Title Page

Genome analysis using Single Nucleotide Polymorphisms in the fungal plant pathogen *Leptosphaeria maculans*

Manuel Zander[1,2], Dhwani A. Patel[1], Angela Van de Wouw[3], Kaitao Lai[2], Michal T. Lorenc[2], Emma Campbell[1], Alice Hayward[1], David Edwards[2], Harsh Raman[4], Jacqueline Batley[1]

1 School of Agriculture and Food Sciences and Centre for Integrative Legume Research, University of Queensland, Brisbane, QLD 4072, Australia

2 Australian Centre for Plant Functional Genomics and School of Agriculture and Food Sciences, University of Queensland, Brisbane, QLD 4072, Australia

3 School of Botany, University of Melbourne, VIC 3010, Australia

4 NSW Department of Primary Industries, Wagga Wagga Agricultural Institute, PMB, Wagga Wagga, NSW 2650, Australia

Contact

Email: j.batley@uq.edu.au

Phone: +61 (0)7 3346 9534

Abstract

Next generation sequencing technology allows rapid re-sequencing of individuals, as well as the discovery of single nucleotide polymorphisms (SNPs), for genomic diversity and evolutionary analyses. By sequencing two Isolates of the fungal plant pathogen *Leptosphaeria maculans,* the causal agent of blackleg disease in *Brassica* crops, we have generated a resource of over 76 million sequence reads aligned to the reference genome. We identified over 21,000 SNPs with an overall SNP frequency of one SNP every 2,065 bp. Sequence validation of a selection of these SNPs in additional isolates collected throughout Australia indicates a high degree of polymorphism in the Australian population. In preliminary phylogenetic analysis isolates from Western Australia clustered together and those collected from *B. juncea* stubble were identical. . These SNPs provide a novel marker resource to study the genetic diversity of this pathogen. We demonstrate that re-sequencing provides a method of validating previously characterised SNPs and analysing differences in important genes, such as the disease related avirulence genes of *L. maculans*. Understanding the genetic characteristics of this devastating pathogen is vital in developing long-term solutions to managing blackleg disease in *Brassica* crops.

Keywords

Leptosphaeria maculans – SNPs – Re-sequencing - Molecular markers – Blackleg disease - Brassica

Introduction

The most economically damaging pathogen of *Brassica* species, particularly canola (*Brassica napus*; rapeseed; oilseed rape), is *Leptosphaeria maculans* (Howlett et al. 2001). *L. maculans*, a filamentous Ascomycete, is the causal agent of phoma stem canker, commonly referred to as blackleg. In late stages of infection it spreads through the stem vasculature causing lesions and leading to poor growth, lodging and eventually plant death. This fungus is found in canola-growing regions world-wide such as Australia, Canada and Europe. Increased production of canola in these regions has led to a rise in the severity of the disease (Marcroft and Bluett 2008). In Australia alone *L. maculans* infection is responsible for an estimated $100 million (AUD) in crop losses each year, with average losses ranging from 15 – 48 % (AOF 2012) and significant efforts are underway to improve resistance to this disease (Tollenaere et al. 2012; Kaur et al. 2009).

Understanding the characteristics of *L. maculans* is vital for developing an effective and sustainable approach to the management of blackleg disease on *Brassica* species. Given the threat imposed by this pathogen on the global rapeseed industry, it is imperative to gain more insight into the genetic structure of this fungus. Such information will assist in developing resistant *Brassicas* and eventually reduce crop losses. The completion of the *L. maculans* genome sequence (Rouxel et al. 2011) was a significant development in the study of this fungal pathogen and provides a reference genome to which molecular markers can be physically mapped. This has been highly useful in other plant pathogens with sequenced genomes, such as the wheat pathogen *Stagonospora nodorum* (Hane et al. 2007) and the cereal pathogen *Fusarium graminearum* (Cuomo et al. 2007). Importantly, a reference genome also allows mapping of whole genome re-sequencing data, which is becoming a high-throughput, cost-effective method to study genome-wide diversity, particularly for the relatively small, lower complexity genomes of many fungal species. By re-sequencing the genome of different *L. maculans* Isolates variations in genome sequence and structure can be elucidated.

Previously genetic analyses of blackleg populations were performed using AFLP (Barrins et al. 2004), microsatellite and minisatellite markers (Hayden et al. 2007). SNPs (single nucleotide polymorphisms) are increasingly popular for studying genomic evolution and diversity (Lai et al. 2012; Batley and Edwards 2007; Duran et al. 2009). Sequence variation can have a major impact on how an organism develops and responds to the environment. The high heritability of SNPs makes them an excellent indicator of genetic diversity and phylogeny in pathogenic fungal species, such as *L. maculans*. Additional features of SNPs such as stability during evolution, low mutation rate and their high density in the genome also make them the tool of choice for such analyses (Appleby et al. 2009; Hayward et al. 2012b; Duran et al. 2010). Current large-scale SNP genotyping assays, such as the Illumina GoldenGate assay, are cost-effective and flexible for analysing large numbers of SNPs across multiple individuals (Edwards et al. 2012). To date, no previous studies have implemented the use of SNPs in genetic analyses of *L. maculans.*

Advances in genome sequencing technologies have revolutionised plant and fungal genomics (Edwards et al. 2013; Lee et al. 2012; Marshall et al. 2010). They have made genome sequencing, re-

sequencing and SNP discovery highly accessible, high-throughput and cost-effective (Batley and Edwards 2009; Hayward et al. 2012a; Hayward et al. 2012c). The process of whole genome re-sequencing involves aligning millions of short sequence reads generated on a single next-generation sequencing run to a reference genome sequence. Once this has been achieved, it is possible to identify genetic variation between individuals, which can be linked to variation in phenotype to provide molecular genetic markers and insights into gene function (Imelfort et al. 2009; Batley and Edwards 2009).

Here we describe the re-sequencing of two *L. maculans* Isolates for the identification of 21,814 SNPs. We demonstrate the application of a novel SNP calling method, SGSautoSNP (Lorenc et al. 2012) and its robustness and sensitivity in identifying polymorphisms in *L. maculans*. We describe the use of these SNPs for phylogenetic analysis, genome analysis, including SNP properties and density in relation to genomic position and predicted function.

Experimental methods

Fungal samples

The Isolates 04MGPP021 and 06MGPP041 (Raman et al., 2012), were used for next-generation sequencing, initial SNP prediction and validation. For simplicity, these are referred to as Isolates 21 and 41 respectively herein. A further set of 22 Isolates from various spatiotemporal sources, were used for validation of the SNPs and initial assessment of polymorphism (Table 1).

Growth conditions

L. maculans Isolates were grown on 20% Campbell's V8 juice (Australia), 2% agar with 10 µg/mL Rifampicin and 30 µg/mL Chloramphenicol in a growth cabinet at room temperature in a 12 hr/12 hr dark/light cycle. Light was provided by one GE 18 Watt (Tri-tech F18T8/865) bulb and one Crompton 18 Watt (F18T8 BLB blacklight blue) bulb. For long term storage at room temperature, 5 mm agar plugs and/or filter discs (Grade AA, Whatman^{TM}) of each Isolate, derived from culture plates, were stored in sterile water or in microfilm-sealed (Parafilm) 1.5 mL microcentrifuge tubes (Interpath Services) placed in a container of silica beads, respectively.

Microprep DNA extraction

For DNA extractions, 50 mL liquid cultures of 10% V8 juice supplemented with 10 µg/mL Rifampicin and 30 µg/mL Chloramphenicol were inoculated with a filter disc or plug and grown as above for 2-3 weeks. Fungal tissue was harvested by draining each liquid culture through an autoclaved piece of miracloth (Calbiochem) placed in a filter funnel. Mycelial tissue in the miracloth was squeezed to get rid of excess liquid and stored in sterile 15 mL centrifuge tubes (Cellstar) at -80 ℃ for DNA extractions.

A microprep protocol (Fulton et al. 1995) was followed for fungal genomic DNA extraction, with the following modifications. After incubation at 65 ℃, each sample was washed with 750 µL of Phenol: Chloroform and centrifuged for 5 mins at 10,000 rpm. The DNA samples were resuspended in 50 µL of DNase-RNase free water, quantified using a Nanodrop (ND-1000 spectrophotometer) and visualized for integrity on a 1% TAE-agarose gel containing ethidium bromide.

Next generation sequencing

Sequence libraries were prepared for the Isolates 21 and 41 using the Illumina paired end (PE) and mate pair (MP) Genomic DNA Sample Prep Kit (Illumina, San Diego, USA) according to the manufacturer's instructions. One PE and one MP library was made for Isolate 41, as well as one MP library from Isolate 21. The insert sizes for the libraries were selected to be ~ 500 bp for PE libraries and ~2,500 – 3,000 bp for MP libraries. The Illumina GAIIx platform provided by the AGRF (Australian Genome Research Facility) was used to generate paired end and mate pair short sequence reads of 100 bp length. The three libraries were run on seperate lanes of the sequencer. Upon completion of

sequencing, the sequence reads were quality trimmed to ensure accurate mapping and SNP prediction.

SNP prediction

Illumina paired sequence reads for Isolates 21 and 41 were aligned to the *L. maculans* ('brassicae') Isolate v23.1.3 reference genome (Rouxel et al. 2011) using SOAPaligner 2 (Li et al. 2009). The SOAP –r 0 parameter was used to allow a sequence read to align to the genome only once, as well as the –M 2 parameter, allowing only two mismatches to the reference genome. The software SGSautoSNP (Lorenc et al. 2012), was used for SNP prediction between the two *L. maculans* Isolates with default parameters.

SnpEff variant analysis

The snpEff variant annotation tool (Cingolani et al. 2012) was used to predict the effect of the identified SNPs on the annotated genome of *L. maculans* within different genomic DNA sequences, including putative exons, introns, and gene upstream and downstream sequences. The patterns of codon usage and the ratio of transitions/transversions resulting from SNPs were also recorded.

