



This is a repository copy of *Identification of candidate virulence loci in Striga hermonthica, a devastating parasite of African cereal crops.*

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/193655/>

Version: Submitted Version

---

**Preprint:**

Qiu, S. [orcid.org/0000-0003-4913-2015](https://orcid.org/0000-0003-4913-2015), Bradley, J.M. [orcid.org/0000-0001-9579-6803](https://orcid.org/0000-0001-9579-6803), Zhang, P. et al. (4 more authors) (Submitted: 2022) Identification of candidate virulence loci in *Striga hermonthica*, a devastating parasite of African cereal crops. [Preprint - bioRxiv] (Submitted)

<https://doi.org/10.1101/2022.01.13.476148>

---

**Reuse**

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.



[eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk)  
<https://eprints.whiterose.ac.uk/>

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

3

4 Suo Qiu<sup>1a\*\*</sup>, James M. Bradley<sup>1b\*\*</sup>, Peijun Zhang<sup>1</sup>, Roy Chaudhuri<sup>1</sup>, Mark Blaxter<sup>2,c</sup>,  
5 Roger K. Butlin<sup>1,3\*</sup>, Julie D. Scholes<sup>1\*</sup>

6 <sup>1</sup>School of Biosciences, University of Sheffield, Western Bank, Sheffield, S10 2TN, UK.

7 <sup>2</sup> Institute of Evolutionary Biology, School of Biological Sciences, The University of  
8 Edinburgh, Ashworth Laboratories, Charlotte Auerbach Road, Edinburgh, EH9 3FL, UK.

9 <sup>3</sup>Department of Marine Sciences, University of Gothenburg, S-405 30 Gothenburg,  
10 Sweden.

11

12 Current address:

13 <sup>a</sup>Institute of Evolutionary Biology, School of Biological Sciences, The University of  
14 Edinburgh, Ashworth Laboratories, Charlotte Auerbach Road, Edinburgh, EH9 3FL, UK.

15 <sup>b</sup>Cell and Systems Biology, University of Toronto, 25 Willcocks St., Toronto, ON M5S  
16 3B2, Canada

17 <sup>c</sup>Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10  
18 1SA, UK.

19

20 \*\*Contributed equally to the work

21

22 \*Corresponding authors:

23 **Professor Julie Scholes**

24 Tel: +44 07557766335; Email: [j.scholes@sheffield.ac.uk](mailto:j.scholes@sheffield.ac.uk)

25 **Professor Roger Butlin**

26 Tel: +44 0114 222 0097

27 Email: [r.k.butlin@sheffield.ac.uk](mailto:r.k.butlin@sheffield.ac.uk).

## 28 Summary

- 29
- 30 • Parasites have evolved proteins, Virulence Factors (VFs), that facilitate plant  
31 colonization, yet VFs mediating parasitic plant-host interactions are poorly  
32 understood. *Striga hermonthica* is an obligate, root-parasitic plant of cereal hosts  
33 in sub-Saharan Africa, causing devastating yield losses. Understanding the  
34 molecular nature and allelic variation of VFs in *S. hermonthica* is essential for  
35 breeding resistance and delaying the evolution of parasite virulence.
  - 36 • We assembled the *S. hermonthica* genome and identified secreted proteins by *in*  
37 *silico* prediction. Pooled sequencing of parasites growing on a susceptible and a  
38 strongly resistant rice host allowed us to scan for loci where selection imposed by  
39 the resistant host had elevated the frequency of alleles contributing to successful  
40 colonisation.
  - 41 • Thirty-eight putatively secreted VFs had extremely different allele frequencies with  
42 functions including host cell wall modification, protease inhibitors, oxidoreductase  
43 and kinase activities. These candidate loci had significantly higher Tajima's D than  
44 the genomic background, consistent with balancing selection.
  - 45 • Our results reveal diverse strategies used by *S. hermonthica* to overcome different  
46 layers of host resistance. Understanding the maintenance of variation at virulence  
47 loci by balancing selection will be critical to managing the evolution of virulence  
48 as a part of a sustainable control strategy.

49 **Key words:** Parasitic plants; *Striga hermonthica*; Virulence Factors (VFs); Striga  
50 genome; secretome; population genomics.

## 52 Introduction

53 Plants are constantly challenged by parasites from across all kingdoms of life (Win *et al.*,  
54 2012; Mitsumasu *et al.*, 2015). As a consequence, they have evolved sophisticated  
55 surveillance systems to detect and protect themselves against parasite invasion (Cook  
56 *et al.*, 2015; Wu *et al.*, 2018; Kanyuka *et al.*, 2019). In turn, plant parasites have evolved  
57 suites of proteins, miRNAs, or other molecules which are delivered into host plants to  
58 facilitate colonisation (Virulence Factors (VFs)) (Win *et al.*, 2012; Giraldo *et al.*, 2013;

59 Zheng *et al.*, 2013; Mitsumasu *et al.*, 2015). These VFs are pivotal in determining the  
60 outcome of a parasite-plant interaction. Despite substantial advances in understanding  
61 the identity and mode of action of VFs in plant interactions with fungal, bacterial and  
62 nematode parasites (Win *et al.*, 2012; Giraldo *et al.*, 2013; Zheng *et al.*, 2013) much less  
63 is known about VFs mediating parasitic plant interactions with their plant hosts  
64 (Westwood *et al.*, 2010, 2012; Timko *et al.*, 2012). Parasitic plants occur in almost all  
65 terrestrial habitats and have evolved independently at least 12 times (Kuijt 1969;  
66 Westwood *et al.*, 2010; Clarke *et al.*, 2019). Regardless of evolutionary origin, parasitic  
67 plants possess a multicellular organ called the 'haustorium', through which direct  
68 structural and physiological connections are formed with their host plant (Westwood *et*  
69 *al.*, 2010; Yoshida *et al.*, 2016). This allows them to abstract water, organic and inorganic  
70 nutrients. In addition, the haustorium is increasingly recognised to play a role in host  
71 manipulation, through the movement of parasite-derived proteins, miRNAs and other  
72 small molecules into the host plant (Aly *et al.*, 2011; Timko *et al.*, 2012; Westwood 2013;  
73 Yoshida *et al.*, 2016; Shahid *et al.*, 2018; Clarke *et al.*, 2019).

74 *Striga* is a genus of obligate, root parasitic plants within the Orobanchaceae (Parker &  
75 Riches 1993; Spallek *et al.*, 2013). One species in particular, *Striga hermonthica* (Del.)  
76 Benth., is a notorious parasite of rain-fed rice, maize, sorghum and millets, leading to  
77 devastating losses in crop yields for resource-poor farmers in sub-Saharan Africa (SSA)  
78 (Scholes & Press 2008; Rodenburg *et al.*, 2016). Control of *S. hermonthica* is extremely  
79 difficult as the parasite is an obligate outbreeder, with high fecundity, wide dispersal and  
80 a persistent, long-lived seed bank (Parker & Riches 1993) leading to a large effective  
81 population size (Huang *et al.*, 2012). Resistant crop varieties are a crucial component of  
82 successful control strategies (Scholes & Press 2008) yet, even for crop varieties  
83 considered highly resistant, genetic variation within parasite populations is such that a  
84 few individuals can overcome host resistance responses and form successful  
85 attachments (Gurney *et al.*, 2006; Cissoko *et al.*, 2011). To develop crop varieties with  
86 durable resistance against *S. hermonthica*, it is vital to understanding fully, the repertoire,  
87 mode of action and genetic variability of parasite VFs that suppress or circumvent host  
88 defences (Timko *et al.*, 2012; Rodenburg *et al.*, 2017). Given the highly polymorphic  
89 populations of *S. hermonthica* and genetic diversity of the seed bank, we hypothesised  
90 that *S. hermonthica* is likely to possess suites of VFs that allow it to overcome layers of  
91 resistance in multiple host plant varieties. The aim of this study was to identify candidate  
92 genes encoding polymorphic VFs in *S. hermonthica*.

93 To achieve our aims we combined two complementary approaches. First, we assembled  
94 and annotated the genome of *S. hermonthica*, and developed a pipeline for  
95 computational prediction of putative secreted proteins (the secretome) and candidate  
96 VFs. The assembled genome was then used as a reference for an experimental,  
97 population genomics analysis, to compare DNA sequence variants in bulked (pooled)  
98 samples of *S. hermonthica* grown on a susceptible (NERICA-7) or resistant (NERICA-  
99 17) rice host (Fig. 1a i-ii). This allowed us to scan for loci in the *S. hermonthica* genome  
100 where the selection imposed by the resistant host had elevated the frequency of alleles  
101 contributing to successful colonisation (termed ‘virulence’ alleles) (Fig. 1 b-d). A similar  
102 approach was used to identify candidate genomic regions associated with resistance in  
103 *Solanum vernei* to the potato cyst nematode, *Globodera pallida* (Eoche-Bosy *et al.*,  
104 2017). The intersection between genes encoding predicted VFs and genes with highly  
105 significant allele frequency differences in the genome scan of *S. hermonthica*, revealed  
106 a set of candidate virulence loci encoding proteins with many functions, including cell  
107 wall modification, protease, or protease inhibitor, oxidoreductase and putative receptor-  
108 like protein kinase activities. Our results show that diverse strategies are used by *S.*  
109 *hermonthica* to overcome different layers of host resistance and suggest a polygenic  
110 basis of virulence in this parasite.

