Genome-wide DNA Variants Identify Genetic Diversity and Population Structure of Leptosphaeria maculans in Western Canada

Qilin Chen  
Saskatoon Research and Development Centre

Gary Peng  
Saskatoon Research and Development Centre

Randy Kutcher  
University of Saskatchewan

Fengqun Yu (✉ fengqun.yu@canada.ca)  
Agriculture and Agri-Food Canada  https://orcid.org/0000-0002-7957-7632

Research article

Keywords: Leptosphaeria maculans, genome-wide DNA variants, Western Canada, genetic diversity

DOI: https://doi.org/10.21203/rs.3.rs-24766/v2

License: © This work is licensed under a Creative Commons Attribution 4.0 International License.  Read Full License
Abstract

Background: *Leptosphaeria maculans* is a serious concern for canola production in Canada. For effective management, knowledge of the pathogen's genetic variability and population structure is a prerequisite. Despite some information on race dynamics of the western Canadian *L. maculans* population in recent years, genetic diversity based on a large number of genome-wide DNA variants has not been investigated.

Results: From 1,590 *L. maculans* isolates collected from 23 field sites in three provinces: Manitoba, Saskatchewan and Alberta, Canada, in the years 2007-2008 and 2012-2014, 150 representative isolates were selected and whole-genome sequenced, and 31,870 polymorphic DNA variants (SNPs and InDels) were used to study *L. maculans* genetic diversity and population structure. Cluster analysis showed that the genetic diversity levels and isolate groupings varied with the number and genomic regions of the variants involved; isolates collected in 2012-2014 were more genetically diverse than those collected in 2007-2008 when genome-wide variants were considered. The genome wide association study (GWAS) detected variants in *egn4_Lema_T86290* (AvrLm4-7), *egn4_Lema_T86300* and *egn4_Lema_T86310* associated with the year of collection, but no variants was found to be associated with the province or specific location from which the isolates were collected. Population structure analysis indicated the presence of three distinct sub-populations in western Canada. While isolates from Saskatchewan were mainly of one sub-population (sub-pop1), the Alberta isolates comprised two sub-populations (sub-pop1 and sub-pop2), and all the 3 subpopulations were found in Manitoba.

Conclusion: The genetic diversity of the western Canadian *L. maculans* population varied among provinces. It was highly admixed in Manitoba, followed by that in Alberta. The Saskatchewan population had the lowest genetic diversity. Significant genome variation between 2007-2008 and 2012-2014 occurred in the genes *egn4_Lema_T86290* (AvrLm4-7), *egn4_Lema_T86300* and *egn4_Lema_T86310*, with AvrLm4-7 becoming much more common in the *L. maculans* population in the later period.

Background

Plants and pathogens co-evolve after interactions with each other. This co-evolution is fueled by DNA variation underlying the recognition of pathogen proteins by the host and the defeat of host defenses by the pathogen; together with new mutations, genetic diversity in populations of both the host and pathogen represent a pool of possible variants to maintain adaptation via natural selection [1].

The fungal pathogen *Leptosphaeria maculans* Ces. & de Not., a filamentous ascomycete, is the causal agent of blackleg (phoma stem canker) of canola/oilseed rape (*Brassica napus* L.). It occurs in many regions where *B. napus* is grown, and is a serious concern for canola/rapeseed production, especially in Canada, Australia and Europe. Sustainable canola production requires effective management of blackleg using an integrated approach, including crop rotation, fungicide seed treatment [2], and resistant cultivar development [3]. The use of resistant cultivars and a diverse crop rotation are considered the most effective and economical ways to control the disease. However, the pathogen has both sexual and asexual reproduction systems in its life cycle, which enables it to combine the most fitting genotypes and quickly increase their frequencies in the population through clonal reproduction [4, 5, 6]. As a result, the high genetic variability in the pathogen population may lead to races that reduce the lifespan of the resistant varieties, such as the breakdown of *Rlm1* in France [7], *LepR3* in Australia [8] and *Rlm3* in Canada [9].

Knowledge of the population genetic structure of the pathogen may offer insight into the best breeding strategy for durable resistance [6]. Genetic diversity of *L. maculans* has been studied in different geographical regions with various types of DNA markers, including Random Amplified Polymorphic DNA (RAPD), Amplified Fragment Length Polymorphism (AFLP), minisatellite and microsatellite and SNPs. Cluster analysis of RAPD patterns grouped 93 virulent isolates of *L. maculans* collected from canola fields in southern Ontario Canada into two populations based on the estimates of genetic distance [10]. Analysis of AFLP markers performed on 100 isolates collected in Australia, Europe and North American suggested that the Australian and European populations were separated in adjacent clusters, while the North American population partially overlapped both of the others [11]. Use of AFLP markers found a low degree of genetic differentiation among *L. maculans* isolates collected from different regions in Australia [12, 13]. Using six microsatellite and two minisatellite markers, the genetic diversity of 13 subpopulations of *L. maculans* was identified among 513 isolates collected from multiple sites over 2 years in Australia [14]. Use of several minisatellite markers
found a high degree of gene exchange among field *L. maculans* populations of in France [4, 15] or a rapid countrywide diffusion of novel virulence alleles wherever novel resistance sources were used [15]. Fourteen minisatellite markers were used to characterize a worldwide collection of 465 isolates of *L. maculans*. Clustering analyses partitioned genotypes into distinct populations corresponding to major geographic regions, along with two differentiated populations in western Canada [16].

In recent years, DNA variants, including SNPs (Single Nucleotide Polymorphisms) and InDels (Insertions/Deletions), have become popular molecular markers for genetic diversity studies. SNPs can be categorized into transitions (C to T, G to A) and transversions (C to G, A to T, T to G or C to A), although transitions are more common than transversions. The Illumina GoldenGate genotyping assay can be used to analyse 384-3072 SNP loci across multiple individuals simultaneously [17]. A GoldenGate assay was used to assess the genetic diversity of 59 Australian *L. maculans* isolates collected from different years, regions and cultivars, and there was no correlation between SNP haplotypes and cultivars or the year of collection, indicating a high rate of sexual reproduction and evolutionary diversity in the pathogen [18].

Next generation sequencing (NGS) has ignited a revolution in the life sciences; it enables the inexpensive and rapid discovery of DNA variants in many organisms. Although NGS has found broad application in fundamental genome and genetic research, there are limited reports on the identification of genome-wide DNA variants in plant pathogens, including *L. maculans*. Zander et al. [19] reported 21,814 genome-wide SNPs in *L. maculans* based on a study of two Australian isolates. In a preliminary phylogenetic analysis, isolates from Western Australia clustered together and those collected from *Brassica juncea* stubble were identical. These SNPs provide a novel characterization of the genetic diversity for this pathogen population [19].

Genome sequencing revealed that the *L. maculans* genome has an isochore-like structure [20], which can be divided into AT and GC-rich blocks, probably caused by the amplification of transposable elements and repeat-induced point (RIP) mutations. The RIP mechanism causes nucleotide substitutions from C to T and G to A and is a premeiotic repeat-inactivation mechanism specific to fungi that creates genetic diversity in the fungal genome [20].

Methods to detect genetic diversity and population structure fall into two categories: model-based parametric approaches and non-parametric approaches. Model-based methods, exemplified by STRUCTURE [21] and *snmf*, a function of the R package LEA [22], assumes hypothetical K populations characterized by allele frequencies at a locus, probabilistically assigning individuals to each of these populations and allowing for admixture estimation of the genome originating from a given population. While other programs differ from STRUCTURE in using variation inference (fastStructure) [23] or a maximum likelihood framework (ADMITTURE) [24], they proposed an explicit causal model assuming linkage equilibrium between loci and Hardy Weinberg equilibrium between alleles. Nonparametric methods such as PCA first reduces the dimensionality of a variant frequency matrix and then identifies groups of individuals.

In this study, we applied NGS technology to sequence 162 of 1590 characterized isolates collected from different years and regions. Taking advantage of the abundance of genome-wide DNA variants, we investigated the genetic diversity and population structure of *L. maculans* isolates collected in western Canada in the years of 2009~2009 and 2012~2014 using both parametric (*snmf*) and non-parametric methods including PCA and clustering. The results could provide valuable information and resources for blackleg management.

