**Preprint:**

Qiu, S. orcid.org/0000-0003-4913-2015, Bradley, J.M. 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

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

Suo Qiu[1a]**, James M. Bradley[1b]**, 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.

**Contributed equally to the work

*Corresponding authors:

**Professor Julie Scholes**

Tel: +44 07557766335; Email:j.scholes@sheffield.ac.uk

**Professor Roger Butlin**

Tel: +44 0114 222 0097

Email:r.k.butlin@sheffield.ac.uk.

## Summary

- Parasites have evolved proteins, Virulence Factors (VFs), that facilitate plant colonization, yet VFs mediating parasitic plant-host interactions are poorly understood. *Striga hermonthica* is an obligate, root-parasitic plant of cereal hosts in sub-Saharan Africa, causing devastating yield losses. Understanding the molecular nature and allelic variation of VFs in *S. hermonthica* is essential for breeding resistance and delaying the evolution of parasite virulence.
- We assembled the *S. hermonthica* genome and identified secreted proteins by *in silico* prediction. Pooled sequencing of parasites growing on a susceptible and a strongly resistant rice host allowed us to scan for loci where selection imposed by the resistant host had elevated the frequency of alleles contributing to successful colonisation.
- Thirty-eight putatively secreted VFs had extremely different allele frequencies with functions including host cell wall modification, protease inhibitors, oxidoreductase and kinase activities. These candidate loci had significantly higher Tajima's D than the genomic background, consistent with balancing selection.
- Our results reveal diverse strategies 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.

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

## Introduction

Plants are constantly challenged by parasites from across all kingdoms of life (Win *et al.,* 2012; Mitsumasu *et al.,* 2015). As a consequence, they have evolved sophisticated surveillance systems to detect and protect themselves against parasite invasion (Cook *et al.,* 2015; Wu *et al.,* 2018; Kanyuka *et al.,* 2019). 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)) (Win *et al.,* 2012; Giraldo *et al.,* 2013;

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

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

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

# Materials and Methods

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

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

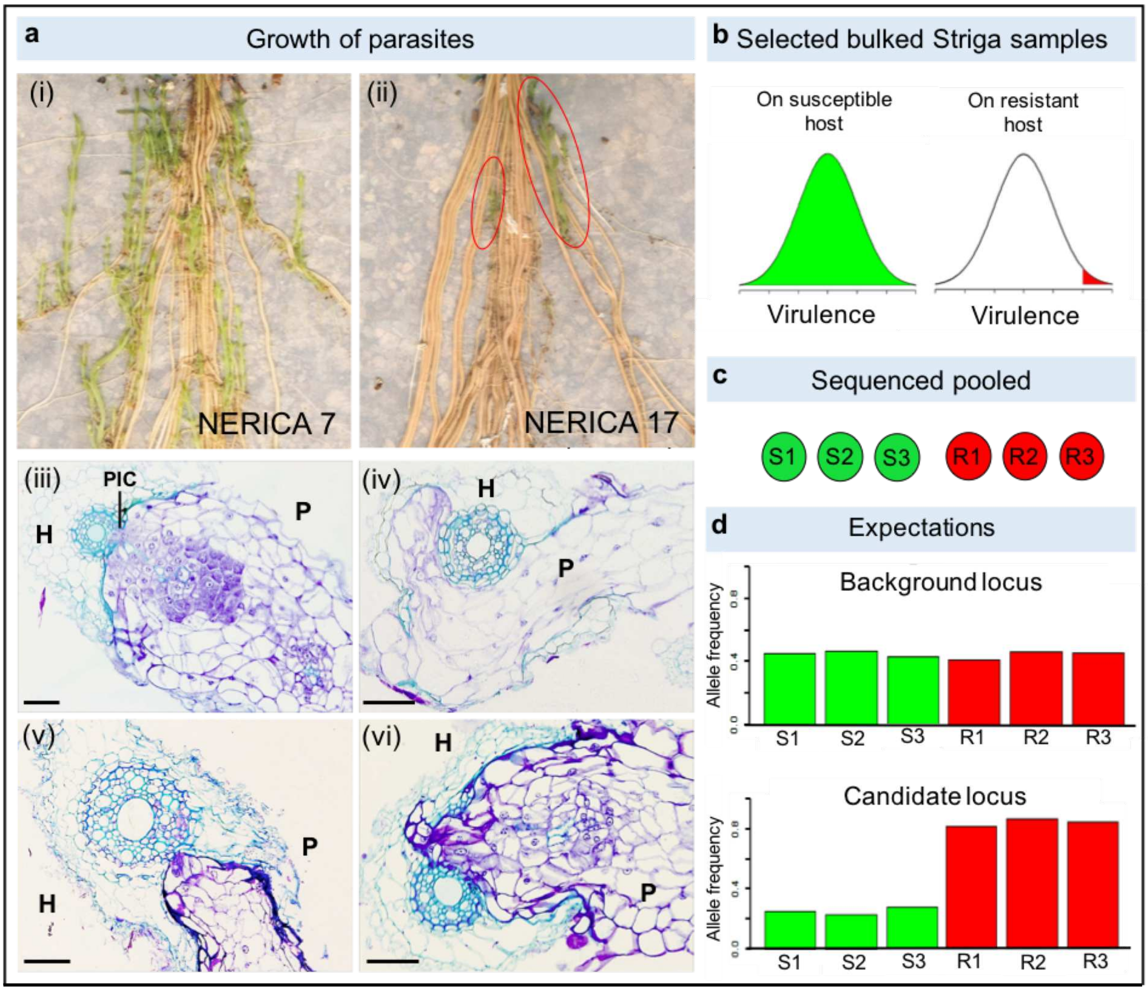


**Figure. 1. | Experimental strategy for the identification of *Striga hermonthica* virulence loci.** *Striga hermonthica* (Kibos accession) were grown on susceptible (NERICA 7) and resistant (NERICA 17) rice hosts (**a**). The whole rice root systems show many *S. hermonthica* individuals parasitising 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 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**).