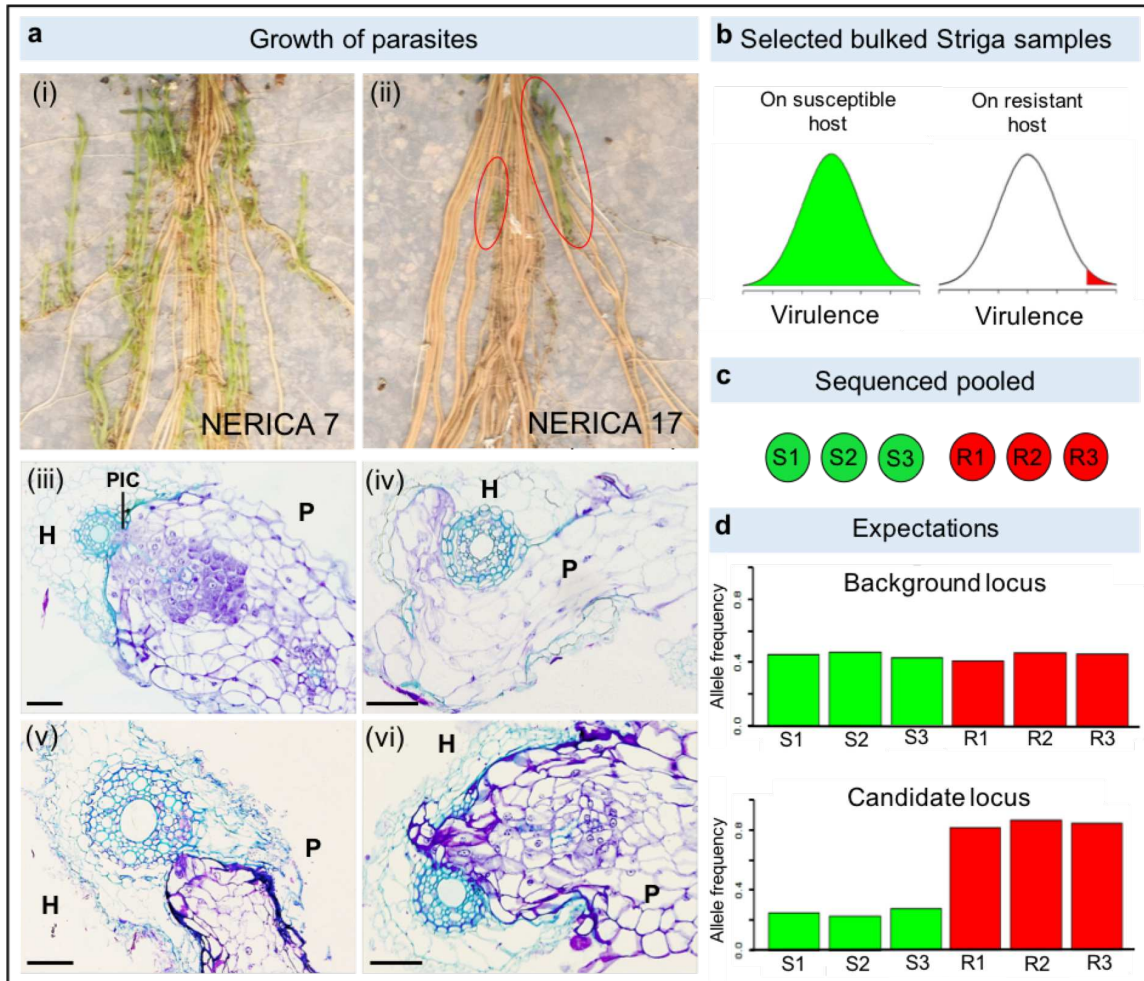
111

## 112 **Materials and Methods**

### 113 **Collection and extraction of *S. hermonthica* DNA for genome and** 114 **pooled sequencing**

115 An accession (population sample) of *S. hermonthica* seeds was collected from  
116 individuals’ parasitising maize in farmers’ fields in the Kibos region of Kenya (0° 5’  
117 30.1272” S; 34° 46’ 4.6416” E). To obtain *S. hermonthica* for genome sequencing and  
118 the bulked sample analysis (BSA), rice seedlings of the varieties, NERICA-7 and  
119 NERICA-17, were grown in rhizotrons and infected with germinated *S. hermonthica*  
120 seeds as described in (Gurney *et al.*, 2006). Plants were grown in a controlled  
121 environment room with a 12 h photoperiod, a photon-flux density of  
122 500  $\mu\text{mol}.\text{quanta}.\text{m}^{-2}.\text{s}^{-1}$  at plant height, a day / night temperature of 28 / 25 °C and 60  
123 % relative humidity. For the construction of a reference genome, one *S. hermonthica*  
124 individual was randomly harvested from NERICA-7. For the pooled sequencing, 300 *S.*





**Figure. 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  $\mu$ m. 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).

125 *hermonthica* individuals (> 30 mg in weight) were harvested from NERICA-7 and from  
126 NERICA-17, divided into 20 mg aliquots and immediately frozen in liquid nitrogen. The  
127 300 individuals from NERICA-7 or NERICA-17 were divided into three pools of 100  
128 individuals (three biological replicates). DNA was extracted from the six pools (see  
129 Methods S1) and samples were subjected to paired-end sequencing using an Illumina  
130 HiSeq machine at the Beijing Genomics Institute (BGI), China. The libraries, insert sizes  
131 and sequencing depth are shown in Table S1. DNA from the individual harvested from  
132 NERICA-7 for the production of a reference genome was sequenced on an Illumina  
133 HiSeq2500 sequencer at Edinburgh Genomics, UK. Six paired-end DNA libraries were  
134 constructed with different insert sizes (Table S1).

### 135 **De novo assembly of the *S. hermonthica* genome**

136 Reads were cleaned and filtered as described in Methods S1. After filtering, ~2.7 billion  
137 reads were generated from the short insert libraries and 0.76 billion reads from mate-pair  
138 libraries. This corresponded to ~230 X and ~54 X coverage of the *S. hermonthica*  
139 genome, respectively. The cleaned and filtered reads were used to assess the *S.*  
140 *hermonthica* genome size, repetitiveness and heterozygosity, compared with 12 other  
141 plant species (Table S2), in the module preQC, implemented in the software sga  
142 (<https://github.com/jts/sga>). This analysis showed *S. hermonthica* was highly  
143 heterozygous and therefore the software Platanus, which is specifically designed for  
144 highly heterozygous genomes, was chosen to assemble the *S. hermonthica* genome  
145 (Kajitani *et al.*, 2014) (Table S3).

146 To further improve the *S. hermonthica* genome assembly, Chicago and Dovetail Hi-C  
147 libraries were prepared and sequenced at Dovetail Genomics, California, USA  
148 (<https://dovetailgenomics.com/plant-animal/>) (Table S3). For construction of Chicago  
149 libraries, DNA from the same *S. hermonthica* individual (used for initial sequencing) was  
150 sequenced on an Illumina HiSeq 2500 platform. For the Hi-C libraries, plant tissues from  
151 an F1 individual from a cross between the sequenced individual and another *S.*  
152 *hermonthica* individual (Kibos accession) were used for the library construction and  
153 sequencing. Sequences from both the Chicago and Hi-C libraries were used only to  
154 improve the contiguity of the initial genome assembly using the Dovetail HiRise  
155 Assembler software. RepeatModeler was used to generate a *S. hermonthica*-specific  
156 repeat library and RepeatMasker was then used to classify repeat elements in the

157 genome. A repeat-masked version of the genome was used for annotation (Smit *et al.*,  
158 2008; 2013).

## 159 **Annotation of the *S. hermonthica* genome**

160 The genome was annotated using three methods. Firstly, gene structures were inferred  
161 using a *S. hermonthica* transcriptome dataset of cDNAs collected from *S. hermonthica*  
162 individuals at eight developmental stages, generated by the Parasitic Plant Genome  
163 Project (PPGP) (Westwood *et al.*, 2012; Yang *et al.*, 2015). The reads were mapped onto  
164 the *S. hermonthica* genome assembly using TopHat to identify exon regions and splice  
165 positions (Trapnell *et al.*, 2009). Transcriptome-based gene structures were predicted  
166 using Cufflinks (Trapnell *et al.*, 2012) and candidate coding regions were then  
167 constructed in Transdecoder (<https://github.com/TransDecoder/>). Secondly, protein  
168 sequences from *Arabidopsis thaliana* (TAIR10), *Mimulus guttatus* (v2.0), *Solanum*  
169 *lycopersicum* (ITAG2.4), *Oryza sativa* (IRGSP1.0) and *Sorghum bicolor* (79), were used  
170 to determine consensus gene models in the genome. The protein sequences were  
171 mapped onto the *S. hermonthica* genome using TBLASTN and pairwise alignments were  
172 then input into Genewise (Birney 2004) to predict gene models in *S. hermonthica*. Thirdly,  
173 an *ab initio* method was used for *de novo* prediction of genes in the *S. hermonthica*  
174 genome using the software, Braker (Hoff *et al.*, 2016). Finally, Evidence Gene Modeler  
175 was used to integrate various gene models from the transcript data, mapped proteins,  
176 and the predicted gene models from the *ab initio* method (Haas *et al.*, 2008). The  
177 completeness of the gene set was assessed using BUSCO v5 using the 2,326 core  
178 orthologs from eudicots\_odb10, with default settings.

## 179 **Functional annotation of the *S. hermonthica* proteome**

180 Putative protein functions were assigned to *S. hermonthica* proteins using BLASTp  
181 analyses against the SwissProt and TrEMBL databases, and against the proteomes of  
182 *Arabidopsis thaliana* (version 30) and *Oryza sativa* (version 7). A BLASTp analysis was  
183 also conducted against the pathogen-host interaction database (PHI-base, version 4.2)  
184 (<http://www.phi-base.org/index.jsp>). BLASTp analyses were run locally using the NCBI  
185 BLAST package (version: ncbi-2.3.0+) and a hit was taken to be significant if e-value <  
186  $10^{-5}$ , bit score and percentage identity > 30. Protein motifs and domains were determined  
187 by searching databases including Pfam, PATHER, GENE3D, CDD, PRINTS, PROSITE,  
188 ProDom and SMART with InterProScan Gene Ontology (GO) terms for individual  
189 proteins retrieved from the corresponding InterPro descriptions.



## 190 **Inference of orthogroups (OG)**

191 Orthologous gene groups (OGs) were inferred using the software OrthoFinder v2  
192 (Emms & Kelly, 2015). The number of genes per species for each OG was transformed  
193 into a matrix of Z-scores to quantify gene family expansion / contraction. The significance  
194 of expansion or contraction was determined using CAFE v4.2 (Han *et al.*, 2013).  
195 Functional annotation of OGs was predicted based on sequence similarity to the InterPro  
196 protein family database. See full details in Methods S1.

## 197 **Prediction, analysis and refinement of the *S. hermonthica* secretome**

198 Secreted *S. hermonthica* proteins were predicted using SignalP v 3.0 and 4.1 (Bendtsen  
199 *et al.*, 2004; Petersen *et al.*, 2011) (Fig. S1). Transmembrane spanning regions were  
200 identified using TMHMM2.0 (Krogh *et al.*, 2001). Proteins with a secretion signal but  
201 without a predicted transmembrane helix were retained as the 'secretome'. Pfam  
202 domains enriched in the *S. hermonthica* secretome compared with the rest of the  
203 proteome (non-secretome) were significant when the corrected p value was  $< 0.1$ ,  
204 according to a Chi-squared test with a false discovery rate (FDR) correction for multiple  
205 testing (Benjamini *et al.*, 1995). The initial secretome was then refined into subsets based  
206 on a series of structural and functional characteristics (Fig. S1) See Methods S1.

## 207 **Identification and analysis of candidate virulence loci using pooled** 208 **sequencing data**

209 The raw sequence reads from the six pools were trimmed and filtered for coverage (see  
210 Methods S1). The likelihood of the observed read counts for the two most common  
211 alleles, across the six pools was calculated according to equation 3 from Gompert and  
212 Buerkle (2011) to allow for the two levels of sampling associated with pooled sequencing  
213 data (sampling of reads and of individuals). We compared three allele-frequency models  
214 for each SNP using the Akaike information criterion (AIC): a null model with a single allele  
215 frequency for all pools, a control-virulent model with one frequency for the control pools  
216 (from the NERICA-7 host) and one for the virulent pools (from the NERICA-17 host) and  
217 a replicate model with a different allele frequency for each of the three pairs of pools (one  
218 control and one virulent) that were sequenced together. The control-virulent model was  
219 the model of interest while the replicate model was intended to check for consistency  
220 across pairs of pools. Therefore, two  $\Delta AIC$  values were obtained:  $\Delta AIC_{cv} = AIC_{null} -$   
221  $AIC_{control-virulent}$  and  $\Delta AIC_{rep} = AIC_{control-virulent} - AIC_{replicate}$ . High positive

222 values of  $\Delta\text{AIC}_{\text{cv}}$  represent better fits than the null model and indicate significant  
223 differences between control and virulent pool types. SNPs with positive  $\Delta\text{AIC}_{\text{rep}}$  values  
224 were likely to be affected by artefacts caused by sequencing methods and were excluded  
225 from the following analyses. All analysis steps were repeated independently for SNPs  
226 based on BWA and NOVOALIGN mapping as recommended by Kofler *et al.*, (2016).

227 The effective population size in *Striga* is likely to be large (Parker & Riches 1993) and  
228 this is consistent with high diversity in our samples (overall mean  $\pi = 0.011$ ). Therefore,  
229 we also expected that linkage disequilibrium would break down quickly. To define a  
230 suitable window size to search for regions potentially implicated in virulence, the extent  
231 of linkage disequilibrium in *S. hermonthica* was investigated (see Methods S1 for details).  
232 On the basis of this analysis, 1 kbp windows were used to detect genomic regions  
233 potentially associated with virulence on the basis of allele frequency differences between  
234 pools from the susceptible and resistant hosts.

