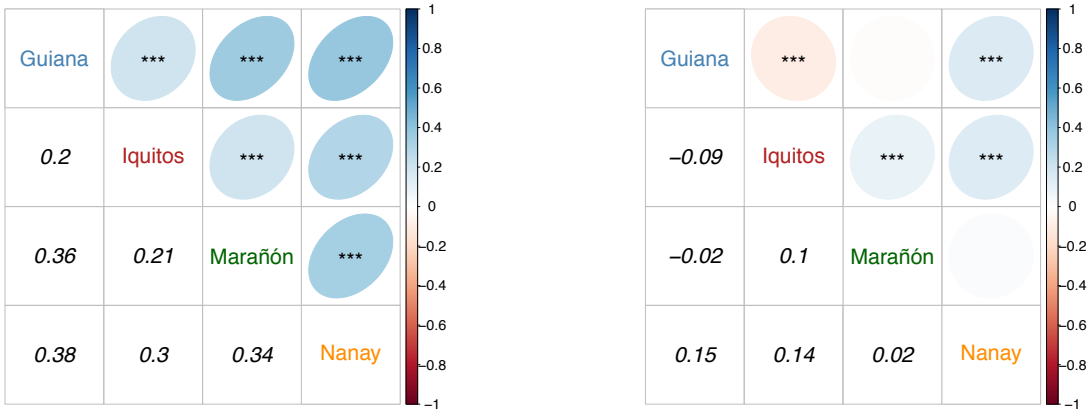
Supplemental Figure S2-1: Proportion of genes that are unique to each population for various sized subsamples, ranging from 200 to 2000 genes, for *P. palmivora* treatment (left) or R/S phenotype (right). Genes were either ranked by $|\log_2$ fold change| before subsampling (blue), or subsampled at random (orange). Each dot represents one of four populations sampled. Means are represented as crosses. For every sample size, the proportion of genes unique to each population was significantly higher when the genes were drawn at random (one-way ANOVA, Proportion Unique Genes ~ Sample Size + Subsample Method + Sample Size:Sample Method: p-values < 0.001).



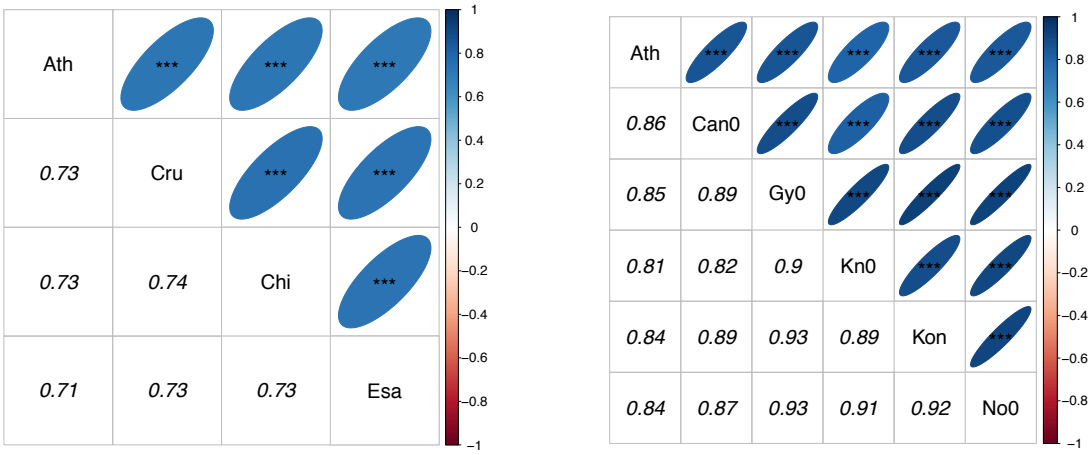
Supplemental Figure S2-2: Overlap of differentially expressed paralogs (i.e. paralogous genes with ≥ 95% identity).The blue, red, green, and orange bars represent GO terms that are only enriched in Guiana Iquitos, Marañón, or Nanay, respectively. The pink bar indicates orthogroups that are significantly enriched across all four populations. Numbers above the bars indicate the number of orthogroups in that specific intersection.



