Haplotype-phased and chromosome-level genome assembly of *Puccinia polysora*, a giga-scale fungal pathogen causing southern corn rust

Junmin Liang[1], Yuanjie Li [1,2], Peter N. Dodds[3], Melania Figueroa[3], Jana Sperschneider[3], Shiling Han[1,2], Clement K.M. Tsui[4,5], Keyu Zhang[6], Leifu Li[6], Zhanhong Ma[6], Lei Cai[1,2*]

[1] State Key Laboratory of Mycology, Institute of Microbiology, Chinese Academy of Sciences, Beijing, 100101, China

[2] College of Life Sciences, University of Chinese Academy of Sciences, Beijing, 100049, China

[3] Commonwealth Scientific and Industrial Research Organization, Agriculture and Food, Canberra, 10 ACT, Australia

[4] Division of Infectious Diseases, Faculty of Medicine, University of British Columbia, Vancouver, BC V6T 1Z3, Canada

[5] National Centre for Infectious Diseases, Tan Tock Seng Hospital, 308433, Singapore

[6] Department of Plant Pathology, China Agricultural University, Beijing, 100193, China

Correspondence: Lei Cai, State Key Laboratory of Mycology, Institute of Microbiology, Chinese Academy of Sciences, Beijing, 100101, China.

Email: cail@im.ac.cn

Abstract

Rust fungi are characterized by large genomes with high repeat content and have two haploid nuclei in most life stages, which makes achieving high-quality genome assemblies challenging. Here, we describe a pipeline using HiFi reads and Hi-C data to assemble a gigabase-sized fungal pathogen, *Puccinia polysora* f.sp. *zeae*, to haplotype-phased and chromosome-scale. The final assembled genome is 1.71 Gbp, with ~850 Mbp and 18 chromosomes in each haplotype, being currently the largest fungal genome assembled to chromosome scale. Transcript-based annotation identified 47,512 genes with a similar number for each haplotype. A high level of interhaplotype variation was found with 10% haplotype-specific BUSCO genes, 5.8 SNPs/kbp, and structural variation accounting for 3% of the genome size. The *P. polysora* genome displayed over 85% repeat content, with genome-size expansion, gene losses, and gene family expansions suggested by multiple copies of species-specific orthogroups. Interestingly, these features did not affect overall synteny with other *Puccinia* species with smaller genomes. Fine-time-point transcriptomics revealed seven clusters of co-expressed secreted proteins that are conserved between two haplotypes. The fact that candidate effectors interspersed with all genes indicated the absence of a "two-speed genome" evolution in *P. polysora*. Genome resequencing

of 79 additional isolates revealed a clonal population structure of *P. polysora* in China with low geographic differentiation. Nevertheless, a minor population drifted from the major population by having mutations on secreted proteins including *AvrRppC*, indicating the ongoing evolution and population differentiation. The high-quality assembly provides valuable genomic resources for future studies on the evolution of *P. polysora*.

Keywords: rust fungi, *de novo* genome assembly, phasing, population genomics, *AvrRppC*

1 | INTRODUCTION

Rust fungi (Pucciniales) constitute one of the largest orders in the kingdom of fungi, with more than 8,000 species grouped into 18 families and approximately 170 accepted genera (Zhao et al. 2020, 2021; Aime and McTaggart 2021). Rust fungi are obligate biotrophs, which are usually recalcitrant to in vitro culturing and show host-specificity to particular species, genera, or families of vascular plants. Their life cycles are diverse and complex, including up to five types of spores (spermatia, aeciospores, urediniospores, teliospores, and basidiospores) (Zhao et al. 2021). Spores of the most abundant life stage are dikaryon, which contain two physically separated and genetically different haploid nuclei. The Pucciniales contains fungal species with the largest known genomes, with an average haplotype genome size of 380 Mbp estimated by flow cytometry, far larger than that of all fungi (37.7 Mbp) (Tavares et al. 2014). Based on flow cytometry, the largest known rust genome (*Uromyces bidentis*) was estimated up to 2.4 Gbp (Ramos et al. 2015). The repeat contents of rust genomes range from 30% to 91% (Amie et al. 2017; Tobias et al. 2021). Their dikaryotic nature and highly repetitive sequences pose substantial challenges for high-quality genome assembly as well as in-depth understanding of evolution within the rust group.

In addition to above intriguing biological features, rust fungi have also received wide attention for causing major diseases on agricultural and forest crops worldwide. The most devastating rust pathogen include *Puccinia striiformis*, *P. graminis* and *P. triticina*, causing three of the world's most serious wheat rust diseases (Kolmer et al. 2005); *Puccinia sorghi* and *P. polysora,* also seriously threaten global crop production by causing common and southern corn rust of maize (Crouch and Szabo 2011; Ramirez-Cabral et al. 2017); and *Melampsora larici-populina* and *Austropuccinia psidii*, two notorious forest pathogens leading to myrtle rust and

poplar rust, respectively (Pinon and Frey 1997; Winzer et al. 2019). Effective strategies for management of rust disease rely heavily on host resistance breeding programmes. In host-rust interactions, rust fungi deliver specific effectors (avirulence genes, *Avr*) into host cells, which can be recognized by resistance proteins from resistant plants and trigger defense responses (Dodds and Rathjen 2010). However, rust pathogens can acquire mutations in the *Avr* gene to avoid this recognition and facilitate the pathogen infection (Cui et al. 2015). For instance, the Ug99 race (TTKSK) of *P. graminis* evolved virulence to the widely deployed *Sr31* resistance gene in wheat has become a big threat to global wheat production (Singh et al. 2011). Therefore, understanding the molecular mechanism of *Avr*-*R*, or gene-for-gene relationship, is critical for durable disease control. Given their biotrophic lifestyle, the genetic transformation of rust fungi is difficult and challenging. Nevertheless, genome sequencing data of rust species has enabled the prediction of candidate effectors which facilitate the identification of *Avr* genes (Figueroa et al. 2016; Anderson et al. 2016; Maia et al. 2017; Salcedo et al. 2017; Chen et al. 2017; Miller et al. 2018; 2020; Upadhyaya et al 2021).

Early rust genome assemblies were haploid representations and different homologous haplotypes were collapsed into a consensus assembly, which did not fully capture sequences from both nuclei. Using single-molecule real-time (SMRT) sequencing technology and haplotype resolution software, FALCON-Unzip (Chin et al. 2016), allowed higher contiguity and partially haplotype-phased genomes of a few *Puccinia* species (Miller et al. 2018; Schwessinger et al. 2018, Wu et al. 2021). Other softwares have been developed to obtain sub-assemblies from diploid assemblies, such as HaploMerger2 (Huang et al. 2017) and Purge_Haplotigs (Roach et al. 2018), but many duplicate contigs cannot be correctly switched leading to mis-joins in the final sub-assemblies. A recent pipeline, NuclearPhaser, was designed to phase two haplotypes from the diploid genome using Hi-C data (Duan et al. 2022). Three *Puccinia* species, causing wheat or oat rust have been assembled to a fully-phased and chromosome level using NuclearPhaser (Li et al. 2019; Duan et al. 2022; Henningsen et al. 2022). Unlike wheat-like rusts, which have been studied extensively, the genome information of corn rust pathogens is largely unknown. In changed global cereal settings, maize has become the world's most productive food crop (FAO, https://www.fao.org/faostat/en/#data/QCL), meanwhile, the distribution of the two major rust diseases (common corn rust and southern corn rust) in corn is projected to expand to temperate regions with increasing global temperatures (Ramirez-Cabral et al. 2017), which poses a greater

threat to global food safety. A draft genome sequence of common corn rust (*P. sorghi*) was released in 2016, although it was highly fragmented (Rochi et al. 2016), however, genomic information of southern corn rust pathogens has not been reported, which seriously hampers avirulence gene identification and breeding of disease resistance. The genome sizes of recent phased *Puccinia* species range from ~170 Mbp to 250 Mbp (Li et al. 2019; Miller et al. 2018; Wu et al. 2021), whereas the genome of *P. polysora* was estimated over gigabase in our preliminary test (Figure S1), which poses new challenges for genome assembly.

In this study, we assembled the genome of *P. polysora* to haplotype-resolved and chromosome-scale levels using a modified haplotype-phasing assembly pipeline. To our knowledge, this is the largest fungal genome as well as the first giga-scale fungal genome ever assembled to complete-chromosome (telomere to telomere) level. The *P. polysora* genome was annotated by fine-time-point gene expression data from germinated spores and infected issues, which supplied robust data for coexpression analyses of secreted proteins and prediction of candidate secreted effectors. Also, we investigated the genetic divergence and population structure of *P. polysora* using the genome resequencing of 79 additional samples from different parts of China. The high-quality genome information represents a valuable resource for better understanding genome evolution, host adaption, and genes involved in host-pathogen interactions.

2 | MATERIALS AND METHODS

2.1 | Fungal isolates and plant inoculation

The isolate, GD1913, collected from Guangdong province in the 2019 annual rust survey, was selected for reference genome sequencing. A total of 79 isolates, representing three possible populations of *P. polysora* from the south, central, and north China were selected and used for genome resequencing (Table S1). All isolates were purified by selecting a single pustule from infected leaves and amplified by 2–3 rounds of infection on highly susceptible variety Zhengdan 958. Corn seeds immersed in Chlormequat chloride (1.5 g/L) were sown in 10-cm wide square pots 10 days before inoculation. When 10-cm tall, seedling growth was reduced by adding 15 mL maleic hydrazide acid (1.5 g/L) per pot. Urediniospores were amplified by spraying on 10-day-old corn seedlings. The inoculated seedlings were incubated with dew at 27℃ in the dark for 24h before transferring to a climate-controlled chamber (12h, 25℃ dark period and a 12h, 27℃ light period). To prevent airborne contamination, 75% ethyl alcohol was sprayed on the area

after each inoculation and each pot was protected by a cellophane bag. Once the pots were heavily infected (ca. 15–20 days), spores were harvested to clean cellophane by scraping them with sterile needles. For long-term storage, fresh spores were dried (10% relative humidity) in a desiccator for 1 day at 4℃ and then maintained at -80℃ or in liquid nitrogen.