SNP validation

In order to validate the SNP prediction and ensure no ascertainment bias in the isolates used for SNP identification, a selection of 20 SNPs, with a SNP score of 2 or more from SGSautoSNP, were chosen for PCR and sequencing (Table 2). These were located on different SuperContigs of the v23.1.3 reference genome ranging from the largest (SC0 – 4.26 Mbp) to one of the smallest (SC60 – 6,199 bp). These were subsequently annotated as SNP 1-20 (Table 2). Primers were designed to achieve a product size of 200 – 400 bp.

PCR reactions were carried out in 20 µL containing 1 x Reaction Buffer with 20 mM $MgCl_2$ (Scientifix, Victoria, Australia), 10 mM dNTPs (Scientifix, Victoria, Australia), 2 units of iTAQ DNA Polymerase (Scientifix, Victoria, Australia), 0.1 mM of each primer and 20 ng of DNA template. The cycling conditions were as follows: 94 ℃ for 3 mins, followed by 30 cycles of 94℃ for 30 s, 58℃ for 40 s and 72℃ for 30 s and a final extension at 72℃ for 4 mins. PCR products were visualized on a 1% TAE-Agarose gel containing ethidium bromide (0.5 µg/mL). The GeneRuler™ 1 kb DNA Ladder or the GeneRuler™ 100 bp DNA Ladder were used as size standards (both ThermoScientific, MA, USA; 0.5 µg/ µL, 50 µg) .

PCR products were silica-purified (Boyle and Lew 1995) and eluted in 20 µL of DNAse RNAse free UltraPure™ H_2O (GIBCO-Invitrogen, CA, USA). An aliquot was then run on a 1% TAE-Agarose gel containing ethidium bromide (0.5 µg/mL) for visual quantification. For all downstream applications, tubes were centrifuged at 13,200 rpm for 5 mins prior to use to pellet silica beads. Purified PCR products were directly sequenced using the PD Sanger sequencing service at the Australian Genome Research Faculty (AGRF), Brisbane. Geneious Pro (Drummond et al. 2012) was used to align and

assemble all sequenced SNP regions per Isolate to the reference SuperContig used to identify the SNPs.

Phylogenetic analysis

Data from SNPs 1, 2, 3, 5, 6, 7, 8, 9, 11, 14, 15, 16, 17 and 18 for each isolate, including the reference, was used to produce an unrooted dendrogram computed based on the Euclidean distance of the Isolates, using the R project for statistical computing (The R Foundation for Statistical Computing version 2.13.0, 2009). Isolate 7 was eliminated from the tree due to insufficient sequence data. Both alignments and assemblies were scanned manually to detect novel SNPs in the flanking regions of the existing annotated SNPs. A 'dudi.pca' principal component analysis was conducted using the R project for Statistical Computing (The R Foundation for Statistical Computing version 2.13.0, 2009).

Results

Sequencing

To generate a SNP resource for *L. maculans* genome analysis we performed whole genome next generation sequencing (NGS) on two blackleg Isolates. In total, 76,313,272 short sequence reads of 100 bp were generated, of these 29,051,680 reads were from Isolate 21 (average insert size 3,000 bp) and 47,261,592 reads were from both libraries of Isolate 41 (19,120,582 from MP (average insert size 2,500 bp) and 28,141,010 from PE (average insert size 480 bp). Predicted genome coverage for these Isolates was high when sequence tags were aligned to the reference genome (Rouxel et al. 2011): 140x coverage for Isolate 41 and 86x coverage for Isolate 21. The minimum genome coverage was 4x and the maximum was 602x with the mean coverage being 28.308. The highest coverage was seen in SuperContigs 1 – 40. No significant difference in coverage was observed between AT and GC rich regions, with the average read coverage being uniform across the genome.

SNP discovery

SNP prediction using SGSAutoSNP (Lorenc et al. 2012) revealed 21,814 SNPs between Isolates 21 and 41 (Supplementary Table 1). This indicates an abundance of SNPs in the *L. maculans* genome, with an average of one SNP every 2,065 bp. SGSAutoSNP records only those SNPs where the polymorphism is present in a minimum of two unique, overlapping sequence reads per individual, minimising false-positive SNP calls due to possible sequence errors. To further provide a measure of confidence in SNP prediction this software also generates a SNP score: representing the minimum number of sequences per individual containing the polymorphism. In this study, the SNP score ranged from 2 to 43, with an average of 6.96.

For each SNP, the directionality of the change cannot be inferred from the data, thus polymorphisms were grouped alphabetically, that is, A>G and G>A were grouped as A>G and so on (Table 3). A

greater number of transitions (A>G or C>T) (17,663) than transversions (A>C, A>T, C>G or G>T) (4,151) were identified. This was consistent across all SuperContigs, with an average ratio of 4:1 transitions to transversions (Table 4).

SNP validation

As an initial validation of the predicted 21,814 SNPs between Isolates 21 and 41, 20 SNPs were selected for amplification and direct Sanger sequencing of PCR products. These SNPs were chosen to represent a range of SGSautoSNP scores, including the lowest possible score of 2, up to 23, with an average of 10.5. Amplification was successful for 18 (90%) of the predicted SNPs and 14 of these were amplified and sequenced in a further 22 Isolates (Table 1) to check for ascertainment bias.

Of the 18 SNPs amplified in Isolates 21 and 41 all were shown to be true polymorphisms, following alignments of the obtained sequences to the *L. maculans* whole genome shotgun sequences using Geneious Pro (Drummond et al. 2012). This gave an accuracy of 100% for SGSAutoSNP-based SNP prediction. Of the 14 SNPs which were subsequently analysed across 22 samples, six were within coding regions and eight within intergenic regions. Eight of the 14 SNPs were transitions, five were transversions, and one was a tri-allelic variant which was both a transition and a transversion. The polymorphism information content (PIC) of these, representing the usefulness of each SNP as a marker for population genetic and phylogenetic analysis, ranged from 0.06 to around 0.5 for different SNPs with an average of 0.33. This indicates that the tested SNPs provide a good level of polymorphic information for use in downstream phylogenetic studies. There was no observable difference in the PIC values according to SNP type (transition/transversion) or location in the genome (Table 4). Similarly no observable difference in PIC values could be seen based on the SNP score.

Phylogenetic analysis

In order to determine the suitability of the SNPs for phylogenetic analysis and to understand the genetic relationships among the *L. maculans* isolates studied, cluster analysis was performed, based on the combined sequences of the 14 SNP loci. Isolate 7 failed to amplify and was discarded from further analysis. The Euclidean distances between pairs of accessions were then used to generate a dendrogram (Figure 1). On the phylogenetic tree, as indicated in Figure 1, seven of the nine isolates collected from Western Australia grouped together in one clade. The phylogenetic tree indicates that, based on genetic distance, Isolates 2 and 3 are identical with these SNPs. These isolates were both collected from *B. juncea*. Isolates '21' and '41' that the SNP discovery was performed on are considerably genetically different. Isolate 14 is shown to be an outlier. No association can be observed between Isolate similarity and place of collection, year of collection or cultivar source.

The results from the Principal Component Analysis confirmed the results obtained from the phylogenetic analysis. There is a random spread of Isolates along the two principal component axes, with no relatedness between isolates. Isolates 21 and 41 are again shown to be different and 14 is an

outlier (Figure 2). In contrast, Isolates 2 and 3 group slightly differently. Axis 1 represents principal component 1 and axis 2 represents principle component 2.

SNP properties

Analysis indicates that there is, for the most part, a consistent average SNP frequency across the genome of *L. maculans*, with the majority of SuperContigs over 500,000 bp in length ranging from one SNP every 1,166 bp (SuperContig 22) to one SNP every 3,168 bp (SuperContig 10). Exceptions to this finding are SuperContigs 30 and 37, which display an unusually low number of polymorphisms; being one SNP every 38,715 bp and 17,397 bp, respectively (Figure 3).

To predict the possible functional effect of SNPs in *L. maculans*, SnpEff was used to determine modifying SNP effects, based on the predicted coding regions of the annotated genome. The total number of effects that the predicted SNPs have on the genome of *L. maculans* was 58,434, with a ratio of missense/silent mutations of 1.0263. The predicted modifying effects of these SNPs is summarised in Table 5. The highest number of SNPs can be seen in upstream and downstream regions (~75%) of predicted genes (Figure 4). Coding regions are less affected with only 6.03 % of modifying effects being predicted within coding regions.

SNP density

SNP density can be seen to vary across the supercontigs (Figure 5), including in flanking regions to *AvrLm* genes. SC6 shows a low SNP frequency between bases 190,000 and 245,000 (Figure 5). This region contains ~22 predicted genes, including deoxyribonuclease tatD gene, asparagine synthetase, exportin-1 and transport-protein sec23, all of which play important roles in cell function and metabolism. Isolates 21 and 41 were highly similar to the reference in all these genes, further highlighting the likely conservation of these putative genes. No difference in sequence read coverage was seen between any of these regions (30-50 x coverage throughout), further suggesting that these polymorphisms are valid. The high peak seen at base 970,000 of SC6 shows a SNP frequency of 41 SNPs/10Kb, which is around 1 SNP every 250 bp. This region contains no predicted genes and may show this level of diversity due to low selection pressure in this region. SC12 exhibits a SNP frequency of up to 16 SNPs/10Kb at the AvrLm4-7 gene locus.

