Identification of candidate virulence loci in *Striga hermonthica*, a devastating parasite of African cereal crops

Suo Qiu[1¶#a], James M. Bradley[1¶#b], Peijun Zhang[1], Roy Chaudhuri[1], Mark Blaxter[2,#c], Roger K Butlin[1,3*], Julie D. Scholes[1*]

[1]School of Biosciences, University of Sheffield, Western Bank, Sheffield, S10 2TN, UK.

[2] Institute of Evolutionary Biology, School of Biological Sciences, The University of Edinburgh, Ashworth Laboratories, Charlotte Auerbach Road, Edinburgh, EH9 3FL,UK.

[3]Department of Marine Sciences, University of Gothenburg, S-405 30 Gothenburg, Sweden.

Current address:

[a#]Institute of Evolutionary Biology, School of Biological Sciences, The University of Edinburgh, Ashworth Laboratories, Charlotte Auerbach Road, Edinburgh, EH9 3FL,UK.

[b#]Cell and Systems Biology, University of Toronto, 25 Willcocks St., Toronto, ON M5S 3B2, Canada

[c#]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK.

[¶]Joint first authors

*Corresponding authors: Professor Julie Scholes (E-mail: j.scholes@sheffield.ac.uk); Professor Roger Butlin (r.k.butlin@sheffield.ac.uk).

Correspondence to:

Professor Julie Scholes

School of Biosciences

University of Sheffield

Sheffield, S10 2TN, UK

Abstract

Parasites have evolved suites of proteins, Virulence Factors (VFs), that are delivered into host plants to facilitate colonization. Whilst VFs mediating plant-microbe and plant-nematode interactions have been characterised extensively, less is known about VFs mediating parasitic plant interactions with their hosts. *Striga hermonthica* is an obligate, root-parasitic plant capable of parasitizing multiple cereal hosts in sub-Saharan Africa, causing devastating losses in yields. An understanding of the molecular nature and allelic variation of VFs in *S. hermonthica* is essential for breeding durable resistance and delaying the evolution of parasite virulence. To address this issue, we assembled a genome for *Striga hermonthica* and identified candidate VFs by combining *in silico* prediction of secreted proteins with pooled sequencing of parasites growing on a susceptible and a strongly resistant rice host. Consistent with predictions for parasites, like *S. hermonthica*, that can interact with multiple hosts, we identified multiple loci, potentially with a wide range of functions, implicated in overcoming host resistance. Extremely different allele frequencies were observed at 152 non-secreted and 38 putatively secreted VFs between *S. hermonthica* parasitising the resistant and susceptible rice varieties. Our candidate, secreted VFs encompassed functions such as host cell wall modification, protease inhibitors, oxidoreductase and kinase activities, as well as several with unknown functions. Consistent with maintenance of variation at virulence loci by balancing selection the candidate loci had significantly higher Tajima's D on average than the genomic background. Our results show that diverse strategies are used by *S. hermonthica* to overcome different layers of host resistance. Understanding the maintenance of variation at virulence loci by balancing selection will be critical to managing the evolution of virulence as a part of a sustainable control strategy.

Author statement

The parasitic weed *S. hermonthica* infects the staple cereal crops of sub Saharan Africa inflicting huge agricultural damage on poor subsistence farmers. Understanding coevolution between Striga and its hosts is essential for effective control strategies, yet little is known about the molecular or population genetic basis of virulence variation in this parasite. Here we combine the sequencing of the *S. hermonthica* genome with an experimental population genetic analysis to identify suites of genes that encode candidate virulence factors enabling some Striga individuals to overcome host resistance. This study revealed variation in multiple functional categories of genes that contribute towards parasite virulence. It provides strong evidence that any single mechanism of host resistance will be rapidly overcome by the allelic diversity in the parasite population, and thus highlights the pitfalls of relying on few resistant genotypes. Only by identifying this allelic diversity will it be possible to combine resistance alleles, in host varieties that are suitable for different agro-ecological zones, and in ways that achieve sustained control by delaying the evolution of virulence. Our data provide a key resource to aid researchers, plant breeders and plant pathologists in developing much-needed durable control against this parasitic weed.

Introduction

Plants are constantly challenged by parasites from across all kingdoms of life [1,2]. As a consequence, they have evolved sophisticated surveillance systems to detect and protect themselves against parasite invasion [3,4,5]. In turn, plant parasites have evolved suites of proteins, miRNAs, or other molecules which are delivered into host plants to facilitate colonisation (Virulence Factors (VFs)) [1,2,6,7] and they are pivotal in determining the outcome of a parasite-plant interaction. Despite substantial advances in understanding the

78 identity and mode of action of VFs in plant interactions with fungal, bacterial and nematode
79 parasites [1,6,7] much less is known about VFs mediating parasitic plant interactions with
80 their plant hosts [8-10]. Parasitic plants occur in almost all terrestrial habitats and have
81 evolved independently at least 12 times [8,11,12]. Regardless of evolutionary origin,
82 parasitic plants possess a multicellular organ called the 'haustorium', through which direct
83 structural and physiological connections are formed with their host plant [8,13]. This allows
84 them to abstract water, organic and inorganic nutrients. In addition, the haustorium is
85 increasingly recognised to play a role in host manipulation, through the movement of
86 parasite-derived proteins, miRNAs and other small molecules into the host plant [10,12-
87 16].

88 *Striga* is a genus of obligate, root parasitic plants within the Orobanchaceae [17,18]. One
89 species in particular, *Striga hermonthica* (Del.) Benth., is a notorious parasite of rain-fed
90 rice, maize, sorghum and millets, leading to devastating losses in crop yields for resource-
91 poor farmers in sub-Saharan Africa (SSA) [19,20]. Control of this parasite is extremely
92 difficult as *S. hermonthica* is an obligate outbreeder, with high fecundity, wide dispersal
93 and a persistent, long-lived seed bank [18] leading to a large effective population size [21].
94 Resistant crop varieties are a crucial component of successful control strategies [19] yet,
95 even for crop varieties considered highly resistant, genetic variation within parasite
96 populations is such that a few individuals can overcome host resistance responses and
97 form successful attachments [22,23]. In order to develop crop varieties that display durable
98 resistance against highly polymorphic populations of *S. hermonthica*, it is vital to gain a
99 comprehensive understanding of the repertoire, mode of action and genetic variability of
100 parasite VFs that suppress or circumvent host defence responses [10,24]

101 Here we have combined two complementary approaches to identify candidate genes
102 encoding VFs in *S. hermonthica.* First, we assembled and annotated the genome of *S.*

103 *hermonthica*, and developed a pipeline for computational prediction of putative secreted
104 proteins (the secretome) and candidate VFs. The assembled genome was then used as a
105 reference for an experimental, population genomics analysis, to compare DNA sequence
106 variants in bulked (pooled) samples of *S. hermonthica* grown on a susceptible (NERICA
107 7) or resistant (NERICA 17) rice host (Fig 1a i-ii). This allowed us to scan for loci in the *S.*
108 *hermonthica* genome where the selection imposed by the resistant host had elevated the
109 frequency of alleles contributing to successful colonisation (termed 'virulence' alleles) (Fig
110 1 b-d). A similar approach was used to identify candidate genomic regions associated with
111 resistance in *Solanum vernei* to the potato cyst nematode, *Globodera pallida* [25]. The
112 intersection between genes encoding predicted VFs and genes with highly significant
113 allele frequency differences in the genome scan of *S. hermonthica*, revealed a set of
114 candidate virulence loci encoding proteins with many functions, including cell wall
115 modification, protease, or protease inhibitor, oxidoreductase and putative receptor-like
116 protein kinase activities. Our results show that diverse strategies are used by *S.*
117 *hermonthica* to overcome different layers of host resistance and suggest a polygenic basis
118 of virulence in this parasite.

119 Results

120 The Striga hermonthica genome is very heterozygous

121 We obtained a single population of *S. hermonthica* seeds from farmer's fields in Kibos,
122 Kenya and infected a highly susceptible rice variety, NERICA 7 (Fig 1a). The genetic
123 diversity of the seed population is reflected in the subtle differences of flower colour and
124 morphology of attached parasites (Fig 2a). We sequenced, assembled and characterised
125 the genome of a single individual from this population, which to our knowledge, represents
126 the first genome assembly for *S. hermonthica*. The genome size was estimated by K-mer

127 analysis to be 1,475 Mbp, (Fig 2b). This agrees closely with a previous flow cytometry-
128 based estimate [26] and is more than twice the size of the recently sequenced genome of
129 *S. asiatica* [27]. The assembly consisted of 34,907 scaffolds > 1 kbp in length, with an N50
130 of 10.0 Mbp and 29 scaffolds making up half of the genome size (S1 Table). The *S.*
131 *hermonthica* genome was remarkably heterozygous (overall mean $\pi = 0.011$) (Fig 2 b,c)
132 when compared with other parasitic and non-parasitic plant genomes, likely reflecting the
133 fact that it is an obligate outbreeding species. In addition, the genome contained a large
134 proportion (69%) of repetitive DNA (Fig 2 b,c) which was dominated by long terminal repeat
135 (LTR) elements (S2 Table), a pattern also found for the leafless shoot-parasitic plants,
136 *Cuscuta australis* and *C. campestris* [28, 29] and the closely related parasitic plant *S.*
137 *asiatica* [27]. As expected, the density of repetitive elements along each scaffold negatively
138 correlated with the density of protein-coding genes (Fig 2c). In total, 29,518 protein-coding
139 genes were predicted from the *S. hermonthica* genome, which was comparable to *S.*
140 *asiatica* (34,577), the closely related non-parasitic plant *Mimulus guttatus* (28,140) and to
141 *Arabidopsis thaliana* (27,416) (S3 Table). Through our BUSCO analysis of gene set
142 completeness [30], we found 87.3% of 2,326 conserved single-copy orthologs in
143 eudiocotyledons were complete in the *S. hermonthica* genome, which is very similar to
144 that found in *S. asiatica* (88.7%) (Fig 3; S4 Table). Of the BUSCOs not found in the *S.*
145 *hermonthica* genome, over half were also absent from the *S. asiatica* genome (S4 Table).
146 Both *Striga* spp. share missing BUSCOs that are present in the genome of the closely
147 related non-parasitic plant, *Mimulus guttatus* (Fig 3b; S4 Table). Similarly, two shoot
148 holoparasites, *C. australis* and *C. campestris*, with a BUSCO completeness of 81.0 and
149 81.7% respectively, also shared many missing BUSCOs that were present in the genome
150 of the non-parasitic close relative, *Ipomea nil* (Fig 3c). This is consistent with previous
151 findings suggesting some missing BUSCOs are likely to be a result of the parasitic lifestyle
152 [27-29, 31].

153 **The Striga hermonthica secretome**

154 One way that parasite proteins can interact with host biology is through parasite-directed
155 secretion. We identified 3,375 putatively secreted proteins in *S. hermonthica* (11.4 % of
156 the proteome) (S1 Fig), many of which were homologous to *A. thaliana* secreted proteins
157 (S5 Table), providing experimental evidence for secretion into the extracellular space. On
158 average, the *S. hermonthica* secreted proteins were both significantly smaller and had a
159 higher percentage of cysteine residues compared with the rest of the proteome (Fig 5 a,b).
160 Genes encoding secreted proteins tended to be more clustered (within 15 kbp of their
161 nearest neighbour) compared to all genes in the genome ($p < 10^{-5}$, 10^5 permutations) (S2
162 Fig) suggesting they are likely to be arrayed in tandem and belong to large gene families
163 [35]. Functionally, the secretome was rich in protein domains involved in cell wall
164 modification (e.g. endoglucanases, cellulases, pectinesterases, expansins, and pectate
165 lyases), protease activity (e.g. papain-like cysteine proteases, aspartic proteases, and
166 subtilase proteases) and oxidoreductase activity (peroxidases, copper oxidases, and
167 cytochrome p450 proteins) (Fig 5c, S3 and S4 Figs). The cytochrome P450 domain, for
168 example, was particularly frequent in the *S. hermonthica* secretome (3.13% of protein
169 domains) compared with the rest of the proteome (0.25% of protein domains) (S3 Fig).
170 Three other highly-abundant protein domains in the secretome were described as copper
171 oxidases (S3 Fig) and are commonly found in laccases that are involved in the generation
172 or breakdown of phenolic components, such as lignin [36]. Small cysteine-rich proteins are
173 common characteristics of VFs from a range of phytoparasites [37, 38]. In *S. hermonthica*,
174 183 such proteins were identified (Fig 5a) and were similar to proteins annotated as
175 carbohydrate binding X8 domain-containing proteins, protease inhibitor/lipid transfer
176 proteins, PAR1-like proteins, pectinesterases, RALF-like proteins and thaumatin-like
177 proteins (S4 Fig), many of which are likely to play a role in host-Striga interactions [27,39].

178 We identified several protein domains in the *S. hermonthica* secretome that were enriched
179 to a higher degree than observed in the secretome of the closely-related non-parasitic
180 plant, *M. guttatus* (Fig 5c, S3 Fig , S1 Data), suggesting these functions are relevant to the
181 parasitic lifestyle of *S. hermonthica*. The xyloglucan endotransglycosylase (PF06955)
182 domain, for example, was found in 17 *S. hermonthica* proteins (Fig 5c, S4 Fig). Xyloglucan
183 endotransglucosylases / hydrolases (XETs) have the potential to modify either the parasite
184 or host cell walls (or both) during parasitism [40]. XETs are secreted from the haustoria of
185 the parasitic plant *Cuscuta reflexa* during a susceptible interaction on its host *Pelargonium*
186 *zonale,* and contribute towards the pathogenicity of this parasitic plant [40]. Pectate lyase
187 superfamily (PF12708) and pectinesterase (PF01095) domains were enriched in the
188 secretome of *S. hermonthica* compared to *M. guttatus* and may act as VFs to modify host,
189 or parasite, pectin during penetration. We found a battery of different carbohydrate-active
190 glycosyl hydrolase (GH) domains that were enriched in the *S. hermonthica* secretome (Fig.
191 5c, S3 Fig). Eight *S. hermonthica* proteins were annotated as cellulases of the GH5 family
192 (containing domain PF00150) (S4 Fig) and were similar to secreted cellulases that function
193 as VFs in some phytoparasitic nematodes [41]. The degradation of cellulosic β-1,4-glucans
194 has been observed in susceptible sorghum roots infected by *S. hermonthica* [42] and may
195 be mediated by these secreted enzymes to facilitate the migration of *S. hermonthica*
196 intrusive cells between host root cortical cells. The identification of many putatively
197 secreted VFs in the *S. hermonthica* genome, that are likely to modify host plant cell walls,
198 raises an interesting question about how such proteins are targeted to avoid damaging the
199 parasite's own cell walls.