235 Regions starting from 5kbp upstream of the start codon and ending no further than 2 kbp  
236 downstream of the stop codon of a gene were divided into 1 kbp-windows and the mean  
237  $\Delta\text{AIC}_{\text{cv}}$  across all the SNPs in each window was calculated. A permutation test was  
238 performed to obtain the probability of observing the mean  $\Delta\text{AIC}_{\text{cv}}$  value, or higher, for  
239 each window based on the distribution of  $\Delta\text{AIC}_{\text{cv}}$  across the regions as a whole (see  
240 Methods S1 for details). Finally, we retained genic regions (defined as regions from 2  
241 kbp upstream of the start codon to the 1 kbp window containing the stop codon) for which  
242 this probability was less than or equal to  $2 \times 10^{-5}$  for both the BWA and NOVOALIGN  
243 analyses in any window. This cut-off was chosen to provide experiment-wide significance  
244 given the number of protein-coding genes in the analysis (29,518). In the secretome, a  
245 more relaxed cut-off of  $1 \times 10^{-4}$  was used to reflect the prior expectation that the secretome  
246 would be enriched with pathogenicity-related genes and the smaller number of genes in  
247 this set (3,375). Thirty-two genes met this criterion for both Novoalign and BWA (Data  
248 S1). In addition, six genes encoding putative secreted proteins that passed the  $1 \times 10^{-4}$   
249 cut-off for either Novoalign or BWA were included in the candidate set because they  
250 either contained large numbers of non-synonymous SNPs or contained high impact  
251 SNPs that can alter protein structure (e.g. due to protein truncation) (Data S1).

252 Two population statistics were calculated for each genic region in the control pool using  
253 the software Popoolation (Kofler *et al.*, 2011). These were nucleotide diversity ( $\pi$ ) and  
254 Tajima's D, a statistic describing the allele frequency spectrum used for testing whether

255 a DNA sequence is evolving under a process that departs from the standard neutral  
256 model, such as selection or demographic change (Tajima, 1989). See Methods S1 for  
257 details.

## 258 **Analyses of candidate virulence genes**

259 The candidate virulence genes were categorised into functional groups based on the  
260 annotations of the closest matching homologs from the *A. thaliana* and *O. sativa*  
261 proteomes, as well as the Pfam domain annotations. For each gene, the numbers of  
262 SNPs were counted for the promoter region (within 2 kbp upstream of the start codon),  
263 the intronic region and coding region, and the numbers of non-synonymous SNPs were  
264 determined. To quantify the allele frequency differences between control and virulent  
265 pools for these candidate virulence genes, the proportion of SNPs with high fixation index  
266 ( $F_{ST}$ ) values in the significant window was calculated (see Methods S1).

## 267 **Expression profiling of candidate virulence genes**

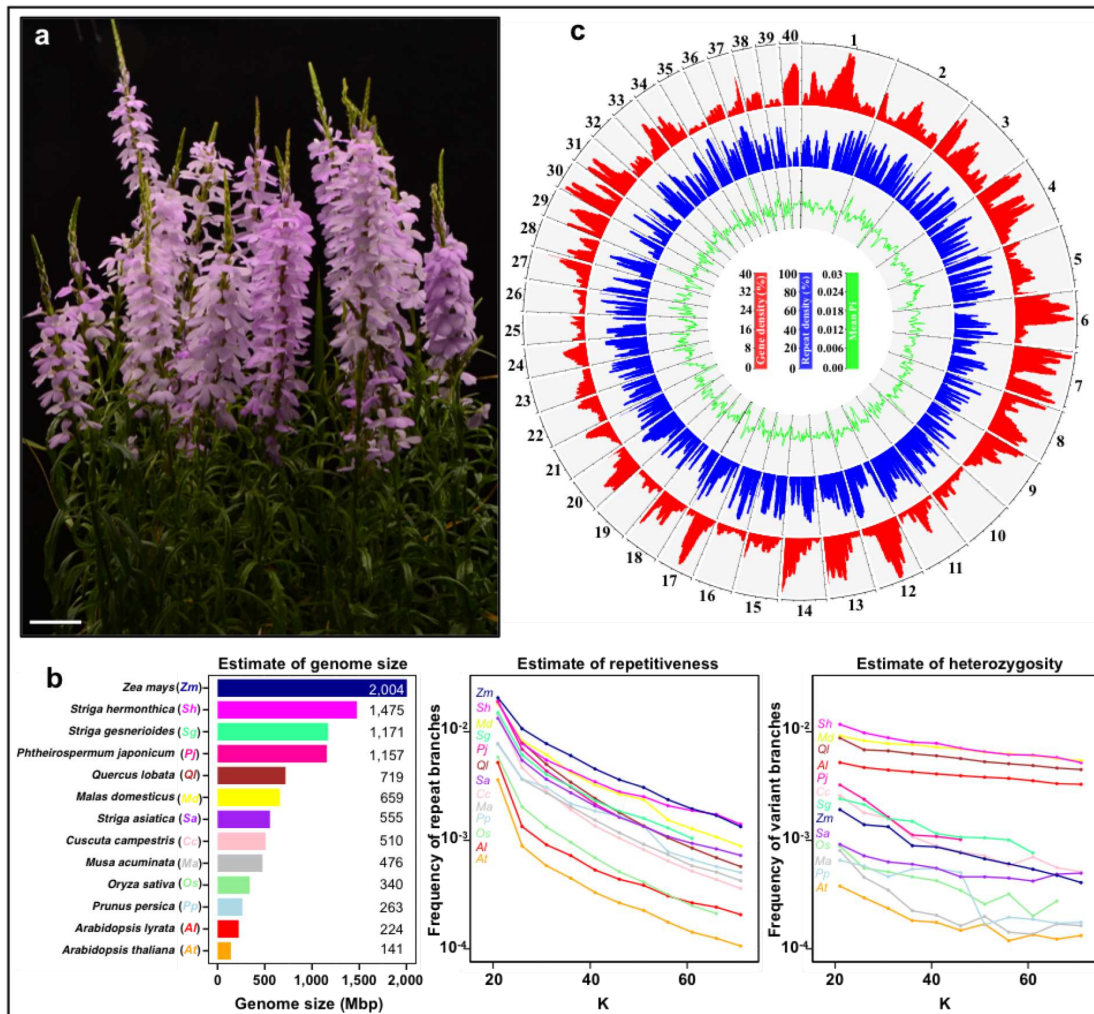
268 Expression profiles for candidate virulence genes were determined for *S. hermonthica*  
269 collected at 2, 4, or 7 days post infection from the roots of NERICA-7 rice plants (full  
270 details are provided in Methods S1). In addition, unattached *S. hermonthica* haustoria  
271 were induced *in vitro* by the addition of 10  $\mu$ M DMBQ (Fernández-Aparicio *et al.*, 2013).  
272 Cleaned reads were mapped to the *S. hermonthica* genome using Tophat2, version  
273 v2.0.12 and quantified with HTSeq (version 0.6.1). FPKM values for each gene at each  
274 time point were used to calculate a fold change in expression relative to the haustorial  
275 sample and significance assessed with a one-way ANOVA. For each gene, log2 fold  
276 expression values, across the time points, were centred around 0 and scaled by the  
277 standard deviation for plotting as a heatmap using the pheatmap function in R. Further  
278 details are provided in Methods S1.

279

## 280 **Results**

### 281 **The *S. hermonthica* genome is very heterozygous**

We obtained a single population of *S. hermonthica* seeds from farmer's fields in Kibos, Kenya and infected a highly susceptible rice variety, NERICA-7 (Fig. 1a). The genetic diversity of the seed population is reflected in the subtle differences of flower colour and morphology of attached parasites (Fig. 2a). We sequenced, assembled and



**Figure 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.

282 characterised the genome of a single individual from this population, which to our  
 283 knowledge, represents the first genome assembly for *S. hermonthica*. The genome size  
 284 was estimated by K-mer analysis to be 1,475 Mbp, (Fig. 2b). This agrees closely with a  
 285 previous flow cytometry-based estimate (Estep *et al.*, 2012) and is more than twice the  
 286 size of the recently sequenced genome of *S. asiatica* (Yoshida *et al.*, 2019). The



287 assembly consisted of 34,907 scaffolds > 1 kbp in length, with an N50 of 10.0 Mbp and  
288 29 scaffolds making up half of the genome size (Table S3). The *S. hermonthica* genome  
289 was remarkably heterozygous (overall mean  $\pi = 0.011$ ) (Fig. 2 b,c) when compared with  
290 other parasitic and non-parasitic plant genomes, likely reflecting the fact that it is an  
291 obligate outbreeding species. In addition, the genome contained a large proportion (69%)  
292 of repetitive DNA (Fig 2 b,c), dominated by long terminal repeat (LTR) elements (Table  
293 S4), a pattern also found for the shoot-parasitic plants, *Cuscuta australis* and *C.*  
294 *campestris* (Sun *et al.*, 2018; Vogel *et al.*, 2018) and the closely related parasitic plant *S.*  
295 *asiatica* (Yoshida *et al.*, 2019). As expected, the density of repetitive elements along each  
296 scaffold negatively correlated with the density of protein-coding genes (Fig 2c). In total,  
297 29,518 protein-coding genes were predicted from the *S. hermonthica* genome, which  
298 was comparable to *S. asiatica* (34,577), the closely related non-parasitic plant *Mimulus*  
299 *guttatus* (28,140) and to *Arabidopsis thaliana* (27,416) (Table S5).

