Gene loss in the fungal canola pathogen *Leptosphaeria maculans*

Authors:

Agnieszka A. Golicz[1,2]

Paula A. Martinez[1,2]

Manuel Zander[1,2]

Dhwani A. Patel[1,2]

Angela P Van De Wouw[3]

Paul Visendi[1,2]

Timothy L. Fitzgerald[4]

David Edwards[1,2,5]

Jacqueline Batley[1,5,6]

(1) School of Agriculture and Food Sciences, University of Queensland, Brisbane, Queensland, 4072, Australia

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

(3) School of Botany, University of Melbourne, Melbourne, Victoria, 3010, Australia

(4) CSIRO Plant Industry, St Lucia , QLD , Australia

(5) School of Plant Biology, University of Western Australia, WA, 6009, Australia

(6) Corresponding Author: jacqueline.batley@uwa.edu.au, Tel: +61 (0)7 3346 9534

Abstract

Recent comparisons of the increasing number of genome sequences have revealed that variation in gene content is considerably more prevalent than previously thought. This variation is likely to have a pronounced effect on phenotypic diversity, and represents a crucial target for the assessment of genomic diversity. *Leptoshaeria maculans,* a causative agent of phoma stem canker, is the most devastating fungal pathogen of *Brassica napus* (oilseed rape/canola). A number of *L. maculans* genes are known to present in some isolates but lost in the others. We analyse gene content variation within three *L. maculans* isolates, using a hybrid mapping and genome assembly approach and identify genes which are present in one of the isolates but missing in the others. In total, 57 genes are shown to be missing in at least one isolate. The genes encode proteins involved in a range of processes including oxidative processes, DNA maintenance, cell signalling and sexual reproduction. The results demonstrate the effectiveness of the method and provide new insight into genomic diversity in *L. maculans*.

Key words: NGS, re-sequencing, gene loss, gene content variation, *Leptosphaeria maculans*

Introduction

The field of genomics is rapidly advancing, assisted by revolutions in DNA sequencing technology, with an increasing number of genomes being sequenced to varying degrees of completeness. The availability of increasingly complete references for a rapidly expanding number of genomes is providing detailed insight into the variation in gene content between organisms separated by a range of evolutionary distances. Over the last two decades, comparisons of a number of genome sequences have revealed considerable variation in gene content between and within species. Variation in gene content is known to be common in humans (Mills et al. 2011) fungi (Huang et al. 2014; McDonald et al. 2013; Syme et al. 2013) and many plant species (Morgante et al., 2005; Lai et al. 2010) (Tan et al. 2012; Zhang et al. 2014).

One of the main sources of variation in gene content is gene loss, which is known to occur via several mechanisms. Pseudogenization appears to commonly result in gene loss on an evolutionary time-scale (Zheng et al. 2007), with mutation leading to loss of function, removing selection pressure and resulting in ongoing genetic deterioration. The movement of transposable elements also contributes substantially to gene inactivation and subsequent loss in many organisms (Oliver and Greene 2009; Oliver 2012; Raffaele and Kamoun 2012), while unequal or illegitimate recombination events can result in large-scale gene loss (Devos et al. 2002). Mechanisms by which new genes can arise include duplication or mutation of existing genes, stable incorporation of foreign sequences (horizontal gene transfer) (Keeling and Palmer 2008; Ma et al. 2010; Oliver 2012; Richards 2011; Rouxel et al. 2011), or formation of novel genes from non-coding genomic regions (*de novo* gene birth) (Carvunis et al. 2012; Tautz and Domazet-Lošo 2011); gene duplication is the most extensively-studied of these processes. Duplicated genes within an organism are generally referred to as paralogues and groups of paralogues are often referred to as ‘gene families’. A substantial proportion of genes within organisms from all major forms of life are paralogues, i.e. they appear to share their origin with at least one other gene within the genome (Oliver 2012; Zhang 2003).

Variation in gene content is expected to influence many biological processes. In some instances the biological significance of such variation is well defined; for example, the presence of pathogen ‘effector’ or ‘avirulence’ (*Avr*) genes and corresponding host ‘resistance’ (*R*) genes is known to be a critically important factor for disease development in a range of host-pathogen interactions. Recognition of a specific pathogen *Avr* gene by a complementary *R* gene leads to disease resistance in the host, and the diversification or loss of *Avr* genes can therefore benefit the pathogen (Parker and Gilbert 2004). Structural rearrangements (deletions) and repeat-induced point (RIP) mutations (Fudal et al. 2009) are important mechanisms leading to the loss of *Avr* genes, via either deletion of the whole genomic region containing the gene, or substantial change to the coding sequence which prevents *R* gene recognition (Fudal et al. 2009; Gout et al. 2007). The susceptibility of canola (*Brassica napus*) to the fungal pathogen *Leptosphaeria maculans*, the causative agent of phoma stem canker, commonly known as blackleg disease, is in part determined by *Avr*-*R* gene interactions.

Here we present results of genome wide gene loss detection in three *L. maculans* isolates. We utilize a hybrid mapping and genome assembly approach to investigate gene content variation in three isolates of *L. maculans* an important pathogen of canola. The results suggest that genes involved in a number of

biological processes can be influenced by gene loss. These include genes involved in oxidative processes, DNA maintenance, cell signalling and sexual reproduction.

Methods

Mapping the Illumina resequencing data

The reference sequence (v2) for *L. maculans* and corresponding annotation (Rouxel et al. 2011) were downloaded (URGI Unite Recherche Genomique Info 2010).

Illumina DNA sequencing data for two *L. maculans* isolates 04MGPS021 and 06MGPP041, hereafter referred to as isolates 21 and 41, were previously described by (Zander et al. 2013). Sequence reads were mapped to the reference sequence using BWA v.0.6.2 (Li and Durbin 2009) with the following parameters: bwa aln –n 5 –k 2 –t 4; bwa sampe –n 100.

