Secreted protein gene derived-single nucleotide polymorphisms (SP-SNPs) reveal population diversity and differentiation of *Puccinia striiformis* f. sp. *tritici* in the United States

Chongjing XIA^a, Anmin WAN^a, Meinan WANG^a, Derick A. JIWAN^a, Deven R. SEE^{a,b},
Xianming CHEN^{a,b,*}

^aDepartment of Plant Pathology, Washington State University, Pullman, WA 99164-6430, USA

^bUS Department of Agriculture, Agricultural Research Service, Wheat Health, Genetics, and Quality Research Unit, Pullman, WA 99164-6430, USA

*Corresponding author: Xianming Chen, Tel: +1 509-335-8086; Fax: (509)-335-9581
E-mail addresses: chongjing.xia@wsu.edu (C.J. Xia); awan@wsu.edu (A.M. Wan);
meinan_wang@wsu.edu (M.N. Wang); djiwan@wsu.edu (D.A. Jiwan);
deven_see@wsu.edu (D.R. See) and xianming@wsu.edu (X.M. Chen).

Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture. USDA is an equal opportunity provider and employer.

ABSTRACT

Single nucleotide polymorphism (SNP) is a powerful molecular marker technique that has been widely used in population genetics and molecular mapping studies for various organisms. However, the technique has not been used for studying *Puccinia striiformis* f. sp. *tritici* (*Pst*), the causal agent of stripe rust, one of the most destructive pathogens in wheat production worldwide, due to the lack of SNP markers. In the present study, we developed over a hundred secreted protein gene-derived SNP (SP-SNP) markers and used 92 markers to study the population structure of the fungal pathogen. From 352 *Pst* isolates collected over all stripe rust epidemiological regions in the United States, we identified 242 multi-locus genotypes. The SP-SNP genotypes had a moderate, but significant correlation ($r = 0.34$; $P < 0.0001$) with the virulence phenotype data. Clustering of the multi-locus genotypes was consistent by multivariate discriminant analysis of principal components, neighbor-joining method, Bayesian clustering algorithm and cluster matching and permutation program, revealing distinct genetic groups. Analysis of molecular variance (AMOVA) detected significant differences between the eastern and western US *Pst* populations. High heterozygosity was found in the US population with significant differences identified among epidemiological regions. Analysis of population differentiation revealed that populations between the eastern and western US were highly differentiated while moderate differentiation was found in populations within the western or eastern US. Isolates from the western US were more diverse than isolates from the eastern US. The information is useful for guiding the disease management in different epidemiological regions.

Keywords: heterozygosity, population diversity, population differentiation, *Puccinia striiformis* f. sp. *tritici*, single nucleotide polymorphism, wheat stripe rust

Introduction

Newly introduced and re-emerged plant pathogens can cause severe epidemics in agro-ecosystems, resulting in significant yield losses (Stukenbrock & McDonald 2008). Among plant pathogens, *Puccinia striiformis* Westend. f. sp. *tritici* Erikss. (*Pst*) is notorious as a highly destructive pathogen that can be long-distance dispersed on both national and continental scales and can damage cereal crops in large production areas (Brown & Hovmøller 2002). *Pst* causes wheat stripe rust (also called yellow rust) throughout the wheat-production regions of the world (Stubbs 1985; Chen 2005; Wellings 2011). In the United States (US), the disease is historically destructive in the regions west of the Rocky Mountains (western US), but has become increasingly important also in the eastern US (Chen 2005; Chen *et al.* 2010). The heterokaryotic fungal pathogen is highly diverse for virulence to stripe rust resistance genes in wheat, and *Pst* populations have continued to evolve (Chen *et al.* 2010). Annually, more than 20 races are identified from several hundred *Pst* isolates based on virulence and avirulence patterns on a set of wheat genotypes with different resistance genes (Chen 2005; Chen *et al.* 2010; Wan & Chen 2012, 2014; Wan *et al.* 2016). *Pst* races with virulence to newly deployed resistance genes are often observed and detected within few years after the release of wheat cultivars with race specific resistance. Genetic

variation behind the virulence diversity in *Pst* populations is shaped by different mechanisms, including mutation, introduction, host and environment selection and genetic recombination (Little & Manners 1969; Johnson & Taylor 1976; Wellings & McIntosh 1990; Hovmøller *et al.* 2002, 2008, 2011, 2015; Park & Wellings 2012). The alternate hosts of *Pst*, *Berberis* spp. and *Mahonia* spp., had been identified under controlled conditions (Jin *et al.* 2010; Wang & Chen 2013). However, the low infection rate of *Pst* on *Berberis* and the short dormancy and rapid germination of teliospores under natural conditions indicate that sexual reproduction is likely infrequent (Zhao *et al.* 2013). In fact, no *Pst* aecia were found on barberry plants, and barberry does not function as an alternate host for *Pst* in the US Pacific Northwest due to teliospore degradation and barberry phenology (Wang & Chen 2015; Wang *et al.* 2015). Therefore, *Pst* is considered to reproduce mostly by the clonal reproduction of urediniospores, cycling primarily on wheat and grasses (Cheng & Chen 2014; Cheng *et al.* 2016).

The US is delimited into twelve wheat stripe rust epidemiological regions based on the geographical barriers, prevailing winds and other weather conditions, crop systems and occurrence of different races of *Pst* (Line & Qayoum 1992; Chen *et al.* 2010). The Rocky Mountains range serves as a major geographical barrier limiting gene flow between the western and eastern US populations. Wheat stripe rust causes more frequent epidemics in the regions west of the Rocky Mountains because of the mild winters and cool springs which are conducive for disease development, but are generally less frequent in the regions east of Rocky Mountains. Since 2000, however, the disease has become relatively frequent compared to the years prior to 2000 (Chen *et al.* 2002; Chen 2005; Chen *et al.* 2010; Wan & Chen 2012, 2014; Wan *et al.* 2016).

The differentiation of *Pst* populations among regions in the eastern and western US is demonstrated in terms of disease severity, yield loss and number of races detected every year. The selection for virulence by growing wheat cultivars with race specific resistance genes in each region increases the race diversity (Chen *et al.* 1993; Chen *et al.* 2010; Wan & Chen 2012, 2014; Wan *et al.* 2016). However, the genetic diversity and differentiation of *Pst* populations need to be quantified. The paucity of knowledge about the genetic differentiation among epidemiological regions may limit our capacities to deploy proper resistance genes.

A number of molecular polymorphism techniques have been used to study the genetic diversity of *Pst* populations. Conventionally, dominant markers such as random amplified polymorphic DNA (RAPD) and amplified fragment length polymorphism (AFLP) were used to characterize *Pst* populations in North America and Europe (Chen *et al.* 1993; Justesen *et al.* 2002). Later, the development of microsatellite/simple sequence repeat (SSR) markers for *Pst* have enabled to determine multi-locus dikaryotic genotypes for population studies of this fungus (Enjalbert *et al.* 2002; Bahri *et al.* 2009; Chen *et al.* 2009; Cheng *et al.* 2012; Bailey *et al.* 2013; Luo *et al.* 2015; Zhan *et al.* 2015). Co-dominant markers, such as SSR, are more informative, especially for dikaryotic rust fungi like *Pst*, because they can detect heterozygosity of polymorphic loci. More recently, the single nucleotide polymorphisms (SNPs) showed a promising potential for applications in population genetics and disease-causal allele identification because of its high numbers, genome-wide coverage, co-dominance and suitability for use of high-throughput technologies (Morin *et al.* 2009; Kirk & Freeland 2011). It was estimated that around 30 SNPs would be sufficient to detect a moderate level of

differentiation ($F_{ST} = 0.01$) in population diversity studies, and 80 or more for determining demographic independence ($F_{ST} = 0.005$) (Morin *et al.* 2009).

Planting resistant wheat cultivars is the most economically and environmentally friendly strategy for sustainable management of stripe rust. Our goal of research is to provide useful information to develop appropriate strategies for the management of stripe rust in various epidemiological regions. Studying *Pst* population structure will provide valuable knowledge for deploying various resistance genes in wheat cultivars against the pathogen populations in different regions. As many avirulence genes in filamentous fungi have been found to encode secreted proteins (SP) (Catanzariti *et al.* 2006; Kamoun 2007; Petre *et al.* 2014), markers developed from SP genes may indicate virulence changes in the pathogen population. In a separate study, we found significant associations of 30 SP loci with 9 avirulence genes in the *Pst* fungus (Xia *et al.* 2016). The specific objectives of this study were to 1) develop SP-SNPs markers for studying the *Pst* fungus; 2) quantify the genetic diversity in the recent *Pst* populations of various epidemiological regions in the US and the differentiation between the populations of the eastern and western US; 3) determine migration levels among different epidemiological regions of *Pst* populations; and 4) estimate the heterozygosity of *Pst* based on SP-SNP loci.

Materials and Methods

Isolate selection and urediniospore multiplication

The isolates used in this study were selected from the *Pst* collections made in 2010 and 2011 when wheat stripe rust was widespread and caused severe epidemics in the US (Wan & Chen 2012, 2014; Chen 2014; Wan *et al.* 2016). A total of 352 *Pst* isolates were selected to represent all virulence races detected in 2010 and 2011, and to cover all epidemiological regions in the US (Table 1; Fig 1). Urediniospores were inoculated on susceptible wheat 'Nugaines' for multiplication, as described previously (Wan & Chen 2014). Virulence phenotyping was conducted by inoculating isolates on the current set of 18 *Yr* single-gene lines of wheat for identifying *Pst* races, as described by Wan & Chen (2014).

DNA extraction

For each isolate, 10 mg dried urediniospores were used for genomic DNA extraction. A mixture of urediniospores and 50 μ L 0.1 mm sterile glass beads in each tube was shaken for 2 min in Mini Beadbeater (Biospec products, Inc.), followed by a modified CTAB procedure as described by Chen *et al.* (1993). The DNA concentration in the stock solution was qualified by Nanodrop 1000 (Thermo Fisher Scientific Inc., MA, USA). A working solution of 5 ng/ μ L was made by diluting the stock solution with sterilized ddH₂O and used for polymerase chain reactions (PCR). The quality of the DNA working solution were further checked by PCR amplification with the internal transcribed spacer (ITS) primers ITS3puc (5'-ACATCGATGAAGAACACAGT-3') (Wang *et al.* 2015) and ITS4B (Gardes & Bruns 1993) for all 352 isolates. The PCR amplicons were checked on 0.8% agarose gel. The 304 bp target bands should be clearly present for each isolate. If

not, the stock solution was re-purified and/or working solution was re-made and checked before SP-SNP genotyping.