### Results

#### Selection of isolates for re-sequencing

Cluster analysis was conducted using phenotypic data from bioassays on a set of differential hosts (Table S1) inoculated individually with 1590 *L. maculans* isolates from western Canada. Six clusters were identified, containing 125, 325, 179, 143, 195 and 623 isolates, respectively (S Fig. 1). A total of 162 isolates (Table S2), with 27, 31, 19, 23, 14 and 48 isolates from each of the six cluster (S Fig. 1) were selected for DNA sequencing. Temporally 98 isolates were from years 2007-2008 and 54 from 2012-2014, or spatially 31 isolates from four locations in Alberta (AB), 63 from 10 locations in Manitoba (MB), and 68 from nine locations in Saskatchewan (SK).
The NGS short reads were assembled on the reference genome of *L. maculans* ‘brassicae’ isolate v23.1.3. Breadth and depth of coverage were examined to evaluate the quality of sequencing. Both coverages varied with isolates and supercontigs (SCs); four isolates with low depth and breadth of coverage were removed from the analysis. The remaining 158 isolates showed breadth of coverage of >80% (S Fig. 2 A), and five SCs (SC 27, 28, 33, 34, 36) had lower breadth of coverage ranging from 49.6 to 73.8% (S Fig. 2 B). For depth of coverage, 88% of the isolates displayed values >20. The coverage for SCs ranged from 12 to 26 except for SC30, which was as high as 621. In summary, the average breadth of coverage was 91% and the average depth of coverage was 36.

### Variant discovery

Compared with the reference genome of *L. maculans*, about 920,000 DNA variants (SNPs and InDels) were identified; among them were 104,744 polymorphic biallelic variants with MAF ≥ 1 and miss call rates ≤ 50%. Considering cluster analysis and PCA were to be performed in this study, miss call rates >5% would be considered statistically consequential [25], therefore, additional eight isolates were removed. After filtering against the criteria of phred quality (Q call ≥ 25), minimum allele frequency (MAF ≥ 5), and miss call rate (≤ 5%), 31,870 polymorphic variants were selected for subsequent analysis. Variant numbers in each isolate ranged from 8,591 to 14,997 with an average of 9,575, as compared with the reference genome.

To understand DNA variation among the isolates, the sequences of 150 retained isolates were paired to each other to find polymorphic variant numbers between isolates. The numbers varied dramatically with 11,175 isolate pairs in the range 508 to 17,708. About 4,000 isolate pairs showed 4,000~5,000 polymorphic variants, and 610 pairs displayed 15,000~18,000 polymorphic variants (S Fig. 3). This would suggest that approximately 56% of the 31,870 total polymorphic variants identified may be found in a single isolate pair.

### Uneven variant distribution among genomic regions

As expected, variant numbers were highly correlated with SC size (R²=0.95, data not shown). There were 2,884 variants detected in the longest supercontig SC0, but only 14 variants in the shortest SC40 (Table 1). The *L. maculans* genome has an isochore-like structure, characterized by alternating AT- and GC-blocks [20]. Variant density in the two blocks was compared along with coding and non-coding regions. Variant density in the AT blocks and the non-coding regions was 0.88 and 0.78 variants/Kb, higher than the average for the whole genome (0.71 variants/Kb) (Table 1). In contrast, GC blocks and coding regions had lower variant density (0.62 and 0.61 variants/Kb). This suggested uneven variant distribution in the contrasting regions, which became more evident when individual supercontigs were examined. For instance,
Table 1. Variant distribution and density in supercontigs and different genomic regions of *L. maculans*.

| SCs\(^a\) | Size (base) | Gene numbers | Variant numbers | AT-blocks coverage\(^b\) (%) | Coding regions coverage\(^c\) (%) | Variant density\(^d\) (variants per kb) |
|-----------|-------------|--------------|-----------------|-----------------------------|----------------------------------|----------------------------------------|
|           |             |              |                 | Whole SC AT blocks GC Blocks \(b\) Non-coding regions Coding regions |
| SC0       | 4258568     | 1263         | 2884            | 31.4                        | 45.67                            | 0.68 0.95 0.55 0.79 0.54             |
| SC1       | 3378610     | 1080         | 2712            | 26.3                        | 45.96                            | 0.80 1.03 0.72 0.84 0.76             |
| SC2       | 2939989     | 916          | 1928            | 25.0                        | 46.09                            | 0.66 1.10 0.51 0.79 0.50             |
| SC3       | 2348246     | 565          | 2039            | 44.6                        | 34.61                            | 0.87 0.91 0.83 0.88 0.84             |
| SC4       | 1918205     | 662          | 1171            | 28.7                        | 46.87                            | 0.61 0.97 0.47 0.77 0.43             |
| SC5       | 1869450     | 633          | 1516            | 25.9                        | 50.65                            | 0.68 1.01 0.57 1.01 0.62             |
| SC6       | 1888674     | 510          | 1274            | 41.9                        | 33.79                            | 0.80 0.68 0.89 0.61 0.81             |
| SC7       | 1769547     | 567          | 1267            | 32.4                        | 44.69                            | 0.86 1.18 0.70 0.80 0.61             |
| SC8       | 1809296     | 618          | 1019            | 25.8                        | 45.84                            | 0.70 1.04 0.58 0.67 0.44             |
| SC9       | 1772623     | 474          | 1514            | 34.8                        | 39.70                            | 0.57 0.59 0.57 1.02 0.60             |
| SC10      | 1758670     | 434          | 1052            | 33.9                        | 35.57                            | 0.60 0.85 0.47 0.67 0.47             |
| SC11      | 1590160     | 285          | 954             | 47.5                        | 25.52                            | 0.77 0.67 0.86 0.72 0.25             |
| SC12      | 1631710     | 439          | 1161            | 35.0                        | 38.01                            | 0.71 1.03 0.54 0.81 0.55             |
| SC13      | 1634580     | 492          | 1225            | 30.4                        | 42.29                            | 0.58 1.21 0.31 0.76 0.73             |
| SC14      | 1533332     | 512          | 1018            | 26.0                        | 51.78                            | 0.69 1.10 0.54 0.86 0.48             |
| SC15      | 1560629     | 442          | 1053            | 36.5                        | 39.64                            | 0.65 0.76 0.59 0.72 0.60             |
| SC16      | 1397653     | 353          | 1428            | 41.2                        | 39.83                            | 0.52 0.61 0.46 1.11 0.89             |
| SC17      | 1445693     | 415          | 723             | 29.7                        | 41.99                            | 0.99 1.39 0.82 0.54 0.44             |
| SC18      | 1351976     | 366          | 1148            | 37.8                        | 40.36                            | 0.85 0.85 0.85 0.88 0.80             |
| SC19      | 1186800     | 288          | 716             | 41.2                        | 36.34                            | 0.60 0.75 0.50 0.66 0.51             |
| SC20      | 1087932     | 292          | 1088            | 39.3                        | 40.47                            | 1.00 0.87 1.08 0.92 1.12             |
| SC21      | 1020521     | 296          | 700             | 33.2                        | 40.20                            | 0.69 0.85 0.60 0.73 0.62             |
| SC22      | 731443      | 35           | 309             | 84.4                        | 5.53                             | 0.42 0.40 0.56 0.44 0.15             |
| SC23      | 521426      | 157          | 473             | 32.1                        | 42.97                            | 0.91 1.10 0.81 0.98 0.81             |
| SC24      | 475869      | 116          | 334             | 39.5                        | 40.12                            | 0.70 0.82 0.63 0.84 0.50             |
| SC25      | 318058      | 99           | 249             | 28.7                        | 48.58                            | 0.78 1.16 0.63 0.92 0.64             |
| SC26      | 261540      | 74           | 172             | 34.1                        | 43.28                            | 0.66 0.69 0.64 0.71 0.59             |
| SC27      | 250629      | 3            | 186             | 77.1                        | 0.38                             | 0.47 0.52 0.28 0.62 31.50            |
| SC28      | 236098      | 1            | 8               | 96.1                        | 0.45                             | 0.45 0.42 1.08 0.02 2.82             |
| SC29      | 200940      | 7            | 72              | 41.0                        | 3.22                             | 0.93 1.58 0.47 0.28 2.78             |
| SC30      | 154863      | 35           | 117             | 63.8                        | 12.49                            | 0.05 0.00 0.14 0.86 0.00             |
| SC31      | 143268      | 37           | 106             | 69.7                        | 40.50                            | 0.50 0.24 1.10 1.24 0.00             |
for SC2, SC4, SC13 and SC35, variant density in AT blocks were 1.10, 0.97, 1.21 and 0.35 variants/Kb, respectively, nearly 2-4 times of that in GC blocks; similarly, in SC11 and SC22, non-coding regions had variant intensity (0.75 and 0.44 variants/kb) almost twice as high as in coding regions (Table 1).