Coverage calculation and gene loss identification

A custom Java program SGSGeneLoss was used to calculate position-wise coverage across the whole genome, and then within exons of each individual gene. For gene-specific analysis, both 'horizontal' and 'vertical' coverage was calculated. Horizontal coverage was measured using *frac_exons_covered*, defined as the proportion of exon regions of a gene covered by mapped reads. Values for *frac_exons_covered* are in the range of 0 to 1; for a given gene a value of 0 would be returned where no reads map to any exon position, whereas a value of 1 would be returned where reads map to all exon positions. Vertical coverage is measured using *cov_cat*, equivalent to the average coverage depth across exon positions of a gene. *cov_cat* shows a continuous distribution, but is split into categories for visualisation purposes, with categories specified by the user. The categories used for the examples provided here are as follows: 0; > 0x <= 10x; > 10x <= 20x; > 20x <= 40x; > 40x <= 70x; > 70x. Outline of the methods and details of calculations used to determine *frac_exons_covered* and *cov_cat* are shown in Online Resource 1 and Online Resource 2.

The results of the exon coverage calculation are visualised as scatter plots by script graph_chromosomes.R using ggplot2 (Wickham 2009) and as circular plots by script graph_circles.R using ggbio (Yin et al. 2012).

For minimum coverage estimation, genes were considered lost if *frac_exons_covered* was equal to zero, indicating that no reads map to the predicted exonic regions. For actual gene loss calling in isolates 21 and 41 the parameters were relaxed and genes were considered lost if *frac_exons_covered* was <=0.05 Lost genes were functionally annotated using Blast2GO (Conesa et al. 2005).

The software package is freely available from: http://www.appliedbioinformatics.com.au/index.php/SGSGeneLoss

Isolate 21 and 41 genome assembly and annotation

Sequencing reads from isolates 21 and 41 were separately assembled using Velvet (Zerbino and Birney 2008) v1.2.10. Prior to the assembly reads were quality trimmed using Trimmomatric (Bolger et al. 2014) v0.32 (SLIDINGWINDOW:5:20 MINLEN:80). The k-mer values used for the assembly were 55 for isolate 21 and 51 for isolate 41. Resulting genome assemblies (contigs) were compared to the reference sequence using MEGABLAST (Camacho et al. 2009) (BLAST+ v.2.2.28). Contigs or parts of contigs which did not match the reference sequence were extracted and the sequences at least 300 bp in length were annotated using Eugene web server (http://bioinformatics.psb.ugent.be/webtools/EuGene/) (Foissac et al. 2008). Predicted genes were functionally annotated using Blast2GO. The annotated genes in isolates 21 and 41 were compared using BLASTP (BLAST+ v2.2.28)

PCR validation

Fragments of the mating type alleles *MAT1-1* and *MAT1-2* were amplified from each isolate (v23.1.2, 21 and 41) using PCR primers previously described by (Cozijnsen and Howlett 2003). The three primers used were: one with a sequence identical to that of the region the flanking idiomorph (primer 1), one identical to sequences within the alpha-box of *MAT1-1* (primer 2) and one identical to sequences in the HMG domain of *MAT1-2* (primer 3). The expected amplified fragment sizes were: 688 bp for the *MAT1-1* locus and 443 bp for the *MAT1-2* locus.

primer 1	TGGCGAATTAAGGGATTGCTG
primer 2	CTCGATGCAATGTACTTGG
primer 3	AGCCGGAGGTGAAGTTGAAGCCG

Results

Gene loss between isolates 21, 41 and v23.1.3 (isolate for which a published reference sequence was available) was investigated. The method used was a combination of read mapping and genome de novo assembly.

Estimating minimal coverage requirements

The reliability of read mapping as gene loss detection tool is dependent on the sequencing depth and the resulting coverage of the reference genome. If sequencing coverage is too low, segments of the genome may not be represented by chance, resulting in false calling of gene loss. To estimate the minimum coverage required to achieve reliable results, dataset 41 was split to produce datasets representing 1x, 2x, 5x, 10x, 15x, 20x, 30x, 40x and 60x coverage. Each dataset was used to predict gene loss and the number of genes lost plotted as the function of coverage (Fig. 1). The resulting curve plateaus between 10x and 15x coverage, which can be considered an appropriate coverage necessary to achieve accurate results. This is in agreement with the Lander-Waterman statistics (Lander and Waterman 1988); which predicts that at 5x sequencing coverage 0.67% of the genome is not represented, whereas at 10x coverage only 0.0045% of the genome is not represented.

Gene loss in isolates 21 and 41

Isolates 21 and 41 were sequenced to 64.6x and 62.5x coverage respectively. The total number of genes predicted to be lost, compared to the reference sequence, was 20 for isolate 21, and 33 for isolate 41. The list of all predicted lost genes is presented in Table 1 and visualized in Fig. 2. Functional annotation of the lost genes in isolate 21 suggests that they include *Avr1* (Fig. 3), cytochromes P450 and glycosyltransferese (Table 1). Functional annotation of genes lost in isolate 41 suggests the genes lost include: *Avr1*, *Avr6*, *MAT1-2*, cytochromes P450, helicases, kinase and phosphotransferase (Table 1). The function of many (51.4%) genes lost in isolates 21 and 41 is unknown. In total 16 genes were absent in both isolate 21 and 41: four genes were absent only in isolate 21 and 17 genes were absent only in isolate 41.

The spatial distribution of genes lost suggests that genes were lost either in isolation or as contiguous losses of multiple genes. Out of 20 gene lost in isolate 21, 8 (40%) neighboured another gene lost. Out of 33 genes lost in isolate 41, 18 (54.5%) also neighboured another gene lost.

