Genome wide association study reveals possible SNPs associated with ToxC in Pyrenophora tritici-repentis (tan spot of wheat)

Ryan Gourlie[1], Megan McDonald[2], and Reem Aboukhaddour[1]

[1]*Agriculture and Agri-Food Canada, LRDC, Lethbridge, AB, Canada*
[2]*University of Birmingham, School of Biosciences, Edgbaston, Birmingham, United Kingdom*

Agriculture and Agri-Food Canada

Abstract

Pyrenophora tritici-repentis (Died.) Drechs. (Ptr) is the causal agent of tan spot, one of the most destructive foliar pathogens of wheat worldwide. Ptr is known to secrete three necrotrophic effectors: ToxA, ToxB, and ToxC. The exact nature of ToxC remains unknown, though recently the ToxC1 gene was identified and shown to be required but insufficient to produce ToxC (1). We have utilized 37 fully sequenced genomes of Ptr, of which 22 are ToxC-producing to correlate SNP data with the ToxC phenotype via a genome wide association study (GWAS). SNP variant calls were generated with Genome Analysis Toolkit using isolate M4 (v2.2) as a reference genome. Variants were filtered based on quality scores, minor allele frequency, and percentage of genomes with missed calls. Tassel (v5.2.58) was used to create distance and kinship matrices (identity by state; centered IBS) and perform principle component analysis to account for population structure. GWAS was performed with Tassel using a Mixed-Linear-Model algorithm. A peak *-log10(p-value)* of 3.4 was observed near the 5' end of chromosome 6 in M4. The SNP associated with the peak is located within the coding sequence of a gene with no known function (DUF1479), based on the Pfam database. Further work with GAPIT, another GWAS program is currently being used to test the robustness of the SNP peak by utilizing different methods for population structure correction (e.g. VanRaden, EMMA) and alternative GWAS modelling (e.g. BLINK, MLMM). Comparative analysis of the region surrounding the peak has not yet provided concrete differences between ToxC-producing and ToxC-non-producing isolates.

Results and Discussion

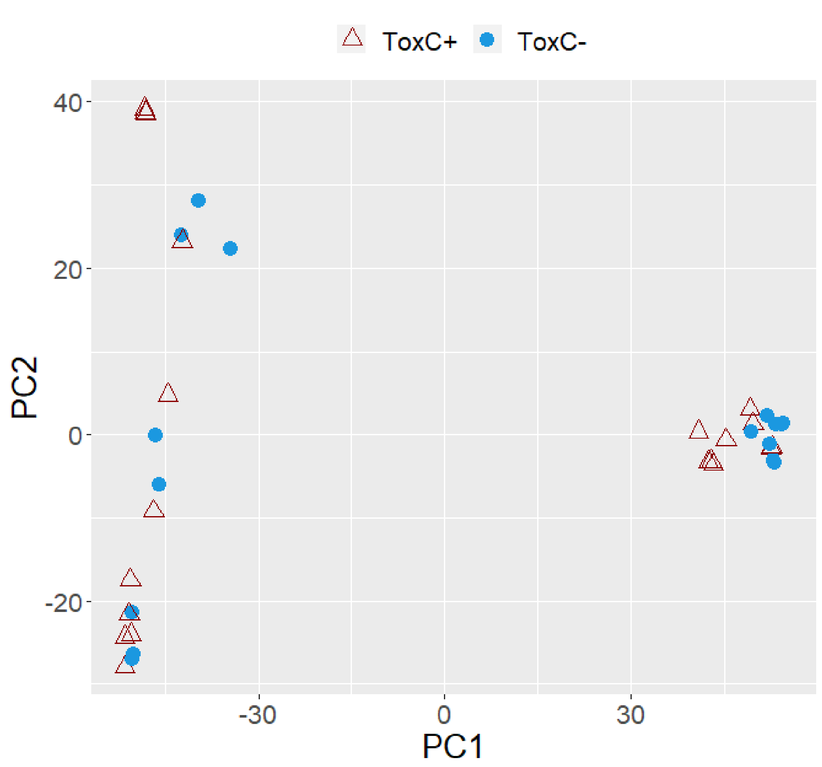

Figure 1 Principal component analysis of SNP data revealed population structure within the utilized Ptr genomes.