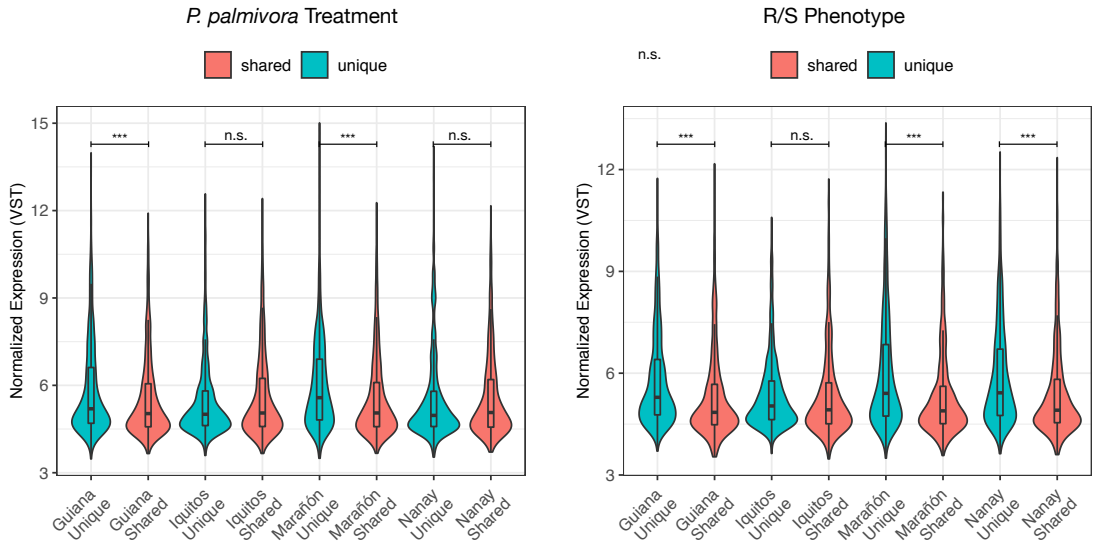
Supplemental Figure S2-3: Overlap of differentially expressed orthogroups (i.e. orthogroups containing 1 or more differentially expressed genes).The blue, red, green, and orange bars represent GO terms that are only enriched in Guiana Iquitos, Marañón, or Nanay, respectively. The pink bar indicates orthogroups that are significantly enriched across all four populations. Numbers above the bars indicate the number of orthogroups in that specific intersection.



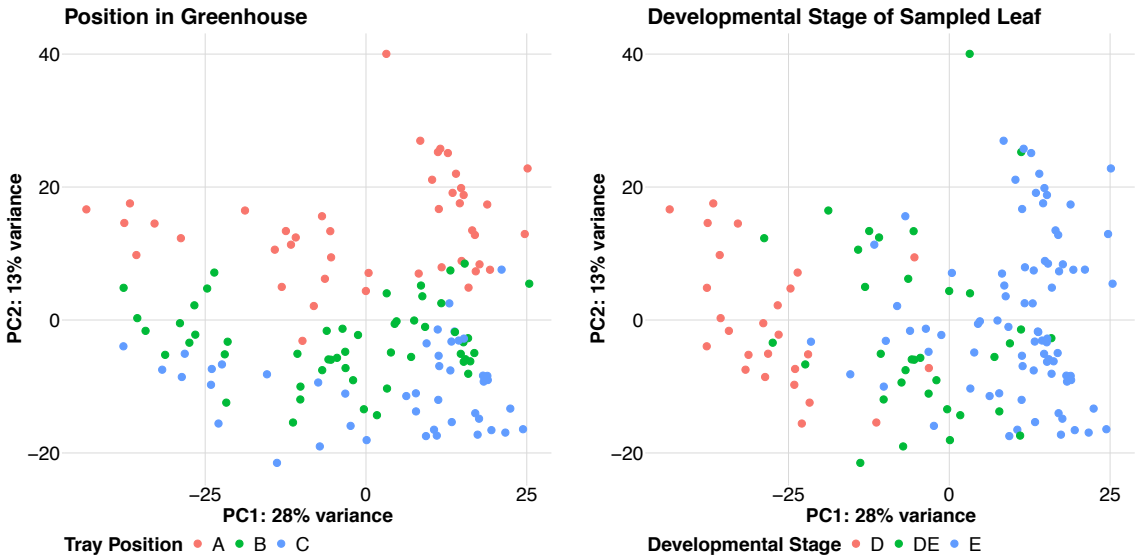
Supplemental Figure S2-4: Pairwise Spearman correlations of mean $\log_2$ fold changes for all orthogroups included in this study. All genes were first classified into orthogroups, then mean $\log_2$ fold change for each orthogroup and population were then calculated. The bottom triangle is the Spearman correlation coefficient. The top triangle is the correlation coefficient depicted as an ellipse, the shape of which depends on the size of the coefficient. Stars indicate statistical significance ($p < 0.001$), tested using Spearman's rho.



