A genotyping by target sequencing chip for population genetic analysis of the wheat stripe rust pathogen (*Puccinia striiformis*)

Haohao Yan¹, Zhe Ma¹, Qiang Yao², Shiqin Cao³, Quzhen Jia⁴, Jiaqi Li¹, Jie Zhao¹, Weiyi Yan¹, Juhong Ma¹, Wen Chen⁵, Bo Zhang⁴, Dejun Han⁶, Zhensheng Kang¹, Lili Huang¹, and Qingdong Zeng¹

¹Northwest A&F University College of Plant Protection  
²Academy of Agriculture and Forestry Science, Qinghai University  
³Wheat Research Institute, Gansu Academy of Agricultural Sciences  
⁴Plant Protection Institute, Gansu Academy of Agricultural Sciences  
⁵Plant Protection Institute, Guizhou Academy of Agricultural Sciences  
⁶Northwest A&F University College of Agronomy

June 13, 2024

Abstract

Puccinia striiformis f. sp. tritici (Pst), the causative agent of wheat stripe rust, poses a significant threat to wheat production due to its rapid long-distance migration and epidemic properties. Understanding the genetic structure and dynamics of the Pst population is crucial for early prediction and establishment of effective control strategies. The types of molecular marker analyses used in previous population genetic studies are often costly, time-consuming, and labor-intensive. We developed a genotyping by target sequencing (GBTS) chip for Pst designed with candidate secretion proteins and highly polymorphic single nucleotide polymorphism (SNP) sites identified from genome resequencing. The chip can be used directly with diseased leaves, saving time and avoiding cross-contamination between samples. The feasibility and efficiency of the chip was tested using 225 infected leaf samples collected from the northwest oversummering region of China. This test yielded 1,293,150 high-quality SNPs with a maximum gap of 99,512 bp. Strict quality controls produced 19,139 SNPs, comprising the final Pst 20K GBTS chip. Population genetic analysis revealed frequent gene flow and similar genetic diversity of Pst between epidemic regions, consistent with wind field analysis, trajectory tracking, and field monitoring. The results demonstrated that the GBTS chip is more efficient, convenient, and lower in cost than previous methods. This study provides new insights into stripe rust population dynamics. Furthermore, the newly established chip offers a valuable method for enriching epidemiological recognition, guiding future research into inter-regional or continental transmission of an important plant pathogen.
Wheat Research Institute, Gansu Academy of Agricultural Sciences, Lanzhou, Gansu 730070, China
Plant Protection Institute, Gansu Academy of Agricultural Sciences, Lanzhou, Gansu 730070, China
Plant Protection Institute, Guizhou Academy of Agricultural Sciences, Guiyang, Guizhou 550006, China

*Correspondence: Qingdong Zeng, email: zengqd@nwafu.edu.cn; Zhensheng Kang email: kangzs@nwafu.edu.cn; Lili Huang email: huanglili@nwafu.edu.cn.

Abstract

*Puccinia striiformis* f. sp. *tritici* (*Pst*), the causative agent of wheat stripe rust, poses a significant threat to wheat production due to its rapid long-distance migration and epidemic properties. Understanding the genetic structure and dynamics of the *Pst* population is crucial for early prediction and establishment of effective control strategies. The types of molecular marker analyses used in previous population genetic studies are often costly, time-consuming, and labor-intensive. We developed a genotyping by target sequencing (GBTS) chip for *Pst* designed with candidate secretion proteins and highly polymorphic single nucleotide polymorphism (SNP) sites identified from genome resequencing. The chip can be used directly with diseased leaves, saving time and avoiding cross-contamination between samples. The feasibility and efficiency of the chip was tested using 225 infected leaf samples collected from the northwest oversummering region of China. This test yielded 1,293,150 high-quality SNPs with a maximum gap of 99,512 bp. Strict quality controls produced 19,139 SNPs, comprising the final *Pst* 20K GBTS chip. Population genetic analysis revealed frequent gene flow and similar genetic diversity of *Pst* between epidemic regions, consistent with wind field analysis, trajectory tracking, and field monitoring. The results demonstrated that the GBTS chip is more efficient, convenient, and lower in cost than previous methods. This study provides new insights into stripe rust population dynamics. Furthermore, the newly established chip offers a valuable method for enriching epidemiological recognition, guiding future research into inter-regional or continental transmission of an important plant pathogen.

KEYWORDS: *Puccinia striiformis* f. sp. *tritici*, *Pst* 20K GBTS chip, gene exchange, genetic diversity, wind field analysis, trajectory tracking

1 INTRODUCTION

Wheat stripe rust is an airborne fungal disease that occurs in wheat at all growth stages. The causative agent is *Puccinia striiformis* f. sp. *tritici* (*Pst*), a member of the phylum Basidiomycota. The disease is widely distributed across nearly all continents, seriously threatening wheat production worldwide (Chen, 2005; Hovmøller et al., 2011; Li & Zeng, 2002; Wellings, 2011). Due to the importance of wheat as a staple food crop, this disease is a threat to both economic prosperity and food security. Globally, wheat stripe rust causes estimated yield losses of ~2.1% of total production (Savary et al., 2019), corresponding to ~979 million US dollars annually (Beddow et al., 2015).