To understand if genes in AT-blocks and GC-blocks mutate differently, we calculated variant density in genes in the two bipartite regions. There were 12,469 genes annotated in the *L. maculans* genome. We sorted genes in AT-rich regions into three classes: class A genes resided entirely in an AT-rich block, class B genes were those in which only part of their sequences overlapped with an AT-block, and class C genes were located solely within an GC-equilibrated region with one of its ends falling within 2,000 bases from an AT-block (Table S3). In the end, 1,158 genes were assigned to AT-blocks, leaving 11,311 genes assigned to GC-blocks. There were 10,042 variants detected in 3,971 genes in the GC-blocks, and 757 variants in 341 genes in the AT-blocks, with an average of 2.5 variants/gene in the GC-blocks and 2.2 variants/gene in the AT-blocks. Class A had 28 genes, with 25 located in SC30. Despite the higher variant density in AT blocks, none of the DNA variants was found in genes with full sequences residing in an AT-block. Class B was comprised of 248 genes; 80 of them had 274 variants distributed in 28 SCs. Class C were made up of 881 genes, including avirulence genes previously characterized, such as *AvrLm4-7* [26], *AvrLm3* [27], *AvrLm2* [28] and *AvrLm6* [29], suggesting class C might serve as a pool of effectors. Class C displayed 746 variants in 259 genes.

*L. maculans* genes could be classified into small, secreted protein-encoding genes (SSP-genes) and non-SSP genes. We inspected coding regions to determine if DNA mutation was biased towards SSP genes or non-SSP-genes. Rouxel et al. [20] identified 652 SSPs, from which we selected 587 annotated genes, carrying gene names beginning with *egn4_Lema*, for variant density comparison. In this study, 327 variants were found in 153 SSP genes, and 10,668 variants were detected in 4364 non-SSP genes. On average, SSP genes had 2.1 variants/gene and non-SSP genes had 2.4 variants/gene. Variant numbers fluctuated considerably among genes (Fig. 1). The top 2 SSP genes, *egn4_Lema_T086290.1*(*AvrLm4-7*) and *egn4_Lema_T015100.1* had
only 37 and 23 variants, respectively, while the top 2 non-SSP genes, egn4_Lema_T015070.1 and egn4_Lema_T015110.1, had 101 and 100 variants, respectively. Next to AvrLm4-7, AvrLm3 was detected with nine SNPs (six nonsynonymous SNPs and three synonymous), and the other avirulence genes carried three SNPs at most, indicating that some non-SSP genes might have mutated more frequently than SSP genes.

**Dominance of transition variants**

Variants are classified into transition, transversion and InDels (insertions and deletions). The variant composition of these four types was examined. Transition variants were dominant in the *L. maculans* genome, accounting for 68.0% of all variants, followed by transversion variants at 24.5%. Variant composition in different genomic regions was also investigated. When compared at the genome level, AT blocks contained more transition variants (75.0%) and fewer transversion (19.9%) and InDel variants (5.1%). In contrast, GC blocks had a reduced proportion of transition (63.2%) but higher percentage of transversion (28%) variants and Indels (8.8%) (S Fig. 4). In other words, when compared with GC blocks, AT blocks had 19% higher transition variants, but 29% fewer transversion variants and 42% fewer InDels.

Comparatively, the difference in variant composition between coding and non-coding regions, and SSP and non-SSP genes showed was less significant that between AT- and GC- blocks (S Fig. 4). Of course, the ratio varied considerably depending on the genes considered. For an example, gene egn4_Lema_T086290.1 (AvrLm4-7) had 36 transitions but only one transversion and no InDels; on the other hand, gene egn4_Lema_T007850.1 had five transversions but only one transition.

**Phylogeny of *L. maculans* isolates**

This study involved 26 isolates from AB, 59 from SK and 65 from MB. Considering the abundance of variants identified, genetic diversity among these isolates was investigated first.

The variants were discovered in different genomic regions or genes of interest, such as AT-blocks (13,851 variants), GC-blocks (18,019), coding regions (10,994), non-coding regions (20,876), SSP-genes (952), non-SSP genes (10,042), genes in AT-blocks (326), genes in GC blocks (10,668) and avirulence genes (48). Given the difference in genetic diversity as measured by variants in different regions, hierarchical cluster analysis was performed using the variants from different sources and their dendrograms were compared. Visual side-by-side comparison of overcrowded dendrograms was inconvenient due to the number of isolates involved, dendrograms were then compared to each other by their cophenetic matrices instead since a phylogenetic dendrogram is the graphical representation of a cophenetic matrix [30]; a coefficient value close to 1 would represent a significant similarity between the two dendrograms in the comparison (Fig. 2). The similarity of dendrograms between genome-wide variants and variants from different genomic regions could be classified into three tiers, the first tier included AT-blocks, GC-blocks, coding-regions, non-coding regions and non SSP genes, with correlation coefficients of >0.94. The second tier encompassed SSP genes, and genes in AT and GC blocks, from which the correlation coefficient ranged from 0.60 to 0.87. The third tier was *Avr* genes, with a correlation coefficient of only 0.14. It was found, however, that the similarity between any of these dendrograms with whole-genome dendrogram was only partly correlated with the number of variants involved. For instance, the genes in AT-blocks had almost 14 times as many variants as SSP genes, but its correlation coefficient (0.60) was less than the latter (0.73) when the two dendrograms were compared with that of genome-wide variants (Fig. 2).

We then chose to describe three dendrograms based on our interest in pathogen-host interaction. The first was created by 48 SNPs in avirulence genes, the second constructed by the 952 variants in SSP genes, and the third was from genome-wide variants, to illustrate their difference in measuring genetic diversity. The three dendrograms are hereafter referred to as the Avr dendrogram, the SSP dendrogram and the whole-genome dendrogram.

The Avr dendrogram consisted of two main groups (I, II), each with two subgroups (Fig. 3 A). Subgroups A and B in Group I were made up of 72 and 22 isolates, respectively. For group II, There were 18 isolates in subgroup C and 38 in subgroup D. Each subgroup contained isolates collected from the three provinces, but it was noticeable that subgroups C and D were composed of isolates only from years 2007-2008, while isolates from both 2007-2008 and 2012-2014 were found in subgroups A and B. While the Avr dendrogram showed avirulence gene diversity, it also revealed that some isolates shared an identical set of variants in multiple avirulence genes because the Euclidean distance among them was 0. For example, 33 isolates were clustered as leaf A in
subgroup A (Fig. 3 A); they were literally the same in terms of avirulence gene sequences. So were the 25 isolates as leaf B in subgroup D. It was noticed, however, that isolates collected from the same location were observed in different clads, for instance, 19 isolates from Melfort, SK, were found in subgroup A, C and D, which denoted avirulence gene diversity for a specific location.

The SSP dendrogram had two main groups (I, II) (Fig. 3B). Group I contained 127 isolates divided into two subgroups (A, B), and group II had only 23 isolates further divided into two subgroups (C, D). Group II consisted of isolates from MB and AB, but not SK. Subgroup C contained 10 isolates, all collected from five locations (Holland, Cartwright, Lowe Farm, Carberry, and Carman) in MB in 2012-2014, except that one isolate CB07-100 was sampled in 2007-2008. Subgroup D had 13 isolates from five locations (Holland, Cartwright, Lowe Farm, Winker, and Plum Coulee) in MB, and two locations (Olds and Camrose) in AB. Group I could be divided into two subgroups: A and B. Subgroup A had 55 isolates collected exclusively in 2007-2008, while subgroup B had 72 isolates from 19 locations collected from both time periods. It was noticeable that isolates collected in 2007-2008 appeared in all four subgroups, but isolates from 2012-2014 were observed only in three subgroups (B, C, D), so isolates collected in 2007-2008 were more genotypically diverse in terms of SSP gene variations than those collected in 2012-2014. With respect to geographic locations, MB isolates were of the highest diversity in SSP genes as compared with SK and AB isolates, because these isolates appeared in the four subgroups (A, B, C, D); followed by AB isolates, which were found in three subgroups (A, B, D). Saskatchewan isolates from nine locations had the least variability, and were clustered in only two subgroups (A, B).

