Title

The genome of the oomycete Peronosclerospora sorghi, a cosmopolitan pathogen of maize and sorghum, is inflated with dispersed pseudogenes

Journal

G3: Genes, Genomes, Genetics, 13(3)

ISSN

2160-1836

Authors

Fletcher, Kyle

Martin, Frank

Isakeit, Thomas

et al.

Publication Date

2023-03-09

DOI

10.1093/g3journal/jkac340

G3, 2023, 13(3), jkac340
https://doi.org/10.1093/g3journal/jkac340
Advance Access Publication Date: 2 January 2023
Investigation

The genome of the oomycete *Peronosclerospora sorghi*, a cosmopolitan pathogen of maize and sorghum, is inflated with dispersed pseudogenes

Kyle Fletcher [1], Frank Martin,[2] Thomas Isakeit,[3] Keri Cavanaugh,[1] Clint Magill,[3] Richard Michelmore [1,4,*]

[1]The Genome Center, University of California, Davis, CA 95616, USA
[2]U.S. Department of Agriculture–Agriculture Research Service, Salinas, CA, 93905, USA
[3]Department of Plant Pathology and Microbiology, Texas A&M University, College Station, TX 77843, USA
[4]Departments of Plant Sciences, Molecular & Cellular Biology, Medical Microbiology & Immunology, University of California, Davis, CA 95616, USA

*Corresponding author: University of California, Davis, The Genome Center, One Shields Avenue, Davis, California 95616 USA. Email: rwmichelmore@ucdavis.edu

Abstract

Several species in the oomycete genus *Peronosclerospora* cause downy mildew on maize and can result in significant yield losses in Asia. Bio-surveillance of these pathogens is a high priority to prevent epidemics on maize in the United States and consequent damage to the US economy. The unresolved taxonomy and dearth of molecular resources for *Peronosclerospora* spp. hinder these efforts. *P. sorghi* is a pathogen of sorghum and maize with a global distribution, for which limited diversity has been detected in the southern USA. We characterized the genome, transcriptome, and mitogenome of an isolate, representing the US pathotype 6. The highly homozygous genome was assembled using 10× Genomics linked reads and scaffolded using Hi-C into 13 chromosomes. The total assembled length was 303.2 Mb, larger than any other oomycete previously assembled. The mitogenome was 38 kb, similar in size to other oomycetes, although it had a unique gene order. Nearly 20,000 genes were annotated in the nuclear genome, more than described for other downy mildew causing oomycetes. The 13 chromosomes of *P. sorghi* were highly syntenic with the 17 chromosomes of *Peronospora effusa* with conserved centromeric regions and distinct chromosomal fusions. The increased assembly size and gene count of *P. sorghi* is due to extensive retrotransposition, resulting in putative pseudogenization. Ancestral genes had higher transcript abundance and were enriched for differential expression. This study provides foundational resources for analysis of *Peronosclerospora* and comparisons to other oomycete genera. Further genomic studies of global *Peronosclerospora* spp. will determine the suitability of the mitogenome, ancestral genes, and putative pseudogenes for marker development and taxonomic relationships.

Keywords: chromosome, centromere, whole-genome sequencing, mitogenome, RNAseq, synteny

Introduction

Oomycetes are destructive pathogens of plants and animals that have caused several historical and contemporary epidemics (Derevnina *et al.* 2016). The genus *Peronosclerospora* contains multiple, poorly resolved species of oomycetes, which cause downy mildews on graminaceous plants including grain crops and sugarcane (Perumal *et al.* 2008; Suharjo *et al.* 2020; Crouch *et al.* 2022). Maize is a host to many of these species and downy mildew outbreaks can result in significant yield losses (Anaso *et al.* 1989; Bock *et al.* 1998; Thakur and Mathur 2002; Crouch *et al.* 2022). Maize production plays a major role in the economy of the United States and was valued at over $61 billion dollars by the USDA National Agricultural Statistics Service for 2020. Consequently, measures for surveillance and strategies for response to species that cause major downy mildew epidemics on maize elsewhere, but are not established in the United States, are desirable (Futrell and Frederiksen 1970). This threat to US agriculture resulted in *P. philippinensis* being placed on the select agent list (Perumal *et al.* 2008); however, the taxonomic relationships of *Peronosclerospora* spp., including *P. philippinensis*, is unresolved (Bonde *et al.* 1984; Micales *et al.* 1988; Perumal *et al.* 2008; Crouch *et al.* 2022). Therefore, it may be prudent to evaluate whether the introduction of any species that causes downy mildew on maize could threaten US production.

Currently, the only *Peronosclerospora* species reported in the United States is *P. sorghi*, which causes downy mildew on maize and sorghum (Crouch *et al.* 2022). This pathogen was first reported in Texas, United States in 1961 and later reported from ten states. At the time, *P. sorghi* was regarded as invasive and estimated to cause a $2.5 million loss in sorghum, broomcorn, and maize production in Texas alone in 1969 (Futrell and Frederiksen 1970). The variability of *P. sorghi* in the United States is believed to be low (Prom *et al.* 2015); however, this has not been determined at the genomic level. *Peronosclerospora sorghi* can complete its asexual cycle on both sorghum and maize. The species is homothallic;

Received: June 27, 2022. **Accepted:** December 5, 2022

however, oospores generated during the sexual cycle have only been reported in sorghum in the United States (Pawar 1986; Perumal *et al.* 2008). Previously, *P. sorghi* has been managed through crop rotation, resistant hybrids, and seed treatment with metalaxyl (methyl N-(2,6-dimethylphenyl)-N-(methoxyacetyl)-DL-alaninate). Outbreaks of metalaxyl resistant downy mildew on sorghum occurred in Texas in 2001 and 2002 (Isakeit *et al.* 2003; Perumal *et al.* 2008), and in subsequent years, associated with sorghum monoculture. After growers in Texas switched to monoculturing with resistant hybrids in response to metalaxyl resistance, a new pathotype, 6, was detected (Isakeit and Jaster 2005), which was also resistant to metalaxyl. Consequently, constant monitoring of this pathogen is required. At present, six pathotypes have been described from the Americas, based on virulence phenotypes on 10 differential inbred lines of sorghum (Prom *et al.* 2015). Pathotypes 1 to 3 and 6 were identified from Texas, pathotype 4 from Brazil, and pathotype 5 from Honduras (Prom *et al.* 2015). The line SC155 was initially thought to differentiate pathotype 6 from 3 (Isakeit and Jaster 2005), but this was not subsequently confirmed (Prom *et al.* 2015). Pathotype 6 isolates overcome resistance of proprietary Pioneer brand germplasm resistant to pathotype 3 isolates (Prom *et al.* 2015). American cultivars of maize are highly resistant to *P. sorghi* but are reported to be susceptible to other *Peronosclerospora* spp. found in Asia (Duck *et al.* 1987; Perumal *et al.* 2008). Generating genomic resources for *P. sorghi* will enable the development of molecular tools to monitor and distinguish populations of the pathogen globally.

Genomic resources are foundational for accurately defining the variation present in populations of a pathogen, conducting biosurveillance, and identifying vulnerabilities and opportunities that can be utilized for disease management. Recently, two 17-chromosome-scale genome assemblies have been reported for the oomycetes *Bremia lactucae* and *Peronospora effusa,* which cause lettuce and spinach downy mildew, respectively (Fletcher *et al.* 2021; Fletcher *et al.* 2022). Despite being taxonomically distant, these oomycetes share a high degree of synteny. Because downy mildew oomycetes are polyphyletic (McCarthy and Fitzpatrick 2017; Bourret *et al.* 2018; Fletcher *et al.* 2018; Fletcher *et al.* 2019; Fletcher *et al.* 2022), this high level of synteny suggests that similar or derived chromosome architectures are expected for all downy mildew causing oomycetes, and several clades of *Phytophthora* spp. (Fletcher *et al.* 2022).

The present study describes the chromosome-scale genome assembly of *P. sorghi*. This is the first genome assembly for any species in the genus *Peronosclerospora* and will be a foundational resource for surveillance of this pathogen globally. *P. sorghi* has the largest genome of any oomycete sequenced to date. Comparative genomics revealed a unique genomic architecture with 13 chromosomes derived from the ancestral architecture. In addition, the transcriptome of *P. sorghi* was characterized to generate reliable gene models and investigate differential gene expression (DGE) of infected leaves pre- and post-sporulation of the pathogen.

Methods

Isolate propagation and collection

The P6 isolate of *Peronosclerospora sorghi* was previously collected from Wharton County, Texas (Isakeit and Jaster 2005) and maintained on inoculated plants in a greenhouse at Texas A&M as described previously (Radwan *et al.* 2011). Sporangia were collected into distilled water from multiple plants and shipped frozen to UC Davis for DNA extraction and to Dovetail Genomics to generate Hi-C libraries. For RNAseq, pairs of segments from leaves of an infected plant were split longitudinally prior to sporulation. One segment of each pair was placed immediately in RNAlater (Thermo Fisher Scientific, Waltham, MA) to characterize the transcriptome pre-sporulation. The second segment was placed in a 20°C incubator overnight to trigger sporulation and then placed in RNAlater. This material was sent to UC Davis for RNA extraction.

DNA extraction, sequencing, and genome assembly

Genomic DNA was extracted from a pellet of sporangia to generate 10× Genomics linked-read libraries. Briefly, the pellet was vortexed for 2 min in a microcentrifuge tube with ~200 µl of Rainex-treated beads and 500 µl of 2× extraction buffer [100 mM Tris-HCl pH 8.0, 1.4 M NaCl, 20 mM EDTA, 2% (wt/vol) cetyltrimethylammonium bromide, and B-mercaptoethanol at 20 µl/ml]. After vortexing, the material was transferred to a fresh 2 ml tube and subsequently treated with RNase (20 µl/ml) at 65°C for 30 min. Next, an equal volume of 1:1 phenol/chloroform was added, mixed, and centrifuged at maximum speed (5,200×g) for 15 min. The aqueous phase was retained, mixed with 24:1 chloroform/isoamyl alcohol, and again centrifuged as maximum speed for 15 min. The aqueous phase was mixed with 0.7 volumes of isopropanol and DNA precipitated at −20°C for 1 h. DNA was pelleted by centrifuging at maximum speed for 30 min, washed with 70% ethanol, dried, and suspended in 10 mM Tris-HCl. Genomic libraries were constructed using the 10× Genomics Chromium (Weisenfeld *et al.* 2017) and sequenced on an Illumina×10 at Novogene (Sacramento, CA). All raw data, plus the subsequent assemblies and annotation, are available under NCBI BioProject PRJNA845776, accession JAPDHQ000000000.

Genomic sequence data generated to study the genome of *P. sorghi* included 449,879,160 linked-read pairs (10× Genomics) for isolate P6, equivalent to 294× genome coverage. Genome assemblies were constructed using the 10× Genomics SuperNova v2.0 pipeline (Weisenfeld *et al.* 2017) and positively filtered for oomycete scaffolds using BLASTn (Altschul *et al.* 1990; Fletcher and Michelmore 2018). Optimal assembly conditions using different barcode fractions were determined empirically (Supplementary Table 1). Hi-C libraries were prepared by Dovetail Genomics. The Hi-C libraries were sequenced on a HiSeq 4,000 at the UC Davis DNA Technologies Core to yield 321,243,925 Hi-C paired-end reads that were used for scaffolding the Supernova assembly with the Dovetail Genomics Inc. HiRise pipeline (Putnam *et al.* 2016). Hi-C interactions were visualized by re-aligning Hi-C reads using bwa mem and plotting contact matrices with HiCExplorer (Ramírez *et al.* 2018). Centromeres were identified in regions of the assembly enriched for short-distance cis-interactions and regional trans-interactions, resulting in crosses on the Hi-C plot. This is because centromeres of distinct chromosomes may be physically close to one another in Rabl-like conformations (Varoquaux *et al.* 2015). In the initial assembly, some chromosome-scale scaffolds were inconsistent with the expectation of one centromere per scaffold (Supplementary Fig. 1). These potential misjoins were investigated with Hi-C reads and broken at gaps (strings of unknown bases; N). Coordinates of Hi-C reads were recalculated and enrichment of read pairs aligned to chromosome arms was sought. Sequences were rejoined so the centromeres formed expected crosses when contact matrices were plotted. Centromere coordinates in the final assembly were defined by identifying 100 kb windows of the genome with short mean cis-interaction distance of Hi-C, first-mate reads. These windows were visualized as scatter plots using ggplot2 (Wickham 2016). Telomeric repeats

(CCCTAAA) were identified by searching for three copies of the repeat unit in the first and final 50 kb of each sequence and validated manually. Assembly completeness was evaluated using KAT (Mapleson *et al.* 2017) and BUSCO (Simao *et al.* 2015). Results from the BUSCO analysis were added to a previous phylogenomic analysis (Fletcher *et al.* 2022). Briefly, peptide sequences for each BUSCO prediction found to be single copy in all 32 species surveyed were aligned independently with MAFFT v7.245 (Katoh and Standley 2013). The aligned sequences were concatenated, and a Maximum Likelihood phylogeny was produced with RAxML v8.2.9 (Stamatakis 2014) using the PROTGAMMAAUTO model and 1,000 bootstraps. The phylogenetic tree was visualized using Geneious v8.0.5 (Kearse *et al.* 2012) and labels were italicized in Microsoft PowerPoint. The Newick tree produced by RAxML is available as Supplementary File 1.

Mitochondrial assembly

Contigs from a de novo genomic assembly in CLC Genomics Workbench (v9; Qiagen, Redwood City, CA) were identified as mitochondrial due to sequence similarity with *P. tabacina* mitochondrial sequences (KT893455) by BLAST analysis (Derevnina *et al.* 2015). These were used as templates for further assembly with SeqMan NGen (v16.0.0, DNASTAR, Madison, WI, USA). The resulting assemblies were evaluated for uniformity and depth of coverage. Contigs were broken when gaps/low coverage or inconsistencies were observed and the set of smaller contigs reassembled using the reference-guided assembly—special workflows assembly option of SeqMan NGen to extend the ends of the contigs and close the gaps. Open reading frames (ORFs) were predicted and annotated with Geneious v9.1.8 (Biomatters, New Zealand) using the universal genetic code. Encoded products of genes were identified using BLAST (Altschul *et al.* 1990) analysis against mitochondrial genome sequences published for *Peronospora tabacina* (Derevnina *et al.* 2015). Genes encoding tRNAs were identified using tRNAscan-SE v1.3.1 (Lowe and Eddy 1997).

RNA extraction, sequencing, and transcript assembly

Total RNA was extracted from pairs of infected leaf segments pre- and post-sporulation of *P. sorghi* collected at Texas A&M, using a Qiagen RNeasy kit (Cat. No./ID: 74904), followed by poly-A cDNA generation. Strand-specific libraries were generated using the KAPA mRNA HyperPrep Kit (KR1352 v5.17) per supplier instructions. Fragments were 150 bp paired-end sequenced on a HiSeq 4,000 at the UC Davis DNA Technologies Core. Between 37,570,743 and 123,050,654 strand-specific cDNA read-pairs were generated across three biological replicates of the two conditions for a total of 290,220,630 cDNA read pairs. The resulting reads were trimmed using bbduk.sh (Bushnell 2016), assembled with Trinity v2.4.0 in strand-specific mode (Grabherr *et al.* 2011), and mapped to a combined reference containing the *S. bicolor* (Paterson *et al.* 2009) and *P. sorghi* assemblies using minimap2 (Li 2018). The pathogen's transcriptome was defined as those that primarily aligned to *P. sorghi* scaffolds, while the host's transcriptome was defined as those that primarily aligned to *S. bicolor* scaffolds. Transcriptome completeness for the host and pathogen was evaluated by BUSCO (Simao *et al.* 2015) in transcriptome mode. The sense (top) strand of the *P. sorghi* transcriptome was translated into a set of longest ORFs using TransDecoder.LongOrfs v5.5.0 (Haas *et al.* 2013), and conserved Pfam domains were identified using InterProScan (Finn *et al.* 2014; Jones *et al.* 2014). The assembled transcriptome is available under NCBI BioProject PRJNA845776, accession GKCP00000000. Unmapped transcripts were evaluated with Kraken2 (Wood *et al.* 2019) using the NCBI non-redundant protein (nr) database. Results were visualized using KronaTools (Ondov *et al.* 2011). For the pathogen, the coordinates of transcripts were compared with the coordinates of genes annotated (see Annotation section below) using BEDtools v2.29.2 intersect (Quinlan 2014). Transcripts that did not overlap genes were investigated for ORFs using TransDecoder.LongOrfs v5.5.0 (Haas *et al.* 2013) and surveyed for the presence of conserved Pfam domains using InterProScan (Finn *et al.* 2014; Jones *et al.* 2014).

