Parental assigned chromosomes for cultivated cacao provides insights into genetic architecture underlying resistance to vascular streak dieback

Authors: Peri A. Tobias[1]*, Jacob Downs[1], Peter Epaina[1,5], Gurpreet Singh[1], Robert F. Park[1], Richard J. Edwards[2], Eirene Brugman[3], Andi Zulkifli[6], Junaid Muhammad[3], Agus Purwantara[4], David I. Guest[1]

*Author for correspondence: peri.tobias@sydney.edu.au

Affiliations: [1]School of Life and Environmental Sciences, The University of Sydney, Camperdown, NSW, Australia, [2] Minderoo OceanOmics Centre at UWA, Oceans Institute, The University of Western Australia, Perth, WA, Australia, [3]Cocoa Research Center Faculty of Agriculture, Universitas Hasanuddin, Makassar, Indonesia, [4]Mars Cocoa Research Station, Pangkep, Indonesia, [5]Cocoa Board of Papua New Guinea, Kokopo, East New Britain Province, Papua New Guinea, [6]Megarezky University, Makassar, Indonesia

Keywords: *Theobroma cacao, Ceratobasidium theobromae,* VSD, chromosome-level genome, resistance genes, NLR

Abstract

Diseases of *Theobroma cacao* disrupt cocoa bean supply and economically impact growers. Vascular streak dieback (VSD), caused by *Ceratobasidium theobromae,* is a new encounter disease of cacao currently contained to southeast Asia and Melanesia. Resistance to VSD has been tested with large progeny trials in Sulawesi, Indonesia, and in Papua New Guinea with the identification of informative quantitative trait loci (QTL). Using a VSD susceptible progeny tree (clone 26), derived from a resistant and susceptible parental cross, we assembled the genome to chromosome-level and discriminated alleles inherited from either resistant or susceptible parents. The parentally phased genomes were annotated for all predicted genes and then specifically for resistance genes of the *nucleotide-binding site leucine-rich repeat*

class (NLR). On investigation, we determined the presence of NLR clusters at informative QTLs, as well as other potential disease response gene candidates. Understanding the genetics underlying resistance and susceptibility to VSD will accelerate the breeding cycle by providing clear targets for molecular screening. Additionally, we provide the first diploid, fully scaffolded and parentally phased genome resource for *Theobroma cacao* L.

Introduction

Cacao, *Theobroma cacao* L. (Malvaceae) is an important perennial tree crop producing beans that are used as raw materials in the chocolate, beverage, and cosmetic industries (Schnell *et al.*, 2007). Much of the world's cocoa bean production comes from small-holder, mixed cropping farms in some fifty countries that depend on trade with developed countries where the raw product is manufactured into chocolate (Daymond *et al.*, 2022). Cacao is a difficult crop to grow profitably due to several endemic problems. Alongside other problems, diseases disrupt the supply of cocoa and have serious economic impacts to growers (Marelli *et al.*, 2019). Responses to disease can lead to tree removal and replacement with more profitable and reliable crops (Daymond *et al.*, 2022). The centre of origin for *T. cacao* is the Upper Amazon, South America, where evolutionary studies indicate divergence from close common ancestors around 9.9 million years ago (Richardson *et al.*, 2015). While genetic diversity exists within the geographic centre of origin, cultivation largely relies on a limited genetic pool of *T. cacao* (Motamayor *et al.*, 2008). Southeast (SE) Asia is one of the major suppliers of cocoa beans and cultivation has historically been relatively disease-free due to the lack of co-evolved pathogens (Bailey and Meinhardt, 2016). However in the 1960s, vascular streak dieback (VSD) emerged as a new encounter disease of cacao and caused devastating crop losses in Papua New Guinea (PNG) (Guest and Keane, 2007).

The causal agent of VSD is *Ceratobasidium theobromae*, a fastidious biotrophic fungus, first discovered and described by Philip Keane (1972). *Ceratobasidium theobromae* then spread to other SE Asian and Melanesian cacao regions as an emergent disease on cacao (Guest and Keane, 2007). In Indonesia, the pathogen has spread into Sumatera Regions (Trisno *et al.*, 2016) and was subsequently confirmed in a new area of Barru District, Sulawesi (Junaid *et al.*, 2020). The pathogen is currently contained to the SE Asia region but presents a serious biosecurity threat to cacao in West Africa and Latin America. Symptoms include green spotted chlorosis in older leaves and vascular browning in the petioles and stems that often lead to

61 rapid branch or whole tree death in susceptible genotypes (Guest and Keane, 2007; Samuels
62 *et al.*, 2012). Some genotypes of *T. cacao* that were resistant to disease 50 years ago remain
63 resistant, indicating that durable resistance is highly heritable and therefore may be exploited
64 in breeding to introgress resistance into new, improved cacao genotypes (Tan and Tan, 1988).
65 Varieties of cacao that show resilience to infection by *C. theobromae* have been studied in
66 both PNG and Sulawesi, Indonesia. Results of these studies identified informative quantitative
67 trait loci (QTL) that correlate with both qualitative and quantitative resistance in progeny
68 populations (Epaina, 2012; Singh, 2021). The two independent studies used progenies derived
69 from crosses between Trinitario **KA2-101** (resistant to VSD) and **K82** (susceptible to VSD) in
70 PNG (Epaina, 2012), and between Amelonado **S1** (resistant to VSD) with Iquitos, Criollo and
71 Amelonado **RUQ 1347** (susceptible to VSD and formerly known as CCN 51) in Sulawesi,
72 Indonesia (Singh, 2021). The progeny trials used 346 and 130 progeny that segregated in
73 response to VSD in PNG and Sulawesi, respectively. The informative QTLs associated with
74 resistance indicated a likely polygenic response and were mapped onto nine of the ten *T.*
75 *cacao* chromosomes from the PNG progeny trial (Epaina, 2012) and three in the Sulawesi trial
76 (Singh, 2021). The two studies used either simple sequence repeat (SSR) or single nucleotide
77 polymorphism (SNP) markers and compared linkage maps with established disease resistance
78 mapping developed by Lanaud et al. (2009). The available genomes for Matina 1-6
79 (Motamayor *et al.*, 2013) and Criollo (Argout *et al.*, 2011) were used to inform the linkage
80 maps. Interestingly, VSD QTLs co-localized with a region identified for resistance to
81 Phytophthora Pod Rot (PPR) and Frosty Pod Rot (FP) (causal agents *Phytophthora palmivora*
82 and *Moniliophthora roreri*, respectively) on chromosome 8, while others were common for
83 response to both VSD and PPR on chromosomes 3, 4 and 5 in only one study. Both studies
84 identified VSD QTLs within similar regions on chromosomes 3, 8 and 9. These apparent shared
85 QTL regions indicate hotspots for disease resistance.
86 Classical selection and breeding for resistance, particularly in perennial trees, can be a long
87 and difficult process (Sniezko, 2006). The research is further complicated for resistance to *C.*
88 *theobromae* as the pathogen cannot be cultured in vitro and is extremely slow growing within
89 the host plant before visible symptoms develop. Traditional Koch's postulates cannot be
90 satisfied for biotrophic organisms, however molecular approaches are becoming increasingly
91 important in validation studies (Bhunjun *et al.*, 2021). Understanding the genetics underlying

92 resistance and susceptibility to VSD is likely to accelerate the breeding cycle by providing clear
93 targets for molecular screening.
94 Effective plant responses to disease at the molecular level are well established as involving
95 resistance genes of the *nucleotide-binding site leucine-rich repeat* family (NLRs) (Fluhr *et al.*,
96 2001). NLR genes are known to be numerous, clustered, and highly polymorphic both
97 between and within species (Barragan and Weigel, 2020; Santos *et al.*, 2022; Weyer *et al.*,
98 2019; Van de Weyer *et al.*, 2019). We therefore approached the problem of understanding
99 the underlying genetics of resistance to VSD by building on the structural knowledge of NLR
100 genes and from the progeny QTL studies of Epaina (2012) and Singh (2021). We aimed to
101 leverage the latest sequencing and bioinformatic approaches to gain a clearer insight into the
102 genetic regions of interest. Our premise for the work was that (i) QTLs are informative for
103 predicting resistance, (ii) genetic resistance hotspots are present, (iii) QTLs may indicate the
104 presence of NLR gene clusters, (iv) polymorphic NLR genes within clusters, from contrasting
105 genotypes, may explain disease phenotype. The questions we therefore posed were, can we
106 assemble the cacao genome and successfully discriminate alleles inherited from parents of
107 known VSD phenotype within an F1 tree progeny, and if so, can we determine differences in
108 NLR clusters, type, and specific gene sequences at the QTL locations, based on parentage? An
109 additional question that we tested was, using new sequencing technologies, can we assemble
110 and scaffold to chromosome level both pathogen (~30 Mb) and host (~380 Mb) genomes from
111 a single sample? This would circumvent the inability to culture the biotrophic *C. theobromae*
112 and provide insight into the genomics of a compatible interaction. This final aim was not
113 successful due to the limited pathogen DNA we were able to extract, but future studies could
114 work to improve on this approach.
115 Taking these premises and goals, we built a high-quality parental phased genome resource
116 for a susceptible cacao (clone 26) of known parental cross, **S1 maternal** and **RUQ 1347**
117 **paternal**. We annotated the genome for all predicted genes, and then specifically for only
118 NLR-type genes, and conducted fine scale investigation of genomic regions associated with
119 resistance to *C. theobromae* to determine the resistance gene complements inherited from
120 both parents. We made comparative studies into the NLR genes, structurally, quantitatively,
121 and qualitatively. Our investigations show structural differences in NLR presence within
122 parental chromosomes and provide insights into resistance to VSD and potentially other

serious diseases of cacao. Additionally, we provide the first diploid, fully scaffolded and phased genome resource for *Theobroma cacao* L.

Results and Discussion

Near complete parental assigned chromosomes for cultivated *Theobroma cacao*

We assembled and parentally phased the chromosomes of a VSD susceptible *T. cacao* (clone 26), the progeny of a cross between a maternal resistant (S1, haplotype A) with a paternal susceptible (RUQ 1347, haplotype B) tree. All but three inherited chromosomes (two S1, one RUQ 1347) show telomeric caps (black dots in Figure 1). The genomes show few gaps (red plus symbol in Figure 1), high levels of contiguity and completeness, as well as chromosomal synteny (Edwards *et al.*, 2022) based on conserved single copy orthologs (Simão *et al.*, 2015). Minor inverted translocations on chromosomes four and nine (pink synteny lines) are likely to be biologically real considering the lack of flanking gaps that would indicate mis-scaffolding. The phased genomes are both highly complete according to assessment of conserved single copy orthologs and basic statistics (Table 1).