The whole-genome dendrogram formed two main groups (I and II) (Fig. 3C). Group I was comprised of 143 isolates partitioned into four subgroups: A, B, C and D. Subgroups A, B and D each contained 22 isolates, with 19 isolates originating from seven locations in MB, two isolates from AB, and one from SK. The major subgroup C was comprised of 118 isolates, accounting for 79% of the isolates from the three provinces, which represented the majority of the L. maculans population gathered in western Canada. Group II contained only seven isolates, which were all collected from MB except for CR07-97 from Camrose, AB. Group II were divided into two subgroups (E,F). The two isolates in subgroup E were collected from AB and MB, and the five isolates in subgroup F were all from MB. All SK isolates, regardless of the year of collection, were all clustered in this subgroup.

The Avr, SSP and whole-genome dendrograms resulted in different isolate groupings. This suggested that the level and complexity of genetic diversity varied with the set of variants used to examine these isolates. Regardless, the three dendrograms indicated high genetic diversity among these isolates. As shown by SSP and whole-genome dendrograms, the level of genetic diversity varied with the source of the isolates, with MB isolates being most diverse, followed by AB isolates. SK isolates appeared least diverse. Considering that use of variants in Avr or SSP genes alone would only reflect genomic diversity with respect to the L. maculans interaction with the differential hosts, and that genome-wide variants would be more informative and inclusive, we used genome-wide variants for further analysis.

PCA analysis of L. maculans isolates

To further illustrate the relationships of the isolates collected at different locations and times, PCA was performed using genome-wide variants. PC1, accounting for 17.8% of variation, and PC2 for 6.1%, were used to map the isolates (Fig. 4). Except for five outliers (CR07-48, CR07-58, CR07-97, 12CC-481 and 12CC-419) from AB, all SK and AB isolates clustered together close to the origin on the negative side along with approximately half of the MB isolates. The other isolates from MB, however, segregated widely on the right side of the PCA map (Fig. 4). This indicated that isolates from SK and AB were of similar genetic composition, while MB isolates were highly diverse, which was consistent with the previously described cluster analysis. The genetic diversity among isolates seemed to increase with years, as exemplified by SK and MB isolates. For examples, some isolates collected in 2012-2014, such as 12CC-387, 13CCSK19-07 and 12CC-329 from SK, and 12CC-481 and 12CC-419 from AB, were dispersed from the main cluster. The genetic variation also became greater with time for MB isolates, as over half of the isolates shifted away from the isolates sampled earlier in 2007-2008 (Fig. 4).

The difference in genetic diversity between years and locations may be reflected in variant changes. In order to find DNA variants associated with years and regions, we carried out a GWAS study using TASSEL 5.2.31 [31] with the variants as genotype source and collection years, locations and provinces as traits. TASSEL offered two different models, the generalized linear model (GLM) and the multiple linear model (MLM) for marker-trait association. We tested the validity of GLM (PCA) and MLM (PCA+K) by evaluating their quantile-quantile plots (QQ-plots). A QQ-plot deviating from the null hypothesis across the entire standard
distribution reflects undetected population stratification or cryptic relatedness. Consequently, seemingly significant variants could be statistically unreliable, whereas truly significant variants would generate deviations only at the end of plot range [32]. While GLM (PCA) employs principal components only as covariates to control the population structure, MLM (PCA+K) uses both principle component and kinship to eliminate or reduce the confounding effect of a real association. The MLM (PCA+K) test proved to have higher correction efficiency because the base of its QQ-plot was consistent with the reference line (Fig. 5B), as compared with the GLM (PCA) approach (Fig. 5A). Therefore, Manhattan plots of MLM (PCA+K) tests were examined to find variants associated with years, locations and provinces, and the variants were mapped to SCs with significant p values for potentially associated traits. There were the variants associated with the year (Fig. 5C), province (Fig. 5D) and location of the field (Fig. 5E) of collections. To reduce false positives, the false discovery rate (FDR) approach [33] was adopted for selection of variants. As a result, none of the variants was associated with the province or field site of collection, while 143 variants in SC12 were associated with the year of collection significantly. Of the 143 variants, 20 were located in egn4_Lema_T086290 (AvrLm4-7), 29 in egn4_Lema_T086300 and six in egn4_Lema_T086310. The rest of the variants were located in non-coding regions in proximity to these three genes. The other four significant variants were found in SC20, SC8, SC10 and SC11.

Population structure of L. maculans

Initially, the popular Bayesian-based STRUCTURE software was tested for population structure analysis; it imposed tremendous computational burden due to the large number of variants involved, making implementation extremely inefficient. Landscape and ecological association (LEA), an R package [22], was used to perform STRUCTURE-like analysis and estimate admixture coefficients efficiently with improved algorithms [21, 34]. LEA was run first to establish a curve of cross-entropy versus number of ancestral populations, the optimal number K=3 indicated three ancestral populations in western Canada (Fig. 6A), which are hereafter referred to as sub-pop1, sub-pop2 and sub-pop3. The function snmf of LEA differentiated the 150 isolates by partitioning their genome into the three ancestral populations, with probability denoted by an admixture coefficient (Fig. 6B). Sub-pop1 appeared to be the dominant sub-population, consisting of 76% of the 150 isolates with a correlation coefficient of >0.8. Four isolates, one from AB (CR07-97) and three from MB (PC07-33, PC07-29 and 13CCMB02-06), constituted sub-pop2 with the coefficient at also >0.8. In addition, five isolates from MB (PC07-24, PC07-30, PC07-67, PC07-97 and MT07-31) were members of sub-pop3 with a correlation coefficient of 0.9998. All the other isolates were mixes of these sub-populations to a greater or lesser extent.

For a better understanding of the temporal and spatial distribution of the three sub-populations, the means of the admixture coefficients for isolates collected from the same field locations were calculated and mapped to these locations for 2007-2008 (Fig. 6C) and 2012-2014 (Fig. 6D), separately. The L. maculans population admixture varied among provinces; all SK isolates from 10 localities, regardless of the year of collection, were mainly composed of sub-pop1. The AB isolates were made up of two subpopulations (sub-pop1 and sub-pop2), exemplified by isolates from Camrose in 2007-2008 and Olds in 2012-2014, with admixture proportions of sub-pop1/sub-pop2/sub-pop3 being 85/13/1% and 42/51/7%, respectively (Table 2). This indicates the expansion of sub-pop2 between 2007-2008 and 2012-2014. The MB isolates turned out to have the greatest admix at both collection times. In 2007-2008, sub-pop1 was comprised of about 90% of the L. maculans population in Brandon and Carberry, but the admixture proportion (sub-pop1/sub-pop2/sub-pop3) in Plum Coulee, collected in 2007-2008, was 23/29/48%, indicating three major subpopulations. In 2012-2014, sub-pop2 in four locations (Carman, Winkler, Low Farm and Cartwright) increased noticeably, ranging between 39.2 and 55.7%, similar to that observed at Olds, AB. At the same time, sub-pop3 accounted for 5.5-11.6% of the isolates at all seven locations in MB, but sub-pop1 was significantly lower as compared with Brandon and Carberry in 2007-2008.
Table 2. Admixture coefficient of isolates