Annotation

Repeats in the genome assembly of *P. sorghi* were defined with RepeatModeler v1.73 (Smit and Hubley 2008) and masked with RepeatMasker v4.0.9 (Smit *et al.* 2013). The same library was used to identify repeats in the transcriptome assembly. Gene models were annotated in the genome assembly using MAKER (Cantarel *et al.* 2008), with additional putative effectors identified using hidden Markov models (HMM) with HMMER (Eddy 2011) and regular expression string searches of ORFs (Fletcher and Michelmore 2018). The MAKER pipeline was provided with the RepeatModeler profile as well as assembled transcripts and translated ORFs from the transcriptome of *P. sorghi*, all described above, plus ESTs (option: altest) and protein sequences of other oomycete species available from NCBI. MAKER was initially run without a SNAP HMM, inferring genes using est2genome and protein2genome. These predictions were used to train a SNAP HMM (Korf 2004) that was used for a subsequent run of MAKER with both est2genome and protein2genome set to 0. The predicted proteins were again used to train a new SNAP HMM (Campbell *et al.* 2014). This process was repeated twice to generate three SNAP HMMs, which were used sequentially in three independent runs of MAKER. The annotations produced were evaluated as previously described (Fletcher *et al.* 2018) to select a single optimal run. This involved contrasting the number of gene models predicted, mean protein length, BLASTp hits to other oomycete annotations, and Pfam domains annotated by InterProScan (Finn *et al.* 2014; Jones *et al.* 2014). Annotation of genes encoding putative effectors was performed as previously described (Fletcher *et al.* 2018). Briefly, the entire genome was translated into ORFs. These ORFs were surveyed for secretion signals using SignalP3.1 and SignalP4.0, and crinkler (CRN) motifs of LWY domains using HMMs. For peptides with secretion signals, the 60 residues beyond the predicted cleavage site were surveyed for an RXLR or RXLR-like motif and subsequently for a downstream EER or EER-like motif. ORFs encoding peptides that were predicted to be secreted and contained an (L)WY domain or a CRN motif were considered high-confidence putative effectors (HCPEs). ORFs encoding peptides that were predicted to be secreted and encoded an RXLR and EER domain, but did contain an (L)WY domain, or encoding peptides not predicted to be secreted, but contained an (L)WY domain, or a CRN motif were considered low-confidence putative effectors (LCPEs). The putative effectors and MAKER annotations were reconciled so that annotations did not overlap on the same strand. This was performed so that (1) any HQE or LQE annotations that did not overlap a MAKER annotation were added to the master annotation; for *P. sorghi* this was 12 HCPEs and 122 LCPEs. (2) HCPEs that overlapped MAKER annotations with the same start coordinates but earlier stop coordinates were discarded; for *P. sorghi* this was six peptides. (3) HCPEs that overlapped MAKER annotations with the same start coordinates but later stop coordinates replaced the model proposed by MAKER if they had a higher BLASTp score to the NCBI nr database than

the overlapping MAKER model; for *P. sorghi* this was six peptides. (4) HCPEs that overlapped MAKER annotations but had different start coordinates and later or identical stop coordinates were retained over proposed MAKER models; for *P. sorghi* this was 27 peptides. (5) HCPEs that overlapped MAKER annotations but had different start coordinates and earlier stop coordinates were investigated to determine if the MAKER model should have a modified start coordinate; for *P. sorghi* this was six peptides. (6) Any LCPEs that overlapped MAKER annotations were discarded; for *P. sorghi* this was 142 ORFs. The same effector prediction workflow was then applied to the reconciled annotation set to determine the reported effector counts. Tracks for repeats, transcript coverage, annotation, and effector annotations were generated in 100 kb windows along each chromosome using Bedtools v2.29.2 and plotted using Circos (Krzywinski *et al.* 2009).

Comparative genomics

Predicted peptide sequences of *P. sorghi* were used in an orthology analysis including another 39 genome assemblies of 34 different species (Supplementary Table 2). The number of genes for each assembly surveyed was calculated by counting the number of FASTA entries in the peptide file. Orthology was calculated using OrthoFinder v2.2.1 (Emms and Kelly 2015). Resulting orthogroups were filtered for interspecies orthogroups, which required assignment of proteins from at least two different oomycete species. The number of interspecific orthogroups assigned to each assembly and the number of proteins represented in these orthogroups was calculated. Interspecies orthogroups were then summarized at different taxonomic levels for visualization. For the purposes of this study, core orthogroups had proteins from *P. sorghi* and at least one other downy mildew clade 2 species; one downy mildew clade 1 species; one *Phytophthora* spp.; one species in the Family Pythiaceae; one species in the Family Albuginaceae; one species in the Order Saprolegniales. For all orthogroups, a ratio of *P. sorghi* to *P. effusa* gene assignment was calculated, ordered, and plotted as a heatmap using ggplot2 (Wickham 2016). Single-copy orthologs between *P. sorghi* and *P. effusa* were identified as members of orthogroups assigned exactly one protein sequence from *P. sorghi*, one from *P. effusa*. Coordinates of the genes encoding these protein sequences were extracted from GFF files and plotted as links using Circos (Krzywinski *et al.* 2009), scatter plots using ggplot2 to establish synteny, or vertical lines using ggplot2 to establish centromere conservation. Genes expanded in *P. sorghi* relative to *P. effusa* (notated throughout as >1:1) were identified as members of interspecific orthogroups assigned more than one *P. sorghi* sequence and exactly one *P. effusa* sequence. Genes expanded in *P. effusa* relative to *P. sorghi* (notated throughout as 1:>1) were identified as members of interspecific orthogroups assigned exactly one *P. sorghi* sequence and more than one *P. effusa* sequence. Genes expanded in both (>1:>1) were identified as members of interspecific orthogroups containing multiple *P. sorghi* and multiple *P. effusa* sequences. Genes absent in *P. effusa* but present as single-copy (1:0) or multicopy in *P. sorghi* (>1:0) were identified as members of interspecific orthogroups lacking *P. effusa* annotations with either one or multiple *P. sorghi* annotations, respectively. Genes not assigned to interspecific orthogroups were classified as lacking orthology for the purpose of this study.

Genes expression was investigated using the six RNAseq data sets generated here (three replicates of two conditions; see above). A single reference containing the assembled sequence and annotation of *P. sorghi* and *S. bicolor* (GCF_000003195.3; Paterson *et al.* 2009) using STAR v2.7.9 in runMode genomeGenerate. Paired-end RNAseq data was aligned to this reference using STAR with in quantMode GeneCounts setting sjdbOverhang to 99. Transcript abundance was determined by calculating the FPKM across all generated RNAseq reads. The total exon length of each gene was calculated by summing the length of annotated exons in the GFF file. The raw counts generated by STAR were summed for each gene and scaled to counts per million (CPM). This value was further scaled by the total exon length for each gene to obtain the FPKM. Analysis of variance (ANOVA) was conducted in R using aov(FPKM ~ Gene Classification) and pairwise comparisons generated using Tukey's honestly significant difference (HSD) (Abdi and Williams 2010) in R using TukeyHSD(). Groups of significance were identified using multcompView (Graves *et al.* 2015) in R. Dot plots showing the synteny and transcript abundance of *P. sorghi* genes were plotted in R ggplot() and geom_point() (Wickham 2016). Scatter plots showing the transcript abundance for secreted and non-secreted genes in each category were plotted using ggplot() and geom_jitter() (Wickham 2016) setting height to 0. DGE analysis was conducted using edgeR (Robinson *et al.* 2010). Read counts of pathogen and host genes were independently analyzed in R. Genes with fewer than five CPM were dropped from the analysis and normalization was conducted on the reduced data set using calcNormFactors(). Multi-dimensional scaling coordinates were obtained using plotMDS() and plotted using ggplot() (Wickham 2016). A design matrix was constructed using limma (Ritchie *et al.* 2015) model.-matrix(), dispersion estimated using estimateDisp() from edgeR (Robinson *et al.* 2010), and a negative binomial generalized linear model was fit using glmFit(). The two conditions were then compared using makeContrasts() and a negative binomial generalized log-linear model was fit using glmLRT(). Differentially regulated genes were identified as those that have an adjusted *P*-value <0.05. For each gene, the log fold change and -log10 adjusted *P*-value was plotted in R using ggplot() and geom_point(). For the host, the function of differentially regulated genes was investigated by identifying the annotated product produced in the GFF file. For the pathogen, the function of differentially regulated genes was investigated by identifying conserved Pfam domains annotated by InterProScan (Finn *et al.* 2014; Jones *et al.* 2014). Chi-squared tests were run in R using chisq.test().

The dN/dS for genes annotated in *P. sorghi* that had single-copy orthologs in *P. effusa* was performed if the orthologs had a blastp alignment coverage greater than 20%. Pairwise codon-based alignments were conducted for each pair of orthologs using PRANK. The dN/dS for each alignment was calculated using codeml in pairwise mode (Yang 2007). This analysis was run as pairwise instead of per orthogroup because the *P. sorghi* genes assigned to the same orthogroup may cover different parts of the *P. effusa* ortholog, which resulted in codeml using a smaller portion of sequence for the calculation. An ANOVA for dN/dS of different gene classifications was calculated as above using Tukey's HSD. Data were plotted as a raincloud plot in R using ggplot() (Wickham 2016; Allen *et al.* 2021).

The length of each peptide sequence annotated in the genome of *P. sorghi* was obtained by building an index file with SAMtools faidx (Li *et al.* 2009). The mean exon count of each gene was obtained by counting the number of exon entries per annotation in the GFF file. An ANOVA was run to compare the mean encoded peptide length and mean exon count for each of the gene categories as described above. The length of every intron encoded in the genome of *P. sorghi* was determined from the GFF file and plotted as a histogram using ggplot() (Wickham 2016). Histograms were colored depending on whether transcripts

could be detected (FPKM >0) for the gene from which the intron originates.

Compartmentalization of the *P. sorghi* genome assembly was determined by calculating the 5' and 3' intergenic distances between genes on chromosomal scaffolds (Dong *et al.* 2015; Frantzeskakis *et al.* 2019). This information was extracted and oriented from the GFF file. Density of intergenic distances for genes with FPKM equal to 0 and FPKM >0 were plotted on a log10-scaled heatmap using ggplot() and geom_bin2d() (Wickham 2016) with 40 bins on each axis. A histogram of intergenic distance for each flank was plotted for gene categories defined by orthology with *P. effusa* using ggplot() and geom_histogram() (Wickham 2016). Histograms were colored by the presence or absence of transcript detection.

Results

Our initial genome assembly of *P. sorghi* contained 342 Mb of nucleotide sequence that was scaffolded using 10× Genomics linked and Hi-C reads into 12 large superscaffolds. This genome assembly was manually corrected by inspecting the Hi-C contact matrix (Supplementary Fig. 1). Centromeres were identified on each superscaffold by short-distance cis contacts, which resulted in crosses on the Hi-C plot (Fig. 1a). The initial Hi-C plot indicated that there were three mis-assembled superscaffolds (Supplementary Fig. 1). The arms of Chromosome (Chr.) 11 were joined at the telomeric regions not the centromere; the chromosomes arms were therefore reoriented and scaffolded over the centromere. The two arms of Chr. 8 were split in the initial assembly and joined to Chr. 1 and Chr. 2; the false joins were evident in the Hi-C plot because one end of each scaffold was enriched for short-distance cis contacts in addition to those at the centromere. The two arms of Chr. 8 were separated and then scaffolded together. Hi-C reads were remapped to validate the revised assembly (Fig. 1a). These manual corrections were subsequently confirmed by synteny analysis (see below). Small unplaced scaffolds were removed because they were flagged as identical by NCBI during submission. Removal of these scaffolds was supported by *k*-mer analysis (Fig. 1b). Full assembly statistics are available in Supplementary Table 1.

Our final assembly contained thirteen chromosomal pseudomolecules with a total scaffolded length of 283.5 Mb; there were 6,731 unplaced small scaffolds (Chr. 0) with an average length of 6.1 kb and a total length of 41.7 Mb (Table 1). Analysis of *k*-mers determined that the genome of *P. sorghi* was highly homozygous because only a single peak was present in the *k*-mer histogram that approximated a Poisson distribution (Fig. 1b). The high-quality of the genome assembly was evident because most of the single-copy *k*-mers calculated from the read set were present only once in the genome assembly. Short telomeric-like sequences were assembled on both ends of Chr. 1, one end of Chr. 6, one end of Chr. 11, and four unplaced scaffolds. Benchmarking with BUSCO demonstrated that the assembly was 97.9% complete with little duplication (Fig. 1c), supporting a high level of completeness. All but one BUSCO was predicted from the assembled chromosomes of *P. sorghi*. The single BUSCO gene located on Chr. 0 was scored as duplicate and occurred three twice on two separate scaffolds.

A subset of single-copy orthologs identified in *P. sorghi* by BUSCO were added to a previously reported phylogenetic analysis using 18 BUSCO genes found to be single-copy across 31 genome assemblies (Fletcher *et al.* 2022). This determined that the two graminicolous downy mildew assemblies, those of *P. sorghi* and *Sclerospora graminicola*, formed a monophyletic clade within downy mildew clade 2 (Fig. 2). This phylogeny supports that these two graminicolous downy mildews share a common biotrophic ancestor with other genera of downy mildews in this clade, including *Peronospora*, *Pseudoperonospora*, and *Hyaloperonospora* (McCarthy and Fitzpatrick 2017; Bourret *et al.* 2018; Fletcher *et al.* 2018; Fletcher *et al.* 2019; Fletcher *et al.* 2022).

Over 83% (254 Mb) of the assembled sequence was repetitive. The most common repeat elements were long terminal-repeat retrotransposons, which comprised 221 Mb of the assembly. The density of repeats was higher in Chr. 0 compared with the chromosome-scale scaffolds (Table 1). Repeat density in 100 kb windows along the chromosomal pseudomolecules ranged from 0.29 to 1.00, and the average repeat density of windows was 0.82 (Fig. 3).

The mitochondrial genome had a read depth of 2,359× and was assembled into a 38,497 bp circular contig with a GC content of 22.4% (Supplementary Fig. 2; GenBank accessions OP873122). Coding regions constituted 90.3% of the genome with 6.0% of this total representing hypothetical genes. A total of 35 known genes (encoding 18 respiratory chain proteins, 16 ribosomal proteins, and an import protein *ymf16* of the *sec*Y-independent pathway), the *rnl* and *rns*, and 25 tRNA genes encoding for 19 amino acids were present. In addition, there were five hypothetical proteins (*ymf98*, *ymf99*, *ymf100*, *ymf101*, and *orf32*) in common with other oomycete mitochondrial genomes (Fletcher *et al.* 2018; Cai and Scofield 2020) and one putative ORF (*orf161*) that was present in *P. sorghi* downstream from the *cox1* gene (Fig. 4). BLAST queries to GenBank identified no significant sequence similarities for *orf161*. The mitochondrial gene order in *P. sorghi* was similar to *Peronospora* species with some exceptions (Fig. 4). There was a large inversion in the mitogenome of *P. sorghi*, relative to all other oomycete taxa. This inversion split the mitogenome into two parts representing ~71 and 29% of the sequence. The larger portion encoded 31 genes plus four putative ORFs (*ymf100* to *cob*) and spanned a region with a predominantly conserved gene order across several downy mildew taxa and *Phytophthora* species. Another difference in gene order for *P. sorghi* was the placement of the *nad5-nad6-trnR* genes, which were in the same location but inverted and encoded on the opposite strand rather than downstream from the *atp1* gene (Fig. 4). The inversion of genes *cob* to *atp1* (highlighted blue in Fig. 4) common to *B. lactucae*, *Plasmopara viticola*, and *P. infestans*, relative to the other taxa sampled, is consistent with polyphyly of oomycetes that cause downy mildew diseases (Bourret *et al.* 2018; Fletcher *et al.* 2018).

