Scientific
Research
Publishing

Association Analysis of SP-SNPs and Avirulence Genes in *Puccinia striiformis* f. sp. *tritici*, the Wheat Stripe Rust Pathogen

Chongjing Xia[1], Meinan Wang[1], Anmin Wan[1], Derick A. Jiwan[1], Deven R. See[1,2], Xianming Chen[1,2]*

[1]Department of Plant Pathology, Washington State University, Pullman, USA
[2]United States Department of Agriculture, Agricultural Research Service, Wheat Health, Genetics, and Quality Research Unit, Pullman, USA
Email: *xianming@wsu.edu

Abstract

Puccinia striiformis f. sp. *tritici* (*Pst*) is one of the pathogenic fungi on wheat, caused stripe rust that is a great threat for wheat production all over the world. Intensive efforts have been made to study genetics of wheat resistance to this disease, but few on avirulence of the pathogen due mainly to the nature of obligate biotrophism and the lack of systems for studying its genetics and molecular manipulations. To overcome these limitations, a natural *Pst* population comprising 352 isolates representative of a diverse virulence spectrum was genotyped using 97 secreted protein-single nucleotide polymorphism (SP-SNP) markers to identify candidate avirulence genes using association analysis. Among avirulence genes corresponding to 19 resistance genes, significantly associated SP-SNP markers were detected for avirulence genes *AvYr*1, *AvYr*2, *AvYr*6, *AvYr*7, *AvYr*8, *AvYr*44, *AvYrExp*2, *AvYrSP*, and *AvYrTye*. These results indicate that association analysis can be used to identify markers for avirulence genes. This study has laid the foundation for developing more SP-SNPs for mapping avirulence genes using segregating populations that can be generated through sexual reproduction on alternate hosts of the pathogen.

Keywords

---

*Corresponding author.

1. Introduction

Puccinia striiformis Westend. f. sp. *tritici* Erikss. (*Pst*) is an obligate biotrophic fungus causing stripe (yellow) rust mainly on wheat. Significant yield losses have been reported in most wheat production regions, and the pathogen still threatens wheat production worldwide [1] [2]. Intensive efforts have focused on monitoring the disease, identifying and incorporating resistance genes into wheat cultivars, and use of fungicides to manage the disease [3] [4]. However, no genes for pathogenicity or virulence have been characterized for *Pst*.

To understand the mechanism of the "boom and bust" cycle of the battle between avirulent/virulent pathogens and resistant/susceptible cultivars, in the middle 20th century Flor made a breakthrough by demonstrating the gene-for-gene concept that "for each *R* gene that conditions resistance in the host there is a corresponding *Avr* gene that conditions pathogenicity in the parasite" [5]. Since then, scientists have begun to search for the biochemical and molecular basis for the gene-for-gene concept, and the elicitor-receptor model has been proposed. Based on this model, the recognition of an elicitor (product of an *Avr* gene) in the pathogen by a receptor (product of the matching *R* gene) in the resistant host plant activates a signal transduction pathway that leads to a resistant response, such as a hypersensitive response (HR), in the host [6]. Interestingly, most of AVR proteins among filamentous pathogen effectors have a signal peptide region [7]. Particularly in flax rust, *Melampsora lini*, four putative *Avr* genes (*AvrL*567, *AvrM*, *Avr*4, and *AvrP*123) have been identified from 21 haustorially expressed secreted proteins (HESPs), and transient expression of these genes in flax causes resistance gene-mediated cell death, thus demonstrating their avirulence activities [8]. There are some exceptions. For example, two protein effectors in the stem rust fungus (*P. graminis* f. sp. *tritici*) may activate the R protein long before haustoria formation [9]. The *Pst*-wheat pathosystem well fits the "gene-for-gene" concept. Although numerous resistance genes in wheat have been identified with some cloned [10], no avirulence genes have been molecularly characterized.

Limited studies were conducted to identify candidate avirulence gene (*Avr*) in *Pst*. For example, Huang *et al.* [11] studied the expression profiling of 442 gene transcripts of *Pst* in *in vitro* and *Pst* infected wheat leaves using a custom-made *Pst* gene chip including secreted protein (SP) genes. The differentially expressed transcripts were involved in various biological processes. Several transcripts were expressed differentially between compatible and incompatible interactions, and thus could be virulence/avirulence related genes. Using quantitative RT-PCR analysis, Cantu *et al.* [12] identified 22 effector candidate genes which were significantly expressed during the infection in wheat. These candidate genes encoded secreted proteins whose transcripts were enriched in haustoria. Similarly, potential *Avr* genes were identified in cDNA libraries made from germinating urediniospores and fungal cells harvested during the infection processes by examining expression profiles of putatively secreted protein genes [13] [14]. All of these studies to identify candidate *Avr* genes were based on gene expression levels in germinated urediniospores and in infected wheat tissues, while phenotypic virulence profiles were rarely considered. In the present study, we attempted to identify *Avr* genes, corresponding to resistance genes in wheat, through association analysis between genetic variation and virulence/avirulence phenotypes taking the advantages of SP-single nucleotide polymorphisms (SP-SNPs).

SP-SNPs derived from expressed sequence tags (ESTs) would be promising markers for identifying avirulence genes since some SP-SNP markers may be closely associated with or directly in the coding regions, and thus more likely to be involved in causal alleles or highly correlated with the causal alleles. It is possible that the effectors in *Pst* are also highly expressed secreted proteins (HESPs) as in the study with *M. lini* [8], because *M. lini* and *Pst* are related rust fungal species, and both fungi are obligate biotrophic parasites and obtain their nutrient from living hosts through haustoria. Moreover, most of the identified filamentous pathogen effectors have a signal peptide domain [7]. This would be the first study to identify *Avr* genes in *Pst* using an association mapping approach based on SP-SNP markers, and may lead to a feasible approach for identification and characterization of *Avr* genes in *Pst*, which is genetically and molecularly hard to manipulate.

2. Materials and Methods

2.1. Isolate Selection

A total of 352 *Pst* isolates used in the present study were selected from the *P. striiformis* collections made in 2010 and 2011 in the US. These isolates were selected to cover as many races as possible, thus representing a diverse combination of avirulence/virulence profiles.

2.2. Virulence Tests

The *Pst* isolates were tested for their avirulence/virulence patterns using the 18 *Yr* single-gene lines and supplementary differentials, as described by Wan and Chen [15]. The avirulence or virulence of isolates to a particular *Yr* resistance gene line was represented by an infection type (IT), which was scored using a 0-to-9 scale, with 0 as the most avirulent and 9 as the most virulent, as described previously [16]. ITs were recorded at 20 to 22 days after inoculation. For association analysis in this study, a phenotypic trait of an isolate to a particular *Yr* gene was defined as avirulence when IT was 0 - 4 and virulent when IT was 5 - 9. Totally, 19 phenotypic traits (18 *Yr* single-gene differentials plus one supplementary differential supposed to have a single stripe rust resistance gene, *Yr*2) were analyzed for associations with SP-SNP markers. These 19 resistance genes were: *Yr*1, *Yr*2, *Yr*5, *Yr*6, *Yr*7, *Yr*8, *Yr*9, *Yr*10, *Yr*15, *Yr*17, *Y*24, *Yr*27, *Yr*32, *Yr*43, *Yr*44, *YrExp*2, *YrSP*, *YrTr*1, and *YrTye*. Their corresponding avirulence genes were symbolized as *AvYr*1, *AvYr*2, and so on.