| years     | Provinces | Locations   | number of isolates | Admixture coefficient |
|-----------|-----------|-------------|--------------------|-----------------------|
|           |           |             |                    | subpop1 | subpop2 | subpop3 |
| 2007-2008 | AB        | Vegreville  | 4                  | 99.2    | 0.3     | 0.5     |
| 2007-2008 | AB        | Vermilion   | 8                  | 95.7    | 0.3     | 4.0     |
| 2007-2008 | AB        | Camrose     | 12                 | 85.8    | 13.1    | 1.1     |
| 2007-2008 | MB        | Plum Coulee | 9                  | 23.3    | 28.7    | 48.0    |
| 2007-2008 | MB        | Brandon     | 10                 | 89.4    | 8.7     | 1.9     |
| 2007-2008 | MB        | Carberry    | 11                 | 84.1    | 11.1    | 4.8     |
| 2007-2008 | SK        | Indian head | 8                  | 96.1    | 0.8     | 3.0     |
| 2007-2008 | SK        | Scott       | 9                  | 98.5    | 0.9     | 0.6     |
| 2007-2008 | SK        | Melfort     | 19                 | 93.6    | 0.2     | 6.2     |
| 2012-2014 | AB        | Olds        | 2                  | 42.3    | 51.0    | 6.6     |
| 2012-2014 | MB        | Carman      | 1                  | 43.7    | 40.8    | 15.5    |
| 2012-2014 | MB        | Winkler     | 2                  | 41.8    | 46.2    | 11.9    |
| 2012-2014 | MB        | Killarney   | 2                  | 75.7    | 14.0    | 10.3    |
| 2012-2014 | MB        | Lowe Farm   | 3                  | 27.7    | 55.7    | 16.6    |
| 2012-2014 | MB        | Franklin    | 3                  | 76.3    | 18.1    | 5.5     |
| 2012-2014 | MB        | Holland     | 8                  | 67.9    | 25.4    | 6.6     |
| 2012-2014 | MB        | Cartwright  | 10                 | 49.2    | 39.2    | 11.6    |
| 2012-2014 | SK        | Scott       | 1                  | 97.5    | 0.0     | 2.5     |
| 2012-2014 | SK        | Speers      | 3                  | 91.2    | 8.2     | 0.7     |
| 2012-2014 | SK        | Biggar      | 3                  | 91.7    | 5.4     | 2.9     |
| 2012-2014 | SK        | Watrous     | 4                  | 96.8    | 0.1     | 3.1     |
| 2012-2014 | SK        | Langenburg  | 4                  | 89.5    | 7.3     | 3.2     |
| 2012-2014 | SK        | Meota       | 5                  | 97.0    | 0.2     | 2.8     |
| 2012-2014 | SK        | Goodeve     | 9                  | 95.8    | 2.8     | 1.4     |

Discussion

This study characterized genome wide DNA variants of the *L. maculans* isolates collected in western Canada and investigated the genetic diversity and population structure of *L. maculans* using 31,870 genome-wide DNA variants identified from 150 isolates based on temporal and spatial samplings. These isolates represented the *L. maculans* population in western Canada in 2007-2008 and 2012-2014. To the best of our knowledge, this was the first time that the abundance of genome-wide DNA variants was applied to the study of genetic diversity in *L. maculans*.

Our study found that transition variants were the dominant type, especially in AT-blocks, which had 19% more transition variants than in GC-blocks. This confirmed the presence in *L. maculans* of RIP mutations, caused by a premeiotic repeat-inactivation mechanism, leading to C-to-T and G-to-A changes [35], which resulted in a bipartite structure, namely, alternating AT- and GC-blocks [20]. The SNPs in *AvrLm4-7* were reportedly affected by RIP mutations [5] because the *AvrLm4-7* gene is located in an AT-
block border. With this in mind, we examined the full sequences of all 28 genes contained in AT-blocks, but no variant was detected in these genes. Of the 28 genes, 25 were located in SC30. The BLAST annotation revealed that these genes were related to mitochondrial functions (data not shown), suggesting that SC30 is located in mitochondria. The DNA mutations in these mitochondrial genes might not be inherited due to their extremely detrimental effect on L. maculans fundamental biological function, or the RIP mechanism may not work on mitochondrial DNA.

Pathologists and breeders have long understood the importance of genetic variation in L. maculans with regard to the effectiveness and durability of disease resistance; the outcome of host-pathogen interaction often leads to avirulence gene mutation in the pathogen and consequently resistance breakdown in canola/rapeseed cultivars within short periods of time [9, 36]. Avirulence genes are members of SSP-encoding genes that play key roles in interactions with hosts [37]. Selection pressure tends to indirectly target the genes localized in rapidly-evolving transposable element-rich regions [38]. For this reason, SSP genes are supposed to mutate at a higher frequency than non-SSP genes. Counter-intuitively, some non-SSP genes accumulated more mutations than SSP genes showed by this study.

Genetic diversity of L. maculans has been studied with a number of molecular markers, including those use for mating types [4] and avirulence genes [18,19] studied the genetic diversity of the L. maculans population in Australia using 384 SNPs selected from the list of 21,814 SNPs described by Zander et al. [19], and determined that the SNPs were located in 76 SCs. This raised the question of how variants should be chosen for a study. We carefully selected 31,870 variants (SNPs and InDels) by applying stringent quality parameters without the need for further validation because it is common practice to use NGS-derived variants on fundamental studies of humans [39], plants [40], and fungi [41]. In this study, we categorized variants by their genomic regions including AT-blocks, GC-blocks and SSP genes, and used these variants to perform cluster analyses. It was shown that dendrograms using variants from different regions differed; genetic diversity measured by variants selected from different genomic regions could only partially represent that based on genome-wide variants, with cophenetic correlation coefficients ranging from 0.14 to 0.96. We compared Avr and SSP dendrograms with genome-wide dendrograms, and reached three conclusions. (1) The Avr dendrograms exhibited the lowest level of genetic diversity, because it had only 53 leaves. This was due to the fact that some isolates shared the same leaf, such as those in leaf A or leaf B (Fig. 3 A), while both SSP and whole-genome dendrograms had 150 leaves. (2) Both Avr and SSP dendrograms showed that isolates in 2007-2008 had higher genetic diversity than those from 2012-2014; isolates in 2007-2008 were found in four subgroups, while isolates in 2012-2014 were placed in three subgroups. (3) For isolates collected in the same years or from the same locations, isolates were highly dispersed across whole-genome dendrograms. Therefore, it would be beneficial to provide variant distribution information for a phylogenetic tree so that the genetic diversity could be better understood. For example, a dendrogram constructed from variants in SSP genes identified the similarity of isolates as related to their interaction with the host.

The L. maculans clonal population had been previously concluded according to mating type proportion and identical genotypes [16,42]. We also observed that isolates from a location clustered together in the Avr dendrogram based on 48 variants. For example, there were 19 isolates collected from Melfort, SK, in 2007-2008 used in this study, and 11 of them shared the same genotype and cladded together in leaf B (Fig. 3A). When 31,870 variants were considered in the analysis, the whole genome dendrogram depicted a different picture; despite the 11 isolates that were clustered in the same subgroup (Fig. 3C), they were distributed into different clads. From the perspective of this study, clonal multiplication might be true, but to a small geographical area; it still remains unclear how the area should be defined reasonably.

Our GWAS result showed that there were no variants associated with province or specific location, which meant that the variability identified were not geographically specific. Interestingly, 143 variants were found to be associated with the year of collection, with 20 of them located in AvrLm4-7. It was also found that AvrLm4-7 was among the most mutated of the SSP genes, indicating AvrLm4-7 mutated significantly between 2007-2008 and 2012-2014. Theoretically, an Avr gene mutates in response to selection pressure imposed by its cognate R genes [43]. In western Canada, however, the major gene integrated into blackleg resistant cultivars was Rlm3 [44], therefore, it is understandable that AvrLm3 exhibited nine SNPs. It was bit surprising that AvrLm4-7 had five times as many SNPs as in AvrLm3, despite the reality that the corresponding R genes had not be used commercially in the region. All isolates selected from 2012-2014 carried Avr4-7, compared to only 50% of the isolates from 2008-2009 carried Avr4-7. This would show a substantial increase of AvrLm4-7 in the L. maculans population between 2007-2008 and 2012-2014, and this is consistent with previous race-profiling results that indicated that the frequency of AvrLm3 decreased from 2008 to 2012 in
western Canada, while \textit{AvrLm4-7} increased based on phenotypical testing on host differentials \cite{44}. \textit{AvrLm7} was proven to have a masking effect on \textit{AvrLm3} \cite{45}; thus, these mutations in \textit{AvrLm4-7} might be an effort for \textit{L. maculans} to disrupt \textit{AvrLm3} function. Therefore, the loss of function of \textit{AvrLm3} or the overcoming of \textit{Rlm3} by \textit{L. maculans} \cite{46}, was realized likely through the activation of \textit{AvrLm4-7} because the six non-synonymous SNPs in \textit{AvrLm3} do not compromise \textit{AvrLm3} function in the absence of \textit{AvrLm7} \cite{27}. In addition, \textit{AvrLm4-7} was reported to contribute to pathogen fitness \cite{47}; any environmental stress which threatens pathogen fitness might trigger DNA mutations in the gene.