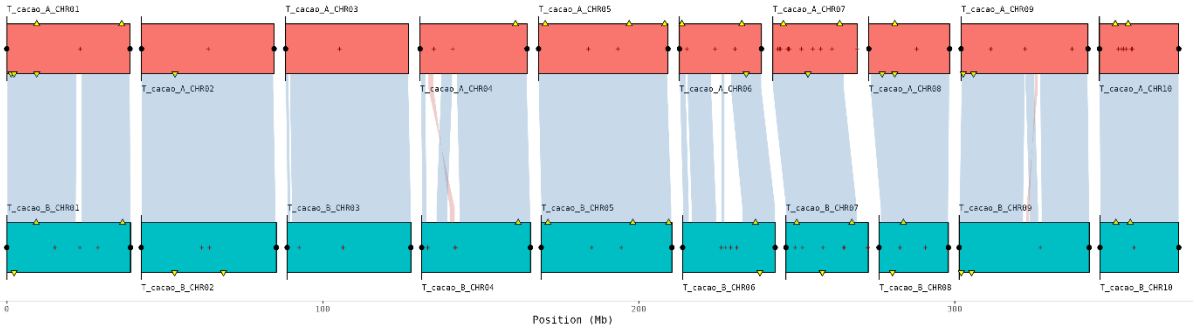


Figure 1 Chromsyn (Edwards *et al.*, 2022) **synteny plots of Theobroma cacao parentally phased genome, haplotype A (top) and B (below). Synteny blocks of collinear "Complete" BUSCO genes** (Simão *et al.*, 2015) **link scaffolds from adjacent assemblies: blue, same strand; red, inverse strand. Yellow triangles mark "Duplicated" BUSCOs. Filled circles mark telomere predictions from Telociraptor (black) (Edwards, 2023). Assembly gaps are marked as dark red + signs.**

The goal for genome assemblies of gapless and telomere-to-telomere (T2T) chromosomes has recently been achievable with the latest sequencing technologies (Sergey *et al.*, 2022). The resulting genome assemblies permit greater depth in understanding structural and evolutionary biology (Mao and Zhang, 2022). Recently, chromosome-level genomes for three wild cacao species from the Upper Amazon were made public (Nousias *et al.*, 2024). These provide the first cacao genomes produced using long read sequence technology and indicate an evolutionary divergence, based on conserved orthologs, from cultivated Criollo and Matina reference genomes (Argout *et al.*, 2011; Motamayor *et al.*, 2013) at ~ 1.83 and ~ 1.34 million

150 years ago. The wild cacao assemblies represent the collapsed haploid state for genomes and
151 are larger in both chromosome and genome base pair (bp) lengths compared with earlier
152 reference genomes. Each of our parentally phased genome assemblies are comparable in size
153 to the wild cacao, at around 380 Mb, however we annotated more predicted protein coding
154 genes at around 26,000 (Table 1) compared to around 21,000, and higher percentage of
155 repetitive sequences at 62% compared to 53% (Nousias *et al.*, 2024). We annotated all genes
156 for the predicted nucleotide binding site (NBARC) domain type as well as predicted 'complete'
157 NLRs that incorporate NBARC plus leucine-rich repeat (LRR) domains (Table 1). Of the
158 complete NLRs we determined only 11 and 15 Toll/Interleukin-1 receptor-type TNL-type, and
159 only 5 and 12 novel integrated domains (ID) within NLRs per haplotype.

160 **Table 1 Raw data statistics from NanoPlot (De Coster and Rademakers, 2023) and genome statistics for both sets of**
161 **parentally inherited *Theombroma cacao* chromosomes using Quast and BUSCOs** (Gurevich *et al.*, 2013; Simão *et al.*, 2015)**.**
162 **Haplotype A chromosomes derived from maternal plant S1 and haplotype B from paternal plant RUQ 1347. Annotation**
163 **statistics based on FindPlantNLRs and Funannotate** (Chen *et al.*, 2023; Palmer and Stajich, 2023)**. NLR annotated resistance**
164 **genes were classified and largely conformed to coiled-coil class with only small numbers of Toll/Interleukin-1 receptor-**
165 **type (TIR) and integrated domain type NLRs.**

Raw HiFi data		
Total bases	30,259,224,851	
Mean read length	14,413.4	
Mean read quality	35.6	
Genome statistic	**Haplotype A**	**Haplotype B**
Total genome length (bp)	374,523,129	381,631,648
chromosome length (bp)	338,223,858	340,140,478
No. of chromosomes	10	10
N50 (bp)	38,945,508	39,156,173
L50	5	5
No. of contigs	165	315
GC (%)	33.73	33.67
BUSCO complete (genome; *n* = 1,614)	**98.2%**	**98.4%**
Single-copy	97.5%	97.8%
Duplicated	0.7%	0.6%
BUSCO fragmented	0.6%	0.9%
BUSCO missing	0.3%	0.7%

Protein-coding genes (funannotate)	26,199	26,321
NBARCs (FindPlantNLRs annotation)	342	370
NLRs	293	326
TIR-NLR (TNL)	11	15
Integrated domain (ID) NLRs	5	12
Repeats	**62.0%**	**62.4%**

Low structural diversity of NLRs between *Theobroma cacao* alleles

The NLR-type resistance genes were annotated, as described in the experimental procedures, and the parentally inherited chromosomes were compared, noting that haplotype A (red annotations (Figure 2Figure 2) aligned with the resistant maternal and haplotype B (yellow annotations) with susceptible paternal trees. We found that the predicted NLR-type genes varied in quantity and types between the two sets of chromosomes. Of the 342 and 370 predicted genes that contained NBARC domains (Table 1), one of the key identifiers of this class of NLR-type resistance genes, our research determined a discrepancy of 293 and 326 complete gene models within Haplotype A and B respectively. Using the web tool version of OrthoVenn3 (Sun *et al.*, 2023) with the OrthoMLC algorithm, using an expect value of 1e-15 and inflation value 1.5, we found that NLR amino acid sequences clustered into 83 orthologous groups (orthogroups). Of these orthogroups, 82 clusters were shared between the two haplotypes with the largest cluster including 55 proteins, 28 and 27 from haplotype A and B respectively. Forty-four orthogroups contained a single NLR protein from each haplotype, meaning that the matching homolog was present in each assembly. Nevertheless, aligned sequences for homologous proteins showed some variations and there were also 2.4% single NLRs, numbering 4 and 11 from haplotype A and B respectively.

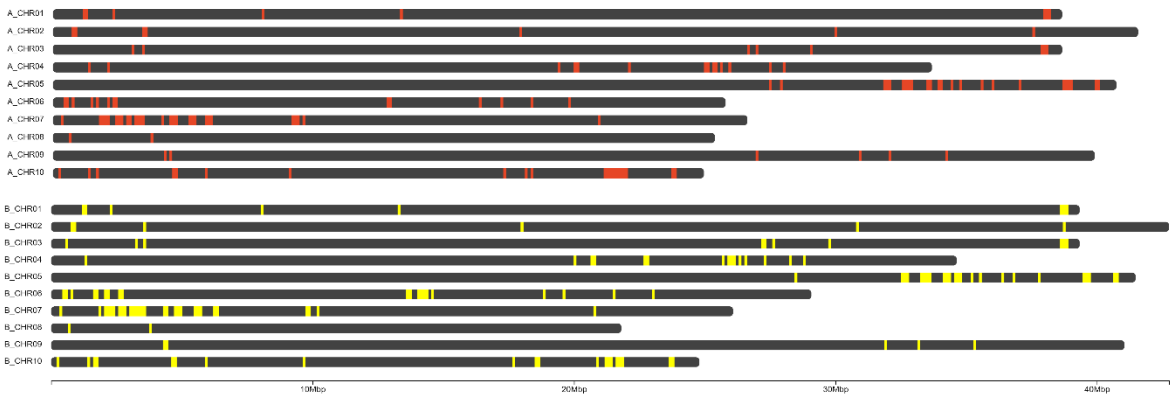
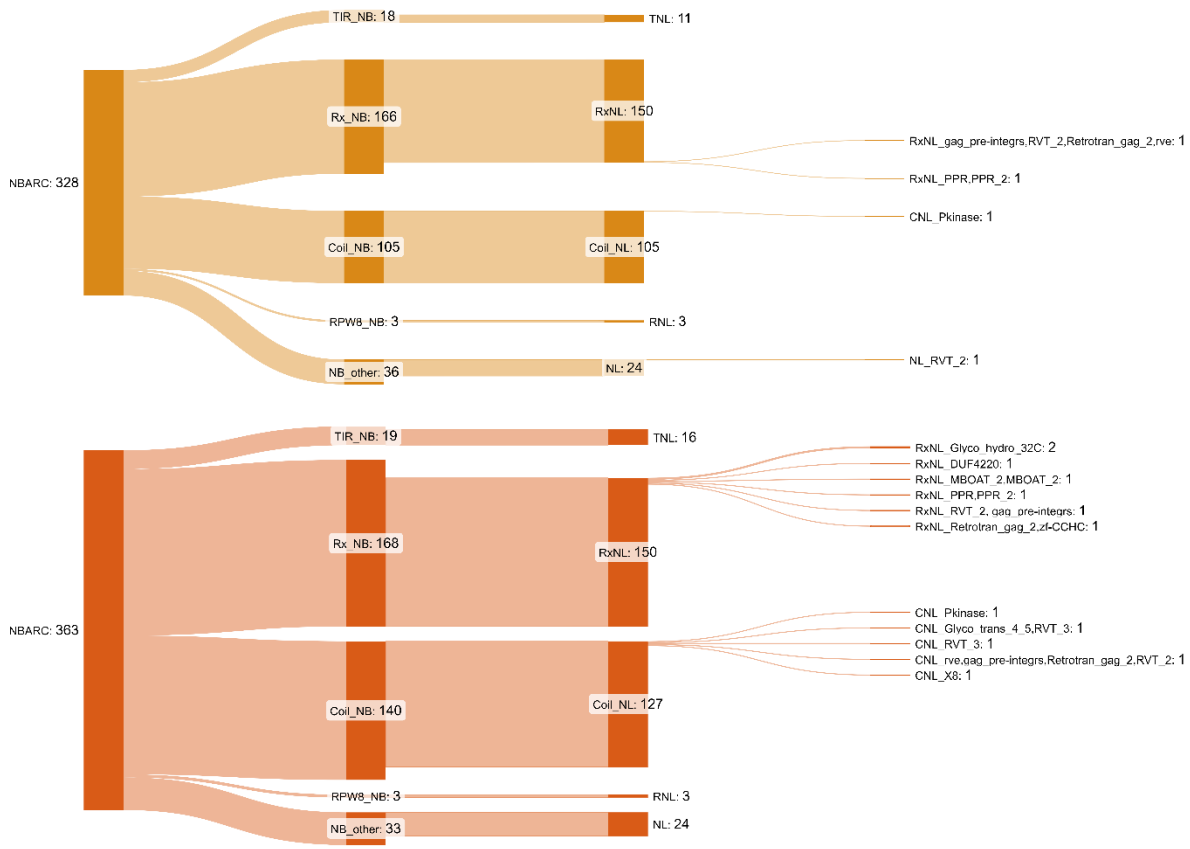