SP-SNP marker development

A total of 442 *Pst* expressed sequence tags (ESTs), which were previously characterized as genes putatively coding **SPs** (Huang *et al.* 2011), were selected for SP-SNP development. Sequence data from three cDNA libraries, the urediniospore (Ling *et al.* 2007), germinated urediniospore (Zhang *et al.* 2008) and haustorial cDNA library (Yin *et al.* 2009), were obtained from the NCBI EST database and translated into amino acid sequences using EMBOSS Sixpack (v1.14), which gave three forward and three reverse translations using the standard code (http://www.ebi.ac.uk/Tools/st/emboss_sixpack/). Three software programs were used to verify the EST sequences as SPs. SignalP 4.1 predicts the presence of signal peptides and their locations of cleavage sites based on a combination of several artificial neural networks (Petersen *et al.* 2011). TMHMM 2.0 predicts transmembrane helices in proteins based on a hidden Markov model (Krogh *et al.* 2001). TargetP 1.1 predicts a secreted protein based on the predicted presence of secretory pathway signal peptide in the N-terminal sequence (Emanuelsson *et al.* 2000). The ESTs containing signal peptides were assembled referring the PST-78 genome sequence contigs, and the exon/intron boundaries were determined. The PST-78 genomic sequence was obtained from the **Puccinia** Group Sequencing Project from the Broad Institute of Harvard and Massachusetts Institute of Technology

(http://www.broadinstitute.org/annotation/genome/puccinia_group/MultiHome.html). The assembled genes in PST-78 genomic sequences were aligned with the PST-130 genomic sequences at GenBank under accession AEEW000000000 (Cantu *et al.* 2011) to identify mismatches. To develop SP-SNP markers, base substitutions were used, instead of insertions/deletions which could be resulted from low-quality sequence data. Also, for high-quality candidate SP-SNPs, we eliminated mismatches that are clustered together since clustered mismatch sites are not suitable for primer design. For this study, one polymorphic site was selected for each gene. Primer oligonucleotides were designed for each SP-SNP site using MassARRAY[®] Typer Designer (SEQUENOM, Inc. San Diego, California).

SP-SNP genotyping

Genotyping of all isolates over SP-SNP loci was conducted on a Sequenom MassARRAY iPLEX platform. In the initial PCR reaction, the locus-specific region about 100 bp containing a SN-SNP site was amplified. PCR was performed in a total volume of 4.0 μ L, containing 1.8 μ L of HPLC grade water, 0.5 μ L of 10 $\times$ PCR buffer with 20 mM MgCl₂, 0.4 μ L of 25 mM MgCl₂, 0.1 μ L of 25 mM dNTP mix, 0.5 μ L of 1.0 μ M forward primer, 0.5 μ L of 1.0 μ M reverse primer, 0.2 μ L of 5 units/ μ L of PCR enzyme and 1.0 μ L of 5 ng/ μ L DNA template. The PCR enzyme was ordered from Roche Diagnostic Corporation (Indianapolis, IN, USA), and other reagents were from the PCR Accessory Set from Sequenom Inc. (San Diego, CA, USA). All PCR amplifications were run in a Veriti[™] 384-Well Thermal Cycler (Applied Biosystems, CA, USA) using the following

208 program: 2 min at 95 °C for initial denaturation; 45 cycles of 30 s at 95 °C, 30 s at
209 annealing temperature of 56 °C and 1 min at 72 °C; followed by a final extension step
210 for 5 min at 72 °C and cooling down to 4 °C. Then the iPLEX Gold Reagent Kit
211 (Sequenom Inc., San Diego, California) was used for shrimp alkaline phosphatase (SAP)
212 reaction and single base extension reaction. The SAP reaction was conducted in which
213 phosphatase was used to eliminate unincorporated dNTPs from the previous reaction.
214 The reaction was performed with additional 2 µL of SAP master mix, containing 1.53 µL
215 of HPLC grade water, 0.17 µL of 10× SAP buffer, and 0.3 µL of 1.7 units/µL SAP
216 enzyme to a PCR product. The reaction was run in a Veriti™ 384-Well Thermal Cycler
217 (Applied Biosystems, CA, USA) using the following program: 40 min at 37 °C, followed
218 by 5 min at 85 °C, and cooling down to 4 °C. The single base extension reaction was
219 conducted using modified dideoxynucleoside terminators and thermosequencase
220 enzyme, which produced two extension products differing only in the last one base for
221 each SNP locus. The reaction was performed with additional 2 µL of extension mix,
222 containing 0.619 µL of HPLC (high performance liquid chromatography) grade water,
223 0.2 µL of 10× iPlex buffer plus, 0.2 µL of iPlex terminator, 0.94 µL of extension primer
224 mix with concentration of 8, 10 or 15 µM depending on the primer mass and 0.041 µL of
225 iPlex enzyme. The reaction was run in a Veriti™ 384-Well Thermal Cycler (Applied
226 Biosystems, CA, USA) using the following program: 30 s at 94 °C for initial denaturation;
227 40 cycles of 5 s at 94 °C for denaturation, 5 s at 52 °C for annealing and 5 s at 80 °C for
228 extension for 10 cycles; followed by 3 min at 72 °C for final extension and cooling down
229 to 4 °C. Different alleles were identified by the distinct masses of the extended products
230 using MALDI-TOF mass spectrometry in a Sequenom MassARRAY system. The SNP

231 allele calling and clustering were conducted manually using MassARRAY[®]
232 TYPERANALYZER software (Sequenom).

233

234 **Data analyses**

235

236 A multivariate discriminant analysis of principal components (DAPC) was applied to
237 identify putative clusters of genetically related *Pst* isolates. This method does not rely
238 specific population genetics models, and assumptions about the Hardy-Weinberg
239 equilibrium (HWE) or linkage equilibrium are not required (Jombart *et al.* 2010). The
240 DAPC multivariate analysis was conducted using the Adegenet package in the R
241 environment. The optimal number of clusters was determined on the basis of the
242 Bayesian Information Criterion (BIC). A Bayesian clustering algorithm incorporated in
243 software STRUCTURE version 2.3.4 (Pritchard *et al.* 2000) was also used to validate
244 population subdivisions of the *Pst* isolates in the DAPC analysis. An admixture model
245 was tested with sampling locations provided as population identifications. This
246 admixture model is reasonable because *Pst* isolates may have mixed ancestors. The
247 admixture model assigns individuals into populations based on the fraction of their
248 genome inherited from ancestors in that population. The outputs are the posterior mean
249 estimates of these proportions. Thus, this model can deal with somatic recombinant
250 isolates. The program was run with a 100,000 burnin periods and then 100,000 Markov
251 Chain Monte Carlo (MCMC) repetitions. In the simulation scenarios, estimating K was
252 set from 1 to 15 and allele frequencies were set as independent among populations.
253 The best number of population subdivisions was determined based on the second order

rate of changes in the log probability of data between successive K values, $\Delta K = m(|L(K+1) - 2L(K) + L(K-1)|) / s[L(K)]$ as described by Evanno *et al.* (2005). The peak in the distribution of ΔK showed the best K. The robustness of different clusters from K=2 to K=5 was assessed by repeating 100 independent runs. For each value of K, the LargeKGreedy algorithm implemented in the program CLUMPP version 1.1 was used to search the best alignment of multiple replicate cluster analyses (Jakobsson & Rosenberg 2007). The population structure was visualized using DISTRUCT version 1.1 (Rosenberg 2004).

As an alternative way for clustering isolates, the sequential, agglomerative, hierarchical and nested (SAHN) clustering method was used for constructing a dendrogram for the *Pst* isolates by the unweighted pair group method with the arithmetic means (UPGMA) in the NTSYSpc 2.21L program (Rohlf 2009). Using this program, a similarity matrix was generated based on the simple matching (SM) coefficient in the SIMQUAL module. A UPGMA dendrogram tree was generated from the SM matrix using the SAHN module. The existence of clusters was checked by generating a cophenetic value matrix using the tree matrix produced by the SAHN module. Then the cophenetic value matrix was compared with the original SM matrix in the MXCOMP module in NTSYSpc. Also, a principal component analysis (PCA) was performed to check the existence of the clusters by plotting the first 3 components. Besides the UPGMA method, the neighbor-joining (NJ) method based on parsimony was also performed for constructing a phylogenetic tree. The SM matrix was transformed in a DCENTER module and generating a double-centered matrix which was further to compute eigenvectors in EIGEN module. Then the distance matrix was

computed from eigenvectors in SIMINT module and used to generate a NJ phylogenetic tree using the unweighted method in the NJoin module of the NTSYSpc. Principal component analysis was conducted based on the eigenvectors. The correlation between the double-center similarity matrix and the distance matrix used by the neighbor-joining method was determined by the Mantel test in the NTSYSpc.

Multi-locus genotypes (MLGs) for all isolates over all tested SP-SNP loci and number of MLGs per race were determined manually using the SP-SNP data in EXCEL of Microsoft Office (Microsoft, Edmonds, WA, USA). For the MLGs that were detected from more than one isolate, the program GenClone 2.0 was used to estimate 1) the probability of a given MLG under the assumption of the HWE, and 2) the probability of MLGs resulting from independent sexual reproductive events taking into account departures from the HWE (Arnaud-Haond & Belkhir 2007; Arnaud-Haond *et al.* 2007).

GenAlEx 6.5 software was used to estimate several parameters for describing population genetics (Peakall & Smouse 2006, 2012). Such parameters were the numbers of observed (N_a) and effective (A_e) alleles, Shannon's information index (I), observed heterozygosity (H_o), expected heterozygosity (H_e), unbiased expected heterozygosity (uH_e), fixation index (F_{is}), number of effective migrants (N_m) and percentage of polymorphic loci. Indices A_e , I , H_o , H_e , uH_e and F were calculated over loci for each population in epidemiological regions, and their grand mean were calculated over loci and all populations. F_{is} and N_m were calculated over all isolates for each locus. The percentage of polymorphic loci was calculated over loci for each population. Also, Shannon's mutual information index (sH_{ua}) was calculated over loci for all pairwise population comparison. Allelic richness was computed using FSTAT

v2.9.3.2 (J. Goudet, Lausanne, Switzerland). The null hypothesis of HWE in the *Pst* population in the US was tested using exact tests implemented in Genepop version 4.2 (Raymond & Rousset 1995). The Markov chain parameters were set as 1,000 dememorizations, 100 batches with 1,000 iterations per batch.

Collections of *Pst* isolates were defined as 12 populations with each epidemiological region (R) as one population with R1 to R6 in the western and R7 to R12 in the eastern US (Chen 2005; Chen *et al.* 2010). Population genetic variance was determined by analysis of molecular variance (AMOVA) in GenAlEx 6.5 with 999 permutations of the data set. Both F_{ST} (Wright 1965) and R_{ST} (Slatkin 1995) statistics were conducted for SP-SNP genotype variation among isolates within populations, variation between populations within the western or eastern US regions, and variation between the western and eastern US regions; while Φ_{ST} (Michalakis & Excoffier 1996) statistics, which are applicable to binary data, was conducted for virulence phenotype variation of isolates within populations, variation between populations within the western/eastern US regions and variation between the western and eastern US regions. To analyze population differentiation, pairwise comparisons of the 12 region populations were conducted based on F_{ST} and R_{ST} statistics for SP-SNP genotypes, and Φ_{PT} statistics for virulence phenotypes. Also calculated were Nei's genetic distances (D) (Nei 1972) between the populations based on the SP-SNP genotypes and virulence phenotypes. The correlation coefficients between F_{ST} , R_{ST} , Φ_{PT} and Nei's D genetic distance values from the pairwise comparisons of the 12 region populations were tested using the Mantel coefficient performed in GenAlEx, with 999 permutations of the dataset.