2.3. DNA Extraction

Genomic DNA was extracted using a modified CTAB method, as described previously [17]. The DNA stock solutions were diluted to 5 ng/µL with sterilized ddH$_2$O. The DNA quality and concentration of the diluted working solution were checked by PCR amplification using the internal transcribed spacer (ITS) primers ITS3 and ITS4. The quantified working solution was used for SP-SNP genotyping.

2.4. Isolates Genotyping

Genotyping of 352 *Pst* 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. Then shrimp alkaline phosphatase (SAP) reaction was conducted in which phosphatase was used to eliminate unincorporated dNTPs from the previous reaction. A single base extension reaction was conducted using modified dideoxynucleotide terminators and thermosequenase enzyme, which produced two extension products differing only in the last one base for each SNP locus. The SP-SNP alleles were identified based on the distinct masses of the extended primers using the matrix-assisted laser desorption ionization-time-of-flight (MALDI-TOF) mass spectrometry. The automatically generated genotype calls were checked manually and overwritten based on the intensity peaks of the alleles in the mass spectrum using Spectro TYPER (SEQUENOM, Inc. San Diego, California). Generally, the locus having an intensity peak of one allele which was 2 times higher than the other one was considered as a homozygous locus, otherwise as heterozygous. Also, the background noise was taken into consideration for overwriting the automatic calls.

2.5. Data Analyses

The SP-SNP data were examined to determine the minor allele frequency (MAF), number of heterozygotes, and number of missing alleles. Correlations between SP-SNPs and the 19 virulence/avirulence traits of the *Pst* isolates were analyzed, and a Spearman correlation matrix was generated in the program Genome Association and Prediction Integrated Tool (GAPIT) [18]. The IT distribution of isolates of the 19 virulence/avirulence traits was also determined.

The mixed linear model (MLM) including both fixed and random effects was incorporated in the program GAPIT and used for association analyses. By including individuals as random effects, the MLM model was able to incorporate relationships among isolates in the association analysis. The relationships among isolates were represented as a kinship matrix computed using the VanRaden method. Specifically, two MLM scenarios were performed in this study, a naive scenario and a basic scenario with principal component analysis (PCA) and kinship. The naive scenario was achieved by setting to group using commend "group. from = 0 and group. to = 0". By doing this, the scenario was less powerful by ignoring the fixed effects such as population structure, but served as a control for comparison with the basic scenario with PCA and kinship. While the basic scenario incorporated PCs as fixed effects and used to group kinship matrix for isolates relationships [18]. The commands were: "myGAPIT <- GAPIT(Y = myY, G = myG, group. to = 0, group. from = 0)" and "myGAPIT <- GAPIT(Y = myY, G = myG, PCA. total = 3, Model. selection = T, group. from = 351, group. to = 351)" for naive scenario and basic scenario with PCA and kinship, respectively. The false discovery rate-adjusted *P* values were applied to reduce the false-positive rate in the multiple testing carried out in the association analyses [19].

3. Results

3.1. Infection Type Distribution

IT was used to qualitatively describe the virulence of *Pst* isolates. However, the IT data was not completely qualitative or quantitative, and no normal distribution was expected, as shown in **Table 1**. The balanced distribution of avirulence (e.g. IT 0 - 4) and virulence (e.g. IT 5 - 9) was expected for association and would avoid spurious association to some extent. For example, 193 out of 352 *Pst* isolates in the present study were highly avirulent (IT 1 - 2) to *Yr*1, and 157 were highly virulent to *Yr*1 (IT 8) while only 2 had an intermediate IT. In this case, the avirulence/virulence distribution for *Yr*1 was relatively balanced, nearly 1:1 for avirulent to virulent isolates. A similar case was observed for *YrTye* (**Table 1**). However for some virulence phenotypes, the avirulence/virulence distributions were not balanced. For example, all isolates used in this study and those collected thus far from the US were avirulent to *Yr*5 and *Yr*15 [1] [15]. Therefore, these two traits were not included for further analyses. Even though the virulence/avirulence distributions of 15 other traits were not exactly balanced, they were included in the association analysis as either avirulent or virulent isolates were more than 5%.

3.2. Allele Tests

The frequency distributions of minor alleles (MAF, minor allele frequency), heterozygous, and missing alleles were shown in **Figure 1**. The mean frequency of minor alleles was around 40%. The *Pst* population used in the present study had a high heterozygosity (data not shown), but largely within an acceptable range. All alleles with a frequency less than 0.05 were filtered prior to analysis. There were very few missing data points. Generally, these estimates indicated that the genotypes of the *Pst* population in this study were suitable for further analyses.

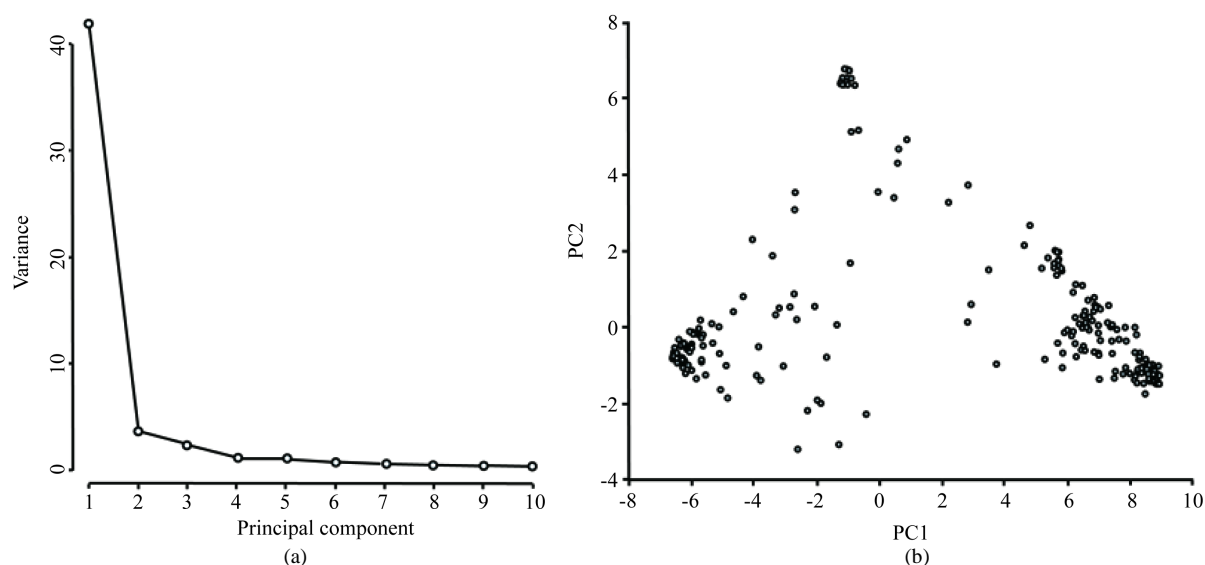### 3.3. Principal Component Analysis (PCA) and Kinship Analysis

Principal component analysis conducted using the GAPIT program indicated that the first two PCs explained the largest portions of the genetic variance, as shown in **Figure 2**. The Efficient Mixed Model Association (EMMA) algorithm was used to establish a kinship matrix, a heat map of values in the kinship matrix is shown in **Figure 3**.