From the 37 Ptr genomes, a total of 74,025 SNP variants were catalogued by Genome Analysis Toolkit after read quality filtering. Of these SNPs, 23,117 remained after further filtering within Tassel. Principal component analysis of SNP data revealed significant population structure within the Ptr genomes along both axes (**Figure 1**). Population structure can produce false correlations between SNPs and phenotypes. After population structure was accounted for, GWAS using a MLM model revealed a possible correlation between a group of SNPs near the 5' end of chromosome 6 (M4) and the ToxC phenotype (**Figure 2**). Generally, a *–log10(p-value)* of 5 is considered significant. The highest value reported here is 3.4, however due to the large number of SNPs filtered and the relatively low number of genomes used, this peak may still be significant. Luckily, a further 20+ genomes recently became available and will be included to increase the statistical power of the GWAS. Furthermore, alternative methods for controlling population structure and/or GWAS algorithms may increase confidence. Several SNPs were located in coding genes with no known functions however as of yet we've been unable to convincingly link differences larger than SNPs to the ToxC phenotype.

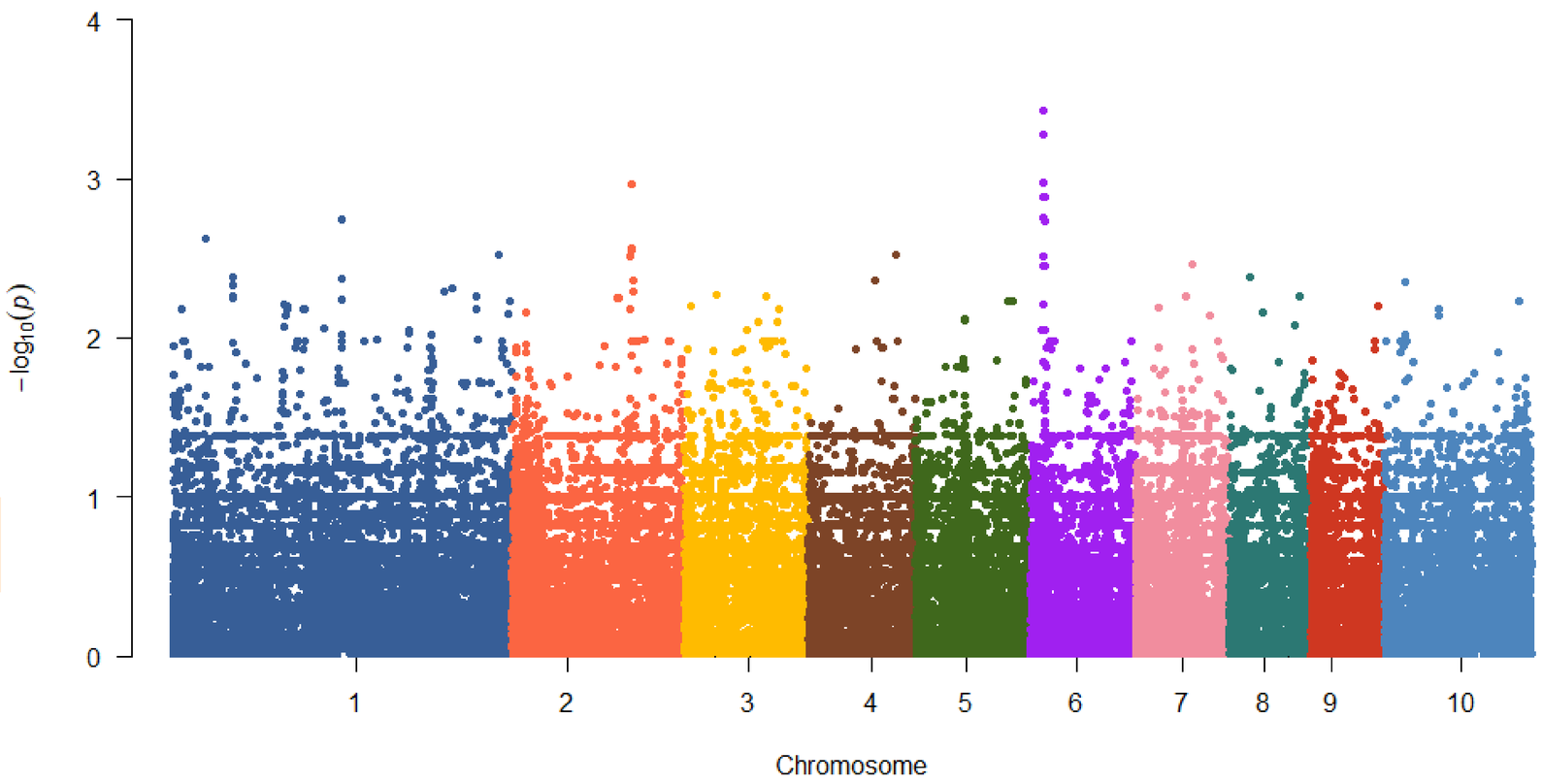

Figure 2 Manhattan plot of SNPs from 37 Ptr genomes, divided by chromosomes of reference isolate M4v2.2. A peak *–log10(p-value)* of 3.4 is visible in chromosome 6 (purple dots).

Materials and Methods

Isolate collection, DNA extraction, sequencing, and de novo genome assembly details described in Gourlie et al., 2022.

Variant calling with Genome Analysis Toolkit

Variant calls were generated with Genome Analysis Toolkit (v4.2.0.0)(GATK) (2), Pt-1C-BFP (3) was used as a reference initially and analysis we done again using M4v2.2 (4) as a reference. The reference genome was indexed using bowtie2 (5) (-build) and SAMtools (6) (-faidx), and a sequence dictionary made with GATK (-CreateSequenceDictionary). Raw Illumina reads of 40 isolates, generated in previous work (7), were aligned to the reference with