The correlation coefficient between the SP-SNP genotypes and virulence phenotypes was computed using their similarity data in NTSYSpc. Since 30 SP-SNPs were significantly associated with 9 avirulence loci (Xia *et al.* 2016), correlation analyses for avirulence-associated and non-associated SP-SNPs were performed separately to estimate whether SP-SNPs were under selection. To compare SP-SNPs with SSR markers, 180 isolates from 2010 which were genotyped by both types of markers were used to determine the correlations between the SSR/SP-SNP genotypes and the virulence phenotypes separately and also the correlation between the SSR and SP-SNP genotypes.

Results

Development of SP-SNP markers

From the 446 *Pst* putative SP ESTs, signal peptides were detected in 384 sequences by at least two of the prediction programs. A total of 441 mismatches were identified in 275 sequences when aligned between PST-78 and PST-130 genomic sequences. Mismatches were chosen for primer design when no substitutions occur in upstream and downstream 100 bp away from the mismatch sites. A total of 183 pairs of oligonucleotide primers were designed. These SP-SNPs have been deposited to the Single Nucleotide Polymorphism Database (dbSNP) in the National Center for Biotechnology Information (NCBI) with accession numbers from ss1583581412 to ss1583581509. After screening with 60 randomly selected *Pst* isolates, 92 polymorphic

SP-SNPs were selected (Table D1), based on the minor allele frequency $\geq 5\%$, to study the *Pst* population.

Multi-locus genotypes

A total of 242 multi-locus genotypes (MLGs) were detected from the 352 *Pst* isolates genotyped with the 92 SP-SNP markers (Table 1; Table D2). Of the 242 MLGs, 21 were presented by two or more isolates and the remaining were each presented by a single isolate. MLGs with most numbers of isolates included MLG015 (43 isolates), MLG002 (19 isolates), MLG037 (16 isolates) and MLG010 (12 isolates).

Among the epidemiological regions, the number of MLGs ranged from 4, detected from 4 isolates in R10, to 102, detected from 111 isolates in R1 (Table 1). Private MLGs, which were present only in one epidemiological region ranged from 1 in R10 to 94 in R1. The western regions had a total of 178 private MLGs with 83.1% of the isolates, whereas the eastern regions had only 45 private MLGs with 44.5% of the isolates. Identical MLGs were shared within races from different epidemiological regions.

When the 242 MLGs were organized by races, the number of MLGs per race ranged from 1 for many races to 49 in race PSTv-37 (Table S1). In addition to PSTv-37, races with multiple MLGs were PSTv-11 (43 MLGs) and PSTv-14 (32 MLGs). These three races were the most predominant ones in 2010 and 2011 (Wan & Chen 2014; Wan *et al.* 2016), indicating that these predominant races each consisted of genetically different isolates.

Genetic clusters

The 352 isolates were grouped into three genetic lineages (GLs) based on the DAPC multivariate analysis (Fig 2A, 2B). In the DAPC analysis, 164 (46.6%) isolates were clustered into GL1, 139 (39.5%) into GL2 and 49 (13.9%) into the Admix group. The isolates from the western US population were in both GL1 (60 isolates, 25%) and GL2 (136 isolates, 56%), with 46 isolates in Admix (19%). Except 6 isolates (4 in GL2 and 2 in Admix), the isolates from the eastern US population were all clustered in GL1 together with isolates from the western US population (Table 1). The PCA analysis revealed similar clustering groups as the DAPC analysis. The first three eigenvectors from the PCA analysis explained 75.32% of the total variation in the SP-SNP genotypes, with the first eigenvector explaining 64.51%, the second 6.47% and third 4.34% (Fig 3).

In general, the clustering by the neighbor-joining method was consistent with that by the DAPC analysis. $_N$ GL1 and $_N$ GL3 were two major groups containing 179 (50.85%) and 149 (42.33%) isolates, respectively while $_N$ GL2 had 24 (6.82%) isolates (Table 2, Fig S1). $_N$ GL1A and $_N$ GL3A from the neighbor-joining method corresponded to GL1 and GL2 of the DAPC results, respectively, while the Admix group in the DAPC analysis was further separated into three subgroups using the neighbor-joining method, $_N$ GL1B, $_N$ GL2 and $_N$ GL3B (Table 2; Fig S1). These clusters were similar to the results from the UPGMA method (Fig S2).

When the model-based Bayesian clustering analysis in the STRUCTURE program was used, the optimal grouping number $K = 2$ was obtained (Fig 2B; Fig S3A). The admixture model was used in STRUCTURE analysis. Individuals were assigned

into an admix population when the estimated membership in either population was less than 0.80. Two genetic lineages (GL1 and GL2) and one admix group (Admix) were determined, consistent with the results from the DAPC analysis. These three clusters were corresponding to the two clusters from STRUCTURE when the Admix group was counted (Fig 2B, $K = 2$). However, when $K = 3, 4$, and 5 were analyzed, more detailed structures were identified, mostly in the western populations (Fig 2B). All of the clustering and structural analyses indicated that the western regions (R1-R6) have more genetic populations than the eastern regions (R7-R12).

Population diversity

The overall genetic diversity of 352 *Pst* isolates was estimated based on the 92 SP-SNP loci (Table 3). Two alleles of each SP-SNP locus were detected in the populations from all epidemiological regions, while no private alleles were identified. All populations had an observed number of alleles more than 1.9, except for populations from R4 and R12, which included relatively small numbers of isolates (6 and 10). The number of effective alleles ranged from 1.24 per locus in R4 to 1.71 per locus in R10. There was little variation among the populations in the number of observed alleles and the number of effective alleles because only two alleles were detected at each SNP locus. The population from R7 had the lowest allelic richness at 1.60 per locus while population from R2 had the highest allelic richness at 1.91 per locus. The population from R10 was most diverse with a Shannon's information index I of 0.57 and the R4 population had the lowest I statistic of 0.20.

Genetic variation and population differentiation

The genetic variance for the epidemiological region populations of *Pst* was partitioned by AMOVA based on allele frequencies at the SP-SNP loci and virulence to 18 Yr single-gene line differentials (Table 4). In the AMOVA analysis, the collection of isolates from each of the 12 epidemiological regions was treated as a population, and two geographic regions were defined as the western US including R1 to R6 and the eastern US including R7 to R12. The F_{ST} and R_{ST} values were calculated on the basis of allele frequencies at the SP-SNP loci and revealed 78% and 77% ($P < 0.001$), respectively of the total variation within populations. Both F_{ST} and R_{ST} revealed 16% ($P < 0.001$) of the total variation between the western and eastern US regions. The SP-SNP variation among populations was only 6% and 7% of the total variation based on F_{ST} and R_{ST} , respectively, but were all significant ($P < 0.001$).

The variation of virulence to the 18 Yr single-gene differentials had a similar distribution based on the Φ_{PT} statistics (Table S2). The variation within populations was 65% ($P < 0.001$) of the total variation, the variation between the western and eastern US regions 21% ($P < 0.001$), and the variation among populations 14% ($P < 0.001$). Of the 47 races, 27 had isolates in a single genotype group or subgroups and 20 had isolates in two or more groups or subgroups (Table S1). There was a moderate, but significant correlation ($r = 0.34$; $P < 0.0001$) between the SP-SNP genotypes (Table D2) and virulence phenotypes (Table D3).

Genetic differentiations between epidemiological regions were determined through population pairwise comparisons using F_{ST} , R_{ST} and Nei's (D) genetic distance for the SP-SNP genotypes and Φ_{PT} statistic and Nei's (D) genetic distance for virulence phenotypes (Table S2). Generally, the populations within either the western or eastern US were low to moderately differentiated ($F_{ST} < 0.15$; $R_{ST} < 0.25$), while populations between the two regions were highly differentiated ($F_{ST} > 0.15$; $R_{ST} > 0.25$) with few exceptions. The R6 population was an exception since it was significantly ($P < 0.001$) differentiated from all other epidemiological regions in the western US with F_{ST} values ranging from 0.107 (R2) to 0.365 (R4), but no significant differentiation was detected between R6 with any epidemiological regions in the eastern US with the F_{ST} values ranging from 0.000 (to R10; $P = 0.309$) to 0.007 (to R9; $P = 0.213$). In contrast, significant differentiations (F_{ST} : 0.134 to 0.494; P : 0.002 to 0.001) were detected between a region in the eastern US and all regions in the western US, except with R6 (F_{ST} : 0.000 to 0.007; P : 0.213 to 0.327) and also between R10 and R2 ($F_{ST} = 0.001$, $P = 0.415$) or R5 ($F_{ST} = 0.038$, $P = 0.178$). Similar results were obtained from pairwise comparisons of populations with R_{ST} statistic and Nei's distance for SP-SNP genotypes (Table S2).

Genetic differentiation assessed for pairwise comparisons of populations with Φ_{PT} statistic and Nei's (D) genetic distance for the virulence phenotypes also indicated that in general populations within the western or eastern US were either not significantly or significantly but low to moderately differentiated ($\Phi_{PT} < 0.15$), while the populations between the western and eastern US were highly differentiated ($\Phi_{PT} > 0.15$) (Table S2) with the exceptions of R5 and R6, respectively. Specifically, R1 was significantly

differentiated from any other epidemiological regions in both the western and eastern
 US (Φ_{PT} : 0.032 to 0.351; P : 0.001 to 0.043) except with R4 (Φ_{PT} = 0.034; P = 0.181). R2
 was significantly differentiated from any other epidemiological regions in both the
 western and eastern US (Φ_{PT} : 0.081 to 0.352; P : 0.001 to 0.086). R3 was significantly
 differentiated from all regions in both the western and eastern US (Φ_{PT} : 0.032 to 0.388;
 P : 0.043 to 0.001). R4 was not significantly differentiated from other populations from
 the western US (Φ_{PT} < 0.15) except R5 (Φ_{PT} = 0.348; P < 0.001) and R6 (Φ_{PT} = 0.455; P
 < 0.001). R4 was significantly differentiated from populations from the eastern US (Φ_{PT} >
 0.15; P < 0.05). Interestingly, R5 was significantly differentiated from all other
 populations from both the western and eastern US. All populations from the eastern US
 were significantly **different** from populations from the western US (Φ_{PT} > 0.15; P < 0.05)
 with the exception of R10 with R2 which had Φ_{PT} = 0.188 with P = 0.053. There was no
 significant differentiation among populations **within** the eastern US. Similar results were
 detected from pairwise comparisons of populations with Nei's distance for the virulence
 phenotypes (Table S2).

