The Complete Genome Sequence of the Phytopathogenic Fungus *Sclerotinia sclerotiorum* Reveals Insights into the Genome Architecture of Broad Host Range Pathogens

Mark Derbyshire,[1,*] Matthew Denton-Giles,[1] Dwayne Hegedus,[2] Shirin Seifbarghi,[2] Jeffrey Rollins,[3] Jan van Kan,[4] Michael F. Seidl,[4] Luigi Faino,[4] Malick Mbengue,[5] Olivier Navaud,[5] Sylvain Raffaele,[5] Kim Hammond-Kosack,[6] Stephanie Heard,[3,6] and Richard Oliver[1]

[1]Centre for Crop and Disease Management Department of Environment and Agriculture, Curtin University, Bentley, Perth, Western Australia, Australia

[2]Agriculture and Agri-Food Canada, Saskatoon, Saskatchewan, Canada

[3]Department of Plant Pathology, University of Florida, Gainesville, FL

[4]Laboratory of Phytopathology, Wageningen University, The Netherlands

[5]LIPM Université de Toulouse INRA CNRS, Castanet-Tolosan, France

[6]Department of Plant Biology and Crop Sciences, Rothamsted Research, Harpenden, Hertfordshire, United Kingdom

*Corresponding author: E-mail: mark.derbyshire@curtin.edu.au.

Accepted: February 14, 2017

Data deposition: Details= Chr_1 CP017814 Chr_2 CP017815, Chr_3 CP017816, Chr_4 CP017817, Chr_5 CP017818, Chr_6 CP017819, Chr_7 CP017820, Chr_8 CP017821, Chr_9 CP017822, Chr_10 CP017823, Chr_11 CP017824, Chr_12 CP017825, Chr_13 CP017826, Chr_14 CP017827, Chr_15 CP017828, Chr_16 CP017829.

Abstract

Sclerotinia sclerotiorum is a phytopathogenic fungus with over 400 hosts including numerous economically important cultivated species. This contrasts many economically destructive pathogens that only exhibit a single or very few hosts. Many plant pathogens exhibit a "two-speed" genome. So described because their genomes contain alternating gene rich, repeat sparse and gene poor, repeat-rich regions. In fungi, the repeat-rich regions may be subjected to a process termed repeat-induced point mutation (RIP). Both repeat activity and RIP are thought to play a significant role in evolution of secreted virulence proteins, termed effectors. We present a complete genome sequence of *S. sclerotiorum* generated using Single Molecule Real-Time Sequencing technology with highly accurate annotations produced using an extensive RNA sequencing data set. We identified 70 effector candidates and have highlighted their *in planta* expression profiles. Furthermore, we characterized the genome architecture of *S. sclerotiorum* in comparison to plant pathogens that exhibit "two-speed" genomes. We show that there is a significant association between positions of secreted proteins and regions with a high RIP index in *S. sclerotiorum* but we did not detect a correlation between secreted protein proportion and GC content. Neither did we detect a negative correlation between CDS content and secreted protein proportion across the *S. sclerotiorum* genome. We conclude that *S. sclerotiorum* exhibits subtle signatures of enhanced mutation of secreted proteins in specific genomic compartments as a result of transposition and RIP activity. However, these signatures are not observable at the whole-genome scale.

Key words: *Sclerotinia sclerotiorum*, two-speed, effector, repeat-induced point mutation, PacBio.

Introduction

Sclerotinia sclerotiorum is a plant pathogenic fungus with a remarkably broad host range. An early literature review by Boland and Hall (1994) identified 408 species of plants in 278 genera and 75 families that are susceptible to infection by *S. sclerotiorum*. A number of these host species are economically important crops, for example, *Brassica napus* (canola/oilseed rape), *Glycine max* (soybean), *Beta vulgaris*

© The Author(s) 2017. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

(sugar beet), *Arachis hypogaea* (peanut), and *Lactuca sativa* (garden lettuce) (Derbyshire and Denton-Giles, 2016; Peltier et al. 2012; Naito and Sugimoto, 1986; Heffer, 2007; Chitrampalam et al. 2008). The broad host range of *S. sclerotiorum* is in contrast to most other well-studied plant pathogenic fungi that infect only a single or small number of host species.

Illustrating this point, a recent article listed the "top ten plant pathogenic fungi," based on economic destructiveness and scientific importance (Dean et al. 2012). This article was produced following a survey of 495 international experts on fungal diseases of plants. The list contained eight narrow host range plant infecting fungal species or genera and only two broad host range pathogens, *Botrytis cinerea* and *Fusarium graminearum*. Although the fungus *Fusarium oxysporum* was also included, this species is actually a species complex, which comprises numerous *formae speciales* with different host specificities (Rowe, 1980; Gordon and Martyn, 1997).