2.2 | DNA extraction, library preparation, and sequencing

About 500 mg of fresh urediniospores were ground for 4-5 batches in liquid nitrogen and genomic DNA was extracted using the lysis buffer and cetyltrimethylammonium bromide (CTAB) method (Justesen et al. 2002). The integrity of genomic DNA was assessed by Agilent 4200 Bioanalyzer (Agilent Technologies, Palo Alto, California, USA). The average insert size of 15 kb PacBio library was concentrated with AMPure PB magnetic beads (Pacific Biosciences, California, USA) following the manufacture's instruction. Sequencing of high fidelity (HiFi) long reads was carried out by the Pacific Bioscience Sequel II platform. For genome resequencing, ~10 mg fresh urediniospores were placed in 2 mL screw-gap tubes filled with Lysing Matrix C and ground twice in MP FastPrep-24 TM 5G (Mp Biomedicals, USA) with a speed setting of 4 for 20 seconds. The paired-end library with 150 bp was prepared and sequenced using Illumina Novaseq 6000. All sequencing was performed at Annoroad Gene Technology Co., Ltd, (Beijing, China).

To generate a chromosome-level and haplotype phased assembly of *P. polysora* genome, a Hi-C library was generated following in situ ligation protocols. In brief, fresh urediniospores (~ 100 mg) of GD1913 were used for crosslinking reaction by 2% formaldehyde at room temperature for 15 min. After Glycine quenching, the supernatant was removed and spores were then ground with liquid nitrogen for DNA extraction. The purified DNA was digested with *Mbo* I restriction enzyme (New England Biolabs Inc. Beijing, China) and was labeled by incubating with Biotin-14-dATP (Thermo Fisher Scientific, Massachusetts, USA) and then ligated by T4 DNA Ligase (Thermo Fisher Scientific, Massachusetts, USA). After incubating overnight to reverse crosslinks, the ligated DNA was sheared into ~350 bp fragments. Finally, the Hi-C library was quantified by Bioanalyzer and sequenced on the Illumina NovaSeq 6000 platform using paired-end 150 cycles.

2.3 | *De novo* assembly, haplotype phasing and Hi-C scaffolding

HiFi reads were assembled by Canu 2.1.1 with -pacbio-hifi (Nurk et al. 2020). The coverage depth was calculated by genomeCoverageBed in BEDtools (v.2.29.2) (Quinlan and Hall 2010)

and small contigs (< 20 kbp) with low coverage (< 2×) were excluded from the further assembly. The remaining contigs were examined by BLASTN search (v.2.7.2) against the NCBI nt/nr database (downloaded on May 20, 2021) with E-value set as 1e-10. Contigs with significant matches to mitochondrial, plant rDNA, or chloroplast sequences were discarded in the final assembly.

To obtain a haplotype-phased assembly, HaploMerger2 (Huang et al. 2017), a tool to rebuild both haploid sub-assemblies from the high-heterozygosity diploid genome, was used. The heterozygosity evaluated by Jellyfish v. 2.1.3 (Marcai and Kingsford 2011) and Genomescope.R (Vurture et al. 2017) was 1.08% leading to the identity setting as 95% (Huang et al. 2017). To better distinguish allelic and non-allelic combinations, the scoring scheme for alignments was recalculated by a Perl script (lastz_D_Wrapper.pl) that came with HaploMerger2. The top 22 (in length) contigs accounting for ~10% length of all contigs were assigned to part1.fasta and others were in part2.fasta. Mis-joins were detected by three rounds and all breaks were manually checked combined with Hi-C validation. Based on all-vs-all alignment (hm.new_scaffolds), contigs were assigned separately to haplotype A and haplotype B.
For scaffolding, the raw Hi-C data were trimmed by removing adapters and low-quality bases, and then two-paired reads were separately mapped to two haplotypes independently using BWA-MEM v.0.7.8 (Li and Durbin 2010). To move experimental artifacts, the alignments went through the mapping workflow from the Arima Genomics pipeline (https://github.com/ArimaGenomics/mapping_pipeline/blob/master/01_mapping_arima.sh). Then SALSA v.2.2 (Ghurye et al. 2017) was run to cluster initial contigs into groups. SALSA scaffolding was performed independently on haplotype A and B contigs. Hi-C contact matrices for each scaffold were calculated by HiC-Pro (Servant et al. 2015) and contact maps are visualized by HiCPlotter (Akdemir and Chin 2015). Contig reversal and breaking of mis-joins were corrected manually according to Hi-C contact map and global alignment resulted from HaploMerger2 (hm.new_scaffold). Hicexplorer v.3.4.1 (Wolff et al. 2018)) was used to calculate Hi-C links of read pairs to each haplotype. To guide scaffolds to chromosomes, telomeres were identified by a custom Perl script. The possible tandem repeats, (CCTAAA/TTAGGG)n in most filamentous ascomycete fungi (Lue 2021), and other irregular type reported in closely related rust fungi, (CCCTAA/TTTAGG)n, (CCCCTAA/TTAGGGG)n, (CCCTAAA/TTTAGGG)n, (TTAGGG/CCCTAA)n, were all tested (Li et al. 2019; Tobias et al. 2021). Finally

(CCCTAAA/TTTAGGG)n was identified as the telomere repeat units for *P. polysora*. Then corrected scaffolds were rebuilt to chromosome level by a Perl script. The workflow of chromosome assembly has been illustrated in Figure 1A. The completeness of each haplotype was evaluated using BUSCOs of basidiomycota_odb9 and *Ustilago maydis* as the selected species for AUGUSTUS gene prediction (Stanke and Morgenstern 2005) in BUSCO v. 3.0.2 (Simão et al. 2015).

2.4 | RNA extraction and sequencing

For gene annotation, transcriptome sequencing of *P. polysora* from various infection time points was performed. Corn seedlings of 18 pots (10 seedlings/pot) were inoculated with 10 mg spores/mL mixed of Tween 20 (v/v: 0.05%). About five infected leaves were cut off per biological replicate at 1 day post inoculation (dpi), 2 dpi, 4 dpi, 7dpi, 10 dpi, and 14 dpi. In addition, germinated spore samples were prepared by placing 30 mg of fresh urediniospores on the surface of sterile water at 25–28°C in dark for12 hours. Three biological replicates were performed for each condition. All samples were collected in 4 mL Eppendorf tubes, frozen in liquid nitrogen before storing at -80°C. Before RNA extraction, samples were ground in liquid nitrogen and RNA was extracted using the RNeasy Plant Minikit (Qiagen) according to the manufactuer's protocols (https://www.qiagen.com/us/resources/download.aspx?id=246847e7-0095-43e4-8d1d-41df3f9153dd&lang=en). After checking the RNA quality by Bioanalyzer, ~350 bp library was constructed and sequenced by Illumina Novaseq 6000 platform (Annoroad Gene Technology Co., Ltd, Beijing, China). Before transcriptome assembly, raw reads were trimmed using Trimmomatic v.0.36 (Bolger et al. 2014) with the settings: ILLUMINICLIP 2:30:10 LEADING 3, TRAILING 3 SLIDINGWINDOW 4:10 and MINLEN 50. A reference-guided transcriptome assembly was performed in Trinity v.2.8.5 (Grabherr et al. 2011) using combined reads from germinated spores and infected leaves.

2.5 | Gene prediction and genome annotation

The two haplotypes were annotated independently using Funannotate v1.8.7 (https://github.com/nextgenusfs/funannotate/releases/tag/v1.8.7). A pipeline of core modules was applied as mask-> train->predict->update-> fix->annotate. The repeats of the assembled genome were soft-masked according to a merged library including a self library in RepeatMasker v.4.0.8 (Smit et al. 2015) and a genome-trained library from RepeatModeler v.1.0.11 (Smit and Hubley 2008). The retrotransposons with long terminal repeats (LTR-RTs) were annotated with an in-

house pipeline that uses LTRharvest (Ellinghaus et al. 2008) packaged in LTR_retriever v2.9.0 (Ou and Jiang 2018). LTRharvest was also used to estimate the insert time of LTR-RTs according to the mutation rate of 2.0×10^{-8} of a closely related fungus, *Schizophyllum commune* (Baranova et al. 2015). The prediction step (funannotate predict) was run with transcript evidence from HISAT2 (Kim et al. 2015) RNA-seq alignments and genome-guided Trinity assemblies. Transcript evidence was aligned to two haplotypes separately using Minimap2 (Li 2018) and the protein evidence was aligned to the genome via Diamond (Buchfink et al. 2015)/Exonerate (Slater and Birney 2005) with the default UniProtKb/SwissProt protein database (http://legacy.uniprot.org/uniprot/?query=reviewed%3Ayes) from funannotate. The PASA gene models were parsed to train AUGUSTUS v3.2.3 (Stanke and Morgenstern 2005), snap and GlimmerHMM. In addition, GeneMark_ES (Lomsadze et al. 2005) was self-trained using two haplotypes' sequences. All above evidence was combined with default weight settings using Evidence Modeler (Haas et al. 2008) and filtered by removing genes with short length (< 50 aa), spanning gaps and transposable elements. The tRNA genes were predicted using tRNAscan-SE v1.3.1 (Lowe and Chan 2016). Funannotate update command to add UTR data to the predictions and fix gene models that are in disagreement with the RNA-seq data. Functional annotation was performed based on available databases including Pfam v. 34.0) (Finn et al. 2014), InterPro (v. 86.0) (Jones et al. 2014), eggNOG (v5.0) (Huerta-Cepas et al. 2019), UniProtKB (v. 2021_03), MEROPS (v. 12.3) (Rawlings et al. 2016), carbohydrate hydrolyzing enzymatic domains (CAZymes) (Terrapon et al. 2017) and a set of transcription factors based on InterProScan domains to assign functional annotations.

2.6 | Interhaplotype variation analysis

Small variants including SNPs and indels were identified by mapping trimmed short reads from DNA against haplotype A with BWA-MEM v0.7.8 (Li and Durbin 2010). After removing PCR duplicates using Picard v.2.18.27 (https://github.com/broadinstitute/picard), the bam file was input to GATK 4.1.9 (https://github.com/broadinstitute/gatk) to call SNPs. SNPs and Indels were filtered using gatk VariantFiltration with parameters QD > 2.0 & MQ > 40.0 & FS < 50.0 & SQR < 3.0 for SNPs and QD > 2.0 & QUAL > 30.0 & FS < 200.0 & ReadPosRankSum>-20.0. Variants were annotated from genome location and functional impact using SnpEff v4.3 (Cingolani et al. 2012). To estimate structure variations (SVs) between two haplotigs, we aligned 18 chromosomes of haplotype A to their corresponding chromosomes in haplotype B by using

the nucmer program in Mummer 4 (--maxmatch -t 100 -l 100 -c 500) (Marçais et al. 2018). The chromosome-pair alignments were then analyzed using Assemblytics (http://assemblytics.com/) with the unique sequence length=10,000, maximum variant size=10,000 and minimum variant size=1. To identify the orthologous genes between two haplotypes, we used OrthoFinder v2.5.4 with the default parameters (Emms and Kelly 2015).