From the population structure study, at least three sub-populations were identified in western Canada based on the isolates collected in both 2007-2008 and 2012-2014. Of these sub-populations, sub-pop1 was apparently the predominant group, but genomic integration with subpop2 and subpop3 was detected in isolates from MB and AB. This indicated that the \textit{L. maculans} population in western Canada was also panmictic, as those previously reported in Australia \cite{18} and France \cite{4,15}. A panmictic population allows for high gene flow. Consistent with previous findings using minisatellite markers \cite{16,48}, isolates from MB were the most admixed and the admixture proportion also changed with time. There were three ancestral populations present in MB isolates: in 2007-2008 at Plum Coulee, which is about 150 linear kms from Brandon and Carberry MB, the pathogen population displayed different admixture proportions with subpop3 accounted for 48% of Plum Coulee isolates but only 1.9% of Brandon isolates and 4.8% of Carberry isolates. In 2012-2014, there was no sample available from Plum Coulee, but we were able to select samples from Lowe Farm, which was 24 kms to Plum Coulee. At this location, subpop2 was at 55.7% while subpop3 at only 16.6%. For the same reason, Franklin and Brandon MB, which are only 50 kms apart, were compared, and subpop2 in Franklin increased by 10% in 2012-2014 relative to Brandon in 2007-2008. Similarly, subpop2 in AB was 13.1% at Camrose relative to 51% at Olds. For SK isolates from 2007-2008, subpop2 was detected at very low levels (1%) but the subpopulation was higher (7.3%) in 2012-2014, ranging from 5.3 to 7.3% (Data not shown). Compared with subpop2, subpop3 changes less. Despite a lack of isolates from the far south of AB or far south of SK, there was still a trend for the new subpopulation; subpop2 appeared to migrate from south to north, while the migrating trend for the other subpopulation was less clear. By using 14 minisatellite markers on 12 SCs, Dilmaghani et al. (2012) \cite{16} suggested that the \textit{L. maculans} population in western Canada appeared to be derived from a source in Ontario. This may explain why MB isolates had the highest level of genetic diversity, with a higher proportion of subpop2 and subpop3 than SK isolates. However, it raises the question of why subpop2 was considerably higher at Camrose and Olds in AB than locations in SK. It is possible that subpop2 in AB might have been transported from MB, or there might be another pathway for the subpopulation to spread from MB to AB, while avoiding SK. The data from the current study can not seem to clarify that.

**Conclusion**

In this study, 31,870 polymorphic variants were identified from 150 \textit{L. maculans} isolates collected from western Canada in 2007-2008 and in 2012-2014. For the both collection times, three sub-populations (sub-pop1, sub-pop2, and sub-pop3) were differentiated with uneven spatial distribution across the region. Sub-pop1 was the major sub-population of the \textit{L. maculans} population in SK. The AB \textit{L. maculans} population was composed of sub-pop1 and sub-pop2, while all three sub-populations were prevalent in MB fields.

**Materials And Methods**

**Selection of \textit{L. maculans} isolates**

To monitor changes in race structure of the blackleg pathogen in western Canada, 800 \textit{L. maculans} isolates were collected from trap plots of the susceptible cultivar 'Westar' during 2007-2008. An additional 790 isolates were from commercial canola fields collected between 2012 and 2014. A set of 14 cultivars or breeding lines of \textit{Brassica} spp. (Table S2) were used as differential hosts to determine the presence/absence of \textit{Avr} genes following established methodology \cite{49}. Hierarchical clustering the 1590 isolates was implemented using the squared Euclidean distance and centroid linkage implemented by the function hclust in the R statistical package v3.6.2 (https://www.rdocumentation.org/packages/stats/versions/3.6.2), and isolates were first randomly selected from major branches of the resulting dendrogram to represent different races, then collection locations were considered
to include canola growth fields as many as possible. The number per location varied due to isolate availability. Finally 162 isolates were chosen for sequencing.

**Extraction of fungal DNA**

The isolates selected were cultured in 50 ml sterile centrifuge tubes containing 10 ml liquid V8 medium (800 ml of distilled water, 200 ml of V8 juice, 0.7 g of calcium carbonate, and 100 mg of streptomycin sulfate) in an incubator (Lab Line Orbit Environment shaker) at 20°C and 300 rpm for 7 days after which a mycelial ball of *L. maculans* had formed. The mycelial balls were collected and dried in a Labconco Benchtop Freeze Dryer (Labconco, Kansas City, MO), and then employed for DNA extraction using the EZNA Fungal DNA mini kit (Omega Bio-Tech, Norcross, GA) following the manufacturer's instructions.

**Preparation of DNA libraries and DNA sequencing**

DNA samples of each isolate were quantified using the Quant-iT™ PicoGreen® dsDNA Assay kit (Thermo Fisher Scientific, Waltham, MA). Libraries were generated using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England BioLabs, Whitby, ON) following the manufacturer's instructions. Illumina adapters containing barcodes were purchased from PerkinElmer (Waltham, MA). Size selection of libraries at 350-450 bp was performed using SPRiSelect beads (Beckman Coulter, Brea, CA) prior to PCR amplification (seven cycles). Libraries were quantified using the Kapa Illumina GA with Revised Primers-SYBR Fast Universal kit (Kapa Biosystems, Wilmington, MA). The average size of the fragments was determined using a LabChip GX (PerkinElmer, Waltham, MA) instrument. The libraries were normalized, denatured in 0.05 N NaOH and then diluted to 200 pM, and neutralized using HT1 buffer. Clustering was performed on an Illumina cBot system and the flow cell was run on a HiSeq 4000 for 2 x 100 cycles (paired-end mode) following the manufacturer's instructions. A phiX library was mixed with libraries at 1% as control. The Illumina control software was HCS HD 3.4.0.38, the real-time analysis program was RTA v. 2.7.7. The program bcl2fastq v2.18 was used to demultiplex samples and generate fastq reads.

**Sequence alignment and SNP discovery**

The NGS data was analyzed using the software package Lasergene Genomics Suite 13 including SeqMan NGen, SeqMan Pro and QSeq (DNASTAR, Madison, WI) [50]. Paired-end short reads of 100 bp length were first assembled to the reference genome sequence of *L. maculans 'brassicae'* isolate v23.1.3 using SeqMan NGen following whole genome reference-guided workflow with mer size set at 100 nucleotides and minimum match at 80%. The high minimum match threshold was intended to deal with repetitive sequences in *L. maculans*. Because SeqMan NGen constantly computes match percentage for any alignment of 50 bases to compare with the minimum match threshold in the course of assembly, all alignment with a minimum match greater than the threshold would be dropped. The sequence of wa74_scaffold00338 (GenBank: FO906748.1) containing the gene *AvrLm3*, which was previously reported to be absent from the aforementioned reference genome [20], was included as the reference template to detect variants in *AvrLm3* and its flanking regions. Variants (SNPs and InDels) relative to the *L. maculans* reference isolate were determined using QSeq based on the following filtering criteria: P not ref ≥90%, Q call ≥25, minimum SNP percentage ≥90%, depth ≥10 (https://www.dnastar.com/manuals/ seqman-ngen/17.1/en/topic/part-a-setting-up-the-assembly-in-seqman-ngen). Short InDels were detected in comparison with the reference genome by MAQ algorithm [51]. The algorithm uses mate-pair information in original sequence files and estimates the error probability of each read alignment. The error probabilities are also summarized for the final genotype calls, using a Bayesian statistical model that incorporates the mapping qualities, error probabilities from the raw sequence quality scores. Variant tables of isolates were exported, filtered (Q call ≥25, MAF ≥ 5, depth ≥ 10), and combined for further SNP data mining. The SNP distribution in specific genomic regions and variant composition was investigated. To explore variant density in different blocks, AT- and GC- block segments in supercontigs (SC) were first determined by Excel VBA programming with sliding windows of 120 nucleotides. In this study, any continuous nucleotide stretch meeting the criteria of GC content <33.9% (15) and length ≥1000 base pairs was considered an AT-block. All "N" masked genome regions were excluded from the AT-block assessment. To predict genes located in AT blocks, three criteria were employed: (1) a gene is entirely located in an AT-block, if the full sequence of a gene is contained in AT block, (2) part of the gene sequence overlaps with an AT-block, and (3) a gene entirely resides in a GC-block with one end (5’ end or 3’ end) falls within 1000 bases to an AT-block.