PCR amplification of *MAT* locus

During PCR, the *MAT* locus was amplified in all three isolates; v23.1.3, 21 and 41. A 688 bp fragment was amplified in isolates 21 and v23.1.3 confirming presence of *MAT1-2* allele whilst a 443 bp fragment corresponding to the *MAT1-1* allele was amplified in isolate 41

Gene loss in isolate v23.1.3

Sequencing reads for isolates 21 and 41 were assembled using Velvet. The total assembly length for isolate 21 was 34 Mbp and for isolate 41 was 33.1 Mbp. The N50 values were 2422 and 5743 respectively. Upon comparison with the available reference sequence for isolate v23.1.3 a total of 0.68 Mbp and a total of 0.63 Mbp of sequence were present in isolates 21 and 41 respectively but missing in isolate v23.1.3.

Annotation of sequence found in isolate 21, but not isolate v23.1.3, resulted in discovery of eight genes which are lost in v23.1.3 (Table 2). Annotation of sequence found in isolate 41, but not isolate v23.1.3, resulted in the discovery of 19 genes missing in v23.1.3. Amino acid sequences of genes predicted to be lost in isolate v23.1.3 are available in Online Resource 3. All of the genes annotated for isolates 21 and 41 were found on separate contigs. Sequence comparisons suggest that out of 27 genes found in the re-sequenced isolates, but absent in v23.1.3, seven are present in both isolates. This is further confirmed by functional annotation; the seven corresponding genes found in isolates 21 and 41 have the same functional annotation (Table 2). The corresponding genes are: gene1_21-gene12_41, gene2_21-gene10_41, gene3_21-gene11_41, gene4_21-gene19_41, gene6_21-gene3_41, gene7_21-gene17_41 and gene_8_21-gene8_41. The genome assembly of isolate 21 was found to contain one unique gene, gene5_21, of an unknown function. The genome assembly of isolate 41 was found to contain 12 unique genes, which include genes encoding proteins annotated as eonyl-hydratase, transmembrane protein pft27, coproporphyrinogen III oxidase. Among the genes unique to isolate 41 is the *MAT-1-1* allele. The

predicted sequence is 440 amino acids in length and 99% identical to the published sequence (Cozijnsen and Howlett 2003).

Discussion

We present analysis of gene content variation in three *L. maculans* isolates. Gene content variation was analysed by a combination of read mapping and assembly. The genome assemblies obtained for isolates 21 and 41 and are smaller in size compared to the published reference sequence which is 45 Mbp in length (Rouxel et al. 2011). This may result from different assembly algorithms used. Roughly 30% of the *L. maculans* genome is composed of repetitive elements (transposons) (Rouxel et al. 2011) which are unlikely to be assembled by Velvet.

The v23.1.3 isolate (for which a published reference genome is available) differs by a total of 28 genes with isolate 21 and 52 genes in isolate 41. The higher number of differences between isolate v23.1.3 and isolate 41 is consistent with the published observations, suggesting that isolate 41 is a more distant relative of v23.1.3 than isolate 21 (Zander et al. 2013).

Two classes of genes found in the *L. maculans* genome have previously been shown to display presence/absence variation: the mating type (*MAT*) locus and avirulence (*Avr*) genes, and these genes were used to verify results obtained. *L. maculans* contains a single mating type (*MAT*) locus, with two alternate forms, which must be different for two isolates to mate (Venn 1979). *MAT1-1* is 1,368 bp long, containing a 45-bp intron and encoding a protein of 441 amino. The *MAT1-2* is 1,246 bp, encoding a predicted protein of 397 amino acids (Cozijnsen and Howlett 2003). The reference sequence isolate v23.1.3 is known to contain *MAT1-2* (Fitt et al. 2006) . The results found here suggest that isolate 41 is the opposite *MAT1-1* mating type. The results were validated using PCR. The PCR results confirmed *in silico* gene loss predictions.

Two of the predicted lost genes are classified as avirulence genes, which have previously been shown to display presence/absence variation between isolates (Fudal et al. 2009; Gout et al. 2006; Gout et al. 2007; Van de Wouw et al. 2010; Zander et al. 2013). Our analysis suggests that both isolates 21 and 41 demonstrate differential *AvrLm* gene loss (Fig. 3), with *AvrLm1* predicted to be absent in both isolates, *AvrLm4-7* and *AvrLm11* present in both isolates, while *AvrLm6* is present in isolate 21, but absent in isolate 41. The results are in full agreement with published observations of the virulence specificity of these isolates (Raman et al. 2012; Zander et al. 2013). The loss of *AvrLm1* is most likely due to a larger deletion of a whole genomic region including the *AvrLm1* locus (Fudal et al. 2009; Gout et al. 2007). The loss of *AvrLm6* might be due to a larger genomic deletion or repeat induced point (RIP) mutations, which can lead to functional loss of a gene (Fudal et al. 2009; Van de Wouw et al. 2010).

Additional genes predicted to be lost from one or both isolates include: two cytochrome P450s, two DNA repair helicases and a serine threonine protein kinase (Table 1). Cytochrome P450 enzymes (P450s) are heme-thiolate proteins found in all life forms from prokaryotes (archea, bacteria) to eukaryotes (fungi, insects, plants and animals, including humans). These enzymes catalyse regio- and stereospecific

conversions of a wide range of lipophilic compounds to more hydrophilic derivatives by introducing an oxygen atom derived from molecular oxygen (Crešnar and Petrič 2011). Currently over 8700 different cytochrome P450s have been identified from 113 fungal species (Park et al. 2008). Cytochrome P450 is implicated in a range of fungal processes including, but not limited to; antibiotic detoxification (George et al. 1998), napthalene metabolism and detoxification of polycyclic aromatic hydrocarbons (Cerniglia et al. 1978; da Silva et al. 2004; Sutherland 1992), giberellin biosynthesis (Rojas et al. 2001) and toxin biosynthesis (Bhatnagar et al. 2003). The *L. maculans* reference genome is predicted to contain 62 distinct P450 genes (Park et al. 2008). The two genes found to be lost: G085580.1 (lost in isolate 41 only) and G085620.1 (lost in both isolates) are both annotated as E class P450, group I (Park et al. 2008). Further characterization of these genes may provide information regarding the influence of gene loss on the fungal phenotype.