Wheat stripe rust is particularly destructive in China, which has the largest epidemic area of the disease worldwide. Inter-regional pathogen spread contributes to large-scale epidemics, over the past eight decades there have been eight nationwide outbreaks of wheat stripe rust due to inter-regional spread, causing significant yield losses (up to 14 million metric tons) (Ma, 2018; Wan et al., 2004). The epidemic in China is due to complicated and diverse geographic and meteorological conditions, the multiple cropping systems utilized, and life-cycle features of the pathogen (Ali et al., 2017; Wan et al., 2007; Zeng & Luo, 2006). *Pst* completes the disease cycle through a process of oversummering, infection of autumn-sown wheat seedlings, and overwintering, followed by rapid expansion in the spring (Li & Zeng, 2002). *Pst* is a strict obligate biotrophic fungus and therefore depends on living host plants. Because the pathogen thrives in cool, humid environments, the disease occurs most often in low-latitude areas with low temperatures and high humidity at night, and in high-latitude regions with regularly cool and humid conditions (Chen & Kang, 2017; Li & Zeng, 2002; Rapilly, 1979; Wellings, 2011).
In practice, resistance cultivar deployment or diversification is considered a valuable strategy for the control of wheat stripe rust. However, selective pressure from the host causes rapid pathogen evolution, and most plant resistance (R) genes are quickly circumvented by new pathogen races, resulting in "boom and bust" disease cycles (De Vallavieille-Pope et al., 2012). This process is promoted by the spread of windborne Pst spores that differ in pathogenicity from local strains and are therefore susceptible to regional wheat R genes and causes re-infection of fields. Thereby, an accurate understanding of Pst population structure and migration routes, especially for long-distance migration (Brown, 2015; Brown & Hovmoller, 2002), is very important for the development of disease management strategies such as targeted deployment of wheat cultivars with specific R genes in different epidemic region based on the prevalent pathogens (Campbell, 2001; Perrings et al., 2002).

Since the 1990s, many DNA marker techniques have been utilized to study Pst population structure throughout the world. Restriction fragment length polymorphism (RFLP), which entails Southern blotting of enzyme-digested DNA fragments, is a widely used marker technique (Saiki et al., 1985). Using this approach, identified a family of moderately repetitive DNA sequences specific to Pst (Shan et al., 1998); they revealed significant regional diversity and found that Tianshui City in southern Gansu was a hotspot for variation in virulence genes. RFLP shows relatively high replicability but is complex, time-intensive, costly, and often employs radioactive materials. The random amplified polymorphic DNA (RAPD) technique (Williams et al., 1990) quickly gained popularity as a method for investigating variations and polymorphisms due to its simplicity, user-friendliness, and universal primer applicability. RAPD markers were used to determine the population structure of Pst in the USA through polymorphism characterization of pathogen races (Chen et al., 1993). Despite the advantages of RAPD, it has low replicability and the data it produces are difficult to compare with those produced through other methods. Amplified fragment length polymorphism (AFLP) combines the advantages of RAPD and RFLP in a technique that is simple, efficient, cost-effective, and lacks radioactive substances (Vos et al., 1995). However, all three of these methods use dominant markers and cannot distinguish heterozygous from homozygous alleles.

Single sequence repeats (SSRs), also known as microsatellite markers, are co-dominant sequences identified by Enjalbert et al. (2002). These have been employed extensively to uncover Pst population structure and migration (Awais et al., 2023; Khan et al., 2019). For example, SSRs were used to compare Chinese Pst population with populations worldwide (Ali et al., 2014) and to decipher temporal maintenance in Pst populations in Gansu (Ali et al., 2016). The development of next-generation DNA sequencing has also enabled identification of genome-wide single-nucleotide polymorphism (SNP) markers, which are highly abundant and can be analyzed with automated pipelines (Mardis, 2008). Xia et al. (2016a) developed the first SNP marker panel for studying Pst population structure. Since then, SNP markers have been widely used for biology, genetic, and population studies of Pst (Chen & Kang, 2017). SSR and SNP markers have successfully revealed genetic diversity in Pst and improved our understanding of pathogen migration (Ding et al., 2021; Li et al., 2023). However, neither SSR nor SNP markers cover the whole Pst genome, and both show less polymorphic locus resolution compared to next-generation sequencing technologies. Although application of these markers to representative samples collected across the globe has enabled efficient tracking of invasive lineages and worldwide population structure (Ali et al., 2017; Hovmøller et al., 2016; Thach et al., 2016), the low resolution and incomplete coverage remain critical limitations.

The Pst genome is relatively smaller than 100 Mb in size. However, little is known about it, primarily because dikaryotic urediniospores (n + n) complicate generation of a complete genome assembly. Furthermore, the obligate parasitism of Pst means that it cannot be cultured on artificial media (Hubbard et al., 2015). Ali et al. (2011) conducted the first study using direct DNA extraction from infected lesions to address the constraints inherent in Pst culturing, the extracted DNA was used for microsatellite-based genotyping. Inspired by this study, a field pathogenomics strategy was devised, which focuses on transcriptomic sequencing of Pst-infected wheat leaves (Hubbard et al., 2015). However, most of the data obtained with these strategies are derived from host tissues. Such studies are therefore costly and yield only limited Pst data.

Multiple target-enrichment strategies have recently been developed that share several key benefits: high
throughput, low time requirements, and cost effectiveness (Mamanova et al., 2010). The genotyping by target sequencing (GBTS) strategy is a targeted sequencing approach with notable advantages for the fields of disease epidemiology and pathogen population genetics. In previous study by our lab, GBTS exhibited stability, reliability, flexibility, and cost-effectiveness. The replicability of the approach ensured consistent genotype acquisition regardless of laboratory conditions and sequencing platforms, fostering data comparability, integration, and sharing (Xiang et al., 2023). GBTS can incorporate various marker types, including SSRs, SNPs, and insertion/deletion sites (InDels), for population genetic analyses (Guo et al., 2022), so it is an ideal approach for population genetic characterization of plant pathogens.

To date, GBTS has been applied only to plants and animals (Xu et al., 2020). In the present study, we develop a GBTS-based genomic chip for the fungal pathogen *Pst* based on genome re-sequencing data. Analysis of a panel of wheat stripe rust samples collected from the northwest oversummering region in China was designed to validate the chip and demonstrate genetic relationships between pathogen populations. The inclusion of a significantly greater number of markers than previous methods enhances the resolution of genetic analyses possible with this chip, making it a valuable tool for understanding *Pst* genetic diversity and migration both within and between regions. Ultimately, the *Pst* 20K GBTS chip contributes to the development of comprehensive wheat stripe rust control strategies, ensuring ongoing food security.