The transcriptome of *P. sorghi* isolate P6 was characterized using RNA extracted from lesions on leaves of *Sorghum bicolor* harvested both prior to and post-sporulation. The transcriptome assembly contained 365,281 sequences totaling 482 Mb. Aligning transcripts to the genomes of *P. sorghi* and *S. bicolor* assigned 73,563 (20%) of the transcripts to *P. sorghi* (Fig. 5). The total length of transcripts assigned to *P. sorghi* was 98 Mb; the largest transcript was 16.8 kb. The N_{50} of the *P. sorghi* transcriptome was 2.2 kb, the mean length was 1.3 kb. Over 95% of BUSCO genes were detected in the transcriptome with 49.1% duplicated. Over 10% of the *P. sorghi* transcriptome was derived from repetitive genomic sequences, 184 transcripts of which encoded reverse transcriptase (Pfam domain; PF07727) (Finn *et al.* 2014), ranking it the 10th most frequently identified domain encoded in the transcriptome (Supplementary Table 3). Most of the assembled transcriptome (62%) aligned to the host assembly of *S. bicolor* (225,170 transcripts totaling 356.8 Mb). The largest *S. bicolor* transcript was 18.4 kb, the transcript N_{50} was 2.6 kb, and the mean transcript

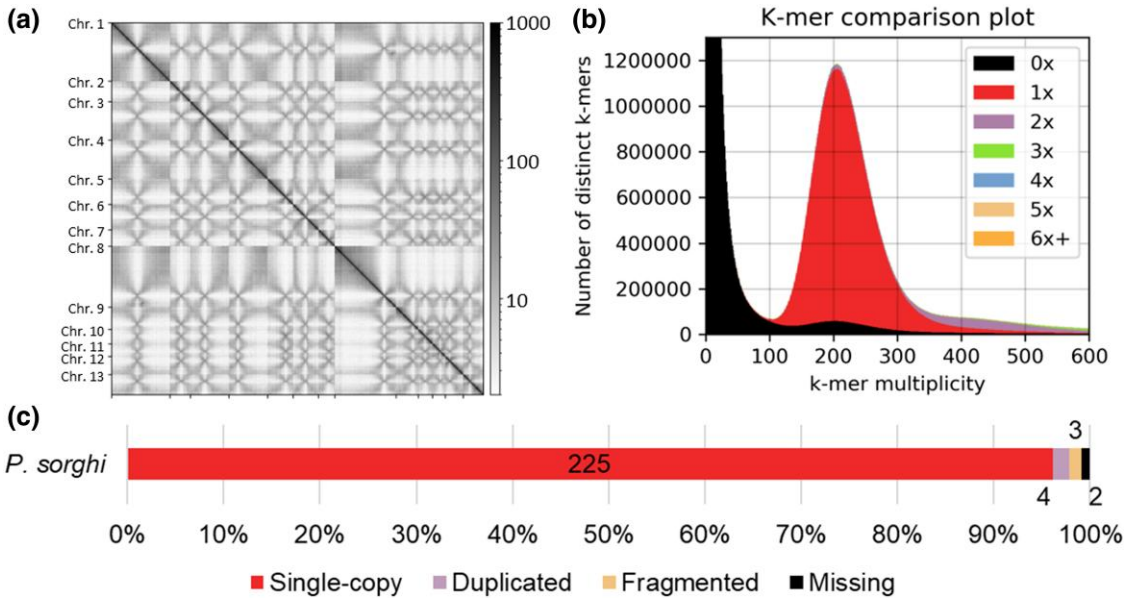

Fig. 1. The genome assembly of *Peronosclerospora sorghi*. a) Hi-C contact matrix of 13 chromosome-scale scaffolds. The strong diagonal reflects the high contact frequency between physically close sequences indicating their correct linear order along each chromosome-scale scaffold. Cross patterns along the x- and y-planes are indicative of high frequencies of trans contacts between centromeres and are likely due to the Rabl-like chromosome configurations. b) Visualization of assembly completeness using *k*-mer inclusion demonstrating that the majority of single-copy *k*-mers are present in the assembly only once (red). The Poisson distribution is consistent with high homozygosity in the genome of *P. sorghi*. c) Stacked bar-chart showing the completeness of the genome assembly as calculated by BUSCO. The color scheme is similar to B so that single-copy BUSCOs are depicted as red. The annotated numbers indicate gene count. The total number of BUSCOs in the protist database was 234.

Table 1. Assembly statistics for the genome of *Peronosclerospora sorghi*.

P. sorghi sequence	Length	Primogenitors[a]	Total gap bases	Percent gaps	AT content	Genes	Genes/Mb	Repeat-masked bases	Percent repeats
Chr. 1	44,403,666	Chr. 1	3,569,330	8.04%	51.0%	2746	61.84	34,200,987	83.76%
Chr. 2	15,566,821	Chr. 2	1,090,110	7.00%	51.6%	996	63.98	11,673,129	80.63%
Chr. 3	29,316,722	Chr. 3, Chr. 5	2,441,500	8.33%	51.2%	1854	63.24	22,246,130	82.78%
Chr. 4	29,556,142	Chr. 4, Chr. 9	2,216,080	7.50%	51.4%	1952	66.04	22,262,399	81.43%
Chr. 5	19,437,638	Chr. 6	1,668,350	8.58%	51.6%	1250	64.30	14,321,970	80.60%
Chr. 6	19,380,668	Chr. 7, Chr. 13	1,459,800	7.53%	51.3%	1317	67.95	14,626,688	81.62%
Chr. 7	12,369,872	Chr. 8	1,079,370	8.73%	51.9%	902	72.92	9,186,674	81.37%
Chr. 8	46,690,312	Chr. 10, Chr. 12	3,572,980	7.65%	51.0%	2667	57.12	35,816,060	83.07%
Chr. 9	16,992,261	Chr. 11	1,138,780	6.70%	51.5%	1043	61.38	13,069,339	82.44%
Chr. 10	10,915,484	Chr. 14	752,700	6.90%	51.8%	747	68.43	8,382,571	82.48%
Chr. 11	9,655,334	Chr. 15	706,550	7.32%	51.8%	565	58.52	7,357,013	82.21%
Chr. 12	13,731,139	Chr. 16	909,030	6.62%	51.5%	1050	76.47	10,492,191	81.83%
Chr. 13	15,489,776	Chr. 17	1,128,150	7.28%	51.7%	1104	71.27	11,615,788	80.88%
Chr. 0	41,745,159	n/a	357,080	0.86%	49.9%	925	22.16	38,862,127	93.90%

[a] Primogenitors retain numbering from the chromosomes of *P. effusa* GCA_021491665.1.

length was 1.6 kb. Over 93% of BUSCO genes were detected in the transcriptome of the host with 58.4% duplicated. An additional 66,548 transcripts totaling 27.3 Mb (18%) aligned to neither pathogen nor host. Of these, 40% were assigned to true fungi, 18% to metazoa, 13% to bacteria, 8% to viridiplantae, 0.3% to Oomycota, 0.2% to archaea, and 0.1% to viruses; Kraken2 was unable to assign a taxonomic identity to 15% of the unaligned transcripts (Supplementary File 2).

A total of 19,118 genes were annotated in the assembly of *P. sorghi*, which covered 18.2% of the genome assembly. The mean length of gene models was 3.0 kb. The density of genes was higher on chromosome-scale scaffolds than unplaced scaffolds, with 18,193 assigned to chromosomes (Table 1). Signal peptides, indicative of extra-cellular secretion, were identified in 1,477 gene models. These included 351 genes encoding secreted RXLR-like effectors (defined in Methods section) and eight secreted CRN effectors. Evidence for transcription was found for 280 of these effector-encoding genes (Supplementary Table 4) and 232 were sufficiently expressed to be included in the subsequent analysis of DGE (see below). In addition, 68 gene models encode peptides containing an LWY domain but lacking a secretion signal and 24 gene models encoded peptides containing a CRN motif but no secretion signal (Table 2). Evidence for transcription was found for 79 of these genes and 63 were included in the subsequent DGE analysis. Clustering peptide sequences of putative effectors identified 19 clusters of genes each encoding three or more high-identity RXLR effectors on 10 of the 13 chromosomes; these clusters spanned 54 kb (Chr. 4) to 28 Mb (Chr. 8). No clusters of genes

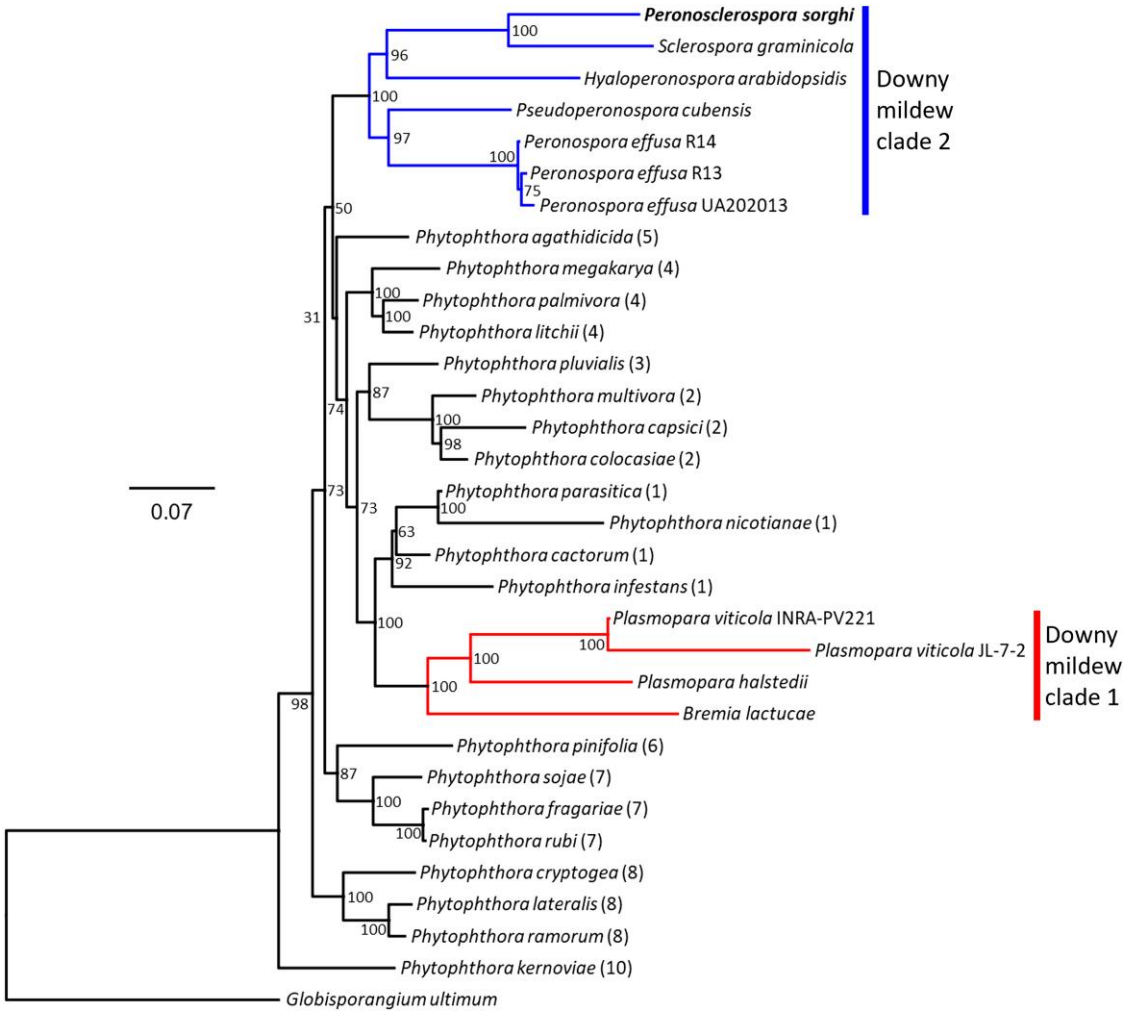

Fig. 2. Phylogenetic analysis of *Peronosclerospora sorghi*. Maximum likelihood phylogenetic tree calculated from the concatenated alignment of 18 BUSCO proteins surveyed in 32 assemblies of 28 oomycete species. All species analyzed belong to the Peronosporaceae, except *Globisporangium ultimum* (Pythiaceae), which was used as an outgroup. Colored branches indicate the two downy mildew clades, consistent with polyphyly of the downy mildews. *P. sorghi* was assigned to downy mildew clade 2 (blue). The scale indicates the mean number of amino acid substitutions per site. Branch support was calculated from 1,000 bootstraps. Most of the data for this figure is the same as used in Fletcher *et al.* 2022, Fig. 5.

encoding CRN effectors were localized to a single chromosome. The physical location of clusters of effector genes were not conserved between *P. sorghi* and *P. effusa* (Supplementary Fig. 3). There were 195 genes encoding RXLR effectors, which clustered with no other RXLR and so were classified as singletons. There were 24 singleton genes encoding CRN effectors. The majority of high-identity pairs of RXLR effectors were not found on the same chromosome (Fig. 6).

Comparing the mapping location of transcripts to the coordinates of genes indicated that the high number of transcripts assembled compared with the number of genes annotated for *P. sorghi* is mostly due to the assembly of putative isoforms. Of the 73,563 transcripts, 40,997 were found to overlap 13,554 annotated genes on the same strand; 35,697 conserved domains were encoded in 19,716 of these 40,997 transcripts, the most abundant domain detected being protein kinase (Pfam: PF000069; Supplementary Table 5). Transcripts were not detected for 5,564 annotated genes. Translated ORFs of the remaining 32,566 transcripts were investigated for conserved Pfam domains. This analysis identified 3,338 conserved domains from 2,267 transcripts. The most abundant conserved domain detected was reverse transcriptase (Pfam: PF07727; Supplementary Table 6), which was encoded on 147 transcripts. In addition, 1,178 of the 2,267 putative transcripts encoding conserved domains were flagged as missing the start codon, stop codon, or both by TransDecoder (Haas *et al.* 2013). Of the final 30,299 transcripts that did not overlap annotations, ORFs encoding peptides larger than 80 amino acids could be identified from 9,648 transcripts, of which 4,851 transcripts encoded ORFs with start and stop codons. The largest complete ORF encoded a 1,369-residue peptide with low homology to other annotated oomycete proteins (e.g. 59.4% identity to GenBank Accession RMX66914.1: *P. effusa*), indicating that the annotation software did not identify all loci in the *P. sorghi* genome. These may be improved in the future as the genomes of more closely related species are assembled. The 20,651 transcripts lacking identifiable ORFs ranged from 201 bp to 4,512 bp. These sequences might represent partial transcripts sequenced to low coverage or transcripts originating from non-coding loci. Therefore, 56% of the assembled transcripts overlapped annotated genes and included putative isoforms. A minority of the transcripts (up to 16%) may encode legitimate genes missed by the annotation software, but this subset also includes transcripts originating from transposons. The final 28% of transcripts lacked detected ORFs and may include non-protein-coding transcripts.