Helicases are enzymes capable of separation of double-stranded nucleic acid (DNA or RNA), and are involved in a number of biological processes including replication, recombination, transcription and translation. The number of helicase genes in eukaryotic organisms is relatively high, with approximately 1% of genes in eukaryotic genomes predicted to encode RNA or DNA helicases (Wu 2012). Furthermore, redundancy of genes coding for helicases has been observed. For example in rice a total of 115 helicase genes have been identified, several of which were redundant (Umate et al. 2011) suggesting that gene loss may affect helicase genes without a pronounced effect on the phenotype.

Genes coding for protein kinases belong to large families and their products possess diverse functions. They are also thought to exhibit a rapid birth-death cycle (Hardie 1999; Rudrabhatla et al. 2006; Zheng et al. 2011). Therefore protein kinases are expected to be affected by gene loss.

Our results support previous observations suggesting that avirulence (*AvrLm*) genes are subject to presence/absence variation. Results presented here also suggest that gene loss influences additional genes in *L. maculans,* including genes possibly involved in oxidative processes, DNA maintenance, cellular signalling and sexual reproduction. The spatial distribution of genes lost suggests that genes can be lost either in isolation or as consecutive losses of multiple genes presumably as a part of larger deletion.

A software package SGSGeneLoss was developed to facilitate gene loss detection in *L. maculans* However, SGSGeneLoss can be used to detect variation in gene content in a number of species. The software allows both detection and visualization of gene loss and is freely available.

Acknowledgements

The authors would like to acknowledge funding support from the Australian Research Council (Projects LP0882095, LP0883462, LP0989200, LP110100200 and DP0985953).

References

Bhatnagar D, Ehrlich KC, Cleveland TE (2003) Molecular genetic analysis and regulation of aflatoxin biosynthesis. Applied microbiology and biotechnology 61:83-93. doi:10.1007/s00253-002-1199-x

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421. doi:10.1186/1471-2105-10-421

Carvunis A-R et al. (2012) Proto-genes and de novo gene birth. Nature 487:370-374. doi:10.1038/nature11184

Cerniglia CE, Hebert RL, Szaniszlo PJ, Gibson DT (1978) Fungal transformation of naphthalene. Archives of microbiology 117:135-143. doi:10.1007/BF00402301

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics (Oxford, England) 21:3674-3676. doi:10.1093/bioinformatics/bti610

Cozijnsen AJ, Howlett BJ (2003) Characterisation of the mating-type locus of the plant pathogenic ascomycete *Leptosphaeria maculans*. Current genetics 43:351-357. doi:10.1007/s00294-003-0391-6

Crešnar B, Petrič S (2011) Cytochrome P450 enzymes in the fungal kingdom. Biochimica et biophysica acta (BBA) - Proteins and proteomics 1814:29-35. doi:10.1016/j.bbapap.2010.06.020

da Silva M, Esposito E, Moody JD, Canhos VP, Cerniglia CE (2004) Metabolism of aromatic hydrocarbons by the filamentous fungus *Cyclothyrium* sp. Chemosphere 57:943-952. doi:10.1016/j.chemosphere.2004.07.051

Devos KM, Brown JKM, Bennetzen JL (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Research 12:1075-1079. doi:10.1101/gr.132102

Fitt BD, Evans N, Howlett BJ, Cooke M (2006) Sustainable strategies for managing Brassica napus (oilseed rape) resistance to Leptosphaeria maculans (phoma stem canker) vol 114. Springer,

Foissac S et al. (2008) Genome annotation in plants and fungi: EuGene as a model platform. Current Bioinformatics 3:87-97

Fudal I et al. (2009) Repeat-induced point mutation (RIP) as an alternative mechanism of evolution toward virulence in *Leptosphaeria maculans*. Molecular Plant-Microbe Interactions 22:932-941. doi:10.1094/MPMI-22-8-0932

George H, Hirschi K, VanEtten H (1998) Biochemical properties of the products of cytochrome P450 genes (PDA) encoding pisatin demethylase activity in *Nectria haematococca*. Archives of microbiology 170:147-154. doi:10.1007/s002030050627

Gout L et al. (2006) Lost in the middle of nowhere: the AvrLm1 avirulence gene of the Dothideomycete *Leptosphaeria maculans*. Molecular microbiology 60:67-80. doi:10.1111/j.1365-2958.2006.05076.x

Gout L et al. (2007) Genome structure impacts molecular evolution at the AvrLm1 avirulence locus of the plant pathogen *Leptosphaeria maculans*. Environmental microbiology 9:2978-2992. doi:10.1111/j.1462-2920.2007.01408.x

Hardie DG (1999) Plant protein serine/threonine kinases: classification and functions. Annual review of plant physiology and plant molecular biology 50:97-131. doi:10.1146/annurev.arplant.50.1.97

Huang J, Si W, Deng Q, Li P, Yang S (2014) Rapid evolution of avirulence genes in rice blast fungus Magnaporthe oryzae. BMC Genetics 15:45

Keeling PJ, Palmer JD (2008) Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet 9:605-618. doi:doi:10.1038/nrg2386

Lander ES, Waterman MS (1988) Genomic mapping by fingerprinting random clones: A mathematical analysis. Genomics 2:231-239. doi:10.1016/0888-7543(88)90007-9

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760. doi:10.1093/bioinformatics/btp324

Ma L-J et al. (2010) Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. Nature 464:367-373. doi:10.1038/nature08850

McDonald MC, Oliver RP, Friesen TL, Brunner PC, McDonald BA (2013) Global diversity and distribution of three necrotrophic effectors in Phaeosphaeria nodorum and related species. New Phytologist 199:241-251. doi:10.1111/nph.12257

Mills RE et al. (2011) Mapping copy number variation by population-scale genome sequencing. Nature 470:59-65. doi:10.1038/nature09708

Oliver KR, Greene WK (2009) Transposable elements: powerful facilitators of evolution. BioEssays 31:703-714. doi:10.1002/bies.200800219