2 MATERIALS AND METHODS

2.1 Identification of genome-wide SNP loci for chip development

The *Pst* GBTS development pipeline is shown in Figure 1. To create a reliable and balanced genotyping chip, the first chromosome-scale assembly of the *Pst* isolate Pst-134E (Schwessinger et al., 2022) was used as a reference genome. Probes were designed using the full genomic DNA sequences (from the start codon to the end codon) encodes all predicted effector proteins. Polymorphism data from 42 *Pst* isolates collected throughout the world were used for basic marker statistics and gap filling. For comparison with previous studies, probes were also designed for 67 previously reported SNP loci (Meng et al., 2020; Xia et al., 2016a) and 94 SSR loci (Bailey et al., 2015; Cheng et al., 2012; Luo et al., 2015). Target locus was included if they met the following criteria: SNPs had a minor allele frequency (MAF) > 0.1; there was < 30% missing data; the heterozygosity was < 30%. At last 8,415 regions were selected, and 94,242 probes (110 bp for each probe) were designed and synthesised.

2.2 Sampling strategy and collection

Samples were collected in Qinghai and Gansu provinces and in Ningxia Autonomous Region from June 2021 to August 2022 during the cropping seasons. For each province/autonomous region, samples were taken from three to four wheat fields in each selected county, there was a distance of at least 15 km between fields. At the early stage of infection, leaves with numerous *Pst* hyphae and urediniospores were collected. There were a total of 225 samples; 134 were collected from June–December of 2021, comprising 44 samples (21GS) from 14 counties in Gansu, 71 samples (21QH) from 13 counties in Qinghai, and 19 samples (21NX) from four counties in Ningxia. The remaining 91 samples were collected from May–August of 2022, consisting of 57 samples (22GS) from 19 counties in Gansu, 27 samples (22QH) from nine counties in Qinghai, and 7 samples (22NX) from two counties in Ningxia. Samples were returned to the laboratory and processed with a freeze dryer (Ningbo Xinzhi Freeze Drying Equipment Co., Ltd; Ningbo, Zhejiang, China) for 2 d, then stored in a desiccator with silica gel at 4 °C until further use. Detailed sample metadata, including sampling site, altitudes, and coordinates, are shown in Figure 2 and Table S1.

2.3 DNA extraction and sequencing

Four tissue samples were removed from the infected portion of each leaf with a leaf punch (6 mm in diameter). DNA was extracted using the CTAB protocol with some modifications as described by (Saghai-Maroof et al., 1984). The extracted DNA was initially treated with the GenoBaits End Repair Kit and sequencing adaptors, then transferred to MolBreeding Biotechnology Co., Ltd. (Shijiazhuang, Hebei, China) for a series of rigorous purification steps including hybridization with specific baits, purification with wash buffer, and
post-PCR amplification. The resulting libraries contained fragments ranging from 350–400 bp in size and were sequenced with the MGISEQ-2000 platform (MGI Tech, Shenzhen, China).

2.4 Variant detection and annotation

Low-quality raw reads were trimmed with fastp v0.20.0 with the following parameters: -n 10 -q 20 -u 40 (Chen et al., 2018). The resulting clean reads were aligned to the Pst-134E reference genome (Schwessinger et al., 2022) using BWA (Li & Durbin, 2009). SNP and InDel variant detection were performed using the Genome Analysis Toolkit (GATK) pipeline (McKenna et al., 2010). Patterning processes were conducted as described by Xiang et al. (2023). The data were first filtered using the VariantFiltration function with the following parameters: QD < 2.0, QUAL < 30, FS > 200.0, and ReadPosRankSum < -20.0. The GATK tools HaplotypeCaller and GenotypeGVCFs were used to variant calling and create raw variant calls. SNPs were annotated in SnpEff v5.1 (Cingolani et al., 2012) using the gene annotations for the Pst-134E16 primary contigs and were visualized in the R package ‘ggplot2’, and SNPs in gene coding sequence (CDS) regions were annotated and statistically analyzed.

To ensure genotyping accuracy, sites with sequencing depth (DP) < 10 were removed. The core SNP set was derived from the remaining SNPs using a linkage disequilibrium (LD) pruning procedure with PLINK v1.9 with the following parameters: ‘--indep pairwise 50 10 0.1’ (Sotiropoulos et al., 2022). For this process, SNPs were removed using a window size of 100 kb, a window step size of 50 SNPs, and an $r^2$ threshold of 0.2.

2.5 Population structure, genetic diversity, and genetic differentiation analyses

SNPs were filtered using the parameters described in a prior study of Puccinia graminis f. sp. tritici at the population scale (Guo et al., 2022). This required removal of loci with more than two alleles, insufficient total or alternative allele read coverage (below 3), and missing data for > 25% of the isolates. Loci with missing values or MAFs < 0.05 were removed, forming a high-quality SNP dataset (deposited to the National Genomics Data Center under project number PRJCA020812). Population structure was determined using STRUCTURE v2.3 (Pritchard et al., 2000). SNPs were filtered by linkage disequilibrium (LD) using PLINK v1.9 as described above (2.4 Variant detection and annotation). STRUCTURE output was processed with CLUMPP (Jakobsson & Rosenberg, 2007). A phylogenetic tree was constructed with the neighbor-joining (NJ) method in the ‘ape’ R package (Alexander et al., 2009) and visualized with iTOL v6 (https://itol.embl.de/) (Purcell et al., 2007). Discriminant analysis of principal components (DAPC) was performed in the ‘adegenet’ R package (Jombart et al., 2010). Kinship heatmaps were generated based on the K-matrix using the R package ‘pheatmap’ v1.0. (Wu et al., 2021). Genetic diversity was analyzed with the ‘POPPR’ R package (Kamvar et al., 2015). Observed and expected heterozygosity (Ho and He , respectively) were estimated in PLINK v1.9 (Chen et al., 2016).

To test genetic differentiation between the six populations collected from three regions in two years, the fixation index ($Fst$) was calculated using the R package ‘hierfstat’ (Takezaki & Nei, 1996). Analysis of Molecular Variance (AMOVA) and significance tests were performed in R to evaluate the molecular differences between samples from different years in the three regions (Excoffier et al., 1992). Nei’s genetic distance (D) (Nei, 1972) was calculated using the dist.genpop function in the ‘adegenet’ R package (v2.1.10). A visual migration network based on relative migrants was generated from the geneflow pattern network for different epidemic populations using the ‘diversity’ R package (Bai et al., 2021).