2.7 | Prediction of candidate effectors

Secreted proteins were predicted based on two rules 1) the presence of a predicted signal peptide using SignalP 4.0 (Petersen et al. 2011); and 2) the absence of predicted transmembrane domains outside the first 60 amino acids (with TMHMM 2.0). To predict candidate effectors, we use EffectorP 3.0 (Sperschneider and Dodds 2021), which applied two machine learning models trained on apoplastic and cytoplasmic effectors. The density plots for genes, repeats, and secreted proteins from chromosomes in two haplotypes were generated using KaryoploteR (Gel and Serra 2017). A recent study identified the first effector of *P. polysora*, *AvrRppC* (Deng et al. 2022). To locate its genome position, we blasted the CDS sequence of $AvrRppC^{ref}$ against genome sequences of two haplotypes in this study.

To understand the expression pattern of secreted proteins, we used all expression data from germinated spores and six timepoints to perform a differential expression analysis. RNA-seq reads were mapped to haplotype A by HISAT2 v2.2.1 (Kim et al. 2015) and FeatureCounts v1.5.3 (Liao et al. 2014) was used to generate read counts for each gene model of secreted protein. Differentially expressed genes were identified by expression in plants relative to germinated spores ($|\log \text{fold change}| > 1.5$; adjusted $P < 0.1$) using the DESeq2 R package (Love et al. 2014). The average rlog-transformed values for each gene were used for clustering using the *k*-means method. The optimal number of clusters was defined using the elbow plot method and circular heatmaps were plotted using the Circlize R package (Gu et al. 2014).

2.8 | Comparative genome analysis

The high quality of *P. polysora* genome provides the opportunity for comparison with other chromosome-level references, such as wheat stem rust (*Pgt*21-0), wheat left rust (*Pt*76) and oat crown rust (*Pca*203) (Li et al. 2019; Duan et al. 2022; Henningsen et al. 2022). Syntenic gene pairs among four *Puccinia* species were identified using the MCSCAN toolkit (Wang et al. 2012). Figures were plotted using MCscan (Python version). (https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version). Orthofinder v2.5.4 (Emms

and Kelly 2015) was applied to compare the orthogroups and the proteome between *P. polysora* and its close rust species with available high-quality genome.

2.9 | Population genetic analyses

To understand the population differentiation of *P. polysora* in China, 79 isolates were resequenced. Reads were qualified by Trimmomatic v.0.36 (Bolger et al. 2014) using the parameters described in the RNA extraction and sequencing paragraph. To assess whether each isolate comprises a single genotype free of contamination, the distribution of read counts for bi-allelic SNPs was calculated by R package vcfR v.1.8.0 (Knaus and Grunwald 2017) and plotted using ggplot2 (v2.2.1) (Wickham 2016). The normal distribution of read allele frequencies at heterozygous positions is expected to rule out contamination from other *Ppz* genotypes and only the pure isolates were applied in the following population genomic analyses. SNPs calling and filtration were performed by GATK v4.1.9 as described in the interhaplotype variation analysis section. Population structure was detected and quantified using principal component analysis (PCA), which was performed using Plink v.1.9 (Purcell et al. 2007). Pairwise FST values were estimated between geographical regions using the PopGenome R package (v2.6.1) (Pfeifer et al. 2014). To assess the effects of variants, the vcf file including selected SNPs was run in SnpEff v4.3 (Cingolani et al. 2012). Because the sexual stage of *P. polysora* has not been reported, we used the standardized index of association (*rd*) to test linkage disequilibrium in the Chinese *Ppz* population. We constructed 100 sets of 10,000 random SNPs by samp.ia function from R package *poppr* (Kamvar et al. 2014) to generate a distribution of *rd* values. The observed *rd* distribution of the Chinese population was compared to the distribution of 10,000 *rd* values constructed using fully randomly simulated datasets with 0%, 50%, 75% and 100% linkage.

3 | RESULTS

3.1 | Chromosome-level genome assembly and haplotype-phasing of *P. polysora*

We generated a total of 53 Gbp of circular consensus reads (32× coverage) from single-molecule real-time sequences on the PacBio Sequel II. A total of 4056 contigs were assembled with a genome size of 1.76 Gbp and a contig N50 of 1.9 Mbp (Table 1). A total of 1627 contigs, accounting for ~2% of the *de novo* assembly size were excluded due to small size, low coverage or high mitochondrial similarity (Table S2). By using Haplomerger2, the remaining contigs were assigned to two haplotypes, A and B, with 613 contigs representing ~862 Mbp and 1321 contigs representing ~852 Mbp, respectively. These contigs were further connected to 173 and 344

scaffolds in haplotypes A and B by Hi-C scaffolding (Table 1), after manual checking of 33 duplicated scaffolds between (17) and within (16) haplotypes (Table S3). By considering both the Hi-C contact maps and the all-vs-all alignment from Haplomerger2, 18 super scaffolds were binned for each haplotype manually. The Hi-C contact map showed evidence of a single centromere on each scaffold and all scaffolds contained telomere sequences at either end (Table 1, Figures 1B, 1C), indicating that each haplotype contains 18 chromosomes. This is consistent with the observations of 18 chromosomes in the genome of other *Puccinia* species (Boehm et al. 1992; Li et al. 2019; Wu et al. 2021; Duan et al. 2022). In addition, 15 (6.5 Mbp)) and 20 (4.5 Mbp) small scaffolds in haplotype A and B respectively could not be assigned to any chromosome. The 18 chromosomes of haplotype A showed an average of over 80% Hi-C links to haplotype A, suggesting high levels of nuclear phasing (Figure 1D). The chromosomes were numbered according to the synteny with the other *Puccinia* species (see below).

3.2 | Assessment of repetitive DNA content

The completeness of the *Ppz* assembly was assessed based on highly conserved Basidiomycete genes (1335 BUSCOs), which suggested that the two haplotypes showed a similar level of completeness, with approximately 90% complete BUSCO genes and an additional 3.5% fragmented BUSCOs (Table 1). However, 126 BUSCO genes were complete in haplotype A but not present in haplotype B. On the contrary, 127 complete BUSCO genes are only present in haplotype B, which suggests interhaplotype variation in *P. polysora*. Considering the two haplotypes in combination, only 11 BUSCOs were missing or fragmented, resulting in 99% complete BUSCOs. That means we assembled a near-complete genome of *P. poplysora*.

Using the *de novo* library of RepeatMasker and genome-trained library of RepeatModeler, RepeatMasker detected repeats accounting for 85% of the diploid assembly (both haplotypes), which is significantly higher than observed in closely related species of *Puccinia* (30%–59%) (Miller et al. 2018; Schwessinger et al. 2018; Wu et al. 2021), but similar to *Austropuccinia psidii* (85% estimated in this study but 91% in its original description) (Tobias et al. 2021). The repeat density and GC content along the two haplotypes were illustrated in Figure 2A. The large percentage of repeats in *P. polysora* was mainly caused by Class I retrotransposons with long terminal repeats (LTR), with the LTR-Gypsy superfamily most abundant (33.5%), and LTR-Copia next (17.9%). The TE family composition was similar to other rust species, with the exception of *A. psidii*, which is LTR-Gypsy dominated with LTR-Copia accounting for only

1.9% (Figure 2B). In addition, about 20% and 16% of repeats of *P. polysora* in two haplotypes were unclassified repeat families based on Repbase (Bao et al. 2015).

Since LTR-RTs occupy the majority of the *P. polysora* genome, we examined whether these repetitive sequences slowly accumulated over time or alternatively were subject to sudden expansion in the life history of *P. polysora*. A total of 41,634 LTR-RT pairs were extracted. An LTR-RT burst was estimated at around 1.7 Mya (Figure 2C).

3.3 | High levels of interhaplotype gene content and structural variation

RNA-seq reads from germinated spores and infected corn leaves at 1, 2, 4, 7, 10, and 14 dpi (Table S4) were pooled and used to generate genome-guided transcriptome assemblies. In total, we annotated 23,270 and 24,242 gene models on haplotype A and haplotype B, respectively (Table 2). However, the gene space accounts for only ~4% of the total genome size. The two haplotypes showed a similar level of functional annotation with about 51% of proteins having at least one functional annotation. Orthofinder identified a total of 13,057 common orthogroups between the two haplotypes, involving 20,802 genes from haplotype A and 21,519 genes from haplotype B, over 90% in each haplotype. In addition, 198 and 235 orthogroups were specific to haplotype A or B respectively. About 70% of haplotype-specific genes have no functional annotations, with the remainder having annotated functions mainly in RNA mediated transposition, transmembrane amino acid transposition or ATP binding (Table S5).

By mapping Illumina reads to only haplotype A, we detected a total heterozygous rate of 6.1 variants/kbp, in which the SNP variant rate is dominated (5.8 SNPs/kbp). About 78% of the SNPs were located in intergenic regions and the rest (22%) were detected in 22,802 protein-coding genes. The results of Assemblytics suggested that structural variation between haplotype A and B comprised about 3.0% (26/849 Mbp) of the haploid genome size (Figure 3A). The full chromosome-pair alignment between two haplotypes is illustrated in Figure 3B, with alignments of chromsome09 and chromsome15 in Figure 3C and D), showing some insertions/deletions and inversions. Among three types of variation, insertions/deletions and repeat expansions/contractions are more prevalent than tandem expansions/contractions. Besides, variation with size bins of 500 to 10,000 bp is the most prevalent, which accounts for 2.8%. Because the maximal variant size was restricted to 10,000 in Assemblytic, the actual difference between the two haplotypes could be even higher than estimated.

3.4 | Prediction of secretome and candidate effectors.