Oliver R (2012) Genomic tillage and the harvest of fungal phytopathogens. New Phytologist 196:1015-1023. doi:10.1111/j.1469-8137.2012.04330.x
Park J et al. (2008) Fungal cytochrome P450 database. BMC genomics 9:402. doi:10.1186/1471-2164-9-402
Parker IM, Gilbert GS (2004) The evolutionary ecology of novel plant-pathogen interactions. Annu Rev Ecol Evol Syst 35:675-700. doi:10.1146/annurev.ecolsys.34.011802.132339
Raffaele S, Kamoun S (2012) Genome evolution in filamentous plant pathogens: why bigger can be better. Nat Rev Micro 10:417-430
Raman R et al. (2012) Molecular mapping of qualitative and quantitative loci for resistance to *Leptosphaeria maculans* causing blackleg disease in canola (*Brassica napus* L.). Theor Appl Genet 125:405-418. doi:10.1007/s00122-012-1842-6
Richards TA (2011) Genome Evolution: Horizontal Movements in the Fungi. Current Biology 21:R166-R168. doi:http://dx.doi.org/10.1016/j.cub.2011.01.028
Rojas MC, Hedden P, Gaskin P, Tudzynski B (2001) The P450-1 gene of *Gibberella fujikuroi* encodes a multifunctional enzyme in gibberellin biosynthesis. Proceedings of the National Academy of Sciences 98:5838-5843. doi:10.1073/pnas.091096298
Rouxel T et al. (2011) Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. Nature communications 2:202. doi:10.1038/ncomms1189
Rudrabhatla P, Reddy MM, Rajasekharan R (2006) Genome-wide analysis and experimentation of plant serine/threonine/tyrosine-specific protein kinases. Plant molecular biology 60:293-319. doi:10.1007/s11103-005-4109-7
Sutherland JB (1992) Detoxification of polycyclic aromatic hydrocarbons by fungi. Journal of industrial microbiology 9:53-61. doi:10.1007/BF01576368
Syme RA, Hane JK, Friesen TL, Oliver RP (2013) Resequencing and Comparative Genomics of *Stagonospora nodorum*: Sectional Gene Absence and Effector Discovery. G3: Genes|Genomes|Genetics 3:959-969
Tan S, Zhong Y, Hou H, Yang S, Tian D (2012) Variation of presence/absence genes among *Arabidopsis* populations. BMC evolutionary biology 12:86. doi:10.1186/1471-2148-12-86
Tautz D, Domazet-Lošo T (2011) The evolutionary origin of orphan genes. Nat Rev Genet 12:692-702
Umate P, Tuteja N, Tuteja R (2011) Genome-wide comprehensive analysis of human helicases. Communicative & integrative biology 4:118-137. doi:10.4161/cib.4.1.13844
URGI Unite Recherche Genomique Info (2010) Plant and fungi data integration. URGI. http://urgi.versailles.inra.fr/Species/Leptosphaeria/Sequences-Databases/Download. 2014
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 Pathogens 6:e1001180. doi:10.1371/journal.ppat.1001180
Venn L (1979) The Genetic Control of Sexual Compatibility in *Leptosphaeria maculans*. Australasian Plant Pathology 8:5-6. doi:10.1071/APP9790005
Wickham H (2009) ggplot2: elegant graphics for data analysis. New York
Wu Y (2012) Unwinding and rewinding: double faces of helicase? Journal of nucleic acids 2012:140601. doi:10.1155/2012/140601
Yin T, Cook D, Lawrence M (2012) ggbio: an R package for extending the grammar of graphics for genomic data. Genome Biology 13:R77. doi:10.1186/gb-2012-13-8-r77
Zander M et al. (2013) Identifying genetic diversity of avirulence genes in *Leptosphaeria maculans* using whole genome sequencing. Funct Integr Genomics 13:295-308. doi:10.1007/s10142-013-0324-5
Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18:821-829
Zhang J (2003) Evolution by gene duplication: an update. Trends in Ecology & Evolution 18:292-298. doi:10.1016/S0169-5347(03)00033-8
Zhang L-M, Luo H, Liu Z-Q, Zhao Y, Luo J-C, Hao D-Y, Jing H-C (2014) Genome-wide patterns of large-size presence/absence variants in sorghum. Journal of Integrative Plant Biology 56:24-37. doi:10.1111/jipb.12121
Zheng D et al. (2007) Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. Genome Research 17:839-851. doi:10.1101/gr.5586307
Zheng L-Y et al. (2011) Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). Genome biology 12:R114. doi:10.1186/gb-2011-12-11-r114

Table 1 List of genes identified as lost in resequenced *Leptosphaeria maculans* isolates relative the reference genome. Genes deemed less confident by the authors (length <300 aa and without EST evidence at the time of prediction pipeline were marked with a *). The *Avr* genes are shown in bold.

Isolate 21	Protein description	Isolate 41	Protein description
		G030170.1*	Predicted protein
		G037480.1*	Predicted protein
		G043120.1*	Predicted protein
G046740.1*	Predicted protein		
G049660.1	M1 protein (*AvrLm1*)	**G049660.1**	M1 protein (*AvrLm1*)
		G049920.1*	Predicted protein
		G049930.1	Predicted protein
		G049940.1	M6 protein (*AvrLm6*)
G076370.1*	Predicted protein		
		G070000.1*	Histidine triad
G079780.1*	Predicted protein	G079780.1*	Predicted protein
G081630.1*	Predicted protein	G081630.1*	Predicted protein
G081640.1*	Predicted protein	G081640.1*	Predicted protein
G082370.1*	Predicted protein		
G082540.1*	Glucose-6-phosphate isomerase	G082540.1*	Glucose-6-phosphate isomerase
G085580.1	Cytochrome p450	G085580.1	Cytochrome p450
G085590.1	Transferase family protein	G085590.1	Transferase family protein
G085610.1*	Predicted protein	G085610.1*	Predicted protein
G085620.1	Benzoate 4-monooxygenase cytochrome p450	G085620.1	Benzoate 4-monooxygenase cytochrome p450
		G085630.1	Carboxylesterase family protein
		G086520.1	DNA repair helicase
		G086530.1	DNA repair helicase
G090230.1*	Predicted protein	G090230.1*	Predicted protein
		G098080.1	Serine threonine protein kinase
		G098090.1	Predicted protein
		G103490.1*	Phosphotransferase family protein
G104510.1*	Predicted protein	G104510.1*	Predicted protein