200 Population genomic analysis to identify candidate virulence loci

201 We next designed an experimental system that would allow us to identify a subset of VFs
202 with genetic variation relevant to the ability to infect some host genotypes and not others.

203 Hundreds of *S. hermonthica* individuals were harvested from either a very resistant
204 (NERICA 17) or susceptible (NERICA 7) rice cultivar, and pools of these individuals were
205 subjected to genome resequencing. After aligning the reads to our reference genome, we
206 detected 1.8 million SNPs in genic regions (defined as regions from 2 kbp upstream of the
207 start codon to the 1 kbp window containing the stop codon). These genic regions were split
208 into 150,741 1 kbp windows and of these, 194 (0.13%) had extreme and consistent allele
209 frequency differences between the bulked pools of *S. hermonthica* selected on the
210 resistant *versus* the susceptible hosts (S5 Fig; S2 Data). These highly differentiated
211 windows were located in 190 genes. These candidate loci potentially encode virulence
212 factors with allelic variants, influencing either structure or expression that contribute to the
213 ability of some individuals to parasitise NERICA 17. As expected for an outbred parasite
214 with a large population that encounters multiple host species and genotypes, many loci
215 were detected and they cover a range of predicted functions. Of these candidate VFs, 152
216 were not predicted to be secreted and were assigned to a wide range of functional
217 categories, including putative transcription factors, hormone signalling pathways,
218 transporters, repeat-containing proteins and a number of proteins of unknown function (Fig
219 6a; S2 Data). Some of these proteins may function to protect the parasite against host
220 defences and facilitate growth on the resistant rice variety. In addition, some may enter
221 the host by non-traditional pathways, for example, via the host-parasite xylem connections.
222 One sixth (24) of these non-secreted proteins had sequence similarity to proteins in the
223 Pathogen-Host Interaction (PHI) database [43]. These included *S. hermonthica* proteins
224 with sequence similarity to a putative leucine-rich repeat protein from *Ralstonia*
225 *solanacearum*, a mitogen-activated protein kinase from *Ustilago maydis*, a calreticulin-like
226 protein from *Magnaporthe oryza* and a cytochrome P450 from *Bursaphelenchus xylophilus*
227 (S2 Data).

228 The remaining 38 VFs were members of the *S. hermonthica* secretome and represent
229 particularly strong candidates associated with the ability to parasitise NERICA 17
230 successfully (Fig 6a,b, S2 Data). These genes were categorised into six functional groups,
231 the largest of which contained 12 genes associated with cell wall modification (Fig 6a,b),
232 including genes encoding an expansin protein, a COBRA-like protein, a β-(1-2)-
233 xylosyltransferase, two trichome birefringence-like (TBL) proteins, a pollen Ole e allergen
234 and two exostosin family proteins, all of which can function to modify the extensibility or
235 other mechanical properties of plant cell walls [2, 44-46] (Fig 6b). Groups of genes
236 annotated as proteases (6 genes including subtilases, aspartyl proteases, and a cysteine
237 proteinase), lipases (3 genes) and kinases (3 genes) were also found. The proteases were
238 always associated with an inhibitor protein domain (Fig 6b). For example, the putative
239 aspartyl proteases possessed one or more xylanase inhibitor domains (Fig 6b). There
240 were also eight genes encoding proteins with a range of different putative functions,
241 including a PAR1-like protein, a probable aquaporin, an α/β-hydrolase and two receptor-
242 like protein kinases (Fig 6b). In addition, a further six genes were annotated as proteins of
243 unknown function (Fig 6b).

244 The 38 candidate VFs were investigated in more detail by quantifying changes in gene
245 expression in haustoria at critical stages of parasite development on the susceptible rice
246 variety NERICA 7, inspecting the distribution of SNPs throughout the promoter and genic
247 regions, and testing for signatures of historical selection. Changes in gene expression
248 were measured in an independent experiment by harvesting *S. hermonthica* haustoria
249 generated *in vitro,* i.e. prior to attachment to the host (Fig 6c). Changes in gene expression
250 of attached haustoria were measured relative to gene expression in haustoria generated
251 *in vitro.* At 2 days after inoculation of the host root, parasite haustoria were attached and
252 parasite intrusive cells had penetrated into the host root cortex. By day 4, the parasite

253 intrusive cells had penetrated between the endodermal cells and by day 7 had formed
254 connections with the xylem vessels of the host, providing direct access to host resources
255 (Fig 1a iii).

256 Prior to attachment to the host, some of the genes encoding candidate VFs were not
257 expressed in haustoria (e.g. subtilase gene (SHERM_16883) and subtilisin-like protease
258 (SHERM_08443) or were expressed at very low levels (e.g the peroxidase
259 (SHERM_00887), glycosyl hydrolase (SHERM_(20042), both aspartyl proteases
260 (SHERM_16482 and SHERM_26730) and an unknown protein (SHERM_03853) (S3
261 Data). However, all 38 genes were expressed in haustoria during the early stages of
262 infection of the susceptible host, NERICA 7 (Fig 6c; S3 Data). There were two main
263 patterns of gene expression. Firstly, 21 genes, including those mentioned above, had low
264 levels of expression in haustoria 2 days post infection, followed by an increase in
265 expression as infection progressed (Fig 6c; S3 Data). In contrast, 17 genes were highly
266 expressed in haustoria 2 days post infection and expression then decreased progressively
267 with time, e.g. genes encoding β-glucosidase, β-(1-2)-xylosyltransferase, and TBL protein
268 SHERM_06484, all of which modify cell walls. The cysteine protease, PAR1, α/β-hydrolase
269 and aquaporin genes also exhibited a similar expression profile (Fig 6c; S3 Data).

270 Most of the 38 genes had significantly differentiating SNPs in their promoter regions (from
271 the start site to 2 kbp up-stream). Some of these SNPs may lead to a change in the
272 regulation of gene expression (Fig 6b). Some genes, for example, the gene encoding the
273 pollen Ole e allergen protein (SHERM_15460), one of the exostosin family proteins
274 SHERM_12118), a probable aquaporin SIP2-1 (SHERM_13450) and one of the two
275 protein TBL genes (SHERM_16484), also had non-synonymous SNPs in the coding region
276 (Fig 6b), that may result in functional differences between the alleles of these genes in

277 individuals infecting NERICA 7 and NERICA 17. Finally, SNPs were also found within
278 predicted intron regions in many of the genes (Fig 6b).

279 The co-evolutionary interactions between hosts and parasites can generate balancing
280 selection [47]. We predicted that genes contributing to virulence would tend to have a
281 history of balancing selection because of the diverse range of hosts used by *S.*
282 *hermonthica*. To test this prediction, we compared Tajima's D between candidate loci and
283 the rest of the genome, expecting to see more positive values for loci with a history of
284 balancing selection [48]. We used the pools from the susceptible host for this comparison
285 because they represented the Striga population as a whole. As predicted, the 152
286 candidate loci in the *S. hermonthica* proteome (S6 Fig) and the 38 candidate loci in the
287 secretome (Fig. 6d) had significantly elevated Tajima's D, on average, compared to all the
288 genes in the genome ($p < 0.0001$ and $p < 0.0003$, respectively; 10^5 permutations). Some
289 loci had particularly high Tajima's D values, for example the two receptor-like protein
290 kinases (Fig. 6b). Interestingly, some loci showed large differences in Tajima's D between
291 the control and virulent *S. hermonthica* pools with the largest difference seen for the TBL
292 gene (SHERM_16484) with a negative ΔD (D_{Vir} - D_{Con}) of -0.9. This suggests strong
293 selection resulting in one common haplotype in the virulent pools in contrast to two or more
294 haplotypes at intermediate frequencies in the control pools. There were also large positive
295 ΔD values: 0.71, 0.16 and 0.20 for one of the putative receptor-like protein kinases
296 SHERM_01541, one of the aspartyl proteases, SHERM_16482, and the peroxidase
297 SHERM_00887, respectively. This suggests that a rare haplotype in the control pools is
298 present at intermediate frequency in the virulent pools. Overall, these changes indicate
299 that selection on the resistant host caused changes in frequency of multi-SNP haplotypes
300 at these loci, haplotypes that may have been created by areas of low recombination or by

301 recent invasion of new variants under positive selection [49] and which underlie the ability
302 of some *S. hermonthica* individuals to overcome resistance in NERICA 17.

303 Discussion

304 Plants secrete proteins involved in many biological functions, from nutrient acquisition, to
305 development and defence [3, 44]. However, unlike most plants, in parasitic plants such as
306 *S. hermonthica* a subset of secreted proteins is likely to function as VFs and contribute
307 towards parasite fitness by facilitating host colonization [10]. We used a combination of *in*
308 *silico* prediction of secreted proteins and pooled sequencing of parasites derived from
309 susceptible and resistant rice hosts, both facilitated by the first available genome
310 assembly, to identify a set of candidate VFs. These are secreted proteins encoded by
311 genes that had extremely different allele frequencies between replicated pools derived
312 from susceptible and resistant hosts, suggesting strong selection for particular variants
313 that facilitate successful colonisation despite host resistance. This experimental approach
314 has not been applied previously to investigate virulence of Striga, or any other parasitic
315 plant. Its success here paves the way to application of similar methods to other host-
316 parasite combinations, providing vital information on virulence mechanisms and their
317 genetic variability within and between parasitic plant populations from different regions of
318 Africa, and so underpinning the development of sustainable control strategies.

319 Our list of 38 candidate, secreted VFs points to key functions involved in pathogenicity,
320 including oxidoreductase, receptor-like protein kinase, protease and protease inhibitor,
321 and cell wall modification activities. The latter is consistent with growing evidence that cell-
322 wall modification is a critical step in plant invasions by many different parasites including
323 parasitic plants. Recently, the structural integrity of lignin was shown to be a crucial
324 component of resistance in roots of the rice variety Nipponbare to infection by *S.*

325 *hermonthica* [50]. In our study it is clear that the host cell wall is involved in resistance in
326 NERICA 17. Most *S. hermonthica* individuals from the Kibos population were unable to
327 penetrate the root endodermis or, if they breached the endodermis, they were unable to
328 establish functional connections to the host xylem vessels (Fig 1a iv-vi). Consistent with
329 this, the largest category of our candidate, secreted VFs included a putative peroxidase,
330 an expansin, pollen allergen-like proteins, a β-glucosidase, a β (1-2) xylosyltransferase,
331 and a TBL protein, all of which function to modify cell walls. The TBL protein,
332 SHERM_16484, had a strikingly different Tajima's D in the control pool compared to the
333 value in the virulent pool. This large difference suggests a change from intermediate to
334 extreme frequencies for multiple SNPs, consistent with selection favouring one haplotype
335 on the resistant NERICA 17, out of several haplotypes present in the population. In *A.*
336 *thaliana* and *O. sativa* TBL proteins belong to large gene families with functions related to
337 cell wall modifications. In *A. thaliana*, At-TBL44 has been implicated in pectin esterification
338 [51, 52], whilst in rice other members of this family appear to be involved in acetylation of
339 xylan moieties in cell walls [53]. In each case, alterations in enzyme activity altered
340 resistance in *A. thaliana* to powdery mildew and in rice to leaf blight [51, 53]. Recently an
341 11 kDa protein was isolated from the cell wall of the shoot parasite *C. reflexa* and identified
342 as a glycine rich protein (GRP) [54]. The protein and its minimal peptide epitope (Crip21)
343 bind to and activate a cell surface resistance gene in tomato (CuRe1), leading to resistance
344 to the parasite, illustrating the importance of cell wall modifications to host resistance.

345 In addition to cell wall modification, several candidate genes were annotated as having
346 protease activity, including two aspartyl proteases, three subtilisin or subtilisin-like genes
347 and a cysteine proteinase. Interestingly, all were predicted to possess a dual-domain
348 structure consisting of a propeptide inhibitor domain and a catalytic protease domain. In
349 other such protease enzymes the propeptide domain auto-inhibits the enzyme activity until

350 cleavage of this inhibitor domain activates the catalytic domain [55]. This provides a
351 mechanism by which the parasite could initially secrete an inactive VF that only becomes
352 active once in the host environment. A similar dual-domain structure was found for a highly
353 expressed, haustorium-specific cysteine protease in the shoot parasitic plant, *C. reflexa*,
354 which positively contributes towards pathogenicity [33]. Although the precise functions of
355 other candidate VFs are unknown, for example the putative aquaporin, PAR1 protein,
356 cytochrome P450 and the 5 proteins with no functional annotation, they provide exciting
357 avenues for further investigation.

358 *S. hermonthica* has extremely high fecundity (>100,000 seeds per plant) [18], a persistent
359 seed bank and is obligate out-crossing [56], leading to a very large effective population
360 size [21]. Therefore, the high heterozygosity that we observed in the *S. hermonthica*
361 genome was not unexpected. *S. hermonthica* parasitizes many different host species and
362 varieties, often within the same geographical area. Populations therefore encounter many
363 different forms of resistance, which they experience as a highly heterogeneous
364 environment. This is expected to maintain genetic diversity at many loci contributing to
365 virulence, which is consistent with observations from field studies that resistant varieties,
366 of any particular crop species, are often parasitized by one or two *S. hermonthica*
367 individuals [22,24]. A typical example is the host-parasite combination used here as a test
368 system; the *S. hermonthica* Kibos population and the strongly resistant upland rice variety,
369 NERICA 17 [23,57], one of 18 NERICA rice varieties grown widely by African farmers [57].

370 This type of parasite interaction with multiple hosts leads to two predictions that are
371 supported by our data. First, multiple loci, potentially with a wide range of functions, are
372 likely to be implicated in overcoming host resistance. We detected 190 strong candidates
373 for contribution to virulence, with extreme allele frequency differences between our control
374 and virulent pools, including many gene families. It is likely that many additional candidate

375 VFs would be revealed, by repeating this comparison on other resistant hosts. An
376 important question for the future will be to determine how individual VFs are implicated in
377 overcoming resistance for specific hosts or across a range of hosts. Second, maintenance
378 of variation at virulence loci by balancing selection will lead to elevated Tajima's D relative
379 to the background, reflecting persistence of multiple alleles at these loci. We found the
380 overall Tajima's D in *S. hermonthica* to be negative, perhaps reflecting population
381 expansion following the spread of agriculture, but our candidate loci had significantly
382 higher Tajima's D on average, consistent with balancing selection on these loci.
383 Understanding the maintenance of variation at virulence loci by balancing selection will be
384 critical to managing the evolution of virulence as a part of a sustainable control strategy
385 [58].

386 Effective control of *S. hermonthica* is essential for food security and poverty alleviation for
387 small-holder subsistence farmers, but it remains elusive. The use of resistance crop
388 varieties is recognised as sustainable and cost effective [19], but the durability of resistant
389 varieties is compromised by the potential for rapid evolution of parasite virulence. Thus,
390 the long-term success of host resistance, as a control strategy for S. *hermonthica* and
391 other parasitic weeds, requires knowledge of the virulence factors involved, their allelic
392 variation within and between Striga populations and their interaction with different host
393 resistance alleles. Only then will it be possible to combine resistance alleles, in host
394 varieties that are suitable for different agro-ecological zones and in ways that achieve
395 sustained control by delaying the evolution of virulence. Our experimental approach and
396 identification of candidate VFs and allelic variation within a *S. hermonthica* population, is
397 a critical first step in this direction.