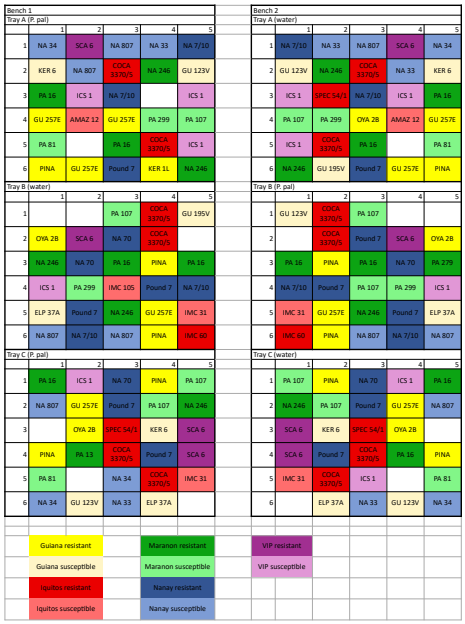
Supplemental Figure S2-5. Pairwise Spearman correlations of $\log_2$ fold changes for 1:1 orthologs between *A. thaliana* and its close relatives (left) and between accessions of *A. thaliana* (right). The bottom triangle is the Spearman correlation coefficient. The top triangle is the correlation coefficient depicted as an ellipse, the shape of which depends on the size of the coefficient. Stars indicate statistical significance ($p < 0.001$), tested using Spearman's rho. Data are from Winkelmüller et al., 2021, *The Plant Cell.*



Supplemental Figure S2-6. Expression of differentially expressed genes that are either unique to a single population (red) or shared across populations, for *P. palmivora* treatment (left) or R/S phenotype (right). Asterisks indicate statistical significance. For treatment, the genes unique to Guiana and Marañón had significantly higher expression than the genes shared among populations (one-way ANOVA, p -value < 2e-16; Tukey's HSD, FDR-adjusted p-value < 0.001). And for phenotype, the genes unique to Guiana, Marañón, and Nanay had significantly higher expression (one-way ANOVA, p -value < 2e-16; Tukey's HSD, FDR-adjusted p-value < 0.001).