The correlations among all statistics for population pairwise values for both SP-
 SNP variation and virulence variation were assessed using Mantel tests with 999
 permutations. There were significant correlations (r = 0.694 to 0.986, P < 0.05) among
 pairwise values of F_{ST} , R_{ST} statistics and Nei's (D) genetic distance for SP-SNP
 variation (Table S3). The correlation (r = 0.839) between the Φ_{PT} statistic and Nei's (D)
 genetic distance for the virulence phenotypes were significant (P < 0.05). Moreover, the
 correlations (r = 0.694 to 0.900) between the values of genetic distance measured by
 SP-SNPs and those by virulence were also significant (P < 0.05).

Heterozygosity and deviation from HWE

The heterozygosity of individual SP-SNP loci among the 352 isolates ranged from 0 to 99% with a mean of 49% (Table D4). The observed heterozygosity (H_o) in the epidemiological region populations differed among epidemiological regions, ranging from 0.24 in R4 to 0.67 in R12 (Table 3). The fixation index was estimated to evaluate the deviation of the observed heterozygosity from the expected heterozygosity. The deviation of the observed heterozygosity value was the **least** ($F_{is} = -0.79$) in the R12 population and the **most** ($F_{is} = -0.048$) in the R1 population. Also, the heterozygosity differed between the genetic groups (Table 5). Overall, the heterozygosity values ranged from 15.6% to 76.3% with a mean of 47.6. GL1 had the highest heterozygosity (53.1 – 76.3%, mean 66.8%) and GL2 had the lowest heterozygosity (17.7 – 42.7%, mean 24.5%). All of the SP-SNP loci showed a significant deviation from HWE except 5 loci, indicating a high clonality in all 12 *Pst* epidemiological region populations. For each region, the deviation from the HWE was highly significant except the population from R10 because of the limited number of isolates in this region. **All of the MLGs were also significantly departure from the HWE ($P_{gen(fis)} < 0.05$), indicating the clonal production of the *Pst* isolates (Table 1). Similarly, the low values of $P_{sex(fis)} (< 0.05)$ of most MLGs did not indicate sexual reproduction, but supported clonal reproduction.**

Gene flow

A high level of gene flow, in terms of the mean number of effective migrants, was detected ($N_m = 58.968$) in the *Pst* populations among the 12 epidemiological regions. The number of effective migrants between regions ranged from 0.328 (between R4 and R12) to 39,030.546 (between R1 and R3) (Table S2). The major MLGs and races shared between or among epidemiological regions were illustrated in Fig. S4.

Discussion

Development of SP-SNP markers

In this study, we developed the first set of SP-SNP markers and applied them to estimate the diversity and structures of the *Pst* populations in the US. The SP-SNPs differentiated the western population from the eastern population as well as populations in each epidemiological region. These results are consistent with the results obtained from virulence phenotypes (Wan and Chen 2014; Wan et al. 2016), and also consistent with the results obtained with SSR markers (A.M. Wan, M.N. Wan & X.M. Chen, unpublished data). The present study demonstrates the usefulness of the SP-SNP markers in studying population genetics of the pathogen.

In the present study, we chose SP genes for developing SNP markers primarily based on the recent findings in other fungi that many cloned avirulence genes coding for secreted protein genes (Catanzariti et al. 2006; Kamoun 2007; Petre et al. 2014). In fact, we have identified 30 markers significantly associated to 9 avirulence loci in *Pst* (Xia et al. 2016). The 92 SP-SNPs from 66 supercontigs selected based on their

polymorphism in an initial screening distribute throughout the *Pst* genome (Table D1). When we analyzed the correlation between the virulence data of the 352 isolates tested in the present study with the 30 SP-SNPs that are significantly associated with 9 avirulence loci (Xia et al. 2016), the correlation coefficient (0.37) was slightly higher than that (0.34) with the 92 markers, while the value (0.32) with the remaining 62 SP-SNPs was slightly lower, as expected (data not shown). Thus, the effects of virulence-associated markers could be neutralized by the large number of markers.

The suitability of SP-SNP markers for population genetic study is also can be seen by comparing with SSR marker results. In a separate study, we have characterized the *Pst* isolates from the US in 2010 with 18 SSR markers (A.M. Wan, M.N. Wang & X.M. Chen, unpublished data). When we compared the results of 180 isolates that were commonly used in that study and the present study, the correlation coefficient of the SSR data with virulence data was 0.48 ($P < 0.0001$) while that of the 92 SP-SNP data was 0.43 ($P < 0.0001$) (data not shown). The correlation coefficient (0.34) obtained in the present study of 352 isolates with 92 SP-SNPs is even slightly lower than, but comparable with the value (0.39) obtained between virulence and marker data with 20 SSR markers in a previous study (Cheng & Chen 2014). The correlation coefficient (0.34) obtained in the present study is also lower, but comparable to the value (0.37) determined using SSR markers for the population of *P. triticina* (the leaf rust pathogen) in France (Goyeau et al. 2007). All of these comparisons indicate that the SP-SNP markers are as suitable as SSR markers for determine genetic relationships between populations, but a large number of SP-SNPs should be used. Because SNP markers have an advantage for high-throughput technology, more

markers can be developed based on *Pst* sequences, and used for population studies as well as tagging genes of interest.

***Pst* population diversity in the US**

In agreement with previous studies (Chen *et al.* 2002; Chen 2005; Wan & Chen 2012, 2014; Wan *et al.* 2016), the present study indicated that the *Pst* populations from the western US had a higher diversity in terms of percentage of polymorphic loci, number of multi-locus genotypes, and ratio of genotype to isolate than the populations from the eastern US. Another evidence to illustrate the more diverse western US population than that of eastern US was that most of the eastern US isolates were clustered together while the isolates from the western US were separated into different genetic groups (Table 1; Fig 1). The fact that the *Pst* population from the western US is more diverse than that of the eastern US is probably due to several factors. First, the cropping system is more diverse in the western US especially the Pacific Northwest (PNW) (Chen *et al.* 2010). Both winter and spring wheat and more wheat cultivars with different resistance types and a large number of stripe rust resistance genes are planted in the western US annually, reducing the direct selection for particular races and race groups and allowing the existence of races with either a short or long virulence spectrum. Second, the weather conditions are more favorable for pathogen survival in the western US, especially in the PNW where winters are mild and late springs and summers are cool, thus providing more opportunities for various races or genotypes to survive (Sharma-Poudyal *et al.* 2014). Third, both winter wheat and spring wheat are planted in the PNW,

plus various auxiliary grass hosts, served as “green bridge” for pathogen survival and spread through continuous infections (Line 2002; Chen *et al.* 2010; Cheng *et al.* 2016).

***Pst* population differentiation in the US**

The results of this study indicated that the *Pst* populations from the western US was highly differentiated with that from the eastern US because of the geographical barriers, cropping systems within regions and weather conditions. However, the population from R6 (covering central and southern California, Arizona, and western New Mexico) was less differentiated from that from the eastern US, indicating that the population in R6 is not highly isolated from the populations in the eastern US. The less differentiation between the populations of R6 and eastern US regions can be explained by the pathogen movement in the US including prevalent eastward wind spreading urediniospores from R6 to R7 and R8, and southeastward spread from California to Mexico, and then northward spread to Texas and other regions in the eastern US (Chen 2005). These movement pathways were evidenced by the detection of similar races in California and the eastern regions, and the sequential appearance of disease from south to north (Chen *et al.* 2002; Chen 2005). The population differentiations between R6 and other regions in the western US were high, and differentiations among other western regions were low to intermediate. In contrast, we were unable to detect significant differentiations among populations within the eastern US. This was probably either because a diverse *Pst* population was not established until recently (Chen *et al.* 2002) with less divergence occurring within each regions, or possibly because more

uniform selection shaped the population. In general, *Pst* cannot overwinter in much of the southern part and cannot overwinter in much of the northern part of the eastern US (Sharma-Poudyal *et al.* 2014). The effect of inoculum sources reduces differentiations in the different regions of the eastern US. This is in agreement with the reports that the old population of *Pst* has been replaced by a new population in the eastern US while in the western US both old and new populations have co-existed since 2000 (Chen *et al.* 2002; Chen 2005; Milus *et al.* 2009; Chen *et al.* 2010; Wan & Chen 2012, 2014; Wan *et al.* 2016).

Analysis of molecular variance with F_{ST} and R_{ST} statistics for the SP-SNP genotypes and Φ_{PT} statistic for virulence consistently showed significant variation among the western and eastern US populations, among epidemiological regions and within the regions. The highest variation was present among isolates within epidemiological regions, which was consistent with results from the PNW population (Cheng & Chen 2014). The variation between the western and eastern US and among epidemiological regions detected from virulence phenotypes were higher than that detected by SP-SNP genotypes, suggesting that the virulence was more influenced by the selection of resistance of wheat cultivars in each region. The overall correlation between the virulence and SP-SNP data was moderate but significant ($r = 0.34$; $P < 0.0001$), suggesting that the SP-SNP polymorphisms are less independent of virulence even though some SP-SNPs may not be neutral. However, as discussed above, it needs to be determined whether some of the SP-SNPs or their genes are significantly associated with virulence and thus influenced by selection.

SP-SNP genotypes associated with virulence phenotypes

Clustering analyses using SP-SNP genotypes showed some association of molecular groups with races. Genetic lineage 1 (GL1) resulted from the DAPC analysis, corresponding to genetic lineage 1A ($_N$ GL1A) from the neighbor-joining method (Table 3; Supporting Fig. S2), was comprised mainly of PSTv-34 (equivalent to PST-98), PSTv-35 (equivalent to PST-78 and PST-80), PSTv-36 (equivalent to PST-98) and PSTv-37 (equivalent to PST-98, PST-99, PST-100 and PST-102) (Wan & Chen 2014) with PSTv-37 being the most predominant race in GL1. These isolates are characterized by avirulence to wheat cultivars carrying *Yr1* and *YrTye* genes. The clustering of these races is consistent with their origin and the hypothesis that they are genetically close (Chen *et al.* 2010). GL2 from the DAPC analysis corresponded to $_N$ GL4A from the neighbor-joining method and was comprised mainly of PSTv-11 (equivalent to PST-127), PSTv-12 (equivalent to PST-127), PSTv-13 (equivalent to PST-143), PSTv-14 (equivalent to PST-141), PSTv-15 (equivalent to PST-144), PSTv-16 (equivalent to PST-151) and PSTv-17 (equivalent to PST-137). These races were also thought to be closely related and characterized by their virulence to *Yr1* and *YrTye* (Wan & Chen 2014).

The admixed group from the DAPC analysis was separated into $_N$ GL1B, $_N$ GL2, $_N$ GL3 and $_N$ GL4B by the neighbor-joining method. $_N$ GL1B and $_N$ GL4B shared the similar characters and clustered together with $_N$ GL1A and $_N$ GL4A, respectively, in both the neighbor-joining dendrogram and PCA plot. $_N$ GL2 was comprised mainly of PSTv-18, PSTv-19, PSTv-20 and PSTv-21. Interestingly, all of these races were collected from

R5 in northwestern Washington. These races are characterized by a narrow virulence spectrum, with PSTv-18 (equivalent to PST-3 and PST-35) avirulent to all 18 *Yr* single-gene differentials, PSTv-19 only virulent to *Yr6*, PSTv-20 only virulent to *Yr17*, and PSTv-21 virulent to both *Yr6* and *Yr17* (Wan & Chen 2014). Isolates of PSTv-18 were obtained from both wheat and grass. Also it can infect some barley cultivars, for example 'Topper' and 'Abed Binder' (data not shown). This can partially explain the high race diversity in R5 (Line 2002; Chen 2005).