Figure 2 Physical locations of predicted NLR genes on the paired chromosomes of *Theobroma cacao* (A) haplotype A (red) and (B) haplotype B (yellow) generated using ChromoMap (Anand and Rodriguez Lopez, 2022) **in RStudio.**

Low numbers of *Theobroma cacao* NLR-type genes compared to other tree species

Resistance genes of the NLR-type are numerous and highly polymorphic both within and between plant species (Van de Weyer *et al.*, 2019). Trees have generally been shown to have higher numbers of NLR-type genes than herbaceous species, and this is suggested to accomodate a changing pathogen environment over the life of a tree (Tobias and Guest, 2014). More recently, analyses of gene content in diploid genomes have been possible due to advances in sequencing and computational technologies. Notably, tree resistance genes have been investigated within the diploid genome for *Melaleuca quinquenervia* (Chen *et al.*, 2023), highlighting variation in numbers (763 and 733) between haplotypes and across chromosomes for this class of gene. Our analysis of the NLR complement within *T. cacao* showed similar representation to an earlier genome study (Argout *et al.*, 2011) that determined just 297 non-redundant NLR-type genes. Our work, however, presents the diploid content of NLRs as 293 and 326 for haplotype A and B respectively, permitting deeper investigation of homologs at loci. We confirmed low numbers and percentages of the Toll/Interleukin-1 receptor (TIR) NLR class (termed TNLs) of resistance genes within *T. cacao* (Figure 3) compared to many trees within the Rosid clade, in which they comprise 40 - 60% (Arya *et al.*, 2014; Christie *et al.*, 2016; Kohler *et al.*, 2008). This was noted in earlier molecular and genomic studies where only 4% of NLRs contained the TIR domain (Argout *et al.*, 2011; Kuhn *et al.*, 2006), with our classification indicating 3.8% and 5% for haplotype A and B. Unlike the previous study with primers for molecular markers that found TNLs on chromosome three and five (Kuhn *et al.*, 2006), we determined TNLs within clusters on chromosome 4 (three genes), 5 (two genes) and 6 (three and five genes on haplotype A and B), plus a single TNL

208 gene on chromosome 2, 4 and 6 of each haplotype. Our finding of only five and 12 NLR-type
209 genes with non-canonical integrated domains in the two haplotypes also suggests that this
210 integration is less important as a strategy for pathogen recognition and response in cultivated
211 *T. cacao*, compared to other species (Marchal *et al.*, 2022).



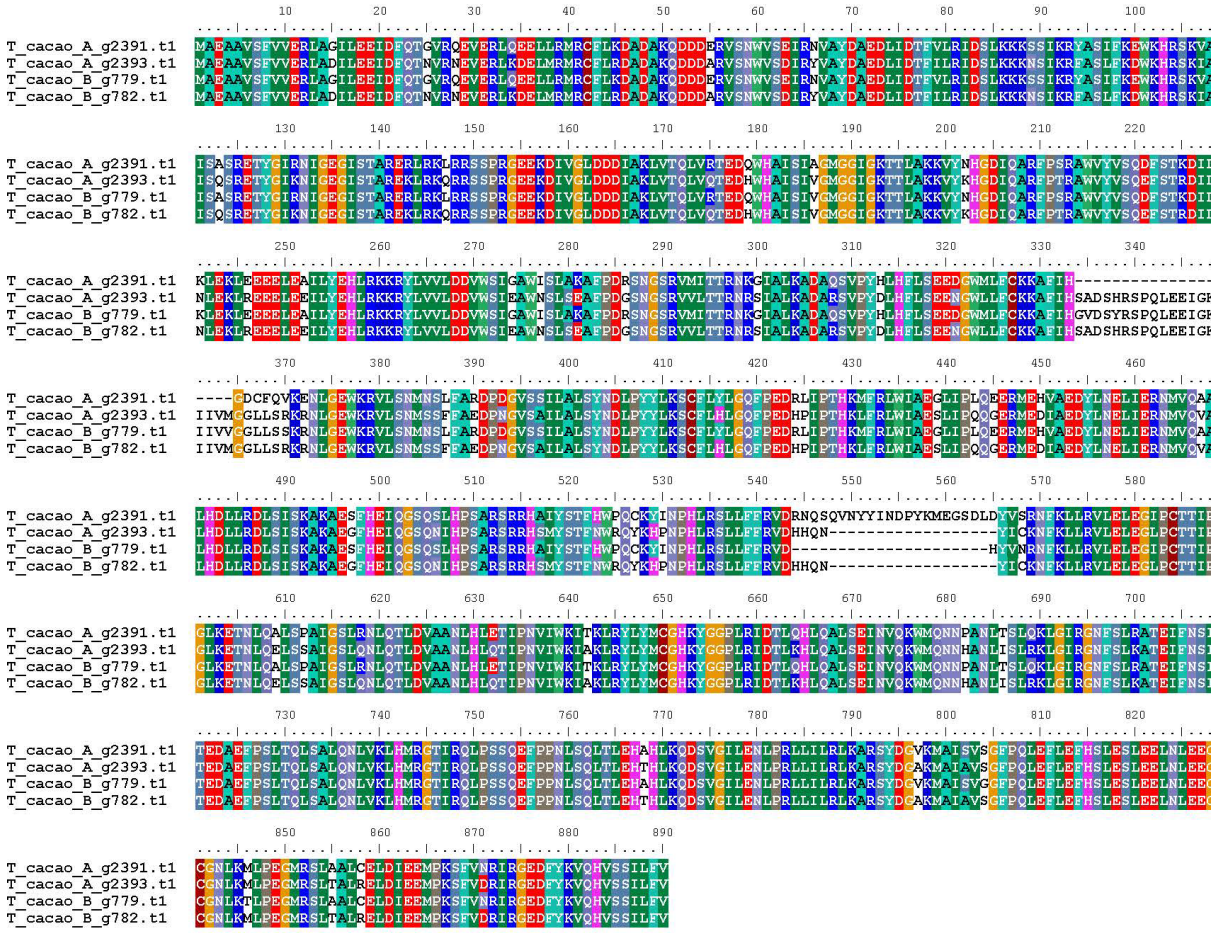
212

213 **Figure 3 The predicted NLR gene complement in the parentally phased *Theobroma cacao* genome. The two sets of**
214 **chromosomes corresponding to (above) Haplotypes A and (below) B were independently classified** (Chen *et al.*, 2023) **and**
215 **visualised to present the domain classes using SankeyMatic (Bogart, 2014) including novel integrated domains (IDs) with**
216 **abbreviations derived from Pfam database** (Mistry *et al.*, 2021)**. NBARC = Nucleotide Binding Domain, TIR =**
217 **Toll/Interleukin-1 receptor, Rx = Potato CC-NB-LRR protein Rx, CC=Coil-Coil Domain, RPW8 = RESISTANCE TO POWDERY**
218 **MILDEW 8-like coiled-coil**

219 *Theobroma cacao* NLR-type genes are present and polymorphic at QTLs for VSD
220 resistance

221 Informative QTL regions were investigated for resistance gene homologs across the parental
222 inherited chromosomes. The most closely aligned VSD QTLs between PNG and Sulawesi *T.*
223 *cacao* progeny trials were on chromosomes 3 and 9, at around 29 and 31 Mb, while
224 chromosome 8 harboured a region between 6 – 7 Mb (Table 3, experimental procedures). We
225 checked these regions based on our annotations for NLRs and identified an NLR cluster of two
226 predicted genes at around 29 Mb on **chromosome 3** within both parentally inherited alleles.

227 The location matches almost exactly with our calculated region based on the reversed start
228 position for the cM QTL (Table 3). When we aligned the four amino acid sequences for these
229 NLRs with ClustalW (Thompson *et al.*, 1994) and visualised in BioEdit (Hall, 1991), we see close
230 homology across the cluster but variation in one haplotype A protein, notably an insertion
231 and deletion in the sequence for gene identity g2391.t1 (Figure 4, above). We found an NLR
232 cluster of three genes on **chromosome 8** at around 3.8 Mb which falls within the 2 cM interval
233 specified by Singh (2021) of 2 - 4 Mb from the QTL locus midpoint. The NLR cluster formed
234 two orthogroups, one group of two head-to-head Rx-NLR predicted proteins from each
235 haplotype (g1318.t1, g1320.t1 and g3681.t1, g3683.t1) and the other of single, nearly exactly
236 homologous proteins, with one amino acid variant, namely proline (g1316.t1, Haplotype A)
237 to arginine (g3679.t1, Haplotype B). Between the region 27 – 36 kb on **chromosome 9** we
238 determined four NLRs from Haplotype A and three from Haplotype B. Paired homologs closest
239 to the calculated QTL mid-point location at around 31 Mb on chromosome 9 were aligned
240 and show an insertion/deletion as well as five single amino acid substitutions (Figure 4,
241 below).The single predicted Rx-NLR gene on chromosome 9 of Haplotype A at 27 Mb has no
242 homolog in Haplotype B.
243 The QTLs on chromosome 3 and 8 accounted for 11 and 15% phenotypic variance in the
244 Sulawesi progeny trials and were informative for VSD and PPR resistance in the PNG trial
245 (Table 3). The chromosome 9 QTL accounted for 16 - 18% phenotypic variance in the Sulawesi
246 trial. All putative resistance genes at QTLs were annotated as Rx NLR-type resistance genes.
247 The Rx-NLR coded protein from potato (*Solanum tuberosum*) has been identified as conferring
248 resistance to the potato virus X by direct recognition. Recent work has clarified the interaction
249 of sensor Rx-NLRs with downstream helper NLR protein oligomerisation (Contreras *et al.*,
250 2023), showing a distinct mechanism for immune responses in plants. While our work has
251 uncovered the arrangement and sequence differences of predicted NLR clusters at QTLs for
252 resistance to VSD, confirmation of these genes is needed across multiple progenies that have
253 clear phenotypes for resistance/susceptibility. Amplification of DNA regions around these
254 clusters will provide additional clarity but was beyond the scope of the current study.
255 Functional validation would also be beneficial to determine mechanisms of recognition and
256 response as it is known that paired head to head NLRs, as we identified on chromosome 8,
257 can determine resistance to more than one pathogen (Xi *et al.*, 2022), potentially explaining
258 the results from the study conducted in PNG.