		G109260.1	Het domain-containing protein
G110260.1*	Predicted protein	G110260.1*	Predicted protein
G113050.1	Tpr domain protein	G113050.1	Tpr domain protein
G113060.1	Predicted protein	G113060.1	Predicted protein
G113080.1	Predicted protein	G113080.1	Predicted protein
		G114280.1	Predicted protein
		G114380.1	Pyridoxamine phosphate oxidase
		G114390.1	Mating-type mat1-2 protein
G116710.1	Glycosyltransferase family 34 protein		
G123850.1*	Predicted protein	G123850.1*	Predicted protein

Table 2 List of genes identified as lost in the reference genome (v23.1.3) relative to the re-sequenced isolates

Isolate 21	Protein description
gene1_21	Extracellular aldonolactonase protein
gene2_21	Hypothetical protein SETTUDRAFT_40572
gene3_21	Integral membrane pth11-like protein
gene4_21	Pyruvate carboxylase subunit a
gene5_21	Hypothetical protein COCCADRAFT_9848
gene6_21	Beta and beta-prime subunits of DNA dependent RNA-polymerase
gene7_21	Scp-like extracellular
gene8_21	DnaJ domain containing protein
Isolate 41	
gene1_41	Enoyl- hydratase
gene2_41	Transmembrane protein pft27
gene3_41	Beta and beta-prime subunits of DNA dependent RNA-polymerase
gene4_41	Chromosome segregation ATPase family protein
gene5_41	Cellular morphogenesis protein
gene6_41	N/A
gene7_41	Het domain-containing protein
gene8_41	DnaJ domain containing protein
gene9_41	Arf GTPase activating protein
gene10_41	Hypothetical protein SETTUDRAFT_40572
gene11_41	Integral membrane pth11-like protein
gene12_41	Extracellular aldonolactonase protein
gene13_41	Hypothetical protein PTT_20052
gene14_41	mRNA binding protein pumilio 2
gene15_41	DNA binding protein with alpha box domain (*MAT-1-1*)
gene16_41	Coproporphyrinogen III oxidase
gene17_41	Scp-like extracellular
gene18_41	Transcription factor bhlh protein
gene19_41	Pyruvate carboxylase subunit a

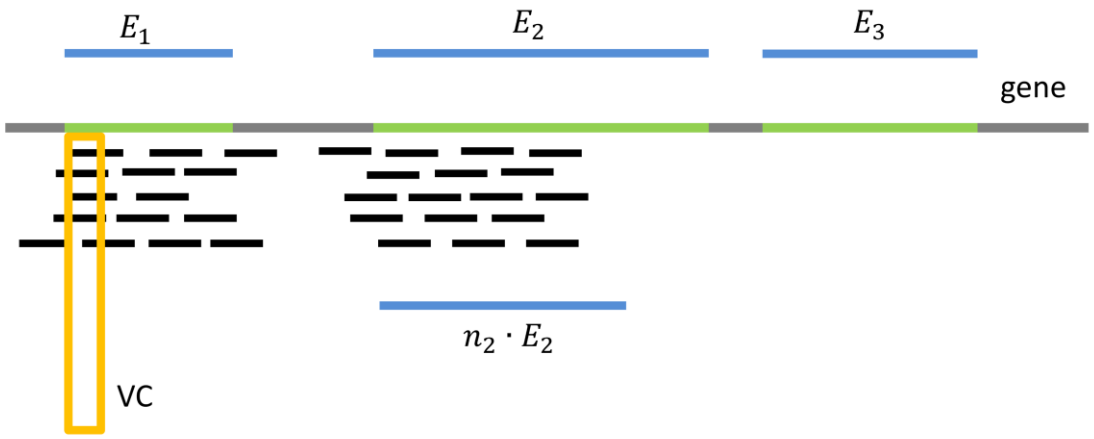
Fig 1 Minimum coverage estimation. Number of lost genes at different levels of coverage (1x, 2x, 5x, 10x, 15x, 20x, 30x, 40x, 60x)

Fig 2 Graphical representation of genes lost for isolates 21 and 41. The picture is composed of three tracks. From the inner track they correspond to: positions of genes lost (blue lines) in isolate 21, positions of genes lost in isolate 41 (blue lines), graphical representations of all contigs (grey rectangles), which have annotated genes

Fig 3 *AvrLm* gene loss in *L. maculans*. Plots visualising lost *AvrLm* genes in isolates 41 and 21. Each dot represents a gene. Coordinate on the x-axis is the position along the contig. Coordinate on the y-axis is the fraction of the total number of positions within exons of the gene covered by mapped reads. Coverage category (*cov_cat*) is the average coverage depth across exons of the gene (see Methods for details). Isolate 41 shows loss of two *AvrLm* genes present on Supercontig 6 – *AvrLm1* and *AvrLm6* (a). Isolate 21 shows loss of one *AvrLm* gene present on Supercontig 6 – *AvrLm1* (b)

Online Resource 1 Details of exon coverage calculation used in gene loss detection

Online Resource Gene loss identification pipeline. First, sequencing reads are mapped to the reference sequence (a). Then, coverage across exons within the annotated genes is calculated (see Methods for details) (b). Finally, genes with sufficiently low coverage are classified as lost (c)



$$\boldsymbol{frac\_exons\_covered} = \frac{\sum_{i=1}^{x} n_i \cdot len(E_i)}{\sum_{i=1}^{x} len(E_i)}$$