We predicted 1183 and 1251 secreted proteins on haplotype A and haplotype B, respectively. By using the machine-learning tool, EffectorP3.0, about 14% and 27% of all secreted proteins were predicted as apoplastic and cytoplasmic effectors, respectively (Table 2). GO enrichment analysis and Interproscan annotation predicted some effectors involved in hydrolase activity, catalytic activity, protein tyrosine phosphatase activity and metal ion binding, but most effectors had no homology with known or predicted functions (Table S6, Figure S2). Then we used gene expression data to predict clusters in the secretome that are differentially expressed during infection. By using *k*-means clustering, seven clusters of genes with different expression profiles were detected in haplotype A (Figure 4A). Genes in cluster 1 showed high expression in germinated urediniospores and early infection (1dpi, 2dpi) but low expression at 4dpi and 7dpi when haustoria form. On the contrary, genes in clusters 2, 3, 4 and 5 showed low expression in germinated spores, and highest in planta expression at days 2-4 (cluster 2), days 2-7 (cluster 3) and from day 7 (cluster 4 and 5). Clusters 6 and 7 were more uniform in expression through these stages, although cluster 6 genes increased later in infection. About 33% to 80% of the secreted proteins in these clusters were predicted as candidate effectors by EffectorP (Table S7). A similar set of expression profile clusters were also detected for genes in haplotype B (Figure 4A). Nevertheless, two haplotypes presented different levels of carbohydrate-active enzymes (CAZymes). A total of 179 and 289 CAZymes were detected, of which 8 vs 77 CAZymes were predicted to be secreted in haplotypes A and B, respectively (Table S7). Among CAZymes subclasses, Glycoside hydrolase (GH) enzymes are abundant, accounting for 48% (88 GH families) and 42% (126 GH families) of all CAZymes of haplotype A and haplotype B, respectively. Of these, 51 GH families were predicted to be secreted. The GH5 (cellulase and other diverse forms being exo-/endo-glucanases and endomannannases) family (Langsto et al. 2011) was observed to be largely expanded in *P. polysora* as well as other *Puccinia* species (Figure 5A). Besides, AA3 (glucose-methanol-choline oxidoreductases), CE4 (chitin and peptidoglycan deacetylases), GH18 (chitinases of classes III and V), CH47(α-mannosidases), GT2 and GT90 are also abundant in *P. polysora*.

We also investigated other factors potentially associated with infection or plant immune inhibition, such as transcription factors (TFs) and peptidases. We found most TF families of *P. polysora* have low abundance except C2H2-type (IPR013087) and CCHC-type (IRP001878) zinc finger class, which have 87 and 67 members in haplotypes A and B (Figure 4B). In total, we

annotated 286 and 304 proteases in haplotypes A and B, with A01A (aspartic proteases), C19 (ubiquitinyl hydrolases), C26 (gamma-glutamyl hydrolase), S08A (subtilisin-like serine protease), S09X (glutamyl endopeptidase C), S10 (carboxypeptidase Y), S33 (prolyl aminopeptidase) and T01A (component peptidases of the proteasome) families expanded in *P. polysora* as well as in other three *Puccinia* species (Figure 4C). A total of 8 (2.8%) and 30 (10.0%) proteases were predicted to be secreted and these proteases showed a discrete distribution in each cluster without an obvious clustering pattern (Table S7).

3.5 | No evidence to support the "two-speed genome" in *P. polysora*

Analysis of the local gene density, measured as flanking distances between neighboring genes showed that the flanking distances in the *P. polysora* genome are generally rather high, with an average distance between genes of 100 kbp (Figure 4B). Accordingly, the surrounding genomic context of most genes in *P. polysora* genome is gene-sparse and repeat-rich. Large flanking distances are not specific to candidate effectors. In line with this pattern, the gene distance density plots revealed very similar distributions between all genes and candidate effectors, including the only known *Avr* effector gene, *AvrRppC* (Figure 4B). We further investigated whether candidate effectors present a different distribution of gene distance density compared to basidiomycete core ortholog genes of *P. polysora*. The results highlighted that candidate effectors are not located in peculiar gene-sparse areas, and the flanking distance centers of both BUSCOs and candidate effectors overlap with that of all genes (Figure 4B).

3.6 | Anchoring the *AvrRppC* in *Ppz*-GD1913

A previous study identified the avirulence effector, *AvrRppC*, and found six allelic variants which were named as $AvrRppC^{ref}$, $AvrRppC^{A}$, $AvrRppC^{C}$, $AvrRppC^{E}$, $AvrRppC^{F}$, $AvrRppC^{J}$ (Deng et al. 2022). By searching the CDS sequence of $AvrRppC^{ref}$ against haplotypes A and B we anchored *AvrRppC* at ~ 9 Mbp on chromosome 14. Although the gene was not present in the original annotation, there was evidence of transcription from RNA-seq reads at 1 dpi and 7 dpi (Figure 4C). Additionally, we found the *AvrRppC* allele on Chromosome14B is $AvrRppC^{ref}$ type whereas the allele in Chromosome 14A represents a new allele type, named $AvrRppC^{1}$ which has two amino acids changes compared with $AvrRppC^{ref}$ and is different from the other five previously reported variants.

3.7 | Comparative genomic analysis

Although *P. polysora* has a 7–10 times larger genome size than other rust species, the gene synteny amongst four *Puccinia* species with full chromosome assemblies is well conserved, with 18 chromosomes corresponding to each other clearly without any chromosome rearrangement (Figure 6A). A few large inversion blocks were detected in chromosomes 1, 3, 4, 7, 8, 9 and 10 between *Ppz*A and *Pt*76 (Figures 6A, S3). Inversion blocks were also detected in chromosomes 5, 7, 10, 13 between *Pt*76 and *Pca*203, and chromosomes 2, 3, 5, 7, 8, 10 and 13 between *Pgt*21-0 and *Pca*203 (Figures 6A, S3).

The protein orthologs among four closely related *Puccinia* species were compared based on haplotype A of each species. It seems that *Puccinia polysora* experienced gene expansion events compared with the other three species (Figure 6B, Table S8). Among 11,840 orthogroups identified from the four species, genes of *P. polysora* were assigned to 7,148 orthogroups which was less than those of over 8,000 orthogroups for the other three species. Nevertheless, *P. polysora* presented more species-specific orthogroups and these orthogroups are featured with a high copy number of gene duplication (Figure 6C). Based on Eggnog annotation result, the orthogroups of *P. polysora* with high copy numbers (>10) are mainly associated with the pathogenicity-related function, energy metabolism, RNA regulation and heat shock reaction (Table S9).

3.8 | Population genetics

To understand the population differentiation of *P. polysora*, we analyzed genome resequencing data from 79 isolates of *Ppz* collected from across maize-growing regions of China and performed a population genetic analysis. The analyses of bi-allelic frequency suggested the expected normal distribution of bi-allelic frequency for most isolates except GD1922-3 and GX1905-2 (Figure S4) which were excluded from the analysis. After removal, a total of 7,147,489 whole-genome SNPs were obtained from the remaining 77 isolates. PCA analysis separated the isolates into 2 groups, with a major group consisting of 73 isolates (blue circle in Figure 7B) and a minor group containing only four isolates (red circle in Figure 7B). However, this separation was independent of the geographic origin of the isolates (North, Central, and South China). Correspondingly, although higher genetic differentiation was revealed between North China and South China populations ($Fst = 8.1\times10^{-4}$) than in other regional pairs (3.9×10^{-4} and 3.6×10^{-4}), all pair-wise *Fst* values were close to zero indicating a lack of geographic differentiation. We used the standardized index of association, *rd*, to estimate the linkage

disequilibrium of *P. polysora* population in China. The observed *rd* distribution for all isolates was between the values calculated from simulated datasets with 0% linkage, a sexual population, and 50% linkage (Figure 7C). The result suggested that, although *Ppz* population in China showed as a clonal population, its evolutionary history may be influenced by some level of sexual reproduction.

Due to limited samples in the minor group, we did not calculate the *F*st between two genetic groups instead of comparing the highly differentiated SNPs between two genetic clusters. A total of 32,281 SNPs (*F*st > 0.9) were filtered and these SNPs were evenly distributed in 18 chromosomes without region specificity. These variants were annotated in 5,975 protein-coding genes, of which 305 genes were affected by the moderate or high impact on amino acid sequences. The known functional annotations of these 305 genes were mainly related to transmembrane transport, signal transduction, zinc ion/protein/nucleic acid binding, catalytic activity and protein kinase activity (Table S10). Interestingly, 16 secreted proteins including *AvrRppC* also showed mutations. The allele types of *AvrRppC* varied between the two genetic groups. In the major group, most isolates (93%) carried allele types of $AvrRppC^{1}$ and $AvrRppC^{ref}$. A few isolates showed the allele combination containing new allele types, $AvrRppC^{1}$ to $AvrRppC^{5}$ as well as $AvrRppC^{ref}$. Whereas in the minor group, all four isolates carried the allele of $AvrRppC^{A}$ and $AvrRppC^{J}$ (Figure 7D). The CDS and amino acid sequences for new allele types are listed in Table S11.

4 | DISCUSSION

As rust fungi are dikaryotic pathogens, obtaining their nuclear phased assembly is critical for pathogenicity studies. In this study, we reported a haplotype-phased and chromosome-scale genome of *P. polysora* based on HiFi reads and Hi-C data. The 18 chromosomes had telomeres at both ends and showed high completeness (~90% complete BUSCOs for each nucleus and 99% complete BUSCOs for two nuclei) and high continuity (N50 of scaffolds = 54 M) as well as over 80% nuclear phasing. These data strongly supported that we generated a high-quality reference genome for *P. polysora*. In addition, this is the first case to assemble a gigabase-sized fungal genome to the telomere-telomere level. The assembly pipeline we illustrated in Figure 1 will inform future genome assembly for other dikaryotic or non-haploid organisms. Although NuclearPhaser has been developed and successfully applied in three rust other species (*Pgt*21-0,

*Pt*76 and *Pca*203) (Li et al. 2019; Duan et al. 2022; Henningsen et al. 2022), whether it can handle large genome size (> 1 G) needs to be further confirmed.

Compared to its close relatives, *P. polysora* has experienced a large genome expansion, with its genome size of 1.71 Gbp equivalent to 7–10 times of four other *Puccinia* species assemblies (*Pst*-104E, *Pgt*21-0, *Pt*76, and *Pca*203) (Schwessinger et al. 2018; Li et al. 2019; Wu et al. 2021; Henningsen et al. 2022). Based on a limited set of rust species, Tavares et al (2014) inferred that rusts infecting monocot Poaceae hosts have considerably smaller genomes than rusts infecting dicot Fabaceae hosts (556.6 Mbp on average). Thus, our study supplied an exception to this observation, indicating it may not be a general model. Gene duplication, repeat content variation and ploidy variation are common mechanisms to account for the genome expansion of fungi (Castanera et al. 2016; Sipos et al. 2017; Todd et al. 2017). Although gene expansion or loss has happened in the evolutionary process from *P. polysora* to other wheat-like *Puccinia* species (Figure 6B) and correspondingly the gene number predicted in *Ppz*-GD1913 (~24000 per haplotype) is much more than those of its two close relatives, *P. coronata* (~18000) and *P. striiformis* (~14000) (Miller et al. 2018; Schwessinger et al. 2018), the expanded genes only contribute to a size difference of 7 Mbp and 17 Mbp (Table S12). On the contrary, repeat analyses suggested that genome expansion of *P. polysora* is due to its large repeat contents (85% of the total genome), notably a proliferation of LTR-RTs, which is consistent with the common evidence for rust genome expansion (Tobias et al. 2021). The evidence of genome expansion has also been detected in other obligate fungi, such as powdery mildew (PM) fungi (Spanu et al. 2010; Frantzeskakis et al. 2020) and symbiotic fungi (Miyauchi et al. 2020). PM fungi lose carbohydrate-active enzymes, transporters, etc. which are probably redundant genes in strict parasitism as tradeoffs (Spanu et al. 2010). The low abundance of PCWDEs and transcription factors in *P. polysora* as well as other *Puccinia* species was in line with the above pattern (Figure 5) (Duplessis et al. 2011, Miller et al. 2018). Despite the massive TE proliferation causing a significant enlargement in genome size, *P. polysora* shows high gene synteny with other *Puccinia* species (Figure 5). This overall chromosome synteny was consistent with the dominant synteny detected by comparing BUSCO genes (Henningsen et al. 2022). In addition, this conservatism of chromosome synteny was also detected between species from different genera (Edwards et al. 2022).