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

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

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

To further improve the *S. hermonthica* genome assembly, Chicago and Dovetail Hi-C libraries were prepared and sequenced at Dovetail Genomics, California, USA (https://dovetailgenomics.com/plant-animal/) (Table S3). For construction of Chicago libraries, DNA from the same *S. hermonthica* individual (used for initial sequencing) was sequenced on an Illumina HiSeq 2500 platform. For the Hi-C libraries, plant tissues from an F1 individual from a cross between the sequenced individual and another *S. hermonthica* individual (Kibos accession) were used for the library construction and sequencing. Sequences from both the Chicago and Hi-C libraries were used only to improve the contiguity of the initial genome assembly using the Dovetail HiRise Assembler software.  RepeatModeler was used to generate a *S. hermonthica*-specific repeat library and RepeatMasker was then used to classify repeat elements in the genome. A repeat-masked version of the genome was used for annotation (Smit *et al.,* 2008; 2013).

## Annotation of the *S. hermonthica* genome

The genome was annotated using three methods. Firstly, gene structures were inferred using a *S. hermonthica* transcriptome dataset of cDNAs collected from *S. hermonthica* individuals at eight developmental stages, generated by the Parasitic Plant Genome Project (PPGP) (Westwood *et al.,* 2012; Yang *et al.,* 2015). The reads were mapped onto the *S. hermonthica* genome assembly using TopHat to identify exon regions and splice positions (Trapnell *et al.,* 2009). Transcriptome-based gene structures were predicted using Cufflinks (Trapnell *et al.,* 2012) and candidate coding regions were then constructed in Transdecoder (https://github.com/TransDecoder/). Secondly, protein sequences from *Arabidopsis thaliana* (TAIR10), *Mimulus guttatus* (v2.0), *Solanum lycopersicum* (ITAG2.4), *Oryza sativa* (IRGSP1.0) and *Sorghum bicolor* (79), were used to determine consensus gene models in the genome. The protein sequences were mapped onto the *S. hermonthica* genome using TBLASTN and pairwise alignments were then input into Genewise (Birney 2004) to predict gene models in *S. hermonthica*. Thirdly, an *ab initio* method was used for *de novo* prediction of genes in the *S. hermonthica* genome using the software, Braker (Hoff *et al.,* 2016). Finally, Evidence Gene Modeler was used to integrate various gene models from the transcript data, mapped proteins, and the predicted gene models from the *ab initio* method (Haas *et al.,* 2008). The completeness of the gene set was assessed using BUSCO v5 using the 2,326 core orthologs from eudicots_odb10, with default settings.

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

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

## Inference of orthogroups (OG)

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

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

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

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

The raw sequence reads from the six pools were trimmed and filtered for coverage (see Methods S1). The likelihood of the observed read counts for the two most common alleles, across the six pools was calculated according to equation 3 from Gompert and Buerkle (2011) to allow for the two levels of sampling associated with pooled sequencing data (sampling of reads and of individuals). We compared three allele-frequency models for each SNP using the Akaike information criterion (AIC): a null model with a single allele frequency for all pools, a control-virulent model with one frequency for the control pools (from the NERICA-7 host) and one for the virulent pools (from the NERICA-17 host) and a replicate model with a different allele frequency for each of the three pairs of pools (one control and one virulent) that were sequenced together. The control-virulent model was the model of interest while the replicate model was intended to check for consistency across pairs of pools. Therefore, two ΔAIC values were obtained: ΔAICcv = AICnull - AICcontrol-virulent and ΔAICrep = AICcontrol-virulent – AICreplicate. High positive values of ΔAICcv represent better fits than the null model and indicate significant differences between control and virulent pool types. SNPs with positive ΔAICrep values were likely to be affected by artefacts caused by sequencing methods and were excluded from the following analyses. All analysis steps were repeated independently for SNPs based on BWA and NOVOALIGN mapping as recommended by Kofler *et al.,* (2016).

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

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

Two population statistics were calculated for each genic region in the control pool using the software Popoolation (Kofler et al., 2011). These were nucleotide diversity (π) and Tajima's D, a statistic describing the allele frequency spectrum used for testing whether a DNA sequence is evolving under a process that departs from the standard neutral model, such as selection or demographic change (Tajima, 1989). See Methods S1 for details.