T_cacao_A_g2391.t1
T_cacao_A_g2393.t1
T_cacao_B_g779.t1
T_cacao_B_g782.t1

259

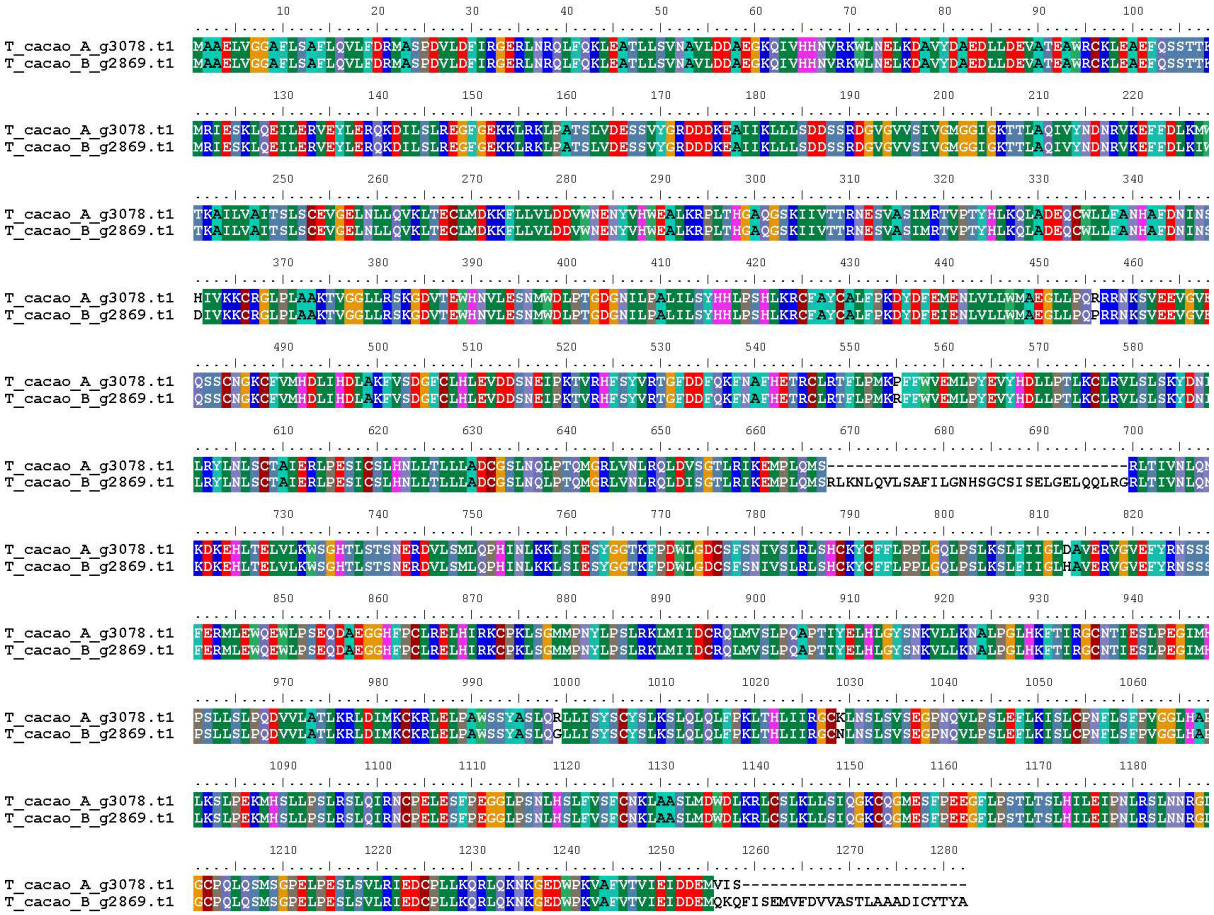


Figure 4 Amino acid sequence alignment of an NLR cluster matching the QTL mapped regions at around 29 Mb on chromosome 3 (above) and single NLR homologs at 31 Mb chromosome 9 (below) within both parentally inherited alleles, A and B, for *Theobroma cacao* clone 26. All predicted NLRs are annotated as Rx-NLRs (Chen *et al.*, 2023) **named for resistance to potato virus X.**

Annotation statistics show high homology between allelic pairs in *Theobroma cacao*

The orthology of all annotated proteins was analysed with OrthoVenn3 as described previously. The predicted proteins indicated 22,486 shared clusters, of which 21,800 included a single predicted homologous gene from each haplotype. The largest cluster had 44 predicted proteins and there were 9.76% singletons. In total there are 2,928 (2,896), 1,865 (1,798), 3,278 (3,315) predicted genes for Haplotype A (and B) on chromosomes 3, 8 and 9 respectively. Except for chromosome 9, with around 76 predicted genes within ~2 Mb of the QTL informing resistance to VSD, the locations on chromosomes 3 and 8 are in highly dense gene-encoding regions, with 325 and 370 predicted genes. We investigated the predicted genes at the QTLs and found that homologs were present between haplotypes (Table 2) and we reviewed their predicted protein functions based on Pfam database (Mistry *et al.*, 2021) annotations. Due to the density of predicted genes, we limited our investigations to

277 immediate proximity around NLRs that corresponded to QTLs. Around the predicted QTL for
278 resistance to VSD on chromosome 3, the Rx-NLR gene cluster is interspersed with two
279 predicted genes coding for protein kinases (Table 2). The kinases have fused LRR domains and
280 predicted transmembrane motifs indicating extracellular receptors and a putative role in
281 defence response. The importance of this class of receptor-like kinase (RLK) has been
282 understood in rice with studies showing *Xa21* confers resistance to *Xanthomonas* sp. and
283 multiple other pathogens (Ercoli *et al.*, 2022). The finding of kinases so closely associated with
284 the Rx-NLR genes on chromosome 3 in both parental alleles, at the predicted region for VSD
285 resistance QTL, suggests a potential role in disease response and a target for further research.
286 The NLR QTL regions associated with resistance on chromosomes 8 and 9 did not have
287 associated disease response-like genes, however it is likely that multiple genes are involved
288 in the VSD resistance response observed.

289 **Table 2 Predicted and annotated genes that are in proximity to the QTLs for resistance to *Ceratobasium theobromae*,**
290 **causal agent of vascular streak dieback in cultivated *Theobroma cacao*. The annotated genomes are available at NCBI:**
291 **PRJNA1083235 and PRJNA1086984.**

Haplotype A	Haplotype B	Pfam ID	Annotation
NLR QTL chromosome 3 (40 kb region)			
V9T21_008377	WDX93_008321	PF00334	Nucleoside diphosphate kinase
V9T21_008378	WDX93_008322	PF01625	Peptide methionine sulfoxide reductase
V9T21_008380	WDX93_008324	PF07714	Protein tyrosine and serine/threonine kinase
V9T21_008380	WDX93_008324	PF13855	Leucine rich repeat
V9T21_008382	WDX93_008326	PF07714	Protein tyrosine and serine/threonine kinase
V9T21_008382	WDX93_008326	PF13855	Leucine rich repeat
V9T21_008382	WDX93_008326	PF08263	Leucine rich repeat N-terminal domain
V9T21_008383	WDX93_008327	PF02298	Plastocyanin-like domain
NLR QTL chromosome 8 (10 kb region)			
V9T21_020396	WDX93_020310	PF00745	Glutamyl-tRNAGlu reductase, dimerisation
V9T21_020396	WDX93_020310	PF01488	Shikimate / quinate 5-dehydrogenase
V9T21_020397	WDX93_020311	PF01293	Phosphoenolpyruvate carboxykinase
V9T21_020399	WDX93_020313	PF08694	Ubiquitin-fold modifier-conjugating enzyme1
V9T21_020401	WDX93_020315	PF12295	Symplekin tight junction protein C terminal
V9T21_020401	WDX93_020315	PF11935	Symplekin/PTA1 N-terminal
V9T21_020402	WDX93_020316	PF11935	Symplekin/PTA1 N-terminal
V9T21_020404	WDX93_020318	PF00719	Inorganic pyrophosphatase
NLR QTL chromosome 9 (3 kb region)			
V9T21_023576	WDX93_023431	PF00514	Armadillo/beta-catenin-like repeat
V9T21_023578	WDX93_023433	PF02341	RbcX protein
V9T21_023579	WDX93_023434	PF00076	RNA recognition motif