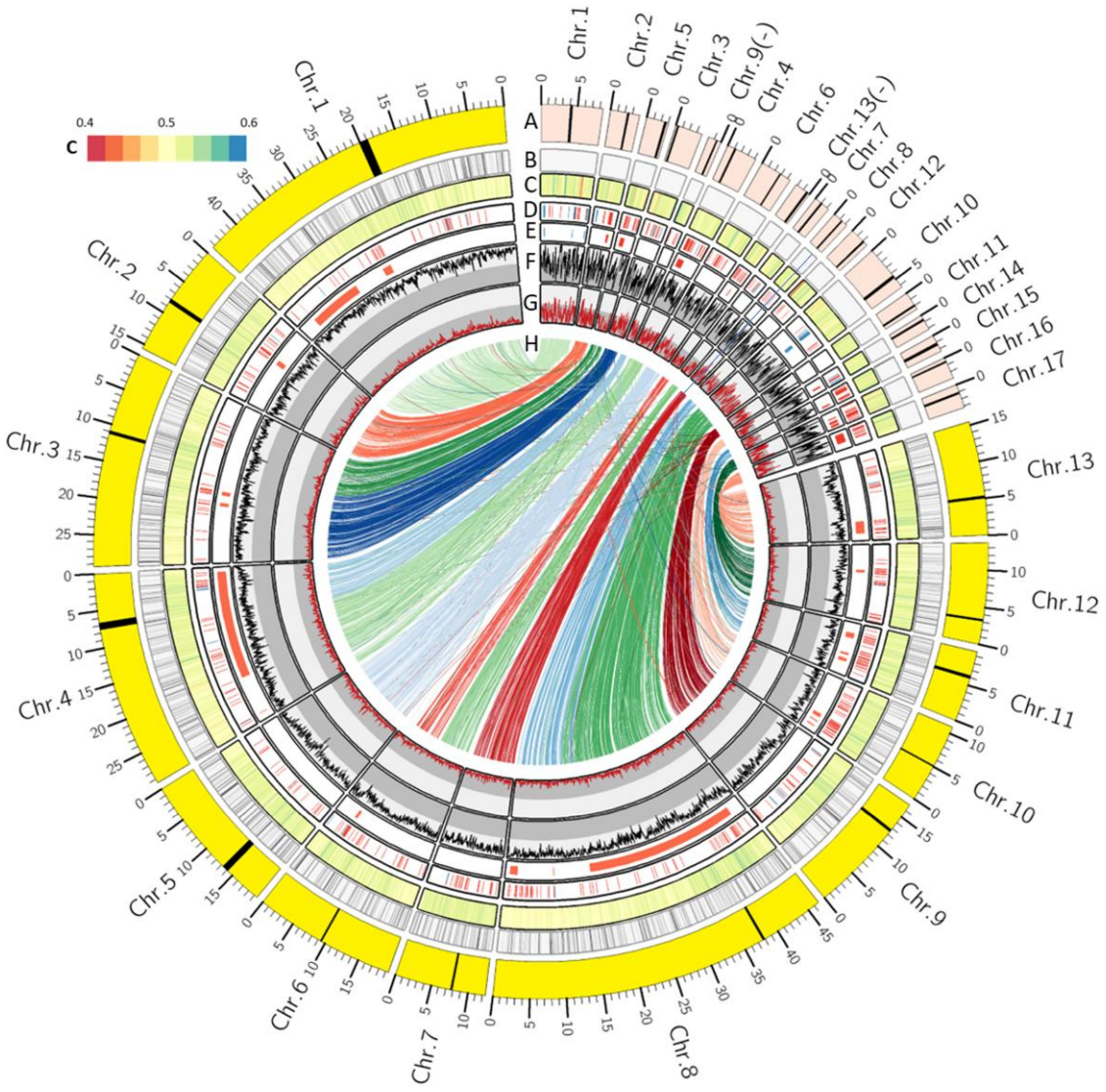

Fig. 3. Comparison of the genome architectures of *Peronosclerospora sorghi* and *Peronospora effusa*. a) Scaled chromosomes showing lengths for *P. sorghi* (yellow) and *P. effusa* (red). Chromosome numbers of *P. sorghi* were ordered based on their designation in *P. effusa*. The scale shows the sizes in Mb. Black bars indicate the putative positions of centromeres (see Fig. 6). b) Distribution of gaps between contigs in the genome assemblies of *P. effusa* (1) and *P. sorghi* (9,804). c) Heatmap of AT content, ranging from 0.4 to 0.6. d) Distribution of annotated effectors. Blue bars represent genes encoding crinklers (98 in *P. effusa* and only 32 in *P. sorghi*). Red bars indicate genes encoding RXLR-like effectors (209 in *P. effusa* and 419 in *P. sorghi*). e) Effector clusters called in the genomes of *P. effusa* and *P. sorghi*. Clusters were called if three or more effectors on the same chromosome had high peptide identity (also see Supplementary Fig. 3). f) Repeat density in 100 kb windows with a 25 kb step. Split in the gray background indicates 0.5. g) Gene density in 100 kb windows with a 25 kb step. Split in the gray background indicates 0.5. h) High levels of synteny depicted by 3,476 links that indicate the genomic positions of single-copy orthologs in the genomes of *P. effusa* and *P. sorghi*. The color of each link reflects the chromosome of *P. effusa* that the ortholog is encoded on.

Of the 19,118 gene models, 14,906 (78.0%) were assigned to 6,229 multi-species orthogroups (Fig. 7a-c). This orthology analysis utilized 39 genome assemblies of 34 species spanning 14 genera across five families of the Oomycota: Peronosporaceae, Pythiaceae, Albuginaceae, Leptolegniaceae, and Saprolegniaceae (Supplementary Table 2). This analysis defined 4,399 core orthogroups common to at least one assembly from each classification of oomycete (Fig. 7e). Core orthogroups accounted for 54.2% of all *P. sorghi* annotations (Fig. 7d). In total, 13,948 *P. sorghi* gene models were assigned to 5,798 orthogroups, which also contained gene models from at least one clade 2 downy mildew species. In contrast, 5,324 of these 5,798 orthogroups contained 7,245 *P. effusa* annotations; the 474 orthogroups, which did not have any *P. effusa* gene models assigned, contained 1,347 gene models from *P. sorghi*. For the majority of the 5,324 orthogroups shared between *P. sorghi* and *P. effusa*, the ratio of genes for each species was 1:1 (3,683 orthogroups; 3,940 genes), for 410 orthogroups (608 genes) it was under 1:1, and for 1,231 orthogroups (8,053 genes) it was greater than 1:1 (Fig. 7f). This indicates that the majority of the orthogroups were not duplicated between the two species and is therefore not consistent with a whole-genome duplication event. An additional 958 *P. sorghi* proteins were assigned to 431 interspecies orthogroups that excluded clade 2 downy mildew species. The 4,212 genes annotated in *P. sorghi* but not assigned to multi-species orthogroups were enriched for putative effectors (164 of 451, $X^2 = 54.381$, $P = 1.651 \times 10^{-13}$) and proteins predicted to be secreted (423/1477, $X^2 = 40.267$ $P = 2.215 \times 10^{-10}$, Table 2). Relative to *Phytophthora* spp., *P. sorghi* was missing similar orthogroups to other downy mildew clade 2 species (Fig. 7), indicating that the lineage has, most likely,

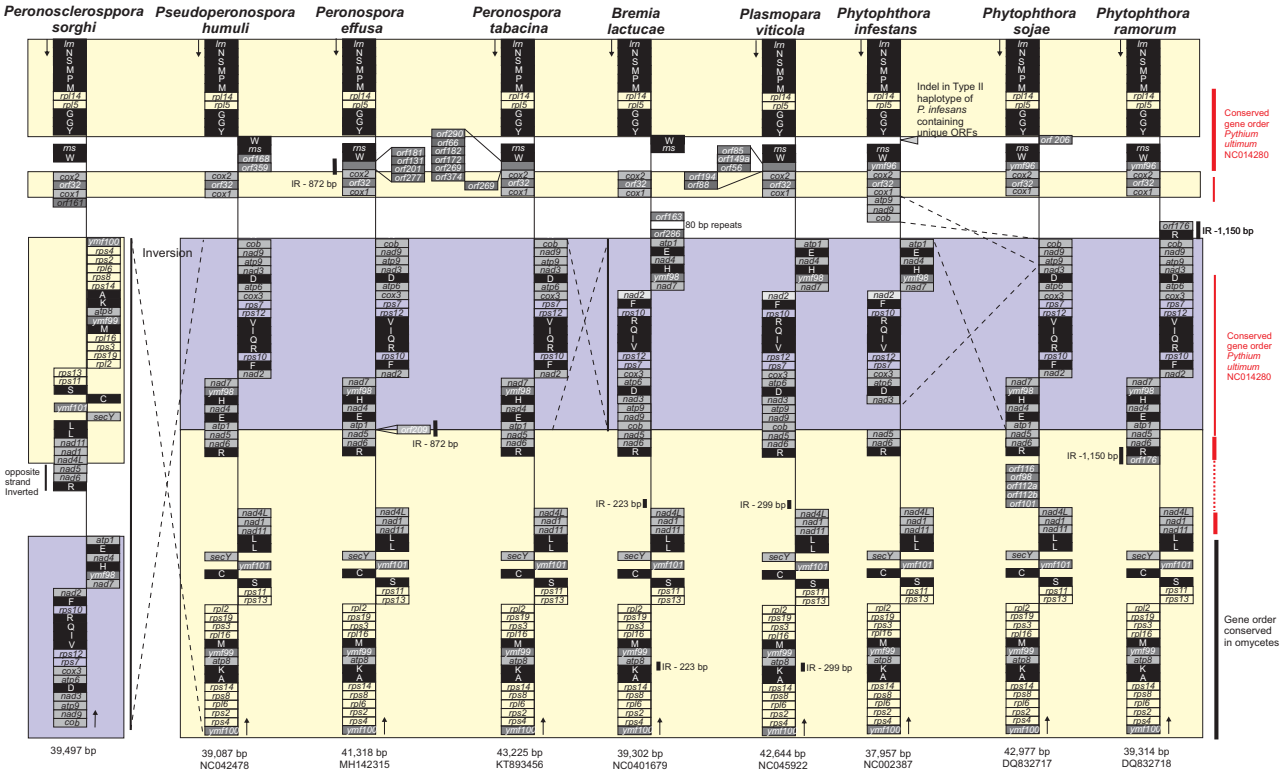

Fig. 4. Comparison of mitochondrial gene order between six genera of the Peronosporaceae. Assemblies of nine species from six genera were analysed. The species are ordered to be consistent with phylogenetics (Fig. 2). Length and NCBI accession of each sequence is at the bottom. All genomes were oriented from the gene encoding the large subunit of the mitochondrial ribosome (*lrn*). Approximate locations of genes are indicated in boxes (not scaled to gene size). Boxes on the left indicate the gene was predicted on the top strand. Boxes on the right indicate the gene was predicted on the bottom (reverse-complemented) strand. Major inversions between assemblies of different species are highlighted by dashed lines. Repeats are annotated by black vertical bars. Colored backgrounds indicate regions of conserved gene order across taxa.

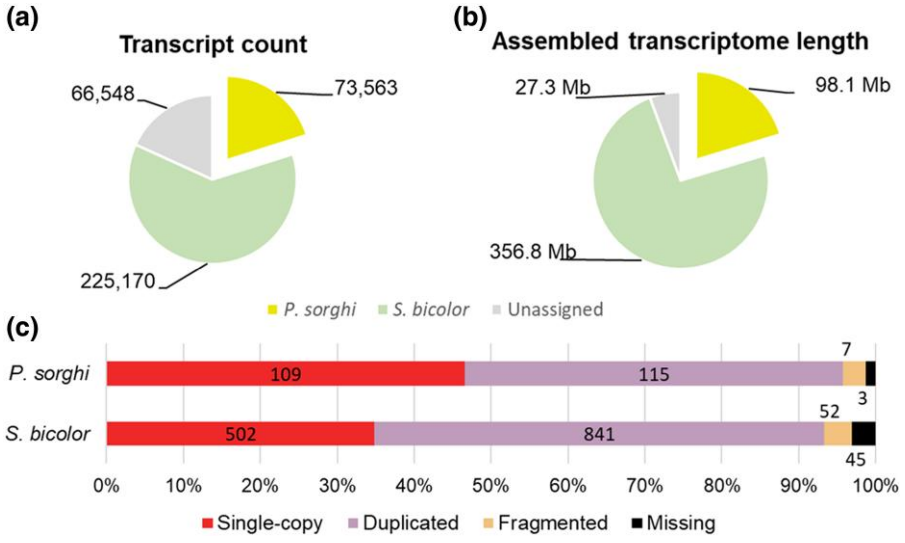

Fig. 5. Transcriptome assembly of the pathogen and host. a) Pie charts showing the proportion and length of sequence assigned to *P. sorghi*, *S. bicolor*, and unassigned to either pathogen or host. b) Stacked bar-chart showing the BUSCO results on the transcriptome assemblies of the pathogen and host.

undergone a similar gene-loss event during the transition to obligate biotrophy. In addition, 677 *P. sorghi* gene models were assigned to 328 orthogroups with *Phytophthora* spp., which were not detected in other downy mildew clade 2 species. These results suggest that the increased gene count in *P. sorghi*, relative to other downy mildew species, is not due to gene retention since divergence from a shared ancestor with *Phytophthora* spp.

The 13 chromosome-scale scaffolds of *P. sorghi* were compared to the 17-chromosome assembly of *P. effusa* based on positions of 3,501 single-copy orthologs. This revealed a high degree of synteny

Table 2. Annotated effectors in the genome of *Peronosclerospora sorghi*.

Signal Peptide	RXLR	EER	(L)WY	CRN	Count	Annotations by MAKER	Annotations from ORFs	Number assigned to orthogroups
+	+	−	−	−	132	132	0	87
+	+	+	−	−	133	58	75	47
+	+	+	+	−	11	4	7	6
+	+	−	+	−	26	15	11	20
+	−	+	+	−	35	21	14	31
+	−	−	+	−	14	6	8	8
−	−	−	+	−	68	31	37	58
+	−	−	−	+	8	3	5	7
−	−	−	−	+	24	17	7	23
+	−	−	−	−	1118	1118	0	848

between the two genomes. The gene order was significantly correlated between the two assemblies ($r = 0.937$, $P < 2.2 \times 10^{-16}$); this clearly validated the assembly of *P. sorghi*. The lineage leading to *P. sorghi* had undergone four chromosome fusions since diverging from the last common ancestor with *P. effusa*. Fusions have occurred between ancestral Chr. 3 and Chr. 5, Chr. 4 and Chr. 9, Chr. 7 and Chr. 13, and Chr. 10 and Chr. 12 of *P. effusa* to form contemporary Chr. 3, Chr. 4, Chr. 6, and Chr. 8, respectively, in *P. sorghi*. The other chromosomes were numbered to retain chromosome order with *P. effusa* (i.e. *P. sorghi* Chr. 1 ≡ *P. effusa* Chr. 1, … *P. sorghi* Chr. 13 ≡ *P. effusa* Chr. 17; Fig. 3h).

Hi-C data allowed the identification of putative coordinates for centromeres. Aligning Hi-C reads back to the genome assembly and calculating the mean cis-interaction distance in 100 kb windows identified chromosomal regions enriched for short-distance interactions (Fig. 8). This demonstrated that the chromosomes were likely organized in Rabl-like configurations (Varoquaux *et al.* 2015; Fig. 1a) that resulted in enriched short-range interactions. Because centromeres have been annotated in *P. effusa*, synteny of flanking single-copy orthologs could be used to validate the coordinates in *P. sorghi*. This analysis showed that orthologs assigned to either chromosome arm in *P. effusa* were located on either side of the putative centromere in *P. sorghi*, indicating that centromere positions were similar between these species. Similar centromere positions could be identified for one of the primogenitors in three of the four chromosome fusions identified in *P. sorghi* relative to *P. effusa*. Chromosome 4 of *P. sorghi* retained the centromere syntenic to *P. effusa* Chr. 9; Chr. 6 of *P. sorghi* retained the centromere syntenic to *P. effusa* Chr. 13, and Chr. 8 retained the centromere syntenic to Chr. 10. For *P. sorghi* Chr. 3, it was not apparent if a centromere was retained from either primogenitor. While the mean cis-interaction distance for bins in the regions syntenic to the centromere of *P. effusa* Chr. 5 was low, the lowest mean cis-interaction distance was at the point of fusion. This could indicate a neocentromere or a false signal because the chromosome arms of the primogenitor are short; the length of regions in *P. sorghi* that contained single-copy orthologs conferring to the chromosome arms was ~2.0 Mb and ~1.3 Mb. Additional evidence for centromere location was sought through identification of Copia-Like Transposons (CoLT). These elements were identified on every chromosome, but they were not enriched in the vicinity of putative centromeres. However, the centromeres of *P. sorghi* were not fully assembled and CoLTs were identified in Chr. 0. Therefore, an improved assembly is required to fully characterize the centromeres of *P. sorghi* and to determine if CoLTs are present in each centromere.

Synteny between *P. sorghi* and *P. effusa* was established using 3,501 genes that were single-copy in both genomes (Fig. 3h, Fig. 9a). We then investigated whether the expansion in gene number in *P. sorghi* was the result of local duplication or genome-wide dispersal. A gene was considered expanded in *P. sorghi* if multiple genes annotated in *P. sorghi* were assigned to an orthogroup containing only one *P. effusa* gene (>1:1); in total, there were 5,445 such multicopy genes in *P. sorghi*. Comparison of the chromosomal positions of these expanded genes in *P. sorghi* established the presence of two broad categories of orthologous genes; 1,511 were located on syntenic chromosomes and 3,532 genes were scattered through the genome on non-syntenic chromosomes (Fig. 9b); in addition, 402 were on Chr. 0. Only a few genes (812) were expanded in *P. effusa* relative to 326 *P. sorghi* genes (Table 3). In total, 667 of these were located on the syntenic chromosome (Fig. 9c). The remaining 9,846 genes in *P. sorghi* were classified into other categories: 3,329 genes were orthologous to multiple *P. sorghi* and multiple *P. effusa* annotations (>1:>1; Table 3), which were located on both syntenic and non-chromosomes (Supplementary Fig. 4); 611 genes annotated in *P. sorghi*, which were assigned as the lone representative of *P. sorghi* to interspecies orthogroups lacking *P. effusa* (1:0); 1,694 genes assigned to interspecies orthogroups containing multiple *P. sorghi* annotations but no *P. effusa* annotations (>1:0); and 4,212 *P. sorghi* annotations lacking orthology with any other oomycete. The pattern of genome-wide dispersal of expanded genes is consistent with duplication by retrotransposition rather than local duplication.