***AvrLm* gene regions**

Avirulence genes are a sub-class of effector proteins that elicit host immune responses, and are involved in resistance and susceptibility of the host to a pathogen such as *L. maculans*. Alignment of sequence tags from Isolates 21 and 41 to the reference genome allowed comparison of previously published avirulence genes in *L. maculans: AvrLm1* (Gout et al. 2006)*, AvrLm4-7* (Parlange et al. 2009), *AvrLm6* (Fudal et al. 2007) and *AvrLm11* (Balesdent et al. 2013). A recently documented C-G

SNP in the *AvrLm4-7* gene at base 358 (Van de Wouw and Howlett 2012) was found between Isolates 21 and 41 using the sequence tags generated in this study (Figure 6). In this study, Isolate 41 was found to contain the virulent C allele (Arginine amino-acid residue) while Isolate 21 has the avirulent G allele (Glycine amino-acid residue). An additional previously characterised SNP (Daverdin et al. 2012) was identified in the coding region of the *AvrLm4-7* gene at base 256 (Figure 6). The SNP lies in the first of two exons of the *AvrLm4-7* gene and changes the base at this locus from a negatively charged D (Aspartic acid) to a neutral N (Asparagine) in Isolate 41.

Sequence comparison of the genomic region surrounding the *AvrLm6* gene showed an absence of sequence tags for Isolate 41 in the coding region. Isolate 41 has previously been shown, to lack the functional *AvrLm6* gene while Isolate 21 possesses it (Raman et al. 2012). An alignment of the non-coding area around this gene has shown that sequence tags from both *L. maculans* Isolates are present in the close proximity of this gene. An alignment of the sequence tags of Isolates 41 and 21 showed that no tags matched the *AvrLm1* locus. The recently identified *AvrLm11* gene (Balesdent et al. 2013) was present in both isolates with 100% sequence identity to the reference genome. This has not been previously characterised in Australian isolates.

Housekeeping gene comparison

Sequence comparison within putative housekeeping genes, between Isolates 21 and 41 showed a high level of conservation in these Isolates compared to the reference. A number of these genes are seen in SuperContig 17 around position 400,000. Isolates 21 and 41 show 100% sequence identity to the reference sequence in the following putative (annotated as ‘similar to‘) genes: actin gene, Hsp70 nucleotide exchange factor fes1, monosaccharide transporter and major facilitator superfamily multidrug-resistance gene. Overall, SuperContig 17 has a lower level of SNP frequency than the neighbouring SuperContigs, indicating that perhaps a range of important conserved genes are located on this scaffold.

Discussion

Our results demonstrate that NGS technology and automated SNP prediction is a viable option for genotyping efforts in *L. maculans*. With new barcoding methods that allow the sequencing of multiple samples in a single lane, NGS can prove a cost-effective method of producing whole genome sequence data in small genomes such as the ~45.12 Mbp genome of *L. maculans*. We demonstrate how this data can be used for genome analysis.

SNP prediction and validation

Combined, the three sequence libraries provided genome coverage of 226x. This is many times the coverage necessary for high confidence SNP calling, with current research suggesting that 10 – 30x coverage for assembled genomes is sufficient (HGP 2013). High levels of coverage also allow for

confidence in SNP prediction, as SGSautoSNP will disregard loci where a difference in SNP call occurs in reads from the same Isolate. Furthermore, re-sequencing of genomes with high coverage can potentially provide information on large-scale recombination events, as well as copy number variation (CNV) and presence/absence variation (PAV) (Li et al. 2012).

The high frequency of SNPs between the two *L. maculans* Isolates is consistent with previous SNP identification studies in fungi; for example average SNP density was one SNP every 3,449 bp in *Fusarium graminearum*, a plant pathogenic fungus that causes head blight disease in wheat and barley (Cuomo et al. 2007). The –r 0 parameter in the SOAPaligner ensures that reads align only once and do not align equally well to multiple regions of the genome, thus minimising false SNP calls in repetitive regions. This provides confidence in the SNP prediction, further making sure that SNPs are only predicted in single-copy regions of the genome.

A higher number of SNPs was observed in upstream and downstream regions of predicted genes, rather than coding regions. We expect a higher level of conservation within coding regions (Castle 2011), as well as even higher conservation at splice sites and start/stop codons. This can also be seen in this data with only 0.11% of modifying SNPs in these regions. The transition to transversion ratio was shown to be similar to SNP prediction in *Candida albicans*, where out of 561 SNPs predicted, 437 were transitions and 126 were transversions (Forche et al. 2004). Repeat-induced point mutations (RIPs) may play a role in the high transition to transversion ratio. Previous studies identified regions of high RIP mutation levels in *L. maculans,* whereby G:C bases are mutated to A:T bases as a protective mechanism against the accumulation of repetitive elements in the fungal genome (Rouxel et al. 2011; Fudal et al. 2009). The greater number of transitions than transversions may also reflect the high frequency of C to T mutation following methylation (Coulondre et al. 1978).

SNP prediction by SGSAutoSNP was found to be correct even for the lowest SNP score of 2, whereby a minimum of two aligned sequence reads per Isolate (minimum of four from the two Isolates) aligned across the polymorphism. This can be useful in non-model organisms where sequence coverage may not be sufficient for genome assembly. On the other hand, high levels of coverage were also shown to provide accurate prediction, with a SNP score of 23 (SNP11) successfully predicted and validated.

The high SNP validation rate seen here is consistent with previous applications of SGSAutoSNP, whereby SNPs from canola and wheat Illumina whole genome sequence reads were predicted with an accuracy of over 96% and 93% respectively (Lorenc et al. 2012). We have now shown that SGSAutoSNP can be successfully implemented in predicting SNPs in the fungal genome of *L. maculans*. The genomes of these organisms vary greatly in size (*L. maculans* – 45.12 Mb, *Triticum aestivum* (Wheat) – 16,000 Mb) and in many orders of complexity, with the *L. maculans* genome being haploid and the wheat genome a hexaploid. However the method described in this research proves to be a robust method of predicting SNPs and thus creating novel molecular markers in small

fungal genomes. A previous study in the ascomycete plant pathogen *Ophiognomonia clavigignenti-juglandacearum* (~16 Mbp) identified only 298 SNPs within 8 Isolates using Roche's GS Mapper, of which only 45 could be found to be true polymorphisms (Broders et al. 2011).

The main advantage of the SGSautoSNP prediction method over existing SNP callers is that it does not consider the reference for SNP discovery (Lorenc et al. 2012). The reference is needed only to assemble the sequence reads, SNPs are called between the assembled reads. Based on the validation results, the SNP calling was successful in large contigs (> 4 Mbp) as well as small contigs, such as SuperContig 60 (6,199 bp). Alignments of small contigs, ESTs and other sequence data could be used in combination to predict SNPs using SGSautoSNP.

Genomic regions of significantly reduced SNP densities can indicate sequence stringency based on high functional conservation of coding or regulatory regions between otherwise diverse individuals. For example, such regions may represent blocks of house-keeping genes or chromosomal regulatory regions required for an organism's reproduction and/or survival (She et al. 2009). SuperContig 30 contains all of the assembled *L. maculans* mitochondrial DNA (Rouxel et al. 2011) thus the low diversity between the two sequenced Isolates, compared to the rest of the genome, could be explained given that mitochondrial DNA is generally highly conserved within a species (Kocher et al. 1989) and encodes vital metabolic genes. A more likely explanation however, is that more than one copy of mitochondrial DNA is present in one or both of the sequenced *L. maculans* Isolates. This is known as heteroplasmy and can be seen in humans (Payne et al. 2012), plants (Tian et al. 2006), animals (Kmiec et al. 2006), as well as fungi (Lesemann et al. 2006). This may explain why the SGSautoSNP software does not call SNPs in these regions, as its stringency does not allow for differences between reads of the same Isolate. This generally works to eliminate false SNP calls due to sequence error, however it may in this case lead to the masking of SNPs between Isolates due to the presence of SNPs within an Isolate.

Population structure and avirulence gene evolution

Interestingly during SNP validation, Isolate 10 was unique in displaying a third allele at the SNP6 locus, making this SNP a tri-allelic polymorphism, thought to be a relatively rare occurrence. Recent SNP research conducted in the human genome has found that tri-allelic SNPs within populations are present more than twice as often as expected (Casci 2010). This Isolate also displays unique polymorphisms for two other SNPs, in an otherwise synchronised polymorphic set. Such recurring differences indicate that Isolate 10 possesses a number of Isolate specific alleles.

Few relationships between isolates and collection parameters could be observed. The isolates from Western Australia tended to group together in one clade of the phylogenetic tree. This is consistent with the findings of Hayden et al. (2007) who also observed geographical separation between western and eastern Australian *L. maculans* populations, using microsatellite and minisatellite markers. In addition the isolates collected from B. juncea grouped together. Relatively low sample sizes in this

study may preclude any definitive interpretation of the relationships between specific populations. In all, the predicted SNPs were highly conserved and polymorphic in the different population Isolates that were included in the analysis. These isolates are from geographically distant locations (up to 3,000 km, from WA to Victoria) and isolated from different growing seasons (from 1987 to 2009). This indicates that these SNPs could be useful in tracking the spatio-temporal spread of loci, and potentially linked traits, in the Australian population.

These initial data suggest that SNPs (RIP induced or non-RIP induced), PAVs, or partial deletions within genes may all play a role in the virulence and avirulence phenotypes seen in previous studies. Looking at these changes using re-sequencing data can be a method to identify previously uncharacterised *AvrLm* genes in *L. maculans.* Identifying regions of the genome that are present in one Isolate yet not in the other may help to determine those areas where the stringency of the alignment software cannot allow alignment of sequence reads due to a high number of polymorphisms, such as in the *AvrLm6* gene region.

Smaller variations, such as SNPs can be rapidly analysed and compared in the re-sequenced individuals by simply mapping the sequence reads to the reference genome. The use of the SGSautoSNP software in this study has identified previously characterised SNPs in *AvrLm4-7*, further highlighting the usefulness of this approach. This can prove to be a simple method of genotyping individuals in important loci of the genome, given that whole-genome sequencing is becoming increasingly accessible and affordable. An interesting observation is the differing SNP frequency within the immediate location of the *AvrLm* genes. The *AvrLm1, AvrLm6* and *AvrLm11* genes exhibit a relatively low SNP frequency within their genomic environment, whereas *AvrLm4-7* resides in a region of higher SNP frequency. This may be due to the high selective pressure from the corresponding plant resistance gene acting at the *AvrLm4-7* locus, as both isolates were collected from *B. napus* cultivars containing the *Rlm4* gene. Prior to 2002 53% cultivars grown in Australia contained *Rlm4*, decreasing to the current 29%. Furthermore, *Rlm4* is the most common resistance gene in Australian cultivars (Marcroft et al. 2012). This will have led to strong selection pressure towards the *AvrLm4* locus and therefore higher SNP frequency. The mechanism of *AvrLm* adaptation to resistance gene pressure may play a role in sequence divergence around these genes.