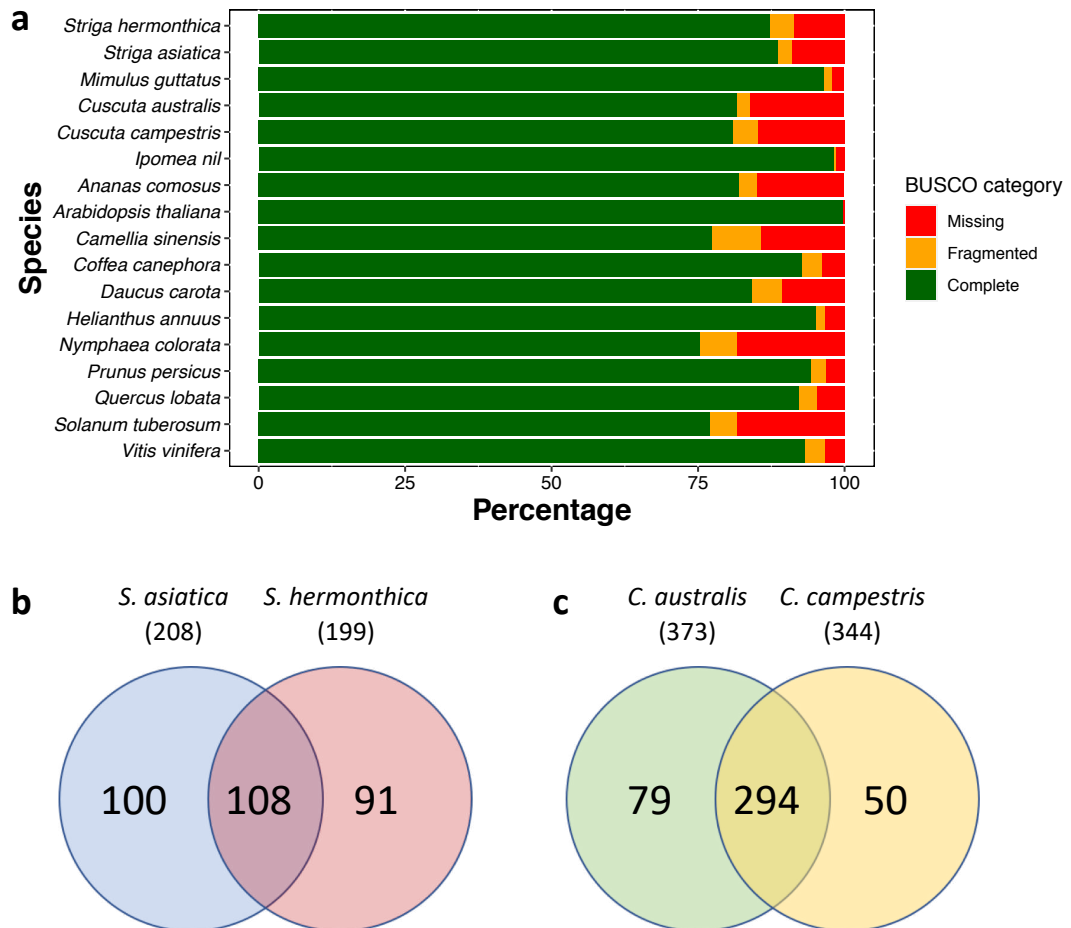
300 BUSCO analysis of gene set completeness (Waterhouse *et al.*, 2018), showed 87.3% of  
301 2,326 conserved single-copy orthologs in eudicotyledons were complete in the *S.*  
302 *hermonthica* genome, similar to that found in *S. asiatica* (88.7%) (Fig. 3; Table S6). Of  
303 the BUSCOs not found in the *S. hermonthica* genome, over half were also absent from  
304 the *S. asiatica* genome (Table S6). Both *Striga* spp. share missing BUSCOs that are  
305 present in the genome of the closely related non-parasitic *Mimulus guttatus* (Fig. 3b;  
306 Table S6). Similarly, two shoot holoparasites, *C. australis* and *C. campestris*, with a  
307 BUSCO completeness of 81.0 and 81.7% respectively, also shared many missing  
308 BUSCOs that were present in the genome of their non-parasitic relative, *Ipomea nil* (Fig.  
309 3c). This is consistent with previous findings suggesting some missing BUSCOs are likely  
310 to be a result of the parasitic lifestyle (Sun *et al.*, 2018; Vogel *et al.*, 2018; Yoshida *et al.*,  
311 2019; Cai *et al.*, 2021).

312 Comparative analysis of orthologous gene groups (orthogroups) between *S.*  
313 *hermonthica* and 12 other plant species identified 22,624 orthogroups in total, of which  
314 12,278 contained *S. hermonthica* genes. Of these, 327 were significantly expanded and  
315 104 were contracted in the *S. hermonthica* genome (Fig. 4a). Expanded orthogroups  
316 included the  $\alpha/\beta$ -hydrolase family, recently shown to have undergone duplication in *S.*  
317 *hermonthica* (Toh *et al.*, 2015), as well as numerous F-box, leucine-rich repeat and  
318 protein kinase domain-containing proteins (Fig. 4b). Of particular interest in the context  
319 of pathogenicity were *S. hermonthica*-specific orthogroups annotated as papain family



320 cysteine proteases, xylanase inhibitors and trypsin and protease inhibitors (Fig. 4b). Both  
321 proteases and protease inhibitors function in a wide range of plant-plant parasite  
322 interactions and may act offensively, by degrading host proteins, or defensively, by  
323 inhibiting host defence enzymes (Bleischwitz *et al.*, 2010; Mueller *et al.*, 2013).

324



**Figure. 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.

325

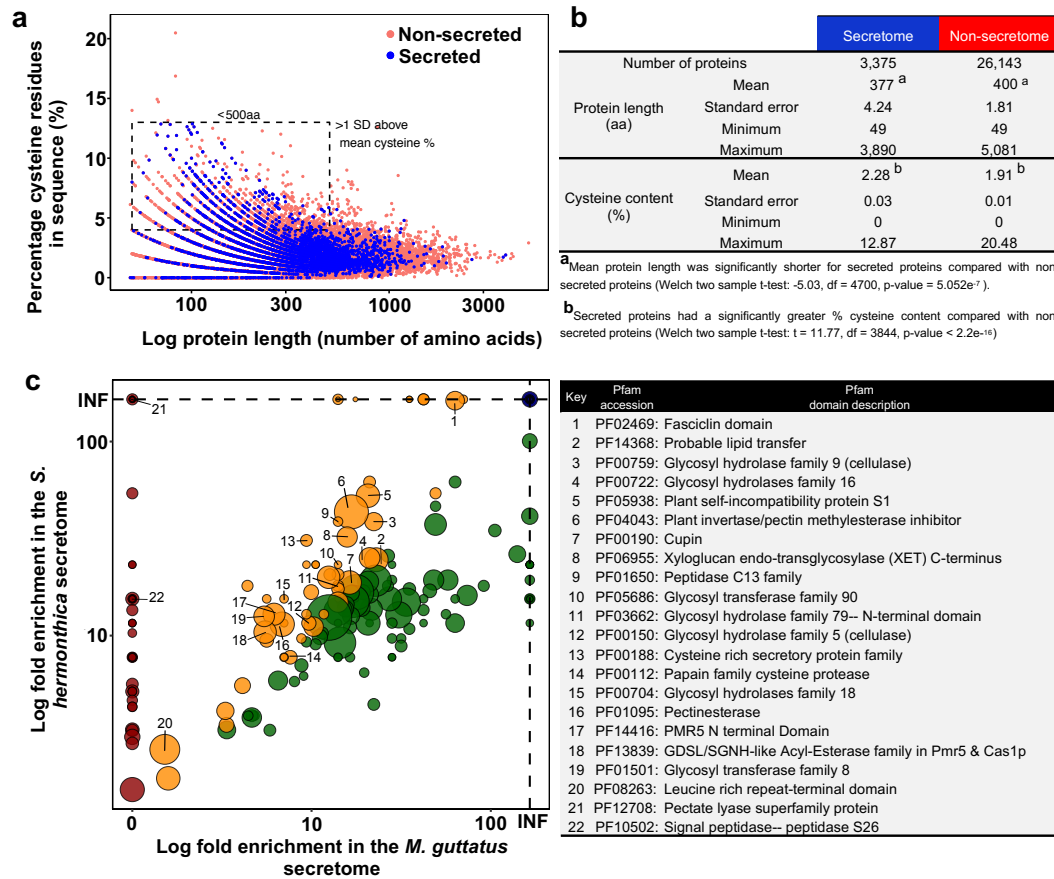


**Figure 4. | Orthogroup analyses.** **a** A time tree for *S. hermonthica* and 12 other species generated in MEGA, based on 42 single-copy genes inferred from OrthoFinder. The number of significantly expanded (red) and contracted (blue) orthogroups based on CAFE analysis are shown above the branches. **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.

## 326 **The *S. hermonthica* secretome**

327 One way that parasite proteins can interact with host biology is through parasite-directed  
328 secretion. We identified 3,375 putatively-secreted proteins in *S. hermonthica* (11.4 % of  
329 the proteome) (Fig. S1), many of which were homologous to *A. thaliana* secreted proteins  
330 (Table S7), providing experimental evidence for secretion into the extracellular space.  
331 On average, the *S. hermonthica* secreted proteins were both significantly smaller and  
332 had a higher percentage of cysteine residues compared with the rest of the proteome  
333 (Fig. 5 a, b). Genes encoding secreted proteins tended to be more clustered (within 15  
334 kbp of their nearest neighbour) compared to all genes in the genome ( $p < 10^{-4}$ ,  $10^5$   
335 permutations) (Fig. S2) suggesting they are likely to be arrayed in tandem and belong to  
336 large gene families (Elizondo *et al.*, 2009). Functionally, the secretome was rich in protein  
337 domains involved in cell wall modification (e.g. endoglucanases, cellulases,  
338 pectinesterases, expansins, and pectate lyases), protease activity (e.g. papain-like  
339 cysteine proteases, aspartic proteases, and subtilase proteases) and oxidoreductase  
340 activity (peroxidases, copper oxidases, and cytochrome p450 proteins) (Fig. 5c, Figs. S3  
341 and S4). The cytochrome P450 domain, for example, was particularly frequent in the *S.*  
342 *hermonthica* secretome (3.13% of protein domains) compared with the rest of the  
343 proteome (0.25% of protein domains) (Fig. S3). Three other highly-abundant protein  
344 domains in the secretome were described as copper oxidases (Fig. S3) and are  
345 commonly found in laccases that are involved in the generation or breakdown of phenolic  
346 components, such as lignin (Kwiatos *et al.*, 2015). Small cysteine-rich proteins are  
347 common characteristics of VFs from a range of phytoparasites (Saunders *et al.*, 2012;  
348 Lu *et al.*, 2016). In *S. hermonthica*, 183 such proteins were identified (Fig. 5a) and were  
349 similar to proteins annotated as carbohydrate binding X8 domain-containing proteins,  
350 protease inhibitor/lipid transfer proteins, PAR1-like proteins, pectinesterases, RALF-like  
351 proteins and thaumatin-like proteins (Fig. S4), many of which are likely to play a role in  
352 host-Striga interactions (Yang *et al.*, 2015; Yoshida *et al.*, 2019).

353 We identified several protein domains in the *S. hermonthica* secretome that were  
354 enriched to a higher degree than observed in the secretome of the closely-related non-  
355 parasitic plant, *M. guttatus* (Fig. 5c, Fig S3, Data S2), suggesting these functions are  
356 relevant to the parasitic lifestyle. The xyloglucan endotransglycosylase (PF06955)  
357 domain, for example, was found in 17 *S. hermonthica* proteins (Fig. 5c, Fig. S4).