Transcript abundance of syntenic and non-syntenic genes was compared to provide evidence for the presence of potential pseudogenes in each set. In total, transcripts were detected for 13,554 of the 19,118 annotated genes. Transcripts were detected for 3,475 of the 3,501 single-copy orthologs identified between *P. sorghi* and *P. effusa* (1:1). Their mean level of abundance was 92.87 fragments per kilobase of exon per million mapped fragments (FPKM; Fig. 9d). No transcripts were identified for the other 26 of these genes, 18 of which were located on non-syntenic chromosomes. Transcripts were detected for 1,350 of 1,511 expanded genes located on syntenic chromosomes relative to *P. effusa*, 2,203 of 3,532 expanded genes on non-syntenic chromosomes, and 192 of 402 expanded genes on Chr. 0 (Table 3). The mean FPKM of expressed, expanded genes in syntenic positions was lower than for genes classified as 1:1 (Fig. 9d; Table 3). Many expanded genes with high transcript abundance were in syntenic positions (Fig. 9b), indicating that these genes were the ancestral copies. The mean FPKM for expanded genes on non-syntenic chromosomes and Chr. 0 was significantly lower than the mean FPKM for both genes classified as 1:1 and expanded genes in syntenic positions (Table 3; Fig. 9d). The distribution of FPKM for genes on non-syntenic chromosomes was bimodal (Fig. 7d; Supplementary Fig. 5), suggesting that some duplicated genes on non-syntenic chromosomes had higher transcript abundance (1,378 genes with mean 30.08

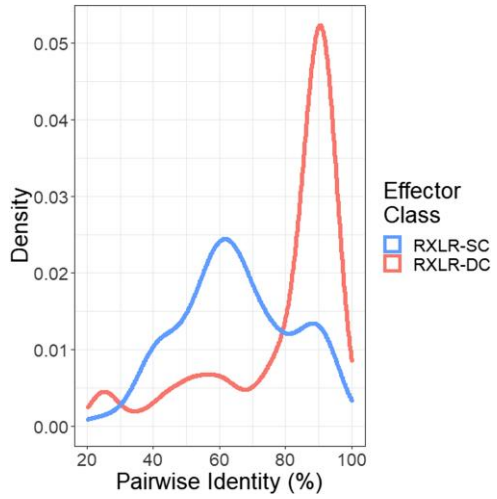

Fig. 6. Effector distribution in the genome assembly of *Peronosclerospora sorghi*. Density of 493 pairwise amino acid identities of 232 of 419 putative RXLR effector proteins assigned to 71 clusters by CD-hit. Pairwise identities of genes were more similar when the genes were located on different chromosomes (RXLR-DC) than when the genes were located on the same chromosome (RXLR-SC). Only eight of the 32 genes encoding CRNs were identified in the three clusters, so pairwise similarities are not plotted.

FPKM), while others had lower transcript abundance (803 genes with mean 0.43 FPKM). Treated independently, both these subsets had significantly lower transcript levels than expanded genes located in syntenic positions ($P = 0.00128$ & $P < 2.2 \times 10^{-16}$). Transcripts were detected for 313 of 326 *P. sorghi* single-copy genes that appeared expanded in the genome of *P. effusa* (Table 3). The transcript abundance of these genes was not significantly different from single-copy genes in *P. sorghi* (Fig. 9d). Genes lacking interspecies orthology had significantly lower transcript abundance than most other genes (Fig. 9d). Similar results were obtained when the CPM was used instead of FPKM (Supplementary Fig. 6). Overall, these results show that genes duplicated to non-syntenic positions in the genome of *P. sorghi* have fewer detectable transcripts than their ancestral, syntenic counterparts.

The transcript abundance of the 1,477 annotated genes encoding predicted secreted proteins was also investigated. Transcripts were detected for 78% of these genes, which was significantly higher than the portion of transcripts detected for genes encoding non-secreted peptides ($X^2 = 41.96$, $P = 9.318 \times 10^{-11}$). The mean transcript abundance of genes predicted to encode a secreted product was also higher than genes not predicted to encode secreted proteins (Table 3). Some of the genes with the most abundant transcripts encoded secreted peptides (Fig. 9d) including annotated effectors. Therefore, transcripts of genes encoding signal peptides, including those on non-syntenic chromosomes or when lacking interspecies orthologs, were detected at higher abundances than the genome-wide average.

Coding sequences were compared to determine if gene expansion in *P. sorghi* had resulted in changes in selection pressures acting on single copy and duplicated genes. The ratio of non-synonymous to synonymous polymorphisms (dN/dS) in *P. sorghi* genes since divergence from the common ancestor with *P. effusa* was calculated and summarized considering chromosomal location of duplicated genes and whether transcripts were detected for duplicated genes. The mean dN/dS of the 3,499 single-copy orthologs (1:1) was 0.125, consistent with these genes undergoing purifying selection since divergence. The mean dN/dS for genes that were probably ancestral (>1:1 on syntenic chromosomes), and for which transcripts were detected, was 0.140. While this is consistent with purifying selection, it was significantly higher than the dN/dS for single-copy orthologs (Fig. 9e), suggesting relaxed purifying selection. The mean dN/dS for expanded genes (>1:1) on non-syntenic chromosomes for which transcripts were detected was 0.150. This was also significantly different from single-copy orthologs (1:1), but not expanded genes (>1:1) on syntenic chromosomes with transcripts (Fig. 9e). The mean dN/dS for expanded genes (>1:1) on non-syntenic chromosomes lacking transcripts was 0.167 and was significantly higher than the means for the other described subset of genes (Fig. 9e). The mean dN/dS for expanded genes on syntenic chromosomes and lacking transcripts was 0.142 and did not differ significantly from the previously described subsets (Fig. 9e). Selection may have been lost on some duplicated genes, consistent with pseudogenization. Therefore, comparative sequence analysis suggests that duplicated genes have significantly different selection pressures than conserved, single-copy genes, and that duplicated genes lacking transcription have been under even weaker purifying selection.

Additional evidence for pseudogenization of expanded genes was sought via global characterization of genes. The mean encoded peptide length (MEPL) for the 3,501 single-copy orthologs was 491.5 residues. This was not significantly different from the MEPL for the 1,511 expanded genes (>1:1) on syntenic chromosomes or the 326 single-copy genes expanded in *P. effusa* (1:>1; Table 3). All other subsets of proteins had significantly shorter MEPLs, including >1:1 genes on non-syntenic chromosomes. Shorter MPELs may indicate nonsense mutations introducing premature stop codons in the ORF. The mean intro count (MIC) of single-copy orthologs was 1.66 and did not significantly differ from the MIC of >1:1 genes on syntenic chromosomes or single-copy *P. sorghi* genes not found in *P. effusa* (1:0; Table 3). The MIC of 1:>1 genes was significantly higher than 1:1 genes. The MIC of all other gene classifications was significantly lower, including >1:1 genes on non-syntenic chromosomes. Reduced intron counts could be due to nonsense mutations or intron loss due to duplication by retrotransposition. The distribution of intron lengths was summarized for each gene category in the context of transcript detection. For 1:1 genes, a major peak was detected with a mode intron length (MIL) of 67 base pair (bp; a minor peak was present with a MIL of 25 bp that may be an annotation artifact). Similar profiles were obtained for >1:1 genes on syntenic chromosomes, 1:>1 genes, and 1:0 genes (Fig. 9f). Transcripts were detected for most of the genes in these categories (Table 3; Fig. 9f). For the other gene categories, two similar peaks could be detected; however, the major and minor peaks had similar counts, and genes with zero transcripts detected were found under both peaks (Fig. 9f). For all annotations, the peak of larger MILs may represent introns of optimal size for splicing in *P. sorghi*; the peak of smaller MILs might represent introns incorrectly predicted by the annotation software. It therefore seems likely that the 19,118 genes annotated for *P. sorghi* contain true protein-coding genes often retaining synteny with *P. effusa* as well as a large number of probable pseudogenes that lacked evidence for transcription and were predicted to encode shorter peptides in fewer exons.

The *P. sorghi* assembly was investigated to determine whether the genome was compartmentalized in relation to transcription and gene classification. Both the 5' and 3' intergenic distances between genes were bimodal, indicative of gene-dense (intergenic distances of less than 6.5 kb either side) and gene-sparse (intergenic distances greater than 6.5 kb on either or both sides) regions in the genome (Fig. 10a). Genes for which transcription was and

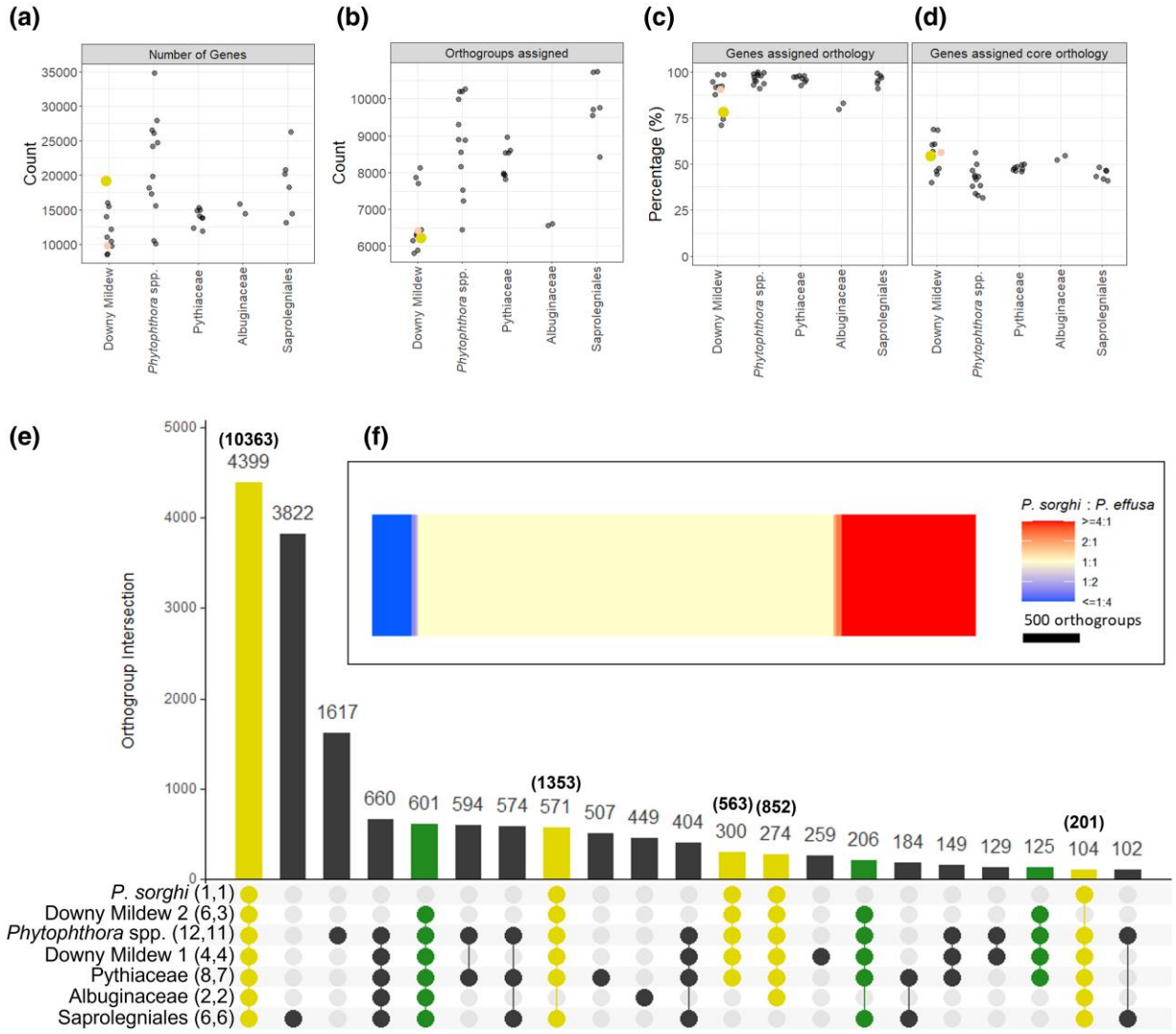

Fig. 7. Orthology analysis of *Peronosclerospora sorghi* with 38 other genome assemblies, representing 31 other species of oomycetes. a) Illustration of the number of genes annotated in the genomes used for orthology analysis. b) Scatter plot showing the number of orthogroups that protein sequences from each species were assigned to. c) Scatter plot showing the percentage of genes for each species assigned to multi-species orthogroups. d) Same as C, but only considering the 4,399 core orthogroups. In b to d, both downy mildew clades 1 and 2 are condensed to a single category on the x-axis. The large yellow dot indicates *P. sorghi*, the red dot indicates *P. effusa* UA202013. e) An UpSet plot showing multi-species intersections of orthogroups between oomycete phylogenetic classifications. Numbers in parentheses next to the classification indicate the number of assemblies and number of species surveyed under that classification. Only intersections totaling more than 100 orthogroups were plotted; therefore, 16,030 of the 17,588 orthogroups calculated are illustrated. The number of orthogroups in each intersection is annotated above the bar. Numbers in parentheses above the bar indicate the number of genes annotated in *P. sorghi* assigned to the orthogroups in the intersection. In total, 13,332 of the 14,906 genes annotated in *P. sorghi* and assigned to multi-species orthogroups are illustrated. Blue bars highlight intersections containing *P. sorghi*. Red bars indicate intersections containing downy mildew clade 2 species, but not *P. sorghi*. Therefore, red bars may indicate unique gene losses in the lineage leading to *P. sorghi* but retained in other related downy mildew clade 2 species. f) Heatmap demonstrating the ratio of sequences annotated in the genomes of *P. sorghi* and *P. effusa* assigned to 5,324 orthogroups that were assigned proteins from both species. The majority of orthogroups are balanced, with the same number of proteins assigned from each species.

was not detected were located in both such gene-dense and the gene-sparse regions. When orthology was considered, all previously ascribed categories were distributed across the different genomic compartments (Fig. 10b). Therefore, conserved single-copy orthologs were annotated in both the gene-sparse and gene-dense region. In addition, annotations that may represent pseudogenes were also located in the gene-sparse and gene-dense regions, regardless of transcriptional status. The dN/dS of genes in the gene-dense region did not significantly differ from the dN/dS of genes in the gene-sparse regions ($P = 0.128$). Therefore, high-confidence protein-coding genes and probable pseudogenes were annotated in both the gene-dense and gene-sparse compartments of the *P. sorghi* genome.