### Analyses of candidate virulence genes

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

### Expression profiling of candidate virulence genes

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

## Results

### 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

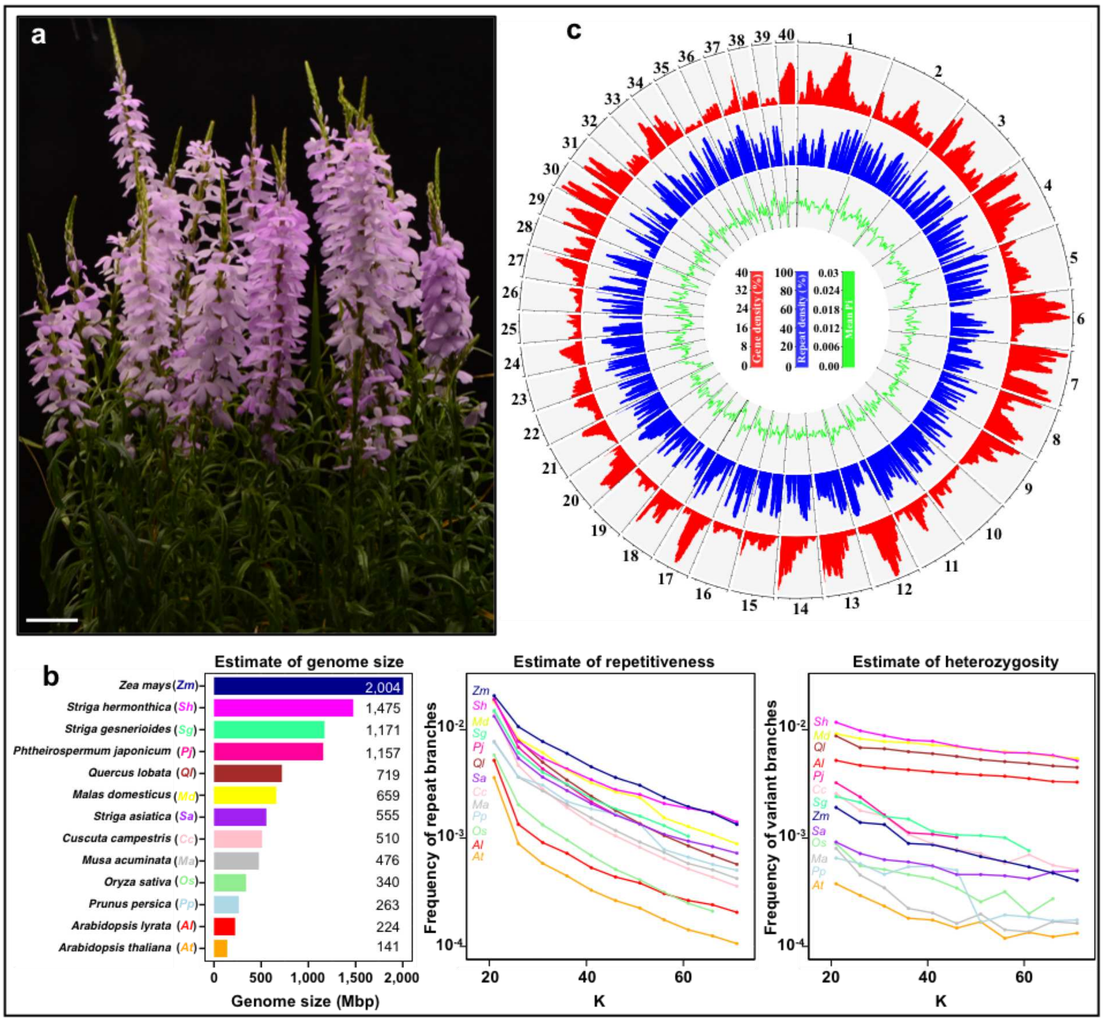


**Figure. 2. | *Striga hermonthica* is an obligate outbreeding parasitic plant with a highly heterozygous and repetitive genome. a,** Flowering *S. hermonthica* growing on the rice host, NERICA 7, derived from a seed batch collected from the Kibos region of Kenya. Scale = 5 cm. **b,** Comparison of genome 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.

characterised the genome of a single individual from this population, which to our knowledge, represents the first genome assembly for *S. hermonthica*. The genome size was estimated by K-mer analysis to be 1,475 Mbp, (Fig. 2b). This agrees closely with a previous flow cytometry-based estimate (Estep *et al.,* 2012) and is more than twice the size of the recently sequenced genome of *S. asiatica* (Yoshida *et al.,* 2019). The assembly consisted of 34,907 scaffolds > 1 kbp in length, with an N50 of 10.0 Mbp and 29 scaffolds making up half of the genome size (Table S3). The *S. hermonthica* genome was remarkably heterozygous (overall mean $\pi = 0.011$) (Fig. 2 b,c) when compared with other parasitic and non-parasitic plant genomes, likely reflecting the fact that it is an obligate outbreeding species. In addition, the genome contained a large proportion (69%) of repetitive DNA (Fig 2 b,c), dominated by long terminal repeat (LTR) elements (Table S4), a pattern also found for the shoot-parasitic plants, *Cuscuta australis* and *C. campestris* (Sun *et al.,* 2018; Vogel *et al.,* 2018) and the closely related parasitic plant *S. asiatica* (Yoshida *et al.,* 2019). As expected, the density of repetitive elements along each scaffold negatively correlated with the density of protein-coding genes (Fig 2c). In total, 29,518 protein-coding genes were predicted from the *S. hermonthica* genome, which was comparable to *S. asiatica* (34,577), the closely related non-parasitic plant *Mimulus guttatus* (28,140) and to *Arabidopsis thaliana* (27,416) (Table S5).

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

Comparative analysis of orthologous gene groups (orthogroups) between *S. hermonthica* and 12 other plant species identified 22,624 orthogroups in total, of which 12,278 contained *S. hermonthica* genes. Of these, 327 were significantly expanded and 104 were contracted in the *S. hermonthica* genome (Fig. 4a). Expanded orthogroups included the $\alpha/\beta$-hydrolase family, recently shown to have undergone duplication in *S. hermonthica* (Toh *et al.,* 2015), as well as numerous F-box, leucine-rich repeat and protein kinase domain-containing proteins (Fig. 4b). Of particular interest in the context of pathogenicity were *S. hermonthica*-specific orthogroups annotated as papain family cysteine proteases, xylanase inhibitors and trypsin and protease inhibitors (Fig. 4b). Both proteases and protease inhibitors function in a wide range of plant-plant parasite interactions and may act offensively, by degrading host proteins, or defensively, by inhibiting host defence enzymes (Bleischwitz *et al.,* 2010; Mueller *et al.,* 2013).