**Figure 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.

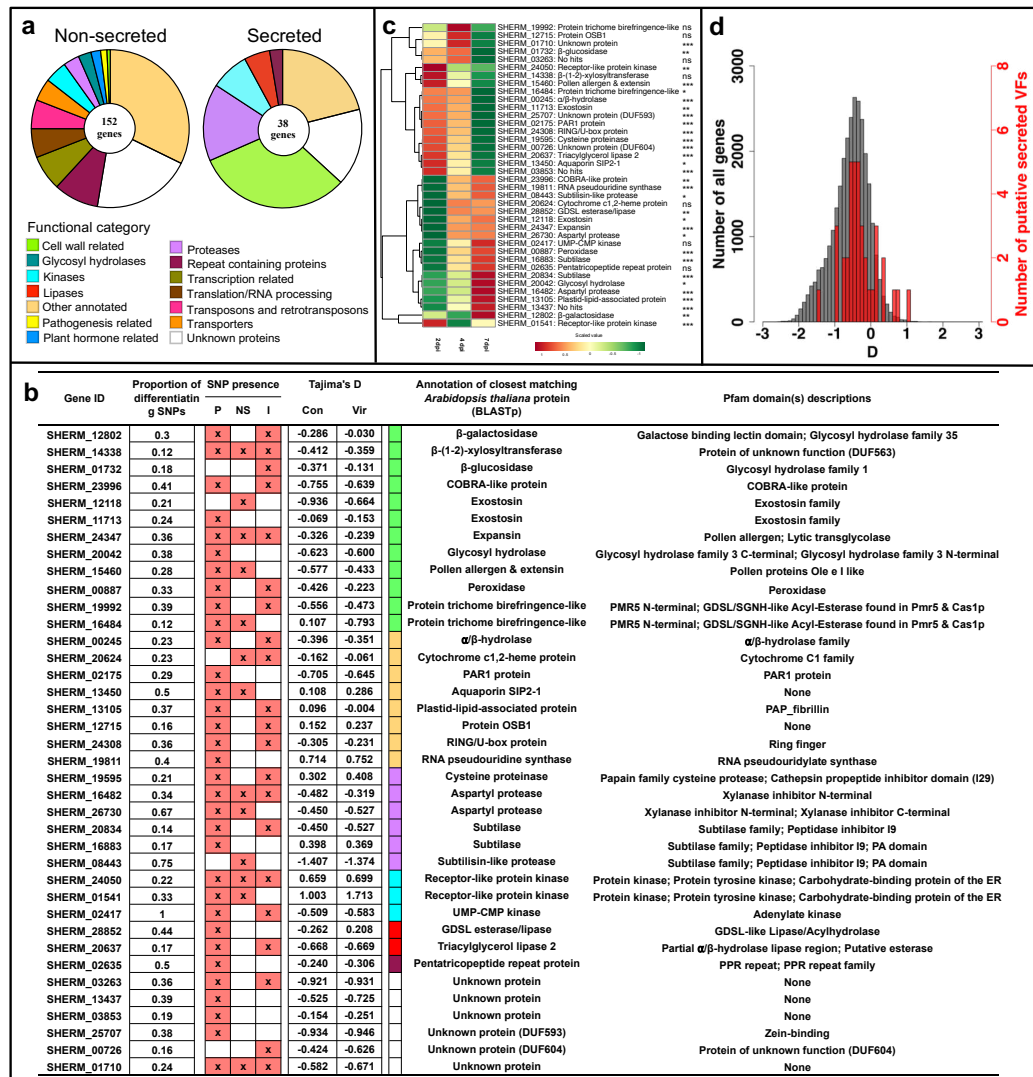
358 Xyloglucan endotransglucosylases / hydrolases (XETs) have the potential to modify  
 359 either the parasite or host cell walls (or both) during parasitism (Olsen & Krause 2017).  
 360 XETs are secreted from the haustoria of the parasitic plant *Cuscuta reflexa* during a  
 361 susceptible interaction on its host *Pelargonium zonale*, contributing towards

362 pathogenicity (Olsen & Krause 2017). Pectate lyase superfamily (PF12708) and  
363 pectinesterase (PF01095) domains were enriched in the secretome of *S. hermonthica*  
364 compared to *M. guttatus* and may act as VFs to modify host, or parasite, pectin during  
365 penetration. We found a battery of different carbohydrate-active glycosyl hydrolase (GH)  
366 domains that were enriched in the *S. hermonthica* secretome (Fig. 5c, Fig. S3). Eight *S.*  
367 *hermonthica* proteins were annotated as cellulases of the GH5 family (containing domain  
368 PF00150) (Fig. S4) and were similar to secreted cellulases that function as VFs in some  
369 phytoparasitic nematodes (Smant *et al.*, 1998). The degradation of cellulosic  $\beta$ -1,4-  
370 glucans has been observed in susceptible sorghum roots infected by *S. hermonthica*  
371 (Olivier *et al.*, 1991) and may be mediated by these secreted enzymes to facilitate the  
372 migration of *S. hermonthica* intrusive cells between host root cortical cells. The  
373 identification of many putatively secreted VFs in the *S. hermonthica* genome, that are  
374 likely to modify host plant cell walls, raises an interesting question about how such  
375 proteins are targeted to avoid damaging the parasite's own cell walls.

### 376 **Population genomic analysis to identify candidate virulence loci**

377 Our experimental system allowed us to identify a subset of VFs with genetic variation  
378 relevant to the ability to infect some host genotypes and not others. Hundreds of *S.*  
379 *hermonthica* individuals were harvested from either a very resistant (NERICA-17) or  
380 susceptible (NERICA-7) rice cultivar, and pools of these individuals were subjected to  
381 genome resequencing. After aligning the reads to our reference genome, we detected  
382 1.8 million SNPs in genic regions. These genic regions were split into 150,741 1 kbp  
383 windows and of these, 194 (0.13%) had extreme and consistent allele frequency  
384 differences between the bulked pools of *S. hermonthica* selected on the resistant *versus*  
385 the susceptible hosts (Fig. S5; Data S1). These highly differentiated windows were  
386 located in 190 genes. These candidate loci potentially encode virulence factors with  
387 allelic variants, influencing either structure or expression that contribute to the ability of  
388 some individuals to parasitise NERICA-17. As expected for an outbred parasite with a  
389 large population that encounters multiple host species and genotypes, many loci were  
390 detected and they cover a range of predicted functions. Of these candidate VFs, 152  
391 were not predicted to be secreted and were assigned to a wide range of functional  
392 categories, including putative transcription factors, hormone signalling pathways,  
393 transporters, repeat-containing proteins and a number of proteins of unknown function  
394 (Fig. 6a; Data S1). Some of these proteins may function to protect the parasite against





**Figure 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<sub>2</sub> 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.

395 host defences and facilitate growth on the resistant rice variety. In addition, some may  
396 enter the host by non-traditional pathways, for example, via the host-parasite xylem  
397 connections. One sixth (24) of these non-secreted proteins had sequence similarity to  
398 proteins in the Pathogen-Host Interaction database (Winnenburg *et al.*, 2007). These  
399 included *S. hermonthica* proteins with sequence similarity to a putative leucine-rich  
400 repeat protein from *Ralstonia solanacearum*, a mitogen-activated protein kinase from  
401 *Ustilago maydis*, a calreticulin-like protein from *Magnaporthe oryza* and a cytochrome  
402 P450 from *Bursaphelenchus xylophilus* (Data S1).

403 The remaining 38 VFs were members of the *S. hermonthica* secretome and represent  
404 particularly strong candidates associated with the ability to parasitise NERICA-17  
405 successfully (Fig. 6a,b, Data S1). These genes were categorised into six functional  
406 groups, the largest of which contained 12 genes associated with cell wall modification  
407 (Fig. 6a,b), including genes encoding an expansin protein, a COBRA-like protein, a  $\beta$ -(1-  
408 2)-xylosyltransferase, two trichome birefringence-like (TBL) proteins, a pollen Ole e  
409 allergen and two exostosin family proteins, all of which can function to modify the  
410 extensibility or other mechanical properties of plant cell walls (Li 2003; Qin *et al.*, 2004;  
411 Honaas *et al.*, 2013; Mitumasu *et al.*, 2015) (Fig. 6b). Groups of genes annotated as  
412 proteases (6 genes including subtilases, aspartyl proteases, and a cysteine proteinase),  
413 lipases (3 genes) and kinases (3 genes) were also found. The proteases were always  
414 associated with an inhibitor protein domain (Fig. 6b). For example, the putative aspartyl  
415 proteases possessed one or more xylanase inhibitor domain(s) (Fig. 6b). There were  
416 also eight genes encoding proteins with a range of putative functions, including a PAR1-  
417 like protein, a probable aquaporin, an  $\alpha/\beta$ -hydrolase and two receptor-like protein  
418 kinases (Fig. 6b). In addition, a further six genes were annotated as proteins of unknown  
419 function (Fig. 6b).

420 The 38 candidate VFs were investigated in more detail by quantifying changes in gene  
421 expression in haustoria at critical stages of parasite development on the susceptible rice  
422 variety NERICA-7 by inspecting the distribution of SNPs throughout the promoter and  
423 genic regions, and testing for signatures of historical selection. Gene expression was  
424 measured in an independent experiment (Fig. 6c). Changes in gene expression of  
425 attached haustoria were measured relative to gene expression in haustoria generated *in*  
426 *vitro*. At 2 days after inoculation of the host root, parasite haustoria were attached and  
427 parasite intrusive cells had penetrated into the host root cortex. By day 4, the parasite

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

431 Prior to attachment to the host, some of the genes encoding candidate VFs were not  
432 expressed in haustoria (e.g. subtilase gene (SHERM\_16883) and subtilisin-like protease  
433 (SHERM\_08443) or were expressed at very low levels (e.g. the peroxidase  
434 (SHERM\_00887), glycosyl hydrolase (SHERM\_(20042), both aspartyl proteases  
435 (SHERM\_16482 and SHERM\_26730) and an unknown protein (SHERM\_03853) (S3  
436 Data). However, all 38 genes were expressed in haustoria during the early stages of  
437 infection of the susceptible host, NERICA-7 (Fig. 6c; Data S3). There were two main  
438 patterns of gene expression. Firstly, 21 genes, including those mentioned above, had  
439 low levels of expression in haustoria 2 days post infection, followed by an increase in  
440 expression as infection progressed (Fig. 6c; Data S3). In contrast, 17 genes were highly  
441 expressed in haustoria 2 days post infection and expression then decreased  
442 progressively with time, e.g. genes encoding  $\beta$ -glucosidase,  $\beta$ -(1-2)-xylosyltransferase,  
443 and TBL protein SHERM\_06484, all of which modify cell walls. The cysteine protease,  
444 PAR1,  $\alpha/\beta$ -hydrolase and aquaporin genes also exhibited a similar expression profile  
445 (Fig. 6c; Data S3).

446 Most of the 38 genes had significantly differentiating SNPs in their promoter regions (from  
447 the start site to 2 kbp up-stream). Some of these SNPs may lead to a change in the  
448 regulation of gene expression (Fig 6b). Some genes, for example, the gene encoding the  
449 pollen Ole e allergen protein (SHERM\_15460), one of the exostosin family proteins  
450 SHERM\_12118), a probable aquaporin SIP2-1 (SHERM\_13450) and one of the two  
451 protein TBL genes (SHERM\_16484), also had non-synonymous SNPs in the coding  
452 region (Fig. 6b) that may result in functional differences between the alleles of these  
453 genes in individuals infecting NERICA-7 and NERICA-17. Finally, SNPs were also found  
454 within predicted intron regions in many of the genes (Fig. 6b).

455 The co-evolutionary interactions between hosts and parasites can generate balancing  
456 selection (Frank 1993). We predicted that genes contributing to virulence would tend to  
457 have a history of balancing selection because of the diverse range of hosts used by *S.*  
458 *hermonthica*. To test this prediction, we compared Tajima's D between candidate loci  
459 and the rest of the genome, expecting to see more positive values (Charlesworth 2006).  
460 We used the pools from the susceptible host for this comparison because they

461 represented the *Striga* population as a whole. As predicted, the 152 candidate loci in the  
462 *S. hermonthica* proteome (Fig S6) and the 38 candidate loci in the secretome (Fig. 6d)  
463 had significantly elevated Tajima's D, on average, compared to all the genes in the  
464 genome ( $p < 0.0001$  and  $p < 0.0003$ , respectively;  $10^5$  permutations). Some loci had  
465 particularly high Tajima's D values, for example the two receptor-like protein kinases (Fig.  
466 6b). Interestingly, some loci showed large differences in Tajima's D between the control  
467 and virulent *S. hermonthica* pools with the largest difference seen for the TBL gene  
468 (SHERM\_16484) with a negative  $\Delta D$  ( $D_{Vir} - D_{Con}$ ) of -0.9. This suggests strong selection  
469 resulting in one common haplotype in the virulent pools in contrast to two or more  
470 haplotypes at intermediate frequencies in the control pools. There were also large  
471 positive  $\Delta D$  values: 0.71, 0.16 and 0.20 for one of the putative receptor-like protein  
472 kinases SHERM\_01541, one of the aspartyl proteases, SHERM\_16482, and the  
473 peroxidase SHERM\_00887, respectively. This suggests that a rare haplotype in the  
474 control pools is present at intermediate frequency in the virulent pools. Overall, these  
475 changes indicate that selection on the resistant host caused changes in frequency of  
476 multi-SNP haplotypes at these loci, haplotypes that may have been created by areas of  
477 low recombination or by recent invasion of new variants under positive selection (Cutter  
478 & Payseur 2013) and which underlie the ability of some *S. hermonthica* individuals to  
479 overcome resistance in NERICA-17.