Table 1. Distribution of infection types of 352 *Puccinia striiformis* f. sp. *tritici* isolates scored on 19 wheat lines with *Yr* single-genes.

Infection type	*Yr* single-genes																		
	*Yr*1	*Yr*2	*Yr*6	*Yr*7	*Yr*9	*Yr*10	*Yr*17	*Yr*24	*Yr*27	*Yr*32	*Yr*43	*Yr*44	*Yr*SP	*Yr*Tr1	*Yr*Exp2	*Yr*5	*Yr*15	*Yr*8	*Yr*Tye
Avirulence																			
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	191	1	3	5	18	181	25	39	6	58	23	2	321	67	3	347	350	10	184
2	2	2	11	40	6	138	15	250	26	215	60	39	5	69	49	5	2	37	20
3	0	5	11	8	0	3	19	38	27	55	21	7	1	6	8	0	0	5	0
4	1	4	3	3	4	1	6	1	7	1	8	3	0	2	0	0	0	2	1
Sub-total	194	12	28	56	28	323	65	328	66	329	112	51	327	144	60	352	352	54	205
Virulence																			
5	1	4	3	1	2	0	47	4	34	5	13	3	0	15	9	0	0	18	0
6	0	12	3	3	1	0	41	2	28	0	6	1	0	3	1	0	0	27	1
7	0	38	9	6	3	6	125	4	74	2	98	10	2	31	4	0	0	98	3
8	157	286	309	286	318	23	74	14	150	16	123	287	23	159	278	0	0	155	143
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Sub-total	158	340	324	296	324	29	287	24	286	23	240	301	25	208	292	0	0	298	147
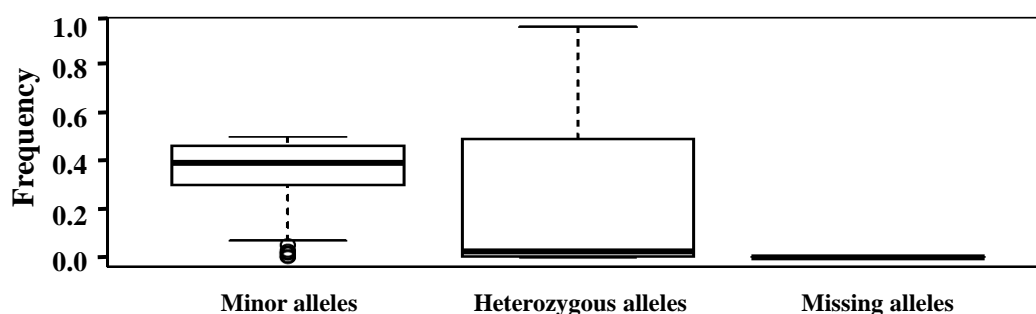
Figure 1. Boxplots of the frequencies of minor alleles and heterozygous loci and missing data for 352 *Puccinia striiformis* f. sp. *tritici* isolates at 97 SP-SNP loci.


Figure 2. Principal component analysis (PCA) of 352 *Puccinia striiformis* f. sp. *tritici* isolates using 97 SP-SNP markers. (a) The genetic variance explained by each principal component, up to 10. (b) Plot of the second principal component (PC2) against the first principal component (PC1).

Generally, two lineages were estimated for the *Pst* population in this study. Within each lineage, the isolates had a closer kinship to other isolates, indicating a strong population structure or stratification.