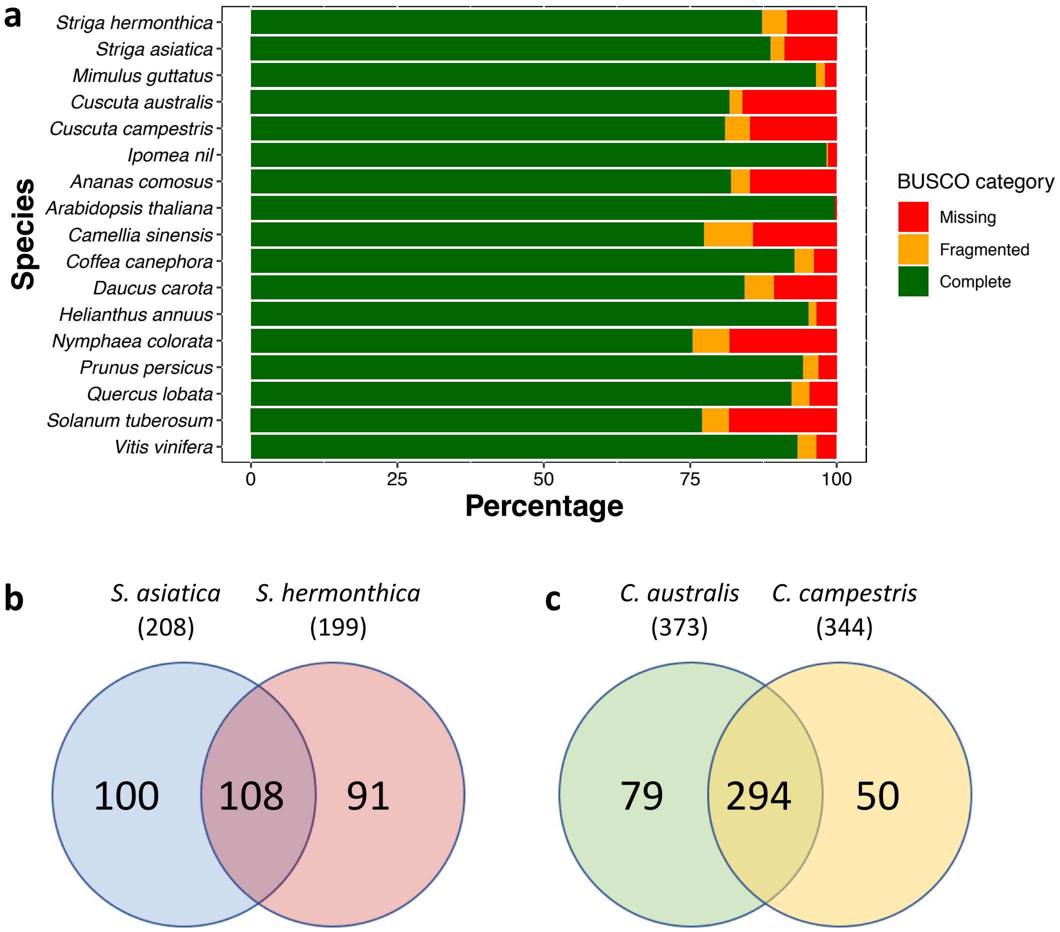


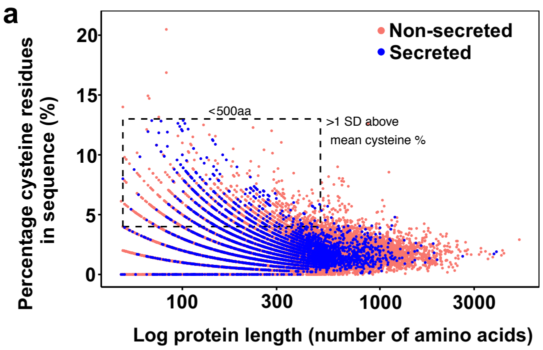
**Figure. 3 | a**, BUSCO completeness analysis for *Striga hermonthica* genome, compared with 16 other published plant genomes. The number of missing BUSCOs for two Striga **b** and two Cuscuta species **c**. The overlap shows genes that are missing from both Striga or Cuscuta species respectively.

**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.

## The *S. hermonthica* secretome

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

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

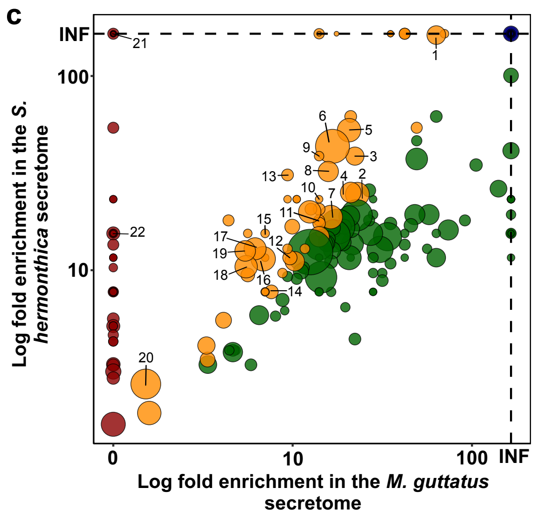


**b**

| | | 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 < $2.2e^{-16}$)



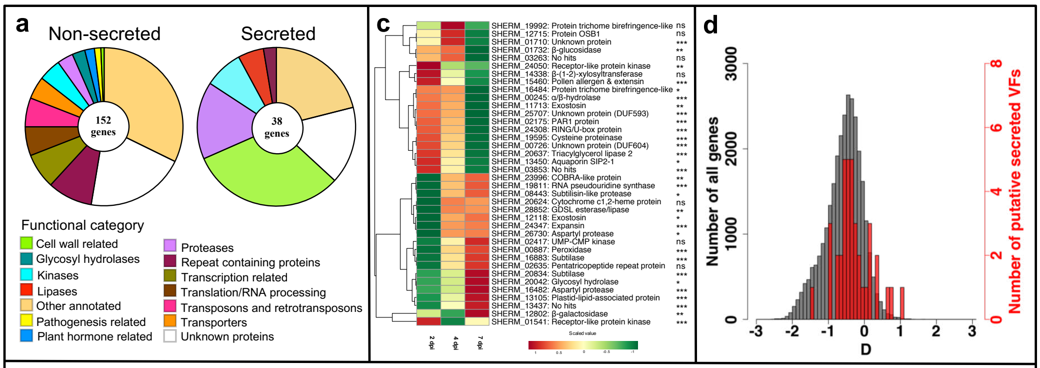
| Key | Pfam accession | Pfam domain description |
|---|---|---|
| 1 | PF02469: | Fasciclin domain |
| 2 | PF14368: | Probable lipid transfer |
| 3 | PF00759: | Glycosyl hydrolase family 9 (cellulase) |
| 4 | PF00722: | Glycosyl hydrolases family 16 |
| 5 | PF05938: | Plant self-incompatibility protein 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 |