480

## 481 Discussion

482 Plants secrete proteins involved in many biological functions, from nutrient acquisition,  
483 to development and defence (Li 2003; Cook *et al.*, 2015). However, unlike most plants,  
484 in parasitic plants such as *S. hermonthica* a subset of secreted proteins is likely to  
485 function as VFs and contribute towards parasite fitness by facilitating host colonization  
486 (Timko *et al.*, 2012). We used a combination of *in silico* prediction of secreted proteins  
487 and pooled sequencing of parasites derived from susceptible and resistant rice hosts,  
488 both facilitated by the first available genome assembly, to identify a set of candidate VFs.  
489 These are secreted proteins encoded by genes that had extremely different allele  
490 frequencies between replicated pools derived from susceptible and resistant hosts,  
491 suggesting strong selection for particular variants that facilitate successful colonisation  
492 despite host resistance. This experimental approach has not been applied previously to

493 investigate virulence of *Striga*, or any other parasitic plant. Its success here paves the  
494 way to application of similar methods to other host-parasite combinations, providing vital  
495 information on virulence mechanisms and their genetic variability within and between  
496 parasitic plant populations from different regions of Africa, and so underpinning the  
497 development of sustainable control strategies.

498 Our list of 38 candidate, secreted, VFs points to key functions involved in pathogenicity,  
499 including oxidoreductase, receptor-like protein kinase, protease and protease inhibitor,  
500 and cell wall modification activities. The latter is consistent with growing evidence that  
501 cell-wall modification is a critical step in plant invasions by many different parasites  
502 including parasitic plants. Recently, the structural integrity of lignin was shown to be a  
503 crucial component of resistance in roots of the rice variety Nipponbare to infection by *S.*  
504 *hermonthica* (Mutuku *et al.*, 2019). In our study the host cell wall is clearly involved in  
505 resistance in NERICA-17. Most *S. hermonthica* individuals from the Kibos population  
506 were unable to penetrate the root endodermis or, if they breached the endodermis, they  
507 were unable to establish functional connections to the host xylem vessels (Fig. 1a iv-vi).  
508 Consistent with this, the largest category of our candidate, secreted VFs included a  
509 putative peroxidase, an expansin, pollen allergen-like proteins, a  $\beta$ -glucosidase, a  $\beta$  (1-  
510 2) xylosyltransferase, and a TBL protein, all of which function to modify cell walls. The  
511 TBL protein, SHERM\_16484, had a strikingly different Tajima's D in the control pool  
512 compared to the value in the virulent pool, consistent with selection favouring one  
513 haplotype on the resistant NERICA-17, out of several haplotypes present in the  
514 population. In *A. thaliana* and *O. sativa* TBL proteins belong to large gene families with  
515 functions related to cell wall modifications. In *A. thaliana*, At-TBL44 has been implicated  
516 in pectin esterification (Vogel *et al.*, 2004; Bacete *et al.*, 2018), whilst in rice other  
517 members of this family appear to be involved in acetylation of xylan moieties in cell walls  
518 (Gao *et al.*, 2017). In each case, alterations in enzyme activity altered resistance in *A.*  
519 *thaliana* to powdery mildew and in rice to leaf blight (Vogel *et al.*, 2004; Gao *et al.*, 2017).  
520 Recently an 11 kDa protein was isolated from the cell wall of the shoot parasite *C. reflexa*  
521 and identified as a glycine rich protein (GRP) (Hegenauer *et al.*, 2020). The protein and  
522 its minimal peptide epitope (Crip21) bind to and activate a cell surface resistance gene  
523 in tomato (CuRe1), leading to resistance to the parasite, illustrating the importance of  
524 cell wall modifications to host resistance.



525 In addition to cell wall modification, several candidate genes were annotated as having  
526 protease activity, including two aspartyl proteases, three subtilisin or subtilisin-like genes  
527 and a cysteine proteinase. Interestingly, all had a dual-domain predicted structure  
528 consisting of a propeptide inhibitor domain and a catalytic protease domain. In other such  
529 protease enzymes, the propeptide domain auto-inhibits the enzyme activity until  
530 cleavage of this inhibitor domain activates the catalytic domain (Shindo & Van Der Hoom  
531 2007). This provides a mechanism by which the parasite could initially secrete an inactive  
532 VF that only becomes active once in the host environment. A similar dual-domain  
533 structure was found for a highly expressed, haustorium-specific cysteine protease in the  
534 shoot parasitic plant, *C. reflexa*, which positively contributes towards pathogenicity  
535 (Bleischwitz *et al.*, 2010) Although the precise functions of other candidate VFs are  
536 unknown, for example the putative aquaporin, PAR1 protein, cytochrome P450 and the  
537 5 proteins with no functional annotation, they provide exciting avenues for further  
538 investigation.

539 *S. hermonthica* has extremely high fecundity (>100,000 seeds per plant) (Parker &  
540 Riches 1993), a persistent seed bank and is obligate out-crossing (Safa *et al.*, 1984),  
541 leading to a very large effective population size (Huang *et al.*, 2012). Therefore, the high  
542 heterozygosity that we observed in the *S. hermonthica* genome was not unexpected. *S.*  
543 *hermonthica* parasitizes many different host species and varieties, often within the same  
544 geographical area. Populations therefore encounter many different forms of resistance,  
545 which they experience as a highly heterogeneous environment. This is expected to  
546 maintain genetic diversity at many loci contributing to virulence, which is consistent with  
547 observations from field studies that resistant varieties, of any particular crop species, are  
548 often parasitized by one or two *S. hermonthica* individuals (Gurney *et al.*, 2006;  
549 Rodenburg *et al.*, 2017). A typical example is the host-parasite combination used here  
550 as a test system; the *S. hermonthica* Kibos population and the strongly resistant upland  
551 rice variety, NERICA-17 one of 18 NERICA rice varieties grown widely by African farmers  
552 (Cissoko *et al.*, 2011; Rodenburg *et al.*, 2015).

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

558 candidate VFs would be revealed, by repeating this comparison on other resistant hosts.  
559 An important question for the future will be to determine how individual VFs are implicated  
560 in overcoming resistance for specific hosts or across a range of hosts. Second,  
561 maintenance of variation at virulence loci by balancing selection will lead to elevated  
562 Tajima's D relative to the background, reflecting persistence of multiple alleles at these  
563 loci. We found the overall Tajima's D in *S. hermonthica* to be negative, perhaps reflecting  
564 population expansion following the spread of agriculture, but our candidate loci had  
565 significantly higher Tajima's D on average, consistent with balancing selection on these  
566 loci. Understanding the maintenance of variation at virulence loci by balancing selection  
567 will be critical to managing the evolution of virulence as a part of a sustainable control  
568 strategy (Mikaberidze *et al.*, 2015).

569 Effective control of *S. hermonthica* is essential for food security and poverty alleviation  
570 for small-holder subsistence farmers, but it remains elusive. The use of resistance crop  
571 varieties is recognised as sustainable and cost effective (Scholes *et al.*, 2008), but the  
572 durability of resistant varieties is compromised by the potential for rapid evolution of  
573 parasite virulence. Thus, the long-term success of host resistance, as a control strategy  
574 for *S. hermonthica* and other parasitic weeds, requires knowledge of the virulence factors  
575 involved, their allelic variation within and between *Striga* populations and their interaction  
576 with different host resistance alleles. Only then will it be possible to combine resistance  
577 alleles, in host varieties that are suitable for different agro-ecological zones and in ways  
578 that achieve sustained control by delaying the evolution of virulence. Our experimental  
579 approach and identification of candidate VFs and allelic variation within a *S. hermonthica*  
580 population, is a critical first step in this direction.

581

## 582 **Acknowledgements**

583 We thank members of the library production, instrumentation and informatics teams at  
584 Edinburgh genomics. We also thank Dr Hernan Morales, University of Gothenburg,  
585 Sweden for providing the R script used to infer the read coverage distribution for each  
586 SNP for each pool of sequenced reads, based on three-component mixture models. We  
587 thank Dr Mamadou Cissoko, University of Sheffield, for help with the production of the  
588 transverse sections through rice roots infected with *S. hermonthica*. This project was  
589 funded by UKRI Biotechnology and Biological Sciences Research Council  
590 (<https://bbsrc.ukri.org>) grants, BB/J011703/1 and BB/P022456/1, awarded to JDS and

591 RK and The Leverhulme Trust (<https://www.leverhulme.ac.uk>) grant (RPG 2013-050)  
592 awarded to JDS and RK.

### 593 **Author contributions**

594 JDS and RKB planned and designed the research. SQ, PZ and JDS contributed to the  
595 production of *S. hermonthica* materials and extraction of DNA for genome and pooled  
596 sequencing. MB carried out library preparation and sequencing of the *S. hermonthica*  
597 genome. SQ led the genome assembly and annotation with contributions from JMB, RC,  
598 JDS and RKB. JMB carried out the prediction and analysis of the *S. hermonthica*  
599 secretome. SQ mapped the pooled *S. hermonthica* sequence reads to the *S.*  
600 *hermonthica* genome. SQ, RKB and JMB contributed to the population genomic  
601 analyses. JMB, PZ and JDS contributed to the analysis of changes in gene expression  
602 in *S. hermonthica* haustoria. SQ and JMB contributed equally. All authors contributed to  
603 writing of the manuscript.

### 604 **Data Availability**

605 Raw reads for the pooled *S. hermonthica* sequences and for the *S. hermonthica* genome  
606 sequence, the assembled genome sequence and annotations have been submitted to  
607 the European Nucleotide Archive (ENA) browser at (<http://www.ebi.ac.uk/ena/data/view/>)  
608 under the following accession numbers: Genome Assembly GCA\_902706635; Project  
609 ID PRJEB35606; Sample ID ERS4058863 and Contig accession CACSLK010000001-  
610 CACSLK010035056.

611

### 612 **References**

613 **Aly R, Hamamouch N, Abu-Nassar J, Wolf S, Joel DM, Eizenberg H, Kaisler E,**  
614 **Cramer C, Gal-On A, Westwood JH. 2011.** Movement of protein and macromolecules  
615 between host plants and the parasitic weed *Phelipanche aegyptiaca* Pers. *Plant Cell*  
616 *Reports* **30**: 2233–2241.

617 **Bacete L, Mélida H, Miedes E, Molina A. 2018.** Plant cell wall-mediated immunity: cell  
618 wall changes trigger disease resistance responses. *The Plant Journal* **93**: 614–636.

619 **Bendtsen J D, Nielsen H, von Heijne G, Brunak S. 2004.** Improved prediction of signal  
620 peptides: SignalP 3.0. *Journal of Molecular Biology* **340**: 783–795.