High heterozygosity in the *Pst* population

A high level of heterozygosity (two alleles present) of dikaryotic urediniospores in the US *Pst* population was revealed in the present study with an average of 51.5% of an isolate estimated over 92 SP-SNP loci. For all populations, the observed heterozygosity values were higher than the expected heterozygosity (H_e) or the unbiased expected heterozygosity (uH_e) values which were calculated based on kinship coefficients for pairs of individuals, thus are unbiased estimators for gene diversity for inbred or genetically related samples (Peakall & Smouse 2006). The high heterozygosity frequency in the US population was consistent with previous reports using different approaches in various *Pst* populations (Liu *et al.* 2012; Cheng & Chen 2014; Holtz *et al.* 2014). The molecular mechanisms underlying the high heterozygosity in the *Pst* population remain unknown. However, several potential processes and plausible mechanisms can be hypothesized for the occurrence and maintenance of the high heterozygosity. Point mutation of one of the two alleles within a homozygous locus in a

666 dikaryotic urediniospore cell is the major source to produce a heterozygous locus
 667 (Hovmøller *et al.* 2011; Liu *et al.* 2012). Although sexual recombination is proposed to
 668 contribute to diversity (Jin *et al.* 2010; Ali *et al.* 2014b), the role of alternate hosts on *Pst*
 669 is limited by host phenology and telial degradation in the US (Wang & Chen 2015;
 670 Wang *et al.* 2015). Hybridization of genetically different asexual lineages is a plausible
 671 mechanism on heterozygosity. Somatic hybridization was reported for *Pst* isolates (Little
 672 & Manners 1969; Johnson & Taylor 1976; Park & Wellings 2012). Recently, Holtz *et al.*
 673 (2014) identified a cluster of *P. striiformis* f. sp. *hordei* (*Psh*) isolates and hypothesized it
 674 to be the result from hybridization between *Pst* and *Psh*. The hybridization of different
 675 clonal lineages may increase the heterozygosity. Heterozygous loci can be maintained
 676 by the “Meselson effect” (Birky 1996; Balloux *et al.* 2003) in which different mutations
 677 can be accumulated and the divergence between two alleles at a locus is promoted by
 678 the mainly or strictly clonal population as in *Pst* populations. Using analytical and
 679 stochastic simulation methods, Balloux *et al.* (2003) concluded that the heterozygosity
 680 was positively affected by high rates of clonal reproduction. Lastly, in the perspectives
 681 of evolution and ecology, the high heterozygous frequency of *Pst* isolates, especially the
 682 heterozygosity in avirulence loci, enables the virulent homozygous locus to emerge
 683 easily from a heterozygous state under the assumption of recessive inheritance of
 684 virulence (Wellings & McIntosh 1990). The recessive character of the virulence of *Pst*
 685 was suggested by previous experiments (Little & Manners 1969; Johnson & Taylor
 686 1976). In this point, the high heterozygosity within pathogenicity loci is responsible, at
 687 least to some extent, for the highly evolving *Pst* population in terms of race dynamics.
 688 Studies on the structures and polymorphisms of these loci will provide detailed

knowledge on the relationship between high heterozygosity and race dynamics in the *Pst* population. However, no avirulence or virulence genes have been molecularly identified and functionalized to date.

One might be confused by the fact that a high heterozygosity was detected from less diverse populations from the eastern US, while the more diverse populations from the western US had a lower heterozygosity. This can be explained by the nature of clonal reproduction of *Pst* populations. **On one hand**, at each single locus the polymorphism is protected within isolates due to the fixed heterozygosity through the 'Meselson effect' because of the clonal reproduction (Balloux *et al.* 2003). On the other hand, the asexual reproduction decreases the genotypic diversity because of the less frequency of recombination which can interrupt the established genotypes. Therefore, the mainly asexual *Pst* population maintains a higher genetic diversity at each single locus, but a lower number of multi-locus genotypes. The clonal reproduction of *Pst* populations in the US was confirmed in the present study. All 92 loci except 5 showed a significant departure from HWE using Fisher's exact tests, indicating lack or infrequent sexual reproduction. In all epidemiological regions, the departure of HWE was significant except in R10, which is more likely due to the limited number of isolates. The predominant clonal reproduction of *Pst* populations indicated by SP-SNPs in the present study was in agreement with the results revealed by SSR and/or other markers for the *Pst* population in the PNW in the US (Cheng & Chen 2014) and the international populations (Hovmøller *et al.* 2002; Bahri *et al.* 2009; Ali *et al.* 2010, 2014a; Holtz *et al.* 2014), even though recombinant populations may exist (Ali *et al.* 2014b). These results also illustrate that the inoculum of *Pst* is mainly from clonal reproduction and rarely from

sexual reproduction, if any. This is in agreement with a previous study focused on the PNW population using co-dominant SSR markers (Cheng & Chen 2014). Even though the sexual stage of *Pst* was observed under greenhouse conditions (Jin *et al.* 2010; Wang & Chen 2013) and occasionally under natural conditions (Zhao *et al.* 2013), the short longevity of germinated basidiospores, the low infection rate on *Berberis* under natural conditions, and the inability of teliospores to survive harsh winters cause sexual reproduction to be undetectable and thus less important for the epidemiology in the aspect of providing inoculum (Wang & Chen 2015; Wang *et al.* 2015).

To our knowledge, the present study is the first to develop SNP markers for the stripe rust pathogen. Using the SP-SNP markers, the results of population analysis indicate that the *Pst* populations in the western US are more diverse, especially in the R1 and R5, than the eastern US. However, the eastern populations are less differentiated from the population in R6, indicating a high migration. Therefore, the breeding program in the eastern US should pay much attention on the *Pst* race dynamics in the western US, especially R6 because the inoculum for stripe rust epidemics in the eastern US can be from R6. Because gene flow also occurs among other regions, the pathogen populations in any other regions should be considered when select resistance sources or genes. In general, diverse types of resistance genes should be incorporated into wheat cultivars to avoid direct selection in the pathogen population and to achieve sustainable control of stripe rust.

Acknowledgements

This research was supported by the U.S. Department of Agriculture, Agricultural Research Service (Project No. 5348-22000-015-00D) and Washington State University (13C-3061-5665) PPNS No. 0690, Department of Plant Pathology, College of Agricultural, Human, and Natural Resource Sciences, Agricultural Research Center, HATCH Project Number WNP00663, Washington State University, Pullman, WA 99164-6430, USA. The Chinese Scholarship Council (CSC) scholarship given to Chongjing Xia is highly appreciated. We are grateful to Drs. Scot Hubert and Tobin Peever for critical review of the manuscript.

References

- Ali S, Gladieux P, Leconte M, Gautier A, Justesen AF, Hovmøller MS, Enjalbert J, de Vallavieille-Pope C, 2014a. Origin, migration routes and worldwide population genetic structure of the wheat yellow rust pathogen *Puccinia striiformis* f. sp. *tritici*. *PLoS Pathogens* **10**: 1-12.
- Ali S, Gladieux P, Rahman H, Saqib MS, Fiaz M, Ahmad H, Leconte M, Gautier A, Justesen AF, Hovmøller MS, Enjalbert J, de Vallavieille-Pope C, 2014b. Inferring the contribution of sexual reproduction, migration and off-season survival to the temporal maintenance of microbial populations: a case study on the wheat fungal pathogen *Puccinia striiformis* f.sp. *tritici*. *Molecular Ecology* **23**: 603-617.
- Ali S, Leconte M, Walker AS, Enjalbert J, de Vallavieille-Pope C, 2010. Reduction in the sex ability of worldwide clonal populations of *Puccinia striiformis* f. sp. *tritici*. *Fungal Genetics and Biology* **47**: 828-838.

- 758 Arnaud-Haond S, Belkhir K, 2007. GENCLONE: a computer program to analyse
 759 genotypic data, test for clonality and describe spatial clonal organization. *Molecular*
 760 *Ecology Notes* **7**: 15-17.
- 761 Arnaud-Haond S, Duarte M, Alberto F, Serrão EA, 2007. Standardizing methods to
 762 address clonality in population studies. *Molecular Ecology* **16**: 5115-5139.
- 763 Bahri B, Leconte M, Ouffroukh A, de Vallavieille-Pope C, Enjalbert J, 2009. Geographic
 764 limits of a clonal population of wheat yellow rust in the Mediterranean region.
 765 *Molecular Ecology* **18**: 4165-4179.
- 766 Bailey J, Karaoglu H, Wellings CR, Park RF, 2013. Isolation and characterisation of 25
 767 genome-derived simple sequence repeat markers for *Puccinia striiformis* f. sp. *tritici*.
 768 *Molecular Ecology Resources* **13**: 760-762.
- 769 Balloux F, Lehmann L, de Meeus T, 2003. The population genetics of clonal and
 770 partially clonal diploids. *Genetics* **164**: 1635-1644.
- 771 Birky CW, 1996. Heterozygosity, heteromorphy, and phylogenetic trees in asexual
 772 eukaryotes. *Genetics* **144**: 427-437.
- 773 Brown JKM, Hovmøller MS, 2002. Aerial dispersal of pathogens on the global and
 774 continental scales and its impact on plant disease. *Science* **297**: 537-541.
- 775 Cantu D, Govindarajulu M, Kozik A, Wang MN, Chen XM, Kojima K, Jurka J,
 776 Michelmore RW, Dubcovsky J, 2011. Next generation sequencing provides rapid
 777 access to the genome of wheat stripe rust. *PLoS ONE* **8**: e24230,
 778 doi:10.1371/journal.pone.0024230.

- 779 Catanzariti AM, Dodds PN, Lawrence GJ, Ayliffe MA, Ellis JG, 2006. Haustorially
 780 expressed secreted proteins from flax rust are highly enriched for avirulence elicitors.
 781 *The Plant Cell* **18**: 243-256.
- 782 Chen CQ, Zheng WM, Buchenauer H, Huang LL, Lu NH, Kang ZS, 2009. Isolation of
 783 microsatellite loci from expressed sequence tag library of *Puccinia striiformis* f.sp.
 784 *tritici*. *Molecular Ecology Resources* **9**: 236-238.
- 785 Chen XM, 2005. Epidemiology and control of stripe rust [*Puccinia striiformis* f. sp. *tritici*]
 786 on wheat. *Canadian Journal of Plant Pathology* **27**: 314-337.
- 787 Chen XM, 2014. Integration of cultivar resistance and fungicide application for control of
 788 wheat stripe rust. *Canadian Journal of Plant Pathology* **36**: 311-326.
- 789 Chen XM, Line RF, Leung H, 1993. Relationship between virulence variation and DNA
 790 polymorphism in *Puccinia striiformis*. *Phytopathology* **83**: 1489-1497.
- 791 Chen XM, Moore MK, Milus EA, Long DL, Line RF, Marshall D, Jackson L, 2002. Wheat
 792 stripe rust epidemics and races of *Puccinia striiformis* f. sp. *tritici* in the United States
 793 in 2000. *Plant Disease* **86**: 39-46.
- 794 Chen XM, Penman L, Wan AM, Cheng P, 2010. Virulence races of *Puccinia striiformis* f.
 795 sp. *tritici* in 2006 and 2007 and development of wheat stripe rust and distributions,
 796 dynamics, and evolutionary relationships of races from 2000 to 2007 in the United
 797 States. *Canadian Journal of Plant Pathology* **35**: 315-333.
- 798 Cheng P, Chen XM, 2014. Virulence and molecular analyses support asexual
 799 reproduction of *Puccinia striiformis* f. sp. *tritici* in the U.S. Pacific Northwest.
 800 *Phytopathology* **104**: 1208-1220.