In fungi, TEs have been implicated in coevolution with the host, genome architecture, plasticity, and adaptation (Lorrain et al. 2021). The "TE-thrust hypothesis" proposed that TEs can act to generate genetic novelties for organisms leading to speciation or may be related to adaptation to new challenges, e.g. new host or rapidly changing climate (Oliver and Greene 2012; Zeh et al. 2009). We detected a TE burst of *P. polysora* at 1.7 Mya (Figure 2) and the phylogenetic tree (Figure 6) suggested that *P. polysora* speciated earlier than the divergence time between *P. coronata* and *P. striiformis* (40 Mya) (Aime et al. 2018). So, TE burst of *P. polysora* may be unlikely for speciation. Other explanation could be the “nearly neutral theory", in which varied genomic features (e.g. large genome size, TE content, introns) are not initially harmful but are passively fixed by mutation or random genetic drift (Lynch and Conery 2003). With more available genome resources in the future, the relationship between effective genome size and the adaptative evolution of *P. polysora* needs to be answered.

The “two-speed genome” concept has been put forward to highlight the over-representation of effector-like genes in the repeat-rich and gene-sparse genome in many filamentous pathogens (Dong et al. 2015, Frantzeskakis et al. 2019). These TE-rich invasion/blocks may contribute to extensive chromosomal reshuffling and the rise of accessory chromosomes (de Jonge et al. 2013). Pathogens can resolve the evolutionary conflict through rapid evolution of effector genes to adapt to changing environment but maintain the housekeeping genes in the core genome with a moderate evolution rate (Presti et al. 2015). However, the situation is clearly different in rust fungi. There was no signal of effector compartments detected in Oat crown rust, wheat stripe rust and myrtle rust pathogens when comparing the intergenic distance between all genes and effectors (Schwessinger et al. 2018; Miller et al. 2018; Tobias et al. 2021). In *P. polysora*, the numerous TEs are evenly dispersed throughout the genome (Figure 2). The intergenic distance distribution of candidate effectors overlapped with that of all genes and BUSCOs (Figure 7). This genomic architecture with TEs and genes interspersed was also reported in another biotrophic fungal group, powdery mildew (PM) fungi (Frantzeskakis et al. 2020). Similarly, AT-rich isochores or large-scale compartmentalization are also missing in PM fungi (Frantzeskakis et al. 2020). When compared with other *Puccinia* species, the high repeat content of *P. polysora* leads to much lower relative gene space (4% in *Ppz* vs 20%–35% for currently reported *Puccinia* species) and large intergenic distance expansion (~100 kbp of *Ppz* vs 1kbp for *Pst* and *Pca*) (Figure 6B) (Miller et al. 2018; Schwessinger et al. 2018).

In the “arms race” between rust species and plants, the *Avr* genes in rust pathogens have been the subject of mutations to avoid the recognition by host resistance genes (Cui et al. 2015). Therefore, monitoring the variation of *Avr* genes in rust populations is a priority for disease management. Based on whole-genome SNPs, we detected two genetic groups of *P. polysora* in China. The highly differentiated SNPs between these two groups suggested mutations in 16 secreted proteins including the only verified avirulence gene, *AvrRppC*. A recent study investigated the allele types of *AvrRppC* in a Chinese *P. polysora* population (Deng et al. 2022). Isolates with allele types of $AvrRppC^{A}$, $AvrRppC^{F}$ and $AvrRppC^{J}$ could escape the recognition of *RppC* causing southern corn rust but $AvrRppC^{ref}$, $AvrRppC^{E}$, and $AvrRppC^{C}$ are avirulence alleles that trigger *RppC*-mediated resistance. In our study, isolates in the major group almost all carried the avirulence allele, $AvrRppC^{ref}$ and four isolates in the minor genetic group carried the virulence allele types, $AvrRppC^{A}$ and $AvrRppC^{J}$ (Figure 7D). These results suggested that differentiation between two genetic groups could be related to virulence evolution and such relation was more prominent in the wheat stripe rust pathogen (Hubbard et al. 2015). Upadhyaya et al (2021) suggested that avirulence genes of *P. graminis* showed a similar expression pattern. In this study, *AvrRppC* was classified in cluster 3 (Figure 4A) with low expression in germinated spores and early infection but the high expression in intermediate infection (4 dpi and 7 dpi). Among 16 secreted proteins with $Fst > 0.9$ between two genetic groups, FUNA_005021 (cluster 2) and FUNA_020823 (cluster 3) showed similar expression patterns to *AvrRppC* and may be candidate avirulence genes. In other rust species, the rust population could drastically shift towards a wider or new spectrum of virulence over time (Miller et al. 2020; Bai et al. 2021), therefore, performing long-term surveillance as well as developing effective virulence markers are particularly critical for *P. polysora*.

Rust fungi have complex life cycles involving asexual as well as sexual reproduction (Figueroa et al. 2020; Zhao et al. 2021). In some species, sexual reproduction is rare or cryptic that it can remain undiscovered. For example, it took a century to discover the alternate host of *P. striiformis* (Jin et al. 2010). In our case, the sexual stage of *P. polysora* is still a mystery; teliospores are rarely found in natural fields and germination tests of teliospores were not successful in past attempts (Cammack 1959). Also, studies on population genetics of *P. polysora* are very limited, possibly due to the lack of effective molecular markers and the reference genome. Our data using whole-genome SNPs supported the clonal Chinese population of *P.*

polysora, however, the low *rd* between those from simulated with 0% and 50% linkage suggested the potential influence of sexual reproduction. The absence of a sexual stage or alternate host may be reflected by unequal gene numbers in the secretome. A supportive example is two species in *Melampsora*. *Melampsora lini* (autoecious, no alternate host, ~800 secreted proteins) was found to have much less secreted proteins and secreted plant cell wall-degrading enzymes (PCWDEs) than its close relative, *M. larici-populina* (heteroecious, with both primary and alternate hosts, ~1800 secreted proteins) (Presti et al. 2015). In our case, we identified 2434 secreted proteins for *P. polysora*, which was comparable to the secretomes of two heteroecious relatives, *P. coronata* (~2500) and *P. graminis* (~2500) (Miller et al. 2018; Li et al. 2019). Of course, this single *Melampsora* example did not allow for making a general conclusion. Varied quality of genome assemblies and different prediction methods can lead to significant bias in secreted protein predictions. However, it will be valuable to test this hypothesis with more available high-quality rust genomes in future studies. If the hypothesis is true, our results may suggest the existence of alternate hosts that have not been discovered yet for *P. polysora.*

5 | CONCLUSION

In conclusion, we successfully obtained the first nuclear phased and chromosome-scale genome assembly of the serious fungal pathogen, *P. polysora*. The high-quality genome assembly and fine-time-point transcriptome facilitate the comparative genomic analyses and The genome expansion of *P. polysora* is driven by TEs, but this did not affect its gene synteny with close relatives. The investigation of genome characteristics and functional features provided a fundamental resource for pathogenicity studies as well as to understand the evolutionary mechanisms of genome expansion in rust fungi. Also, the population genomic data revealed low genetic differentiation in the Chinese *P. polysora* population, which expanded from south to north in recent decades. However, a minor genetic group with low frequency suggested the ongoing virulence evolution to evade recognition by *RppC*, a major resistance gene in Chinese corn cultivars, alarming the need to excavate additional resistance genes as soon as possible.

ACKNOWLEDGEMENTS

We would like to thank Xiufeng Liu, Peng Lu, Gongxian Liao, Shikun Pan, Xingshang Lu, Baoqin Tan, Qiusheng Luo for field diseased sample collection. We also appreciate valuable

suggestions from Dr. Mao Jianfeng and Dr. Qi Wu before submission. Clement K.M. Tsui is grateful to CAS PIFI for the award of visiting scientist fellowship.

AUTHOR CONTRIBUTIONS

The experiment was conceived and managed by Lei Cai and Junmin Liang. Field samples were collected by Junmin Liang, Yuanjie Li, Zhanhong Ma, Leifu Li and Keyu Zhang. Single-spore isolation, reproduction, DNA/RNA extraction and other biological materials used for sequencing were conducted by Junmin Liang and Yuanjie Li. Data analyses were performed by Junmin Liang. Analysis tools were assisted by Peter N. Dodds, Melania Figueroa and Jana Sperschneider. Junmin Liang wrote the manuscript and Peter N. Dodds, Jana Sperschneider, Clement K.M. Tsui, and Lei Cai reviewed the manuscript.

DATA AVAILABILITY. All raw sequence reads (PacBio, Illumina, Hi-C and RNA-seq data) generated in this study are available in the National Microbiology Data Center (NMDC, https://nmdc.cn/) under the BioProject (NMDC10018113) and the accession number for all raw data are listed in Table S13. The assembled genome has been deposited at NMDC under the accession number NMDC60042795. Scripts used to construct figures and annotation files are available at https://github.com/jimie0311/Puccinia-polysora-genome.

FUNDING. This work was financially supported by the National Sciences Foundation of China (NSFC 31972210 and NSFC 31725001) and National Sciences and Technology Fundamental Resources Investigation Program of China (2021FY100900).

REFERENCES

1. Aime, M. C., & McTaggart, A. R. (2021). A higher-rank classification for rust fungi, with notes on genera. *Fungal Systematics and Evolution*, 7, 21–47.
2. Aime, M. C., Bell, C. D., & Wilson, A. W. (2018). Deconstructing the evolutionary complexity between rust fungi (Pucciniales) and their plant hosts. *Studies in Mycology*, 89,143–152.
3. Aime, M. C., McTaggart, A. R., Mondo, S. J., & Duplessis, S. (2017). Phylogenetics and phylogenomics of rust fungi. *Advances in Genetics*, 100, 267–307.