Thus, much of our current knowledge of plant infecting fungi is derived from narrow host range pathosystems, which often exhibit particular characteristics resulting from a pronounced evolutionary arms race between host and pathogen. This arms race often results in pathogens evolving virulence genes that are counteracted by corresponding resistance loci in the host, which are themselves counteracted by new or variant virulence genes in the pathogen, *ad infinitum*. This is termed a "gene-for-gene" interaction, and was first observed by Flor in the interaction between the fungal parasite *Melampsora lini* and its host, flax (*Linum usitatissimum*) (Flor and Comstock, 1972).

Plant pathogenic fungi that coevolve with their hosts in a gene-for-gene manner often exhibit compartmentalized genomes that contain alternating nonrepetitive, gene-rich regions and repetitive, gene sparse regions. Virulence-associated genes are often clustered in repetitive, gene sparse regions, and it is thought that this clustering and compartmentalization allows for their rapid evolution through transposable element (TE)-mediated duplication and mutation, without affecting housekeeping genes. This theory is known as the "two-speed-genome" hypothesis (Haas et al. 2009; Raffaele et al. 2010; Ma et al. 2010; Rouxel et al. 2011; Raffaele and Kamoun, 2012; Croll and McDonald, 2012; Ohm et al. 2012; Dong et al. 2015; Faino et al. 2016).

Enhanced mutation rates in repeat-rich regions in fungi may arise through a process termed repeat-induced point mutation (RIP). This process is a genomic defense against excessive TE activity, which could be deleterious if left uninhibited. It causes point mutations in duplicated sequences through 5′-methylation of cytosine followed by deamination to thymine. As a result, RIP-affected genomic regions often exhibit pronounced A/T content concomitant with a decrease in G/C content (Selker and Stevens, 1985; Hane and Oliver, 2008; Smith et al. 2012). Virulence-associated proteins are often found to have been affected by this process, which may be important for their rapid adaptation (Rouxel et al. 2011).

Many of the virulence-associated genes of plant pathogenic fungi studied to date encode small secreted proteins, whose sole purpose is to manipulate plant defenses to advance fungal infection (Stergiopoulos and de Wit, 2009; Lo Presti et al. 2015). These proteins, termed "effectors," have been discovered in multiple plant pathogenic fungi and exhibit numerous different functions depending on fungal lifestyle. For example, necrotrophic fungi, which require dead tissue on which to feed, often produce effectors that promote cell death, whereas biotrophic fungi, which require living tissue, produce effectors that prevent cell death (Friesen, Faris, et al. 2008; Ciuffetti et al. 2010; de Jonge et al. 2010; Lu et al. 2015; Whigham et al. 2015; Lyu et al. 2016). Hemibiotrophic fungi, which require both living and dead tissue at different life cycle stages, may produce different effectors at different time points during infection (Marshall et al. 2011; Kleemann et al. 2012; Lee et al. 2013; Rudd et al. 2015).

Many effector genes studied are species or even isolate specific and interact directly or indirectly with the products generated by particular intraspecifically polymorphic loci in the host, thus conforming with the gene-for-gene hypothesis (Liu et al. 2006; Bourras et al. 2016; Bourras et al. 2015; Friesen et al. 2006; Friesen, Zhang, et al. 2008). However, several effectors are also conserved across various fungal species and have seemingly similar roles in pathogenesis on plants, interacting with loci conserved within a species and/or between species (Thatcher et al. 2012; de Jonge et al. 2010; Sanz-Martín et al. 2016; Redkar et al. 2015; Hemetsberger et al. 2015; Dallal Bashi et al. 2010; Bae et al. 2006).

In *S. sclerotiorum*, thus far, there have been several proteins described with effector-like properties and/or effector-like activities *in planta*. These proteins include a polygalacturonase enzyme, sspg1d, that interacts with a *B. napus* C2 domain containing protein (Wang et al. 2009); an integrin-like protein, SSITL, that suppresses defense responses in *Arabidopsis thaliana* (although, in addition to the activity of SSITL *in planta*, deletion mutants for the cognate gene exhibited abnormal hyphal tip branching, slower growth *in vitro* and abnormally small sclerotia) (Zhu et al. 2013); two necrosis and ethylene-inducing proteins, SsNep1 and SsNep2, that cause necrosis when heterologously expressed in *Nicotiana benthamiana* (Dallal Bashi et al. 2010); a cutinase, SsCut, that causes cell death in a range of plant species, including the nonhost, wheat (*Triticum aestevum*) (Zhang et al. 2014); a polygalacturonase protein that induces calcium signaling and host cell death in soybean (Zuppini et al. 2005); a hypothetical secreted protein that, when disrupted, attenuates virulence in *S. sclerotiorum* (Liang et al. 2013); a small, cysteine-rich secreted protein with a cyanoviron-N homology (CVNH) domain, which attenuates virulence when deleted (Lyu et al. 2015); and, most recently, a secreted protein with no known

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

functional domains, SsSSVP1, seemingly restricted to *S. sclerotiorum* and the closely related species *B. cinerea*, which causes necrosis in *N. benthamiana.* This protein was shown to interact with *N. benthamiana* QCR8, a subunit of the cytochrome b–c_1 complex of the mitochondrial respiratory chain. It is thought that this activity is what caused necrosis in host tissue (Lyu et al. 2016).