Two previously reported SNPs in the *AvrLm4-7* gene (Daverdin et al. 2012, Van de Wouw and Howlett 2012) was found to be polymorphic between Isolates 21 and 41 using the sequence tags generated in this study. One SNP is known to cause a change from virulence to avirulence against the resistance genes *RLM4* in *B. napus* (Van de Wouw and Howlett 2012; Parlange et al. 2009). Consistent with this, the two Isolates differ in their infection success in *B. napus* plants containing the *Rlm4* resistance locus, with Isolate 41 being virulent and Isolate 21 being avirulent (Raman et al. 2012). The polymorphisms lead to a difference in effector/avirulence protein structure and/or function in the *Brassica* host. It may be possible that the two SNPs (bases 256 and 358) show some association and it is likely that they both play a role in the documented virulence/avirulence phenotypes.

The lack of reads mapping in the *AvrLm6* gene may be due to a deletion or substantial polymorphism in the gene compared to the reference genome preventing the reads from mapping. Substantial polymorphism has previously been shown for Isolate 41 in the *AvrLm6* gene region (Van de Wouw et al. 2010); with 43 documented RIP like mutations, compared to the reference. As no tags were present for the *AvrLm1* locus this indicates that this allele is most likely not present in these Isolates and supports previous observations, using a differential set of *B. napus* lines, that the Isolates 21 and 41 do not possess the *AvrLm1* gene (Harsh Raman, pers. Comm.). The recently identified *AvrLm11* gene (Balesdent et al. 2013) was present in both Isolates with 100% sequence identity to the reference genome. *Rlm11* has not been commercially released in Australian canola cultivars and therefore the gene is unlikely to be under selection pressure in Australian populations.

Conclusions

This study demonstrates the applicability of whole-genome re-sequencing in small fungal genomes such as *L. maculans*. It allows for SNP prediction and validation, and the use of these novel markers as a tool for molecular and phylogenetic analysis. Re-sequencing can also be used to validate known mutations, as well as identify putative mutations with modifying effects on the genome. Furthermore, combining SGSautoSNP and SnpEff, we demonstrate a useful pipeline in determining whole-genome polymorphic trends, such as the transition to transversion ratios, polymorphism directionality and the SNP density across chromosomes or contigs.

Acknowledgements

The authors would like to acknowledge funding support from the Australian Research Council (Projects LP0883462, LP0989200, LP110100200 and DP0985953) and the Grains and Research Development Council (Project DAN00117). Support from the Australian Genome Research Facility (AGRF), the Queensland Cyber Infrastructure Foundation (QCIF) and the Australian Partnership for Advanced Computing (APAC) is gratefully acknowledged.

References

AOF (2012) Australian Oilseeds Federation.

Appleby N, Edwards D, Batley J (2009) New technologies for ultra-high throughput genotyping in plants. In: Somers DJ, Langridge P, Gustafson JP (eds) Plant Genomics Methods and Protocols. Humana Press, UK, pp 19-40

Balesdent MH, Fudal I, Ollivier B, Bally P, Grandaubert J, Eber F, Chevre AM, Leflon M, Rouxel T (2013) The dispensable chromosome of Leptosphaeria maculans shelters an effector gene conferring avirulence towards Brassica rapa. New Phytol 198 (3):887-898. doi:10.1111/nph.12178

Balesedent MH, Attard A, Ansan-Melayah D, Delourme R, Renard M, Rouxel T (2001) Genetic control and host range of avirulence toward Brassica napus cultivars Quinta and Jet Neuf in Leptosphaeria maculans. Phytopathology 91:70-76

Barrins JM, Ades PK, Salisbury PA, Howlett BJ (2004) Genetic diversity of Australian Isolates of Leptosphaeria maculans, the fungus that causes blackleg of canola (Brassica napus). Australas Plant Path 33 (4):529-536. doi:Doi 10.1071/Ap04061

Batley J, Edwards D (2007) SNP applications in plants. In: Oraguzie NC, Rikkerink EHA, Gardiner SE, De Silva HN (eds) Association Mapping in Plants Springer New York, pp 95-102

Batley J, Edwards D (2009) Genome sequence data: management, storage, and visualization. Biotechniques 46 (5):333-+. doi:Doi 10.2144/000113134

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a Genetic-Linkage Map in Man Using Restriction Fragment Length Polymorphisms. Am J Hum Genet 32 (3):314-331

Boyle JS, Lew AM (1995) An Inexpensive Alternative to Glassmilk for DNA Purification. Trends Genet 11 (1):8-8

Broders KD, Woeste KE, SanMiguel PJ, Westerman RP, Boland GJ (2011) Discovery of single-nucleotide polymorphisms (SNPs) in the uncharacterized genome of the ascomycete *Ophiognomonia clavigignenti-juglandacearum* from 454 sequence data. Molecular Ecology Resources 11:693-702. doi:doi: 10.1111/j.1755-0998.2011.02998.x

Casci T (2010) Population Genetics Snps That Come in Threes. Nat Rev Genet 11 (1). doi:Doi 10.1038/Nrg2725

Castle JC (2011) SNPs Occur in Regions with Less Genomic Sequence Conservation. Plos One 6 (6). doi:ARTN e20660 DOI 10.1371/journal.pone.0020660

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu XY, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. Fly 6 (2):80-92. doi:Doi 10.4161/Fly.19695

Coulondre C, Miller JH, Farabaugh PJ, Gilbert W (1978) Molecular basis of base substitution hot spots in Escherichia coli. Nature 274:775-780

Cuomo CA, Gueldener U, Xu JR, Trail F, Turgeon BG, Di Pietro A, Walton JD, Ma LJ, Baker SE, Rep M, Adam G, Antoniw J, Baldwin T, Calvo S, Chang YL, DeCaprio D, Gale LR, Gnerre S, Goswami RS, Hammond-Kosack K, Harris LJ, Hilburn K, Kennell JC, Kroken S, Magnuson JK, Mannhaupt G, Mauceli E, Mewes HW, Mitterbauer R, Muehlbauer G, Munsterkotter M, Nelson D, O'Donnell K, Ouellet T, Qi WH, Quesneville H, Roncero MIG, Seong KY, Tetko IV, Urban M, Waalwijk C, Ward TJ, Yao JQ, Birren BW, Kistler HC (2007) The Fusarium graminearum genome reveals a link between localized polymorphism and pathogen specialization. Science 317 (5843):1400-1402. doi:DOI 10.1126/science.1143708

Daverdin G, Rouxel T, Gout L, Aubertot JN, Fudal I, Meyer M, Parlange F, Carpezat J, Balesdent MH (2012) Genome Structure and Reproductive Behaviour Influence the Evolutionary Potential of a Fungal Phytopathogen. Plos Pathog 8 (11). doi:ARTN e1003020 DOI 10.1371/journal.ppat.1003020

Drummond AJ, Ashton B, Buxton S, Cheung M, Cooper A, Duran C, Field M, Heled J, Kearse M, Markowitz S, Moir R, Stones-Havas S, Sturrock S, Thierer T, Wilson A (2012) Geneious Pro v5.6. http://wwwgeneiouscom

Duran C, Appleby N, Edwards D, Batley J (2009) Molecular genetic markers: discovery, applications, data storage and visualisation. Current Bioinformatics 4:16-27

Duran C, Eales D, Marshall D, Imelfort M, Stiller J, Berkman PJ, Clark T, McKenzie M, Appleby N, Batley J, Basford K, Edwards D (2010) Future tools for association mapping in crop plants. Genome 53 (11):1017-1023. doi:Doi 10.1139/G10-057

Edwards D, Batley J, Snowdon RJ (2013) Accessing complex crop genomes with next-generation sequencing. Theor Appl Genet 126 (1):1-11. doi:DOI 10.1007/s00122-012-1964-x

Edwards D, Zander M, Dalton-Morgan J, Batley J (2012) New technologies for ultra-high throughput genotyping in plant taxonomy. In: Besse P (ed) Molecular Plant Taxonomy.