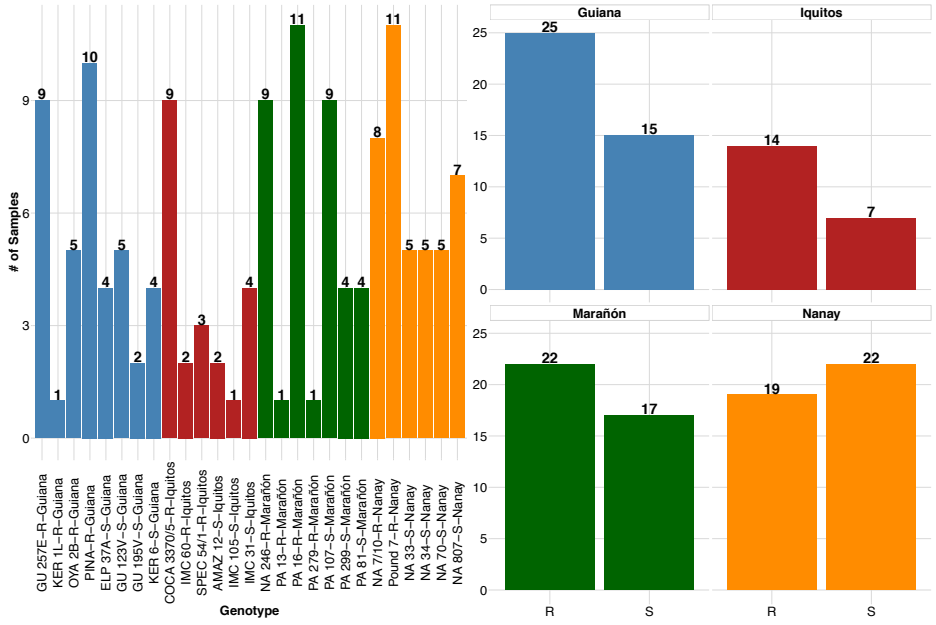


Supplemental Figure S2-7. Environmental covariates included in the GLM used for differential expression. (Left) tray position for each plant in the greenhouse, corresponding to supplemental figure #. (Right) developmental stage of the leaves sampled for the transcriptome experiment.



Supplemental Figure S2-8. Two tables were aligned parallel to one another. On each bench, there were 3 trays, with approx. 30 plants on each tray. To minimize the effect of gradients in

temperature, humidity, and light within the greenhouse, we kept the distance between tables to < 2 ft. We treated the plants in each tray with either pathogen or V8 control, such that parallel trays never experienced the same treatment. We randomized the placement of plants in each tray, with the caveat that the same genotype was in a mirrored position on both tables. Thus for each pair of plants within a genotype, one would receive pathogen treatment and one would receive control treatment. If there was an odd number of plants for a given genotype, or if a genotype only had one representative plant, the odd-numbered individual would be paired with an individual within the same population and resistance/susceptibility class. Lastly, if a genotype within the same population and resistance/susceptibility class was unavailable, we used a genotype in the same resistance/susceptibility class from a different population. Color indicates population membership.



Supplemental Figure S2-9. Distribution of biological replicates for each genotype included in the transcriptome experiment. Color indicates population membership: Guiana (blue), Iquitos (red), Marañón (green), and Nanay (orange). (R) indicates resistant genotypes and (S) indicates susceptible genotypes.

Chapter 3

Table S3-1: Orthogroups that are differentially expressed in response to *P. palmivora* challenge in both wild *Theobroma spp.* and *Theobroma cacao.*