DGE was analyzed to investigate transcriptional differences pre- and post-sporulation of *P. sorghi*. Multi-dimensional scaling indicated that normalized read counts of pathogen and host genes for each of the two biological conditions could be distinguished from one another across the three replicates (Fig. 11, a and b). A total of 10,496 genes passed the five counts per-million threshold. There was evidence for 330 up-regulated and 130 down-regulated genes post-sporulation, compared with pre-sporulation after correction for false discovery (Fig. 11 C). These two subsets of genes were enriched for peptides predicted to be secreted, with 76 up-regulated genes predicted to be secreted ($P < 2{,}2 \times 10^{-16}$) and 25 down-regulated genes ($P = 3.638 \times 10^{-5}$). Pfam domains encoded by differentially expressed genes suggests that up-regulated

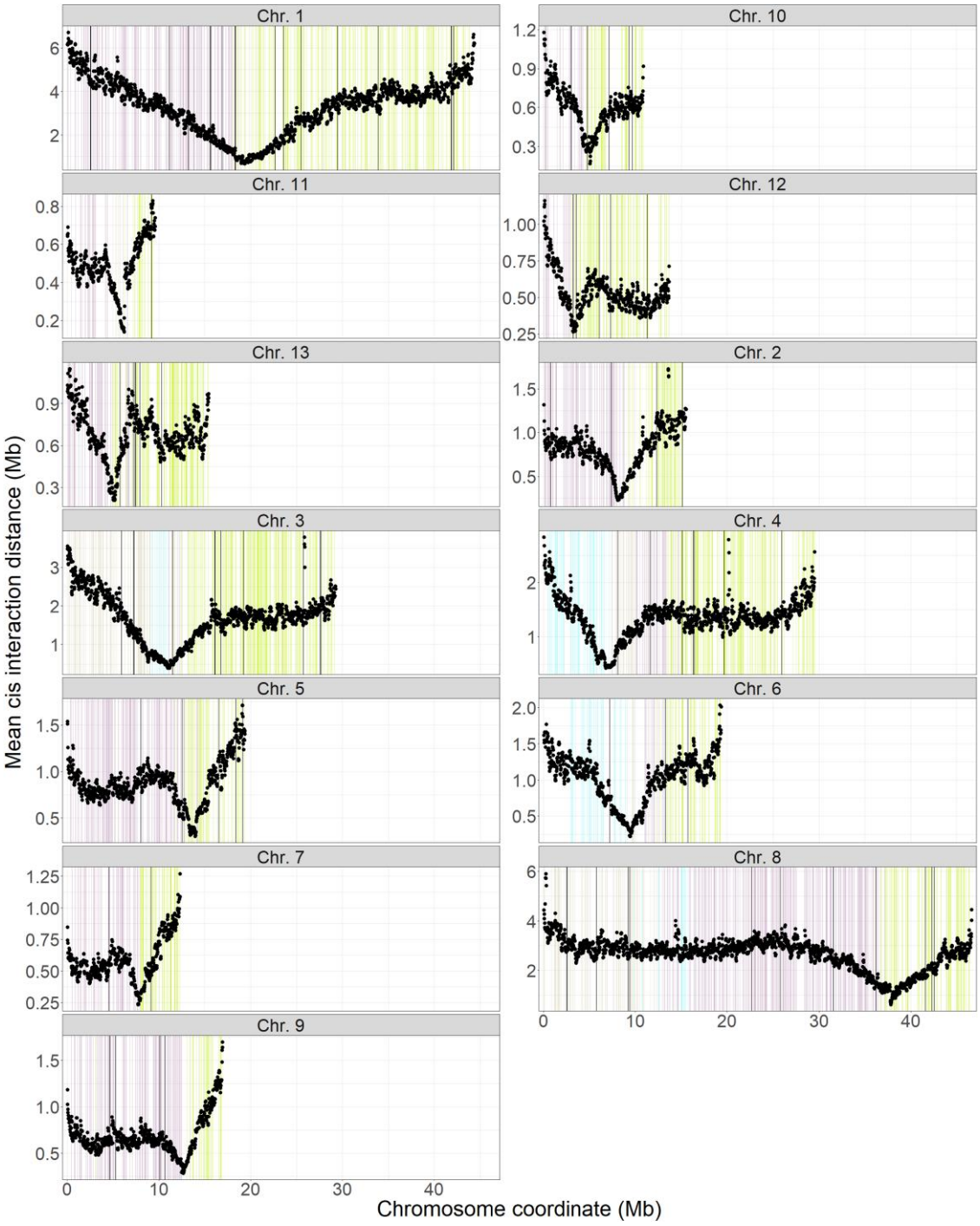

Fig. 8. Conservation of centromeric regions between the genomes of *Peronosclerospora sorghi* and *Peronospora effusa*. The mean cis-interaction distance between Hi-C reads was calculated in 100 kb windows, with a 25 kb step, across the genome of *P. sorghi*. There was a single region enriched for short-distance *cis*-interactions per chromosome, corresponding to the centromeric region. Coordinates of single-copy orthologs in the assembly of *P. sorghi*, used to establish synteny with *P. effusa* (Fig. 5), were plotted in the background. The colors indicate which chromosome arm of *P. effusa* the ortholog is located on; proximal chromosome arms are purple, and distal chromosome arms are green for each chromosome. Most chromosomes have only two colors because they share the same ancestral conformation. Chromosomes 3, 4, 6, and 8 have four colors because these chromosomes have undergone fusions relative to *P. effusa*. Brown depicts orthologs derived from a proximal region, and blue depicts orthologs derived from a distal region. Except for Chr. 3, the region enriched for short-distance *cis*-interactions co-locates with the change in color, indicating that the centromere positions are similar between the two species. In Chr. 3 the putative centromere appears to be at located at the fusion point between primogenitors. Black bars indicate the location of Copia-like transposons in the genome of *P. sorghi*.

proteins include transporters, cutinase, chitinase, cellulose synthase, chitin synthase, and a necrosis inducing protein (Supplementary Table 7). Fewer down-regulated genes could be due to mycelium within the plant tissue after sporulation. In contrast, more genes from the host were down-regulated (1,498) than up-regulated (522) post-sporulation (Fig. 11d). These included six down-regulated genes encoding disease resistance proteins, one up-regulated, 21 down-regulated genes encoding

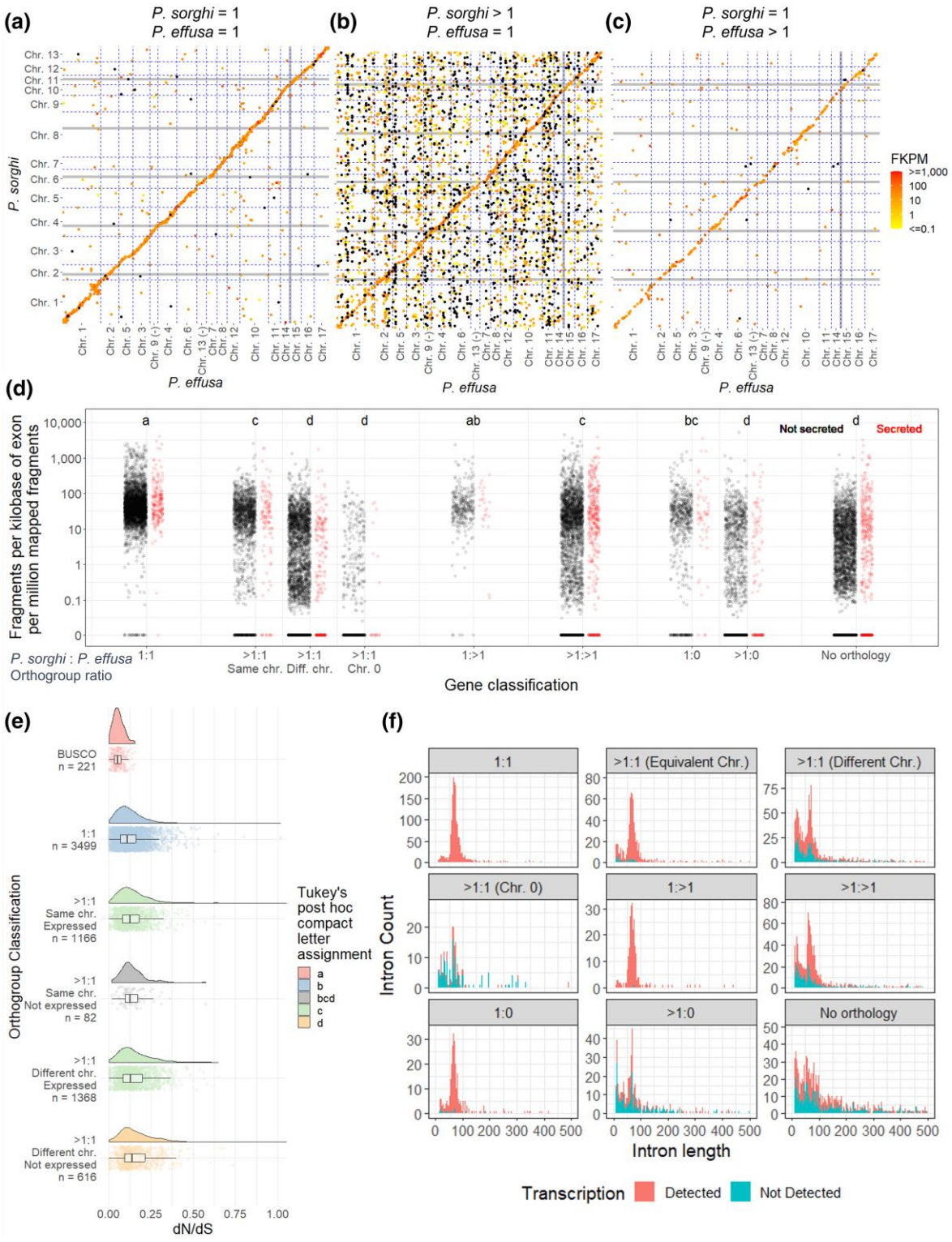

Fig. 9. Multiplication and dispersal of genes in the genome of *Peronosclerospora sorghi*. a) Scatter plot showing the coordinates of single-copy genes in the genomes of *P. effusa* and *P. sorghi*. The x-axis is ordered to demonstrate collinearity, like Fig. 3h. Axes are not scaled to size. Gray gridlines indicate 50 Mb boundaries. Blue dotted lines indicate chromosome boundaries. Points are colored by their transcript abundance (FPKM). Black dots indicate 0 FPKM. b) Like A, except genes were expanded in *P. sorghi*, single copy in *P. effusa*. c) Like A, genes were single copy in *P. sorghi*, expanded in *P. effusa*. d) Scatter plots showing the transcript abundance of genes annotated in *P. sorghi*. Genes were assigned to different categories (x-axis) based on their orthology to *P. effusa*. Genes expanded in *P. sorghi* (>1:1) were sub-categorized based on their synteny with *P. effusa*. Genes encoding peptides predicted to be secreted were plotted in red for each category. Lowercase letters indicate different groups of significance as calculated by Tukey's HSD test. e) Raincloud plots depicting signatures of selection on orthologs single copy in both *P. sorghi* and *P. effusa* and expanded in *P. sorghi* but single copy in *P. effusa*. Genes expanded in *P. sorghi* were split on the basis of synteny with *P. effusa* and detection of transcripts (FPKM > 0). BUSCO genes are shown as representative of highly conserved genes. Colors indicate different groups of significance as calculated by Tukey's test. f) Distribution of intron lengths for genes annotated in *P. sorghi*. Genes were split based on their synteny with *P. effusa*. The color of the stacked histogram indicates the number of introns originating from genes lacking or with detected transcripts.

Table 3. Overview of orthologous gene assignment for *Peronosclerospora sorghi* annotations relative to *Peronospora effusa*.

P. sorghi: P. effusa orthogroup ratio	Assigned genes	Expressed genes	Differential gene expression (up: down).	Mean FPKM	Mean encoded peptide length[a]	Mean intron count[a]	Genes with secreted product	Expressed genes with secreted product	Mean FPKM
1:1	3,501	3,475 (99.3%)	90 : 25	92.87	491.5 (a)	1.66 (a)	202	201 (99.5%)	158.35
>1:1—syntenic chr.	1,511	1,350 (89.3%)	31 : 10	44.63	472.5 (a)	1.67 (a)	113	104 (92.0%)	89.04
>1:1—non-syntenic chr.	3,532	2,203 (62.4%)	18 : 17	12.82	201.0 (d)	1.29 (b)	190	135 (71.1%)	25.39
>1:1 Chr. 0	402	192 (48.8%)	0 : 0	6.89	250.9 (c)	1.16 (bc)	17	7 (41.2%)	9.60
1:>1	326	313 (97.8%)	8 : 5	84.72	514.8 (a)	2.77 (d)	24	22 (91.7%)	76.3
>1:>1	3,329	2,327 (69.9%)	128 : 28	73.77	342.1 (b)	1.21 (b)	356	302 (84.8%)	106.6
1:0	611	539 (88.2%)	14 : 7	69.71	294.2 (c)	1.67 (a)	53	50 (94.3%)	181.45
>1:0	1,694	768 (45.3%)	9 : 12	34.28	198.1 (d)	1.15 (bc)	99	61 (61.6%)	33.36
No orthology	4,212	2,379 (56.5%)	32 : 26	24.21	133.5 (e)	1.10 (c)	423	280 (66.2%)	55.56
Total	19,118	13,554 (70.9%)	330 : 130				1,477	1,155 (78.2%)	

[a] Parenthesized letters indicate significant differences in pairwise analysis (Tukey's HSD).

transporters, 13 up-regulated, and 10 down-regulated genes encoding dehydration response element binding proteins. These differentially regulated host genes could include genes in response to the pathogen or in response to the environment. There are extensive data available for further study of the host complement that is beyond the focus of this paper. The differentially expressed pathogen genes were enriched for genes syntenic with *P. effusa*. Synteny could be established for 356 of the 458 differentially expressed genes, of which 116 were single copy in both *P. sorghi* and *P. effusa*; 72 were expanded in *P. sorghi* relative to *P. effusa*; 13 were single copy in *P. sorghi* but expanded in *P. effusa*; 155 were multicopy in both (Table 3). The expanded, differentially regulated genes in *P. sorghi* were enriched for genes which retained synteny ($X^2 = 8.68$, $P = 0.0032$). Therefore, genes that retain synteny between these two distinct species are also differentially regulated during the lifecycle of *P. sorghi*.

Discussion

This paper describes a third chromosome-scale genome assembly for an oomycete following the genetically oriented genome assembly of *Bremia lactucae* (Fletcher *et al.* 2021) and the telomere-to-telomere (T2T) genome assembly of *Peronospora effusa* (Fletcher *et al.* 2022). All three of these species represent different genera of oomycetes, which cause downy mildew diseases on different plants. *P. sorghi* is phylogenetically closer to *P. effusa* (downy mildew clade 2) than *B. lactucae* (downy mildew clade 1) (Fig. 2). The genome architecture of *P. sorghi* was predominantly syntenic but different from the genomes of *B. lactucae* and *P. effusa* because it was more homozygous, larger, and had fewer chromosomes. Increased homozygosity of isolate P6 may have been beneficial for genome assembly because there is little divergence between the two haplotypes. This may explain why, in combination with Hi-C, it was possible to obtain a chromosome-scale genome assembly of *P. sorghi*. Increased homozygosity of *P. sorghi* may reflect its homothallism, while the other two species are heterothallic. In future studies, it will be interesting to determine how heterozygosity varies between isolates of these species. Previously, the ancestral chromosome-state of downy mildews and many *Phytophthora* spp. has been described as that of *B. lactucae* and *P. effusa*, which share high collinearity between their respective chromosomes (Fletcher *et al.* 2022). While the genome of *P. sorghi* has fewer chromosomes, it retains a high level of synteny with the ancestral state; nine *P. sorghi* chromosomes were colinear with nine chromosomes of *P. effusa* and *B. lactucae* (Fig. 3) (Fletcher *et al.* 2022). The four other *P. sorghi* chromosomes are unique fusions compared to *P. effusa* and *B. lactucae* but retain a high level of synteny with their primogenitors (Fig. 3). Therefore, this genome assembly presents a contemporary state derived from the conserved ancestral configuration.