Forche A, Magee PT, Magee BB, May G (2004) Genome-wide single-nucleotide polymorphism map for Candida albicans. Eukaryot Cell 3 (3):705-714. doi:Doi 10.1128/Ec.3.3.705-714.2004

Fudal I, Ross S, Brun H, Besnard A-L, Ermel M, Kuhn M-L, Balesdent M-H, Rouxel T (2009) Repeat-Induced Point Mutation (RIP) as an Alternative Mechanism of Evolution Toward Virulence in Leptosphaeria maculans. Molecular Plant-Microbe Interactions 22 (8):932-941. doi:10.1094/mpmi-22-8-0932

Fudal I, Ross S, Gout L, Blaise F, Kuhn ML, Eckert MR, Cattolico L, Bernard-Samain S, Balesdent MH, Rouxel T (2007) Heterochromatin-like regions as ecological niches for avirulence genes in the Leptosphaeria maculans genome: Map-based cloning of AvrLm6. Molecular Plant-Microbe Interactions 20 (4):459-470. doi:Doi 10.1094/Mpmi-20-4-0459

Fulton TM, Chunwongse J, Tanksley SD (1995) Microprep Protocol for Extraction of DNA from Tomato and Other Herbaceous Plants. Plant Mol Biol Rep 13 (3):207-209

Gout L, Fudal I, Kuhn ML, Blaise F, Eckert M, Cattolico L, Balesdent MH, Rouxel T (2006) Lost in the middle of nowhere: the AvrLm1 avirulence gene of the Dothideomycete Leptosphaeria maculans. Mol Microbiol 60 (1):67-80. doi:DOI 10.1111/j.1365-2958.2006.05076.x

Hane JK, Lowe RGT, Solomon PS, Tan KC, Schoch CL, Spatafora JW, Crous PW, Kodira C, Birren BW, Galagan JE, Torriani SFF, McDonald BA, Oliver RP (2007) Dothideomycete-plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen Stagonospora nodorum. Plant Cell 19 (11):3347-3368. doi:DOI 10.1105/tpc.107.052829

Hayden HL, Cozijnsen AJ, Howlett BJ (2007) Microsatellite and minisatellite analysis of Leptosphaeria maculans in Australia reveals regional genetic differentiation. Phytopathology 97 (7):879-887. doi:Doi 10.1094/Phyto-97-7-0879

Hayward A, Dalton-Morgan J, Mason A, Zander M, Edwards D, Batley J (2012a) SNP discovery and applications in Brassica napus. Journal of Plant Biotechnology 39:1-12

Hayward A, McLanders J, Campbell E, Edwards D, Batley J (2012b) Genomic advances will herald new insights into the Brassica: Leptosphaeria maculans pathosystem. Plant Biology 14:1-10. doi:DOI 10.1111/j.1438-8677.2011.00481.x

Hayward A, Vighnesh G, Delay C, Samian MR, Manoli S, Stiller J, McKenzie M, Edwards D, Batley J (2012c) Second-generation sequencing for gene discovery in the Brassicaceae. Plant Biotechnol J 10 (6):750-759. doi:DOI 10.1111/j.1467-7652.2012.00719.x

HGP (2013) Human Genome Project http://www.ornl.gov/sci/techresources/Human_Genome/faq/seqfacts.shtml#how. Accessed 30 Jan 2013

Howlett BJ, Idnurm A, Pedras MSC (2001) Leptosphaeria maculans, the causal agent of blackleg disease of Brassicas. Fungal Genet Biol 33 (1):1-14. doi:DOI 10.1006/fgbi.2001.1274

Idnurm A, Howlett BJ (2003) Analysis of loss of pathogenicity mutants reveals that repeat-induced point mutations can occur in the Dothideomycete Leptosphaeria maculans. Fungal Genet Biol 39 (1):31-37. doi:Doi 10.1016/S1087-1845(02)00588-1

Illumina Incorporated (2013) www.illumina.com. Accessed 30 Jan 2013

Imelfort M, Duran C, Batley J, Edwards D (2009) Discovering genetic polymorphisms in next-generation sequencing data. Plant Biotechnol J 7 (4):312-317. doi:DOI 10.1111/j.1467-7652.2009.00406.x

Kaur S, Cogan NOI, Ye G, Baillie RC, Hand ML, Ling AE, Mcgearey AK, Kaur J, Hopkins CJ, Todorovic M, Mountford H, Edwards D, Batley J, Burton W, Salisbury P, Gororo N, Marcroft S, Kearney G, Smith KF, Forster JW, Spangenberg GC (2009) Genetic map construction and QTL mapping of resistance to blackleg (Leptosphaeria maculans) disease in Australian canola (Brassica napus L.) cultivars. Theor Appl Genet 120 (1):71-83. doi:DOI 10.1007/s00122-009-1160-9

Kmiec B, Woloszynska M, Janska H (2006) Heteroplasmy as a common state of mitochondrial genetic information in plants and animals. Curr Genet 50 (3):149-159. doi:DOI 10.1007/s00294-006-0082-1

Kocher TD, Thomas WK, Meyer A, Edwards SV, Paabo S, Villablanca FX, Wilson AC (1989) Dynamics of Mitochondrial-DNA Evolution in Animals - Amplification and Sequencing with Conserved Primers. P Natl Acad Sci USA 86 (16):6196-6200. doi:DOI 10.1073/pnas.86.16.6196

Lai KT, Duran C, Berkman PJ, Lorenc MT, Stiller J, Manoli S, Hayden MJ, Forrest KL, Fleury D, Baumann U, Zander M, Mason AS, Batley J, Edwards D (2012) Single nucleotide polymorphism discovery from wheat next-generation sequence data. Plant Biotechnol J 10 (6):743-749. doi:DOI 10.1111/j.1467-7652.2012.00718.x

Lee HC, Lai KT, Lorenc MT, Imelfort M, Duran C, Edwards D (2012) Bioinformatics tools and databases for analysis of next-generation sequence data. Brief Funct Genomics 11 (1):12-24. doi:Doi 10.1093/Bfgp/Elr037

Lesemann SS, Schimpke S, Dunemann F, Deising HB (2006) Mitochondrial heteroplasmy for the cytochrome b gene controls the level of strobilurin resistance in the apple powdery mildew fungus Podosphaera leucotricha (Ell. & Ev.) ES Salmon. J Plant Dis Protect 113 (6):259-266

Li J, Lupat R, Amarasinghe KC, Thompson ER, Doyle MA, Ryland GL, Tothill RW, Halgamuge SK, Campbell IG, Gorringe KL (2012) CONTRA: copy number analysis for targeted resequencing. Bioinformatics 28 (10):1307-1313. doi:DOI 10.1093/bioinformatics/bts146

Li RQ, Yu C, Li YR, Lam TW, Yiu SM, Kristiansen K, Wang J (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25 (15):1966-1967. doi:DOI 10.1093/bioinformatics/btp336

Lorenc MT, Hayashi S, Stiller J, Hong L, Manoli S, Ruperao P, Visendi P, Berkman PJ, Lai K, Batley J, Edwards D (2012) Discovery of Single Nucleotide Polymorphisms in Complex Genomes Using SGSautoSNP. Biology 1 (2):370-382. doi:10.3390/biology1020370

Marcroft S, Bluett C (2008) Blackleg of canola. Agriculture notes, Department of Primary Industries AG1352

Marcroft SJ, Elliott VL, Cozijnsen AJ, Salisbury PA, Howlett BJ, Van de Wouw AP (2012) Identifying resistance genes to *Leptosphaeria maculans* in Australian *Brassica napus* cultivars based on reactions to isolates with known avirulence genotypes. Crop and Pasture Science 63: 338-350

Marshall DJ, Hayward A, Eales D, Imelfort M, Stiller J, Berkman PJ, Clark T, McKenzie M, Lai KT, Duran C, Batley J, Edwards D (2010) Targeted identification of genomic regions using TAGdb. Plant Methods 6. doi:Artn 19

Parlange F, Daverdin G, Fudal I, Kuhn ML, Balesdent MH, Blaise F, Grezes-Besset B, Rouxel T (2009) 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. Mol Microbiol 71 (4):851-863. doi:DOI 10.1111/j.1365-2958.2008.06547.x

Payne BAI, Wilson IJ, Yu-Wai-Man P, Coxhead J, Deehan D, Horvath R, Taylor RW, Samuels DC, Santibanez-Koref M, Chinnery PF (2012) Universal heteroplasmy of human mitochondrial DNA. Hum Mol Genet 22 (2):384-390. doi:Doi 10.1093/Hmg/Dds435

Raman R, Taylor B, Marcroft S, Stiller J, Eckermann P, Coombes N, Rehman A, Lindbeck K, Luckett D, Wratten N, Batley J, Edwards D, Wang XW, Raman H (2012) Molecular mapping of qualitative and quantitative loci for resistance to Leptosphaeria maculans causing blackleg disease in canola (Brassica napus L.). Theor Appl Genet 125 (2):405-418. doi:DOI 10.1007/s00122-012-1842-6

Rouxel T, Grandaubert J, Hane JK, Hoede C, van de Wouw AP, Couloux A, Dominguez V, Anthouard V, Bally P, Bourras S, Cozijnsen AJ, Ciuffetti LM, Degrave A, Dilmaghani A, Duret L, Fudal I, Goodwin SB, Gout L, Glaser N, Linglin J, Kema GHJ, Lapalu N, Lawrence CB, May K, Meyer M, Ollivier B, Poulain J, Schoch CL, Simon A, Spatafora JW, Stachowiak A, Turgeon BG, Tyler BM, Vincent D, Weissenbach J, Amselem J, Quesneville H, Oliver RP, Wincker P, Balesdent M-H, Howlett BJ (2011) Effector diversification within compartments of the Leptosphaeria maculans genome affected by Repeat-Induced Point mutations. Nature Communications 2. doi:10.1038/ncomms1189

She XW, Rohl CA, Castle JC, Kulkarni AV, Johnson JM, Chen RH (2009) Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. Bmc Genomics 10. doi:Artn 269 Doi 10.1186/1471-2164-10-269

Tian XJ, Zheng J, Hu SN, Yu J (2006) The rice mitochondrial genomes and their variations. Plant Physiol 140 (2):401-410. doi:DOI 10.1104/pp.105.070060

Tollenaere R, Hayward A, Dalton-Morgan J, Campbell E, Lee JRM, Lorenc MT, Manoli S, Stiller J, Raman R, Raman H, Edwards D, Batley J (2012) Identification and characterization of candidate Rlm4 blackleg resistance genes in Brassica napus using next-generation sequencing. Plant Biotechnol J 10 (6):709-715. doi:DOI 10.1111/j.1467-7652.2012.00716.x

Van de Wouw AP, Cozijnsen AJ, Hane JK, Brunner PC, McDonald BA, Oliver RP, Howlett BJ (2010) Evolution of Linked Avirulence Effectors in Leptosphaeria maculans Is Affected by Genomic Environment and Exposure to Resistance Genes in Host Plants. Plos Pathog 6 (11). doi:ARTN e1001180 DOI 10.1371/journal.ppat.1001180

Van de Wouw AP, Howlett BJ (2012) Estimating frequencies of virulent Isolates in field populations of a plant pathogenic fungus, Leptosphaeria maculans, using high-throughput pyrosequencing. J Appl Microbiol 113 (5):1145-1153. doi:DOI 10.1111/j.1365-2672.2012.05413.x