3.4. Thirty SP-SNPs Significantly Associated with Nine *Pst* Virulence Phenotypes

The two scenarios for association analyses in this study were compared in terms of quantile-quantile (QQ)-plot of P values and the optimum compressions. $AvYr1$ was taken as an example to illustrate the comparison between the two scenarios. In the QQ plot, the observed $-\log_{10}(P)$ values were plotted against the expected $-\log_{10}(P)$ values. Thus, the SP-SNPs under the null hypothesis of no association between the SP-SNP and the trait would be laid on the diagonal line in the QQ-plot. The SP-SNPs deviated from the diagonal line indicated potential association, either spurious due to population structure and isolates relatedness, or truly associated with the trait. The comparison of two scenarios with $AvYr1$ as an example was shown in **Figure 4**. Compared to naive scenario which did not take PCA and kinship into account for association analysis, much fewer SP-SNPs were deviated from the null hypothesis of no association in the basic scenario, indicating the spurious associations in the naive scenario. A low level of inflation of the test statistics in the basic scenario with PCA and kinship indicated a high efficiency in controlling spurious associations due potentially to the population structure and cryptic relatedness among the isolates. Similarly, for comparison of the optimum compression for $AvYr1$ using these two scenarios, a low ($-2 \times \log$) likelihood function in the basic scenario with PCA and kinship (**Figure 5**) with all *Pst* isolates grouped into 344 groups for kinship analysis, suggesting a high likelihood function, while only 4 groups
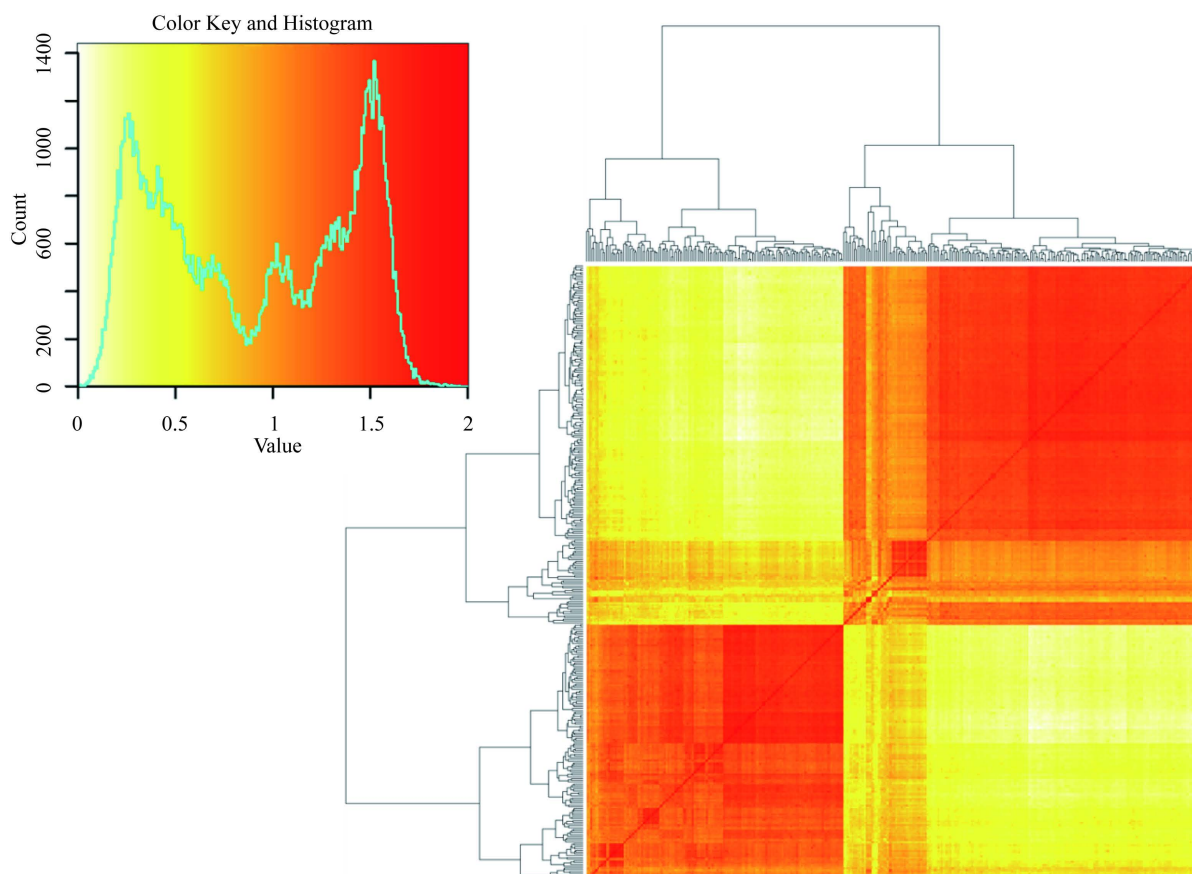
Figure 3. Kinship plot of 352 *Puccinia striiformis* f. sp. *tritici* isolates. The heat map was generated based on the values in the kinship matrix which was computed using the efficient mixed model association (EMMA) algorithm.
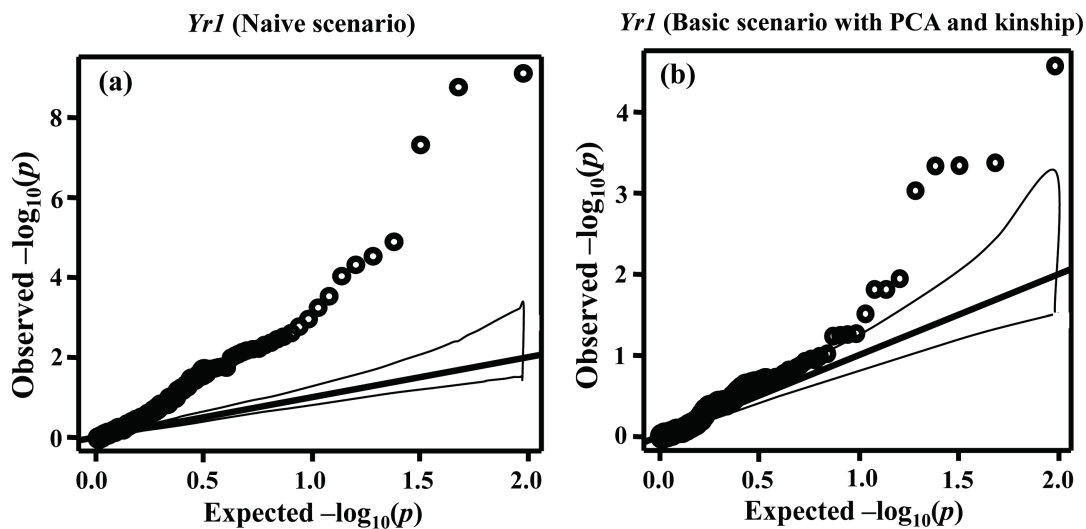
Figure 4. Comparison of quantile-quantile (QQ)-plot of *P*-values using different association models for the *AvYr*1 locus. In the QQ-plot, the observed negative base 10 logarithm of the *P*-values were against the expected observed negative base 10 logarithm of the *P*-values under assumption that the *P*-values follow a uniform [0,1] distribution. The dotted lines represented the 95% confidence interval for the QQ-plot under the null hypothesis of no association between the SP-SNP and the *AvYr* locus. (a) Computed using the naive scenario in the GAPIT program; (b) Computed using the basic scenario with PCA and kinship.
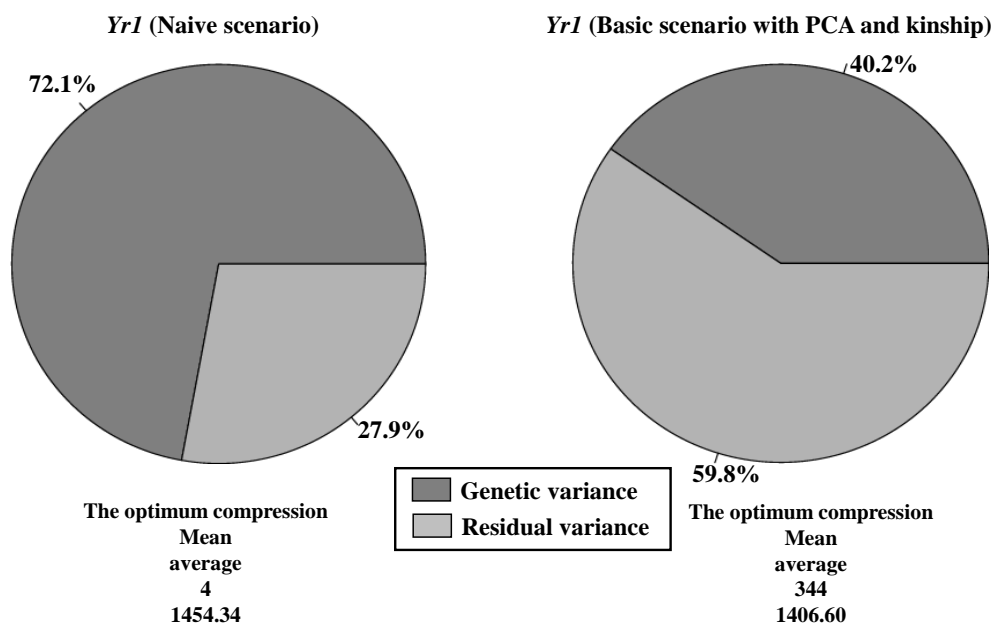
Figure 5. Comparison of the profile for the optimum compression using different association models for *AvYr*1. (Left) Computed using the naive scenario method. The optimal method to estimate group kinship was "Mean", the optimum clustering method was "average", the number of groups was 4, the value of $-2 \times$ log likelihood function was 1454.34, and the genetic variance (heritability) was 0.721. (Right) Computed using the basic scenario method with PCA and kinship. The optimal method to estimate group kinship was "Mean", the optimum clustering method was "average", the number of groups was 344, the value of $-2 \times$ log likelihood function was 1406.60, and the genetic variance (heritability) was 0.402.

were accounted for kinship. Evidence that the basic scenario with PCA and kinship was superior to the naive scenario model was indicated by comparisons of these two scenarios for other virulence traits (data not shown).

The mixed linear model using the basic scenario with PCA and kinship indicated associations between 30 SP-SNP markers and 9 avirulence genes (corresponding to *Yr* genes) with a false discovery rate (FDR)-adjusted *P* value lower than 0.05 (**Table 2**). These 9 genes were *AvYr*1, *AvYr*2, *AvYr*7, *AvYr*8, *AvYr*44, *AvYrExp*2, *AvYrSP*, and *AvYrTye*. For all these avirulence genes, each was significantly associated with two or more SP-SNPs, except *AvYr*2 and *AvYrExp*2. Also, some SP-SNPs were significantly associated with two or more avirulence genes. For example, significant associations were indicated between SP-SNP marker PstSNP00438 and *AvYr*1, *AvYr*7, and *AvYr*44. Specifically for *AvYrTye*, all three significantly associated SP-SNPs were also significantly associated with *AvYr*1. This result was not surprising since the virulence ITs of these traits were not completely independent. To confirm this, the correlation relationships among the 19 traits were estimated, and the results were given in **Figure 6**. The correlations between most of the avirulence genes were low or moderate in terms of correlation coefficients. However, high correlations were suggested between *AvYr*1 and *AvYrTye* (R = 0.90), and between *AvYr*7 and *AvYr*44 (R = 0.89).