- 801 Cheng P, Chen XM, See DR, 2016. Grass hosts harbor more diverse isolates of
 802 *Puccinia striiformis* than cereal crops. *Phytopathology*
 803 <http://apsjournals.apsnet.org/doi/pdf/10.1094/PHTO-07-15-0155-R>
- 804 Cheng P, Chen XM, Xu LS, See DR, 2012. Development and characterization of
 805 expressed sequence tag-derived microsatellite markers for the wheat stripe rust
 806 fungus *Puccinia striiformis* f. sp. *tritici*. *Molecular Ecology Resources* **12**: 779-781.
- 807 Emanuelsson O, Nielsen H, Brunak S, von Heijne G, 2000. Predicting subcellular
 808 localization of proteins based on their N-terminal amino acid sequence. *Journal of*
 809 *Molecular Biology* **300**: 1005-1016.
- 810 Enjalbert J, Duan X, Giraud T, Vautrin D, de Vallavieille-Pope C, Solignac M, 2002.
 811 Isolation of twelve microsatellite loci, using an enrichment protocol, in the
 812 phytopathogenic fungus *Puccinia striiformis* f.sp. *tritici*. *Molecular Ecology Notes* **2**:
 813 563-565.
- 814 Evanno G, Regnaut S, Goudet J, 2005. Detecting the number of clusters of individuals
 815 using the software Structure: a simulation study. *Molecular Ecology* **14**: 2611-2620.
- 816 Gardes M, Bruns D, 1993. ITS primers with enhanced specificity for basidiomycetes -
 817 application to the identification of mycorrhizae and rusts. *Molecular Ecology* **2**: 113-
 818 118.
- 819 Goyeau H, Halkett F, Zapater MF, Carlier J, Lannou C, 2007. Clonality and host
 820 selection in the wheat pathogenic fungus *Puccinia triticina*. *Fungal Genetics and*
 821 *Biology* **44**: 474-483.
- 822 Holtz MD, Kumar K, Zantinge JL, Xi K, 2014. Genetic diversity of *Puccinia striiformis*
 823 from cereals in Alberta, Canada. *Plant Pathology* **63**: 415-424.

- 824 Hovmøller MS, Justesen AF, Brown JKM, 2002. Clonality and long-distance migration of
 825 *Puccinia striiformis* f. sp. *tritici* in north-west Europe. *Plant Pathology* **51**: 24-32.
- 826 Hovmøller MS, Sørensen CK, Walter S, Justesen AF, 2011. Diversity of *Puccinia*
 827 *striiformis* on cereals and grasses. *Annual Review of Phytopathology* **49**: 197-217.
- 828 Hovmøller MS, Walter S, Bayles RA, Hubbard A, Flath K, Sommerfeldt N, Leconte M,
 829 Czembor P, Rodriguez-Algaba J, Thach T, Hansen JG, Lassen P, Justesen AF, Ali S,
 830 de Vallavieille-Pope C, 2015. Replacement of the European wheat yellow rust
 831 population by new races from the centre of diversity in the near-Himalayan region.
 832 *Plant Pathology* doi: 10.1111/ppa.12433.
- 833 Hovmøller MS, Yahyaoui AH, Milus EA, Justesen AF, 2008. Rapid global spread of two
 834 aggressive strains of a wheat rust fungus. *Molecular Ecology* **17**: 3818-3826.
- 835 Huang XL, Chen XM, Coram T, Wang MN, Kang ZS, 2011. Gene expression profiling of
 836 *Puccinia striiformis* f. sp. *tritici* during development reveals a highly dynamic
 837 transcriptome. *Journal of Genetics and Genomics* **38**: 357-371.
- 838 Jakobsson M, Rosenberg NA, 2007. CLUMPP: a cluster matching and permutation
 839 program for dealing with label switching and multimodality in analysis of population
 840 structure. *Bioinformatics* **23**: 1901-1906.
- 841 Jin Y, Szabo L, Carson M, 2010. Century-old mystery of *Puccinia striiformis* life history
 842 solved with the identification of *Berberis* spp. as an alternate host. *Phytopathology*
 843 **100**: 432-435.
- 844 Johnson R, Taylor AJ, 1976. Yellow rust of wheat. Plant Breeding Institute Cambridge.
 845 *Annual Report* 106-109.

- 846 Jombart T, Devillard S, Balloux F, 2010. Discriminant analysis of principal components:
847 a new method for the analysis of genetically structured populations. *BMC Genetics*
848 **11**: 94. doi:10.1186/1471-2156-11-94
- 849 Justesen AF, Ridout CJ, Hovmøller MS, 2002. The recent history of *Puccinia striiformis*
850 f. sp. *tritici* in Denmark as revealed by disease incidence and AFLP markers. *Plant*
851 *Pathology* **51**: 13-23.
- 852 Kamoun S, 2007. Groovy times: Filamentous pathogen effectors revealed. *Current*
853 *Opinion in Plant Biology* **10**: 358-365.
- 854 Kirk H, Freeland JR, 2011. Applications and implications of neutral versus non-neutral
855 markers in molecular ecology. *International Journal of Molecular Sciences* **12**: 3966-
856 3988.
- 857 Krogh A, Larsson B, von Heijne G, Sonnhammer EL, 2001. Predicting transmembrane
858 protein topology with a hidden Markov model: application to complete genomes.
859 *Journal of Molecular Biology* **305**: 567-580.
- 860 Line RF, 2002. Stripe rust of wheat and barley in North America: a retrospective
861 historical review. *Annual Review of Phytopathology* **40**: 75-118.
- 862 Line RF, Qayoum A, 1992. Virulence, aggressiveness, evolution, and distribution of
863 races of *Puccinia striiformis* (the cause of stripe rust of wheat) in North America,
864 1968-87. Technical Bulletin Number 1788. United States Department of Agriculture,
865 Agricultural Research Service.
- 866 Ling P, Wang M, Chen XM, Campbell KG, 2007. Construction and characterization of a
867 full-length cDNA library for the wheat stripe rust pathogen (*Puccinia striiformis* f. sp.
868 *tritici*). *BMC Genomics* **8**: 145. doi:10.1186/1471-2164-8-145

- 869 Little R, Manners JG, 1969. Somatic recombination in yellow rust of wheat (*Puccinia*
870 *striiformis*).I. Production and possible origin of 2 new physiologic races. *Transactions*
871 *of the British Mycological Society* **53**: 251-258.
- 872 Liu B, Chen XM, Kang Z, 2012. Gene sequences reveal heterokaryotic variations and
873 evolutionary mechanisms in *Puccinia striiformis*, the stripe rust pathogen. *Open*
874 *Journal of Genomics* **1**: 1.
- 875 Luo HY, Wang XJ, Zhan GM, Wei GR, Zhou XL, Zhao J, Huang LL, Kang ZS, 2015.
876 Genome-wide analysis of simple sequence repeats and efficient development of
877 polymorphic SSR markers based on whole genome re-sequencing of multiple
878 isolates of the wheat stripe rust fungus. *PLoS one* **10**: e0130362.
- 879 Michalakis Y, Excoffier L, 1996. A generic estimation of population subdivision using
880 distances between alleles with special reference for microsatellite loci. *Genetics* **142**:
881 1061-1064.
- 882 Milus EA, Kristensen K, Hovmøller MS, 2009. Evidence for increased aggressiveness in
883 a recent widespread strain of *Puccinia striiformis* f. sp. *tritici* causing stripe rust of
884 wheat. *Phytopathology* **99**: 89-94.
- 885 Morin PA, Martien KK, Taylor BL, 2009. Assessing statistical power of SNPs for
886 population structure and conservation studies. *Molecular Ecology Resources* **9**: 66-
887 73.
- 888 Nei M, 1972. Genetic distance between populations. *The American Naturalist* **106**: 283-
889 292.
- 890 Park RF, Wellings CR, 2012. Somatic hybridization in the Uredinales. *Annual Review of*
891 *Phytopathology* **50**: 219-239.

- 892 Peakall R, Smouse PE, 2006. GENALEX 6: genetic analysis in Excel. Population
893 genetic software for teaching and research. *Molecular Ecology Notes* **6**: 288-295.
- 894 Peakall R, Smouse PE, 2012. GENALEX 6: genetic analysis in Excel. Population
895 genetic software for teaching and research-an update. *Bioinformatics* **28**: 2537-2539.
- 896 Petersen TN, Brunak S, von Heijne G, Nielsen H, 2011. SignalP 4.0: discriminating
897 signal peptides from transmembrane regions. *Nature Methods* **8**: 785-786.
- 898 Petre B, Joly DL, Duplessis S, 2014. Effector proteins of rust fungi. *Frontiers in Plant*
899 *Science* **5**: Article 416 (1-7).
- 900 Pritchard JK, Stephens M, Donnelly P, 2000. Inference of population structure using
901 multilocus genotype data. *Genetics* **155**: 945-959.
- 902 Raymond M, Rousset F, 1995. GENEPOP (version 1.2): population genetics software
903 for exact tests and ecumenicism. *The Journal of Heredity* **86**: 248-249.
- 904 Rohlf FJ, 2009. NTSYSpc: numerical taxonomy system, ver. 2.21c. Exeter Software:
905 Setauket: New York.
- 906 Rosenberg NA, 2004. DISTRUCT: a program for the graphical display of population
907 structure. *Molecular Ecology Notes* **4**: 137-138.
- 908 Sharma-Poudyal D, Chen XM, Rupp R, 2014. Potential oversummering and
909 overwintering regions for the wheat stripe rust pathogen in the contiguous United
910 States. *International Journal of Biometeorology* **58**: 987-997.
- 911 Slatkin M, 1995. A measure of population subdivision based on microsatellite allele
912 frequencies. *Genetics* **139**: 1463.