Identification of centromeres also supports the 13-chromosome architecture of *P. sorghi*. Analysis of Hi-C data identified one centromeric region on each chromosome (Fig. 1a, 8). Aligning the assembly of *P. sorghi* to the assembly of *P. effusa* showed which centromeres were lost after chromosome fusions and that the retained centromeres were in approximately syntenic positions (Fig. 8). It was not possible to compare centromere sequences between the two species because the current assembly of *P. sorghi* is based on Illumina 10× Genomics sequencing in contrast to the gapless chromosomes of *P. effusa* that is based on Pacific BioSciences HiFi reads (Fletcher *et al.* 2022). It was also not possible to assemble telomeric sequences for all 13 chromosomes of *P. sorghi*. In the future, it will be helpful to generate long read assemblies

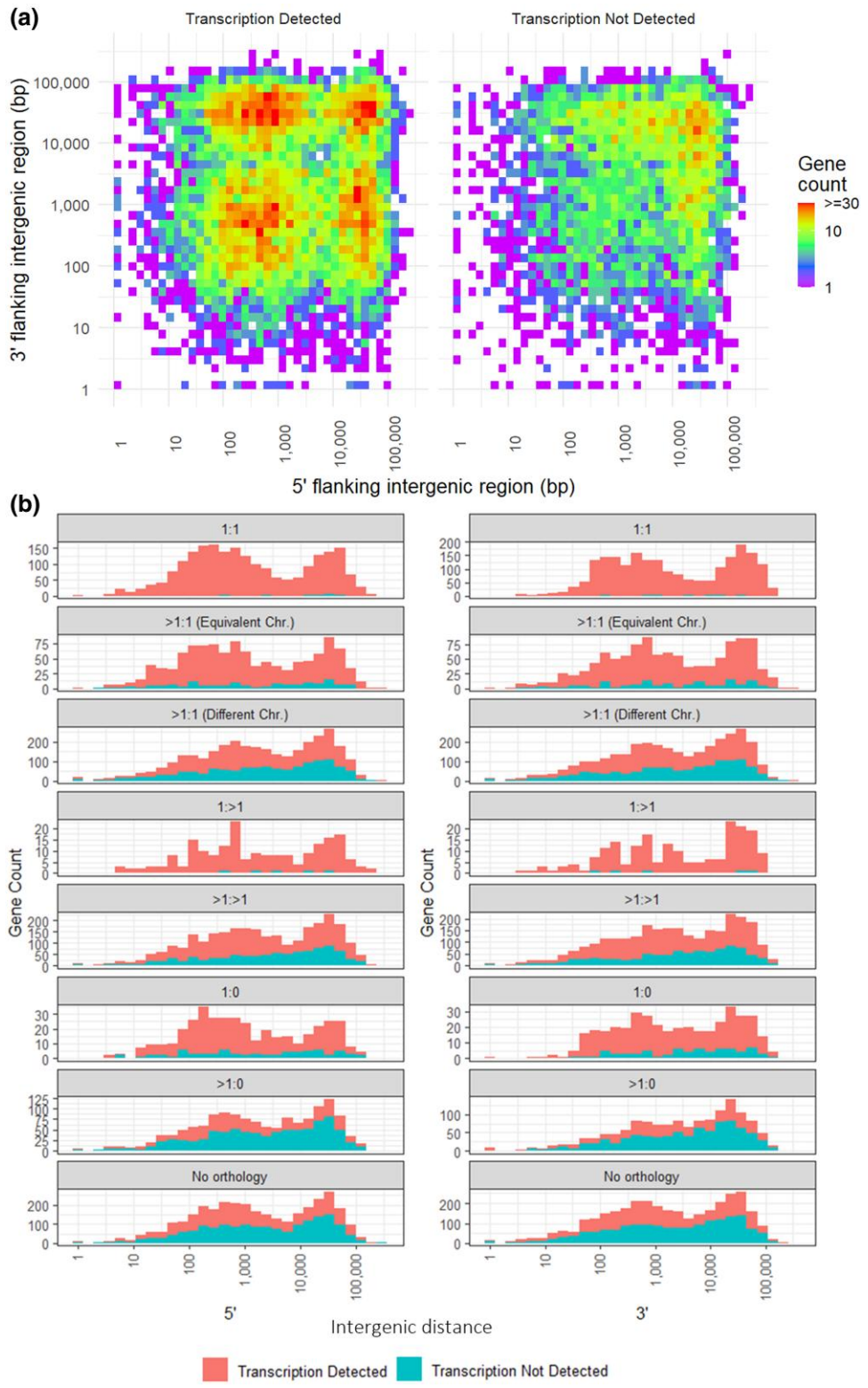

Fig. 10. Compartmentalization of the *Peronosclerospora sorghi* genome based on intergenic distances. a) For each annotated gene, the 5' distance (bp) to the next gene was plotted on the y-axis, and the 3' distance (bp) was plotted on the y-axis. A heatmap was plotted for genes both with and without detectable transcripts. The heatmaps demonstrate that intergenic distances on both flanks were bimodal. Some genes had short intergenic distances (<6.5 kb) on both flanks, consistent with high gene density. Others had long intergenic distances (>6.5 kb) on both or either flank, consistent with gene sparsity. b) Stacked histograms on the left show the 5' intergenic distance, while the right show the 3' intergenic distance. Separate histograms were plotted for genes based on their synteny with *P. effusa*. All gene categories had genes consistent with high gene density and gene sparsity. The stacked histogram indicated the number of genes for which transcripts could and could not be detected. Genes with and without detectable transcripts had intergenic distances consistent with high gene density and gene sparsity.

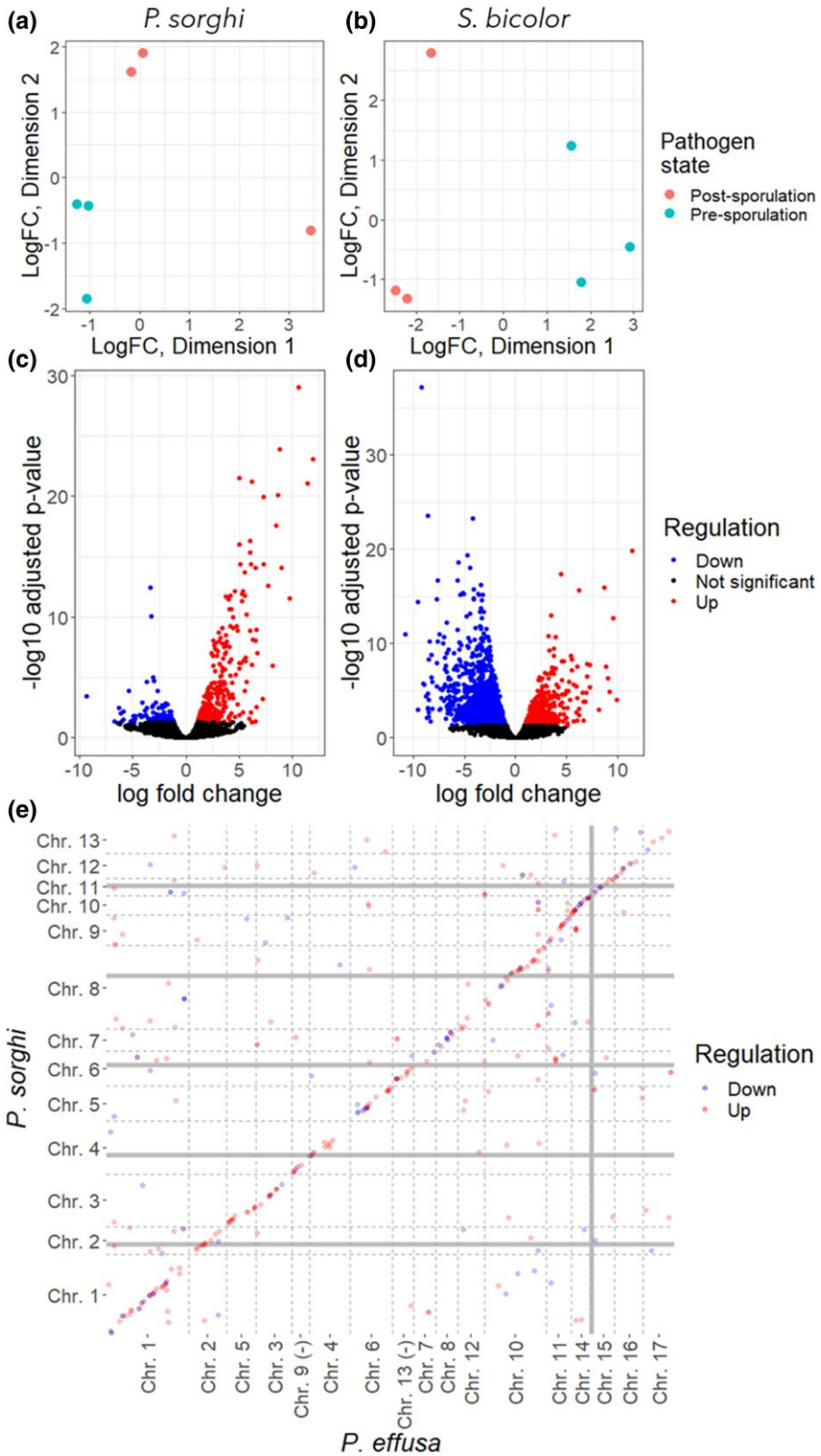

Fig. 11. Differential gene expression of *Peronoscleropsora sorghi* and *Sorghum bicolor*. a) Multi-dimensional scaling (MDS) plots demonstrating that the gene counts of the pathogen can be distinguished based on their pre- or post-sporulation source. b) MDS plot demonstrating that the gene counts of the host could be distinguished based on their pre- or post-sporulation source. c) Volcano plot demonstrating up- and down-regulated genes of the pathogen. d) Same as C, but for the host. e) Scatter plot of differentially expressed genes showing enrichment of genes syntenic with *P. effusa*.

of *Peronosclerospora* spp. to assemble the centromeres and telomeres completely and determine whether they harbor sequences conserved with *P. effusa*.

Genome-wide synteny indicates that proliferation of repeat sequences underlies the increased genome size of *P. sorghi* relative to *P. effusa* rather than a whole-genome duplication event. The assembled length of *P. sorghi* was more than six times that of the T2T assembly of *P. effusa*, although their chromosomes were highly syntenic. Therefore, the inflated genome size was not caused by auto- or allopolyploidy. Instead, since diverging from their most recent common ancestor, the genome of *P. sorghi* has accumulated or retained more transposable elements than the genome of *P. effusa*. Comparing 100 kb windows across the two assemblies, *P. sorghi* had a higher density of repeats and a lower density of genes annotated than *P. effusa* (Fig. 3, f and g). This is consistent with the previously reported correlation of repeat content with genome size across multiple oomycete species, including downy mildews (Fletcher *et al.* 2019). Inflation of genome size exclusively by retention of transposable elements is different from some *Phytophthora* spp., where whole-genome duplication in addition to proliferation of transposable elements have resulted in increased genome sizes (Morales-Cruz *et al.* 2020; Ayala-Usma *et al.* 2021).

Despite its high gene-model count, *P. sorghi* probably underwent similar gene-loss events to other downy mildews due to its biotrophic and non-flagellate lifestyle. Evidence for this comes from the high number of orthogroups that are present in *Phytophthora* spp. and species in the Pythiaceae, but absent in *P. sorghi* and other downy mildew species; only 104 interspecies orthogroups were assigned a total of 201 sequences from *P. sorghi* but no other downy mildew clade 2 species (Fig. 7). This is consistent with *P. sorghi* undergoing similar gene-loss events as other clade 2 downy mildews. More contiguous assemblies of *Phytophthora* spp. from multiple *Phytophthora* clades as well as more downy mildew species, may reveal genes lost in downy mildew species, which encode proteins dispensable in the biotrophic lifecycle, but are required for the necrotrophic phase of the lifecycles of *Phytophthora* species. Since *Phytophthora* species produce flagellated zoospores, the flagellated state is likely ancestral to all downy mildews. Loss of flagella genes likely occurred multiple times in clade 2 downy mildews because *Sclerospora* and *Pseudoperonospora* spp. retain flagellated zoospores, but *Peronosclerospora*, *Hyaloperonospora*, and *Peronospora* spp. do not (Fig. 2) (Palti and Cohen 1980; Fletcher *et al.* 2018; Crouch *et al.* 2022).

The increased number of genes in *P. sorghi* relative to other downy mildews is likely due to gene expansion through retrotransposition, which has led to the accumulation of pseudogenes (Fig. 9) and dispersed, transcribed genes, including genes encoding putative effectors (Fig. 3). For the majority of orthogroups, the ratio of genes assigned from *P. sorghi* compared to *P. effusa* was 1:1 (Fig. 7f), indicating that the majority of orthologs were conserved. Further, the majority of these orthogroups were single copy in *P. effusa* and *P. sorghi*, syntenic, transcribed, had signals of purifying selection, and a normal distribution of intron lengths (Fig. 9). A minority of orthogroups (1,231) were expanded in *P. sorghi* relative to *P. effusa*, accounting for the majority of orthologous genes (8,053 genes) annotated in *P. sorghi* (Fig. 7); a minority of these genes were syntenic between *P. sorghi* and *P. effusa* and retained similar characteristics with the single-copy set, suggesting that they were ancestral (Fig. 9, Table 3). The remainder of these genes were on non-syntenic chromosomes, suggesting duplication by retrotransposition. Transcripts could not be detected for a larger portion of these genes (Table 3) and when transcripts were detected, their transcript abundance was significantly lower (Fig. 9). In addition, genes on non-syntenic chromosomes encoded shorter peptides in fewer exons and had a large number of short introns predicted, characteristic of pseudogenes. Of the expanded genes, those that were transcribed had significantly weaker signals of selection than single-copy genes; however, they had similar signals of selection regardless of whether they were syntenic, suggesting that both the ancestral and transposed copies were under similar relaxed selection (Zhang 2003). Some of the transcribed non-syntenic genes had short introns predicted, suggesting nonsense alleles or transcribed pseudogenes (Fig. 9f). Expanded genes with undetected transcription had weaker signals of purifying selection consistent with pseudogenization (Fig. 9e). While the genome of *P. sorghi* was consistent with compartmentalization (Fig. 10a), the compartments did not have distinct complements of putative pseudogenes and protein-coding genes; syntenic genes annotated as single copy in both *P. sorghi* and *P. effusa* were located in both gene-dense and gene-sparse compartments, as were the syntenic and non-syntenic subset of the expanded genes (Fig. 10b). Genes lacking transcription, expanded in *P. sorghi*, which had weaker signals of purifying selection (Fig. 9e, i.e. pseudogenes), were also present in both compartments (Fig. 10b). Although both single copy and expanded genes were differentially expressed (Table 3), the majority were syntenic (Fig. 11 e). In summary, the majority of orthologs between *P. sorghi* and *P. effusa* are single copy, but the majority of annotations in the *P. sorghi* genome originate from a minor portion of expanded orthologs. Some of the expanded genes may be functional paralogs, while others are likely pseudogenes. These findings should be considered when analyzing other oomycete assemblies, especially assemblies with high gene counts because they may also harbor a significant portion of pseudogenes. Identification of the syntenic, ancestral gene pairs will also be important if attempting reference-based scaffolding across species.

The evolutionary advantage of increased genome sizes in currently reported assemblies of graminicolous downy mildews compared with other species remains unknown. One consequence of genome-inflation by retrotransposition in *P. sorghi* is the presence of highly similar genes encoding effectors on distinct chromosomes, which is not present in *P. effusa* (Fletcher *et al.* 2022). In *P. effusa*, gene duplications, including of effector-encoding genes, appear to have more often resulted in clusters of genes at similar chromosome positions (Fig. 9c). Local gene duplication is likely the result of unequal crossing over or non-homologous recombination, rather than retrotransposition (Hurles 2004). The coenocytic nature of oomycetes means that within their mycelia there may be rapid somatic evolution with the fittest nuclei proliferating and being preferentially represented in the next asexual or sexual generation. For *P. effusa*, this has resulted in an organism with a compact genome and tight clusters of effectors. For *P. sorghi* this has resulted in an organism with a genome bloated with repeats, putative pseudogenes, and widely dispersed effector genes (Figs. 3 and 5). Comparative genomics with other oomycetes, including downy mildews and *Phytophthora* spp., will determine whether there is a relationship between genome size and effector distribution.

The mitochondrial genome of *P. sorghi* is circular in orientation and at ~38.5 kb is similar in size to other oomycetes (Fig. 4). The mitogenome encoded the same common suite of genes observed in other oomycetes, including 35 genes, ribosomal RNAs, tRNAs, and the putative ORFs *ymf16, ymf98, ymf99, ymf100*, and *ymf101*. There was also an additional putative ORF (*orf161*) of unknown function encoded between *cox1* and *ymf100* that is unique to *P.*

sorghi. The presence of species-specific putative ORFs is a common phenomenon and was observed in all nine of the taxa illustrated in Fig. 4. The mitochondrial gene order between the nine Peronosporaceae taxa representing six genera was largely conserved with the exceptions of inversions of gene-blocks which retained internal gene order (Fig. 4). This finding parallels a previous study examining a broader array of taxa, in particular *Phytophthora* species (Winkworth *et al.* 2022). The mechanisms driving this type of genome evolution has yet to be determined. Recently, inverted repeats were reported as flanking a region of the mitogenome, which was inverted in some, but not all isolates of *P. effusa* (Skiadas *et al.* 2022). Recombination between the small 1,150 bp inverted repeats in *P. ramorum* generated isomers of the mitochondrial genome where the region between the repeats was also present in an inverted orientation (Martin 2008), but this did not lead to structural changes in the mitochondrial genome. However, some taxa, including *P. sorghi*, lack inverted repeats flanking inversions/translocations relative to other taxa (Fig. 4). Therefore, it is possible that other mechanisms may underlie these differences.

The genome of *P. sorghi* provides the molecular foundation to characterize taxonomic relationships within and between species of *Peronosclerospora* and to deploy diagnostic molecular markers. Establishing molecular diagnostics to detect and distinguish potential immigration of graminicolous downy mildews, including *P. sorghi* and other *Peronosclerospora* spp., such as *P. philippinensis*, is a priority to prevent epidemics on corn in the United States. This is desirable since US corn is potentially highly susceptible to tropical downy mildew causing pathogens (Duck *et al.* 1987; Perumal *et al.* 2008). Highly conserved markers may provide genus-level resolution suitable for screening at ports of entry. The mitogenome is likely to be a good source for molecular marker development given its unique configuration in *P. sorghi* compared to other Peronosporaceae species (Fig. 4) (Winkworth *et al.* 2022) and its high-copy number relative to the nuclear genome. Variation of the mitogenome within and between *Peronosclerospora* species should be assayed to confirm its suitability as a diagnostic marker. Single-copy protein-coding genes in the nucleus may provide additional sequences for conserved markers; however, high sequence conservation could result in false species assignment. The expanded paralogs and pseudogenes may provide opportunities for lineage-specific markers that are useful for taxonomic and population studies of *Peronosclerospora* spp. (Devos *et al.* 1995) because sequences under weaker purifying selection may accumulate unique polymorphisms quicker than sequences under stronger purifying selection. All marker types should be validated to determine their variation within and between species; consequently, additional studies of the mitochondrial and nuclear genome of *Peronosclerospora* spp., including global isolates of *P. sorghi*, are required to efficiently monitor these important pathogens.