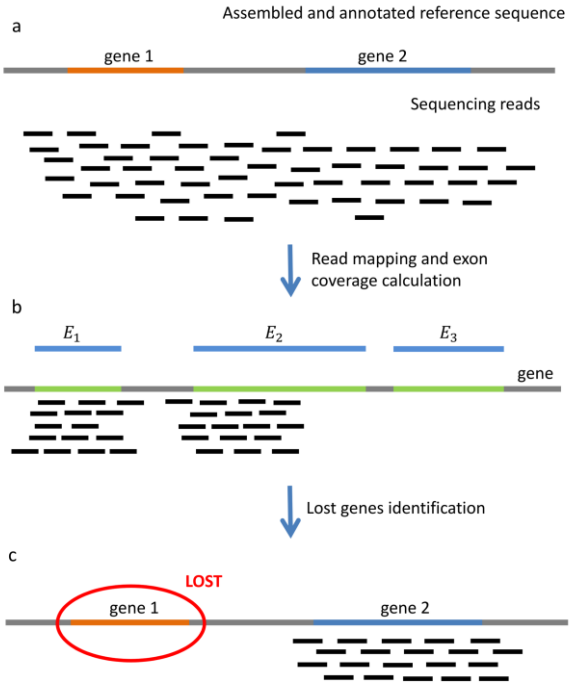
$x -$ *number of exons in a gene*

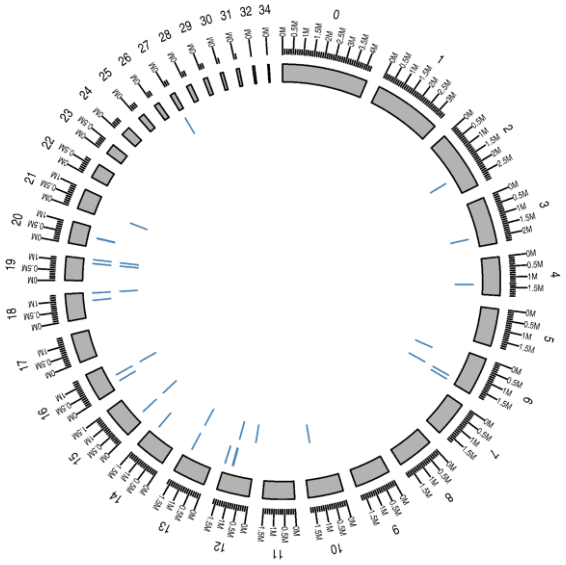
$n -$ *fraction of exon covered* $n \in [0,1]$