4. Discussion

The preliminary results in the present study showed that association analysis can be a powerful tool for relating molecular markers to avirulence genes of *Pst*. Compared with previous gene expression studies [11]-[14], the association analysis takes the virulence phenotypes into consideration, instead of focusing only on differentiation of gene expression of the pathogen in the *Pst*-wheat interaction process. Therefore, it may provide direct genetic evidence for virulence variation. Even though only 97 SP-SNPs were used in this study, we found significant associations (P_FDR_adj < 0.05) for 41 SP-SNP *Pst* virulence associations involved in 9 *AvYr* genes and 30 SP-SNP loci (**Table 2**). Most of the genes used in this study for SP-SNP development were studied previously by Huang *et al.* [11] who analyzed the gene expression profiling from both compatible and incompatible interactions in which "Avocet Susceptible" and resistant "AvSYr5NIL" were inoculated with race PST-78. Six of SP genes identified in the present study were also differentially expressed in their studies. These genes are

Table 2. SP-SNPs associated with avirulence genes of *Puccinia striiformis* f. sp. *tritici*.

Trait[a]	SNP ID	Supercontig[b]	Position in supercontig[b]	R of model with SNP	P	P_FDR_adj[b]	Protein ID[c]	Protein function[d]	Substitution[e]	Amino acid change
*AvYr*1	00438	1.38	88249	0.83	2.67×10^{-5}	2.53×10^{-3}	PSTG06534	HP	Intron	None
	00309	1.69	420957	0.83	4.17×10^{-4}	0.01	PSTG09436	HP	NS	Leu/Arg
	00011	1.96	157476	0.83	4.49×10^{-4}	0.01	PSTG11251	Mago nashi protein	Intron	None
	00222	1.40	615965	0.83	4.63×10^{-4}	0.01	PSTG06861	HP	NS	Ser/Phe
	00405	1.110	97515	0.83	9.03×10^{-4}	0.02	PSTG11949	HP	NS	Cys/Tyr
*AvYr*2	00007	1.187	131220	0.44	1.77×10^{-4}	0.02	PSTG14871	HP	UTR	None
*AvYr*6	00396	1.30	431009	0.52	4.81×10^{-7}	4.57×10^{-5}	PSTG05563	Glycerol kinase	UTR	None
	00441	1.85	83076	0.50	2.31×10^{-5}	1.09×10^{-3}	PSTG10535	U6 snRNA-associated Sm-like protein LSm2	Intron	None
	00015	1.21	57897	0.49	2.36×10^{-4}	7.46×10^{-3}	PSTG04138	Alpha, alphatrehalase	UTR	None
	00132	1.13	708524	0.48	1.00×10^{-3}	0.02	PSTG02790	HP	UTR	None
	00399	1.13	1112078	0.48	2.00×10^{-3}	0.03	PSTG02835	HP	NS	His/Gln
	00014	1.93	309281	0.48	2.00×10^{-3}	0.03	PSTG11109	HP	NS	Ala/Gly
*AvYr*7	00064	1.11	387678	0.79	9.02×10^{-5}	8.57×10^{-3}	PSTG02350	HP	S	None
	00392	1.31	481120	0.79	3.90×10^{-4}	0.01	PSTG05710	GDP-mannose 4,6 dehydratase	Intron	None
	00438	1.38	88249	0.79	4.40×10^{-4}	0.01	PSTG06534	HP	Intron	None
	00004	1.145	12799	0.78	1.87×10^{-3}	0.04	PSTG13532	high mobility group protein B1	UTR	None
	00024	1.145	87717	0.78	2.44×10^{-3}	0.04	PSTG13552	HP	S	None
	00414	1.42	197066	0.78	3.40×10^{-3}	0.05	PSTG07002	ADP-ribosylation factor-like protein 8A	S	None
	00258	1.37	626929	0.78	3.64×10^{-3}	0.05	PSTG06507	HP	S	None
*AvYr*8	00432	1.12	573559	0.57	1.80×10^{-4}	0.02	PSTG02592	HP	S	None
	00367	1.3	1557927	0.56	9.70×10^{-4}	0.04	PSTG00836	HP	Intron	None
*AvYr*44	00414	1.42	197066	0.79	1.70×10^{-4}	0.01	PSTG07002	ADP-ribosylation factor-like protein 8A	S	None
	00392	1.31	481120	0.79	2.50×10^{-4}	0.01	PSTG05710	GDP-mannose 4,6 dehydratase	Intron	None
	00438	1.38	88249	0.79	6.50×10^{-4}	0.01	PSTG06534	HP	Intron	None
	00017	1.5	799856	0.79	1.28×10^{-3}	0.03	PSTG01177	HP	UTR	None
	00258	1.37	626929	0.78	1.94×10^{-3}	0.03	PSTG06507	HP	S	None
	00024	1.145	87717	0.78	2.26×10^{-3}	0.03	PSTG13552	HP	S	None
	00438	1.38	88249	0.79	6.50×10^{-4}	0.01	PSTG06534	HP	Intron	None
	00004	1.145	12799	0.78	2.48×10^{-3}	0.03	PSTG13532	high mobility group protein B1	UTR	None
	00282	1.91	34788	0.78	3.81×10^{-3}	0.04	PSTG10918	HP	S	None
	00363	1.7	931271	0.78	4.21×10^{-3}	0.04	PSTG01573	HP	S	None
*AvYrExp*2	00376	1.72	436595	0.67	4.92×10^{-5}	4.68×10^{-3}	PSTG09673	Histone H4	S	None
AvYrSP	00126	1.208	34854	0.59	9.81×10^{-9}	9.32×10^{-7}	PSTG15248	HP	NS	Thr/Ala
	00043	1.13	1111649	0.58	1.68×10^{-7}	7.99×10^{-6}	PSTG02836	HP	UTR	None
	00309	1.69	420957	0.56	2.41×10^{-5}	7.60×10^{-4}	PSTG09436	HP	NS	Leu/Arg
	00021	1.230	1502	0.55	1.90×10^{-4}	4.61×10^{-3}	PSTG15691	HP	UTR	None
	00182	1.4	935429	0.55	5.50×10^{-4}	0.01	PSTG00958	HP	NS	Thr/Ile
	00058	1.15	718034	0.54	1.74×10^{-3}	0.03	PSTG03194	HP	NS	Leu/Ser
	00368	1.22	210296	0.54	2.67×10^{-3}	0.04	PSTG04299	HP	Intron	None
AvYrTye	00011	1.96	157476	0.80	1.50×10^{-4}	7.70×10^{-3}	PSTG11251	Mago nashi protein	Intron	None
	00222	1.40	615965	0.80	1.60×10^{-4}	7.70×10^{-3}	PSTG06861	HP	NS	Ser/Phe
	00405	1.110	97515	0.80	6.40×10^{-4}	0.02	PSTG11949	HP	NS	Cys/Tyr