2.6 High-altitude wind analysis

Upper-air wind field data were obtained from the European Centre for Medium-Range Weather Forecasts (https://www.ecmwf.int; downloaded from https://CDs.climate.copernicus.eu). Data for the northwest overwintering region of China were analyzed in Python using the U component and V component data with the pressure level at 700 hPa. The propagation track and diffusion range of Pst uredospores were simulated with HYSPLIT-4 models. HYSPLIT-4 simulations were conducted at altitudes of 1900 m and 1000 m, corresponding to the sampling sites in Salar Autonomous County, Xunhua, Haidong City, Qinghai Province and Maiji District, Tianshui City, Gansu Province, respectively. For this experiment, the locations of first
sample collection in each season in Qinghai and Gansu were set as the simulated central fungus source sites, with a tracking time of 5 d.

2.7 Pipeline integration
To facilitate analysis by users who are unfamiliar with bioinformatics methods, a locus subsampling and SNP calling pipeline was integrated into the sequencing pipeline of MolBreeding Biotechnology Co., Ltd. (Shijiazhuang, Hebei, China).

3 RESULTS

3.1 Variant validation
Sequencing of the 225 *Pst* samples with the *Pst* 20K GBTS chip yielded 361.9 Gb of raw data, corresponding to 323.9 Gb of clean reads. The effective rate (the ratio of filtered clean data to raw data) was 80.78–94.79% and the mapping rate to the primary contig of the *Pst*-134E reference genome ranged from 45.88–95.89%, averaging 85.12% per sample (Table S2). After processing, there were 1,293,150 SNPs in the raw VCF output with a maximum gap of 99,512 bp; only 40 regions were larger than 20,000 bp (Figure 3a; Table 1). Of the mutation regions examined, the highest proportion of variants (up to 33.63%) was found in the upstream region, followed closely by the downstream region (33.41%). This demonstrated a minimal difference in SNP number between the upstream and downstream regions. SNPs in intergenic and exon regions accounted for 20.80% and 7.85%, respectively; other regions exhibited fewer variants (Figure 3b; Figure S1a). To measure genetic diversity and high-resolution differences between isolates using high-density molecular markers, SNPs in the CDS region were annotated and statistically analyzed. Among the variants, non-synonymous single nucleotide variants (SNVs) and synonymous SNVs exhibit the highest proportion, whereas stop gain and stop loss variants constitute the lowest proportion (Figure S1b). To ensure the accuracy of further analyses, loci were removed if they had more than two alleles, if the minimum depth of total or alternative allele read coverage was below 3, or if > 25% of the isolates had missing data for the region of interest. A total of 105,542 SNPs passed this quality control (QC) step. For population structure, phylogenetic, and other genetic analyses, loci with missing values or an MAF < 0.05 were removed, yielding 19,139 SNPs. In addition to these SNP loci, 90 published SSR markers were successfully genotyped with the chip (Table S3), enabling comparisons of these data to previous findings. Having validated this approach and the tested markers, the new tool was named the *Pst* 20K GBTS chip.

3.2 Comprehensive cost analysis
The per-sample expenses for the *Pst* 20K GBTS chip amounted to ~15 US dollars, encompassing DNA extraction (~$0.8), library construction (~$5), probe hybridization (~$4), sequencing (~$3), bioinformatics analysis (~$0.5), labor (~$1.5), and depreciation costs (~$0.2). The cost per sample of SSR marker is ~1 US dollars (Sangon, Shanghai, China). The cost per sample per KASP marker is ~0.2 US dollars. Thus, in comparison to PCR markers, the cost of sequencing for a single marker was reduced by at least three orders of magnitude. In addition, compared to high-throughput resequencing, the costs of this method were similar, and the acquired data were of comparable quality and quantity. However, the chip method reduced the time and human resources required for spore reproduction. These results validated the feasibility of the *Pst* 20K GBTS chip method for high-throughput population-scale *Pst* genotyping.

3.3 *Pst* exchange dynamics in the northwest oversummering region
After validating the quality and cost-effectiveness of the *Pst* 20K GBTS chip, analyses of the *Pst* population structure and genetic diversity were conducted to reveal the dynamics of pathogen exchange in the northwest oversummering region. Population structural analysis indicated that the *Pst* populations in this region in 2021 and 2022 were optimally represented at K = 3 (Figure S2a). At K = 2, the isolates from Qinghai, Gansu, and Ningxia were divided into two groups distributed across the six populations (Figure 4a). At K = 3, Group 1 (which contained 153 isolates) was the dominant cluster, followed by Group 2 (48 isolates), then
Group 3 (24 isolates) (Table S4). Notably, STRUCTURE results exhibited striking similarities in population structure among the Qinghai, Gansu, and Ningxia samples across both years. Moreover, the population from the Longdong area of 22GS showed genetic divergence from the other 22GS samples, but was consistent with the structure of 21NX, suggesting a potentially close relationship between 22GS samples from the Longdong area and 21NX samples. Then, we increased the value of K to 9. The same result of two different ranges of K is that Delta K supports the optimal number of groups is 3 (Figure S2a, b), and Group 3 (yellow bars) remains independent (Figure S2c).

DAPC analysis reinforced the findings of STRUCTURE, categorizing all genotypes into three groups (Figure 4b). Group 1 represented a significant portion of the six populations and emerged as the primary component. Group 2 was most abundant in the 21GS, 22GS, and 21QH populations. Group 3 was primarily found in the 21NX, 22QH, and 22GS populations, but was also present at small proportions in the 21GS and 21QH populations. In 2022, the Ningxia population displayed a notable shift; it retained genotypes from only Group 1 in 2021 and lost the Group 2 and Group 3 genotypes, indicating a significantly altered genetic landscape (Figure S3).

Phylogenetic analysis validated the divisions between populations, yielding three lineages (Group 1, Group 2, and Group 3) (Figure 4c). This analysis highlighted the close genetic relationships between isolates from Group 1 and Group 2, underscoring their separation from those in Group 3, consistent with the DAPC results. Despite the presence of just three primary clusters across multiple analyses, each group exhibited internal subdivisions, indicating smaller-scale genetic exchanges among isolates in each of the three epidemic regions. Finally, genetic relationships were visualized with a kinship heatmap, which corroborated division of the 225 isolates into three groups (Group 1, Group 2, and Group 3). Thus, the results of STRUCTURE, DAPC, phylogenetic, and kinship analyses emphasized the dynamic nature of pathogen exchange throughout the northwest oversummering region and supported the existence of three main genotypic groups (Figure 4b).