- 621 **Benjamini Y, Hochberg Y. 1995.** Controlling the False Discovery Rate: A Practical and  
622 Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B*  
623 *(Methodological)* **57**: 289–300.
- 624 **Birney E. 2004.** GeneWise and genomewise. *Genome Research* **14**: 988–995.
- 625 **Bleischwitz M, Albert M, Fuchsbauer H-L, Kaldenhoff R. 2010.** Significance of  
626 Cuscutain, a cysteine protease from *Cuscuta reflexa*, in host-parasite interactions. *BMC*  
627 *Plant Biology* **10**: 8.
- 628 **Cai L, Arnold BJ, Xi Z, Khost DE, Patel N, Hartmann CB, Manickam S, Sasirat S,**  
629 **Nikolov LA, Mathews S, et al. 2021.** Deeply altered genome architecture in the  
630 endoparasitic flowering plant *Sapria himalayana* Griff. (Rafflesiaceae). *Current Biology*  
631 **31**: 1002-1011.e9.
- 632 **Charlesworth D. 2006.** Balancing selection and its effects on sequences in nearby  
633 genome regions. *PLoS Genetics* **2**: 6.
- 634 **Cissoko M, Boissard A, Rodenburg J, Press MC, Scholes JD. 2011.** New Rice for  
635 Africa (NERICA) cultivars exhibit different levels of post-attachment resistance against  
636 the parasitic weeds *Striga hermonthica* and *Striga asiatica*. *New Phytologist* **192**: 952–  
637 963.
- 638 **Clarke CR, Timko MP, Yoder JI, Axtell MJ, Westwood JH. 2019.** Molecular dialog  
639 between parasitic plants and their Hosts. *Annual Review of Phytopathology* **57**.
- 640 **Cook DE, Mesarich CH, Thomma BPHJ. 2015.** Understanding plant immunity as a  
641 surveillance system to detect invasion. *Annual Review of Phytopathology* **53**: 541–563.
- 642 **Cutter AD, Payseur BA. 2013.** Genomic signatures of selection at linked sites: unifying  
643 the disparity among species. *Nature Reviews Genetics* **14**: 262–274.
- 644 **Elizondo LI, Jafar-Nejad P, Clewing JM, Boerkoel CF 2009.** Gene clusters, molecular  
645 evolution and disease: A speculation. *Current Genetics*. **12**: 64-75.
- 646 **Emms DM, Kelly S. 2015.** OrthoFinder: solving fundamental biases in whole genome  
647 comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**.
- 648 **Eoche-Bosy D, Gautier M, Esquibet M, Legeai F, Bretaudeau A, Bouchez O, et al. 2017**  
649 Genomic scans on experimentally evolved populations reveal candidate regions for adaptation  
650 to plant resistance in the potato cyst nematode *Globodera pallida*. *Molecular Ecology* **26**: 4700-  
651 4711.

- 652 **Estep MC, Gowda BS, Huang K, Timko MP, Bennetzen JL. 2012.** Genomic  
653 characterization for parasitic weeds of the genus by sample sequence analysis. *The*  
654 *Plant Genome Journal* **5**: 30.
- 655 **Fernández-Aparicio M, Huang K, Wafula EK, Honaas LA, Wickett NJ, Timko MP,**  
656 **dePamphilis CW, Yoder JI, Westwood JH. 2013.** Application of qRT-PCR and RNA-  
657 Seq analysis for the identification of housekeeping genes useful for normalization of gene  
658 expression values during *Striga hermonthica* development. *Molecular Biology Reports*  
659 **40**: 3395–3407.
- 660 **Frank SA. 1993.** Coevolutionary genetics of plants and pathogens. *Evolutionary Ecology*  
661 **7**: 45–75.
- 662 **Gao Y, He C, Zhang D, Liu X, Xu Z, Tian Y, Liu X-H, Zang S, Pauly M, Zhou Y, et al.**  
663 **2017.** Two trichome birefringence-like proteins mediate xylan acetylation, which is  
664 essential for leaf blight resistance in rice. *Plant Physiology* **173**: 470–481.
- 665 **Giraldo MC, Dagdas YF, Gupta YK, Mentlak TA, Yi M, Martinez-Rocha AL, Saitoh**  
666 **H, Terauchi R, Talbot NJ, Valent B. 2013.** Two distinct secretion systems facilitate  
667 tissue invasion by the rice blast fungus *Magnaporthe oryzae*. *Nature Communications* **4**.
- 668 **Gompert Z, Buerkle CA. 2011.** A hierarchical bayesian model for next-generation  
669 population genomics. *Genetics* **187**: 903–917.
- 670 **Gurney AL, Slate J, Press MC, Scholes JD. 2006.** A novel form of resistance in rice to  
671 the angiosperm parasite *Striga hermonthica*. *New Phytologist* **169**: 199–208.
- 672 **Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR,**  
673 **Wortman JR. 2008.** Automated eukaryotic gene structure annotation using  
674 EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biology*  
675 **9**: R7.
- 676 **gHan MV, Thomas GWC, Lugo-Martinez J, Hahn MW. 2013.** Estimating gene gain and  
677 loss rates in the presence of error in genome assembly and annotation using CAFE 3.  
678 *Molecular Biology and Evolution* **30**: 1987–1997.
- 679 **Hegenauer V, Slabby P, Körner M, Bruckmüller J-A, Burggraf R, Albert I, et al. 2020.** The  
680 tomato receptor CuRe1 senses a cell wall protein to identify *Cuscuta* as a pathogen. *Nature*  
681 *Communications* **11**: 5299.



- 682 **Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016.** BRAKER1:  
683 Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS:  
684 *Bioinformatics* **32**: 767–769.
- 685 **Honaas LA, Wafula EK, Yang Z, Der JP, Wickett NJ, Altman NS, Taylor CG, Yoder**  
686 **Jl, Timko MP, Westwood JH, et al. 2013.** Functional genomics of a generalist parasitic  
687 plant: Laser microdissection of host-parasite interface reveals host-specific patterns of  
688 parasite gene expression. *BMC Plant Biology* **13**: 9.
- 689 **Huang K, Whitlock R, Press MC, Scholes JD. 2012.** Variation for host range within  
690 and among populations of the parasitic plant *Striga hermonthica*. *Heredity* **108**: 96–104.
- 691 **Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M,**  
692 **Harada M, Nagayasu E, Maruyama H, et al. 2014.** Efficient *de novo* assembly of highly  
693 heterozygous genomes from whole-genome shotgun short reads. *Genome Research* **24**:  
694 1384–1395.
- 695 **Kanyuka K, Rudd JJ. 2019.** Cell surface immune receptors: the guardians of the plant's  
696 extracellular spaces. *Current Opinion in Plant Biology* **50**: 1–8.
- 697 **Kofler R, Langmuller AM, Nouhaud P, Otte KA, Schlotterer C. 2016.** Suitability of  
698 different mapping algorithms for genome-wide polymorphism scans with pool-seq data.  
699 *Genes|Genomes|Genetics* **6**: 3507–3515.
- 700 **Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A, Kosiol**  
701 **C, Schlotterer C. 2011.** PoPoolation: A toolbox for population genetic analysis of next  
702 generation sequencing data from pooled individuals. *PLoS ONE* **6**: e15925.
- 703 **Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. 2001.** Predicting  
704 transmembrane protein topology with a hidden markov model: application to complete  
705 genomes. *Journal of Molecular Biology* **305**: 567–580.
- 706 **Kuijt J. 1969.** *The Biology of Parasitic Plants*. CA: University of California Press,  
707 Berkeley.
- 708 **Kwiatos N, Ryngajłło M, Bielecki S. 2015.** Diversity of laccase-coding genes in  
709 *Fusarium oxysporum* genomes. *Frontiers in Microbiology* **6**.
- 710 **Li Y. 2003.** BRITTLE CULM1, which encodes a COBRA-Like protein, affects the  
711 mechanical properties of rice plants. *The Plant Cell* **15**: 2020–2031.

- 712 **Lu S, Edwards MC. 2016.** Genome-wide analysis of small secreted cysteine-rich  
713 proteins identifies candidate effector proteins potentially involved in *Fusarium*  
714 *graminearum* –wheat interactions. *Phytopathology* **106**: 166–176.
- 715 **Mikaberidze A, McDonald BA, Bonhoeffer S. 2015.** Developing smarter host mixtures  
716 to control plant disease. *Plant Pathology* **64**: 996–1004.
- 717 **Mitsumasu K, Seto Y, Yoshida S. 2015.** Apoplastic interactions between plants and  
718 plant root intruders. *Frontiers in Plant Science* **6**.
- 719 **Mueller AN, Ziemann S, Treitschke S, Aßmann D, Doehlemann G. 2013.**  
720 Compatibility in the *Ustilago maydis*–Maize interaction requires inhibition of host cysteine  
721 proteases by the fungal effector Pit2. *PLoS Pathogens* **9**: e1003177.
- 722 **Mutuku JM, Cui S, Hori C, Takeda Y, Tobimatsu Y, Nakabayashi R, Mori T, Saito K,**  
723 **Demura T, Umezawa T, et al. 2019.** The structural integrity of lignin is crucial for  
724 resistance against *Striga hermonthica* parasitism in rice. *Plant Physiology* **179**: 1796-  
725 1809.
- 726 **Olivier A, Benhamou N, Leroux GD. 1991.** Cell surface interactions between sorghum  
727 roots and the parasitic weed *Striga hermonthica*: cytochemical aspects of cellulose  
728 distribution in resistant and susceptible host tissues. *Canadian Journal of Botany* **69**:  
729 1679–1690.
- 730 **Olsen S, Krause K. 2017.** Activity of xyloglucan endotransglucosylases/hydrolases  
731 suggests a role during host invasion by the parasitic plant *Cuscuta reflexa*. *PLOS ONE*  
732 **12**: e0176754.
- 733 **Parker C, Riches CR. 1993.** *Parasitic weeds of the world*. UK: CAB International.
- 734 **Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011.** SignalP 4.0: discriminating  
735 signal peptides from transmembrane regions. *Nature Methods* **8**: 785.
- 736 **Qin L, Kudla, U, Roze, A.H. E, Goverse A, Popeijus H, Nieuwland J, Overmars H,**  
737 **Jones JT, Schots A, Smant G, et al. 2004.** A nematode expansin acting on plants.  
738 *Nature Brief Communications* **427**: 30.
- 739 **Rodenburg J, Cissoko M, Kayeke J, Dieng I, Khan ZR, Midega CAO, Onyuka EA,**  
740 **Scholes JD. 2015.** Do NERICA rice cultivars express resistance to *Striga hermonthica*  
741 (Del.) Benth. and *Striga asiatica* (L.) Kuntze under field conditions? *Field Crops Research*  
742 **170**: 83–94.

- 743 **Rodenburg J, Demont M, Zwart SJ, Bastiaans L. 2016.** Parasitic weed incidence and  
744 related economic losses in rice in Africa. *Agriculture, Ecosystems & Environment* **235**:  
745 306–317.
- 746 **Rodenburg J, Cissoko M, Kayongo N, Dieng I, Bisikwa J, Irakiza R, Masoka I,**  
747 **Midtega CAO, Scholes JD. 2017.** Genetic variation and host–parasite specificity of  
748 *Striga* resistance and tolerance in rice: the need for predictive breeding. *New Phytologist*  
749 **214**: 1267–1280.
- 750 **Safa SB, Jones BMG, Musselman LJ. 1984.** Mechanisms favoring outbreeding in  
751 *Striga hermonthica* [Scrophulariaceae]. *New Phytologist* **96**: 299–305.
- 752 **Saunders DGO, Win J, Cano LM, Szabo LJ, Kamoun S, Raffaele S. 2012.** Using  
753 hierarchical clustering of secreted protein families to classify and rank candidate effectors  
754 of rust fungi. *PLoS ONE* **7**: e29847.
- 755 **Scholes JD, Press MC. 2008.** *Striga* infestation of cereal crops – an unsolved problem  
756 in resource limited agriculture. *Current Opinion in Plant Biology* **11**: 180–186.
- 757 **Shahid S, Kim G, Johnson NR, Wafula E, Wang F, Coruh C, Bernal-Galeano V,**  
758 **Phifer T, dePamphilis CW, Westwood JH, et al. 2018.** MicroRNAs from the parasitic  
759 plant *Cuscuta campestris* target host messenger RNAs. *Nature* **553**: 82–85.
- 760 **Shindo T, Van Der Hoorn RAL. 2007.** Papain-like cysteine proteases: key players at  
761 molecular battlefields employed by both plants and their invaders. *Molecular Plant*  
762 *Pathology* **9**: 5299.
- 763 **Smant G, Stokkermans JPWG, Yan Y, de Boer JM, Baum TJ, Wang X, Hussey RS,**  
764 **Gommers FJ, Henrissat B, Davis EL, et al. 1998.** Endogenous cellulases in animals:  
765 Isolation of -1,4-endoglucanase genes from two species of plant-parasitic cyst  
766 nematodes. *Proceedings of the National Academy of Sciences* **95**: 4906–4911.
- 767 **Smit A, Hubley R. 2008.** RepeatModeler Open-1.0. <http://www.repeatmasker.org>.
- 768 **Smit A, Hubley R, Green P. 2013.** RepeatMasker Open-4.0. [http://www.Repeatmasker](http://www.Repeatmasker.org)  
769 [.org](http://www.Repeatmasker.org).
- 770 **Spallek T, Mutuku M, Shirasu K. 2013.** The genus *Striga* : A witch profile. *Molecular*  
771 *Plant Pathology* **14**: 861–869.