Tables

Table 1 Information about all Isolates used in the study

Isolate Number	Reference number	Year Cultured	Brassica species Isolate cultured from	Cultivar/line Isolate cultured from	Resistance genes in cultivar[#]	Stubble collection site	State	Reference
Reference	v23.1.3.	Mid-1990	N/A*	N/A	N/A	Europe	N/A	(Balesedent et al. 2001)
21'	04MGPP021	2004	*B. napus*	AG-Emblem	*Rlm4, Rlm9*	Bordertown	South Australia	
41'	06MGPP041	2006	*B. napus*	Skipton	*Rlm4, Rlm9*	Lake Bolac	Victoria	
1	04MGP-P029	2004	*B. napus*	TI1 Pinnacle	*Rlm3, Rlm9*	Geelong	Victoria	
2	07VTJH020	2007	*B. juncea*	JC05007	Unknown	Horsham	Victoria	
3	09SMJ087	2009	*B. juncea*	EXCEED™ OasisCL	Unknown	Kaniva	Victoria	
4	05MGPP033	2005	*B. napus*	Skipton	*Rlm4, Rlm9*	Yeelanna	South Australia	(Van de Wouw et al. 2010)
5	89/C16	1988	*B. napus*	Unknown	N/A	Millicent	South Australia	
6	04MGP-S006	2004	*B. napus*	Surpass400	*Rlm1, RlmS*	Eyre Peninsula	South Australia	
7	10SMJ041	2009	*B. juncea*	EXCEED™ OasisCL	Unknown	Tamworth	New South Wales	
8	06MGPP019	2006	*B. napus*	ATR-Beacon	*Rlm3, Rlm4*	Wagga Wagga	New South Wales	
9	LM300	2002	*B. napus*	TI1 Pinnacle	*Rlm3, Rlm9*	Mt Barker	Western Australia	
10	LM592	2003	*B. napus*	TI1 Pinnacle	*Rlm3, Rlm9*	Mt Barker	Western Australia	
11	06S012	2006	*B. napus*	ATR-Beacon	*Rlm3, Rlm4*	Bordertown	South Australia	
12	06J085	2006	*B. napus*	Unknown	N/A	Horsham	Victoria	
13	05P032	2005	*B. napus*	Skipton	*Rlm4, Rlm9*	Yeelanna	South Australia	(Van de Wouw et al. 2010)
14	IBCN13 (D1)	1991	*B. napus*	Unknown	N/A	Mt Barker	Western Australia	(Purwantara et al. 2000)
15	IBCN15 (D2)	1988	*B. napus*	Unknown	N/A	Streatham	Victoria	(Purwantara et al. 2000)
16	IBCN16 (D3)	1988	*B. napus*	Unknown	N/A	Mt Barker	Western Australia	(Purwantara et al. 2000)
17	IBCN17 (D4)	1988	*B. napus*	Unknown	N/A	Millicent	South Australia	(Purwantara et al. 2000)
18	IBCN18 (D5)	1988	*B. napus*	Unknown	N/A	Penshurst	Victoria	(Purwantara et al. 2000)
19	IBCN75 (D6)	1987	*B. napus*	Unknown	N/A	Mt Barker	Western Australia	(Purwantara et al. 2000)
20	IBCN76 (D7)	1987	*B. napus*	Unknown	N/A	Mt Barker	Western Australia	(Purwantara et al. 2000)
21	D8	2005	*B. napus*	Surpass501TT	*Rlm1, RlmS*	Mt Barker	Western Australia	(Marcroft et al. 2012)
22	D9	2005	*B. napus*	ATR-Beacon	*Rlm3, Rlm4*	Mt Barker	Western Australia	(Marcroft et al. 2012)
23	PHW1223 (D10)	1987	*B. napus*	Unknown	N/A	Mt Barker	Western Australia	(Purwantara et al. 2000)

* Isolate v23.1.3 is the result of a series of *in vitro* crosses between European field Isolates (Balesedent et al. 2001). Not all data on these Isolates was available, "-" denotes an unknown variable. IBCN numbers represent the IDs of "International Blackleg Collection Network" Isolates

[#] Resistance genes as determined by Marcroft et al. (2012)

Table 2 Predicted SNPs chosen for validation

SNP #	Super Contig	Position	SNP Score	Base Change	Forward Primer (5' to 3')	Reverse Primer (5' to 3')
1	0	369650	15	**A/G**	CATACTGCAGATTGAGGATGC	GCTACGTTTGCCCTCATTC
2	20	433691	8	**C/T**	ACGCCCTCGGTTAGCTC	ATCGGTCTCGGAGTGGTC
3	25	259169	5	**A/T**	GAGGCTTGAAATAGTGAGGTACG	CATCCATCACAGCGCAG
4	30	115	6	**T/G**	GGAGCTAGTAAGCGATGTCG	AGGTAAACCATAGGTTACCATACCC
5	5	996624	4	**A/G**	CGTCAACTCTACCGACTCG	ACGCTGCCAGCTTGG
6	60	506	2	**A/T**	GCTCTTGGGAACCTATCTTCC	CGACGCTGACGTGCC
7	9	1524477	8	**C/T**	CGACTCTGCAAGACACCG	CCTGATGGGACTTGAGCC
8	9	789721	11	**A/G**	CTTGTATCGAGATAGGTGGACTG	GGTGGATTACTCTACTTGGAGGC
9	2	838930	13	**A/C**	AGGGATTGTACACTGTGATTACCAG	GATGAGCGTGGACTCGG
10	4	1281059	15	**C/T**	CATACACTGCAACTTGATGGATG	TGAATGTTCTTGCGTTGCC
11	4	1362010	23	**G/T**	GCCGAGCACAGAAGAGC	CGTTGAGGTTAGAGGTGGAGTG
12	8	169231	14	**C/T**	CCCGAGCCACTACCACTAC	GCGATATAGGAGAACATGTGCC
13	2	889136	13	**A/G**	CTAGAGAACATGGACGGTACTCTAG	ACACACTCTTAGTCTCACGCAC
14	8	1296096	14	**A/G**	CTTGGACAATAGAGACATTCAGTG	GGGGCGGCTGGATTG
15	12	352844	10	**A/G**	GGGCAAGAGAGCGAATG	GCACAAACACCGTAGGCC
16	12	838085	18	**C/G**	TCGCCCTGTATAGCTATTTCG	GCATCACTGTATACCGCCG
17	12	1107690	5	**C/G**	GCGGGTTGAGGTTGCC	CGCACGGGTGAAGGAAG
18	14	1268037	11	**A/G**	CGGCGCGTCAAAGAC	CCTCACCGCTGTTGGC
19	14	852465	6	**G/T**	CGTCGTCGGCTACCACAG	GCACATGCAATCGCAAGAG
20	3	562718	5	**C/T**	CGTCTCACGACCCTTCATG	CTCGTCCACCTCTACAACAACG

Table 3 SNP details for all SNPs detected

SNP type	Base change	SNP Count	%
Transition	**A > G**	8905	40.82
Transition	**C > T**	8758	40.15
Transversion	**A> C**	1088	4.99
Transversion	**A > T**	1021	4.68
Transversion	**C > G**	1038	4.76
Transversion	**G > T**	1004	4.6

SNP Count - Number of each SNP type predicted; % - percent value overall

Table 4 Properties of the validated SNPs

SNP number	SuperContig	Position	Genic location	Base changes	PIC	SNP Score
SNP1	0	369650	CDS - Syn	**G >A**	0.4352	15
SNP2	20	433691	CDS - Syn	**C > T**	0.1472	8
SNP3	25	259169	Intergenic	**A > T**	0.4862	5
SNP5	5	996624	CDS - Nonsyn	**G > A**	0.1472	4
SNP6	60	506	Intergenic	**A > T/G**	0.4966	2
SNP7	9	1524477	Genic - Syn	**C > T**	0.3085	8
SNP8	9	789721	Intergenic	**A > G**	0.0671	11
SNP9	2	838930	Intergenic	**A > C**	0.3299	13
SNP11	4	1362010	Intergenic	**G > T**	0.4898	23
SNP14	8	1296096	Intergenic	**A > G**	0.4959	14
SNP15	12	352844	CDS - Nonsyn	**G > A**	0.4445	10
SNP16	12	838085	Intergenic	**G > C**	0.2188	18
SNP17	12	1107690	Intergenic	**C > G**	0.4133	5
SNP18	14	1268037	CDS – Nonsyn	**G > A**	0.0643	11

Position - nucleotide position on the supercontig; CDS - coding region of a gene; PIC - Polymorphism Information Content (Botstein et al. 1980); Syn – synonymous change; Nonsyn – non-synonymous change)

Table 5 Modifying SNP effects

Type (alphabetical order)	Count	Percent
Downstream	22,583	38.65
Intergenic	10,234	17.51
Intron	825	1.41
Non Synonymous Coding	1,782	3.05
Splice Site Acceptor	5	0.01
Splice Site Donor	5	0.01
Start Lost	8	0.01
Stop Gained	39	0.07
Stop Lost	8	0.01
Synonymous Coding	1,743	2.98
Synonymous Stop	9	0.02
Upstream	21,193	36.27

Figures

Fig. 1 Dendrogram showing the relationship between 24 Australian isolates as well as the reference (Ref) *Leptosphaeria maculans* isolate, v23.1.3. Highlighted in red are two isolates cultured from *Brassica juncea* rather than *B. napus*. Highlighted in blue are the isolates that were cultured from sites in WA.

Fig. 2 PCA showing relationship between blackleg Isolates used in this study. Highlighted in red are two isolates cultured from *Brassica juncea* rather than *B. napus*. Highlighted in blue are the isolates that were cultured from sites in WA. Ref refers to the reference isolate, v23.1.3..