### 3.4 Genetic diversity of Pst populations

The Pst 20K GBTS chip was used to closely examine all 225 isolates, identifying unique multi-locus genotypes (MLGs) for each individual. The overall populations were stratified into distinct MLGs, corresponding to the collection sites in each year: 21GS (44 MLGs), 21NX (19), 21QH (71), 22GS (57), 22NX (7), and 22QH (27). After post-clonal sample size correction, the 22NX population exhibited the lowest expected MLG (eMLG) count at 7, whereas the other five populations all had eMLGs of 10. Genetic diversity analyses underscored the remarkable polymorphism within the SNP sites (ranging from 96.75–100.00%), affirming the comprehensive polymorphic coverage of the chip. All six populations displayed high observed allele numbers, averaging 1.976. Remarkably, the 21GS, 21QH, and 22GS populations all had the highest observed number of alleles ($N_a = 2.00$), whereas the 22NX population had the lowest ($N_a = 1.868$). Similar patterns were found for the effective allele numbers, with 21NX exhibiting the highest value ($N_e = 1.665$) and 22NX having the lowest ($N_e = 1.547$).

The Shannon diversity index ($I$) showcased variations between the populations; diversity was highest in the 21NX population ($I = 0.556$), followed by the 21QH population ($I = 0.542$), then the 22NX population ($I = 0.466$). $H_o$ surpassed $H_e$ in all populations, with the 21NX population showing the highest $H_o$(0.522) and $H_e$ (0.378). Heterozygosity was lowest in the 22NX population ($H_o = 0.445$, $H_e = 0.313$) (Table 2). Overall, these results demonstrated high genetic diversity in the 21NX population and low genetic diversity in the 22NX population (post-winter), consistent with the DAPC results.

### 3.5 Frequent gene flow between populations

$F_{st}$ values were calculated for the six populations and ranged from 0.005–0.053. AMOVA indicated that the differences between sample sites and sampling times were statistically significant ($p< 0.05$) (Table 3, Table S5). $D$ values ranged from 0.006–0.061, further demonstrating a lack of genetic divergence among these populations. Gene flow ($N_m$) values ranged from 0.06–1, indicating highly variable levels of genetic exchange (Figure 5). In 2021, robust gene exchange was noted across all three populations, particularly
between the 21QH and 21GS populations. Conversely, genetic relationships between the Ningxia and Gansu populations remained relatively weak in 2021, despite the geographical adjacency of these locations. This intriguing pattern indicated a unique link between the Qinghai and Ningxia populations, consistent with the STRUCTURE results.

Although substantial genetic exchange persisted among the populations in 2022, these interactions were significantly diminished compared to 2021. There was clear proximity between the 22QH and 22GS populations, whereas both of these populations had more distant relationships with the 22NX population. Notably, five of the populations (all except 22NX) exhibited robust gene flow. The affinity between the 21QH and 22GS populations surpassed the affinity between the 21GS and 22QH populations, suggesting a significant contribution from 21QH to the 22GS population in the post-summer period (Figure 5). In summary, during 2021 and 2022, populations in the northwest China epidemic regions underwent frequent gene exchange, with the Qinghai and Gansu populations demonstrating a closer genetic relationship than either shared with the Ningxia population. These results provide valuable insights into the intricate dynamics of *Pst* dispersal.

### 3.6 Wind movement and trajectory models of the *Pst* migration route

The general trend of high-altitude wind movement in the northwest oversummering region from September to December 2021 was from west to east (Figure 6a). Theoretically, this would allow *Pst* to spread from Qinghai Province to Gansu Province and to Ningxia Autonomous Region with the airflow. A HYSPLIT-4 trajectory model showed movement from the central fungus source (in Salar Autonomous County, Xunhua, Haidong City, Qinghai Province) to the wheat growing areas of Gansu Province and Ningxia Autonomous Region (probability > 10%). There was also a high probability that spores from the central source would spread via airflow to the Huang-huai-hai wheat area (Figure 6b). From April to July 2022, there were two general wind trends in western and central Qinghai: one from Xinjiang to Qinghai Province and the other from Guangyuan City, Sichuan Province to eastern Qinghai via Gansu Province (Figure 6c). A trajectory model for this period showed a likely spread from the central fungus source in Maiji District, Tianshui City, Gansu Province from east to west, covering part of Ningxia Autonomous Region and the eastern wheat area of Qinghai Province (probability > 30%) (Figure 6d). Thus, the upper air wind field data in these two periods were highly consistent with the results of the HYSPLIT-4 trajectory models.

### 3.7 Temporal dynamics of the wheat stripe rust epidemic in northwestern China

To unravel the epidemic pattern of *Pst* in the northwest oversummering region of China, field surveys were conducted for the three sampling regions in the 2021 and 2022 cropping seasons. Observations from 2021 revealed a notable contrast in disease occurrence between Gansu and Qinghai provinces. Gansu experienced earlier disease occurrence than Qinghai, beginning after the wheat harvesting period from August to October. Interestingly, fields in Gansu exhibited a lower incidence of stripe rust in volunteer seedlings and a subsequent decrease in disease prevalence. Conversely, Qinghai Province sustained more spontaneous seedlings, preserving a significant reservoir of urediniospores. Consequently, disease occurrence among winter wheat in both Qinghai and Gansu transpired synchronously in mid-November of 2021, whereas disease onset lagged in Ningxia winter wheat, beginning in late November. In 2022, the temporal progression of the wheat stripe rust epidemic displayed a dynamic pattern. It emerged first in Longnan City in mid-May, gradually spread to Tianshui City and Longdong area in late May, then progressed to Linxia City in mid-June. Disease onset did not begin in Qinghai and Ningxia until early June, marking a significant shift in the regional landscape of the epidemic (Table S6). These findings underscore the nuanced interplay between environmental factors and pathogen dynamics, shedding light on the complex epidemiological factors contributing to *Pst* infections in this region.