398 # Materials and Methods

399 ## Collection and extraction of S. hermonthica DNA for genome
400 and pooled sequencing

401 An accession (population sample) of *S. hermonthica* seeds was collected from individuals'
402 parasitising maize in farmers' fields in the Kibos region of Kenya (0° 5' 30.1272" S; 34° 46'
403 4.6416" E). To obtain *S. hermonthica* for genome sequencing and the bulked sample
404 analysis (BSA), rice seedlings of the varieties, NERICA 7 and NERICA 17, were grown in
405 rhizotrons and infected with germinated *S. hermonthica* seeds (to ensure synchronous
406 attachment) as described in [22]. Plants were grown in a controlled environment room with
407 a 12 h photoperiod, a photon-flux density of 500 µmol.quanta.m^{-2}.s^{-1} at plant height, a day
408 / night temperature of 28 / 24 °C and 60 % relative humidity. For the construction of a
409 reference genome, one *S. hermonthica* individual was randomly harvested from NERICA
410 7. For the pooled sequencing, 300 *S. hermonthica* individuals were harvested from
411 NERICA 7 and from NERICA 17. Individual parasites were harvested when > 30 mg in
412 weight, divided into 20 mg aliquots and immediately frozen in liquid nitrogen. The 300
413 individuals collected from NERICA 7 or NERICA 17 were divided into three pools of 100
414 individuals (three biological replicates). Each pool was randomly divided into 25 groups of
415 four individuals (20 mg of tissue per individual) for DNA extraction. To remove high levels
416 of secondary metabolites and polysaccharides in *S. hermonthica*, lysed plant tissues were
417 first washed in a buffer containing 2 % polyvinylpyrrolidone, 0.25 M NaCl, 0.2 M Tris-HCl
418 and 50 mM EDTA. DNA was then extracted following the CTAB protocol [59]. The quality
419 of DNA was checked by gel electrophoresis and quantified by Nanodrop. An equal quantity
420 of DNA from each of the 25 samples was then combined to form one biological replicate.
421 The six DNA samples (three from *S. hermonthica* individuals from NERICA 7 and three
422 from NERICA 17) were then subjected to paired-end sequencing using an Illumina HiSeq

423 machine at the Beijing Genomics Institute (BGI), China. The libraries, insert sizes and
424 sequencing depth are shown in S6 Table. DNA from the individual harvested from NERICA
425 7 for the production of a reference genome was sequenced on an Illumina HiSeq2500
426 sequencer at Edinburgh Genomics, UK. Six paired-end DNA libraries were constructed
427 with different insert sizes as described in S6 Table.

428 De novo assembly of the *S. hermonthica* genome

429 For *de novo* genome assembly, adapters and low quality reads (Q < 20) were removed
430 using CutAdapt [60]. Only reads longer than 50 bp for the paired-end libraries and 30 bp
431 for the mate-pair libraries were retained for subsequent analyses. Duplicated reads
432 generated by PCR amplification in the library construction process were removed using
433 FASTUNIQ [61]. Sequencing errors, which can create difficulties for the short-read
434 assembly, were corrected using the software BLESS, a *k*-mer spectrum-based method
435 designed to remove *k*-mers with a small number of occurrences [62]. After filtering, ~2.7
436 billion reads were generated from the short insert libraries and 0.76 billion reads from
437 mate-pair libraries. This corresponded to ~230 X and ~54 X coverage of the *S.*
438 *hermonthica* genome, respectively.

439 The cleaned and filtered reads were used to assess the *S. hermonthica* genome size,
440 repetitiveness and heterozygosity, compared with 12 other plant species (S6 Table), in the
441 module preQC, implemented in the software sga (https://github.com/jts/sga). The Illumina
442 sequencing reads of these 12 species were downloaded from the NCBI Sequence Read
443 Archive (SRA; https://www.ncbi.nlm.nih.gov/sra/) with the source project numbers given in
444 S6 Table. This analysis showed *S. hermonthica* was highly heterozygous and therefore
445 the software Platanus, which is specifically designed for highly heterozygous genomes,
446 was chosen to assemble the *S. hermonthica* genome [63] (S1 Table).

447 To further improve the *S. hermonthica* genome assembly, Chicago and Dovetail Hi-C
448 libraries were prepared and sequenced at Dovetail Genomics, California, USA
449 (https://dovetailgenomics.com/plant-animal/) (S1 Table). For construction of Chicago
450 libraries, DNA from the same *S. hermonthica* individual (as used for initial sequencing)
451 was sequenced on a standard Illumina HiSeq 2500 platform. For the Hi-C libraries, plant
452 tissues from an F1 individual from a cross between the sequenced individual and another
453 *S. hermonthica* individual (Kibos accession) were used for the library construction and
454 sequencing. Sequences from both the Chicago and Hi-C libraries were then used to
455 improve the contiguity of the initial genome assembly using the Dovetail HiRise Assembler
456 software. RepeatModeler was used to generate a *S. hermonthica*-specific repeat library
457 and RepeatMasker was then used to classify repeat elements in the genome. A repeat-
458 masked version of the genome was used for annotation [64, 65].

459 Annotation of the S. hermonthica genome

460 The genome was annotated using three methods. Firstly, gene structures were inferred
461 using a *S. hermonthica* transcriptome dataset generated by the Parasitic Plant Genome
462 Project (PPGP) [9]. This dataset included Illumina sequencing of cDNAs collected from *S.*
463 *hermonthica* individuals at eight developmental stages [39]. The reads were mapped onto
464 the *S. hermonthica* genome assembly using TopHat to identify exon regions and splice
465 positions [66]. Transcriptome-based gene structures were predicted using Cufflinks [67]
466 and candidate coding regions were then constructed in Transdecoder
467 (https://github.com/TransDecoder/). Secondly, the protein sequences from five species:
468 *Arabidopsis thaliana* (TAIR10), *Mimulus guttatus* (v2.0), *Solanum lycopersicum* (ITAG2.4),
469 *Oryza sativa* (IRGSP1.0) and *Sorghum bicolor* (79), were used to determine consensus
470 gene models in the genome. The protein sequences from rice and sorghum were included
471 because they are natural hosts of *S. hermonthica* and previous studies showed that

472 horizontal gene transfer is common between the host species and *Striga* [68, 69]. The
473 protein sequences were mapped onto the *S. hermonthica* genome using TBLASTN and
474 pairwise alignments were then input into Genewise [70] to predict gene models in *S.*
475 *hermonthica*. Thirdly, an *ab initio* method was used for *de novo* prediction of genes in the
476 *S. hermonthica* genome by the software, Braker [71]. Braker combines the advantages of
477 the two gene prediction software packages: GeneMark-ET and AUGUSTUS, and
478 integrates RNA-Seq read information into final gene predictions [72, 73]. Finally, Evidence
479 Gene Modeler was used to integrate various gene models from the transcript data,
480 mapped proteins, and the predicted gene models from the *ab initio* method [74]. The
481 completeness of the gene set was assessed using BUSCO v5 using the 2,326 core
482 orthologs from eudicots_odb10, with default settings.

483 Functional annotation of the S. hermonthica proteome

484 Putative protein functions were assigned to *S. hermonthica* proteins using BLASTp
485 analyses against the SwissProt and TrEMBL databases, and against the proteomes of
486 *Arabidopsis thaliana* (version 30 downloaded from TAIR) and *Oryza sativa* (version 7
487 downloaded from MSU); which were chosen for their extensive functional annotations. A
488 BLASTp analysis was also conducted against the pathogen-host interaction database
489 (PHI-base, version 4.2), which contains curated biological information on host and parasite
490 genes shown to be involved in pathogen-host interactions (maintained by Rothamsted
491 Research, UK: http://www.phi-base.org/index.jsp). BLASTp analyses were run locally
492 using the NCBI BLAST package (version: ncbi-2.3.0+) and a hit was taken to be significant
493 if e-value < 10^{-5}, bit score > 30 and percentage identity > 30. Protein motifs and domains
494 were determined by searching databases including Pfam, PATHER, GENE3D, CDD,
495 PRINTS, PROSITE, ProDom and SMART with InterProScan (Zdobnov and Apweiler

496 2001). Gene Ontology (GO) terms for individual proteins were retrieved from the
497 corresponding InterPro descriptions.

498 Inference of Orthogroups (OG)

499 Orthologous gene groups (Orthogroups or OGs) among 13 plant species (*Amborella*
500 *trichopoda*, *Musa acuminata*, *Oryza sativa*, *Zea mays*, *Sorghum bicolor*, *Aquilegia*
501 *coerulea*, *Solanum lycopersicum*, *Mimulus guttatus*, *Striga hermonthica*, *Arabidopsis*
502 *thaliana*, *Medicago truncatula*, *Populus trichocarpa* and *Vitis vinifera*) were inferred using
503 the software OrthoFinder v2 [75]. Except for *S. hermonthica*, the proteomes of these
504 species were downloaded from Phytozome V12.1 (https://phytozome.jgi.doe.gov/
505 pz/portal.html). The number of genes per species for each OG was transformed into a
506 matrix of *Z*-scores to quantify gene family expansion / contraction in comparison with other
507 plant genomes. The significance of expansion or contraction was determined using CAFE
508 v4.2 [76]. Specifically, multi-species alignments of 42 single-copy genes from the
509 OrthoFinder results were first constructed using MAFFT v7.407 [77]. A maximum likelihood
510 (ML) tree using the alignments of the 42 single-copy genes was constructed in MEGA X
511 78], then converted into an ultrametric tree and calibrated based on a divergence time of
512 82.8 Ma–127.2 Ma between *A. thaliana* and *P. trichocarpa* [79]. The ultrametric tree and
513 the gene numbers in each orthogroup from the OrthoFinder results were then used for the
514 CAFE analysis. The functional annotation of each OG was predicted based on sequence
515 similarity to the InterPro protein family database where the majority of the proteins in the
516 OG shared the same Pfam annotation.

517 Prediction and analysis of the S. hermonthica secretome

518 Secreted *S. hermonthica* proteins were predicted according to the default thresholds using
519 SignalP3.0 and SignalP4.1 [80, 81] (S1 Fig). SignalP3.0 predicts a secretion signal peptide

520 according to both neural network (SignalP3.0-NN) and hidden Markov model (SignalP3.0-
521 HMM) algorithms, whilst SignalP4.1 uses a neural network algorithm. Putative secreted
522 proteins were searched for transmembrane spanning regions in the remaining portion of
523 the protein sequence (without secretion signal) using TMHMM2.0 [82]. Proteins with a
524 secretion signal but without a predicted transmembrane helix were retained as the
525 'secretome'. Putative secreted proteins were ranked according to the agreement between
526 the different SignalP algorithms: '3' = predicted by all three SignalP algorithms
527 (SignalP3.0-NN, SignalP3.0-HMM and SignalP4.1); '2' = predicted by at least two
528 algorithms and '1' = predicted by one algorithm. TargetP1.1 was used to provide additional
529 predictions on the cellular localization of the putative secreted proteins ('S' = extracellular;
530 'C' = chloroplast; 'M' = mitochondrial) [83]. The closest matching *Arabidopsis thaliana*
531 protein for each *Striga hermonthica* putative secreted protein was queried against the
532 SUBA4 online database, which combines bioinformatic and experimental evidence to
533 predict subcellular localisation of *A. thaliana* proteins (https://suba.live) [84]. The
534 consensus location of each queried protein was recorded and the number of each
535 subcellular location was expressed as a percentage of the total. Using this approach, the
536 most common subcellular location predicted in the secretome was 'extracellular',
537 supporting the validity of our secretome prediction pipeline for *S. hermonthica* (S5 Table).

538 To determine whether a putative secreted protein was expressed during early stages of
539 parasitism a BLASTn search using the protein's coding sequence was conducted against
540 the stage 3 (~ 48 hours post infection) and stage 4 (~ 72 hours post infection) transcriptome
541 dataset downloaded from the PPGP [38]. BLASTn hits were taken to be significant if they
542 returned an e-value < 10^{-10}, percentage identity of > 95 % and a bit score of > 60. To
543 identify enriched Pfam domains in the *S. hermonthica* secretome compared with the rest
544 of the proteome (non-secretome), the frequency of each Pfam domain was first calculated

545 as the number of occurrences of each domain over the total number of domains in the
546 secretome or non-secretome. Enrichment fold for each domain was then calculated as the
547 frequency of hits in the secretome over the frequency in the non-secretome. The
548 significance of enrichment was assessed using a Chi-squared test with a false discovery
549 rate (FDR) correction for multiple testing [85] and was taken to be significant when the
550 corrected p value was < 0.1.

551 To test whether genes in the secretome clustered together in the genome, a permutation
552 test was performed. For each gene, 'MinDis', defined as the distance between the start
553 positions of the focal gene and its nearest neighbouring gene, was calculated. Genes
554 without a neighbouring gene on the same scaffold were excluded. MinDis was also
555 calculated for 10,000 randomly sampled gene sets with the same size as the secretome.
556 The numbers of clustered genes with MinDis falling into a series of distance intervals were
557 counted and were then compared with the numbers in the same distance interval in the
558 10,000 random gene sets.

559 Refining the secretome to select subsets of potential 560 virulence factors

561 The initial secretome was refined into subsets of VFs based on a series of structural and
562 functional characteristics (S1 Fig). First, those putative secreted proteins that met the
563 following criteria were retained: SignalP ranking = 3; TargetP location = S; and have a
564 BLASTn hit against either stage 3 or 4 *S. hermonthica* transcriptome libraries (suggesting
565 gene expression during early stages of parasitism). From these, four subsets of proteins
566 were identified. Subset 1 comprised small, secreted cysteine-rich proteins (SSCRPs) <
567 500 amino acids in length and containing ≥ 4 % cysteine residues. Subset 2 consisted of
568 proteins containing Pfam domains that were significantly enriched in the *S. hermonthica*
569 secretome. Only those Pfam domains that were enriched to a greater extent in the *S.*

570 *hermonthica* secretome compared with their enrichment in the *M. guttatus* secretome
571 (determined in exactly the same way as for *S. hermonthica*) were used to obtain this
572 subset. Subset 3 comprised proteins with no known Pfam domains or orthogroup
573 assignment. Subset 4 contained proteins with similarity to proteins involved in host-
574 pathogen interactions based on having a match against the PHI-base (S1 Fig).