292 ## *Concluding remarks*

293 A recent study, using single nucleotide polymporphisms, confirmed low genetic diversity due
294 to selection for domestication in cultivated cacao (Orduña *et al.*, 2024). The results of that
295 work may explain why similar loci are apparent across varieties for resistance to different
296 diseases, as was found in the two studies, in Indonesia and in PNG, that informed the current
297 research. Genome resources have, until recently, been presented as collapsed haploid
298 genomes and the allocation of QTLs has been limited due to loss of valuable information from
299 the two allele sets. Technologies are rapidly improving and permitting far more detailed
300 comparative studies than were possible only a few years ago. Our original goal of assembling
301 both the host and compatible pathogen genomes from one sample was unsuccessful due to
302 low sequence coverage, only 1X HiFi reads, for *C. theobromae* reads. However future methods
303 could use adaptive sequencing with the Oxford Nanopore Technologies metagenome
304 sampling tools (Martin *et al.*, 2022). This approach is rapidly developing, with new models
305 incorporating Bayesian dynamic adaptive sampling (Weilguny *et al.*, 2023), and providing
306 promising tools to investigate the biology of unculturable biotrophic pathogens.
307 For this research we employed recent advances in technology and software to assemble and
308 assign parental chromosomes for a VSD susceptible *T. cacao* genotype (clone 26) from the
309 Mars Cocoa Research Centre in Pangkep, Sulawesi, Indonesia. We annotated the two parental
310 chromosome sets independently and specifically we characterised all the predicted NLR-type
311 resistance genes. The resulting curated and annotated genomes were compared at QTL
312 locations that were identified to explain resistant phenotypes to VSD and PPR. Our results
313 show that the QTLs may indeed resolve to NLR-type resistance genes, and that these may be
314 important to investigate further for breeding trees with resistance to VSD, PPR, and
315 potentially other diseases of cacao. Our work provides a more comprehensive analysis of the
316 predicted NLRs and other genes on each chromosome and shows the variation in structure
317 and type of these genes within a VSD susceptible individual.
318 Future work could develop molecular screening around these specific genes across both
319 resistant and susceptible trees for a range of diseases of cacao. The current work has
320 developed foundational genome resources to facilitate such analyses that may be used to
321 assist future breeding programs for resistance.

322 ## *Experimental procedures*

323 ### High Molecular Weight DNA extraction protocol testing

324 Prior to our work in Sulawesi, Indonesia, we ran exhaustive high molecular weight (HMW)
325 DNA extraction protocol testing on cacao leaf samples from plants maintained in the
326 University of Sydney greenhouses. A persistent problem remained for clean extractions due
327 to high contaminant levels that had the properties of a jelly-like substance, presumably
328 pectin, though not confirmed. We attempted to resolve the nature of the contaminant with
329 enzymatic testing using pectinase. We also tested our contaminant DNA with nuclear
330 magnetic resonance spectrometer (NMR). Data was collected on a 600 MHz spectrometer
331 equipped with triple resonance CryopProbe and the following buffer conditions: 10 mM Tris;
332 10% D2O; pH 8.0, in 3 mm short NMR tubes (volume ~160 µl). We ran controls of 10 mM Tris
333 buffer and clean HMW DNA in Tris buffer. Test samples included pectin, pectin with pectinase
334 treatment, contaminated DNA with pectinase treatment, contaminated DNA with
335 macerozyme (incl. pectinase) treatment. Unfortunately the signals in the aromatic/amide
336 regions were beyond detection, suggesting a large DNA-contaminant (potentially covalent)
337 complex form in this sample. Young leaves were found to produce cleaner DNA but did not
338 show any symptoms of VSD, so we aimed to extract samples from older infected leaves using
339 a combined protocol described below that improved on but did not resolve the contaminant
340 problem.

341 ### Cacao leaf material collected from field trial tree

342 We collected fully expanded leaf material on the 10 January 2023, during wet weather, from
343 the Mars Cocoa Research Institute field trials in Pangkep, South Sulawesi, Indonesia. The
344 leaves were found on T. *cacao* F1 generation progeny, VSD susceptible clone number 26, from
345 Sulawesi 1 – S1 (resistant maternal), crossed with RUQ 1347 (MIS_GBR207 CCN 51)
346 (susceptible paternal) parental plants (Turnbull et al., 2017). The leaves were symptomatic of
347 VSD with browning of the leaf apex and vascular death visible in the petiole (Figure 5). Leaves
348 were placed in plastic bags and transported to the laboratories at Universitas Hasannudin,
349 Makassar and refrigerated overnight.

Figure 5 Leaf from *Theobroma cacao*, VSD susceptible clone 26, growing at the Mars Cocoa Research Institute field trials, Pangkep South Sulawesi, Indonesia. (Left) The symptomatic leaf apex dieback and some chlorosis indicating *Ceratobasium theobromae*. (Right) Visible vascular death caused by the pathogen. Scale bar left ~5 cm.

High Molecular Weight DNA extracted from cacao leaves

The following day while leaves were fresh, HMW DNA was extracted using a protocol adapted from the sorbitol wash method developed by Jones et al. (2021), and the HMW plant DNA CTAB extraction method developed by Hilario (2018). Small (1 cm^2) pieces of the leaves were placed into 1.5 mL centrifuge tubes with a small amount (a few mg) of PVP40000. The tissue was flash frozen in liquid nitrogen and ground into a powder with micropestle while frozen. Three or more rounds of sorbitol wash were used on the samples, in accordance with the protocol (Jones et al., 2021), until the supernatant was no longer viscous. Then DNA was extracted, according to the protocol (Hilario, 2018), using CTAB buffer with an incubation step at 56 °C for 2 hours, in a water bath. Apart from an extended RNase A incubation step, of 10 mins, the protocol was followed until the precipitation step. The DNA formed a visible gelatinous precipitate at this stage and was left to continue precipitating at -20 °C overnight. The following day the DNA was collected by centrifugation at 3000 rpm for 25 min at room temperature and cleanup steps followed according to the protocol before dissolving in 10 mM Tris-HCL pH 8.0. The DNA was quantified, and purity was checked using a NanoDrop™ Lite Plus spectrophotometer and a QuBit™ 2.0 Fluorometer. Purified HMW DNA (10 x 1.5 µL tubes) was couriered under an Australian Department of Agriculture, Water, and the Environment import permit to the Australian Genome Research Facility (AGRF), Brisbane, Australia, for cleanup, size selection, library prep and sequencing on Pacific Biosciences of California, Inc. (PacBio) Sequel II.

374 ## Formaldehyde cross-linking of leaves for HiC sequencing

375 We prepared samples according to the manufacturer's protocols for chromatin crosslinking
376 library prep and sequencing at Phase Genomics (Seattle, Washington, US) to obtain HiC reads.
377 Five grams of leaf material was immersed in 2% (v/v) sodium hypochlorite for 2 mins for
378 surface sterilisation and then washed with ultrapure water. The leaves were cut into strips
379 and immersed in 1% (v/v) formaldehyde for 40 minutes with intermittent mixing. The
380 crosslinking reaction was quenched with the addition of glycine to a final concentration of
381 125 mM for 15 minutes. The crosslinked leaves were washed with ultrapure water, dried,
382 then ground to a powder after freezing with liquid nitrogen.

383 ## Data processing and genome assembly

384 Details of all the software, versions, parameters and scripts used to process the data are
385 available here: https://github.com/peritob/Theobroma-cacao-genome/tree/main

386 ## Preparing the raw data for processing

387 We downloaded the 23 GB (ccs.bam) HiFi sequence read data from AGRF, filtered the reads
388 using HiFiAdapterFilt (Sim *et al.*, 2022) and checked the statistics using Nanoplot (De Coster
389 and Rademakers, 2023). The HiC paired-end read (25 GB) mapping statistics against the cacao
390 Criollo reference genome (GCF_000208745.1) were good for informative read pairs (12.18%)
391 but not against the pathogen genome (GCA_012932095.1) at (0.03%), although 4.28% same
392 strand high quality reads were mapped. We mapped all the filtered HiFi reads to the *C.*
393 *theobromae* genome (Ali *et al.*, 2019) with Minimap2 (Li, 2018) to filter the pathogen from
394 host. We presumed all the mapped reads to be pathogen and all the unmapped reads to be
395 from the host and proceeded accordingly. Additionally, we ran a local blast with internal
396 transcribed spacer (ITS) primers on the VSD mapped raw sequence reads. This confirmed a
397 single exact *in silico* hit using the ITS regions using forward primers ITS4 (White *et al.*, 1990)
398 and reverse ITS5A (Stanford *et al.*, 2000). Raw Illumina data available for the parental plants
399 was trimmed of adaptors and low quality reads with Fastp (Chen *et al.*, 2018).

400 ## Genome assembly and chromosome scaffolding

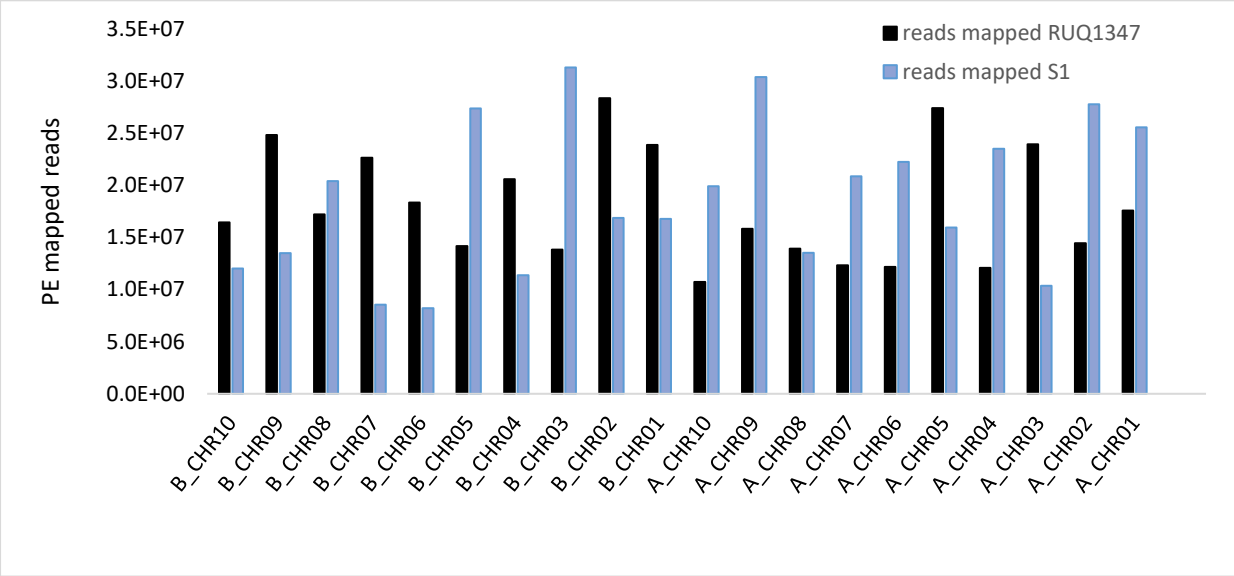
401 We attempted several assembly approaches with Hifiasm software (Cheng *et al.*, 2021).
402 Previous experience with HiC scaffolding has shown the possibility to phase the separate

403 nuclear compartments (Tobias *et al.*, 2022). We therefore attempted to build a plant-
404 pathogen hybrid genome to scaffold with HiC and visually identify and separate the two
405 organisms using the heatmaps. The hybrid genome did not complete, likely due to insufficient
406 overlaps for the pathogen reads. We tried to assemble all the pathogen reads into consensus
407 sequences using Flye (Kolmogorov *et al.*, 2019) and Hifiasm without success. We then tested
408 the trio-binning option for Hifiasm of the diploid *T. cacao* incorporating the Illumina parental
409 reads after running Yak (Li, 2020). This was unsuccessful, so we assembled the diploid *T. cacao*
410 genome using HiC integration with Hifiasm followed by HiC scaffolding of the resulting
411 genomes independently with Juicer (Neva C. Durand *et al.*, 2016), 3D-DNA (Dudchenko *et al.*,
412 2017) and manual scaffolding with Juicebox (Neva C Durand *et al.*, 2016).