### 4 DISCUSSION

Over the past three decades, methodologies used in research of *Pst* populations have undergone a profound evolution, progressing from RFLP (Shan et al., 1998) to PCR-based DNA fingerprinting (Mardis, 2008;
Welsh & McClelland, 1990; Williams et al., 1990), and finally to advanced techniques such as DNA sequencing (Chen, 2005; Hovmöller et al., 2011). Next-generation sequencing technology (Ding et al., 2021; Li et al., 2023) and SNP detection (Xia et al., 2016a) in particular revolutionized efforts to characterize genetic diversity in this pathogen. As molecular detection techniques continually evolve, a plethora of molecular markers and detection approaches have kept pace with technological advances. However, efforts to comprehensively understand wheat stripe rust dynamics have outstripped the capabilities of traditional markers. Although genome-wide sequencing offers comprehensive insights into genomic variations, the applicability of this approach is limited by its high costs, especially for large-scale sample analyses. Furthermore, traditional genotyping methods that rely on PCR markers or next-generation sequencing of \( Pst \) necessitate a four-time spore multiplication process, which takes approximately three months.

To fulfill the unmet need for a high-throughput, low-cost \( Pst \) genotyping solution, we here leveraged GBTS technology to develop the \( Pst \) 20K GBTS chip. This innovative chip includes various molecular marker types, including SNPs, SSRs, and InDelbs, providing an extensive repertoire of markers accessible through solution capture. The \( Pst \) 20K GBTS chip allows for direct use of diseased leaves collected from the field, saving a great deal of time and personnel costs. Overall, the chip is significantly less expensive per sample than PCR-based methods. Thus, the \( Pst \) 20K GBTS chip allows identification of a rich spectrum of genomic variations in stripe rust while boasting high coverage, precision, and efficiency. Importantly, the analysis pipeline was provided to the MolBreeding Biotechnology company and is publicly accessible at https://github.com/zengqd/PopulationGenetics/tree/main/Fungi/PstGTBSChip/Step1_1fq2wcf.sh. The pipeline requires only simple genotype data as the input and includes filtering processes and standard genetic analyses, such as PCA, kinship analysis, and phylogenetic tree construction. This enables broad use of our method by researchers without expertise in bioinformatics.

Using the \( Pst \) 20K GBTS chip, we conducted a pioneering genetic analysis of \( Pst \) interactions within and between stripe rust populations in the oversummering region of China, shedding light on the population dynamics of this important pathogen. The genetic exchanges observed between Qinghai, Gansu, and Ningxia \( Pst \) populations underscore the complex environmental factors that influence the spread of wind-borne spores (Chen et al., 2014; Kong et al., 2014; Yao et al., 2014). Group differentiation, gene flow, and phylogenetic analyses were performed for the different populations in each province, revealing the dynamics of pathogen spread. Notably, the presence of \( Pst \) genetic Group 3 (which appeared to have originated in Ningxia) in Qinghai and Gansu suggested potential cross-regional transmission. Our results thus warrant further investigation into the origins and ecological implications of Group 3.

Extensive field investigations spanning Qinghai and Gansu provinces and Ningxia Autonomous Region during 2021 and 2022 yielded intriguing insights into the survival strategies of wheat stripe rust in the face of changing ecological dynamics. Our in-depth analysis of wheat stripe rust epidemiology in the northwest oversummering region of China represents a departure from previous reports. Longnan City, Gansu Province has historically been identified as a pivotal hotspot for generating and maintaining variations in \( Pst \) virulence, although wheat stripe rust consistently encroaches upon the entire wheat planting area (Chen et al., 2013). Our investigation from June 2021 to August 2022 revealed shifting patterns in epidemic dynamics. In 2021, disease prevalence followed a distinctive order: Qinghai exhibited the highest proportion of agricultural field plots affected by \( Pst \) (84.04%), followed by Gansu (75.34%), then Ningxia (71.43%). Notably, the disease prevalence rate in Qinghai Province ranged from 5–50%, slightly surpassing Gansu (5–40%) and significantly outpacing Ningxia (1–20%) (Table S6). Moreover, field disease severity and prevalence rates peaked in Qinghai, highlighting the unique susceptibility of this region. Wheat stripe rust is quite common in Qinghai, and the amount of urediospores was no less than that in Gansu and Ningxia. Unexpected \( Pst \) resilience was observed in Qinghai, likely due to the presence of late-maturing spring wheat and abundant volunteer wheat, which provide a conducive environment for \( Pst \) post-harvest survival. This contrasts sharply with the declining suitability of fields in Gansu and Ningxia for \( Pst \) post-harvest survival due to climatic shifts and agricultural strategies (Chen et al., 2013). This marked shift in host dynamics redefines our understanding of \( Pst \) epidemiology, with significant implications for crop management and resistance strategies.
Our dissection of stripe rust population genetics across diverse regions lays the foundation for strategic crop planning and disease control. Moreover, delineating Pst adaptability to shifting climate patterns provides a valuable strategy for future research, highlighting the necessity of comparable enhancements to host adaptability amid global climate fluctuations. The candidate secretion proteins encoded by genes included in the Pst 20K GBTS chip are expected to play important roles in future Pst genotype and virulence spectrum correlation analyses, facilitating screening and characterization of avirulence genes (Xia et al., 2016b). Thus, our study has significant molecular ecological implications, offering novel insights into wheat stripe rust epidemiology, host–pathogen interactions, and the impacts of the changing climate on disease dynamics.

Crucially, our study underscores the role of wind in Pst transmission. High-altitude wind mapping and trajectory analyses unveiled important prevailing wind patterns, these results suggested potential long-distance dispersal routes, particularly from Xinjiang to the northwest oversummering region. Detection of Group 3 pathogens in disparate locations spanning Qinghai, Ningxia, and the Longdong area of Gansu suggested intriguing cross-regional connections, possibly originating from Xinjiang. These findings illuminate Pst transmission dynamics and underscore an urgent need for in-depth experimental validation of these mechanisms of pathogen spread. Overall, this study provides not only a valuable tool for future population-scale genetic analyses of Pst, but reveals key factors that apparently contribute to the spread and success of this economically important pathogen.