**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.

Xyloglucan endotransglucosylases / hydrolases (XETs) have the potential to modify either the parasite or host cell walls (or both) during parasitism (Olsen & Krause 2017). XETs are secreted from the haustoria of the parasitic plant *Cuscuta reflexa* during a susceptible interaction on its host *Pelargonium zonale,* contributing towards pathogenicity (Olsen & Krause 2017). Pectate lyase superfamily (PF12708) and pectinesterase (PF01095) domains were enriched in the secretome of *S. hermonthica* compared to *M. guttatus* and may act as VFs to modify host, or parasite, pectin during penetration. We found a battery of different carbohydrate-active glycosyl hydrolase (GH) domains that were enriched in the *S. hermonthica* secretome (Fig. 5c, Fig. S3). Eight *S. hermonthica* proteins were annotated as cellulases of the GH5 family (containing domain PF00150) (Fig. S4) and were similar to secreted cellulases that function as VFs in some phytoparasitic nematodes (Smant *et al.,* 1998). The degradation of cellulosic β-1,4-glucans has been observed in susceptible sorghum roots infected by *S. hermonthica* (Olivier *et al.,* 1991) and may be mediated by these secreted enzymes to facilitate the migration of *S. hermonthica* intrusive cells between host root cortical cells. The identification of many putatively secreted VFs in the *S. hermonthica* genome, that are likely to modify host plant cell walls, raises an interesting question about how such proteins are targeted to avoid damaging the parasite's own cell walls.

## Population genomic analysis to identify candidate virulence loci

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



**b**

| Gene ID | Proportion of differentiating SNPs | SNP presence |  |  | Tajima's D |  |  | Annotation of closest matching *Arabidopsis thaliana* protein (BLASTp) | Pfam domain(s) descriptions |
|---|---|---|---|---|---|---|---|---|---|
|  |  | P | NS | I | Con | Vir |  |  |  |
| 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 |

**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). $\text{Log}_2$ 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.

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

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

The 38 candidate VFs were investigated in more detail by quantifying changes in gene expression in haustoria at critical stages of parasite development on the susceptible rice variety NERICA-7 by inspecting the distribution of SNPs throughout the promoter and genic regions, and testing for signatures of historical selection. Gene expression was measured in an independent experiment (Fig. 6c). Changes in gene expression of attached haustoria were measured relative to gene expression in haustoria generated *in vitro.* At 2 days after inoculation of the host root, parasite haustoria were attached and parasite intrusive cells had penetrated into the host root cortex. By day 4, the parasite intrusive cells had penetrated between the endodermal cells and by day 7 had formed connections with the xylem vessels of the host, providing direct access to host resources (Fig. 1a iii).

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

Most of the 38 genes had significantly differentiating SNPs in their promoter regions (from the start site to 2 kbp up-stream). Some of these SNPs may lead to a change in the regulation of gene expression (Fig 6b). Some genes, for example, the gene encoding the pollen Ole e allergen protein (SHERM_15460), one of the exostosin family proteins SHERM_12118), a probable aquaporin SIP2-1 (SHERM_13450) and one of the two protein TBL genes (SHERM_16484), also had non-synonymous SNPs in the coding region (Fig. 6b) that may result in functional differences between the alleles of these genes in individuals infecting NERICA-7 and NERICA-17. Finally, SNPs were also found within predicted intron regions in many of the genes (Fig. 6b).

The co-evolutionary interactions between hosts and parasites can generate balancing selection (Frank 1993). We predicted that genes contributing to virulence would tend to have a history of balancing selection because of the diverse range of hosts used by *S. hermonthica*. To test this prediction, we compared Tajima's D between candidate loci and the rest of the genome, expecting to see more positive values (Charlesworth 2006). We used the pools from the susceptible host for this comparison because they represented the Striga population as a whole. As predicted, the 152 candidate loci in the *S. hermonthica* proteome (Fig S6) and the 38 candidate loci in the secretome (Fig. 6d) had significantly elevated Tajima's D, on average, compared to all the genes in the genome ($p < 0.0001$ and $p < 0.0003$, respectively; $10^5$ permutations). Some loci had particularly high Tajima's D values, for example the two receptor-like protein kinases (Fig. 6b). Interestingly, some loci showed large differences in Tajima's D between the control and virulent *S. hermonthica* pools with the largest difference seen for the TBL gene (SHERM_16484) with a negative ΔD ($D_{Vir}$ - $D_{Con}$) of -0.9. This suggests strong selection resulting in one common haplotype in the virulent pools in contrast to two or more haplotypes at intermediate frequencies in the control pools. There were also large positive ΔD values: 0.71, 0.16 and 0.20 for one of the putative receptor-like protein kinases SHERM_01541, one of the aspartyl proteases, SHERM_16482, and the peroxidase SHERM_00887, respectively. This suggests that a rare haplotype in the control pools is present at intermediate frequency in the virulent pools. Overall, these changes indicate that selection on the resistant host caused changes in frequency of multi-SNP haplotypes at these loci, haplotypes that may have been created by areas of low recombination or by recent invasion of new variants under positive selection (Cutter & Payseur 2013) and which underlie the ability of some *S. hermonthica* individuals to overcome resistance in NERICA-17.

# Discussion

Plants secrete proteins involved in many biological functions, from nutrient acquisition, to development and defence (Li 2003; Cook *et al.,* 2015). However, unlike most plants, in parasitic plants such as *S. hermonthica* a subset of secreted proteins is likely to function as VFs and contribute towards parasite fitness by facilitating host colonization (Timko *et al.,* 2012). We used a combination of *in silico* prediction of secreted proteins and pooled sequencing of parasites derived from susceptible and resistant rice hosts, both facilitated by the first available genome assembly, to identify a set of candidate VFs. These are secreted proteins encoded by genes that had extremely different allele frequencies between replicated pools derived from susceptible and resistant hosts, suggesting strong selection for particular variants that facilitate successful colonisation despite host resistance. This experimental approach has not been applied previously to investigate virulence of Striga, or any other parasitic plant. Its success here paves the way to application of similar methods to other host-parasite combinations, providing vital information on virulence mechanisms and their genetic variability within and between parasitic plant populations from different regions of Africa, and so underpinning the development of sustainable control strategies.