575 ***Identification of candidate virulence loci using pooled sequencing data***

576 The raw sequence reads from the six pools were trimmed to remove low quality bases and
577 adapter sequences using CutAdapt [60]. The cleaned reads were mapped to the *S.*
578 *hermonthica* reference genome using both BWA [86] mem v 0.7.15 and NOVOALIGN
579 (http://www.novocraft.com). Mapped reads were sorted using Samtools [87] and PCR
580 duplicates were removed using Picard (http://broadinstitute.github.io/picard/). Reads
581 around indels were realigned using the modules RealignerTargetCreator and
582 IndelRealigner implemented in gatk [88]. SNPs were called using bcftools
583 (https://samtools. github.io/bcftools/). An R script described in [89] was used to infer the
584 read coverage distribution for each SNP for each pool of sequenced reads, based on
585 three-component mixture models. An upper boundary of the coverage for the intermediate
586 component (single and low copy number regions) was obtained from the fitted distribution
587 for each pool. The high-coverage component is expected to be enriched in repetitive
588 sequences while the low coverage component is enriched in erroneous reads or
589 inadequately sampled genome regions. In order to provide robust data for subsequent
590 analyses, the inference of an allele and its frequency was considered to be reliable only if
591 the following criteria were met. (a) Coverage from each pool was ≥ 4 and $\leq$ the upper
592 boundary inferred from the mixture model distribution. (b) The sum of the coverage across
593 the six DNA pools was ≥ 36. (c) The count of the minor allele frequency, summed over
594 pools, was > 2. (d) The ratio of the coverage of the second highest allele to the total

595 coverage was ≥ 0.1. (e) The ratio of the combined coverage of the third and fourth highest
596 alleles to the total coverage was < 0.02. Only the highest and second highest frequency
597 alleles were considered for the subsequent analyses. Although three or four alleles were
598 present at some positions, the additional alleles were always rare. Cut-offs were
599 determined following exploration of a range of alternative values. They removed only
600 variants that are very unlikely to differ significantly among pools and so they did not
601 influence the ranking of candidate loci.

602 The likelihood of the observed read counts for the two most common alleles, across the
603 six pools was calculated according to equation 3 from Gompert and Buerkle [90] to allow
604 for the two levels of sampling associated with pooled sequencing data (sampling of reads
605 and of individuals). We compared three allele-frequency models for each SNP using the
606 Akaike information criterion (AIC): a null model with a single allele frequency for all pools,
607 a control-virulent model with one frequency for the control pools (from the NERICA 7 host)
608 and one for the virulent pools (from the NERICA 17 host) and a replicate model with a
609 different allele frequency for each of the three pairs of pools (one control and one virulent)
610 that were sequenced together. The control-virulent model was the model of interest while
611 the replicate model was intended to check for consistency across pairs of pools. Therefore,
612 two ΔAIC values were obtained: ΔAICcv = AICnull - AICcontrol-virulent and ΔAICrep =
613 AICcontrol-virulent – AICreplicate. High positive values of ΔAICcv represent better fits than
614 the null model and indicate significant differences between control and virulent pool types.
615 SNPs with positive ΔAICrep values were likely to be affected by artefacts caused by
616 sequencing methods and were excluded from the following analyses. All analysis steps
617 were repeated independently for SNPs based on BWA and NOVOALIGN mapping as
618 recommended by Kofler et al. [91].

619 The effective population size in *Striga* is likely to be large [18] and this is consistent with
620 high diversity in our samples (overall mean $\pi = 0.011$). Therefore, we also expected that
621 linkage disequilibrium would break down quickly. To define a suitable window size to
622 search for regions potentially implicated in virulence, the extent of linkage disequilibrium
623 in *S. hermonthica* was investigated. Firstly, the distribution of the gene lengths, i.e. from
624 the start codon to the stop codon of a gene, for all the genes in the *S. hermonthica* genome
625 was obtained and 4 kbp, which is the 75th percentile of the gene length distribution, was
626 used as the upper limit for the length of region to be considered. The SNP with maximum
627 ΔAICcv within a gene was identified and the region up to 4 kbp from the SNP was divided
628 into eight 500 bp intervals. The mean difference of ΔAICcv, relative to the maximum
629 ΔAICcv, for SNPs in each 500 bp interval was then calculated. The mean difference was
630 typically only elevated in the 500 bp interval containing the SNP with maximum ΔAICcv,
631 so it was clear that linkage disequilibrium above background levels did not normally extend
632 beyond 1 kbp (see S7 Fig). Therefore, 1 kbp windows were used to detect genomic regions
633 potentially associated with virulence on the basis of allele frequency differences between
634 pools from the susceptible and resistant hosts.

635 Regions starting from 5 kbp upstream of the start codon and ending no further than 2 kbp
636 downstream of the stop codon of a gene were divided into 1 kbp-windows and the mean
637 ΔAICcv across all the SNPs in each window was calculated. A permutation test was
638 performed to obtain the probability of observing the mean ΔAICcv value, or higher, for
639 each window based on the distribution of ΔAICcv across the regions as a whole.
640 Specifically, a mean ΔAICcv was calculated for a window containing n SNPs. One hundred
641 thousand samples of n SNPs from the pool of all SNPs within the analysed regions
642 genome-wide, were drawn randomly. For each of these samples, a mean ΔAICcv value
643 was calculated and an expected distribution for the mean in the target window was

644 obtained. The probability that the observed mean ΔAICcv of a window could be matched
645 or exceeded by chance was obtained by counting the number of permuted mean ΔAICcv
646 with values lower than the observed mean ΔAICcv value for that window and then divided
647 by 100,000. Finally, we retained genic regions for which this probability was less than or
648 equal to $2x10^{-5}$ for both the BWA and NOVOALIGN analyses in any window. This cut-off
649 was chosen to provide experiment-wide significance given the number of protein-coding
650 genes in the analysis (29,518). In the secretome, a more relaxed cut-off of $1x10^{-4}$ was
651 used to reflect the prior expectation that the secretome would be enriched with
652 pathogenicity-related genes and the smaller number of genes in this set (3,375). Thirty two
653 genes met this criterion for both Novoalign and BWA (Fig 6; S2 Data). In addition, six
654 genes encoding putative secreted proteins that passed the $1x10^{-4}$ cut-off for either
655 Novoalign or BWA were included in the candidate set because they either contained large
656 numbers of non-synonymous SNPs or contained high impact SNPs that can alter protein
657 structure (e.g. due to protein truncation) (Fig 6; S2 Data).

658 Two population statistics were calculated for each genic region in the control pool using
659 the software Popoolation [92]. These were nucleotide diversity (π) and Tajima's D, a
660 statistic describing the allele frequency spectrum used for testing whether a DNA
661 sequence is evolving under a process that departs from the standard neutral model, such
662 as selection or demographic change [93].

663 Analyses of candidate virulence genes

664 The candidate virulence genes, defined as all genes containing windows that meet the cut-
665 offs for allele-frequency differentiation between control and virulent pool types, were
666 categorised into functional groups based on the annotations of the closest matching
667 homologs from the *A. thaliana* and *O. sativa* proteomes, as well as the Pfam domain
668 annotations. For each gene, the numbers of SNPs were counted for the promoter region

669 (within 2 kbp upstream of the start codon), the intronic region and coding region, and the
670 numbers of non-synonymous SNPs were determined. To quantify the allele frequency
671 differences between control and virulent pools for these candidate virulence genes, the
672 proportion of SNPs with high fixation index (F_{ST}) values in the significant window was
673 calculated. The F_{ST} values between the control and virulent pools for each SNP were
674 calculated using Popoolation2 [94]. The 95th percentile was obtained from the F_{ST} values
675 for all SNPs across the genic regions as a whole. For each genic region, we counted the
676 number of SNPs in its significant window(s) with F_{ST} value higher than the 95th percentile
677 and then divided by the total number of SNPs in the significant window. This provided a
678 measure of differentiation for each candidate virulence gene between the *S. hermonthica*
679 grown on the two host varieties.

680 Expression profiling of candidate virulence genes

681 To determine the expression profiles for candidate virulence genes, an RNA-seq analysis
682 was conducted for *S. hermonthica* (Kibos accession) collected at 2, 4, or 7 days post
683 infection on NERICA 7. Rice plants were grown and infected in the rhizotron system as
684 described previously [22]. *S. hermonthica* were collected from the roots of NERICA 7 by
685 cutting a 1 mm section of rice root around each *S. hermonthica* attachment. In addition,
686 unattached *S. hermonthica* haustoria were induced *in vitro* by the addition of 10 µM DMBQ
687 [95]. For each treatment, four biological replicates were generated. RNA was extracted
688 using the RNeasy plant Minikit (Qiagen) and 2 µg of RNA was treated with DNase to
689 remove DNA contamination (dsDNase, Thermoscientific). The resulting RNA was purified
690 and concentrated using the RNeasy MinElute cleanup kit (Qiagen) and then sent to
691 NovoGene Co. Ltd. (China) for sequencing (https://en.novogene.com/). Cleaned reads
692 were mapped to the *S. hermonthica* genome using Tophat2, version v2.0.12 (default
693 settings except 'mismatch' = 2). Transcript abundance was determined as read count and

694 measured by HTSeq (version 0.6.1) with default settings, except that the '-m' option was
695 set to 'union'. The FPKM values for each gene at each time point were used to calculate
696 a fold change in expression relative to the haustorial sample. To determine whether gene
697 expression changed significantly during infection, a one-way ANOVA was carried out for
698 each gene. If the assumptions of ANOVA were not met, data were log transformed. For
699 each gene, $\log_2$ fold expression values, across the time points, were centred around 0 and
700 scaled by the standard deviation for plotting as a heatmap. Heatmap plotting was
701 conducted using the pheatmap function in R and clustering of genes by expression profiles
702 was conducted using the 'complete' method with the 'Euclidean' distance measure.

703 # Acknowledgements

704 We thank members of the library production, instrumentation and informatics teams at
705 Edinburgh genomics. We also thank Dr Hernan Morales, University of Gothenburg,
706 Sweden for providing the R script used to infer the read coverage distribution for each SNP
707 for each pool of sequenced reads, based on three-component mixture models. We also
708 thank Dr Mamadou Cissoko, University of Sheffield, for help with the production of the
709 transverse sections through rice roots infected with *S. hermonthica*.

710 # Data Availability

711 Raw reads for the pooled *S. hermonthica* sequences and for the *S. hermonthica* genome
712 sequence, the assembled genome sequence and annotations have been submitted to the
713 European Nucleotide Archive (ENA) browser at (http://www.ebi.ac.uk/ena/data/view/)
714 under the following accession numbers: Genome Assembly GCA_902706635; Project ID
715 PRJEB35606; Sample ID ERS4058863 and Contig accession CACSLK010000001-
716 CACSLK010035056.

717 # References

718 1. Win J, Chaparro-Garcia A, Belhaj K, Saunders DGO, Yoshida K, Dong S, et al.
719 Effector Biology of Plant-Associated Organisms: Concepts and Perspectives. Cold
720 Spring Harbor Symposia on Quantitative Biology. 2012;77: 235–247.
721 doi:10.1101/sqb.2012.77.015933

722 2. Mitsumasu K, Seto Y, Yoshida S. Apoplastic interactions between plants and plant
723 root intruders. Frontiers in Plant Science. 2015;6. doi:10.3389/fpls.2015.00617

724 3. Cook DE, Mesarich CH, Thomma BPHJ. Understanding Plant Immunity as a
725 Surveillance System to Detect Invasion. Annual Review of Phytopathology. 2015;53:
726 541–563. doi:10.1146/annurev-phyto-080614-120114

727 4. Wu C-H, Derevnina L, Kamoun S. Receptor networks underpin plant immunity.
728 Science. 2018;360: 1300–1301. doi:10.1126/science.aat2623

729 5. Kanyuka K, Rudd JJ. Cell surface immune receptors: the guardians of the plant's
730 extracellular spaces. Current Opinion in Plant Biology. 2019;50: 1–8.
731 doi:10.1016/j.pbi.2019.02.005

732 6. Giraldo MC, Dagdas YF, Gupta YK, Mentlak TA, Yi M, Martinez-Rocha AL, et al.
733 Two distinct secretion systems facilitate tissue invasion by the rice blast fungus
734 Magnaporthe oryzae. Nature Communications. 2013;4. doi:10.1038/ncomms2996

735 7. Zheng A, Lin R, Zhang D, Qin P, Xu L, Ai P, et al. The evolution and pathogenic
736 mechanisms of the rice sheath blight pathogen. Nature Communications. 2013;4.
737 doi:10.1038/ncomms2427

738 8. Westwood JH, Yoder JI, Timko MP, dePamphilis CW. The evolution of parasitism in
739 plants. Trends in Plant Science. 2010;15: 227–235.
740 doi:10.1016/j.tplants.2010.01.004

741 9. Westwood JH, dePamphilis CW, Das M, Fernández-Aparicio M, Honaas LA, Timko
742 MP, et al. The Parasitic Plant Genome Project: New Tools for Understanding the
743 Biology of Orobanche and Striga. Weed Science. 2012;60: 295–306.
744 doi:10.1614/WS-D-11-00113.1
745 10. Timko MP, Huang K, Lis KE. Host Resistance and Parasite Virulence in Striga–
746 Host Plant Interactions: A Shifting Balance of Power. Weed Science. 2012;60: 307–
747 315. doi:10.1614/WS-D-11-00039.1
748 11. Kuijt J. The Biology of Parasitic Plants. CA: University of California Press,
749 Berkeley.; 1969.
750 12. Clarke CR, Timko MP, Yoder JI, Axtell MJ, Westwood JH. Molecular Dialog
751 Between Parasitic Plants and Their Hosts. Annual Review of Phytopathology.
752 2019;57. doi:10.1146/annurev-phyto-082718-100043
753 13. Yoshida S, Cui S, Ichihashi Y, Shirasu K. The Haustorium, a Specialized Invasive
754 Organ in Parasitic Plants. Annual Review of Plant Biology. 2016;67: 643–667.
755 doi:10.1146/annurev-arplant-043015-111702
756 14. Aly R, Hamamouch N, Abu-Nassar J, Wolf S, Joel DM, Eizenberg H, et al.
757 Movement of protein and macromolecules between host plants and the parasitic
758 weed Phelipanche aegyptiaca Pers. Plant Cell Reports. 2011;30: 2233–2241.
759 doi:10.1007/s00299-011-1128-5
760 15. Westwood JH. The Physiology of the established Parasite-Host association.
761 Parasitic Orobanchaceae. Berlin: Springer-Verlag; 2013.
762 16. Shahid S, Kim G, Johnson NR, Wafula E, Wang F, Coruh C, et al. MicroRNAs from
763 the parasitic plant Cuscuta campestris target host messenger RNAs. Nature.
764 2018;553: 82–85. doi:10.1038/nature25027