**AUTHOR CONTRIBUTIONS**
Z. Kang, L. Huang, and Q. Zeng designed the research; H. Yan, Z. Ma, and Q. Yao performed experiments; S. Cao, Q. Jia, J. Li, J. Zhao, W. Yan, J. Ma, W. Chen, and D. Han collected samples; Q. Zeng, H. Yan, and Z. Ma analyzed the data; Q. Zeng, H. Yan, and Z. Ma wrote the manuscript.

**ACKNOWLEDGEMENTS**
We thank Dr. Guoliang Pei (State Key Laboratory of Crop Stress Biology for Arid Areas) and the HPC of Northwest A&F University for technical server support. This study was financially supported by the National Key R&D Program of China (2021YFD1401000), the Major Program of the National Natural Science Foundation of China (32293240), the Innovation Capability Support Program of Shaanxi (Program No. 2023-CX-TD-56), and the earmarked fund for CARS (CARS-3).

**CONFLICT OF INTEREST STATEMENT**
The authors declare no competing interests.

**DATA ACCESSIBILITY STATEMENT**
Raw sequence reads are deposited in the National Genomics Data Center (NGDC) (BioProjectPRJCA020812).

**BENEFIT-SHARING STATEMENT**
Benefits from this research accrue from the sharing of our data and results on public databases as described above.

**REFERENCES**


**TABLES**

**TABLE 1** Chromosomal distribution of loci included in the *Pst* 20K chip.

<table>
<thead>
<tr>
<th>Chr.</th>
<th>Chr. length</th>
<th>Variant</th>
<th>Variant rate (%)</th>
<th>SNP</th>
<th>Average MAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5,737,078</td>
<td>92,009</td>
<td>62.35</td>
<td>91,130</td>
<td>0.099363135</td>
</tr>
<tr>
<td>2</td>
<td>5,337,683</td>
<td>92,297</td>
<td>60.83</td>
<td>93,524</td>
<td>0.092841351</td>
</tr>
<tr>
<td>3</td>
<td>5,653,929</td>
<td>94,238</td>
<td>61.28</td>
<td>94,614</td>
<td>0.099920464</td>
</tr>
<tr>
<td>4</td>
<td>5,531,554</td>
<td>90,970</td>
<td>60.81</td>
<td>93,358</td>
<td>0.099294646</td>
</tr>
<tr>
<td>5</td>
<td>5,139,519</td>
<td>78,322</td>
<td>65.62</td>
<td>77,594</td>
<td>0.11268899</td>
</tr>
<tr>
<td>6</td>
<td>4,650,038</td>
<td>75,879</td>
<td>61.28</td>
<td>75,204</td>
<td>0.09899443</td>
</tr>
<tr>
<td>7</td>
<td>4,836,860</td>
<td>83,630</td>
<td>57.84</td>
<td>83,513</td>
<td>0.098007027</td>
</tr>
<tr>
<td>8</td>
<td>5,337,734</td>
<td>91,808</td>
<td>60.00</td>
<td>94,701</td>
<td>0.107405409</td>
</tr>
<tr>
<td>9</td>
<td>4,507,480</td>
<td>56,426</td>
<td>79.88</td>
<td>68,210</td>
<td>0.096974223</td>
</tr>
<tr>
<td>10</td>
<td>4,254,770</td>
<td>64,382</td>
<td>66.09</td>
<td>65,959</td>
<td>0.10326453</td>
</tr>
</tbody>
</table>
Chr., chromosome; SNP, single nucleotide polymorphism; MAF, minor allele frequency.

### TABLE 2 Genetic diversity among six populations of *Puccinia striiformis* f. sp. *tritici*.

<table>
<thead>
<tr>
<th>Chr, length</th>
<th>Variant</th>
<th>Variant rate (%)</th>
<th>SNP</th>
<th>Average MAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>4,647,190</td>
<td>80,717</td>
<td>57.57</td>
<td>79,931</td>
</tr>
<tr>
<td>12</td>
<td>4,195,100</td>
<td>64,989</td>
<td>64.55</td>
<td>69,159</td>
</tr>
<tr>
<td>13</td>
<td>3,366,535</td>
<td>59,566</td>
<td>56.52</td>
<td>59,059</td>
</tr>
<tr>
<td>14</td>
<td>3,085,245</td>
<td>49,520</td>
<td>62.30</td>
<td>51,051</td>
</tr>
<tr>
<td>15</td>
<td>2,708,085</td>
<td>43,656</td>
<td>62.03</td>
<td>46,439</td>
</tr>
<tr>
<td>16</td>
<td>2,994,770</td>
<td>49,227</td>
<td>60.84</td>
<td>49,209</td>
</tr>
<tr>
<td>17</td>
<td>3,479,076</td>
<td>46,009</td>
<td>75.62</td>
<td>52,062</td>
</tr>
<tr>
<td>18</td>
<td>2,717,725</td>
<td>48,623</td>
<td>55.89</td>
<td>48,799</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>78,180,371</strong></td>
<td><strong>1,305,032</strong></td>
<td><strong>59.91</strong></td>
<td><strong>1,293,150</strong></td>
</tr>
</tbody>
</table>

Chr., chromosome; SNP, single nucleotide polymorphism; MAF, minor allele frequency.

### TABLE 3 Nei’s genetic distance (D) and fixation index (Fst) for each pairwise comparison of *Puccinia striiformis* f. sp. *Tritici* populations.