413 ### *Theobroma cacao* genome curation

414 We used the Matina 1-6 cacao genome (GCF_000403535.1) to anchor our genome scaffolds
415 with PAFScaff (Field *et al.*, 2020) based on Minimap2 (Li, 2018) mapping. We screened the
416 resulting fasta files for non-nuclear genomic contaminants with FCS-GX (Astashyn *et al.*, 2024)
417 and Tiara (Karlicki *et al.*, 2022). In each case, the two haplotypes were run independently.
418 FCS-GX was run against the NCBI gxdb (build date 2023-01-24; downloaded 2023-09-12).
419 Additional checks of contamination were executed with Taxolotl (Tobias *et al.*, 2022). Taxolotl
420 was provided with the annotation for each haplotype and run with taxwarnrank=class using
421 mmseqs2 against the NCBI nr database (compiled 2022-05-19). In addition to generating
422 results for all sequences in each haplotype, analysis of subsets was performed for (1)
423 chromosome scaffolds only, (2) pure FCS-GX scaffolds, (3) contaminated FCS-GX scaffolds, (4)
424 Tiara "eukarya" scaffolds, and (5) scaffolds not classified as "eukarya" by Tiara. Results were
425 visualised with Pavian. Lists were made of non-plant contaminant contigs from the
426 fcs_gx_reports and were removed, as well as plastid contigs flagged in the tiara outputs. We
427 then ran Telociraptor (Edwards, 2023) followed by Chromsyn (Edwards *et al.*, 2022) on the
428 chromosome-level contigs to compare synteny and scaffolds tweaked according to the contig
429 graphs and figure outputs. We used the parental data available within our lab (Singh, 2021)
430 to assign scaffolds and contigs. We mapped the maternal S1 data to the concatenated
431 haplotype A and B genomes using Hisat2 (Kim *et al.*, 2019) and then similarly mapped the
432 RUQ 1347 data. We assumed that the most reads mapped were the likely parental
433 chromosome/contig and curated each genome. Of the chromosomes, 3, 5 and 8 were

434 switched accordingly (Figure 6). We designated the S1 maternally inherited contigs as genome
435 T_cacao_v1.3_A, and RUQ 1347 paternally inherited as T_cacao_v1.3_B.



436

437 **Figure 6 *Theobroma cacao* resistant (S1) and susceptible (RUQ 1347) parental paired-end (PE) Illumina sequence reads**
438 **mapped to the concatenated Hap A and B progeny chromosomes. The reads that were mapped in greater numbers to**
439 **chromosomes and contigs were used to determine parental assignment and swapped accordingly.**

440 Parental assigned genomes were annotated independently

441 To repeat mask the genome for annotation we ran both genomes through Repeatmasker and
442 Repeatmodeler. We then mapped publicly available RNAseq data from CCN51 *Theobroma*
443 *cacao* from NCBI (PRJNA933172) to the soft-masked genomes independently with hisat2. We
444 trimmed the RNAseq data with Fastp (Chen *et al.*, 2018) prior to mapping and processed the
445 outputs with Samtools (Li *et al.*, 2009) to make bam files. The mapped RNAseq bam files were
446 used as evidence to predict protein coding genes within the funannotate (Palmer and Stajich,
447 2023) *predict* pipeline with the BUSCO embryophyte (Manni *et al.*, 2021; Simão *et al.*, 2015)
448 database. We used Interproscan5 (Jones *et al.*, 2014) independently and then ran
449 funannotate *annotate* to obtain final predicted proteins and coding sequences. In order to
450 make GenBank (Sayers *et al.*, 2023) compatible annotations, we ran the resulting files with
451 Table2asn (GenBank, 2023).

452 NLR-type resistance gene annotation

453 We annotated the resistance gene complement with the FindPlantNLRs pipeline (Chen *et al.*,
454 2023), that incorporates NLR-Annotator (Steuernagel *et al.*, 2020), HMMER (Finn *et al.*, 2015;
455 Johnson *et al.*, 2010) and BLAST (Altschul *et al.*, 1990) components as well as Interproscan5

456 (Jones *et al.*, 2014). Unlike many annotation methods, the pipeline takes an unmasked
457 genome as a starting point to avoid masking, and hence not annotating, repetitive LRR regions
458 that are characteristic of NLR-type genes. The last stage of the pipeline determines NLR gene
459 classes, including predicted integrated domains, and locations within each parental
460 haplotype. All NLRs were plotted to visualise on chromosomes with Chromomap (Anand and
461 Rodriguez Lopez, 2022) and classes, based on domain Pfams, were visualised with
462 Sankeymatic (Bogart, 2014).

463 QTL locations on chromosomes predicted and compared against NLR-type gene 464 predictions

465 Informative QTLs for resistance to VSD were obtained from data of two previous studies
466 (Epaina, 2012; Singh, 2021). Using the chromosome lengths for our parental genome
467 haplotypes we calculated the predicted locations for VSD resistance QTLs. In brief, we took
468 averages of Linkage Group (LG) map lengths (Epaina, 2012), and divided the chromosome
469 base pair length from assembled genomes (current study) by centimorgan (cM). We then
470 determined the expected mid-point (in base pairs) for the resistance QTLs. Singh (2021)
471 predicted QTLs confidence intervals at 3-4 Mb and determined three significant QTLs; on
472 chromosome 3, 8 and 9. Initially our calculations for the QTL locations from the PNG progeny
473 study showed a significant region at around 18.5 Mb on chromosome 8 (Table 3). This was
474 not concordant with the Sulawesi study, at between 5.5 and 8.8 Mb, leading us to investigate
475 if our calculations were misrepresenting the 'start' position for chromosome eight. When
476 recalculated based on the reverse orientation, we found that the QTL matched more closely,
477 at around 7 Mb. We also noted that the QTL identified on chromosome 3, when start positions
478 were reversed, indicated a cluster of NLR-type genes at around 30 Kb. Several significant QTLs
479 were associated with resistance to other pathogens including PPR (Table 3) in the study by
480 Epaina (2012). QTLs that were in similar regions within both studies were regarded as likely
481 harbouring useful candidate genes for breeding resistance. The most closely aligned VSD QTLs
482 between studies were on chromosomes 3, 8 and 9. Detailed comparisons for the VSD
483 significant QTLs across the two studies were undertaken to investigate resistance gene
484 homologs and other predicted genes across the parental chromosomes.

485 **Table 3 QTLs for vascular streak dieback (VSD) resistance were calculated using average *Theobroma cacao* Linkage Group (LG) lengths (Epaina, 2012) and dividing chromosome lengths (bp) by LG lengths (in**
486 **cM) to determine 1 cM in bp (per chromosome). QTLs (cM) were then multiplied by bp length for 1 cM to obtain the chromosome locations for the diploid genome. Green cells and brown cells are VSD and**
487 ***Phytophthora* pod rot (PPR) resistance QTL physical locations on chromosomes.**

***Theobroma cacao* A (S1 inherited)**											
Chromosomes (LG)		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
cM (average)		99.9	107.2	91.6	79.7	89.8	75.8	56.6	60	106.1	62.4
bp per chromosome (this study)		39,005,981	41,965,441	38,945,508	33,857,508	40,959,067	25,938,454	26,721,266	25,579,687	40,159,845	25,091,101
bp per cM		390,450	391,469	425,169	424,812	456,114	342,196	472,107	426,328	378,509	402,101
Singh (2021)	VSD QTL (bp)			9,353,725 **(*29,591,783)**					8,782,359	25,170,874	
Epaina (2012)	VSD QTL (bp)	11,270,737	26,861,405	8,531,872 **(*30,413,636)**	27,163,747	7,976,071			18,514,151 **(*7,065,536)**	31,452,994	
Epaina (2012)	PPR QTL (bp)			8,531,872 **(*30,413,636)**	23,158,620	6,937,499			18,514,151 **(*7065,536)**		
***Theobroma cacao* B (RUQ 1347 inherited)**											
Chromosomes (LG)		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
bp per chromosome (this study)		39,156,173	42,761,000	39,170,102	34,430,140	41,360,500	29,284,707	25,927,970	21,951,902	41,168,245	24,929,739
bp per cM		391,954	398,890	427,621	431,997	460,585	386,342	458,091	365,865	388,014	399,515
Singh (2021)	VSD QTL (bp)			9,407,666 **(*29,762,436)**					7,536,820	25,802,906	
Epaina (2012)	VSD QTL (bp)	11,314,135	27,370,630	8,581,075 **(*30,589,027)**	27,623,167	8,054,243			15,888,421 **(*6,063,481)**	32,242,768	
Epaina (2012)	PPR QTL (bp)			8,581,075 **(*30,589,027)**	23,550,302	7,005,492			15,888,421 **(*6,063,481)**		

488 ***QTL determined with start position on chromosome reversed.**

Acknowledgements

The Joint Cocoa Research Fund, European Cocoa Association supported this research. The authors acknowledge the facilities, and the scientific and technical assistance of the Sydney Informatics Hub at the University of Sydney and, in particular, access to the high-performance computing facility Artemis. We acknowledge the support of Sydney Analytical for running samples and interpretation on the University of Sydney Nuclear Magnetic Resonance (NMR) Core Facilities. Thank you to Mark Powrie who generously assisted during the work in Sulawesi.