**Variant imputation and data analysis**
There were 40,000 missing calls in the variant table to be used for subsequent analysis. PCA and cluster analysis does not support miss data, so miss calls in our SNP panel were imputed prior to the analysis. Though there were various widely-used imputation algorithms such as BEAGLE, IMPUTE and MACH [52], they were found difficult to be applied onto haploid *L. maculans*, for example, a pedigree file was unavailable for our isolates randomly collected from fields. To overcome this, an R script was developed to impute the miss calls based on a reference haplotype. The script was tested on multiple simulation SNP panels with 40,000 miss calls and imputation accuracy reached 90~95%. The imputed variant panel was then converted into a table of 0 and 1 for PCA and cluster analysis [53].

*L. maculans* population structure was assessed using the *snmf* function of the STRUCTURE-like R package LEA [22]. Cross-entropy range of K=1~8 were set to run *snmf* (ploidy=1, iteration=200, tolerance error=0.001%) to test the number of sub-populations, and admixture coefficients of isolates were retrieved and depicted in a barplot. The geographical map of population admixture was created using the R package maptools [54], mapplot [55], ggmap [56], and Mapdata (https://www.rdocumentation.org/packages/mapdata/versions/2.3.0).

PCA analysis was performed using the R package FactoMineR [57] and Factoextra [58]. As a popular tool for multivariate data analysis, FactoMineR offered methods for PCA and CA (correspondence analysis); the latter was specially designed for processing categorical data sets. Considering that the *L. maculans* DNA variant panel was binary consisting of 0 and 1, both PCA and CA methods were compared only to find slight differences in terms of eigenvalues and isolate positions on their maps of the top two principal components. For ease of plotting, PCA was selected to differentiate and group isolates.

GWAS was performed using all variants in the software TASSEL v5.2.31 [31]. The insertion and deletion was converted into “+” and “−” prior to workflow implementation. General linear model (GLM) and mixed linear model (MLM) were primarily tested to find an option fit for our study. Correlation method was selected for the built-in PCA algorithm and normalized_IBS for kinship.

**Declarations**

**Competing interests**

The authors declare that they have no competing interest.

**Ethics approval and consent to participate**

Not applicable

**Consent to publish**

Not applicable

**Availability of data and materials**

The datasets used and/or analyzed during the current study available from the corresponding author on reasonable request.

**Funding**

This work was funded by a competitive grant from Agriculture and Agri-Food Canada. The funding body played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

**Author contributions**

FY conceived of and designed the study; QC conducted the experiments, analyzed data and drafted the manuscript; GP and RK provided important information and materials. All authors reviewed the manuscript and approved the final draft.

**Acknowledgement**
We would like to thank Agriculture and Agri-Food Canada for supporting this research.

Abbreviations

Avr: avirulence; InDel: insertion and deletion; MAF: minimum allele frequency; MC: median coverage; NGS: Next Generation Sequencing; PCR: Polymerase chain reaction; RAPD: random amplified polymorphic DNAs (RAPD); RFLP: restriction fragment length polymorphisms; SC: Supercontig; SNP: Single-nucleotide polymorphism; SSRs: Simple sequence repeats; TC: template coverage; Ts: transition; Tv: transversion; GLM: General Linear Model; MLM: mixed linear model; AB: Alberta; SK: Saskatchewan; MB: Manitoba

References

1. Stukenbrock EH, Bataillon T. A population genomics perspective on the emergence and adaptation of new plant pathogens in agro-ecosystems. *PLoS Pathogens*. 2012, 8(9):e1002893.
2. Peng G, Liu X, McLaren DL, McGregor L, Yu F. Seed treatment with the fungicide flupyradroxyl limits cotyledon infection by *Leptosphaeria maculans* and reduces blackleg of canola. *Canadian Journal of Plant Pathology*. 2020, 2:1-3.
3. Yu F, Lydiate DJ, Gugel R, Sharpe A, Rimmer SR. Intrgression of *Brassica rapa* subsp. *sylvestris* blackleg resistance into *B. napus*. *Molecular breeding* 2012, 30:1495-1506.
4. Gout L, Eckert M, Rouxel T, Balesdent MH. Genetic variability and distribution of mating type alleles in field populations of *Leptosphaeria maculans* from France. *Applied and Environmental Microbiology* 2006, 72:185-191.
5. Daverdin G, Rouxel T, Gout L, Aubertot J-N, Fudal I, Meyer M, Parange F, Carpezat J, Balesdent MH. Genome structure and reproductive behaviour influence the evolutionary potential of a fungal phytopathogen. *PLoS Pathogens* 2012, 8:e1003020.
6. McDonald BA, Linde C. The population genetics of plant pathogens and breeding strategies for durable resistance. *Euphytica* 2002, 124:163-180.
7. Rouxel T, Penaud A, Pиноchet X, Brun H, Gout L, Delourme R, Schmit J, Balesdent MH. A 10-year survey of populations of *Leptosphaeria maculans* in France indicates a rapid adaptation towards the Rlm1 resistance gene of oilseed rape. *European Journal of Plant Pathology* 2003, 109:871-881.
8. Li H, Sivasithamparam K, Barbetti M. Breakdown of a *Brassica rapa* subsp. *sylvestris* single dominant blackleg resistance gene in *B. napus* rapeseed by *Leptosphaeria maculans* field isolates in Australia. *Plant Disease* 2003, 87:752.
9. Zhang X, Peng G, Kutzer HR, Balesdent MH, Delourme R, Fernando WGD. Breakdown of Rlm3 resistance in the *Brassica napus–Leptosphaeria maculans* pathosystem in western Canada. *European Journal of Plant Pathology* 2016, 145:659-674.
10. Mahuku GS, Goodwin PH, Hall R, Hsiang T. Variability in the highly virulent type of *Leptosphaeria maculans* within and between oilseed rape fields. *Canadian Journal of Botany* 1997, 75:1485-1492.
11. Purwantara A, Barrins J, Cozijnsen A, Ades P, Howle, B. Genetic diversity of isolates of the *Leptosphaeria maculans* species complex from Australia, Europe and North America using amplified fragment length polymorphism analysis. *Mycological research* 2000, 104:772-781.
12. Barrins JM, Ades PK, Salisbury PA, Howlett BJ. Genetic diversity of Australian isolates of *Leptosphaeria maculans*, the fungus that causes blackleg of canola (*Brassica napus*). *Australasian Plant Pathology* 2004, 33:529-536.
13. Barrins JM, Purwantara A, Ades PK, Salisbury PA, Howlett A BJ. Genetic diversity of isolates of *Leptosphaeria maculans* from a canola (*Brassica napus*) paddock in Australia. *Australasian Plant Pathology* 2002, 31:129-135.
14. Hayden HL, Cozijnsen AJ, Howlett BJ. Microsatellite and minisatellite analysis of *Leptosphaeria maculans* in Australia reveals regional genetic differentiation. *Phytopathology* 2007, 97:879-887.
15. Travadon R, Sache I, Dutech C, Stachowiak A, Marquer B, Bousset L. Absence of isolation by distance patterns at the regional scale in the fungal plant pathogen *Leptosphaeria maculans*. *Fungal biology* 2011, 115:649-659.
16. Dilmaghani A, Gladieux P, Gout L, Giraud T, Brunner PC, Stachowiak A, Balesdent MH, Rouxel T. Migration patterns and changes in population biology associated with the worldwide spread of the oilseed rape pathogen *Leptosphaeria maculans*. *Molecular Ecology* 2012, 21:2519-2533.
17. Tindall EA, Petersen DC, Nikolaysen S, Miller W, Schuster SC, Hayes VM. Interpretation of custom designed Illumina genotype cluster plots for targeted association studies and next-generation sequence validation. *BMC research notes*. 2010, 3:39.

18. Patel DA, Zander M, Van de Wouw AP, Mason AS, Edwards D, Batley J. Population diversity of *Leptosphaeria maculans* in Australia. *International Journal of Biology*. 2015 7:18.

19. Zander M, Patel DA, Van de Wouw A, Lai K, Lorenc MT, Campbell E, Hayward A, Edwards D, Raman H, Batley J. Identifying genetic diversity of avirulence genes in *Leptosphaeria maculans* using whole genome sequencing. *Functional & integrative genomics* 2013, 13:295-308.

20. Rouxel T, Grandaubert J, Hane JK, Hoede C, Van de Wouw AP, Couloux A, Dominguez V, Anthouard V, Bally P, Bourras S. Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. *Nature communications* 2011, 2:1-10.

21. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*. 2000,155:945-959.