765 17. Spallek T, Mutuku M, Shirasu K. The genus Striga: a witch profile: Profile of Striga,
766 the witchweed. Molecular Plant Pathology. 2013;14: 861–869.
767 doi:10.1111/mpp.12058
768 18. Parker C, Riches CR. Parasitic weeds of the world. UK: CAB International; 1993.
769 19. Scholes JD, Press MC. Striga infestation of cereal crops – an unsolved problem in
770 resource limited agriculture. Current Opinion in Plant Biology. 2008;11: 180–186.
771 doi:10.1016/j.pbi.2008.02.004
772 20. Rodenburg J, Demont M, Zwart SJ, Bastiaans L. Parasitic weed incidence and
773 related economic losses in rice in Africa. Agriculture, Ecosystems & Environment.
774 2016;235: 306–317. doi:10.1016/j.agee.2016.10.020
775 21. Huang K, Whitlock R, Press MC, Scholes JD. Variation for host range within and
776 among populations of the parasitic plant Striga hermonthica. Heredity. 2012;108:
777 96–104. doi:10.1038/hdy.2011.52
778 22. Gurney AL, Slate J, Press MC, Scholes JD. A novel form of resistance in rice to the
779 angiosperm parasite Striga hermonthica. New Phytologist. 2006;169: 199–208.
780 doi:10.1111/j.1469-8137.2005.01560.x
781 23. Cissoko M, Boisnard A, Rodenburg J, Press MC, Scholes JD. New Rice for Africa
782 (NERICA) cultivars exhibit different levels of post-attachment resistance against the
783 parasitic weeds Striga hermonthica and Striga asiatica. New Phytologist. 2011;192:
784 952–963. doi:10.1111/j.1469-8137.2011.03846.x
785 24. Rodenburg J, Cissoko M, Kayongo N, Dieng I, Bisikwa J, Irakiza R, et al. Genetic
786 variation and host–parasite specificity of Striga resistance and tolerance in rice: the
787 need for predictive breeding. New Phytol. 2017;214: 1267–1280.
788 doi:10.1111/nph.14451

789 25. Eoche-Bosy D, Gautier M, Esquibet M, Legeai F, Bretaudeau A, Bouchez O, et al.
790 Genomic scans on experimentally evolved populations reveal candidate regions for
791 adaptation to plant resistance in the potato cyst nematode Globodera pallida.
792 Molecular Ecology. 2017;26: 4700-4711. doi.org/10.1111/mec.14240
793 26. Estep MC, Gowda BS, Huang K, Timko MP, Bennetzen JL. Genomic
794 Characterization for Parasitic Weeds of the Genus by Sample Sequence Analysis.
795 The Plant Genome Journal. 2012;5: 30. doi:10.3835/plantgenome2011.11.0031
796 27. Yoshida S, Kim S, Wafula EK, Tanskanen J, Kim Y-M, Honaas L, et al. Genome
797 Sequence of Striga asiatica Provides Insight into the Evolution of Plant Parasitism.
798 Current Biology. 2019;29: 3041-3052.e4. doi:10.1016/j.cub.2019.07.086
799 28. Sun G, Xu Y, Liu H, Sun T, Zhang J, Hettenhausen C, et al. Large-scale gene
800 losses underlie the genome evolution of parasitic plant Cuscuta australis. Nature
801 Communications. 2018;9. doi:10.1038/s41467-018-04721-8
802 29. Vogel A, Schwacke R, Denton AK, Usadel B, Hollmann J, Fischer K, et al.
803 Footprints of parasitism in the genome of the parasitic flowering plant Cuscuta
804 campestris. Nature Communications. 2018;9. doi:10.1038/s41467-018-04344-z
805 30. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et
806 al. BUSCO Applications from Quality Assessments to Gene Prediction and
807 Phylogenomics. Molecular Biology and Evolution. 2018;35: 543–548.
808 doi:10.1093/molbev/msx319
809 31. Cai L, Arnold BJ, Xi Z, Khost DE, Patel N, Hartmann CB, et al. Deeply Altered
810 Genome Architecture in the Endoparasitic Flowering Plant Sapria himalayana Griff.
811 (Rafflesiaceae). Current Biology. 2021;31: 1002-1011.e9.
812 doi:10.1016/j.cub.2020.12.045

813 32. Toh S, Holbrook-Smith D, Stogios PJ, Onopriyenko O, Lumba S, Tsuchiya Y, et al.
814 Structure-function analysis identifies highly sensitive strigolactone receptors in
815 Striga. Science. 2015;350: 203–207. doi:10.1126/science.aac9476
816 33. Bleischwitz M, Albert M, Fuchsbauer H-L, Kaldenhoff R. Significance of Cuscutain,
817 a cysteine protease from Cuscuta reflexa, in host-parasite interactions. BMC Plant
818 Biology. 2010;10: 227.
819 34. Mueller AN, Ziemann S, Treitschke S, Aßmann D, Doehlemann G. Compatibility in
820 the Ustilago maydis–Maize Interaction Requires Inhibition of Host Cysteine
821 Proteases by the Fungal Effector Pit2. Xu J-R, editor. PLoS Pathogens. 2013;9:
822 e1003177. doi:10.1371/journal.ppat.1003177
823 35. Elizondo LI, Jafar-Nejad P, Clewing JM, Boerkoel CF. Gene Clusters, Molecular
824 Evolution and Disease: A Speculation. Current Genetics. 2009;12:64-75.
825 36. Kwiatos, N., Ryngajłło, M. & Bielecki, S. Diversity of laccase-coding genes in
826 Fusarium oxysporum genomes. Frontiers in Microbiology 6, (2015).
827 37. Saunders DGO, Win J, Cano LM, Szabo LJ, Kamoun S, Raffaele S. Using
828 Hierarchical Clustering of Secreted Protein Families to Classify and Rank Candidate
829 Effectors of Rust Fungi. Stajich JE, editor. PLoS ONE. 2012;7: e29847.
830 doi:10.1371/journal.pone.0029847
831 38. Lu S, Edwards MC. Genome-Wide Analysis of Small Secreted Cysteine-Rich
832 Proteins Identifies Candidate Effector Proteins Potentially Involved in Fusarium
833 graminearum−Wheat Interactions. Phytopathology. 2016;106: 166–176.
834 doi:10.1094/PHYTO-09-15-0215-R
835 39. Yang Z, Wafula EK, Honaas LA, Zhang H, Das M, Fernandez-Aparicio M, et al.
836 Comparative Transcriptome Analyses Reveal Core Parasitism Genes and Suggest

837 Gene Duplication and Repurposing as Sources of Structural Novelty. Molecular
838 Biology and Evolution. 2015;32: 767–790. doi:10.1093/molbev/msu343
839 40. Olsen S, Krause K. Activity of xyloglucan endotransglucosylases/hydrolases
840 suggests a role during host invasion by the parasitic plant Cuscuta reflexa. van
841 Damme EJM, editor. PLOS ONE. 2017;12: e0176754.
842 doi:10.1371/journal.pone.0176754
843 41. Smant G, Stokkermans JPWG, Yan Y, de Boer JM, Baum TJ, Wang X, et al.
844 Endogenous cellulases in animals: Isolation of -1,4-endoglucanase genes from two
845 species of plant-parasitic cyst nematodes. Proceedings of the National Academy of
846 Sciences. 1998;95: 4906–4911. doi:10.1073/pnas.95.9.4906
847 42. Olivier A, Benhamou N, Leroux GD. Cell surface interactions between sorghum
848 roots and the parasitic weed Striga hermonthica : cytochemical aspects of cellulose
849 distribution in resistant and susceptible host tissues. Canadian Journal of Botany.
850 1991;69: 1679–1690. doi:10.1139/b91-213
851 43. Winnenburg R, Urban M, Beacham A, Baldwin TK, Holland S, Lindeberg M, et al.
852 PHI-base update: additions to the pathogen host interaction database. Nucleic Acids
853 Research. 2007;36: D572–D576. doi:10.1093/nar/gkm858
854 44. Li Y. BRITTLE CULM1, Which Encodes a COBRA-Like Protein, Affects the
855 Mechanical Properties of Rice Plants. THE PLANT CELL ONLINE. 2003;15: 2020–
856 2031. doi:10.1105/tpc.011775
857 45. Qin L, Kudla, U, Roze, A.H. E, Goverse A, Popeijus H, Nieuwland J, et al. A
858 nematode expansin acting on plants. Nature Brief Communications. 2004;427: 30.
859 46. Honaas LA, Wafula EK, Yang Z, Der JP, Wickett NJ, Altman NS, et al. Functional
860 genomics of a generalist parasitic plant: Laser microdissection of host-parasite

861 interface reveals host-specific patterns of parasite gene expression. BMC Plant
862 Biology. 2013;13: 9. doi:10.1186/1471-2229-13-9
863 47. Frank SA. Coevolutionary genetics of plants and pathogens. Evolutionary Ecology.
864 1993;7: 45–75.
865 48. Charlesworth D. Balancing Selection and Its Effects on Sequences in Nearby
866 Genome Regions. PLoS Genetics. 2006;2: e64. doi: 10.1371/journal.pgen.0020064.
867 49. Cutter AD, Payseur BA. Genomic signatures of selection at linked sites: unifying the
868 disparity among species. Nat Rev Genet. 2013;14: 262–274. doi:10.1038/nrg3425
869 50. Mutuku JM, Cui S, Hori C, Takeda Y, Tobimatsu Y, Nakabayashi R, et al. The
870 structural integrity of lignin is crucial for resistance against Striga hermonthica
871 parasitism in rice. Plant Physiology. 2019;179: 1796-1809. doi:10.1104/pp.18.01133
872 51. Vogel JP, Raab TK, Somerville CR, Somerville SC. Mutations in PMR5 result in
873 powdery mildew resistance and altered cell wall composition: PMR5 is required for
874 powdery mildew susceptibility. The Plant Journal. 2004;40: 968–978.
875 doi:10.1111/j.1365-313X.2004.02264.x
876 52. Bacete L, Mélida H, Miedes E, Molina A. Plant cell wall-mediated immunity: cell wall
877 changes trigger disease resistance responses. The Plant Journal. 2018;93: 614–
878 636. doi:10.1111/tpj.13807
879 53. Gao Y, He C, Zhang D, Liu X, Xu Z, Tian Y, et al. Two Trichome Birefringence-Like
880 Proteins Mediate Xylan Acetylation, Which Is Essential for Leaf Blight Resistance in
881 Rice. Plant Physiology. 2017;173: 470–481. doi:10.1104/pp.16.01618
882 54. Hegenauer V, Slabby P, Körner M, Bruckmüller J-A, Burggraf R, Albert I, et al. The
883 tomato receptor CuRe1 senses a cell wall protein to identify *Cuscuta* as a pathogen.
884 Nat Comm. 2020; 11: 5299. https://doi.org/10.1038/s41467-020-19147-4

885 55. Shindo T, Van Der Hoorn RAL. Papain-like cysteine proteases: key players at
886 molecular battlefields employed by both plants and their invaders. Molecular Plant
887 Pathology. 2007;9: 5299. doi:10.1111/j.1364-3703.2007.00439.x
888 56. Safa SB, Jones BMG, Musselman LJ. Mechanisms favoring outbreeding in Striga
889 hermonthica[Scrophulariaceae]. New Phytologist. 1984;96: 299–305.
890 57. Rodenburg J, Cissoko M, Kayeke J, Dieng I, Khan ZR, Midega CAO, et al. Do
891 NERICA rice cultivars express resistance to Striga hermonthica (Del.) Benth. and
892 Striga asiatica (L.) Kuntze under field conditions? Field Crops Research. 2015;170:
893 83–94. doi:10.1016/j.fcr.2014.10.010
894 58. Mikaberidze A, McDonald BA, Bonhoeffer S. Developing smarter host mixtures to
895 control plant disease. Plant Pathol. 2015;64: 996–1004. doi:10.1111/ppa.12321
896 59. Stewart CN, Via LE. A rapid CTAB DNA isolation technique useful for RAPD
897 fingerprinting and other PCR applications. Biotechniques. 1993;14: 748–750.
898 60. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing
899 reads. EMBnet.journal17.1 2011; doi:https://doi.org/10.14806/ej.17.1.200
900 61. Xu H, Luo X, Qian J, Pang X, Song J, Qian G, et al. FastUniq: A Fast De Novo
901 Duplicates Removal Tool for Paired Short Reads. PLoS ONE. 2012;7: e52249.
902 doi:10.1371/journal.pone.0052249
903 62. Heo Y, Wu X-L, Chen D, Ma J, Hwu W-M. BLESS: Bloom filter-based error
904 correction solution for high-throughput sequencing reads. Bioinformatics. 2014;30:
905 1354–1362. doi:10.1093/bioinformatics/btu030
906 63. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, et al. Efficient
907 de novo assembly of highly heterozygous genomes from whole-genome shotgun
908 short reads. Genome Research. 2014;24: 1384–1395. doi:10.1101/gr.170720.113
909 64. Smit A, Hubley R. RepeatModeler Open-1.0. http://www.repeatmasker.org. 2008.

910 65. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. http://www.repeatmasker.org.
911 2013.
912 66. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-
913 Seq. Bioinformatics. 2009;25: 1105–1111. doi:10.1093/bioinformatics/btp120
914 67. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene
915 and transcript expression analysis of RNA-seq experiments with TopHat and
916 Cufflinks. Nature Protocols. 2012;7: 562–578. doi:10.1038/nprot.2012.016
917 68. Yoshida S, Maruyama S, Nozaki H, Shirasu K. Horizontal Gene Transfer by the
918 Parasitic Plant Striga hermonthica. Science. 2010;328: 1128–1128.
919 doi:10.1126/science.1187145
920 69. Davis CC, Xi Z. Horizontal gene transfer in parasitic plants. Current Opinion in Plant
921 Biology. 2015;26: 14–19. doi:10.1016/j.pbi.2015.05.008
922 70. Birney E. GeneWise and Genomewise. Genome Research. 2004;14: 988–995.
923 doi:10.1101/gr.1865504
924 71. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1:
925 Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and
926 AUGUSTUS: Table 1. Bioinformatics. 2016;32: 767–769.
927 doi:10.1093/bioinformatics/btv661
928 72. Lomsadze A, Burns PD, Borodovsky M. Integration of mapped RNA-Seq reads into
929 automatic training of eukaryotic gene finding algorithm. Nucleic Acids Research.
930 2014;42: e119–e119. doi:10.1093/nar/gku557
931 73. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically
932 mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24:
933 637–644. doi:10.1093/bioinformatics/btn013

934 74. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated
935 eukaryotic gene structure annotation using EVidenceModeler and the Program to
936 Assemble Spliced Alignments. Genome Biology. 2008;9: R7. doi:10.1186/gb-2008-9-
937 1-r7
938 75. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome
939 comparisons dramatically improves orthogroup inference accuracy. Genome
940 Biology. 2015;16. doi:10.1186/s13059-015-0721-2
941 76. Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW. Estimating Gene Gain and
942 Loss Rates in the Presence of Error in Genome Assembly and Annotation Using
943 CAFE 3. Molecular Biology and Evolution. 2013;30: 1987–1997.
944 doi:10.1093/molbev/mst100
945 77. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7:
946 Improvements in Performance and Usability. Molecular Biology and Evolution.
947 2013;30: 772–780. doi:10.1093/molbev/mst010
948 78. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary
949 Genetics Analysis across Computing Platforms. Battistuzzi FU, editor. Molecular
950 Biology and Evolution. 2018;35: 1547–1549. doi:10.1093/molbev/msy096
951 79. Clarke JT, Warnock RCM, Donoghue PCJ. Establishing a time-scale for plant
952 evolution. New Phytologist. 2011;192: 266–301. doi:10.1111/j.1469-
953 8137.2011.03794.x
954 80. Bendtsen J D, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal
955 peptides: SignalP 3.0. Journal of Molecular Biology. 2004;340: 783–795.
956 81. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal
957 peptides from transmembrane regions. Nature Methods. 2011;8: 785.