Author contributions

DG and PT initiated and led the project. PT and JD collected the material from established clonal trials developed and managed by AP. JM and AP supported and facilitated the research in Sulawesi. JD, PT, AZ, EB extracted HMW DNA and cross-linked samples. PT and EB organised sample deliveries. PE and GS provided results of QTL trials that were supported by DG and RP. GS provided parental data. PT assembled, curated, annotated the genomes, and interpreted results, RE ran contamination screening of genomes. PT submitted the data to NCBI and wrote the manuscript. All authors contributed comments and edits to the manuscript.

Data availability statement

All data that support the findings of this research are publicly available. Raw data is available at National Centre for Biotechnology Information (NCBI) under the following biosample accession: SAMN40241783 and bioproject accession: PRJNA1083235. Annotated diploid *Theobromae cacao* (clone 26) genomes are available here: PRJNA1083235 (maternally inherited), PRJNA1086984 (paternally inherited). NLR predicted coding and amino acid sequences and gff3 data is available at https://github.com/peritob/Theobroma-cacao-genome/tree/main.

References

Ali, S.S., Asman, A., Shao, J., et al. (2019) Draft genome sequence of fastidious pathogen Ceratobasidium theobromae, which causes vascular-streak dieback in Theobroma cacao. *Fungal Biol. Biotechnol.* **6**, 1–10.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–10.

Anand, L. and Rodriguez Lopez, C.M. (2022) ChromoMap: an R package for interactive visualization of multi-omics data and annotation of chromosomes. *BMC Bioinformatics* **23**, 1–9.

Argout, X., Salse, J., Aury, J.-M., et al. (2011) The genome of Theobroma cacao. *Nat. Genet.* **43**, 101–108.

Arya, P., Kumar, G., Acharya, V. and Singh, A.K. (2014) Genome-wide identification and expression analysis of NBS-encoding genes in *Malus* x *domestica* and expansion of *NBS* genes family in Rosaceae. *PLoS One* **9**, e107987.

Astashyn, A., Tvedte, E.S., Sweeney, D., et al. (2024) Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biol.* **25**, 1–25.

Bailey, B.A. and Meinhardt, L.W. eds (2016) Cacao Diseases: A History of Old Enemies and New Encounters. In Cacao Diseases: A History of Old Enemies and New Encounters., Heidelberg: Springer International Publishing.

Barragan, A.C. and Weigel, D. (2020) Plant NLR Diversity : The Known Unknowns of Pan-NLRomes A Brief History of Plant NLRs. *Plant Cell*.

Bhunjun, C.S., Phillips, A.J.L., Jayawardena, R.S., Promputtha, I. and Hyde, K.D. (2021) Importance of molecular data to identify fungal plant pathogens and guidelines for pathogenicity testing based on Koch'S postulates. *Pathogens* **10**, 1–18.

Bogart, S. (2014) SankeyMatic. https://sankeymatic.com/build/

Chen, S., Zhou, Y., Chen, Y. and Gu, J. (2018) Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890.

Chen, S.H., Martino, A.M., Luo, Z., Schwessinger, B., Jones, A., Tolessa, T., Bragg, J.G., Tobias, P.A. and Edwards, R.J. (2023) A high-quality pseudo-phased genome for Melaleuca quinquenervia shows allelic diversity of NLR-type resistance genes. *Gigascience* **12**, 1–18.

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H. and Li, H. (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175.

Christie, N., Tobias, P.A., Naidoo, S. and Külheim, C. (2016) The *Eucalyptus grandis NBS-LRR* gene family: physical clustering and expression hotspots. *Front. Plant Sci.* **6**, 1–16.

Contreras, M.P., Pai, H., Tumtas, Y., et al. (2023) Sensor NLR immune proteins activate oligomerization of their NRC helpers in response to plant pathogens . *EMBO J.* **42**.

Coster, W. De and Rademakers, R. (2023) NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* **39**, 0–2.

Daymond, A.J., Mendez, D.G., Hadley, P., Bastide, P., Abdoellah, S., Acheampong, K., Amores-Puyutaxi, F.M., Anhert, D. and Konan, D.C. (2022) *A Global Review of Cocoa Farming Systems*,.

Dudchenko, O., Batra, S.S., Omer, A.D., et al. (2017) De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science (80-.).* **356**, 92–95.

Durand, Neva C, Robinson, J.T., Shamim, M.S., et al. (2016) Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom Tool. *Cell Syst.* **3**, 99–101.

Durand, Neva C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S. and Aiden, E.L. (2016) Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **3**, 95–98.

Edwards, R.J. (2023) Telociraptor. https://github.com/slimsuite/telociraptor

Edwards, R.J., Dong, C., Park, R.F. and Tobias, P.A. (2022) A phased chromosome-level genome and full mitochondrial sequence for the dikaryotic myrtle rust pathogen, Austropuccinia psidii. *bioRxiv*, 2022.04.22.489119.

566 **Epaina, P.** (2012) *Identification of Molecular Markers and Quantitative Trait Loci linked to*
567 *resistance to Vascular Streak Dieback and Phytophthora Pod Rot of Cacao (Theobroma*
568 *cacao L .)*. The University of Sydney.
569 **Ercoli, M.F., Luu, D.D., Rim, E.Y., et al.** (2022) Plant immunity: Rice XA21-mediated resistance
570 to bacterial infection. *Proc. Natl. Acad. Sci. U. S. A.* **119**.
571 **Field, M.A., Rosen, B.D., Dudchenko, O., et al.** (2020) Canfam-GSD: De novo chromosome-
572 length genome assembly of the German Shepherd Dog (Canis lupus familiaris) using a
573 combination of long reads, optical mapping, and Hi-C. *Gigascience* **9**, 1–12.
574 **Finn, R.D., Clements, J., Arndt, W., Miller, B.L., Wheeler, T.J., Schreiber, F., Bateman, A. and**
575 **Eddy, S.R.** (2015) HMMER web server: 2015 Update. *Nucleic Acids Res.* **43**, W30–W38.
576 **Fluhr, R., King, S.I., Vi, H. and Shakespeare, W.** (2001) Sentinels of Disease . Plant Resistance
577 Genes. **127**, 1367–1374.
578 **GenBank** (2023) Table2asn. https://www.ncbi.nlm.nih.gov/genbank/table2asn
579 **Guest, D. and Keane, P.** (2007) Vascular-streak dieback: A new encounter disease of cacao in
580 Papua New Guinea and Southeast Asia caused by the obligate basidiomycete
581 Oncobasidium theobromae. *Phytopathology* **97**, 1654–1657.
582 **Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G.** (2013) QUAST: Quality assessment tool
583 for genome assemblies. *Bioinformatics* **29**, 1072–1075.
584 **Hall, T.A.** (1991) BioEdit: a user-friendly biological sequence alignment editor and analysis
585 program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41**, 95–98.
586 **Hilario, E.** (2018) Plant nuclear genomic DNA preps. *protocols.io:*
587 **Johnson, L.S., Eddy, S.R. and Portugaly, E.** (2010) Hidden Markov model speed heuristic and
588 iterative HMM search procedure. *BMC Bioinformatics* **11**.
589 **Jones, A., Torkel, C., Stanley, D., Nasim, J., Borevitz, J. and Schwessinger, B.** (2021) High-
590 molecular weight DNA extraction, cleanup and size selection for long-read sequencing.
591 *PLoS One* **16**, 1–6.
592 **Jones, P., Binns, D., Chang, H.Y., et al.** (2014) InterProScan 5: Genome-scale protein function
593 classification. *Bioinformatics* **30**, 1236–1240.
594 **Junaid, M., Purwantara, A. and Guest, D.** (2020) First report of vascular streak dieback
595 symptom of cocoa caused by Ceratobasidium theobromae in Barru District, South
596 Sulawesi. *IOP Conf. Ser. Earth Environ. Sci.* **486**.
597 **Karlicki, M., Antonowicz, S. and Karnkowska, A.** (2022) Tiara: deep learning-based
598 classification system for eukaryotic sequences. *Bioinformatics* **38**, 344–350.
599 **Keane, P.J., Flentje, N.J. and Lamb, K.P.** (1972) Investigation of vascular-streak dieback of
600 cocoa in papua new guinea. *Aust. J. Biol. Sci.* **25**, 553–564.
601 **Kim, D., Paggi, J.M., Park, C., Bennett, C. and Salzberg, S.L.** (2019) Graph-based genome
602 alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–
603 915.
604 **Kohler, A., Rinaldi, C., Duplessis, S., Baucher, M., Geelen, D., Duchaussoy, F., Meyers, B.C.,**
605 **Boerjan, W. and Martin, F.** (2008) Genome-wide identification of NBS resistance genes
606 in Populus trichocarpa. *Plant Mol. Biol.* **66**, 619–636.
607 **Kolmogorov, M., Yuan, J., Lin, Y. and Pevzner, P.A.** (2019) Assembly of long, error-prone
608 reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546.
609 **Kuhn, D.N., Narasimhan, G., Nakamura, K., Brown, J.S., Schnell, R.J. and Meerow, A.W.**
610 (2006) Identification of cacao TIR-NBS-LRR resistance gene homologues and their use as
611 genetic markers. *J. Am. Soc. Hortic. Sci.* **131**, 806–813.
612 **Lanaud, C., Fouet, O., Clément, D., Boccara, M., Risterucci, A.M., Surujdeo-Maharaj, S.,**