- 913 Stubbs RW, 1985. Stripe rust. In: *The cereal rusts, vol. 2, Diseases, Distribution,*
 914 *Epidemiology, and Control* (eds Roelfs AP, Bushnel WR), pp. 61-101. Academic
 915 Press, Orlando.
- 916 Stukenbrock EH, McDonald BA, 2008. The origins of plant pathogens in agro-
 917 ecosystems. *Annual Review of Phytopathology* **46**: 75-100.
- 918 Wan AM, Chen XM, 2012. Virulence, frequency, and distribution of races of *Puccinia*
 919 *striiformis* f. sp. *tritici* and *P. striiformis* f. sp. *hordei* identified in the United States in
 920 2008 and 2009. *Plant Disease* **96**: 67-74.
- 921 Wan AM, Chen XM, 2014. Virulence characterization of *Puccinia striiformis* f. sp. *tritici*
 922 using a new set of *Yr* single-gene line differentials in the United States in 2010. *Plant*
 923 *Disease* **98**: 1534-1542.
- 924 Wan AM, Chen XM, Yuen J, 2016. Races of *Puccinia striiformis* f. sp. *tritici* in the United
 925 States in 2011 and 2012 and comparison with races in 2010. *Plant Disease*
 926 <http://apsjournals.apsnet.org/doi/pdf/10.1094/PDIS-10-15-1122-RE>
- 927 Wang MN, Chen XM, 2013. First report of Oregon grape (*Mahonia aquifolium*) as an
 928 alternate host for the wheat stripe rust pathogen (*Puccinia striiformis* f. sp. *tritici*)
 929 under artificial inoculation. *Plant Disease* **97**: 839.
- 930 Wang MN, Chen XM, 2015. Barberry does not function as an alternate host for *Puccinia*
 931 *striiformis* f. sp. *tritici* in the US Pacific Northwest due to teliospore degradation and
 932 barberry phenology. *Plant Disease* **99**:1500-1506.
- 933 Wang MN, Wan AM, Chen XM, 2015. Barberry as alternate host is important for
 934 *Puccinia graminis* f. sp. *tritici* but not for *Puccinia striiformis* f. sp. *tritici* in the U. S.
 935 Pacific Northwest. *Plant Disease* **99**:1507-1516.

- Wellings CR, 2011. Global status of stripe rust: a review of historical and current threats. *Euphytica* **179**: 129-141.
- Wellings CR, McIntosh RA, 1990. *Puccinia striiformis* f. sp. *tritici* in Australasia-pathogenic changes during the first 10 years. *Plant Pathology* **39**: 316-325.
- Wright S, 1965. The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution* **19**: 395-420.
- Xia CJ, Wang MN, Wan AM, Jiwan DA, See DR, 2016. Association analysis of SP-SNPs and avirulence genes in *Puccinia striiformis* f. sp. *tritici*, the wheat stripe rust pathogen. *American Journal of Plant Sciences* **7**:126-137.
- Yin CT, Chen XM, Wang X, Han Q, Kang Z, Hulbert S, 2009. Generation and analysis of expression sequence tags from haustoria of the wheat stripe rust fungus *Puccinia striiformis* f. sp. *tritici*. *BMC Genomics* **10**: 626.
- Zhan GM, Wang FP, Luo HY, Jiang SC, Zheng WM, Huang LL, Kang ZS, 2015. Screening simple sequence repeat markers in *Puccinia striiformis tritici* based on genomic sequence. *Journal of Zhejiang University-Science B* **16**:727-732.
- Zhang Y, Qu Z, Zheng W, Liu B, Wang X, Xue X, Xu L, Huang L, Han Q, Zhao J, Kang Z, 2008. Stage-specific gene expression during urediniospore germination in *Puccinia striiformis* f. sp. *tritici*. *BMC Genomics* **9**: 203.
- Zhao J, Wang L, Wang ZY, Chen XM, Zhang HC, Yao JN, Zhan GM, Chen W, Huang LL, Kang ZS, 2013. Identification of eighteen *Berberis* species as alternate hosts of *Puccinia striiformis* f. sp. *tritici* and virulence variation in the pathogen isolates from natural infection of barberry plants in China. *Phytopathology* **103**: 927-934.

Figure captions

Fig. 1 Epidemiological regions of stripe rust on wheat, caused by *Puccinia striiformis* f. sp. *tritici* in the United States (adapted from Wan and Chen 2014; for details, refer to Line and Qayoum 1992; Chen 2005; Chen *et al.* 2010). The number in parentheses is the number of isolates from each epidemiological region used in this study. The areas of the circles represent the percentage of total isolates from each region clustered into each genetic lineage based on discriminant analysis of principal components (DAPC) and STRUCTURE analysis.

Fig. 2 Discrimination of *Puccinia striiformis* f. sp. *tritici* (*Pst*) isolates by discriminant analysis of principal components (DAPC) and STRUCTURE analysis. **A)** Three clusters of *Pst* isolates were identified from the DAPC analysis. **B)** Population structure inferred by Bayesian clustering. Each individual is shown as a thin vertical line partitioned into *K* colored components representing inferred membership in *K* genetic clusters. *Pst* isolates are organized by epidemiological regions given at the top and separated by black lines.

Fig. 3 Principal component analysis based on SP-SNP genotypes of 352 *Puccinia striiformis* f. sp. *tritici* isolates. Percentage in brackets indicates the variance explained by each eigenvector. GL = genetic lineage based on the clustering analysis using the DAPC analysis and STRUCTURE program.

Table 1 - Number of *Puccinia striiformis* f. sp. *tritici* isolates, heterozygous marker allele frequency, polymorphic marker loci frequency, private marker alleles, number of multi-locus genotypes (MLGs), frequencies genetic groups and probability values for the Hardy-Weinberg equilibrium ($P_{gen(fis)}$) and sexual reproduction event ($P_{sex(fis)}$) tests for each epidemiological region.

Epidemiology region	No. of isolates	Polymorphic loci (%) ^a	No. of MLGs	Private		Heterozygous frequency (%) ^b	Frequency of genetic group ^c (%)			$P_{gen(fis)}$ ^d	$P_{sex(fis)}$ ^d
				No. of MLG	No. of individuals (%)		GL1	GL2	Admix		
R1	111	97.9	102	94	95 (85.6)	32.3 (15.6-75.0)	15 (13.5)	83 (74.8)	13 (11.7)	$10^{-7} - 10^{-28}$	$10^{-5} - 10^{-208}$
R2	20	95.9	20	16	16 (80)	45.3 (20.6-74.0)	5 (25.0)	10 (50.0)	5 (25.0)	$10^{-10} - 10^{-22}$	$10^{-11} - 10^{-62}$
R3	31	96.9	29	25	26 (83.9)	33.1 (16.1-75.0)	5 (16.1)	23 (74.2)	3 (9.7)	$10^{-7} - 10^{-31}$	$10^{-6} - 10^{-58}$
R4	6	37.1	4	3	5 (83.3)	23.6 (19.6-33.0)	0	6 (100)	0	$10^{-2} - 10^{-5}$	$10^{-1} - 10^{-3}$
R5	32	94.9	27	26	32 (100)	41.2 (20.6-69.1)	2 (6.2)	10 (31.3)	20 (62.5)	$10^{-7} - 10^{-21}$	$10^{-6} - 10^{-62}$
R6	42	95.9	28	22	22 (52.4)	61.5 (22.9-70.8)	33 (78.6)	4 (9.5)	5 (11.9)	$10^{-5} - 10^{-31}$	$10^{-4} - 10^{-115}$
Western US	242	96.9	198	178	201 (83.1)	33.1 (15.6-75.0)	60 (24.8)	136 (56.2)	46 (19.0)	$10^{-11} - 10^{-25}$	$10^{-8} - 0$
R7	29	95.9	19	11	11 (37.9)	66.3 (36.5-76.3)	28 (96.6)	1 (3.4)	0	$10^{-3} - 10^{-37}$	$10^{-1} - 10^{-36}$
R8	25	94.9	15	9	9 (36)	65.5 (20.6-72.6)	24 (96.0)	1 (4.0)	0	$10^{-3} - 10^{-43}$	$10^{-2} - 10^{-34}$
R9	28	92.8	17	11	11 (39.3)	66.2 (53.6-69.8)	27 (96.4)	0	1 (3.6)	$10^{-3} - 10^{-36}$	$10^{-1} - 10^{-31}$
R10	4	92.8	4	1	1 (25)	52.6 (22.7-67.0)	2 (50.0)	1 (25.0)	1 (25.0)	$10^{-11} - 10^{-16}$	$10^{-14} - 10^{-14}$
R11	14	92.8	8	5	5 (35.7)	64.8 (27.8-70.8)	13 (92.9)	1 (7.1)	0	$10^{-4} - 10^{-37}$	$10^{-3} - 10^{-25}$
R12	10	82.5	9	3	3 (30)	67.3 (64.9-72.2)	10 (100)	0	0	$10^{-3} - 10^{-50}$	$10^{-2} - 10^{-13}$
Eastern US	110	90.7	55	45	49 (44.5)	63.8 (20.6-76.3)	104 (94.5)	4 (3.6)	2 (1.9)	$10^{-2} - 10^{-50}$	$10^{-1} - 10^{-108}$
Total (mean)	352	88.8	242			51.5 (15.6-76.3)	164 (46.6)	139 (39.5)	49 (13.9)	$10^{-13} - 10^{-25}$	$10^{-10} - 0$

^a Percentage of polymorphic loci was determined based on the number of loci out of all marker loci showing differences among the isolates in each epidemiological region.

^b Mean of heterozygous loci over 92 SP-SNP loci, values in the parenthesis are range.

^c Genetic groups are those identified using the discrimination analysis by principal components (DAPC) and STRUCTURE program. GL1 = Genetic Group 1, GL2 = Genetic Group 2, and the admix group contains isolates have probabilities belonging to either group less than 80%.

^d The probability values were determined using the program GenClone 2.0 (Arnaud-Haond & Belkhir 2007; Arnaud-Haond *et al.* 2007).

994 **Table 2 - Clusters of SP-SNP genotypes of 352 *Puccinia striiformis* f. sp. *tritici* isolates from the United States in 2010**
 995 **and 2011 on the basis of the discriminant analysis of principal components (DAPC) compared with clustering derived**
 996 **from neighbor joining procedure from NTSYSpc 2.21L.**

DAPC genetic lineage ^a	Neighbor-joining genetic group ^b	Number of isolates	Frequency (%) in epidemiological regions											
			R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12
GL1	_N GL1A	164	13.5	25.0	16.1		6.2	78.6	96.6	96.0	96.4	50.0	92.9	100
Admix	_N GL1B	15	5.3	10.0			3.1	9.5						
Admix	_N GL2	24	1.8	10.0			59.4	2.4						
GL2	_N GL3A	139	74.8	50.0	74.2	100	31.3	9.5	3.4	4.0		25.0	7.1	
Admix	_N GL3B	10	3.6	5.0	9.7				3.4		3.6	25.0		
Total number of isolates		352	111	20	31	6	32	42	29	25	28	4	14	10

997 ^a DAPC genetic lineages are shown in Fig. 2.

998 ^b Neighbor-joining genetic lineages are shown in Supporting Fig. **S1**.