958 82. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane
959 protein topology with a hidden markov model: application to complete genomes.
960 Edited by F. Cohen. Journal of Molecular Biology. 2001;305: 567–580.
961 doi:10.1006/jmbi.2000.4315
962 83. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting Subcellular
963 Localization of Proteins Based on their N-terminal Amino Acid Sequence. Journal of
964 Molecular Biology. 2000;300: 1005–1016. doi:10.1006/jmbi.2000.3903
965 84. Hooper CM, Castleden IR, Tanz SK, Aryamanesh N, Millar AH. SUBA4: the
966 interactive data analysis centre for Arabidopsis subcellular protein locations. Nucleic
967 Acids Res. 2017;45: D1064–D1074. doi:10.1093/nar/gkw1041
968 85. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and
969 Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society:
970 Series B (Methodological). 1995;57: 289–300. doi:10.1111/j.2517-
971 6161.1995.tb02031.x
972 86. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler
973 transform. Bioinformatics. 2010;26: 589–595. doi:10.1093/bioinformatics/btp698
974 87. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
975 Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.
976 doi:10.1093/bioinformatics/btp352
977 88. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine
978 A, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis
979 Toolkit Best Practices Pipeline: The Genome Analysis Toolkit Best Practices
980 Pipeline. In: Bateman A, Pearson WR, Stein LD, Stormo GD, Yates JR, editors.
981 Current Protocols in Bioinformatics. Hoboken, NJ, USA: John Wiley & Sons, Inc.;
982 2013. p. 11.10.1-11.10.33. doi:10.1002/0471250953.bi1110s43

983 89. Morales HE, Faria R, Johannesson K, Larrson T, Panova M, Westram A M, et al.
984 Genomic architecture of parallel ecological divergence: Beyond a single
985 environmental contrast. Science Advances. 2019; eaav9963.
986 90. Gompert Z, Buerkle CA. A Hierarchical Bayesian Model for Next-Generation
987 Population Genomics. Genetics. 2011;187: 903–917.
988 doi:10.1534/genetics.110.124693
989 91. Kofler R, Langmuller AM, Nouhaud P, Otte KA, Schlotterer C. Suitability of Different
990 Mapping Algorithms for Genome-wide Polymorphism Scans with Pool-Seq Data.
991 Genes|Genomes|Genetics. 2016;6: 3507–3515. doi:10.1534/g3.116.034488
992 92. Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A, et al.
993 PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation
994 Sequencing Data from Pooled Individuals. Kayser M, editor. PLoS ONE. 2011;6:
995 e15925. doi:10.1371/journal.pone.0015925
996 93. Tajima F. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA
997 Polymorphism. 1989;123: 585-595.
998 94. Kofler, R, Pandey, RV & Schlötterer, C. PoPoolation2: identifying differentiation
999 between populations using sequencing of pooled DNA samples (Pool-Seq).
1000 Bioinformatics. 2011b; 27: 3435–3436. doi:10.1093/bioinformatics/btr589
1001 95. Fernández-Aparicio M, Huang K, Wafula EK, Honaas LA, Wickett NJ, Timko MP, et
1002 al. Application of qRT-PCR and RNA-Seq analysis for the identification of
1003 housekeeping genes useful for normalization of gene expression values during
1004 Striga hermonthica development. Molecular Biology Reports. 2013;40: 3395–3407.
1005 doi:10.1007/s11033-012-2417-y
1006 96. Hellsten U, Wright KM, Jenkins J, Shu S, Yuan Y, Wessler SR, et al. Fine-scale
1007 variation in meiotic recombination in Mimulus inferred from population shotgun

1008 sequencing. Proceedings of the National Academy of Sciences. 2013;110: 19478–
1009 19482. doi:10.1073/pnas.1319032110
1010 97. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The
1011 Arabidopsis Information Resource (TAIR): improved gene annotation and new tools.
1012 Nucleic Acids Research. 2012;40: D1202–D1210. doi:10.1093/nar/gkr1090
1013 98. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, et al. The TIGR Rice
1014 Genome Annotation Resource: improvements and new features. Nucleic Acids
1015 Research. 2007;35: D883–D887. doi:10.1093/nar/gkl976

1016 Supporting information

1017 Supplementary Figures S1 – S7.pdf

1018 **S1 Fig. Three step pipeline to predict the *Striga hermonthica* secretome and**
1019 **subsets of candidate pathogenicity-related genes.** Step 1: Three SignalP algorithms
1020 were used to identify *S. hermonthica* proteins with a predicted secretion signal at their N-
1021 terminus. These proteins were then checked for a transmembrane helix in the remaining
1022 part of the protein and only retained if no transmembrane domains were identified. Step
1023 2: A range of structural and functional information was gathered for putative secreted
1024 proteins. Step 3: the secretome was refined into sets of candidate pathogenicity-related
1025 genes using the information gleaned in step 2. In addition to the three secretome
1026 subsets, a stand-alone BLASTp analysis was carried out against the pathogen-host
1027 interaction database.

1028 **S2 Fig.Testing gene clustering in the secretome.** The red dot represents the number
1029 of genes in the secretome with MinDis falling into each 5Kb distance interval. The upper
1030 and lower limits of the boxes are the first and third quartiles in the 10,000 permutations in
1031 each distance interval; medians of the data are shown as bands in the boxes; the whiskers

1032 represent the smallest and biggest datum that are still within 1.5 times interquartile range
1033 of the lower and upper quartile, and the outliers are shown as black dots.

1034 **S3 Fig. Relative abundance of Pfam domains in the *Striga hermonthica* secretome**
1035 **or in the rest of the proteome (non-secretome).** The relative abundance of each Pfam
1036 domain is plotted as a percentage of all Pfam domains found in the secretome or the non-
1037 secretome. The 45 Pfam domains with the highest difference in relative abundance
1038 between the two data sets are shown.

1039 **S4 Fig. Functional categorization of 4 subsets of proteins selected from the *Striga***
1040 ***hermonthica* secretome.** *S. hermonthica* proteins were assigned into functional groups
1041 based on the closest BLASTp match with *Arabidopsis thaliana*. If no clear annotation was
1042 available, the closest match with *Oryza sativa* was used. If no clear matches were obtained
1043 for either *A. thaliana* or *O. sativa,* the Pfam domains were used to infer function. Subset 1:
1044 small, cysteine rich proteins. Subset 2: proteins with Pfam domains enriched in the
1045 secretome of *S. hermonthica* to a greater degree than in the secretome of *Mimulus*
1046 *guttatus*. Subset 3: proteins with no known PFAM domain or orthogroup assignment.
1047 Subset 4: proteins with a BLASTp match against a protein in the pathogen-host interaction
1048 database. For subsets 1, 2 and 4, all *S. hermonthica* protein functions which were only
1049 represented once were categorized as 'other'. Subset 3 only contained 25 *S. hermonthica*
1050 proteins and therefore all protein functions were shown, even if they only occurred once.
1051 Venn diagram shows the overlap between the four subsets of secreted *S. hermonthica*
1052 proteins.

1053 **S5 Fig. |** Mean ΔAICcv in relation to the numbers of SNPs in 1kb windows in genic regions.
1054 ΔAICcv is a measure of the magnitude, and consistency across pools, of the allele
1055 frequency difference between *S. hermonthica* growing on resistant NERICA 17 (virulent)
1056 and those growing on susceptible NERICA 7 (control). Significant windows in genic regions

1057 contributing to the secretome (red) and non-secretome (green) were identified by
1058 permutation.

1059 **S6 Fig. Comparison of Tajima's D for the 152 putative VFs (green) and all the genes**
1060 **in the genome (grey) for the control pools.** The 152 candidate loci in the proteome had
1061 significantly elevated D on average ($p < 0.0001$ 10^5 permutations) compared to the rest of
1062 the genome.

1063 **S7 Fig. Distribution of mean ΔAICcv difference to the maximum ΔAICcv ratios in**
1064 **each distance interval.** The upper and lower limits of the boxes are the first and third
1065 quartiles of the ratios in each distance interval; medians of the data are shown as bands;
1066 means are shown in red dots; the whiskers represent the smallest and biggest datum that
1067 are still within 1.5 times interquartile range of the lower and upper quartile, and the outliers
1068 are shown as black dots.

1069 Supplementary Tables 1 – 7.xlsx

1070 **S1 Table. Summary statistics for the Striga hermonthica genome assembly.** Only
1071 contigs / scaffolds with length ≥ 1000 bp are considered). Final assembly was generated
1072 using Chicago and HiC scaffolding, which allowed substantial improvement in the N50.

1073 **S2 Table. Repeat elements identified in the *Striga hermonthica* genome.**

1074 **S3 Table. Comparison of the *Striga hermonthica* genome annotation with other**
1075 **plant species.**

1076 **S4 Table. BUSCO completeness analysis using 2,326 core orthologous genes for**
1077 **eudicots (version: eudicots_odb10).** The parasitic plant genomes have lower BUSCO
1078 completeness compared with their non-parasitic close relatives. *Striga hermonthica*
1079 BUSCO completeness is very similar to that obtained for the recently assembled *S.*
1080 *asiatica* genome.

1081 **S5 Table.** (a) The *Striga hermonthica* putative secreted, or non-secreted proteins with a
1082 hit against an *Arabidopsis thaliana* protein were first identified and then used as the input
1083 into the SUBA4 online database (https://suba.live) to identify the consensus subcellular
1084 location based on bioinformatic and experimental evidence. (b) The number of *A.*
1085 *thaliana* proteins for each subcellular location was recorded and presented as a
1086 percentage of the total.

1087 **S6 Table.** Sequencing information for the *Striga hermonthica* reference genome and
1088 pooled samples for BSA analysis.

1089 **S7 Table.** Plant species included in the analysis of genome size, heterozygosity and
1090 repetitiveness.

1091 Supplementary Data

1092 **S1 Data.xlsx** is associated with Fig 4 and S3 Fig. The file contains all data associated
1093 with the Pfam enrichment analysis of the secretome of *Striga hermonthica* and *Mimulus*
1094 *guttatus*.

1095 **S2 Data.xlsx** is associated with Fig 6. The file contains the list of 152 non secreted and
1096 38 secreted virulence factors with Gene IDs, protein lengths, cysteine % content, details
1097 of SNPs in different regions of the genes, analysis of Tajima's D and Pfam and functional
1098 annotation.

1099 **S3 Data.xlsx** is associated with Fig 6. The file contains the FPKM values used for
1100 calculation of gene expression changes in *S. hermonthica* haustoria following infection of
1101 the rice variety NERICA 7 for the 152 candidate, non-secreted and 38 secreted virulence
1102 factors, together with gene IDs and annotation.

1103

1104

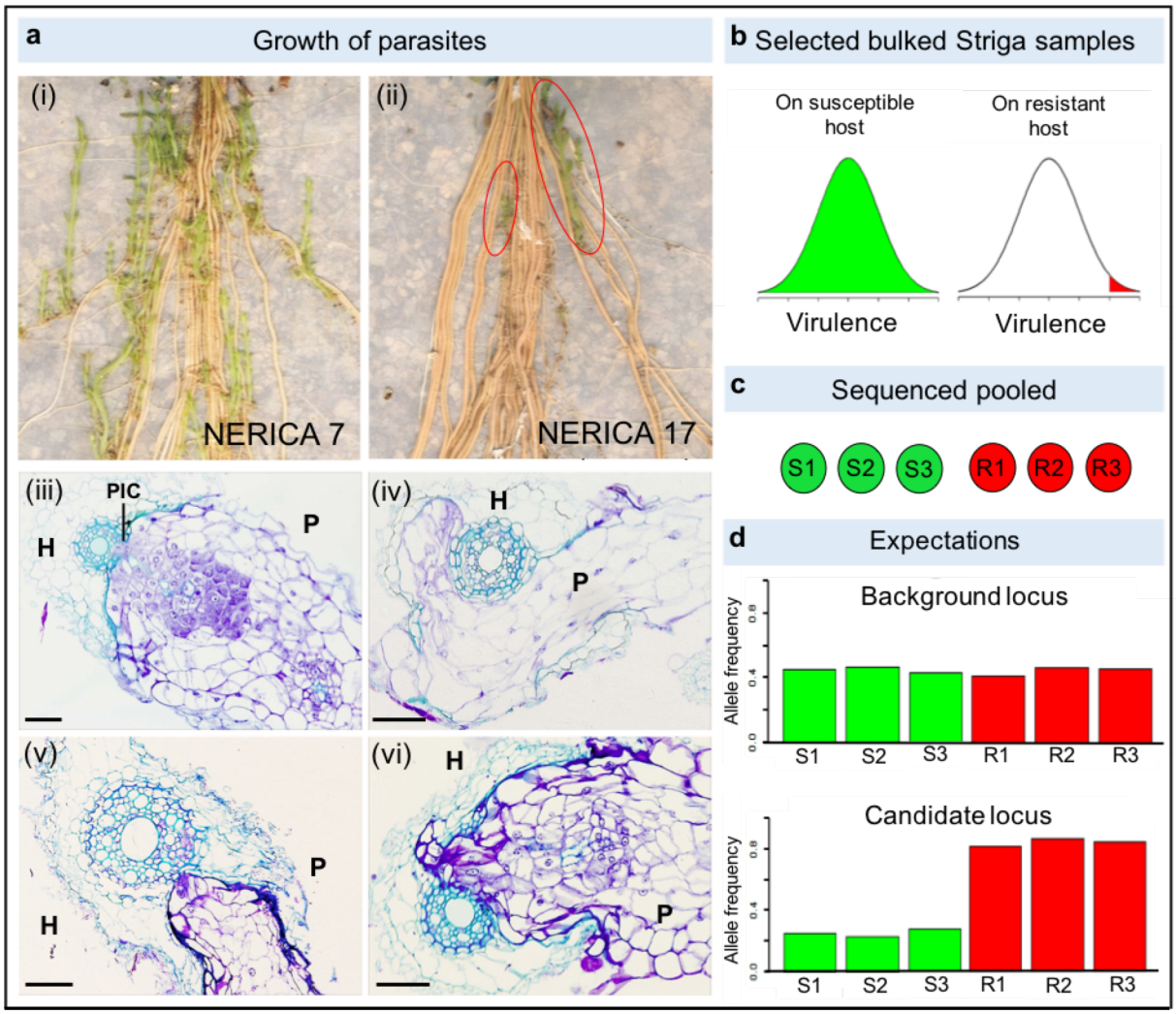


Fig. 1. | Experimental strategy for the identification of *Striga hermonthica* virulence loci. *Striga hermonthica* (Kibos accession) were grown on susceptible (NERICA 7) and resistant (NERICA 17) rice hosts (**a**). The whole rice root systems show many *S. hermonthica* individuals parasitising the roots of NERICA 7 (i) whilst only two individuals (red circles) were able to overcome the resistance response of NERICA 17 (ii) Scale = 1 cm. Transverse sections show *S. hermonthica* invading rice roots for a representative susceptible (iii) and resistant (iv–vi) interaction seven days post inoculation. In the successful host-parasite interaction parasite intrusive cells (PIC) have breached the endodermis and have made connections with the host's xylem (iii). In the resistant rice variety several phenotypes are observed; The parasite invades the host root cortex but is unable to penetrate the suberized endodermis (iv, v); the parasite penetrates the endodermis but is unable to form connections with the host xylem (v). H = host root. P = parasite. Scale = 5 µm. Our experimental strategy was based on the prediction that many *S. hermonthica* genotypes would grow on NERICA 7 but only highly virulent genotypes would grow on NERICA 17 (**b**). Samples of 100 *S. hermonthica* plants were bulked to generate three sequencing pools from each host variety (**c**). We expected that background loci would not differ in allele frequency between pools, but virulence alleles (and neutral alleles in linkage disequilibrium) would have increased frequency in all pools from the resistant host, allowing us to identify candidate loci (**d**).