<table>
<thead>
<tr>
<th>Year</th>
<th>Province</th>
<th>GS</th>
<th>NX</th>
<th>QH</th>
<th>GS</th>
<th>NX</th>
<th>QH</th>
</tr>
</thead>
<tbody>
<tr>
<td>2021</td>
<td>GS</td>
<td>*</td>
<td>0.021</td>
<td>0.006</td>
<td>0.011</td>
<td>0.061</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td>NX</td>
<td>0.011</td>
<td>*</td>
<td>0.014</td>
<td>0.011</td>
<td>0.040</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>QH</td>
<td>0.005</td>
<td>0.009</td>
<td>*</td>
<td>0.008</td>
<td>0.055</td>
<td>0.011</td>
</tr>
<tr>
<td>2022</td>
<td>GS</td>
<td>0.007</td>
<td>0.014</td>
<td>0.007</td>
<td>*</td>
<td>0.058</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>NX</td>
<td>0.047</td>
<td>0.051</td>
<td>0.050</td>
<td>0.052</td>
<td>*</td>
<td>0.061</td>
</tr>
<tr>
<td></td>
<td>QH</td>
<td>0.011</td>
<td>0.020</td>
<td>0.011</td>
<td>0.013</td>
<td>0.053</td>
<td>*</td>
</tr>
</tbody>
</table>

GS, Gansu; NX, Ningxia; QH, Qinghai.

### FIGURE LEGENDS

**FIGURE 1** Application of the genotyping by target sequencing (GBTS) chip to genotyping of *Puccinia striiformis* f. sp. *tritici*.

**FIGURE 2** Wheat stripe rust isolate sampling sites.(a) Geographical location of the northwest
oversummering region in China, comprising Qinghai Province, Gansu Province, and Ningxia Autonomous Region. (b) Elevation map of the northwest oversummering region. (c) Locations of *Puccinia striiformis* f. sp. *tritici* isolate collection in 2021 and 2022 and distribution of wheat planting areas in the northwest oversummering region.

**FIGURE 3** Single nucleotide polymorphism (SNP) data from the genotyping by target sequencing (GBTS) chip. (a) Distribution of SNP markers within a 10-kb window with 18 *Puccinia striiformis* f. sp. *tritici* chromosomes serving as marker panels. Marker density is represented with color. Each bar represents a 10-kb window size. (b) SNP marker annotations. (c) Distribution of SNP types.

**FIGURE 4** Population genetic structure of 225 *Puccinia striiformis* f. sp. *tritici* (*Pst*) isolates collected from the northwest oversummering region of China. (a) Assignment of isolates to genotype clusters using STRUCTURE. Each line represents an isolate and was partitioned into one of K clusters. 21GS, 2021 Gansu; 21NX, 2021 Ningxia; 21QH, 2021 Qinghai; 22GS, 2022 Gansu; 22NX, 2022 Ningxia; 22QH, 2022 Qinghai. (b) Discriminant analysis of principal components (DAPC). (c) Neighbor-joining phylogenetic analysis of the 225 *Pst* isolates using single-nucleotide polymorphisms (SNPs) filtered by linkage disequilibrium (LD). (d) Kinship matrix based on simple matching of genetic similarities.

**FIGURE 5** Migration network showing *Puccinia striiformis* f. sp. *tritici* gene flow (*Nm*) patterns within epidemic regions of China. Line thickness indicates *Nm* intensity. 21GS, 2021 Gansu; 21NX, 2021 Ningxia; 21QH, 2021 Qinghai; 22GS, 2022 Gansu; 22NX, 2022 Ningxia; 22QH, 2022 Qinghai.

**FIGURE 6** Inference of the direction of wheat stripe rust spread based on meteorological data from the northwest oversummering region of China. (a) Upper air wind field map from September–December 2021. S, strong; W, weak. Arrow color corresponds to wind intensity. (b) Inference of the spore movement trajectory in September 2021. Salar Autonomous County, Xunhua, Qinghai Province was designated the central point. (c) Upper air field map from April–July 2022. (d) Inference of the spore movement trajectory in May 2022. Maiji District, Gansu Province was designated the central point.

**SUPPLEMENTAL INFORMATION**

**FIGURE S1** Single nucleotide polymorphism (SNP) distribution and classification. (a) Genic distribution of SNPs. (b) Distribution of SNP types among those within the coding sequence (CDS).

**FIGURE S2** Assignment of isolates to genotype clusters using STRUCTURE. (a) Delta K (K = 2–6) distribution based on the Evanno method. The maximum Delta K value corresponds to the optimal cluster number. (b) Delta K (K = 2–9) value distribution. (c) Predicted isolate distribution among K clusters (represented by color). 21GS, 2021 Gansu; 21NX, 2021 Ningxia; 21QH, 2021 Qinghai; 22GS, 2022 Gansu; 22NX, 2022 Ningxia; 22QH, 2022 Qinghai.

**FIGURE S3** Cluster distribution within the six *Puccinia striiformis* f. sp. *tritici* populations. Isolates were divided with discriminant analysis of principal components (DAPC).

**TABLE S1** Information of the 225 *Puccinia striiformis* f. sp. *tritici* isolates collected for this study.

**TABLE S2** Genotyping by target sequencing (GBTS) chip sequencing results for the 225 *Puccinia striiformis* f. sp. *tritici* isolates.

**TABLE S3** Simple sequence repeat (SSR) markers used in genotyping by target sequencing.

**TABLE S4** Distribution of *Puccinia striiformis* f. sp. *tritici* isolates into genetic groups. Samples were collected from two provinces and one autonomous region during two years. Clustering was performed at K = 3 with STRUCTURE.

**TABLE S5** Analyses of molecular variance among six populations of *Puccinia striiformis* f. sp. *tritici* collected from two provinces and one autonomous region in 2021 and 2022.
TABLE S6 Field investigation statistics for *Puccinia striiformis* f. sp. *tritici* isolates collected from two provinces and one autonomous region in 2021 and 2022.
The Pst GBTS Chip pipeline for this study

FIGURE 1 Application of the genotyping by target sequencing (GBTS) chip to genotyping of Puccinia striiformis f. sp. tritici.
FIGURE 2 Wheat stripe rust isolate sampling sites.
FIGURE 3 Single nucleotide polymorphism (SNP) data from the genotyping by target sequencing (GBTS) chip.
FIGURE 4 Population genetic structure of 225 *Puccinia striiformis* f. sp. *tritici* (*Pst*) isolates collected from the northwest oversummering region of China.
FIGURE 5 Migration network showing *Puccinia striiformis* f. sp. *tritici* gene flow (*Nm*) patterns within epidemic regions of China.
FIGURE 6 Inference of the direction of wheat stripe rust spread based on meteorological data from the northwest oversummering region of China.