Orthogroup	Mean LFC *T. angustifolium*	Mean LFC *T. bicolor*	Mean LFC *T. grandiflorum*	Mean LFC *T. mammosum*	Mean LFC NSF Populations	AHRD Descriptions
324	-1.970	0.917	-0.822	0.610	-1.157	Major facilitator superfamily protein
60	2.596	2.433	2.125	1.325	1.640	FAD-binding Berberine family protein
309	3.548	1.469	1.856	0.924	1.040	Cellulose synthase family
84	2.076	1.675	2.305	2.522	1.050	Glutamate receptor family
931	1.536	-2.214	0.519	0.339	-1.086	Serine acetyltransferase
747	2.459	2.317	1.570	0.639	1.278	Raffinose synthase family
891	0.626	4.915	0.780	0.647	1.308	Pyridoxal phosphate phosphatase protein
36	-0.422	-1.628	-1.765	-0.495	1.138	Cytochrome P450 superfamily
2059	6.148	3.820	4.725	3.048	1.836	Ca-dependent phospholipid-binding, copine
55	-0.088	-0.245	1.263	0.243	1.100	UDP-Glycosyltransferase superfamily
226	0.960	0.588	1.432	1.447	1.165	12-oxophytodienoate reductase
221	-1.576	-0.022	1.207	1.077	1.175	Plant cadmium resistance
264	3.655	2.226	1.414	1.059	1.494	Unknown
54	1.937	2.207	1.280	1.165	1.476	Disease responsive dirigent-like protein
394	-0.499	1.815	1.356	1.246	1.248	Polyphenol oxidase
103	0.693	1.812	1.011	0.744	1.232	Major pollen allergen
266	-1.408	0.588	-1.176	-0.454	-1.089	Cytochrome P450 superfamily
1083	1.675	3.570	1.516	1.466	1.108	GRAM domain family protein
427	1.302	1.467	1.201	1.082	1.117	Chitinase family protein
638	3.272	3.215	1.713	2.587	1.010	Phospholipase D
487	1.854	1.252	1.376	1.139	1.139	Endochitinase
1746	1.024	1.047	1.368	0.947	1.568	Cinnamate-4-hydroxylase
726	1.564	1.520	1.661	1.176	1.001	Aldehyde dehydrogenase
3000	1.597	3.338	-0.789	1.967	1.797	Ornithine decarboxylase
5989	-1.669	-1.218	-1.615	-1.051	-1.491	Arginase
124	0.511	1.453	0.890	1.394	-1.173	Glutaredoxin family protein
1847	2.515	2.516	3.050	1.939	1.345	Unknown
169	1.005	1.645	-0.977	-1.122	1.078	Cinnamyl alcohol dehydrogenase
241	3.432	3.197	3.763	3.459	1.005	Glyoxal oxidase-related protein
680	2.893	3.887	1.814	1.415	1.111	Lactoylglutathione lyase/glyoxalase I
3184	1.042	1.471	1.046	0.762	1.067	LRR protein kinase family protein
463	2.816	3.331	1.831	1.154	1.178	GDSL esterase/lipase
157	-1.113	2.089	-0.995	-1.017	1.262	Beta-galactosidase
435	1.163	3.187	1.722	2.359	1.152	Calmodulin-binding family protein
578	1.190	0.997	0.987	1.005	1.060	Ammonium transporter
3689	-1.389	-1.919	-1.470	-1.037	1.297	ATP synthase gamma chain
802	1.827	3.513	1.964	2.056	1.267	Ankyrin repeat family protein
361	1.077	0.636	1.345	0.677	1.068	WRKY DNA-binding protein
1081	1.697	2.783	-1.118	-0.767	1.092	BCAA aminotransferase
1636	2.207	2.204	2.063	1.339	1.043	Expansin-like protein family
331	4.948	3.103	2.150	1.183	1.277	1-aminocyclopropane-1-carboxylate synthase
2504	-1.441	-2.472	-1.441	-0.824	-1.161	Glycosyltransferase protein family
177	0.775	1.482	1.470	1.075	-1.206	Unknown
434	1.447	1.496	0.495	0.432	1.088	Major facilitator protein, nodulin-like
2250	5.848	2.841	3.566	5.122	1.514	Isoeugenol synthase
2530	-0.522	1.544	0.661	0.936	1.165	NAD(P)H-ubiquinone oxidoreductase
3875	1.938	1.559	1.865	0.916	1.232	Calmodulin like protein family
4976	1.509	2.141	1.386	1.964	1.615	NAD(P)-binding Rossmann-fold

Noah Winters

State College, PA 16801
Email: npw5@psu.edu
Phone: (330)-472-0633

EDUCATION

The Pennsylvania State University

PhD Student, Ecology

Doctoral Minor: Bioinformatics and Genomics

Dissertation: "Evolution of Disease Resistance in *Theobroma cacao* and its Wild Relatives"

Kenyon College

Bachelor of Arts, May 2015

Major: Molecular Biology, *distinction*

AWARDS

USDA National Institute of Food and Agriculture Predoctoral Fellowship (2019 – 2021)

NIH T32 Training Grant in Computational Biology, Informatics, and Statistics (2018 – 2019)

Pennsylvania State University J. Lloyd Huck Life Sciences Graduate Fellowship (2016 – 2017)

PUBLICATIONS

Pokou, D. N., Fister, A. S., **Winters, N**., Tahi, M., Klotioloma, C., Sebastian, A., ... & Guiltinan, M. J. (2019). Resistant and susceptible cacao genotypes exhibit defense gene polymorphism and unique early responses to Phytophthora megakarya inoculation. Plant molecular biology, 99(4-5), 499-516.

PRESENTATIONS

A Combination of Conserved and Divergent Responses Underly Theobroma Cacao L.'s Defense Response to Phytophthora Palmivora. Plant and Animal Genome Conference, 2022

Transcriptome Sequencing of Two *Theobroma cacao* Varieties Reveals Candidate Resistance Genes to *Phytophthora megakarya*. Plant and Animal Genome Conference, 2017.