bowtie2. Alignments were converted from SAM to BAM format and then sorted via SAMtools (-view -S -b -sort). Duplicates were removed from the BAM alignments with picard tools (-MarkDuplicates --REMOVE_DUPLICATES) and read groups added (-AddOrReplaceReadGroups). Read groups were indexed with SAMtools (-index) and variant calls for each isolate were done with GATK HaplotypeCaller. GVCFs were then merged (-GenomicsDBImport --interval_padding 100) and genotyped (-GenotypeGVCFs). Variant calls were then filtered based on various quality scores (QUAL=250; QD=20; FS=0.1;SOR=3;MQ=30;ReadPosRankSum>2; ReadPosRankSum<-2; MQRankSum>2.5; MQRankSum<-2.5) and filtered calls output (-VariantsToTable).

Association of SNPs with ToxC via Tassel

Genome-wide association was done with Tassel (v5.2.58) (7). A phenotype file was created where the ToxC phenotype was treated as binary data with 1 given to isolates which cause ToxC cholorsis symptoms and 0 given to isolates which did not show the ToxC phenotype. The filtered VCF file from GATK was imported and SNPs were filtered again in Tassel based on minor allele frequency, excluding the rarest SNPs (≤5% of isolates) and SNPs with a significant number of missed variant calls (≥10% of isolates). A distance matrix (5 axes) and a kinship matrix (centered IBS; max 5 alleles) was created using the filtered variant calls. Multi-dimentional scaling was performed on the distance matrix and the principal components output from the scaling were intersected with the phenotype and SNP data. A mixed linear model was used to assign p-values to SNPs using the intersected dataset and the kinship matrix as input. The MLM statistics were exported from Tassel and imported into R to produce a Manhattan plot.

References

1. Shi et al. 2022. *Molecular Plant-Microbe Interactions*. 2022(ja)
2. McKenna et al. 2010. *Genome Research*. 20(9):1297-303
3. Manning et al. 2013. *G3*. 3(1):41-63
4. Moolhuijzen et al. 2018. *BMC Genomics*. 19(1):279
5. Langmead et al. 2012. *Nature Methods*. 9(4):357
6. Gourlie et al. 2022. *BMC Biology*. 20(1):1-21
7. Bradbury et al. 2007. *Bioinformatics*. 23(19):2633-5