Fig. 3 SNP density within the *L. maculans* genome. SuperContigs >20,000 bp are included. Change rate denotes the frequency of SNP changes per bp (e.g. 1 change every 3,000 bp)

Fig. 4 SNP distribution variation within the *L. maculans* genome. Percentage (%) of SNPs within different genomic regions

Fig. 5 SNP frequency (Changes/10Kb) plotted against SuperContig position. Position of *AvrLm* genes is shown

Fig. 6 Alignment of AvrLm4-7 genes from the reference, Isolate 21 and Isolate 41. The two SNPs in Isolate 41 are shown

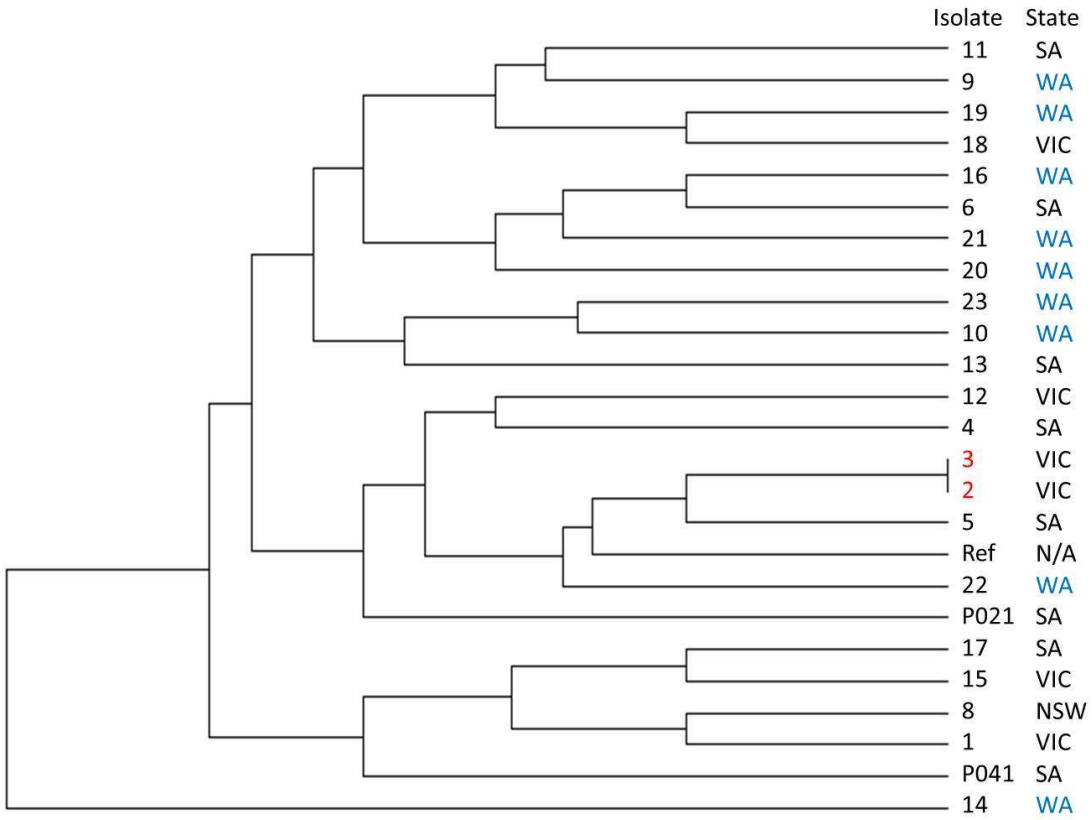

Isolate
State
11 SA
9 WA
19 WA
18 VIC
16 WA
6 SA
21 WA
20 WA
23 WA
10 WA
13 SA
12 VIC
4 SA
3 VIC
2 VIC
5 SA
Ref N/A
22 WA
P021 SA
17 SA
15 VIC
8 NSW
1 VIC
P041 SA
14 WA

Figure 1

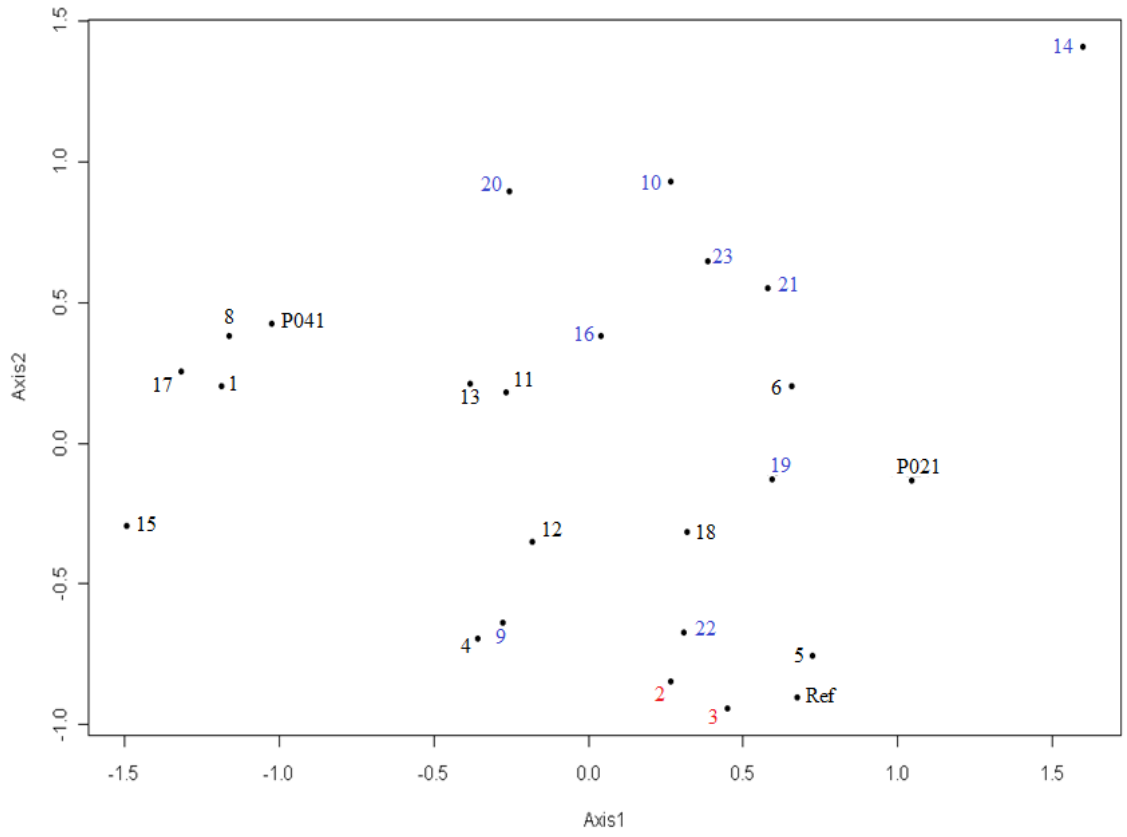
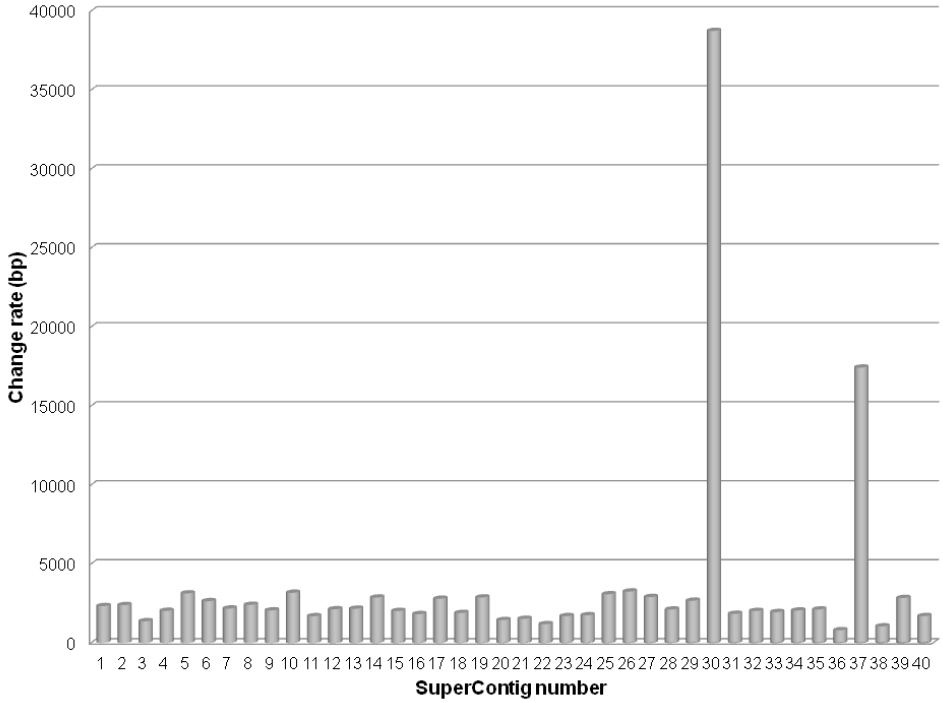
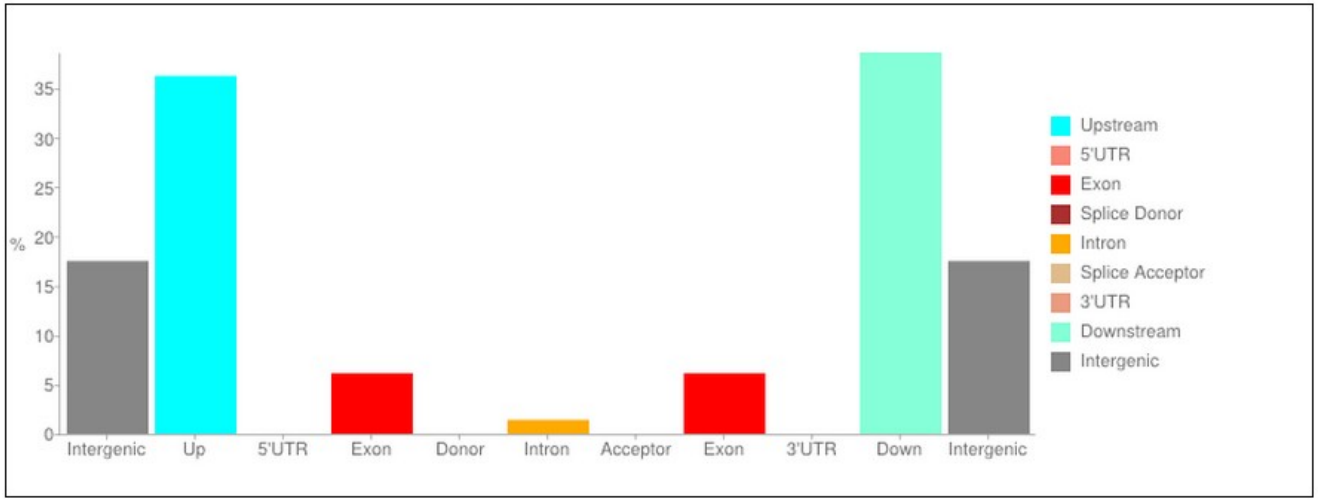
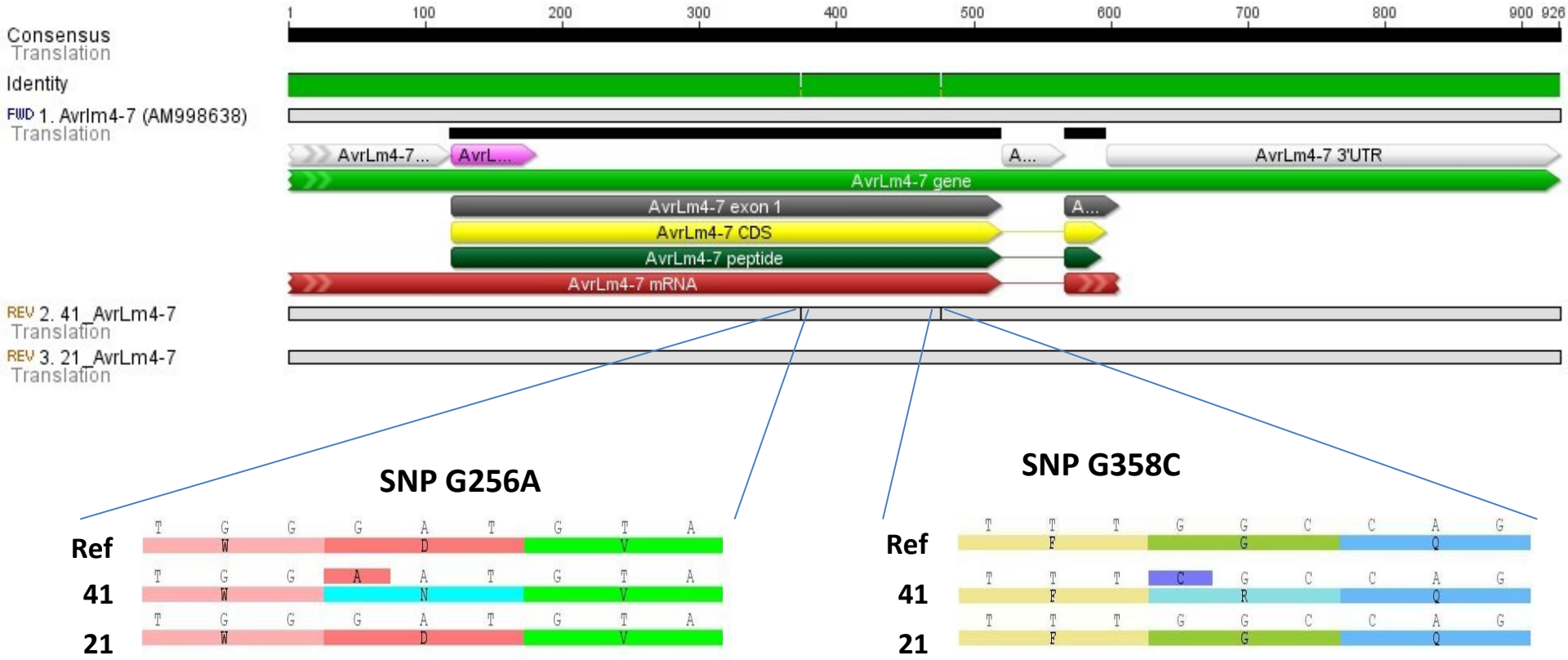