999

1000 **Table 3 - Average of single-locus statistics of *Puccinia striiformis* f. sp. *tritici* populations from 12 epidemiological**
 1001 **regions (R1-R12) in the United States at 92 SP-SNP loci.**

Epid. region	Observed alleles, N_a	Effective alleles, A_e^c	Allelic richness ^d	Shannon's information index, I^e	Observed heterozygosity, H_o^f	Expected heterozygosity, H_e^g	Unbiased expected heterozygosity, uH_e^h	Fixation index, F_{is}^i
R1	1.979 (0.015) ^b	1.490 (0.032)	1.691 (0.218)	0.465 (0.017)	0.323 (0.024)	0.301 (0.014)	0.302 (0.014)	-0.048 (0.041)
R2	1.959 (0.020)	1.629 (0.027)	1.914 (0.229)	0.543 (0.016)	0.453 (0.023)	0.367 (0.012)	0.377 (0.012)	-0.215 (0.039)
R3	1.969 (0.018)	1.486 (0.033)	1.776 (0.327)	0.458 (0.018)	0.331 (0.026)	0.296 (0.015)	0.301 (0.015)	-0.084 (0.038)
R4	1.371 (0.049)	1.244 (0.040)	1.731 (0.398)	0.199 (0.029)	0.235 (0.040)	0.135 (0.020)	0.147 (0.022)	-0.571 (0.048)
R5	1.948 (0.023)	1.577 (0.035)	1.817 (0.185)	0.485 (0.023)	0.410 (0.034)	0.328 (0.018)	0.333 (0.018)	-0.173 (0.056)
R6	1.959 (0.020)	1.706 (0.038)	1.687 (0.235)	0.544 (0.022)	0.613 (0.039)	0.377 (0.017)	0.381 (0.018)	-0.441 (0.059)
R7	1.907 (0.030)	1.691 (0.045)	1.604 (0.277)	0.506 (0.028)	0.660 (0.044)	0.355 (0.021)	0.361 (0.022)	-0.664 (0.046)
R8	1.948 (0.023)	1.698 (0.043)	1.715 (0.312)	0.520 (0.026)	0.654 (0.044)	0.362 (0.020)	0.369 (0.021)	-0.505 (0.070)
R9	1.928 (0.026)	1.692 (0.044)	1.802 (0.257)	0.508 (0.028)	0.661 (0.044)	0.356 (0.021)	0.363 (0.022)	-0.634 (0.052)
R10	1.928 (0.026)	1.714 (0.029)	1.731 (0.381)	0.569 (0.019)	0.525 (0.036)	0.394 (0.014)	0.451 (0.016)	-0.282 (0.066)
R11	1.928 (0.026)	1.703 (0.042)	1.753 (0.347)	0.528 (0.025)	0.647 (0.042)	0.367 (0.019)	0.381 (0.020)	-0.519 (0.066)
R12	1.825 (0.039)	1.683 (0.045)	1.737 (0.373)	0.496 (0.029)	0.671 (0.045)	0.350 (0.022)	0.368 (0.023)	-0.790 (0.037)
Total ^a	1.887 (0.009)	1.609 (0.012)	1.839 (0.157)	0.485 (0.007)	0.515 (0.012)	0.332 (0.006)	0.345 (0.006)	-0.394 (0.017)

1002 ^a Total of single-locus statistics was calculated from 352 isolates from the United States at 92 SP-SNP loci.

1003 ^b Standard error.

1004 ^c Number of effective alleles, $A_e = 1/\sum p_i^2$, where p_i is the frequency of the i th allele.

1005 ^d Allelic richness = $\sum [1 - ((N - N_i)/n)/(N/n)]$ where N_i = number of i th SP-SNP genotype among the N sampled individuals within a population,
 1006 and n = smallest population size

1007 ^e Shannon's information index, $I = -\sum p_i (\ln p_i)$, where $\ln$ is the natural logarithm.

1008 ^f Observed heterozygosity, H_o = number of heterozygotes/ N .

1009 ^g Expected heterozygosity, $H_e = 1 - \sum p_i^2$.

1010 ^h Unbiased expected heterozygosity, $uH_e = 2N(1 - \sum p_i^2)/2N - 1$.

1011 ⁱ Fixation index, $F_{is} = H_e - H_o/H_e$.

1012

Table 4 - Analysis of genetic variance for populations of *Puccinia striiformis* f. sp. *tritici* from 12 epidemiological regions from the United States for genetic differentiation based on allele frequencies at 92 SP-SNP loci and virulence to 18 single-gene differential lines of wheat.

Source of variation ^a	df	Statistics ^b	Sum of squares	Mean squares	Estimated variance	Variation (%)	Statistic value	P value
SP-SNP variation								
Among W/E US	1	F_{RT}	1151.07	1151.07	3.40	16	0.157	0.001
Among regions within W/E US	10	F_{SR}	904.25	90.43	1.41	6	0.078	0.001
Within regions	692	F_{ST}	11626.49	16.80	16.80	78	0.223	0.001
Among W/E US	1	R_{RT}	4702.37	4702.37	13.86	16	0.157	0.001
Among regions within W/E US	10	R_{SR}	3760.92	376.09	5.90	7	0.079	0.001
Within regions	692	R_{ST}	47434.92	68.55	68.55	77	0.224	0.001
Virulence variation								
Among W/E US	1	Φ_{RT}	380.96	380.96	2.12	21	0.209	0.001
Among regions within W/E US	10	Φ_{PR}	437.60	43.76	1.43	14	0.177	0.001
Within regions	340	Φ_{PT}	2252.80	6.63	6.63	65	0.349	0.001

^a W/E US = western United States or eastern United States; ERs = epidemiological regions in the North America.

^b $F_{RT} = V_{AR} / (V_{WP} + V_{AP} + V_{AR})$, in which V_{AR} = variance among regions, V_{AP} = variance among populations within regions, and V_{WP} = variance within populations.

$$F_{SR} = V_{AP} / (V_{AP} + V_{WP}).$$

$$F_{ST} = (V_{AP} + V_{AR}) / (V_{AP} + V_{WP} + V_{AR}).$$

$$R_{RT} = V_{AR} / (V_{WP} + V_{AP} + V_{AR}).$$

$$R_{SR} = V_{AP} / (V_{AP} + V_{WP}).$$

$$R_{ST} = (V_{AP} + V_{AR}) / (V_{AP} + V_{WP} + V_{AR}).$$

$$\Phi_{RT} = V_{AR} / (V_{WP} + V_{AP} + V_{AR}).$$

$$\Phi_{PR} = V_{AP} / (V_{AP} + V_{WP}).$$

$$\Phi_{PT} = (V_{AP} + V_{AR}) / (V_{AP} + V_{WP} + V_{AR}).$$

Table 5 - Heterozygosity in genetic groups of *Puccinia striiformis* f. sp. *tritici* determined by the discrimination analysis by principal components with 352 isolates based on 92 SP-SNP loci.

Genetic group	Isolates		Heterozygosity (%)		P-value
	No.	%	Range	Mean $\pm$ SD	
GL1	164	55.14	53.1 – 76.3	66.8 $\pm$ 2.6	
GL2	139	35.27	17.7 – 42.7	24.5 $\pm$ 4.7	
Admix	49	9.59	15.6 – 75.0	49.0 $\pm$ 13.2	
Overall	352	100.00	15.6 – 76.3	47.6 $\pm$ 20.5	< 0.0001

Supporting information

Supplementary figures

Fig. S1 Dendrogram clustering 352 *Puccinia striiformis* f. sp. *tritici* isolates into three major molecular groups with two sub-groups in each of groups 1 and 3 based on 92 SP-SNP markers using neighbor-joining clustering method in NTSYSpc 2.21L. **(A)** Showing the relationships of the groups and **(B-E)** segments of the dendrogram showing isolates in each group.

Fig. S2 Simplified dendrogram showing two super genetic lineages (GL1 and GL2) and six sub-lineages (GL1A, GL1B, GL1C, GL2A and GL2B) of 352 *Puccinia striiformis* f. sp. *tritici* isolates based on 92 SP-SNP markers using unweighted pair group arithmetic mean clustering method. GL = genetic lineage.

Fig. S3 Clustering of *Puccinia striiformis* f. sp. *tritici* isolates in the United States inferred by STRUCTURE 2.3.4 based on the Bayesian clustering model. **(A)** The best number of population subdivisions was determined based on the second order rate of change of the log probability of data between successive K values. The best K = 2 as indicated by the arrow. **(B-C)** Estimated membership coefficients of the 352 isolates (the X-axis) by STRUCTURE 2.3.4 based on the Bayesian clustering model. Isolates are represented by single vertical bars and listed by epidemiological regions (R); **(B)** R1-R6 in the west US and **(C)** R7-R12 in the eastern US (see Fig. 1 for the regions). The white and dark gray segments stand for the inferred membership fractions in genetic group 1 and

genetic group 2, respectively. The isolates with estimated membership in either population less than 0.80 were indicated by stars above the isolate segments but underneath horizontal solid bars.

Fig. S4. Gene flow of *Puccinia striiformis* f. sp. *tritici* among epidemiological regions (R) in the United States. Numbers within parenthesis are number of isolates. **(A)** Based on shared multi-locus genotypes (MLGs). Three common MLGs are MLG002, MLG015, and MLG037. **(B)** Based on shared races. Two common races are PSTv-11 and PSTv-37.

Supplementary tables

Table S1 Races of 352 *Puccinia striiformis* f. sp. *tritici* isolates from the United States in 2010 and 2011 clustered into molecular groups using the neighbor joining method

Table S2 Genetic distances and differentiations among populations of *Puccinia striiformis* f. sp. *tritici* in epidemiological regions based on the number of effective migrants (N_m), F_{ST} , R_{ST} and Nei's genetic distance (D) assessed with 92 SP-SNP loci and Φ_{PT} and D assessed with avirulence/virulence patterns on 18 Yr single-gene differential lines of wheat

Table S3 Correlation coefficients between F_{ST} , R_{ST} and Φ_{PT} measures of genetic differentiation and Nei's (D) genetic distance based on allelic variation at 92 SP-SNP loci

and for virulence to 18 single-gene differential lines of wheat in *Puccinia striiformis* f. sp. *tritici* populations from the United States

Data tables

Table D1 SP-SNPs and their NCBI accession numbers, EST GenBank accession numbers, genome supercontigs of SNP sites and primer sequences (Excel file)

Table D2 Races, origin (state and region), genotype, multi-locus genotypes (MLG), genetic groups, heterozygous frequency (%), and genetic group of 352 *Puccinia striiformis* f. sp. *tritici* isolates collected from the U.S. and their SNP genotypes at 92 SP-SNP marker loci (Excel file)

Table D3 Infection types of *Puccinia striiformis* f. sp. *tritici* isolates on *Yr* single-gene differentials and their race and origin

Table D4 Observed heterozygosity, expected heterozygosity, and unbiased heterozygosity over 92 SP-SNP loci estimated based on 352 *Puccinia striiformis* f. sp. *tritici* isolates

Fig. 1

Fig. 2

Fig. 3