a. Avirulence genes (*AvYr*) are corresponding to wheat *Yr* resistance genes. b. P_FDR_adj (false discovery rate-adjusted P values) were estimated from association analysis using basic scenario with PCA and kinship model from GAPIT program. c. Supercontig, the SNP locus position in the supercontig, and protein ID are from the whole genome sequence of PST-78 in BROAD Institute Puccinia database (http://www.broadinstitute.org/). d. Protein function is based on BLAST search in NCBI database. e. NS = non-synonymous; S = synonymous; and UTR = untranslated region.
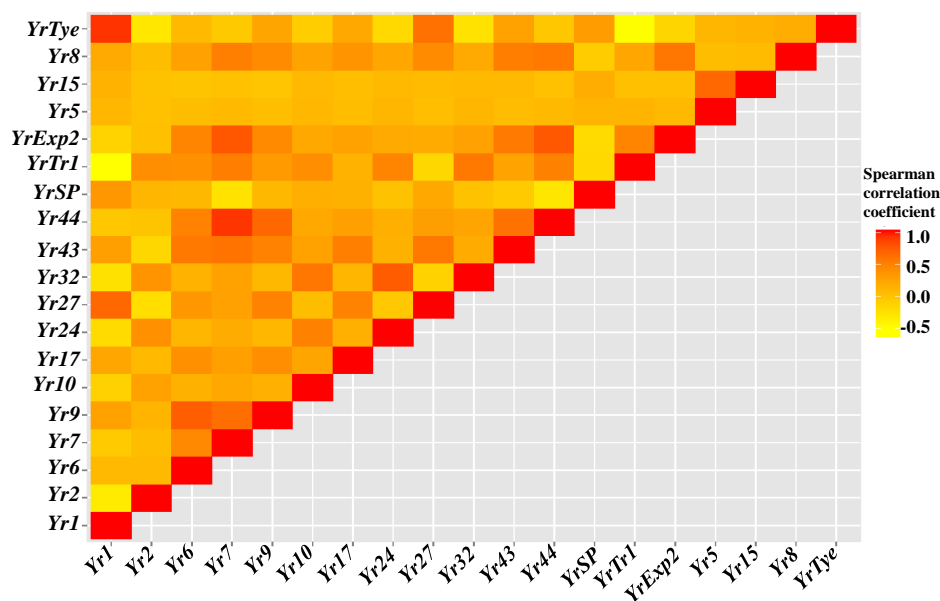
Figure 6. Correlation coefficients between 19 avirulence/virulence loci of *Puccinia striiformis* f. sp. *tritici*. The avirulence genes are represented by their corresponding resistance *Yr* genes to save space.

represented by SP-SNPs 00021, 00024, 00363, 00392, 00414, and 00441. Interestingly, these genes were from either the urediniospore cDNA library or germinated urediniospore cDNA library, but none were from the haustorial cDNA library which was thought as an *Avr* gene reservoir in rust fungi [8]. Most of the identified genes in the present study encode hypothetical proteins with no known functions (**Table 2**), while others are involved in different biological processes, but none of them are known to be involved in fungal pathogenicity. One gene associated with *AvYr*1 and *AvYrTye* has a high homology with the *mago nashi* gene in *Drosophila*. This is an essential gene required for the assembly of germ plasm in *Drosophila*, and nonsense mutations in this gene caused zygotic lethality [20]. A glycerol kinase gene was identified as associated with *AvYr*6. Glycerol kinase has an activity of catalyzing the phosphorylation of glycerol to glycerol-3-phosphate (G3P), which is an energy-producing reaction. However, G3P contributes to basal resistance in plants against *Colletotrichum higginsianum*, a hemibiotrophic pathogen [21]. Another gene associated with *AvYr*6 encodes a Sm-like protein, LSm2 which binds to the U6 small nuclear ribonucleoprotein and may be involved in pre-mRNA splicing [22]. Also associated with *AvYr*6 is an alpha-trehalase gene which is involved in hydrolysis of trehalose, an important storage compound in vegetative cells and spores of fungi [23]. A gene encoding GDP-mannose 4,6-dehydratase was associated with *AvYr*7 and *AvYr*44. This enzyme is involved in bio-synthesis of GDP-L-fucose which is important to the development and strength of stem tissues in plants [24]. A high-mobility-group box1 (HMGB1)-coding gene was identified as associated with *AvYr*7 and *AvYr*44. HMGB1 has multiple functions in DNA architecture and transcriptional regulation. It activates immune cells involved in immune processes [25]. A gene associated with *AvYrExp*2 encodes a histone H4, which is a basic component of the nucleosome. Histone H4 is subject to covalent modification (e.g. acetylation, methylation), therefore it may change the expression of genes that are associated with its parent histone octamer [26]. We need answers for several questions before we can make clear conclusions about the associations of the SP genes and avirulence loci.

The first question is why not all heritability can be explained by the associated SNP markers. Even though significant associations were detected, all associated SP-SNPs accounted for only a limited amount of heritability ranging from 0.00 to 0.63 with an average of 0.26, which was consistent to most genome-wide association studies (GWAS) [27]-[29]. The major reason underlies the hypothesis that common diseases may be caused by common genetic variants. Therefore, in the association analysis the rare variants (for example MAF <5%) are not taken into consideration. However, these rare variants probably also cause common diseases by either acting alone or accumulating effects from many alleles [28]. MAFs of greater than 1% were proposed for SNPs in the GWAS [30]. Particularly in the present study, the SP-SNPs were filtered by MAF > 5%. It is reasonable to speculate that the rare variants may be responsible for the unbalanced virulence distributions which were common

for most virulence loci (**Table 1**). Another possible reason for the unexplained heritability is that only SNP variation is evaluated in most association analyses including the present study. Even though SNP is a common form of variation in most organisms and is a causal variant for some diseases [31], other variations (e.g. copy number variants) should also be incorporated in a GWAS [28].

The second question is what genes they are and what functions they have. For the genes identified in this study, eight were annotated based on homology searches to known genes in the NCBI database. However, none of these genes, as far as we learn from the database, were directly involved in the pathogenicity in the fungi. Such non-causative variations identified in GWAS are not uncommon. A more complicated case is that a lot of markers identified from association analyses are not even in the gene regions. The non-causative associations, probably spurious associations, are mainly due to the linkage disequilibrium (LD) between many loci in the genome [30] [32]. In the LD, the non-causative variants and the causative variants are non-randomly associated, which commonly occurs in structured populations [33]. However, in this study we were unable to estimate the LD decay because of the limited number of SP-SNPs and unknown physical distances between them.