22. Frichot E, François O. LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution* 2015, 6:925-929.

23. Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics* 2014,197:573-589.

24. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* 2009, 19:1655-1664.

25. Schafer JL. Multiple imputation: a primer. *Statistical Methods in Medical Research*1999, 8:3-15. doi: 10.1191/096228099671525676

26. Parlanse F, Daverdin G, Fudal I, Kuhn ML, Balesdent MH, Blaise F, Grezes-Besset B, Rouxel T. *Leptosphaeria maculans* avirulence gene *AvrLm4-7* confers a dual recognition specificity by the *Rlm4* and *Rlm7* resistance genes of oilseed rape, and circumvents *Rlm4*-mediated recognition through a single amino acid change. *Molecular microbiology* 2009, 71:851-863.

27. Plissonneau C, Blaise F, Ollivier B, Leflon M, Carpezat J, Rouxel T, Balesdent MH. Unusual evolutionary mechanisms to escape effector-triggered immunity in the fungal phytopathogen *Leptosphaeria maculans*. *Molecular ecology* 2017, 26:2183-2198.

28. Ghanbarnia K, Fudal I, Larkan NJ, Links MG, Balesdent MH, Profotova B, Fernando WDG, Rouxel T, Borhan MH. Rapid identification of the *Leptosphaeria maculans* avirulence gene *AvrLm2* using an intraspecific comparative genomics approach. *Molecular plant pathology* 2015, 6:225-210.

29. Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*. 1995, 57:289-300.

30. Bianchi D, Brancadoro L, De Lorenzis G. Genetic diversity and population structure in a vitis spp. core collection investigated by SNP markers. *Diversity* 2020, 12:265-282.

31. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 2007 23:2633-2635.

32. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews genetics*. 2008, 9:356-369.

33. Benjamin Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*. 1995, 57:289-300.

34. Bianchi D, Brancadoro L, De Lorenzis G. Genetic diversity and population structure in a vitis spp. core collection investigated by SNP markers. *Diversity* 2020, 12:103.

35. Idnurm A and Howlett BJ. Analysis of loss of pathogenicity mutants reveals that repeat-induced point mutations can occur in the Dothideomycete *Leptosphaeria maculans*. *Fungal Genetics and Biology*. 2003, 39: 31–37.

36. Rouxel T, Balesdent M. The stem canker (blackleg) fungus, *Leptosphaeria maculans*, enters the genomic era. *Molecular Plant Pathology* 2005, 6:225-210.
37. Derbyshire MC. Bioinformatic detection of positive selection pressure in plant pathogens: the neutral theory of molecular sequence evolution in action. *Frontiers in Microbiology* 2020, 11:644. doi:10.3389/fmicb.2020.00644

38. Möller M, Stukenbrock EH. Evolution and genome architecture in fungal plant pathogens. *Nature Reviews Microbiology* 2017, 15:756-771

39. Choudhury A, Ramsay M, Hazelhurst S, Aron S, Bardien S, Botha G, Chimusa ER, Christoffels A, Gamieldien J, Sefid-Dashti MJ, Joubert F. Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nature communications*. 2017, 8(1):1-2.

40. Yano K, Yamamoto E, Aya K, Takeuchi H, Lo PC, Hu L, Yamasaki M, Yoshida S, Kitano H, Hirano K, Matsuoka M. Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nature genetics*. 2016, 48(8):927.

41. Nimmo C, Shaw LP, Doyle R, Williams R, Brien K, Burgess C, Breuer J, Balloux F, Pym AS. Whole genome sequencing *Mycobacterium tuberculosis* directly from sputum identifies more genetic diversity than sequencing from culture. *BMC genomics*. 2019, 20(1):389.

42. Dilmaghani A, Gout L, Moreno-Rico O, Dias JS, Coudard L, Castillo-Torres N, Balesdent MH, Rouxel T. Clonal populations of *Leptosphaeria maculans* contaminating cabbage in Mexico. *Plant Pathology*. 2013, 62(3):520-32.

43. Rouxel T, Balesdent MH. Avirulence genes. eLS. 2010. doi : 10.1002/9780470015902.a0021267

44. Zhang X, Fernando WGD. Insights into fighting against blackleg disease of *Brassica napus* in Canada. *Crop and Pasture Science* 2018, 69:40-47.

45. Plissonneau C, Daverdin G, Ollivier B, Blaise F, Degrave A, Fudal I, Rouxel T, Balesdent MH. A game of hide and seek between avirulence genes AvrLm4-7 and AvrLm3 in *Leptosphaeria maculans*. New Phytologist 2016, 209:1613-1624.

46. Huang YJ, Evans N, Fitt BD, Li ZQ, Rouxel T, Balesdent MH. Fitness cost associated with loss of the AvrLm4 avirulence function in *Leptosphaeria maculans* (phoma stem canker of oilseed rape). In Sustainable strategies for managing Brassica napus (oilseed rape) resistance to *Leptosphaeria maculans* (phoma stem canker) 2006

47. Zou Z, Zhang X, Fernando WGD. Distribution of mating-type alleles and genetic variability in field populations of *Leptosphaeria maculans* in western Canada. *Canadian Journal of Plant Pathology* 2010, 32:77-85.

48. Burland TG. DNASTAR's Lasergene sequence analysis software. In *Bioinformatics methods and protocols* 2000, pp. 71-91. Humana Press, Totowa, NJ.

49. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research* 2008,18:1851-1858.

50. Nothnagel M, Ellingham D, Schreiber S, Krawczak M, Franke A. A comprehensive evaluation of SNP genotype imputation. Human genetics 2009,125:163-171.

51. Gauch HG, Qian S, Piepho HP, Zhou L, Chen R. Effective principal components analysis of SNP data. *bioRxiv* 2018, 1:393611.

52. Bivand R, Lewin-Koh N. maptools: Tools for reading and handling spatial objects. R package version 0.8. 2013; 23.

53. Kahle D, Wickham H. ggmap: Spatial visualization with ggplot2. *The R journal* 2013,5:144-161.

54. Lê S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. *Journal of statistical software* 2008, 25:1-8.

55. Kassambara A, Mundt F. Factoextra: extract and visualize the results of multivariate data analyses. *R package version* 2017,1:337-354.

56. Galili T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics*. 2015, 31:3718-20.
60. Dilmaghani A, Balesdent MH, Didier JP, Wu C, Davey J, Barbetti MJ, Li H, Moreno-Rico O, Phillips D, Despeghel JP, Vincenot L. The *Leptosphaeria maculans–Leptosphaeria biglobosa* species complex in the American continent. *Plant Pathology* 2009, 58:1044-1058.

61. Delourme R, Brun H, Ermel M, Lucas MO, Vallee P, Domin C, Walton G, Li H, Sivasithamparam K, Barbetti MJ. Expression of resistance to *Leptosphaeria maculans* in Brassica napus double haploid lines in France and Australia is influenced by location. *Annals of Applied Biology* 2008, 153:259-269.

62. Kutcher HR, Keri M, McLaren DL, Rimmer SR. Pathogenic variability of *Leptosphaeria maculans* in western Canada. *Canadian Journal of Plant Pathology*. 2007, 29:388-393.

63. Yu F, Gugel RK, Kutcher HR, Peng G, Rimmer SR. Identification and mapping of a novel blackleg resistance locus LepR4 in the progenies from *Brassica napus × B. rapa* subsp. sylvestris. *Theoretical and Applied Genetics*. 2013, 126:307-315.

64. Rouxel T, Willner, E, Coudard L, Balesdent Screening and identification of resistance to *Leptosphaeria maculans*(stem canker) in *Brassica napus* accessions. *Euphytica*. 2003, 133:219–231.

65. Delourme R, Pilet-Nayel ML, Archipiano, M, Horvais R, Tanguay XR, Rouxel T, Brun H, Renard M, Balesdent, MH. A cluster of major specific resistance genes to *Leptosphaeria maculans* in *Brassica napus*. *Phytopathology*. 2004, 94: 578–583.

66. Ansan-Melayah D, Balesdent MH, Buee M, Rouxel T. Genetic characterization of AvrLm1, the first avirulence gene of *Leptosphaeria maculans*. *Phytopathology*. 1995, 85:1525-1529.

67. Balesdent MH, Attard A, Kühn ML, Rouxel T. New avirulence genes in the phytopathogenic fungus *Leptosphaeria maculans*. *Phytopathology*. 2002, 92:1122-1133.