613 **Legavre, T. and Argout, X.** (2009) A meta-QTL analysis of disease resistance traits of
614 Theobroma cacao L. *Mol. Breed.* **24**, 361–374.
615 **Li, H.** (2018) Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**,
616 3094–3100.
617 **Li, H.** (2020) Yak.
618 **Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G.**
619 **and Durbin, R.** (2009) The Sequence Alignment/Map format and SAMtools.
620 *Bioinformatics* **25**, 2078–2079.
621 **Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A. and Zdobnov, E.M.** (2021) BUSCO
622 Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic
623 Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* **38**,
624 4647–4654.
625 **Mao, Y. and Zhang, G.** (2022) A complete, telomere-to-telomere human genome sequence
626 presents new opportunities for evolutionary genomics. *Nat. Methods* **19**, 635–638.
627 **Marchal, C., Michalopoulou, V.A., Zou, Z., Cevik, V. and Sarris, P.F.** (2022) Show me your ID:
628 NLR immune receptors with integrated domains in plants. *Essays Biochem.* **66**, 527–539.
629 **Marelli, J.P., Guest, D.I., Bailey, B.A., Evans, H.C., Brown, J.K., Junaid, M., Barreto, R.W.,**
630 **Lisboa, D.O. and Puig, A.S.** (2019) Chocolate under threat from old and new cacao
631 diseases. *Phytopathology* **109**, 1331–1343.
632 **Martin, S., Heavens, D., Lan, Y., Horsfield, S., Clark, M.D. and Leggett, R.M.** (2022) Nanopore
633 adaptive sampling: a tool for enrichment of low abundance species in metagenomic
634 samples. *Genome Biol.* **23**, 1–27.
635 **Mistry, J., Chuguransky, S., Williams, L., et al.** (2021) Pfam: The protein families database in
636 2021. *Nucleic Acids Res.* **49**, D412–D419.
637 **Motamayor, J.C., Lachenaud, P., Silva e Mota, J.W. da, Loor, R., Kuhn, D.N., Brown, J.S. and**
638 **Schnell, R.J.** (2008) Geographic and genetic population differentiation of the Amazonian
639 chocolate tree (Theobroma cacao L). *PLoS One* **3**.
640 **Motamayor, J.C., Schnell, R., Kuhn, D., et al.** (2013) The genome sequence of the most widely
641 cultivated cacao type and its use to identify candidate genes regulating pod color.
642 *Genome Biol.* **14**, r53.
643 **Nousias, O., Zheng, J., Li, T., et al.** (2024) Three de novo assembled wild cacao genomes from
644 the Upper Amazon. *Sci. data* **11**, 1–9.
645 **Orduña, H.E.N., Müller, M., Krutovsky, K. V and Gailing, O.** (2024) Genotyping of cacao
646 (Theobroma cacao L .) germplasm resources with SNP markers linked to agronomic traits
647 reveals signs of selection. *Tree Genet. Genomes*.
648 **Palmer, J.M. and Stajich, J.E.** (2023) Funannotate.
649 **Richardson, J.E., Whitlock, B.A., Meerow, A.W. and Madriñán, S.** (2015) The age of
650 chocolate: A diversification history of Theobroma and Malvaceae. *Front. Ecol. Evol.* **3**, 1–
651 14.
652 **Samuels, G.J., Ismaiel, A., Rosmana, A., et al.** (2012) Vascular Streak Dieback of cacao in
653 Southeast Asia and Melanesia: In planta detection of the pathogen and a new taxonomy.
654 *Fungal Biol.* **116**, 11–23.
655 **Santos, M. de L., Resende, M.L.V. de, Alves, G.S.C., Huguet-Tapia, J.C., Resende, M.F.R. de**
656 **J. and Brawner, J.T.** (2022) Genome-Wide Identification, Characterization, and
657 Comparative Analysis of NLR Resistance Genes in Coffea spp. *Front. Plant Sci.* **13**, 1–16.
658 **Sayers, E.W., Cavanaugh, M., Clark, K., Pruitt, K.D., Sherry, S.T., Yankie, L. and Karsch-**
659 **Mizrachi, I.** (2023) GenBank 2023 update. *Nucleic Acids Res.* **51**, D141–D144.

Schnell, R.J., Kuhn, D.N., Brown, J.S., Olano, C.T., Amores, F.M. and Motamayor, J.C. (2007) Development of a Marker Assisted Selection Program for Cacao. In Cacao Diseases : Important Threats to Chocolate Production Worldwide., Phytopathology.

Sergey, N., Sergey, K., Arang, R., et al. (2022) The complete sequence of a human genome. *Science (80-.).* **376**, 44–53.

Sim, S.B., Corpuz, R.L., Simmonds, T.J. and Geib, S.M. (2022) HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. *BMC Genomics* **23**.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E. V. and Zdobnov, E.M. (2015) BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212.

Singh, G. (2021) *Identification of Quantitative Trait Loci (QTL) associated with resistance to Vascular Streak Dieback disease of cacao*. The University of Sydney.

Sniezko, R.A. (2006) Resistance breeding against nonnative pathogens in forest trees — current successes in North America 1. *Can. J. Plant Pathol.* **279**, 270–279.

Stanford, A.M., Harden, R. and Parks, C.R. (2000) Phylogeny and biogeography of Juglans (Juglandaceae) based on matK and ITS sequence data. *Am. J. Bot.* **87**, 872–882.

Steuernagel, B., Witek, K., Krattinger, S.G., et al. (2020) The NLR-annotator tool enables annotation of the intracellular immune receptor repertoire. *Plant Physiol.* **183**, 468–482.

Sun, J., Lu, F., Luo, Y., Bie, L., Xu, L. and Wang, Y. (2023) OrthoVenn3: An integrated platform for exploring and visualizing orthologous data across genomes. *Nucleic Acids Res.* **51**, W397–W403.

Tan, G.Y. and Tan, W.K. (1988) Genetic variation in resistance to vascular-streak dieback in cocoa (Theobroma cacao). *Theor. Appl. Genet.* **75**, 761–766.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.

Tobias, P.A., Edwards, R.J., Surana, P., Mangelson, H., Inácio, V., Céu Silva, M. do, Várzea, V., Park, R.F. and Batista, D. (2022) A chromosome-level genome resource for studying virulence mechanisms and evolution of the coffee rust pathogen Hemileia vastatrix *bioRxiv*, 2022.07.29.502101.

Tobias, P.A. and Guest, D.I. (2014) Tree immunity: growing old without antibodies. *Trends Plant Sci.* **19**, 367–370.

Trisno, J., Reflin, R. and Martinius, M. (2016) Vascular Streak Dieback: Penyakit Baru Tanaman Kakao di Sumatera Barat. *J. Fitopatol. Indones.* **12**, 142.

Weilguny, L., Maio, N. De, Munro, R., Manser, C., Birney, E., Loose, M. and Goldman, N. (2023) Dynamic, adaptive sampling during nanopore sequencing using Bayesian experimental design. *Nat. Biotechnol.* **41**, 1018–1025.

Weyer, A.-L. van de, Monteiro, F., Furzer, O.J., et al. (2019) The Arabidopsis thaliana pan-NLRome. *bioRxiv*, 1–32.

Weyer, A.L. Van de, Monteiro, F., Furzer, O.J., et al. (2019) A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis thaliana. *Cell* **178**, 1260-1272.e14.

White, Bruns T, Lee S and Taylor, J. (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In PCR Protocols: A Guide to Methods and Application., pp. 315–322. San Diego, California: Academic Press.

Xi, Y., Cesari, S. and Kroj, T. (2022) Insight into the structure and molecular mode of action of plant paired NLR immune receptors. *Essays Biochem.* **66**, 513–526.

707

708 ## *Figure legends*

709 Figure 1 Chromsyn (Edwards *et al.*, 2022) synteny plots of *Theobroma cacao* parentally
710 phased genome, haplotype A (top) and B (below). Synteny blocks of collinear "Complete"
711 BUSCO genes (Simão *et al.*, 2015) link scaffolds from adjacent assemblies: blue, same strand;
712 red, inverse strand. Yellow triangles mark "Duplicated" BUSCOs. Filled circles mark telomere
713 predictions from Telociraptor (black) (Edwards, 2023). Assembly gaps are marked as dark red
714 + signs**.**

715

716 Figure 2 Physical locations of predicted NLR genes on the paired chromosomes of *Theobroma*
717 *cacao* (A) haplotype A (red) and (B) haplotype B (yellow) generated using ChromoMap (Anand
718 and Rodriguez Lopez, 2022) in RStudio.

719

720 Figure 3 The predicted NLR gene complement in the parentally phased *Theobroma cacao*
721 genome. The two sets of chromosomes corresponding to (above) Haplotypes A and (below)
722 B were independently classified (Chen *et al.*, 2023) and visualised to present the domain
723 classes using SankeyMatic (Bogart, 2014) including novel integrated domains (IDs) with
724 abbreviations derived from Pfam database (Mistry *et al.*, 2021). NBARC = Nucleotide Binding
725 Domain, TIR = Toll/Interleukin-1 receptor, Rx = Potato CC-NB-LRR protein Rx, CC=Coil-Coil
726 Domain, RPW8 = RESISTANCE TO POWDERY MILDEW 8-like coiled-coil**.**

727

728 Figure 4 Amino acid sequence alignment of an NLR cluster matching the QTL mapped regions
729 at around 29 Mb on chromosome 3 (above) and single NLR homologs at 31 Mb chromosome
730 9 (below) within both parentally inherited alleles, A and B, for *Theobroma cacao* clone 26. All
731 predicted NLRs are annotated as Rx-NLRs (Chen *et al.*, 2023) named for resistance to potato
732 virus X.

733

734 Figure 5 Leaf from *Theobroma cacao*, VSD susceptible clone 26, growing at the Mars Cocoa
735 Research Institute field trials, Pangkep South Sulawesi, Indonesia. (Left) The symptomatic leaf
736 apex dieback and some chlorosis indicating *Ceratobasium theobromae*. (Right) Visible
737 vascular death caused by the pathogen. Scale bar left ~5 cm**.**

738

739 Figure 6 *Theobroma cacao* resistant (S1) and susceptible (RUQ 1347) parental paired-end (PE)
740 Illumina sequence reads mapped to the concatenated Hap A and B progeny chromosomes.
741 The reads that were mapped in greater numbers to chromosomes and contigs were used to
742 determine parental assignment and swapped accordingly.

743 *Abbreviations*

744 NLR – nucleotide binding site leucine-rich repeat gene models
745 LRR – leucine-rich repeat domain
746 GB – gigabyte
747 Gb – Gigabase
748 Mb – Megabase
749 kb – kilobase
750 VSD – vascular streak dieback, causal agent *Ceratobasidium theobromae*
751 PPR – Phytophthora Pod Rot, causal agent *Phytophthora palmivora*