The third question is how to address false-positive associations in an association study, especially in which structured populations are used, like in this study. One source of false-positive associations is the statistical fluctuations arisen from chances from multiple-hypothesis testing [32]. To address this problem, some statistical approaches have been proposed including the Bonferroni correction and estimation of the FDR that are two common correction methods [34]. Another source of false-positive associations is from the structured populations in the association studies [32]. If the individuals from subgroups from such a population happen to differ in allele frequency of genetic polymorphisms which are unrelated to the trait of interest, the false-positive association will occur. The population structure, or population stratification, is expected in the *Pst* population because of the mainly clonal reproduction. Two major subgroups were identified, one avirulent to *Yr*1 and *YrTye*, with the other virulent to these genes. Therefore, any SP-SNPs that were different in allele frequency between these two subgroups might appear to be associated to *AvYr*1 and *AvYrTye*. To better rule out the spurious associations caused by population structures in *Pst*, sexually reproduced segregating populations should be developed, and the successful cases in plants, such as in maize, provided an example [29] [35]. To control the effects of population stratification for GWAS and take the advantage of linkage-association analyses, 25 diverse inbred maize lines were crossed to a reference line to capture a large number of recombination events, and 15 RIL families were produced with 200 RILs for each family [35]. This approach was called nested association mapping. Besides these two approaches, replicating genotype-phenotype associations in larger and independent populations was another way for establishing the credibility of GWAS [36].

Fortunately, these obstacles in the GWAS to identify virulence or pathogenicity related genes in the *Pst* population can be conquered, to some extent at least, if not wholly. Firstly, a larger number of isolates should be selected and genotyped through whole genome sequencing approaches. With a large number of isolates, the rare variants are more likely to be detected in the population. The SP-SNPs at the genome coverage should be developed and used for GWAS. The SP-SNP approach will continue focusing on functional genes, instead of non-gene regions. High resolution mapping of an association study depends on the number of markers as well as on the LD decay. Therefore, the LD decay of the *Pst* genome also should be estimated. To develop more SP-SNPs in *Pst*, a project on deep sequencing 12 isolates with different virulence spectra is underway. These deep genomic sequences will enable us to develop more markers for GWAS, including gene absence/presence and deletion/insertion polymorphisms which are probably important in the evolution of virulence of *Avr* genes in plant pathogens. For example, novel avirulence genes in *Magnaporthe oryzae*, the rice blast fungus, were identified from absence regions in the genome [37]. Secondly, the recent discoveries of alternate hosts, on which sexual reproduction of *Pst* can occur [38]-[40] enable us to establish a mapping population. Such sexual *Pst* populations should provide an efficient way to control population stratification by breaking down the LD in the *Pst* genome. Also, with hundreds of thousands of genetic markers, estimation of a linkage map with a high resolution for avirulence genes in *Pst* is possible. The identification of *Avr* genes in the *Pst* will not be the ultimate goal, but the first step to study the function, the evolution, and the genetics of the *Avr* genes, and thus to study the molecular mechanisms underlying the interactions between *Pst* and its hosts.

Acknowledgements

5348-22000-015-00D) and Washington State University (13C-3061-5665) PPNS No. 0707, 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 Hulbert and Tobin Peever for critical review of the manuscript.

References

[1] Chen, X.M. (2005) Epidemiology and Control of Stripe Rust [*Puccinia striiformis* f. sp. *tritici*] on Wheat. *Canadian Journal of Plant Pathology*, **27**, 314-337. http://dx.doi.org/10.1080/07060660509507230

[2] Wellings, C.R. (2011) Global Status of Stripe Rust: A Review of Historical and Current Threats. *Euphytica*, **179**, 129-141. http://dx.doi.org/10.1007/s10681-011-0360-y

[3] Line, R.F. (2002) Stripe Rust of Wheat and Barley in North America: A Retrospective Historical Review. *Annual Review of Phytopathology*, **40**, 75-118. http://dx.doi.org/10.1146/annurev.phyto.40.020102.111645

[4] Chen, X.M. (2014) Integration of Cultivar Resistance and Fungicide Application for Control of Wheat Stripe Rust. *Canadian Journal of Plant Pathology*, **36**, 311-326. http://dx.doi.org/10.1080/07060661.2014.924560

[5] Flor, H.H. (1971) Current Status of the Gene-for-Gene Concept. *Annual Review of Phytopathology*, **9**, 275-296. http://dx.doi.org/10.1146/annurev.py.09.090171.001423

[6] Laugé, R. and De Wit, J.G.M. (1998) Fungal Avirulence Genes: Structure and Possible Functions. *Fungal Genetics and Biology*, **24**, 285-297. http://dx.doi.org/10.1006/fgbi.1998.1076

[7] Kamoun, S. (2007) Groovy Times: Filamentous Pathogen Effectors Revealed. *Current Opinion in Plant Biology*, **10**, 358-365. http://dx.doi.org/10.1016/j.pbi.2007.04.017

[8] Catanzariti, A.M., Dodds, P.N., Lawrence, G.J., Ayliffe, M.A. and Ellis, J.G. (2006) Haustorially Expressed Secreted Proteins from Flax Rust are Highly Enriched for Avirulence Elicitors. *The Plant Cell*, **18**, 243-256. http://dx.doi.org/10.1105/tpc.105.035980

[9] Nirmala, J., Drader, T., Lawrence, P.K., Yin, C., Hulbert, S., Steber, C.M., Steffenson, B.J., Szabo, L.J., Wettstein, D. and Kleinhofs, A. (2011) Concerted Action of Two Avirulent Spore Effectors Activates Reaction to *Puccinia graminis*1 (*Rpg*1)-Mediated Cereal Stem Rust Resistance. *Proceedings of the National Academy of Sciences*, **108**, 14676-14681. http://dx.doi.org/10.1073/pnas.1111771108

[10] Chen, X.M. (2013) High Temperature Adult-Plant Resistance, Key for Sustainable Control of Stripe Rust. *American Journal of Plant Sciences*, **4**, 608-627. http://dx.doi.org/10.4236/ajps.2013.43080

[11] Huang, X.L., Chen, X.M., Coram, T., Wang, M.N. and Kang, Z.S. (2011) Gene Expression Profiling of *Puccinia striiformis* f. sp. *tritici* during Development Reveals a Highly Dynamic Transcriptome. *Journal of Genetics and Genomics*, **38**, 357-371.http://www.sciencedirect.com/science/article/pii/S1673852711001275

[12] Cantu, D., Segovia, V., MacLean, D., Bayles, R., Chen, X.M. Kamoun, S., Dubcovsky, J., Saunders, D.G. and Uauy, C. (2013) Genome Analyses of the Wheat Yellow (Stripe) Rust Pathogen *Puccinia striiformis* f. sp. *tritici* Reveal Polymorphic and Haustorial Expressed Secreted Proteins as Candidate Effectors. *BMC Genomics*, **14**, 270. http://dx.doi.org/10.1186/1471-2164-14-270

[13] Zhang, Y., Qu, Z., Zheng, W., Liu, B., Wang, X., Xue, X., Xu, L., Huang, L., Han, Q., Zhao, J. and Kang, Z. (2008) Stage-Specific Gene Expression During Urediniospore Germination in *Puccinia striiformis* f. sp. *tritici*. *BMC Genomics*, **9**, 203. http://dx.doi.org/10.1186/1471-2164-9-203

[14] Yin, C.T., Chen, X.M., Wang, X.J., Han Q.M., Kang, Z.S. and 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. http://dx.doi.org/10.1186/1471-2164-10-626

[15] Wan, A.M. and Chen, X.M. (2014) Virulence Characterization of *Puccinia striiformis* f. sp. *tritici* Using a New Set of *Yr* Single-gene Line Differentials in the United States in 2010. *Plant Disease*, **98**, 1534-1542. http://dx.doi.org/10.1094/PDIS-01-14-0071-RE

[16] Line, R.F. and Qayoum, A. (1992) Virulence, Aggressiveness, Evolution, and Distribution of Races of *Puccinia striiformis* (the Cause of Stripe Rust of Wheat) in North America, 1968-87. Technical Bulletin Number 1788. United States Department of Agriculture, Agricultural Research Service, Washington DC.