4. Akdemir, K.C., & Chin, L. (2015). HiCPlotter integrates genomic data with interaction matrices. *Genome Biololgy,* 16, 198.
5. Anderson, C., Khan, M. A., Catanzariti, A. M., Jack, C. A., Nemri, A., Lawrence, G. J., Upadhyaya, N. M., Hardham, A. R., Ellis, J. G., Dodds, P. N., Jones, D. A. (2016). Genome analysis and avirulence gene cloning using a high-density RADseq linkage map of the flax rust fungus, *Melampsora lini*. *BMC Genomics,* 17, 667.
6. Bai, Q., Wan, A. M., Wang, M. N., See, D. R., & Chen, X. M. (2021). Molecular characterization of wheat stripe rust pathogen (*Puccinia striiformis* f.sp. *tritici*) Collections from Nine countries. *International Journal of Molecular Sciences*, 22, 9457.
7. Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6, 4–9.
8. Baranova, M. A., Logacheva, M. D., Penin, A. A., Seplyarskiy, V. B., Safonova, Y. Y., Naumenko, S. A., Klepikova, A. V., Geraso,pv. E. D., Bazykin, G. A., James, T. Y., & Kondrashov, A. S. (2015). Extraordinary genetic diversity in a wood decay mushroom. *Molecular Biology and Evolution*, 32, 2775–2783.
9. Boehm, E. W. A., Wenstrom, J. C., McLaughlin, D. J., Szabo, L. J., Roelfs, A. P., & Bushnell, W. R. (1992). An ultrastructural pachytene karyotype for *Puccinia graminis* f.sp. *tritici*. *Canadian Journal of Botany*, 70, 401–413.
10. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114–2120.
11. Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12, 59.
12. Cammack, R. H. (1959). Studies on *Puccinia polysora* Underw. III. Description and life cycle of *P. polysora* in West Africa. *Transactions of the British Mycological Society*, 42, 55–58.
13. Castanera, R., López-Varas, L., Borgognone, A., LaButti, K., Lapidus, A., Schmutz, J., Grimwood, J., Pérez, G., Pisabarro, A. G., Grigoriev, I. V., Stajich, J. E., & Ramírez, L. (2016). Transposable elements versus the fungal genome: impact on whole genome architecture and transcriptional profiles. *PLoS Genetics*, 12, e1006108.

14. Chen, J., Upadhyaya, N. M., Ortiz, D., Sperschneider, J., Li, F., Bouton, C., ... & Dodds, P. N. (2017). Loss of *AvrSr50* by somatic exchange in stem rust leads to virulence for *Sr50* resistance in wheat. *Science*, 358, 1607–1610.
15. Chin, C. S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion. G. T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G. R., Delledonne, M., Luo, C., Ecker, J. R., Cantum D., Ranmk, D, R., & Schatz, M. C. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*. 13, 1050–1054.
16. Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Ruden, D. M., & Lu, X. Y. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)*, 6, 80–92.
17. Crouch, J. A., & Szabo, L. J. (2011). Real-time PCR detection and discrimination of the southern and common corn rust pathogens *Puccinia polysora* and *Puccinia sorghi*. *Plant Disease*, 295, 624–632.
18. Cui, H., Tsuda, K., & Parker, J. E. (2015). Effector-triggered immunity: from pathogen perception to robust defense. *Annual Review of Plant Biology,* 66, 487–511.
19. de Jonge, R., Bolton, M. D., Kombrink, A., van den Berg, G. C. M., Yadeta, K. A., Thomma, B. P. H. J. (2013). Extensive chromosomal reshuffling drives evolution of virulence in an asexual pathogen. *Genome Research*, 23,1271–1282.
20. Deng, C., Leonard, A., Cahill, J. M., Meng, L., Li, Y.R., Thatcher, S. Li, X. Y., Zhao, X. D., Du, W. J., Li, Z., Li, H. M., Llaca, V., Fengler, K., Marshall, L., Harris, C., Tabor, G., Li, Z. M., Tian, Z. Q., Yang, Q. H., Chen, Y. H., Tang, J. H., Wang, X. T., Hao, J. J., Yan, J. B., Lai, Z. B., Fei, X. H., Song, W. B., Lai, J. S., Zhang, X. C., Shu, G. P., Wang, Y. B., Chang, Y. X., Zhu, W. L., Xiong, W., Sun, J., Li, B. L., & Ding, J. Q. (2022).The RppC-AvrRppC NLR-effector interaction mediates the resistance to southern corn rust in maize. *Molecular Plant,* 15, 904–912.
21. Dodds, P. N., & Rathjen, J. P. (2010). Plant immunity: towards an integrated view of plant-pathogen interactions. *Nature Reviews Genetics*. 11, 539–548.
22. Dong, S., Raffaele, S., & Kamoun, S. (2015). The two-speed genomes of filamentous pathogens: waltz with plants. *Current Opinion in Plant Biology*, 35, 57–65.

23. Duan, H., Jones, A. W., Hewitt, T., Mackenzie, A., Hu, Y., Sharp, A., Lewis, D., Mago, R., Upadhyaya, N. M., Rathjen, J. P, Stone, W. A., Schwessinger, B., Figueroa, M., Dodds, P. N., Periyannan, S., & Sperschneider J. (2022). Identification and correction of phase switches with Hi-C data in the Nanopore and HiFi chromosome-scale assemblies of the dikaryotic leaf rust fungus *Puccinia triticina*. *Current Biology*, 23, 84.
24. Duplessis S, Cuomo CA, Lin YC, Aerts A, Tisserant E, Veneault-Fourrey C., Joly, D. L., Hacquard, S., Amselem, J., Cantarel, B. L., Chiu, R., Coutinho, P. M., Feau, N., Field, M., Frey, P., Gelhaye, E., Goldberg, J., Grabherr, M. G., Kodira, C. D., Kohler, A., Kües, U., Lindquist, E. A., Lucas, S. M., Mago, R., Mauceli, E., Morin, E., Murat, C.,Pangilinan, J. L., Park, R., Pearson, M., Quesneville, H., Rouhier, N., Sakthikumar, S., Salamov, A. A., Schmutz, J., Selles, B., Shapiro, H., Tanguay, P., Tuskan, G. A., Henrissat, B., Van de, Peer. Y., Rouze, P., Ellis, J. G., Dodds, P. N., Schein, J. E., Zhong, S., Hamelin, R. C., Grigoriev, I. V., Szabo, L. J., & Martin, F. (2011). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proceedings of the National Academy of Sciences*, 108, 9166–9171.
25. Edwards, R. J., Dong, C. M., Park, R. F., Tobias, P. A. (2022) A phased chromosome-level genome and full mitochondrial sequence for the dikaryotic myrtle rust pathogen, *Austropuccinia psidii*. https://doi.org/10.1101/2022.04.22.489119
26. Ellinghaus, D., Kurtz, S. & Willhoeft, U. (2008). *LTRharvest*, an efficient and flexible software for *de novo* detection of LTR retrotransposons. BMC Bioinformatics 9, 18.
27. Emms, D.M. & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology,* 16, 157.
28. Figueroa, M., Dodds, P. N., & Henningsen, E. C. (2020). Evolution of virulence in rust fungi–multiple solutions to one problem. *Current Opinion in Plant Biology*, 56, 20–27.
29. Figueroa, M., Upadhyaya, N. M, Sperschneider, J., Park, R. F., Szabo, L. J., Steffenson, B., Eills, J. G., Dodds, P. N. (2016). Changing the game: using integrative genomics to probe virulence mechanisms of the stem rust pathogen *Puccinia graminis* f. sp. *tritici*. *Frontiers in Plant Science*, 7, 205.

30. Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L., Tate, J., & Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Research,* 42, D222–D230.
31. Frantzeskakis, L., Barsoum, M., Panstruga, R., Kusch, S., Kiss, L., Takamatsu, S., & Panstruga, R. (2019). The *Parauncinula polyspora* draft genome provides insights into patterns of gene erosion and genome expansion in powdery mildew fungi. *MBio*, 10, 1–17.
32. Frantzeskakis, L., Pietro, A. D., Rep, M., Schirawski, J., Wu, C. H., & Panstruga, R. (2020). Rapid evolution in plant–microbe interactions—a molecular genomics perspective. *New Phytologist*, 225, 1134–1142.
33. Gel, B., & Serra, E. (2017). karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* 33, 3088–3090.
34. Ghurye, J., Pop, M., Koren, S., Bickhart, D., & Chin, C. S. (2017). Scaffolding of long read assemblies using long range contact information. *BMC genomics*, 18, 527.
35. Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., & Regev, A. (2011). Trinity: reconstructing a full-length transcriptome assembly from RNA-Seq data. *Nature biotechnology*, 29, 644–652.
36. Gu, Z. (2014). Circlize implements and enhances circular visualization in R. *Bioinformatics* 30, 2811–2812.
37. Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., & Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome biology, 9, R7.
38. Henningsen, E. C., Hewitt, T., Dugyala, S., Nazareno, E. S., Gilbert, E., Li, F., Kianian, S. F, Steffenson, B. J., Dodds, P. N., Sperschneider, J., & Figueroa, M. (2022). A chromosome-level, fully phased genome assembly of the oat crown rust fungus *Puccinia coronata* f. sp. *avenae*: a resource to enable comparative genomics in the cereal rusts. https://doi.org/10.1101/2022.01.26.477636.

765 39. Huang, S. F., Kang, M. J., & Xu, A. L. (2017). HaploMerger2: rebuilding both haploid sub-
766 assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics*, 33,2577–
767 2579.
768 40. Hubbard, A., Lewis, C. M., Yoshida, K., Ramirez-Gonzalez, R. H., de Vallavieille-Pope, C.,
769 Thomas, J., Kamoun, S., Bayles, R., Uauy, C., & Saunders, D. G. (2015). Field
770 pathogenomics reveals the emergence of a diverse wheat yellow rust population. *Genome*
771 *Biology*, 16, 23. https://doi.org/10.1186/s13059-015-0590-8
772 41. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H.,
773 Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019).
774 eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource
775 based on 5090 organisms and 2502 viruses. Nucleic acids research, 47(D1), D309–D314.
776 42. Jin, Y., Szabo, L. J., & Carson, M. (2010). Century-old mystery of *Puccinia striiformis* life
777 history solved with the identification of *Berberis* as an alternate host. *Phytopathology*. 100,
778 432–435.
779 43. Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H.,
780 Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A.,
781 Scheremetjew, M., Yong, S. Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: genome-
782 scale protein function classification. Bioinformatics (Oxford, England), 30(9), 1236–1240.
783 https://doi.org/10.1093/bioinformatics/btu031
784 44. Justesen, A. F., Ridout, C. J., & Hovmøller, M. S. (2002). The recent history of *Puccinia*
785 *striiformis* f. sp. *tritici* in Denmark as revealed by disease incidence and AFLP markers.
786 *Plant pathology*, *51*, 13–23.
787 45. Kamvar, Z N., Tabima, J. F., & Grünwald, N. J. (2014) Poppr: an R package for genetic
788 analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ,* 2,
789 e281.
790 46. Kim, D., Langmead, B. & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low
791 memory requirements. *Nature Methods*, 12, 357–360.
792 47. Knaus, B. J., & Grünwald, N, J. (2017). vcfr: a package to manipulate and visualize variant
793 call format data in R. *Molecular Ecology Resources*. 17, 44–53.
794 48. Kolmer, J. A. (2005). Tracking wheat rust on a continental scale. *Current Opinion in Plant*
795 *Biology*, 8, 441–449.