Our list of 38 candidate, secreted, VFs points to key functions involved in pathogenicity, including oxidoreductase, receptor-like protein kinase, protease and protease inhibitor, and cell wall modification activities. The latter is consistent with growing evidence that cell-wall modification is a critical step in plant invasions by many different parasites including parasitic plants. Recently, the structural integrity of lignin was shown to be a crucial component of resistance in roots of the rice variety Nipponbare to infection by *S. hermonthica* (Mutuku *et al.,* 2019). In our study the host cell wall is clearly involved in resistance in NERICA-17. Most *S. hermonthica* individuals from the Kibos population were unable to penetrate the root endodermis or, if they breached the endodermis, they were unable to establish functional connections to the host xylem vessels (Fig. 1a iv-vi). Consistent with this, the largest category of our candidate, secreted VFs included a putative peroxidase, an expansin, pollen allergen-like proteins, a β-glucosidase, a β (1-2) xylosyltransferase, and a TBL protein, all of which function to modify cell walls. The TBL protein, SHERM_16484, had a strikingly different Tajima's D in the control pool compared to the value in the virulent pool, consistent with selection favouring one haplotype on the resistant NERICA-17, out of several haplotypes present in the population. In *A. thaliana* and *O. sativa* TBL proteins belong to large gene families with functions related to cell wall modifications. In *A. thaliana*, At-TBL44 has been implicated in pectin esterification (Vogel *et al.,* 2004; Bacete *et al.,* 2018), whilst in rice other members of this family appear to be involved in acetylation of xylan moieties in cell walls (Gao *et al.,* 2017). In each case, alterations in enzyme activity altered resistance in *A. thaliana* to powdery mildew and in rice to leaf blight (Vogel *et al.,* 2004; Gao *et al.,* 2017). Recently an 11 kDa protein was isolated from the cell wall of the shoot parasite *C. reflexa* and identified as a glycine rich protein (GRP) (Hegenauer *et al.,* 2020). The protein and its minimal peptide epitope (Crip21) bind to and activate a cell surface resistance gene in tomato (CuRe1), leading to resistance to the parasite, illustrating the importance of cell wall modifications to host resistance.

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

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

This type of parasite interaction with multiple hosts leads to two predictions that are supported by our data. First, multiple loci, potentially with a wide range of functions, are likely to be implicated in overcoming host resistance. We detected 190 strong candidates for contribution to virulence, with extreme allele frequency differences between our control and virulent pools, including many gene families. It is likely that many additional candidate VFs would be revealed, by repeating this comparison on other resistant hosts. An important question for the future will be to determine how individual VFs are implicated in overcoming resistance for specific hosts or across a range of hosts. Second, maintenance of variation at virulence loci by balancing selection will lead to elevated Tajima's D relative to the background, reflecting persistence of multiple alleles at these loci. We found the overall Tajima's D in *S. hermonthica* to be negative, perhaps reflecting population expansion following the spread of agriculture, but our candidate loci had significantly higher Tajima's D on average, consistent with balancing selection on these loci. 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 (Mikaberidze *et al.,* 2015).

Effective control of *S. hermonthica* is essential for food security and poverty alleviation for small-holder subsistence farmers, but it remains elusive.  The use of resistance crop varieties is recognised as sustainable and cost effective (Scholes *et al.,* 2008), but the durability of resistant varieties is compromised by the potential for rapid evolution of parasite virulence. Thus, the long-term success of host resistance, as a control strategy for S*. hermonthica* and other parasitic weeds, requires knowledge of the virulence factors involved, their allelic variation within and between Striga populations and their interaction with different host resistance alleles. Only then 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 experimental approach and identification of candidate VFs and allelic variation within a *S. hermonthica* population, is a critical first step in this direction.

## Acknowledgements

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

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

## Author contributions

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

## Data Availability

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

## References

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

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

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

**Benjamini Y, Hochberg Y. 1995**. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**: 289–300.

**Birney E. 2004**. GeneWise and genomewise. *Genome Research* **14**: 988–995.

**Bleischwitz M, Albert M, Fuchsbauer H-L, Kaldenhoff R. 2010**. Significance of Cuscutain, a cysteine protease from *Cuscuta reflexa*, in host-parasite interactions. *BMC Plant Biology* **10**: 8.

**Cai L, Arnold BJ, Xi Z, Khost DE, Patel N, Hartmann CB, Manickam S, Sasirat S, Nikolov LA, Mathews S, *et al.* 2021**. Deeply altered genome architecture in the endoparasitic flowering plant *Sapria himalayana* Griff. (Rafflesiaceae). *Current Biology* **31**: 1002-1011.e9.

**Charlesworth D. 2006**. Balancing selection and its effects on eequences in nearby genome regions. *PLoS Genetics* **2**: 6.

**Cissoko M, Boisnard A, Rodenburg J, Press MC, Scholes JD. 2011**. New Rice for Africa (NERICA) cultivars exhibit different levels of post-attachment resistance against the parasitic weeds *Striga hermonthica* and *Striga asiatica*. *New Phytologist* **192**: 952–963.

**Clarke CR, Timko MP, Yoder JI, Axtell MJ, Westwood JH. 2019**. Molecular dialog between parasitic plants and their Hosts. *Annual Review of Phytopathology* **57**.

**Cook DE, Mesarich CH, Thomma BPHJ. 2015**. Understanding plant immunity as a surveillance system to detect invasion. *Annual Review of Phytopathology* **53**: 541–563.

**Cutter AD, Payseur BA. 2013**. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Reviews Genetics* **14**: 262–274.

**Elizondo LI, Jafar-Nejad P, Clewing JM, Boerkoel CF 2009**. Gene clusters, molecular evolution and disease: A speculation. *Current Genetics.* **12**: 64-75.