[17] Chen, X.M, Line, R.F. and Leung, H. (1993) Relationship between Virulence Variation and DNA Polymorphism in *Puccinia striiformis*. *Phytopathology*, **83**, 1489-1497. http://dx.doi.org/10.1094/Phyto-83-1489

[18] Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E. and Zhang, Z. (2012) GAPIT: Genome Association and Prediction Integrated Tool. *Bioinformatics*, **28**, 2397-2399. http://dx.doi.org/10.1093/bioinformatics/bts444

[19] Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, **57**, 289-300.

[20] Newmark, P.A. and Boswell, R.E. (1994) The *Mago Nashi* Locus Encodes an Essential Product Required for Germ Plasm Assembly in Drosophila. *Development*, **120**, 1303-1313.

[21] Venugopal, S.C., Chanda, B., Vaillancourt, L., Kachroo, A. and Kachroo, P. (2009) The Common Metabolite Glycerol-3-Phosphate Is a Novel Regulator of Plant Defense Signaling. *Plant Signaling & Behavior*, **4**, 746-749. http://dx.doi.org/10.4161/psb.4.8.9111

[22] Achsel, T., Brahms, H., Kastner, B., Bachi, A., Wilm, M. and Luhrmann, R. (1999) A Doughnut-Shaped Heteromer of Human Sm-Like Proteins Binds to the 3'-End of U6 snRNA, Thereby Facilitating U4/U6 Duplex Formation *in Vitro*. *The EMBO Journal*, **18**, 5789-5802. http://dx.doi.org/10.1093/emboj/18.20.5789

[23] Thevelein, J.M. (1984) Regulation of Trehalose Mobilization in Fungi. *Microbiological Reviews*, **48**, 42-59.

[24] Mulichak, A.M., Bonin, C.P., Reiter, W.D. and Garavito, R.M. (2002) Structure of the MUR1 GDP-Mannose 4,6-Dehydratase from *Arabidopsis thaliana*: Implications for Ligand Binding and Specificity. *Biochemistry*, **41**, 15578-15589. http://dx.doi.org/10.1021/bi0266683

[25] Klune, J.R., Dhupar, R., Cardinal, J., Billiar, T.R. and Tsung, A. (2008) HMGB1: Endogenous Danger Signaling. *Molecular Medicine*, **14**, 476-484.

[26] Bhasin, M., Reinherz, E.L. and Reche, P.A. (2006) Recognition and Classification of Histones Using Support Vector Machine. *Journal of Computational Biology*, **13**, 102-112. http://dx.doi.org/10.1089/cmb.2006.13.102

[27] Maher, B. (2008) Personal Genomes: The Case of the Missing Heritability. *Nature*, **456**, 18-21. http://dx.doi.org/10.1038/456018a

[28] Witte, J.S. (2010) Genome-Wide Association Studies and Beyond. *Annual Review of Public Health*, **31**, 9-20. http://dx.doi.org/10.1146/annurev.publhealth.012809.103723

[29] Brachi, B., Morris, G.P. and Borevitz, J.O. (2011) Genome-Wide Association Studies in Plants: The Missing Heritability Is in the Field. *Genome Biology*, **12**, 232. http://dx.doi.org/10.1186/gb-2011-12-10-232

[30] Wang, W.Y.S., Barratt, B.J., Clayton, D.G. and Todd, J.A. (2005) Genome-Wide Association Studies: Theoretical and Practical Concerns. *Nature Review of Genetics*, **6**, 109-118. http://dx.doi.org/10.1038/nrg1522

[31] Joosten, M.H.A.J., Cozijnsen, T.J. and De Wit, P.J.G.M. (1994) Host Resistance to a Fungal Tomato Pathogen Lost by a Single Base-Pair Change in an Avirulence Gene. *Nature*, **367**, 384-386. http://dx.doi.org/10.1038/367384a0

[32] Hirschhorn, J.N. and Daly, M.J. (2005) Genome-Wide Association Studies for Common Diseases and Complex Traits. *Nature Review of Genetics*, **6**, 95-108. http://dx.doi.org/10.1038/nrg1521

[33] Flint-Garcia, S.A., Thornsberry, J.M. and Buckler IV, E.S. (2003) Structure of Linkage Disequilibrium in Plants. *Annual Review of Plant Biology*, **54**, 357-374. http://dx.doi.org/10.1146/annurev.arplant.54.031902.134907

[34] Pearson, T.A. and Manolio, T.A. (2008) How to Interpret a Genome-Wide Association Study. *The Journal of the American Medical Association*, **299**, 1335-1344. http://dx.doi.org/10.1001/jama.299.11.1335

[35] McMullen, M.D., Kresovich, S., Villeda, H.S., Bradbury, P., Li, H., Sun, Q., Flint-Garcia, S., Thornsberry, J., Acharya, C., Bottoms, C., Brown, P., *et al*. (2009) Genetic Properties of the Maize Nested Association Mapping Population. *Science*, **325**, 737-740. http://dx.doi.org/10.1126/science.1174320

[36] Chanock, S.J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D.J., Thomas, G., Hirschhorn, J.N., Abecasis, G., Altshuler, D., Bailey-Wilson, J.E., *et al*., NCI-NHGRI Working Group on Replication in Association Studies (2007) Replicating Genotype-Phenotype Associations. *Nature*, **447**, 655-660. http://dx.doi.org/10.1038/447655a

[37] Yoshida, K., Saitoh, H., Fujisawa, S., Kanzaki, H., Matsumura, H., Yoshida, K., Tosa, Y., Chuma, I., Takano, Y., Win, J., Kamoun, S. and Terauchi, R. (2009) Association Genetics Reveals Three Novel Avirulence Genes from the Rice Blast Fungal Pathogen *Magnaporthe oryzae*. *The Plant Cell*, **21**, 1573-1591. http://dx.doi.org/10.1105/tpc.109.066324

[38] Jin, Y., Szabo, L. and Carson, M. (2010) Century-Old Mystery of *Puccinia striiformis* Life History Solved with the Identification of *Berberis* spp. as an Alternate Host. *Phytopathology*, **100**, 432-435. http://dx.doi.org/10.1094/PHYTO-100-5-0432

[39] Wang, M.N. and Chen, X.M. (2013) First Report of Oregon Grape (*Mahonia aquifolium*) as an Alternate Host for the Wheat Stripe Rust Pathogen (*Puccinia striiformis* f. sp. *tritici*) under Artificial Inoculation. *Plant Disease*, **97**, 839. http://dx.doi.org/10.1094/PDIS-09-12-0864-PDN

[40] Wang, M.N. and Chen, X.M. (2015) Barberry Does Not Function as an Alternate Host for *Puccinia striiformis* f. sp. *tritici* in the US Pacific Northwest Due to Teliospore Degradation and Barberry Phenology. *Plant Disease*, **99**, 1500-1506. http://dx.doi.org/10.1094/PDIS-12-14-1280-RE