- 772 **Sun G, Xu Y, Liu H, Sun T, Zhang J, Hettenhausen C, Shen G, Qi J, Qin Y, Li J, et**  
773 **al. 2018.** Large-scale gene losses underlie the genome evolution of parasitic plant  
774 *Cuscuta australis*. *Nature Communications* **9**.
- 775 **Tajima F. 1989** Statistical method for testing the neutral mutation hypothesis by DNA  
776 polymorphism. *Genetics*. **123**: 585-595.
- 777 **Timko MP, Huang K, Lis KE. 2012.** Host resistance and parasite virulence in Striga–  
778 host plant interactions: A shifting balance of power. *Weed Science* **60**: 307–315.
- 779 **Toh S, Holbrook-Smith D, Stogios PJ, Onopriyenko O, Lumba S, Tsuchiya Y,**  
780 **Savchenko A, McCourt P. 2015.** Structure-function analysis identifies highly sensitive  
781 strigolactone receptors in *Striga*. *Science* **350**: 203–207.
- 782 **Trapnell C, Pachter L, Salzberg SL. 2009.** TopHat: discovering splice junctions with  
783 RNA-Seq. *Bioinformatics* **25**: 1105–1111.
- 784 **Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL,**  
785 **Rinn JL, Pachter L. 2012.** Differential gene and transcript expression analysis of RNA-  
786 seq experiments with TopHat and Cufflinks. *Nature Protocols* **7**: 562–578.
- 787 **Vogel JP, Raab TK, Somerville CR, Somerville SC. 2004.** Mutations in PMR5 result in  
788 powdery mildew resistance and altered cell wall composition: PMR5 is required for  
789 powdery mildew susceptibility. *The Plant Journal* **40**: 968–978.
- 790 **Vogel A, Schwacke R, Denton AK, Usadel B, Hollmann J, Fischer K, Bolger A,**  
791 **Schmidt MH-W, Bolger ME, Gundlach H, et al. 2018.** Footprints of parasitism in the  
792 genome of the parasitic flowering plant *Cuscuta campestris*. *Nature Communications* **9**.
- 793 **Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G,**  
794 **Kriventseva EV, Zdobnov EM. 2018.** BUSCO Applications from quality assessments to  
795 gene prediction and phylogenomics. *Molecular Biology and Evolution* **35**: 543–548.
- 796 **Westwood JH, Yoder JI, Timko MP, dePamphilis CW. 2010.** The evolution of  
797 parasitism in plants. *Trends in Plant Science* **15**: 227–235.
- 798 **Westwood JH, dePamphilis CW, Das M, Fernández-Aparicio M, Honaas LA, Timko**  
799 **MP, Wafula EK, Wickett NJ, Yoder JI. 2012.** The parasitic plant genome project: New  
800 tools for understanding the biology of *Orobanchaceae* and *Striga*. *Weed Science* **60**: 295–  
801 306.

802 **Westwood JH. 2013.** The physiology of the established parasite-host association. In:  
803 Parasitic Orobanchaceae. Berlin: Springer-Verlag.

804 **Win J, Chaparro-Garcia A, Belhaj K, Saunders DGO, Yoshida K, Dong S, Schornack**  
805 **S, Zipfel C, Robatzek S, Hogenhout SA, et al. 2012.** Effector biology of plant-  
806 associated organisms: Concepts and perspectives. *Cold Spring Harbor Symposia on*  
807 *Quantitative Biology* **77**: 235–247.

808 **Winnenburg R, Urban M, Beacham A, Baldwin TK, Holland S, Lindeberg M, Hansen**  
809 **H, Rawlings C, Hammond-Kosack KE, Kohler J. 2007.** PHI-base update: additions to  
810 the pathogen host interaction database. *Nucleic Acids Research* **36**: D572–D576.

811 **Wu C-H, Derevnina L, Kamoun S. 2018.** Receptor networks underpin plant immunity.  
812 *Science* **360**: 1300–1301.

813 **Yang Z, Wafula EK, Honaas LA, Zhang H, Das M, Fernandez-Aparicio M, Huang K,**  
814 **Bandaranayake PCG, Wu B, Der JP, et al. 2015.** Comparative transcriptome analyses  
815 reveal core parasitism genes and suggest gene duplication and repurposing as sources  
816 of structural novelty. *Molecular Biology and Evolution* **32**: 767–790.

817 **Yoshida S, Cui S, Ichihashi Y, Shirasu K. 2016.** The haustorium, a specialized invasive  
818 organ in parasitic plants. *Annual Review of Plant Biology* **67**: 643–667.

819 **Yoshida S, Kim S, Wafula EK, Tanskanen J, Kim Y-M, Honaas L, Yang Z, Spallek T,**  
820 **Conn CE, Ichihashi Y, et al. 2019.** Genome sequence of *Striga asiatica* provides insight  
821 into the evolution of plant parasitism. *Current Biology* **29**: 3041-3052.e4.

822 **Zheng A, Lin R, Zhang D, Qin P, Xu L, Ai P, Ding L, Wang Y, Chen Y, Liu Y, et al.**  
823 **2013.** The evolution and pathogenic mechanisms of the rice sheath blight pathogen.  
824 *Nature Communications* **4**.

825

## 826 **Figure legends**

827 **Figure. 1. Experimental strategy for the identification of *Striga hermonthica***  
828 **virulence loci.** *Striga hermonthica* (Kibos accession) were grown on susceptible  
829 (NERICA 7) and resistant (NERICA 17) rice hosts (a). The whole rice root systems show  
830 many *S. hermonthica* individuals parasitising the roots of NERICA 7 (i) whilst only two  
831 individuals (red circles) were able to overcome the resistance response of NERICA 17  
832 (ii) Scale = 1 cm. Transverse sections show *S. hermonthica* invading rice roots for a



833 representative susceptible (iii) and resistant (iv–vi) interaction seven days post  
834 inoculation. In the successful host-parasite interaction parasite intrusive cells (PIC) have  
835 breached the endodermis and have made connections with the host's xylem (iii). In the  
836 resistant rice variety several phenotypes are observed; The parasite invades the host  
837 root cortex but is unable to penetrate the suberized endodermis (iv, v); the parasite  
838 penetrates the endodermis but is unable to form connections with the host xylem (v). H  
839 = host root. P = parasite. Scale = 5  $\mu$ m. Our experimental strategy was based on the  
840 prediction that many *S. hermonthica* genotypes would grow on NERICA 7 but only highly  
841 virulent genotypes would grow on NERICA 17 (**b**). Samples of 100 *S. hermonthica* plants  
842 were bulked to generate three sequencing pools from each host variety (**c**). We expected  
843 that background loci would not differ in allele frequency between pools, but virulence  
844 alleles (and neutral alleles in linkage disequilibrium) would have increased frequency in  
845 all pools from the resistant host, allowing us to identify candidate loci (**d**).

846

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

861

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

866

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

875

876 **Figure. 5. *Striga hermonthica* secretome.** **a**, Relationship between protein length (log  
877 scale) and cysteine content (as a % of total amino acid number) for putatively-secreted  
878 (blue) and non-secreted (red) proteins in the *S. hermonthica* proteome. Secreted proteins  
879 < 500 amino acids in length and with a cysteine % > 1 standard deviation above the  
880 mean, were selected as a subset of small, cysteine rich proteins. **b**, Descriptive statistics  
881 for length and cysteine content for secreted and non-secreted proteins. **c**, Pfam domains  
882 enrichment (log fold-change) in the *S. hermonthica* secretome, relative to the proteome  
883 as a whole, compared to the corresponding enrichment in the *Mimulus guttatus*  
884 secretome. INF denotes infinite enrichment (Pfam domain only found in the secretome).  
885 Points above the 1:1 diagonal were enriched more in the *S. hermonthica* secretome  
886 relative to *M. guttatus* and have been coloured accordingly. Red symbol: domains only  
887 enriched in the *S. hermonthica* secretome. Yellow symbol: domains enriched more in the  
888 *S. hermonthica* secretome than in the *M. guttatus* secretome. Green symbol: domains  
889 enriched more in the *M. guttatus* secretome than in the *S. hermonthica* secretome. Blue  
890 symbol: domains present only in the secretome in both species. Sizes of the points were  
891 weighted according to the frequency of occurrence of each Pfam domain in the *S.*  
892 *hermonthica* secretome. Annotations for the most significantly enriched of the Pfam  
893 domains ( $p < 0.01$ ) that were also enriched more in the *S. hermonthica* secretome  
894 relative to the *M. guttatus* secretome, are given in the accompanying table with their  
895 functional descriptions.

896

897 **Figure. 6. Identification of *Striga hermonthica* genes that display significant allele**  
898 **frequency differences between pools of individuals parasitising the susceptible**

899 **rice variety (NERICA 7) and those that successfully parasitise the resistant rice**  
900 **variety (NERICA 17).** **a** Functional categorisation of non-secreted proteins and secreted,  
901 candidate virulence factors (VFs). **b** The 38 genes encoding putative secreted *S.*  
902 *hermonthica* proteins with their associated measure of differentiation (proportion of  
903 differentiating SNPs within the significant window) between the control and virulent sets  
904 of pools. The presence of SNPs in the promoter region (P), non-synonymous SNPs in  
905 the coding region (NS) and those in the intronic regions (I) are indicated with an X. The  
906 annotation of the closest matching *Arabidopsis thaliana* protein is shown along with  
907 coloured boxes that correspond to the functional category assigned in the pie chart in **a**.  
908 Tajima's D was calculated for individuals grown on NERICA 7 (Con) or NERICA 17 (Vir).  
909 **c.** Clustered gene expression profiles of the 38 candidate VFs in *S. hermonthica*  
910 haustoria parasitising NERICA 7 at 2, 4 and 7 days post-inoculation (dpi). Log<sub>2</sub> fold  
911 change in expression is shown relative to expression levels in haustoria induced *in vitro*.  
912 The gene IDs and putative functions based on best BLASTp hit against the *A. thaliana*  
913 proteome correspond with part **b**. Significant changes in gene expression in haustoria  
914 during the infection time course are shown \*\*\* (p < 0.001); \*\* (p < 0.01); \* (p < 0.05); ns  
915 non-significant (ANOVA). **d.** Comparison of Tajima's D for the 38 putative VFs (red) and  
916 all the genes in the genome (grey) for the control pools.  
917