**Emms DM, Kelly S. 2015**. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**.

**Eoche-Bosy D, Gautier M, Esquibet M, Legeai F, Bretaudeau A, Bouchez O, et al. 2017** Genomic scans on experimentally evolved populations reveal candidate regions for adaptation to plant resistance in the potato cyst nematode *Globodera pallida*. *Molecular Ecology* **26**: 4700-4711.

**Estep MC, Gowda BS, Huang K, Timko MP, Bennetzen JL**. **2012**. Genomic characterization for parasitic weeds of the genus by sample sequence analysis. *The Plant Genome Journal* **5**: 30.

**Fernández-Aparicio M, Huang K, Wafula EK, Honaas LA, Wickett NJ, Timko MP, dePamphilis CW, Yoder JI, Westwood JH**. **2013**. Application of qRT-PCR and RNA-Seq analysis for the identification of housekeeping genes useful for normalization of gene expression values during *Striga hermonthica* development. *Molecular Biology Reports* **40**: 3395–3407.

**Frank SA**. **1993**. Coevolutionary genetics of plants and pathogens. *Evolutionary Ecology* **7**: 45–75.

**Gao Y, He C, Zhang D, Liu X, Xu Z, Tian Y, Liu X-H, Zang S, Pauly M, Zhou Y, *et al.*** **2017**. Two trichome birefringence-like proteins mediate xylan acetylation, which Is essential for leaf blight resistance in rice. *Plant Physiology* **173**: 470–481.

**Giraldo MC, Dagdas YF, Gupta YK, Mentlak TA, Yi M, Martinez-Rocha AL, Saitoh H, Terauchi R, Talbot NJ, Valent B**. **2013**. Two distinct secretion systems facilitate tissue invasion by the rice blast fungus *Magnaporthe oryzae*. *Nature Communications* **4**.

**Gompert Z, Buerkle CA**. **2011**. A hierarchical bayesian model for next-generation population genomics. *Genetics* **187**: 903–917.

**Gurney AL, Slate J, Press MC, Scholes JD**. **2006**. A novel form of resistance in rice to the angiosperm parasite *Striga hermonthica*. *New Phytologist* **169**: 199–208.

**Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR**. **2008**. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**: R7.

**gHan MV, Thomas GWC, Lugo-Martinez J, Hahn MW**. **2013**. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecular Biology and Evolution* **30**: 1987–1997.

**Hegenauer V, Slabby P, Körner M, Bruckmüller J-A, Burggraf R, Albert I, et al. 2020**. The tomato receptor CuRe1 senses a cell wall protein to identify *Cuscuta* as a pathogen. *Nature Communications* **11**: 5299.

**Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M**. **2016**. BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS: *Bioinformatics* **32**: 767–769.

**Honaas LA, Wafula EK, Yang Z, Der JP, Wickett NJ, Altman NS, Taylor CG, Yoder JI, Timko MP, Westwood JH, *et al.* 2013**. Functional genomics of a generalist parasitic plant: Laser microdissection of host-parasite interface reveals host-specific patterns of parasite gene expression. *BMC Plant Biology* **13**: 9.

**Huang K, Whitlock R, Press MC, Scholes JD**. **2012**. Variation for host range within and among populations of the parasitic plant *Striga hermonthica*. *Heredity* **108**: 96–104.

**Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, *et al.* 2014**. Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research* **24**: 1384–1395.

**Kanyuka K, Rudd JJ**. **2019**. Cell surface immune receptors: the guardians of the plant's extracellular spaces. *Current Opinion in Plant Biology* **50**: 1–8.

**Kofler R, Langmuller AM, Nouhaud P, Otte KA, Schlotterer C. 2016.** Suitability of different mapping algorithms for genome-wide polymorphism scans with pool-seq data. *Genes|Genomes|Genetics* **6**: 3507–3515.

**Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A, Kosiol C, Schlötterer C**. **2011**. PoPoolation: A toolbox for population genetic gnalysis of next generation sequencing data from pooled Individuals. *PLoS ONE* **6**: e15925.

**Krogh A, Larsson B, von Heijne G, Sonnhammer ELL**. **2001**. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology* **305**: 567–580.

**Kuijt J**. **1969**. *The Biology of Parasitic Plants*. CA: University of California Press, Berkeley.

**Kwiatos N, Ryngajłło M, Bielecki S**. **2015**. Diversity of laccase-coding genes in *Fusarium oxysporum* genomes. *Frontiers in Microbiology* **6**.

**Li Y**. **2003**. BRITTLE CULM1, which encodes a COBRA-Like protein, affects the mechanical properties of rice plants. *The Plant Cell* **15**: 2020–2031.

**Lu S, Edwards MC. 2016**. Genome-wide analysis of small secreted cysteine-rich proteins identifies candidate effector proteins potentially involved in *Fusarium graminearum* −wheat interactions. *Phytopathology* **106**: 166–176.

**Mikaberidze A, McDonald BA, Bonhoeffer S. 2015**. Developing smarter host mixtures to control plant disease. *Plant Pathology* **64**: 996–1004.

**Mitsumasu K, Seto Y, Yoshida S. 2015**. Apoplastic interactions between plants and plant root intruders. *Frontiers in Plant Science* **6**.

**Mueller AN, Ziemann S, Treitschke S, Aßmann D, Doehlemann G. 2013**. Compatibility in the *Ustilago maydis*–Maize interaction requires inhibition of host cysteine proteases by the fungal effector Pit2. *PLoS Pathogens* **9**: e1003177.

**Mutuku JM, Cui S, Hori C, Takeda Y, Tobimatsu Y, Nakabayashi R, Mori T, Saito K, Demura T, Umezawa T, *et al.* 2019**. The structural integrity of lignin is crucial for resistance against *Striga hermonthica* parasitism in rice. *Plant Physiology* 179: 1796-1809.