14
20
10
23
21
8
P041
16
17
1
13
11
6
19
P021
15
12
18
4
9
22
5
2
Ref
3
Axis2
Axis1
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

Figure 2

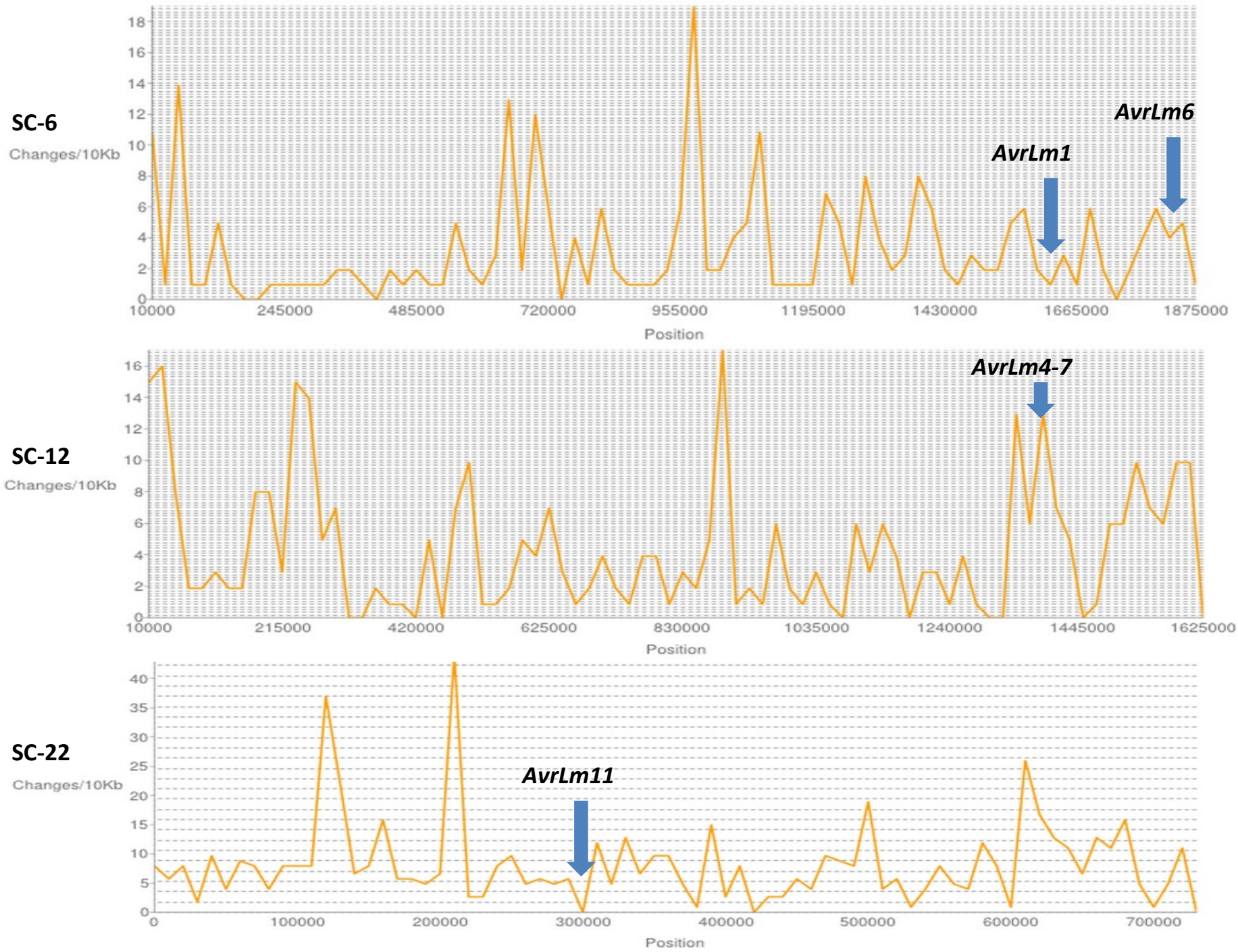

40000
35000
30000
25000
20000
15000
10000
5000
0
Change rate (bp)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
SuperContig number


35
30
25
%
20
15
10
5
0
Intergenic
Up
5'UTR
Exon
Donor
Intron
Acceptor
Exon
3'UTR
Down
Intergenic
Upstream
5'UTR
Exon
Splice Donor
Intron
Splice Acceptor
3'UTR
Downstream
Intergenic



SNP G256A

SNP G358C


SC-6
Changes/10Kb
AvrLm6
AvrLm1
Position
SC-12
Changes/10Kb
AvrLm4-7
Position
SC-22
Changes/10Kb
AvrLm11
Position

Author/s:
Zander, M;Patel, DA;Van de Wouw, A;Lai, K;Lorenc, MT;Campbell, E;Hayward, A;Edwards, D;Raman, H;Batley, J

Title:
Identifying genetic diversity of avirulence genes in Leptosphaeria maculans using whole genome sequencing

Date:
2013-08-01

Citation:
Zander, M., Patel, D. A., Van de Wouw, A., Lai, K., Lorenc, M. T., Campbell, E., Hayward, A., Edwards, D., Raman, H. & Batley, J. (2013). Identifying genetic diversity of avirulence genes in Leptosphaeria maculans using whole genome sequencing. FUNCTIONAL & INTEGRATIVE GENOMICS, 13 (3), pp.294-308. https://doi.org/10.1007/s10142-013-0324-5.