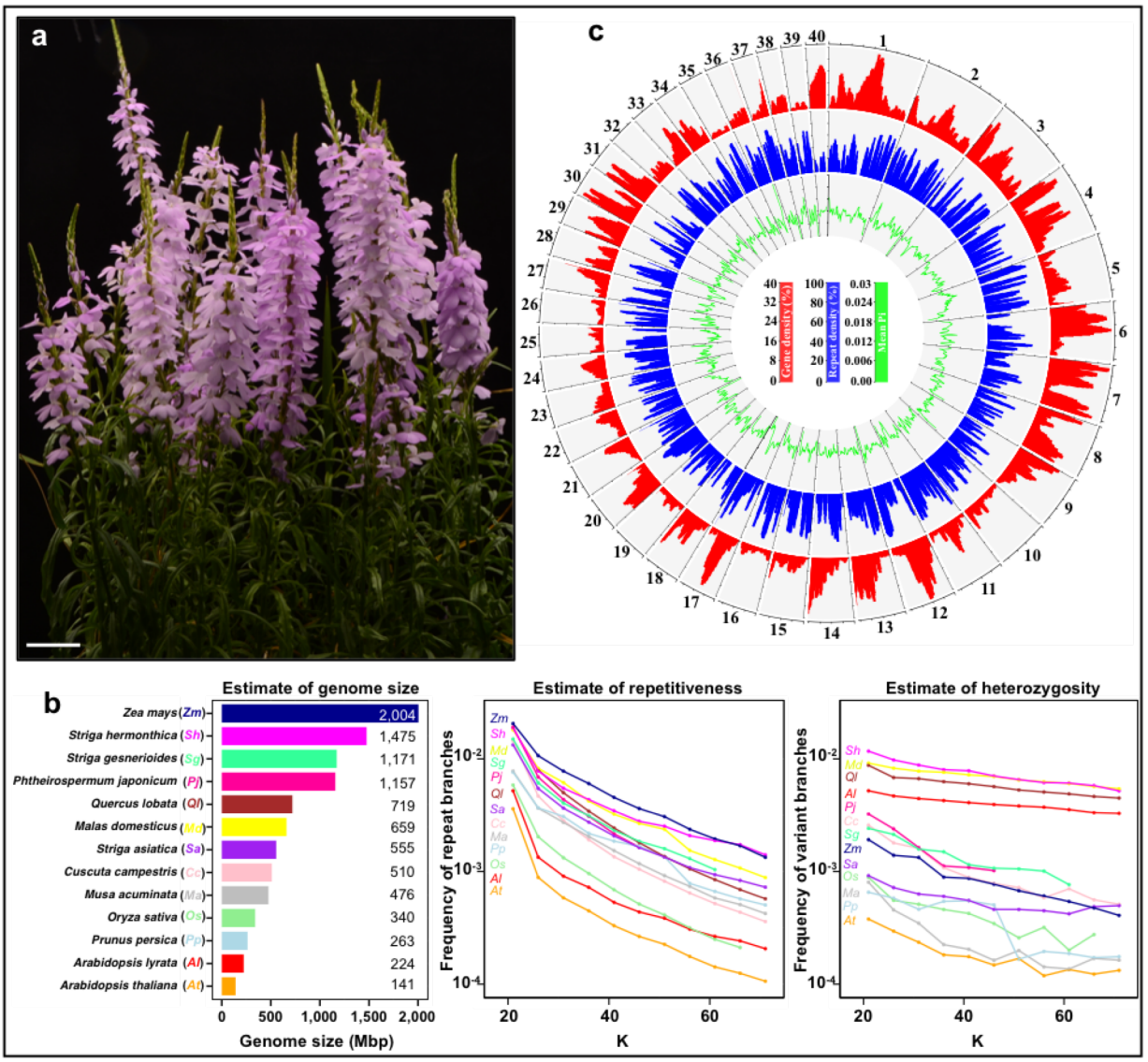


Fig. 2. | *Striga hermonthica* is an obligate outbreeding parasitic plant with a highly heterozygous and repetitive genome. a, Flowering *S. hermonthica* growing on the rice host, NERICA 7, derived from a seed batch collected from the Kibos region of Kenya. Scale = 5 cm. **b,** Comparison of genome size, heterozygosity and repetitiveness between *S. hermonthica* and 12 other plants. The estimate of the genome size (Mbp) was based on k-mer count statistics. The estimate of heterozygosity was based on variant branches in the k-de Bruijn graph. The repetitiveness of the genomes was based on frequency of repeat branches in the k-de Bruijn graph. K: k-mer length. **c,** Genomic features calculated in 1 Mbp windows with a slide of 250 kbp for the largest 40 scaffolds in the *S. hermonthica* genome assembly. Outer bar plot (red): gene density (percentage of the window comprised of genic regions). Mid bar plot (blue): repeat density (percentage of window comprised of repetitive sequence). Inner line plot (green): nucleotide diversity (mean Pi for genic regions). Axes tick marks around plot circumference denote 4 Mbp. Vertical axis tick marks are defined in the centre.

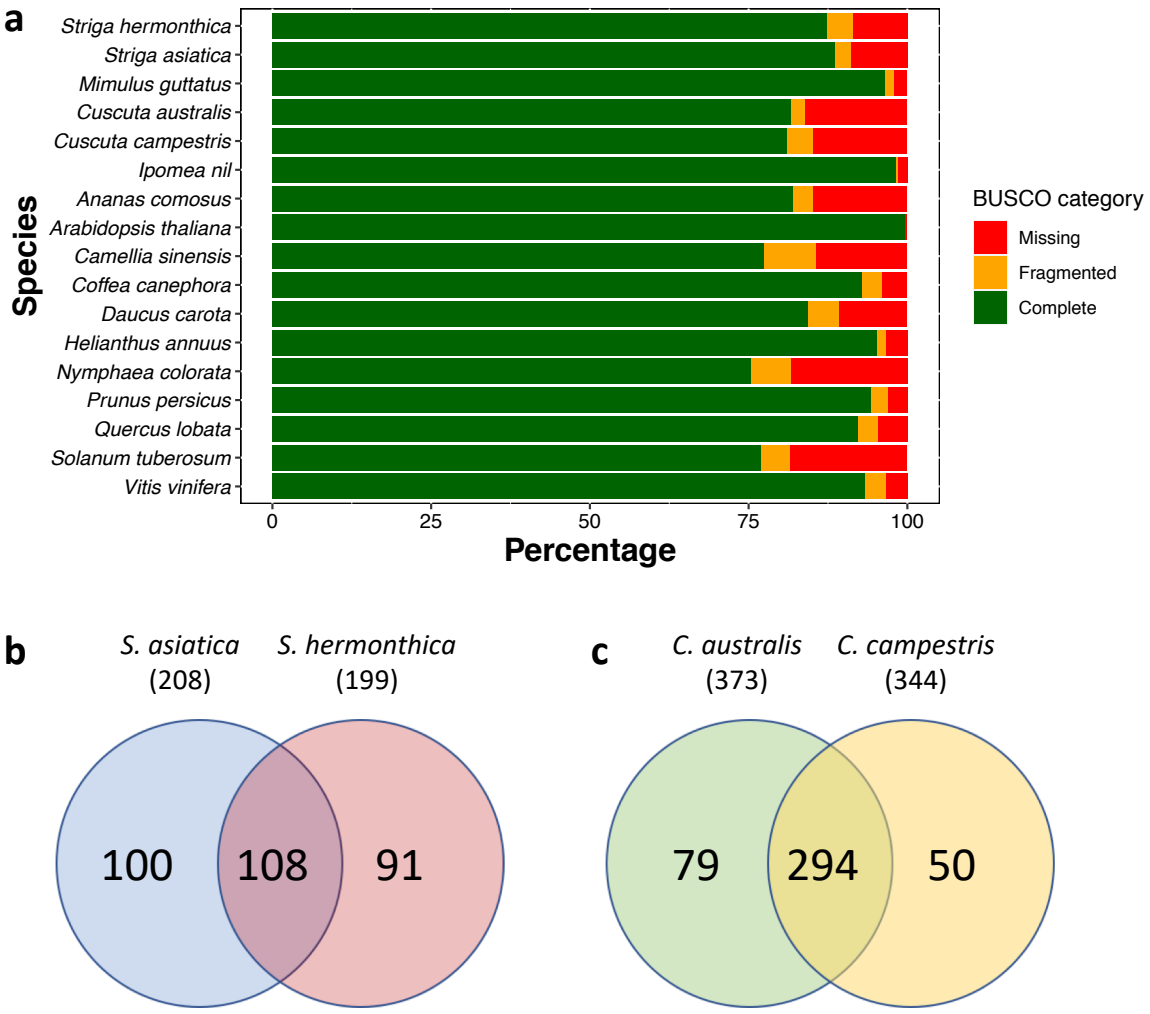
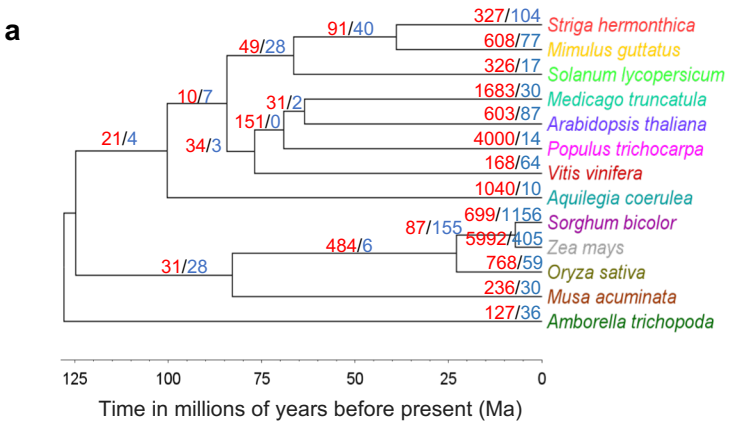


Fig. 3 | a, BUSCO completeness analysis for *Striga hermonthica* genome, compared with 16 other published plant genomes. The number of missing BUSCOs for two Striga **b** and two Cuscuta species **c**. The overlap shows genes that are missing from both Striga or Cuscuta species respectively.



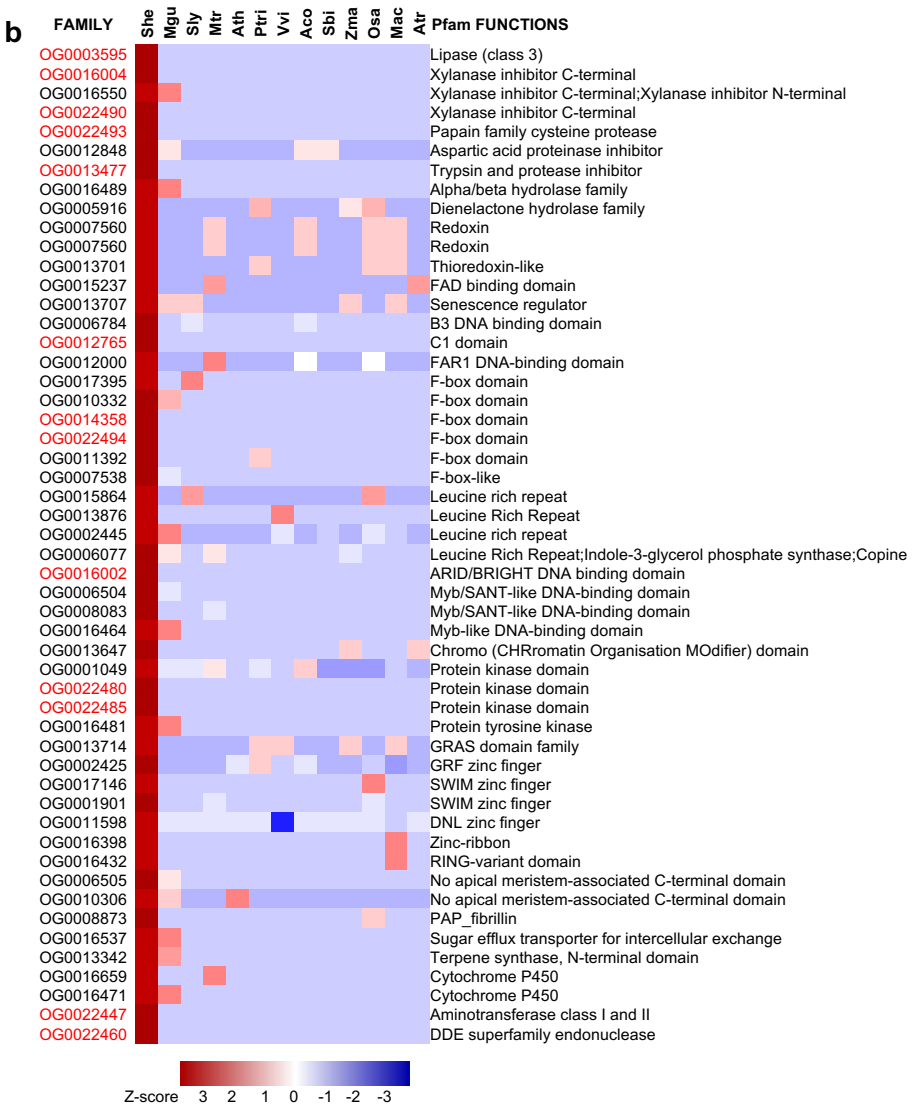
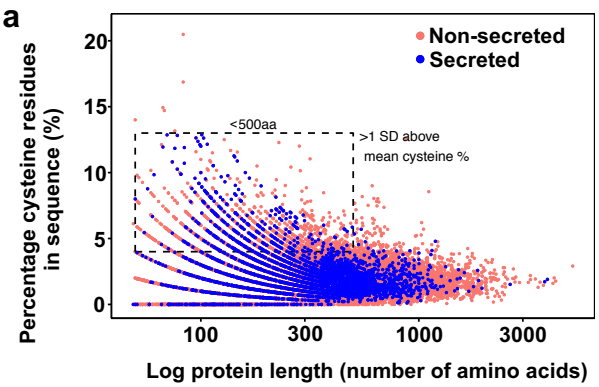


Fig. 4. | Orthogroup analyses. a A time tree for *S. hermonthica* and 12 other species generated in MEGA, based on 42 single-copy genes inferred from OrthoFinder. The number of significantly expanded (red) and contracted (blue) orthogroups based on CAFE analysis are shown above the branches. **b** Significantly expanded orthogroups in *S. hermonthica*, after removing proteins encoded as transposable elements, compared to 12 other plant species. Orthogroups only found in *S. hermonthica*, have family names in red. Higher Z-scores indicate the orthogroups are more expanded in a species while lower Z-scores indicate the orthogroups are more contracted in a species.