**Olivier A, Benhamou N, Leroux GD. 1991**. Cell surface interactions between sorghum roots and the parasitic weed *Striga hermonthica* : cytochemical aspects of cellulose distribution in resistant and susceptible host tissues. *Canadian Journal of Botany* **69**: 1679–1690.

**Olsen S, Krause K. 2017**. Activity of xyloglucan endotransglucosylases/hydrolases suggests a role during host invasion by the parasitic plant *Cuscuta reflexa*. *PLOS ONE* **12**: e0176754.

**Parker C, Riches CR. 1993**. *Parasitic weeds of the world*. UK: CAB International.

**Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011**. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* **8**: 785.

**Qin L, Kudla, U, Roze, A.H. E, Goverse A, Popeijus H, Nieuwland J, Overmars H, Jones JT, Schots A, Smant G, *et al.* 2004**. A nematode expansin acting on plants. *Nature Brief Communications* **427**: 30.

**Rodenburg J, Cissoko M, Kayeke J, Dieng I, Khan ZR, Midega CAO, Onyuka EA, Scholes JD. 2015**. Do NERICA rice cultivars express resistance to *Striga hermonthica* (Del.) Benth. and *Striga asiatica* (L.) Kuntze under field conditions? *Field Crops Research* **170**: 83–94.

**Rodenburg J, Demont M, Zwart SJ, Bastiaans L. 2016**. Parasitic weed incidence and related economic losses in rice in Africa. *Agriculture, Ecosystems & Environment* **235**: 306–317.

**Rodenburg J, Cissoko M, Kayongo N, Dieng I, Bisikwa J, Irakiza R, Masoka I, Midega CAO, Scholes JD. 2017**. Genetic variation and host–parasite specificity of *Striga* resistance and tolerance in rice: the need for predictive breeding. *New Phytologist* **214**: 1267–1280.

**Safa SB, Jones BMG, Musselman LJ. 1984**. Mechanisms favoring outbreeding in *Striga hermonthica* [Scrophulariaceae]. *New Phytologist* **96**: 299–305.

**Saunders DGO, Win J, Cano LM, Szabo LJ, Kamoun S, Raffaele S. 2012**. Using hierarchical clustering of secreted protein families to classify and rank candidate effectors of rust fungi. *PLoS ONE* **7**: e29847.

**Scholes JD, Press MC. 2008**. Striga infestation of cereal crops – an unsolved problem in resource limited agriculture. *Current Opinion in Plant Biology* **11**: 180–186.

**Shahid S, Kim G, Johnson NR, Wafula E, Wang F, Coruh C, Bernal-Galeano V, Phifer T, dePamphilis CW, Westwood JH, *et al.* 2018**. MicroRNAs from the parasitic plant *Cuscuta campestris* target host messenger RNAs. *Nature* **553**: 82–85.

**Shindo T, Van Der Hoorn RAL. 2007**. Papain-like cysteine proteases: key players at molecular battlefields employed by both plants and their invaders. *Molecular Plant Pathology* **9**: 5299.

**Smant G, Stokkermans JPWG, Yan Y, de Boer JM, Baum TJ, Wang X, Hussey RS, Gommers FJ, Henrissat B, Davis EL, *et al.* 1998**. Endogenous cellulases in animals: Isolation of -1,4-endoglucanase genes from two species of plant-parasitic cyst nematodes. *Proceedings of the National Academy of Sciences* **95**: 4906–4911.

**Smit A, Hubley R. 2008**. RepeatModeler Open-1.0. *http://www.repeatmasker.org.*

**Smit A, Hubley R, Green P. 2013**. RepeatMasker Open-4.0. *http://www. Repeatmasker .org*.

**Spallek T, Mutuku M, Shirasu K. 2013**. The genus *Striga* : A witch profile. *Molecular Plant Pathology* **14**: 861–869.

**Sun G, Xu Y, Liu H, Sun T, Zhang J, Hettenhausen C, Shen G, Qi J, Qin Y, Li J, *et al.* 2018**. Large-scale gene losses underlie the genome evolution of parasitic plant *Cuscuta australis*. *Nature Communications* **9**.

**Tajima F**. **1989** Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. **123**: 585-595.

**Timko MP, Huang K, Lis KE**. **2012**. Host resistance and parasite virulence in Striga–host plant interactions: A shifting balance of power. *Weed Science* **60**: 307–315.

**Toh S, Holbrook-Smith D, Stogios PJ, Onopriyenko O, Lumba S, Tsuchiya Y, Savchenko A, McCourt P**. **2015**. Structure-function analysis identifies highly sensitive strigolactone receptors in Striga. *Science* **350**: 203–207.

**Trapnell C, Pachter L, Salzberg SL**. **2009**. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105–1111.

**Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L**. **2012**. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7**: 562–578.

**Vogel JP, Raab TK, Somerville CR, Somerville SC**. **2004**. Mutations in PMR5 result in powdery mildew resistance and altered cell wall composition: PMR5 is required for powdery mildew susceptibility. *The Plant Journal* **40**: 968–978.

**Vogel A, Schwacke R, Denton AK, Usadel B, Hollmann J, Fischer K, Bolger A, Schmidt MH-W, Bolger ME, Gundlach H, *et al.* 2018**. Footprints of parasitism in the genome of the parasitic flowering plant *Cuscuta campestris*. *Nature Communications* **9**.

**Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM**. **2018**. BUSCO Applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution* **35**: 543–548.

**Westwood JH, Yoder JI, Timko MP, dePamphilis CW**. **2010**. The evolution of parasitism in plants. *Trends in Plant Science* **15**: 227–235.

**Westwood JH, dePamphilis CW, Das M, Fernández-Aparicio M, Honaas LA, Timko MP, Wafula EK, Wickett NJ, Yoder JI**. **2012**. The parasitic plant genome project: New tools for understanding the biology of Orobanche and Striga. *Weed Science* **60**: 295–306.

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

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

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

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

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

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

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

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

## Figure legends

**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 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**).

**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.

**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.

**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.

**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.

**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). $\text{Log}_2$ 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.