Data availability

All raw reads and the nuclear genome assembly and annotation and the assembled transcriptome are available at NCBI under BioProject PRJNA845776. The mitochondrial genome and annotation is available at NCBI under accession OP873122.

Supplemental material available at G3 online.

Acknowledgments

We thank H. Xu (UC Davis) for raw data submissions to NCBI and E. Georgian (UC Davis) for editorial services. The sequencing was carried out by the DNA Technologies and Expression Analysis Cores at the UC Davis Genome Center, supported by NIH Shared Instrumentation Grant 1S10OD010786-01. The bioinformatic analysis was carried out using the UC Davis LSSC0 High Performance Computing cluster maintained by the UC Davis Bioinformatics Core.

Funding

K.F., F.M., and R.M. are grateful for support from USDA-APHIS award numbers (AP17PPQS&T00C153, AP18PPQS&T00C117, AP19PPQS&T00C200, AP20PPQS&T00C147, & AP21PPQS&T00C125).

Conflicts of interest

None declared.

Literature cited

Abdi H, Williams LJ. Tukey's honestly significant difference (HSD) test. Encycl Res Design. 2010;3(1):1–5. doi:10.4135/9781412961288.

Allen M, Poggiali D, Whitaker K, Marshall T, van Langen J, Kievit RA. Raincloud plots: a multi-platform tool for robust data visualization [version 2; peer review: 2 approved]. Wellcome Open Res. 2021;4:63. doi:10.12688/wellcomeopenres.15191.2.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–410. doi:10.1016/S0022-2836(05)80360-2.

Anaso AB, Emechebe AM, Tyagi PD, Manzo SK. Assessment of loss in yield due to sorghum downy mildew (*Peronosclerospora sorghi*) of maize in Nigerian Guinea savanna. Tropical Pest Management. 1989;35(3):301–303. doi:10.1080/09670878909371385.

Ayala-Usma DA, Cárdenas M, Guyot R, Mares MCD, Bernal A, Muñoz AR, Restrepo S. A whole genome duplication drives the genome evolution of *Phytophthora betacei*, a closely related species to phytophthora infestans. BMC Genomics. 2021;22(1):795. doi:10.1186/s12864-021-08079-y.

Bock CH, Jeger MJ, Mughoho LK, Cardwell KF, Adenle V, Mtisi E, Akpa AD, Kaula G, Mukasambina D, Blair-Myers C. Occurrence and distribution of *Peronosclerospora sorghi* [weston and uppal (shaw)] in selected countries of west and Southern Africa. Crop Prot. 1998;17(5):427–439. doi:10.1016/S0261-2194(98)00037-4.

Bonde M, Peterson G, Dowler W, May B. Isozyme analysis to differentiate species of *Peronosclerospora* causing downy mildews of maize. Phytopathology. 1984;74(11):1278–1283. doi:10.1094/Phyto-74-1278.

Bourret TB, Choudhury RA, Mehl HK, Blomquist CL, McRoberts N, Rizzo DM. Multiple origins of downy mildews and mito-nuclear discordance within the paraphyletic genus *Phytophthora*. PLoS One. 2018;13(3):e0192502. doi:10.1371/journal.pone.0192502.

Bushnell B. 2016. BBMap Short Read Aligner. University of California, Berkeley, California. URL http://sourceforge.net/projects/bbmap.

Cai G, Scofield SR. Mitochondrial genome sequence of *Phytophthora sansomeana* and comparative analysis of *Phytophthora* mitochondrial genomes. PLoS One. 2020;15(5):e0231296–e0231296. doi:10.1371/journal.pone.0231296.

Campbell MS, Holt C, Moore B, Yandell M. Genome annotation and curation using MAKER and MAKER-P. Curr Protoc Bioinform. 2014;48(1):4.11.11–14.11.39. doi:10.1002/0471250953.bi0411s48.

Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A, Yandell M. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008;18(1):188–196. doi:10.1101/gr.6743907.

Crouch J, Davis W, Shishkoff N, Castroagudín V, Martin F, Michelmore R, Thines M. Peronosporaceae species causing downy mildew diseases of Poaceae, including nomenclature revisions and diagnostic resources. Fungal Syst Evol. 2022;9(1):43–86. doi:10.3114/fuse.2022.09.05.

Derevnina L, Chin-Wo-Reyes S, Martin F, Wood K, Froenicke L, *et al.* Genome sequence and architecture of the tobacco downy mildew pathogen *Peronospora tabacina*. Mol Plant Microbe Interact. 2015; 28(11):1198–1215. doi:10.1094/MPMI-05-15-0112-R.

Derevnina L, Petre B, Kellner R, Dagdas YF, Sarowar MN, Giannakopoulou A, De la Concepcion JC, Chaparro-Garcia A, Pennington HG, van West P, *et al.* Emerging oomycete threats to plants and animals. Philos Trans R Soc Lond B Biol Sci. 2016;371-(1709):20150459. doi:10.1098/rstb.2015.0459.

Devos KM, Bryan GJ, Collins AJ, Stephenson P, Gale MD. Application of two microsatellite sequences in wheat storage proteins as molecular markers. Theor Appl Genet. 1995;90(2):247–252. doi:10.1007/BF00222209.

Dong S, Raffaele S, Kamoun S. The two-speed genomes of filamentous pathogens: waltz with plants. Curr Opin Genet Dev. 2015; 35:57–65. doi:10.1016/j.gde.2015.09.001.

Duck N, Bonde M, Peterson G, Bean G. Sporulation of *Peronosclerospora sorghi*, *P. sacchari*, and *P. philippinensis* on maize. Phytopathology. 1987;77(3):438–441. doi:10.1094/Phyto-77-438.

Eddy SR. Accelerated profile HMM searches. PLOS Comput Biol 2011; 7(10):e1002195. doi:10.1371/journal.pcbi.1002195

Emms DM, Kelly S. Orthofinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16(1):157. doi:10.1186/s13059-015-0721-2.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, *et al.* Pfam: the protein families database. Nucleic Acids Res. 2014;42(D1):D222–D230. doi:10.1093/nar/gkt1223.

Fletcher K, Gil J, Bertier LD, Kenefick A, Wood KJ, Zhang L, Reyes-Chin-Wo S, Cavanaugh K, Tsuchida C, Wong J, *et al.* Genomic signatures of heterokaryosis in the oomycete pathogen *Bremia lactucae*. Nat Commun. 2019;10(1):2645. doi:10.1038/s41467-019-10550-0.

Fletcher K, Klosterman SJ, Derevnina L, Martin F, Bertier LD, Koike S, Reyes-Chin-Wo S, Mou B, Michelmore R, *et al.* Comparative genomics of downy mildews reveals potential adaptations to biotrophy. BMC Genomics. 2018;19(1):851. doi:10.1186/s12864-018-5214-8.

Fletcher K, Michelmore R. From short reads to chromosome-scale genome assemblies. In: Ma W, Wolpert T editors. Plant Pathogenic Fungi and Oomycetes: Methods and Protocols. Springer New York, New York, NY; 2018. p. 151–197.

Fletcher K, Shin O-H, Clark KJ, Feng C, Putman AI, Correll JC, Klosterman SJ, Van Deynze A, Michelmore RW, *et al.* Ancestral chromosomes for the peronosporaceae inferred from a telomere-to-telomere genome assembly of *Peronospora effusa*. Mol Plant Microbe Interact. 2022;35(6):450–463. doi:10.1094/MPMI-09-21-0227-R.

Fletcher K, Zhang L, Gil J, Han R, Cavanaugh K, Michelmore R, *et al.* AFLAP: assembly-free linkage analysis pipeline using k-mers from genome sequencing data. Genome Biol. 2021;22(1):115. doi:10.1186/s13059-021-02326-x.

Frantzeskakis L, Kusch S, Panstruga R. The need for speed: compartmentalized genome evolution in filamentous phytopathogens. Mol Plant Pathol. 2019;20(1):3–7. doi:10.1111/mpp.12738.

Futrell M, Frederiksen R. Distribution of sorghum downy mildew (*Sclerospora sorghi*) in the U.S.A. Plant Dis Rep. 1970;54(4):311–314.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, *et al.* Trinity: reconstructing a full-length transcriptome without a genome from RNA-seq data. Nat Biotechnol. 2011;29(7):644–652. doi:10.1038/nbt.1883.

Graves S, Piepho H-P, Selzer ML. 2015. Package 'multcompView'. Visualizations of paired comparisons.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, *et al.* De novo transcript sequence reconstruction from RNA-seq: reference generation and analysis with Trinity. Nat Protocols. 2013;8(8):1494–1512. doi:10.1038/nprot.2013.1084.

Hurles M. Gene duplication: the genomic trade in spare parts. PLoS Biol. 2004;2(7):e206. doi:10.1371/journal.pbio.0020206.

Isakeit T, Jaster J. Texas Has a new pathotype of *Peronosclerospora sorghi*, the cause of Sorghum downy mildew. Plant Dis. 2005;89(5): 529. doi:10.1094/PD-89-0529A.

Isakeit T, Odvody G, Jahn R, Deconini L. *Peronosclerospora sorghi* resistant to metalaxyl treatment of sorghum seed in Texas. Phytopathology. 2003;93:S39.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C., McWilliam H, Maslen J, Mitchell A, Nuka G, *et al.* Interproscan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236–1240. doi:10.1093/bioinformatics/btu031.

Katoh K, Standley DM. MAFFT Multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–780. doi:10.1093/molbev/mst010.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, *et al.* Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28(12):1647–1649. doi:10.1093/bioinformatics/bts199.

Korf I. Gene finding in novel genomes. BMC Bioinform. 2004;5(1):59. doi:10.1186/1471-2105-5-59.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19(9):1639–1645. doi:10.1101/gr.092759.109.

Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–3100. doi:10.1093/bioinformatics/bty191.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–2079. doi:10.1093/bioinformatics/btp352.

Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25(5):955–964. doi:10.1093/nar/25.5.955.

Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics. 2017;33(4):574–576.

Martin FN. Mitochondrial haplotype determination in the oomycete plant pathogen *Phytophthora ramorum*. Curr Genet. 2008;54(1): 23–34. doi:10.1007/s00294-008-0196-8.

McCarthy CGP, Fitzpatrick DA. Phylogenomic reconstruction of the oomycete phylogeny derived from 37 genomes. mSphere. 2017; 2(2):e00095–17. doi:10.1128/mSphere.00095-17.

Micales J, Bonde M, Peterson G. Isozyme analysis and aminopeptidase activities within the genus *Peronosclerospora*. Phytopathology. 1988; 78(11):1396–1402. doi:10.1094/Phyto-78-1396.

Morales-Cruz A*et al.* Independent whole-genome duplications define the architecture of the genomes of the devastating west African Cacao black pod pathogen *Phytophthora megakarya* and its close relative *Phytophthora palmivora*. G3 (Bethesda). 2020; 10(7):2241–2255. doi:10.1534/g3.120.401014.

Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a web browser. BMC Bioinform. 2011;12(1):385. doi:10.1186/1471-2105-12-385.

Palti J, Cohen Y. Downy mildew of cucurbits (*Pseudoperonospora cubensis*): the fungus and its hosts, distribution, epidemiology and control. Phytoparasitica. 1980;8(2):109–147. doi:10.1007/BF02994506.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, *et al.* The *Sorghum bicolor* genome and the diversification of grasses. Nature. 2009;457(7229):551–556. doi:10.1038/nature07723.

Pawar MN. Pathogenic Variability and Sexuality in Peronosclerospora Sorghi (Weston and Uppal) Shaw, and Comparative Nuclear Cytology of Peronosclerospora species: Texas A&M University; 1986.

Perumal R, Nimmakayala P, Erattaimuthu SR, No E-G, Reddy UK, Prom LK, Odvody GN, Luster DG, Magill CW. Simple sequence repeat markers useful for sorghum downy mildew (*Peronosclerospora sorghi*) and related species. BMC Genet. 2008; 9(1):77–77. doi:10.1186/1471-2156-9-77.

Prom LK, Perumal R, Montes-Garcia N, Isakeit T, Odvody GN, Rooney WL, Little CR, Magill C. Evaluation of gambian and Malian sorghum germplasm against downy mildew pathogen, *Peronosclerospora sorghi*, in Mexico and the USA. J Gen Plant Pathol. 2015;81(1):24–31. doi:10.1007/s10327-014-0557-8.

Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 2016;26(3):342–350. doi:10.1101/gr.193474.115.

Quinlan AR. BEDTools: the Swiss-army tool for genome feature analysis. Curr Protoc Bioinform. 2014;47(1):11.12.11–11.12.34. doi:10.1002/0471250953.bi1112s47.

Radwan GL, Perumal R, Isakeit T, Magill CW, Prom LK, Little CR *et al.* Screening exotic Sorghum germplasm, hybrids, and elite lines for resistance to a new virulent pathotype (P6) of *Peronosclerospora sorghi* causing downy mildew. Plant Health Prog. 2011;12(1):17. doi:10.1094/PHP-2011-0323-01-RS.

Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, Habermann B, Akhtar A, Manke T. High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat Commun. 2018;9(1):189. doi:10.1038/s41467-017-02525-w.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47–e47. doi:10.1093/nar/gkv007.

Robinson MD, McCarthy DJ, Smyth GK. Edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England). 2010;26(1):139–140. doi:10.1093/bioinformatics/btp616.

Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19): 3210–3212. doi:10.1093/bioinformatics/btv351.

Skiadas P, Klein J, Quiroz-Monnens T, Elberse J, de Jonge R, Van den Ackerveken G, Seidl MF. Sexual reproduction contributes to the evolution of resistance-breaking isolates of the spinach pathogen *Peronospora effusa*. Environ Microbiol. 2022;24(3):1622–1637. doi:10.1111/1462-2920.15944.

Smit A, Hubley R. RepeatModeler Open-1.0 2008.

Smit A, Hubley R, Green P. RepeatMasker open-4.0 2013.

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9): 1312–1313. doi:10.1093/bioinformatics/btu033.

Suharjo R, Swibawa IG, Prasetyo J, Fitriana Y, Danaatmadja Y, Budiawan A, Roberts S, Noorhajati N, Amad M, Thines M. *Peronosclerospora australiensis* is a synonym of *P. maydis*, which is widespread on sumatra, and distinct from the most prevalent Java maize downy mildew pathogen. Mycol Prog. 2020;19(11): 1309–1315. doi:10.1007/s11557-020-01628-x.

Thakur R, Mathur K. Downy mildews of India. Crop Protection. 2002; 21(4):333–345. doi:10.1016/S0261-2194(01)00097-7.

Varoquaux N, Liachko I, Ay F, Burton JN, Shendure J, Dunham MJ, Vert J-P, Noble WS. Accurate identification of centromere locations in yeast genomes using hi-C. Nucleic Acids Res. 2015;43-(11):5331–5339. doi:10.1093/nar/gkv424.

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid genome sequences. Genome Res. 2017;27(5): 757–767. doi:10.1101/gr.214874.116.

Wickham H. ggplot2: Elegant Graphics for Data Analysis: Springer; 2016.

Winkworth RC, Neal G, Ogas RA, Nelson BCW, McLenachan PA, Bellgard SE, Lockhart PJ. Comparative analyses of complete peronosporaceae (oomycota) mitogenome sequences—insights into structural evolution and phylogeny. Genome Biol Evol. 2022; 14(4):evac049. doi:10.1093/gbe/evac049.

Wood DE, Lu J, Langmead B. Improved metagenomic analysis with kraken 2. Genome Biol. 2019;20(1):257. doi:10.1186/s13059-019-1891-0.

Yang Z. PAML 4: phylogenetic analysis by Maximum likelihood. Mol Biol Evol. 2007;24(8):1586–1591. doi:10.1093/molbev/msm088.

Zhang J. Evolution by gene duplication: an update. Trends Ecol Evol (Amst). 2003;18(6):292–298. doi: 10.1016/S0169-5347(03)00033-8.

Communicating editor: T. Jamann