49. Li, F., Upadhyaya, N. M., Sperschneider, J., Matny, O., Nguyen-Phuc, H., Mago, R., Raley, C., Miller, M. E., Silverstein, K., Henningsen, E., Hirsch, C. D., Visser, B., Pretorius, Z. A., Steffenson, B. J., Schwessinger, B., Dodds, P. N., & Figueroa, M. (2019). Emergence of the Ug99 lineage of the wheat stem rust pathogen through somatic hybridisation. *Nature communications*, 10, 5068. https://doi.org/10.1038/s41467-019-12927-7
50. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics,* 34, 3094–3100.
51. Li, H., & Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26, 589–595.
52. Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30, 923–930.
53. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O., & Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm, *Nucleic Acids Research,* 33, 6494–6506.
54. Lorrain, C., Feurtey, A., Mller, M., Haueisen, J., & Stukenbrock, E. (2021). Dynamics of transposable elements in recently diverged fungal pathogens: lineage-specific transposable element content and efficiency of genome defences. *G3-Genes Genomes Genetics,* 11, jkab068.
55. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15, 550.
56. Lowe, T. M., & Chan, P. P. (2016). TRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Research*, 44, W54 –W57.
57. Lue, N. F. (2021). Duplex telomere-binding proteins in fungi with canonical telomere repeats: new lessons in the rapid evolution of telomere proteins. *Frontier in Genetics*, 12,638790.
58. Lynch, M., & Conery, J. S. (2003). The origins of genome complexity. *Science,* 302, 1401–1404.
59. Maia, T,, Badel, J. L., Marin-Ramirez, G., Rocha, C. M., Fernandes, M. B., da Silva, J. C, de Azevedo-Junior, G. M., Brommonschenkel, S. H. (2017). The *Hemileia vastatrix* effector HvEC016 suppresses bacterial blight symptoms in coffee genotypes with the SH1 rust resistance gene. *New Phytologist*, 213,1315–1329.

60. Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S.L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, 14, e1005944.
61. Marcais, M., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics,* 27, 764–770.
62. Miller, M. E., Nazareno, E. S., Rottschaefer, S. M., Riddle, J., Dos Santos Pereira, D., Li, F., Nguyen-Phuc, H., Henningsen, E. C., Persoons, A., Saunders, D., Stukenbrock, E., Dodds, P. N., Kianian, S. F., & Figueroa, M. (2020). Increased virulence of *Puccinia coronata* f. sp. *avenae* populations through allele frequency changes at multiple putative *Avr* loci. *PLoS genetics*, 16, e1009291. https://doi.org/10.1371/journal.pgen.1009291
63. Miller, M. E., Zhang, Y., Omidvar, V., Sperschneider, J., Schwessinger, B., Raley, C., Palmer, J. M., Garnica, D., Upadhyaya, N., Rathjen, J., Taylor, J. M., Park, R. F., Dodds, P. N., Hirsch, C. D., Kianian, S. F., & Figueroa, M. (2018). *De Novo* Assembly and Phasing of Dikaryotic Genomes from Two Isolates of *Puccinia coronata* f. sp. *avenae*, the Causal Agent of Oat Crown Rust. *mBio*, 9, e01650-17. https://doi.org/10.1128/mBio.01650-17
64. Miyauchi, S., Kiss, E., Kuo, A., Drula, E., Kohler, A., Sánchez-García, M., Morin, E., Andreopoulos, B., Barry, K. W., Bonito, G., Buée, M., Carver, A., Chen, C., Cichocki, N., Clum, A., Culley, D., Crous, P. W., Fauchery, L., Girlanda, M., Hayes, R. D., ⋯ Martin, F. M. (2020). Large-scale genome sequencing of mycorrhizal fungi provides insights into the early evolution of symbiotic traits. *Nature Communication,* 11, 5125. https://doi.org/10.1038/s41467-020-18795-w
65. Nurk, S., Walenz, B. P., Rhie, A., Vollger, M. R., Logsdon, G. A., Grothe, R., ... & Koren, S. (2020). HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome research*, *30*, 1291–1305.
66. Oliver, K. R.,& Greene, W. K. (2012). Transposable elements and viruses as factors in adaptation and evolution: an expansion and strenthening of the TE thrust hypothesis. *Ecology and Evolution*, 2, 2912–2933.
67. Ou, S., & Jiang, N. (2018). LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiology*, 176, 1410–1422.

68. Petersen, T. N., Brunak, S., von Heijne, G., & Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, 8, 785–786.
69. Pfeifer, B., Wittelsbuerger, U., Ramos-Onsins, S. E., & Lercher, M. J. (2014). PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. *Molecular Biology and Evolution*, 31, 1929–1936.
70. Pinon, J., & Frey, P. (1997). Structure of *Melampsora larici-populina* populations on wild and cultivated poplar. *European Journal of Plant Pathology*. 103,159–173.
71. Presti, L. L., Lanver, D., Schweizer, G., Tanaka, S., Liang, L., & Tollot, M. (2015). Fungal effectors and plant susceptibility. *Annual Review of Plant biology,* 66, 513–545.
72. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., maller, J.m Sklar, P., Bakker, P. I. W. D., & Daly, M. J (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. The *American Journal of Human Genetics*, 81, 559–575.
73. Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 26, 841–842.
74. Ramirez-Cabral, N. Y. Z., Kumar, L., & Shabani, F. (2017). Global risk levels for corn rusts (*Puccinia sorghi* and *Puccinia polysora*) under climate change projections. Journal of Phytopathology, 165, 563–574.
75. Ramos, A. P., Tavares, S., Tavares, D., Silva, M. D. C, Loureiro, J., & Talhinhas, P. (2015). Flow cytometry reveals that the rust fungus, *Uromyces bidentis* (Pucciniales), possesses the largest fungal genome reported-2489Mbp. *Molecular Plant Pathology*, 16, 1006–1010.
76. Rawlings, N. D., Barrett, A. J., & Finn, R. (2016). Twenty years of the Merops database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Research*, 44, D343–D350.
77. Roach, M. J., Schmidt, S. A., & Borneman, A R. (2018). Purge Haplotigs: allelic contig ressignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, 19, 460.
78. Rochi, L., Diéguez, M. J., Burguener, G., Darino, M. A., Pergolesi, M. F., Ingala, L. R., Cuyeu, A. R., Turjanski, A., Kreff, E. D., & Sacco, F. (2018). Characterization and comparative analysis of the genome of *Puccinia sorghi* Schwein, the causal agent of maize common rust. *Fungal Genetics and Biology*, 112, 31–39.

79. Salcedo, A., Rutter, W., Wang, S., Akhunova, A., Bolus, S., Chao, S., Anderson, N., De Soto, M. F., Rouse, M.,Szabo, L., Bowden, R. L., Dubcovsky. J., & Akhunov, E. (2017). Variation in the *AvrSr35* gene determines *Sr35* resistance against wheat stem rust race Ug99. *Science*, 358, 1604–1606.
80. Schwessinger, B., Sperschneider, J., Cuddy, W. S., Garnica, D. P., Miller, M. E., Taylor, J. M., Dodds, P. N., Figueroa, M., Park, R. F., & Rathjen, J. P. (2018). A near-complete haplotype-phased genome of the dikaryotic wheat stripe rust fungus *Puccinia striiformis* f. sp. *tritici* reveals high interhaplotype diversity. *mBio*, 9, e0227517.
81. Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C. J., Vert, J. P., Heard, E., Dekker, J., & Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology*, 16, 259.
82. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics,* 31, 3210–3212.
83. Singh, R. P., Hodson, D. P., Huerta-Espino, J., Jin, Y., Bhavani, S., Njau, P., Herrera-Foessel, S., Singh, P. K., Singh, S., & Govindan, V. (2011). The emergence of Ug99 races of the stem rust fungus is a threat to world wheat production. *Annual Review of Phytopathology*, 49, 465–481.
84. Sipos, G., Prasanna, A. N., Walter, M. C., O'Connor, E., Bálint, B., Krizsán, K., Kiss, B., Hess, J., Varga. T., Slot, J., Riley, R., Bóka, B., Rigling, F., Barry, K., Lee, J., Mihaltcheva, S., LaButti, K., Lipzen, A., Moloney, N. M., Sperisen, C., Kredics, L., Vágvölgyi, C., Patrignani, A., Fitzpatrick, D., Nagy, I., Soyle. S., Anderson, J. B., Grigoriev, I. V., Güldener, U., Münsterkötter, M., & Nagy, L. G. (2017). Genome expansion and lineage-specific genetic innovations in the forest pathogenic fungi Armillaria. Nature Ecology & Evolution, 1, 1931–1941.
85. Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6, 1–11.
86. Smit, A., & Hubley, R. (2008). RepeatModeler open 1.0. Institute for Systems Biology, Seattle, WA. http://www.repeatmasker.org/
87. Smit, A., Hubley, R., & Green, P. (2015). RepeatMasker open 4.0. 2013–2015. Institute for Systems Biology, Seattle, WA. http://www.repeatmasker.org/