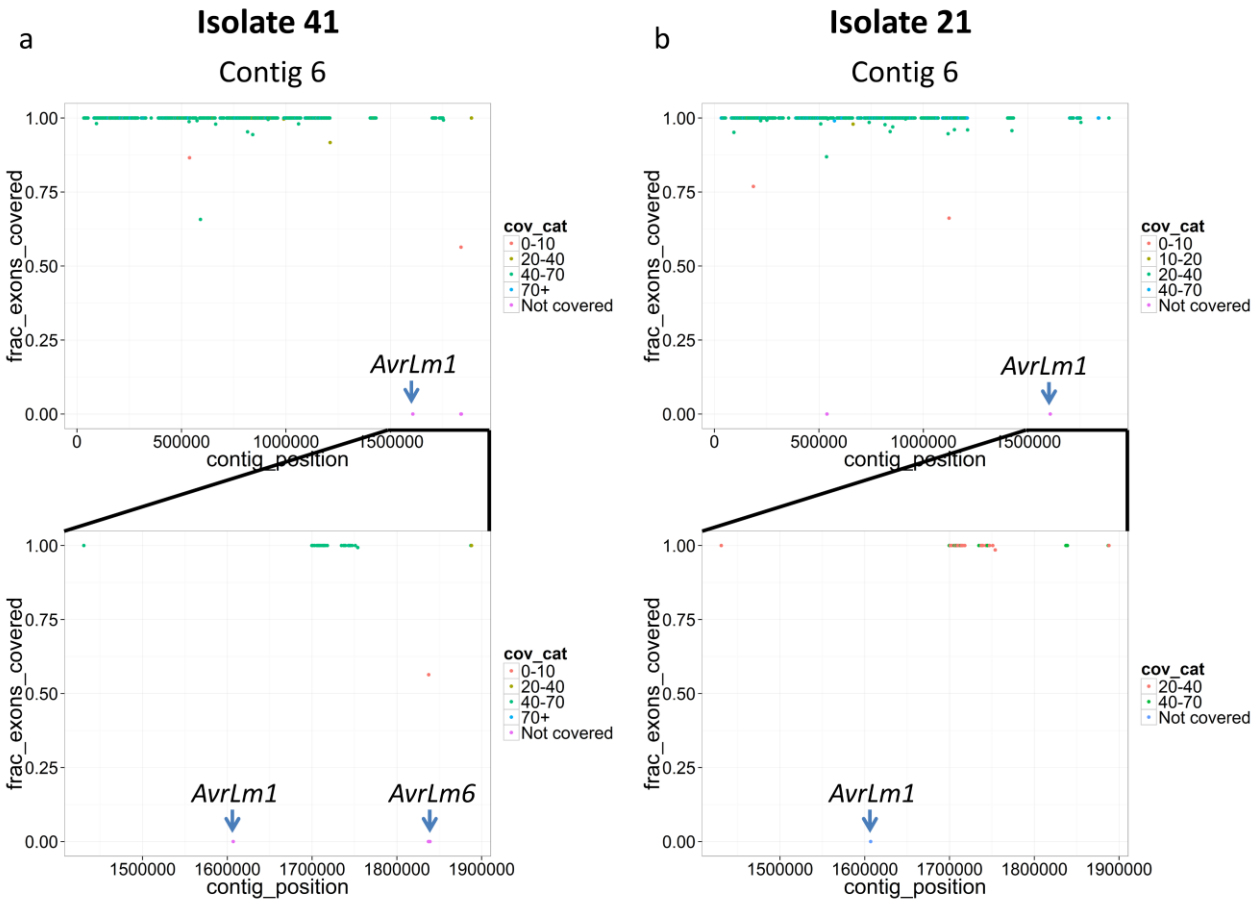
$E_i -$ *ith exon of the gene*

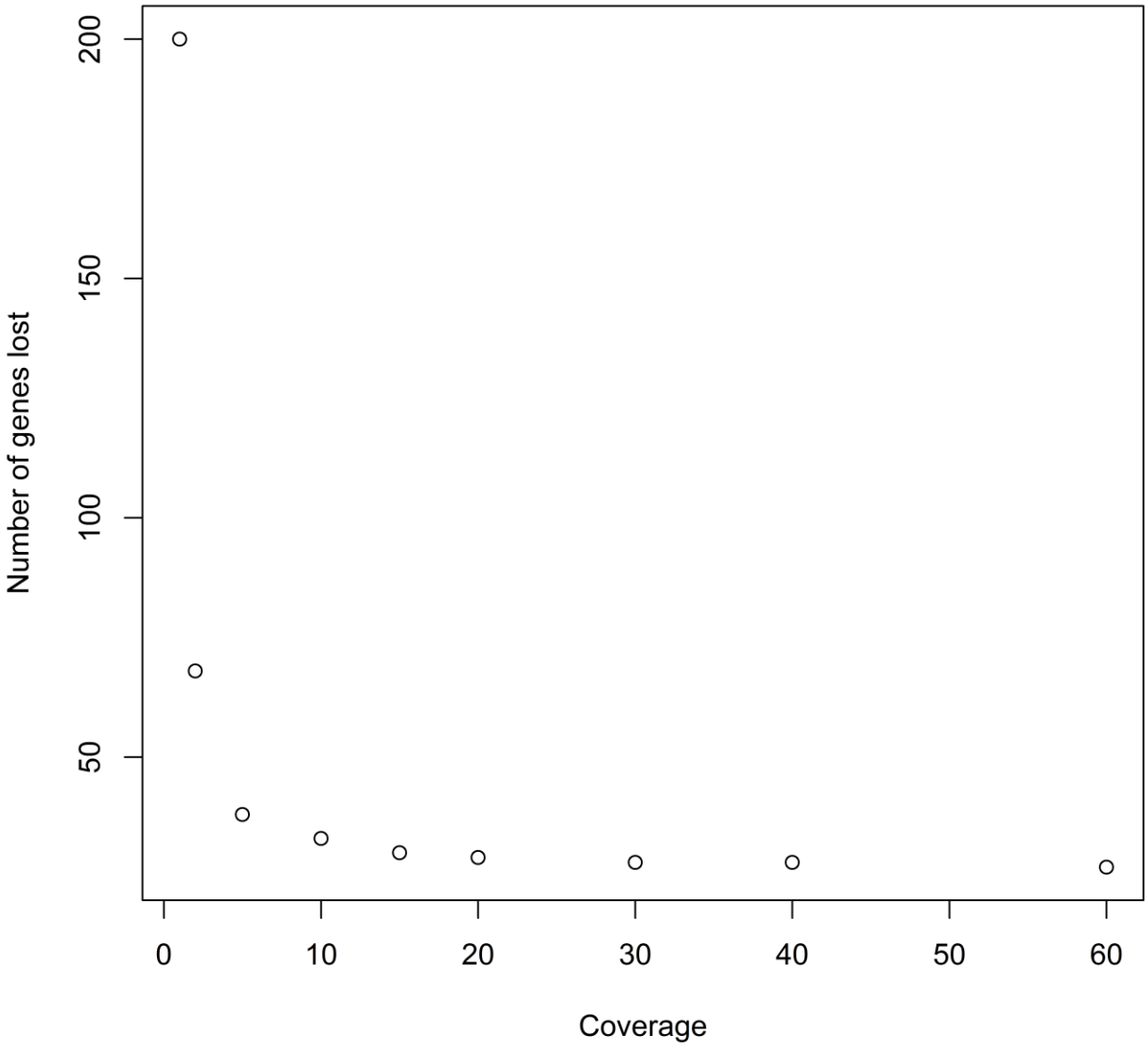
$$\boldsymbol{cov\_cat} = \frac{\sum_{i=1}^{x} \sum_{j=1}^{l} VC_{ij}}{\sum_{i=1}^{x} n_i \cdot len(E_i)}$$

$l -$ *length of the exon*


a
Assembled and annotated reference sequence
gene 1
gene 2
Sequencing reads
Read mapping and exon coverage calculation
b
E_1
E_2
E_3
gene
Lost genes identification
c
LOST
gene 1
gene 2




a
Isolate 41
Contig 6
frac_exons_covered
1.00
0.75
0.50
0.25
0.00
0
500000
1000000
1500000
contig_position
AvrLm1
cov_cat
0-10
20-40
40-70
70+
Not covered
1500000
1600000
1700000
1800000
1900000
AvrLm1
AvrLm6
b
Isolate 21
Contig 6
cov_cat
0-10
10-20
20-40
40-70
Not covered


Number of genes lost
200
150
100
50
0
10
20
30
40
50
60
Coverage

Author/s:
Golicz, AA; Martinez, PA; Zander, M; Patel, DA; Van De Wouw, AP; Visendi, P; Fitzgerald, TL; Edwards, D; Batley, J

Title:
Gene loss in the fungal canola pathogen Leptosphaeria maculans

Date:
2015-03-01

Citation:
Golicz, A. A., Martinez, P. A., Zander, M., Patel, D. A., Van De Wouw, A. P., Visendi, P., Fitzgerald, T. L., Edwards, D. & Batley, J. (2015). Gene loss in the fungal canola pathogen Leptosphaeria maculans. FUNCTIONAL & INTEGRATIVE GENOMICS, 15 (2), pp.189-196. https://doi.org/10.1007/s10142-014-0412-1.