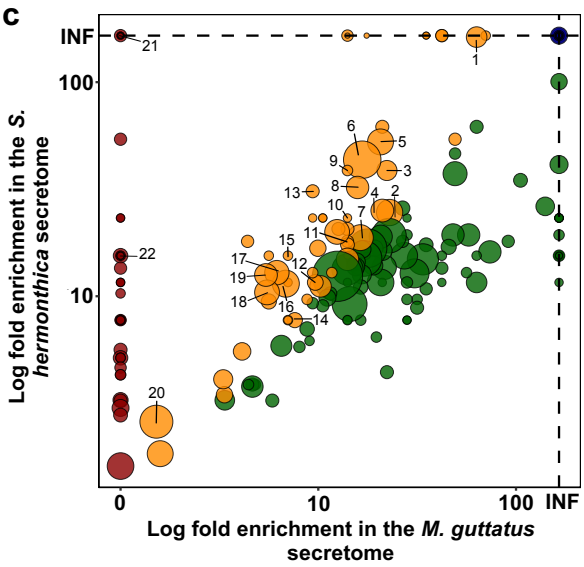


b

		Secretome	Non-secretome
Number of proteins		3,375	26,143
Protein length (aa)	Mean	377 [a]	400 [a]
	Standard error	4.24	1.81
	Minimum	49	49
	Maximum	3,890	5,081
Cysteine content (%)	Mean	2.28 [b]	1.91 [b]
	Standard error	0.03	0.01
	Minimum	0	0
	Maximum	12.87	20.48

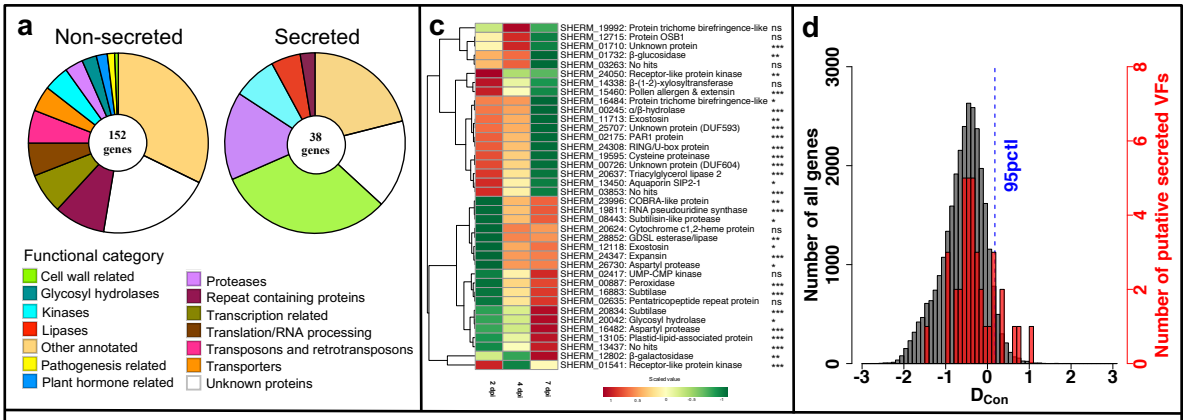
[a] Mean protein length was significantly shorter for secreted proteins compared with non-secreted proteins (Welch two sample t-test: -5.03, df = 4700, p-value = $5.052e^{-7}$).

[b] Secreted proteins had a significantly greater % cysteine content compared with non-secreted proteins (Welch two sample t-test: t = 11.77, df = 3844, p-value < $2.2e^{-16}$)



Key	Pfam accession	Pfam domain description
1	PF02469:	Fasciclin domain
2	PF14368:	Probable lipid transfer
3	PF00759:	Glycosyl hydrolase family 9 (cellulase)
4	PF00722:	Glycosyl hydrolases family 16
5	PF05938:	Plant self-incompatibility protein S1
6	PF04043:	Plant invertase/pectin methylesterase inhibitor
7	PF00190:	Cupin
8	PF06955:	Xyloglucan endo-transglycosylase (XET) C-terminus
9	PF01650:	Peptidase C13 family
10	PF05686:	Glycosyl transferase family 90
11	PF03662:	Glycosyl hydrolase family 79-- N-terminal domain
12	PF00150:	Glycosyl hydrolase family 5 (cellulase)
13	PF00188:	Cysteine rich secretory protein family
14	PF00112:	Papain family cysteine protease
15	PF00704:	Glycosyl hydrolases family 18
16	PF01095:	Pectinesterase
17	PF14416:	PMR5 N terminal Domain
18	PF13839:	GDSL/SGNH-like Acyl-Esterase family in Pmr5 & Cas1p
19	PF01501:	Glycosyl transferase family 8
20	PF08263:	Leucine rich repeat-terminal domain
21	PF12708:	Pectate lyase superfamily protein
22	PF10502:	Signal peptidase-- peptidase S26

Fig. 5. | *Striga hermonthica* secretome. a, Relationship between protein length (log scale) and cysteine content (as a % of total amino acid number) for putatively-secreted (blue) and non-secreted (red) proteins in the *S. hermonthica* proteome. Secreted proteins < 500 amino acids in length and with a cysteine % > 1 standard deviation above the mean, were selected as a subset of small, cysteine rich proteins. **b,** Descriptive statistics for length and cysteine content for secreted and non-secreted proteins. **c,** Pfam domains enrichment (log fold-change) in the *S. hermonthica* secretome, relative to the proteome as a whole, compared to the corresponding enrichment in the *Mimulus guttatus* secretome. INF denotes infinite enrichment (Pfam domain only found in the secretome). Points above the 1:1 diagonal were enriched more in the *S. hermonthica* secretome relative to *M. guttatus* and have been coloured accordingly. Red symbol: domains only enriched in the *S. hermonthica* secretome. Yellow symbol: domains enriched more in the *S. hermonthica* secretome than in the *M. guttatus* secretome. Green symbol: domains enriched more in the *M. guttatus* secretome than in the *S. hermonthica* secretome. Blue symbol: domains present only in the secretome in both species. Sizes of the points were weighted according to the frequency of occurrence of each Pfam domain in the *S. hermonthica* secretome. Annotations for the most significantly enriched of the Pfam domains ($p < 0.01$) that were also enriched more in the *S. hermonthica* secretome relative to the *M. guttatus* secretome, are given in the accompanying table with their functional descriptions.



b

Gene ID	Proportion of differentiating SNPs	SNP presence P	SNP presence NS	SNP presence I	Tajima's D Con	Tajima's D Vir	Annotation of closest matching *Arabidopsis thaliana* protein (BLASTp)	Pfam domain(s) descriptions
SHERM_12802	0.3	x		x	-0.286	-0.030	β-galactosidase	Galactose binding lectin domain; Glycosyl hydrolase family 35
SHERM_14338	0.12	x	x	x	-0.412	-0.359	β-(1-2)-xylosyltransferase	Protein of unknown function (DUF563)
SHERM_01732	0.18			x	-0.371	-0.131	β-glucosidase	Glycosyl hydrolase family 1
SHERM_23996	0.41	x		x	-0.755	-0.639	COBRA-like protein	COBRA-like protein
SHERM_12118	0.21		x		-0.936	-0.664	Exostosin	Exostosin family
SHERM_11713	0.24	x			-0.069	-0.153	Exostosin	Exostosin family
SHERM_24347	0.36	x	x	x	-0.326	-0.239	Expansin	Pollen allergen; Lytic transglycolase
SHERM_20042	0.38	x			-0.623	-0.600	Glycosyl hydrolase	Glycosyl hydrolase family 3 C-terminal; Glycosyl hydrolase family 3 N-terminal
SHERM_15460	0.28	x	x		-0.577	-0.433	Pollen allergen & extensin	Pollen proteins Ole e I like
SHERM_00887	0.33	x		x	-0.426	-0.223	Peroxidase	Peroxidase
SHERM_19992	0.39	x		x	-0.556	-0.473	Protein trichome birefringence-like	PMR5 N-terminal; GDSL/SGNH-like Acyl-Esterase found in Pmr5 & Cas1p
SHERM_16484	0.12	x	x		0.107	-0.793	Protein trichome birefringence-like	PMR5 N-terminal; GDSL/SGNH-like Acyl-Esterase found in Pmr5 & Cas1p
SHERM_00245	0.23	x		x	-0.396	-0.351	α/β-hydrolase	α/β-hydrolase family
SHERM_20624	0.23		x	x	-0.162	-0.061	Cytochrome c1,2-heme protein	Cytochrome C1 family
SHERM_02175	0.29	x			-0.705	-0.645	PAR1 protein	PAR1 protein
SHERM_13450	0.5	x	x		0.108	0.286	Aquaporin SIP2-1	None
SHERM_13105	0.37	x		x	0.096	-0.004	Plastid-lipid-associated protein	PAP_fibrillin
SHERM_12715	0.16	x		x	0.152	0.237	Protein OSB1	None
SHERM_24308	0.36	x		x	-0.305	-0.231	RING/U-box protein	Ring finger
SHERM_19811	0.4	x			0.714	0.752	RNA pseudouridine synthase	RNA pseudouridylate synthase
SHERM_19595	0.21	x		x	0.302	0.408	Cysteine proteinase	Papain family cysteine protease; Cathepsin propeptide inhibitor domain (I29)
SHERM_16482	0.34	x	x	x	-0.482	-0.319	Aspartyl protease	Xylanase inhibitor N-terminal
SHERM_26730	0.67	x	x		-0.450	-0.527	Aspartyl protease	Xylanase inhibitor N-terminal; Xylanase inhibitor C-terminal
SHERM_20834	0.14	x		x	-0.450	-0.527	Subtilase	Subtilase family; Peptidase inhibitor I9
SHERM_16883	0.17	x			0.398	0.369	Subtilase	Subtilase family; Peptidase inhibitor I9; PA domain
SHERM_08443	0.75		x		-1.407	-1.374	Subtilisin-like protease	Subtilase family; Peptidase inhibitor I9; PA domain
SHERM_24050	0.22	x	x	x	0.659	0.699	Receptor-like protein kinase	Protein kinase; Protein tyrosine kinase; Carbohydrate-binding protein of the ER
SHERM_01541	0.33	x	x		1.003	1.713	Receptor-like protein kinase	Protein kinase; Protein tyrosine kinase; Carbohydrate-binding protein of the ER
SHERM_02417	1	x		x	-0.509	-0.583	UMP-CMP kinase	Adenylate kinase
SHERM_28852	0.44	x			-0.262	0.208	GDSL esterase/lipase	GDSL-like Lipase/Acylhydrolase
SHERM_20637	0.17	x		x	-0.668	-0.669	Triacylglycerol lipase 2	Partial α/β-hydrolase lipase region; Putative esterase
SHERM_02635	0.5	x			-0.240	-0.306	Pentatricopeptide repeat protein	PPR repeat; PPR repeat family
SHERM_03263	0.36	x		x	-0.921	-0.931	Unknown protein	None
SHERM_13437	0.39	x			-0.525	-0.725	Unknown protein	None
SHERM_03853	0.19	x			-0.154	-0.251	Unknown protein	None
SHERM_25707	0.38	x			-0.934	-0.946	Unknown protein (DUF593)	Zein-binding
SHERM_00726	0.16			x	-0.424	-0.626	Unknown protein (DUF604)	Protein of unknown function (DUF604)
SHERM_01710	0.24	x	x	x	-0.582	-0.671	Unknown protein	None

Fig. 6. | Identification of *Striga hermonthica* genes that display significant allele frequency differences between pools of individuals parasitising the susceptible rice variety (NERICA 7) and those that successfully parasitise the resistant rice variety (NERICA 17). a Functional categorisation of non-secreted proteins and secreted, candidate virulence factors (VFs). **b** The 38 genes encoding putative secreted *S. hermonthica* proteins with their associated measure of differentiation (proportion of differentiating SNPs within the significant window) between the control and virulent sets of pools. The presence of SNPs in the promoter region (P), non-synonymous SNPs in the coding region (NS) and those in the intronic regions (I) are indicated with an X. The annotation of the closest matching *Arabidopsis thaliana* protein is shown along with coloured boxes that correspond to the functional category assigned in the pie chart in **a**. Tajima's D was calculated for individuals grown on NERICA 7 (Con) or NERICA 17 (Vir). **c**. Clustered gene expression profiles of the 38 candidate VFs in *S. hermonthica* haustoria parasitising NERICA 7 at 2, 4 and 7 days post-inoculation (dpi). Log_2 fold change in expression is shown relative to expression levels in haustoria induced *in vitro*. The gene IDs and putative functions based on best BLASTp hit against the *A. thaliana* proteome correspond with part **b**. Significant changes in gene expression in haustoria during the infection time course are shown *** ($p < 0.001$); ** ($p < 0.01$); * ($p < 0.05$); ns non-significant (ANOVA). **d.** Comparison of Tajima's D for the 38 putative VFs (red) and all the genes in the genome (grey) for the control pools.