88. Spanu, P. D., Abbott, J. C., Amselem, J., Burgis, T. A., Soanes, D. M., Stüber, K., ... & Panstruga, R. (2010). Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science*, 330,1543–1546.
89. Sperschneider, J., & Dodds, P. N. (2022). EffectorP 3.0: Prediction of Apoplastic and Cytoplasmic Wffectors in Fungi and Oomycetes. *Molecular Plant-Microbe Interaction*, 35, 146–156.
90. Stanke, M., & Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research*, 33, W465–W467.
91. Tavares, D., Romos, A. P., Pires, A. S., Azomjeora, H. G., Caldeirinha, P., Link, T., Abranches, R., Silva, M. do. C., Voegele, R. T., Loureiro, J., & Talhinhas, P. (2014). Genome size analyses of Pucciniales reveal the largest fungal genomes. *Frontiers in Plant Sceiences*, 5, 422.
92. Terrapon, N., Lombard, V., Drula, E., Coutinho, P.M., Henrissat, B. (2017). The CAZy Database/the Carbohydrate-Active Enzyme (CAZy) Database: Principles and Usage Guidelines. In: Aoki-Kinoshita, K. (eds) A Practical Guide to Using Glycomics Databases. Springer, Tokyo.
93. Tobias, P. A., Schwessinger, B., Deng, C.H., Wu, C., Dong, C. M., Sperschneider, J., Jones, A. Lou, Z. Y., Zhang, P., Sandhu, K., Smith, G. R., Tibbits, J., Chagne, D., & Park, R. F. (2021). *Austropuccinia psidii*, causing myrtle rust, has a gigabase-sized genome shaped by transposable elements. *G3*, 11, jkaa015.
94. Todd, R. T., Forche, A., & Selmecki, A. (2017). Ploidy Variation in Fungi: Polyploidy, Aneuploidy, and Genome Evolution. *Microbiology Spectrum*, 5,10.
95. Upadhyaya, N. M., Mago, R., Panwar, V., Hewittm, T., Luo, M., Chen, J., Sperschneider, J., Nguyen-Phuc, H., Wang, A.H., Ortiz, D., Hac, L., Bhatt, D., Li, F., Zhang, J. P., Ayliffe, M., Figueroa, M., Kanyuka, K., Wllis, J. G., & Dodds, P. N. (2021). Genomics accelerated isolation of a new stem rust avirulence gene–wheat resistance gene pair. *Nature Plants*, 7, 1220–1228.
96. Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics,* 14,2202–2204.

97. Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., Lee, T. H., Jin, H., Marler, B., Guo, H., Kissinger, J. C., & Paterson, A. H. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research*, 40, e49. https://doi.org/10.1093/nar/gkr1293
98. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag.
99. Winzer, F. L., Berthon, K. A., Carnegie, A. J., Pegg, G. S., & Leishman, M. R. (2019). *Austropuccinia psidii* on the move: survey based insights to its geographical distribution, host species, impacts and management in Australia. *Biological Invasions.* 21, 1215–1225.
100. Wolff, J., Bhardwaj, V., Nothjunge, S., Richard, G., Renschler, G., Gilsbach, R., Manke, T., Backofen, R., Ramírez, F., & Grüning, B. A. (2018). Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization. Nucleic Acids Research, 46, W11–W16.
101. Wu, J. Q., Song, L., Ding, Y., Dong, C. M., Hasan, M., & Park, R. (2021). A chromosome-scale assembly of the wheat leaf rust pathogen *Puccinia triticina* provides insights into structural variations and genetic relationship with haplotype resolution. *Frontiers in Microbiology,* 12, 704253.
102. Zeh, D. W., Zeh, J. A., & Ishida, Y. (2009). Transposable elements and an epigenetic basis for punctuated equilibria. *BioEssays,* 31, 715–726.
103. Zhao, P., Qi, X. H., Crous, P. W., Duan, W. J., & Cai, L. (2020). Gymnosporangium species on Malus: species delineation, diversity and host alternation. *Persoonia*, 45, 68–100.
104. Zhao, P., Zhang, Z. F., Hu, D. M., Tsui, K. M., Qi, X. H., Phurbu, D., Gafforov, Y., & Cai, L. (2021). Contribution to rust flora in China I, tremendous diversity from natural reserves and parks. *Fungal Diversity*, 110, 1–58.

Figure Legends

FIGURE 1 Chromosome-level assembly of GD1913 isolate of *Puccinia polysora*. (A) The dikaryotic phasing pipeline and switch strategies for duplicated contigs. Duplicated contigs were checked based on the all-vs-all sequence alignment from HaploMerger 2 (Huang et al. 2017). (B) Hi-C-based contig anchoring. The heat map showed the density of Hi-C interactions within haplotype A. The 18 chromosomes are highlighted by blue squares. (C) Schematic representation of assembled chromosomes for GD1913 of each haplotype. (D) Percentage of Hi-C links of each chromosome of haplotype A to either haplotype A (blue) or haplotype B (red).

FIGURE 2 Genomic features of *P. polysora* f.sp. *zeae* isolate, GD1913 and repeat comparison analyses with its close relatives. (A) Genomic landscape of 18 chromosomes for haplotype A (hapA) and haplotype B (hapB). From outer to inner circles: (i) chromosomes, (ii) repeats density, (iii) gene density, (iv) candidate effector density, (v) Guanine-cytosine (GC) content. (ii-iv) are drawn in non-overlapping 100 kbp sliding windows. (B) Comparison of genome size (two nuclei) and repeat contents among *Puccinia polysora* f.sp. *zeae* (*Ppz*) and other rust species, *Melampsora larici-populina* (*Mlp*), *Austropuccinia psidii (Ap*), *Puccinia sorghi* f.sp. *zeae* (*Psz*), *Pucccnia coronata* f.sp. *avenae* (*Pca*), *Puccinia graminis* f.sp. *tritici* (*Pgt*), *Puccinia triticina* f.sp. *tritici* (*Ptt*) and *Puccinia striiformis* f.sp. *tritici* (*Pst*). The isolate names are after the hyphen. (C) The insertion time (Mya) distribution of intact LTRs in *P. polysora*.

FIGURE 3 The interhaplotype variation of *Ppz*-GD1913 isolate. (A) Summary of interhaplotype variation between haplotype A and B using Assemblytis. Six types of specific variation were illustrated by schematic plot and each bar chart indicates the number of bases categorized to each type. (B) Synteny plot of haplotype A and B using Mummer. (C) and (D) Two representative whole-chromosome alignments.

FIGURE 4 Analyses of gene expression and gene distance of predicted secretome of *P. polysora*. (A) Clustering analysis of gene expression of secretome on two haplotypes. Heatmaps show rlo-transformed expression values. Cluster numbers are shown outside the graphs, and tracks represent gene expression in germinated spores (GS), and infected tissues at 1 dpi, 2 dpi, 4 dpi, 7 dpi, 10 dpi and 14 dpi. (B) Hexplots for closest-neighbor gene distance density of *P. polysora* in haplotype A (left) and haplotype B (right). Circle dots represent two gene categories, BUSCOs (yellow) and predicted candidate effectors (red). The black dots represent the distribution of the avirulent gene, *AvrRppC.* (C) Gene and repeat density as well as log2TPM (transcripts per

million) of candidate effectors from chromosomes 14A and 14B. The red and blue points represent expression values of candidate effectors at 1 dpi and 7dpi. Positions and protein sequences of *AvrRppC* genes are presented.

FIGURE 5 Functional annotation of CAZymes, transcription factors and Merops proteases in *Ppz*-GD1913 isolate. (A) CAZyme families comparison of *P. polysora* (*Ppz,* GD1913), *P. triticina* (*Ptt*, *Pt76* isolate), *P. graminis* (*Pgt*, *Pgt*21-0 isolate) and *P. coronata* (*Pca*, *Pca*203). Heat maps showing gene numbers annotated in the following classes: auxiliary activities (AA), carbohydrate-binding modules (CBM), carbohydrate esterases (CE), glycoside hydrolases (GH), glycosyltransferases (GTs), and polysaccharide lyases (PL). (B) Percentages of genes from *P. polysora* are predicted to encode members of various fungal transcription factor classes based on InterProScan and Eggnog annotation. (C) Merops families comparison of *Ppz*, *Ptt*, *Pgt* and *Pca*. Heat map showing gene numbers annotated in the classes of aspartic acid (A), cysteine(C), metalloprotease (M), serine protease (S), and threonine protease (T) or peptidase inhibitors (I).

FIGURE 6 Comparative analyses of macro-synteny and orthogroups among *Puccinia polysora* f. sp. *zeae* (GD1913) and its relatives, *Austropuccinia psidii* (Au_3), *P. triticina* f. sp. *tritic*i (*Pt*76), *P. graminis* f.sp. *tritici* (*Pgt*21-0) and *P. coronata* f.sp *avenae* (*Pca*203). (A) Macro-synteny of four species in *Puccinia* with available haplotype-phased and chromosome-scale genome. Only the chromosomes from haplotype A/primary haplotype were compared. Grey lines are homologous proteins identified by MCscan with default parameters and black lines present the genome inversions between species. For better visualization, the genome size ration of *Ppz*B:*Ppz*A:*Pt*76:*Pca*203:*Pgt*21-0 = 0.2:0.2:1:1:1. (B) Numbers of duplicated (+) and lost (-) genes inferred from OrthoFinder. The typical symptoms caused by each species are listed accordingly. (C) Distribution of species-specific orthogroups with different gene copy numbers among four *Puccinia* species.

FIGURE 7 Population genomics analyses of *P. polysora* from China. (A) Geographic distribution of 79 *P. polysora* isolates collected from north, central and south of China. (B) Principal component analysis of Chinese *P. polysora* population showing two genetic clusters. (C) Linkage disequilibrium test of Chinese *P. polysora* population. (D) *AvrRppC* genotypes of *P. polysora* isolates in minor (red) and major (blue) genetic groups.

Supplementary files

Table S1 Information of 79 isolates from China used for genome resequencing.

Table S2 Contig information removed from the final assembly due to low-quality assessment.

Table S3 Treatments for duplicate contigs between/within haplotype A and haplotype B

Table S4 Sample information and data used for transcript analyses.

Table S5 Functional annotation of specific orthologs between two haplotypes.

Table S6 GO and IPR annotation for predicted candidate effectors.

Table S7 Features of secreted proteins in different expression clusters of *P. polysora* f.sp. *zeae.*

Table S8 Proteome comparison of four *Puccinia* species.

Table S9 GO annotation of top 42 *P. polysora* specific orthogroups with high gene copy number.

Table S10 Functional annotation of SNPs with *F*st > 0.9 between two genetic groups.

Table S11 The CDS and amino acid sequences of *AvrRppC* gene.

Table S12 Comparison of Gene number and gene length among three *Puccinia* species.

Table S13 NMDC numbers for sequencing raw data used in this study.

Figure S1 Genome size estimation of *Ppz*-GD1913 using 90 Gb illumina short reads.

Figure S2 GO enrichment analysis of secreted proteins on two haplotypes.

Figure S3 Pair-wise synteny of *P. polysora* and three close relatives.

Figure S4 The bi-allele distribution of 79 isolates used for population genomics. Two isolates (GD1922-3 and GX1905-2) in red squares are removed from the final analysis due to abnormal distribution.