The starting point for investigations into the evolution and molecular activity of fungal effector genes often begins with analysis of the fungal genome. Numerous studies, including two studies on *S. sclerotiorum*, have attempted to identify effector-like or secreted proteins potentially involved in infection (e.g., Amaral et al. 2012; Nemri et al. 2014; Guyon et al. 2014; Heard et al. 2015). These studies have used various criteria to differentiate effector-like from noneffector-like genes. Primarily, the focus has been on small, cysteine-rich proteins with a predicted secretion signal, as many experimentally characterized effectors exhibit these properties (Sperschneider et al. 2015). Recently, a machine learning approach based on numerous criteria derived from properties of known effectors was applied to effector prediction. This approach was shown to be more sensitive and specific than the traditional, and to some degree arbitrary, approach of filtering through manually selected criteria (Sperschneider et al. 2016).

Predicted secretome and effector studies have been facilitated by the huge increase in availability of fungal genome sequences in recent years. However, although there are currently 1,509 fungal genomes that have been sequenced (as of May 26, 2016, source: http://www.ncbi.nlm.nih.gov/genome/browse/), many are fragmented and poorly annotated. The regions missing from these genomes are largely composed of repetitive sequence. This is because the technologies used to sequence them produce reads up to a maximum of 1,000 bp (Sanger) and in many cases only 300 bp (Illumina) (Goodwin et al. 2016). Sequence assemblies based on reads that are substantially shorter than repeated elements do not produce a reliable contiguous sequence as the repeats are often collapsed into a single, short sequence during graph-based assembly (Ekblom and Wolf, 2014). This poses a particular problem to researchers interested in fungal effectors, which, as aforementioned, often cluster within repetitive genomic regions (Thomma et al. 2016).

Recently, Pacific Biosciences developed the single molecule real-time sequencing (PacBio) platform, which produces reads of up to 60 kb (Goodwin et al. 2016). Thus far, two gapless fungal plant pathogen genomes have been assembled using PacBio reads, in combination with optical mapping—a process that uses fluorescent staining in conjunction with restriction enzymes to infer genomic structure based on restriction site patterns across chromosomes (Faino et al. 2015; van Kan et al. 2016).

The existing *S. sclerotiorum* genome, although of high quality compared with many other fungal genomes, is not complete, with an estimated 1.6 Mb of missing sequence, based on comparison of the assembly with an optical map (Amselem et al. 2011). To improve upon the previous genome, its annotations, and subsequent effector analyses, we used the existing optical map in combination with PacBio and Illumina sequencing. Using this improved genome sequence we identified a number of novel effector-like protein encoding genes, several of which were significantly upregulated during infection. We also recharacterized TEs in the *S. sclerotiorum* genome and demonstrated that the five most abundant TEs have been subjected to RIP and exhibit an increased RIP index relative to a randomized set of control sequences. Additionally, comparative analyses with the genomes of several host-specific filamentous plant pathogens revealed a lack of strong association between secreted or effector-like protein encoding genes and repetitive elements or RIP in the genome of *S. sclerotiorum*. We conclude that the broad host range necrotroph *S. sclerotiorum* does not exhibit a classic "two-speed genome" as found in host-specific biotrophic and hemibiotrophic fungi and oomycetes. Although a more subtle signature of potential enhanced selection of secreted proteins was observed, predicted effector-like protein encoding genes specifically were not found to be localized to repeat-rich or RIP-affected regions more than other genes.

Materials and Methods

Growth, nucleic acid extraction, and sequencing of *S. sclerotiorum* Isolate 1980

Various procedures were used to grow *S. sclerotiorum* and extract and sequence nucleic acids for the various analyses presented in this article. All experiments used WT 1980 or a mutant derivative thereof. Five data sets in total were used. Data set 1 consisted of one *in vitro* sample and six *B. napus* infection time points (1, 3, 6, 12, and 24 h postinoculation [hpi]) in triplicate; data set 2 consisted of several *in vitro* samples; and data set 3 consisted of one *in vitro* sample and two *A. thaliana* infection time points for both WT 1980 and the oxaloacetate acetyle hydrolase *S. sclerotiorum* deletion strain Δ*oah* (Liang et al. 2015). These three data sets were used for RNASeq-based gene calling. Data set 1 was also used for differential expression analysis of effector candidates. Data sets 4 and 5 were single *in vitro* samples and were used for Illumina genomic and PacBio sequencing, respectively. Experimental procedures used in the generation of all these data sets are detailed in the supplementary file 1, Supplementary Material online.

Assembly of PacBio Long Reads and Correction with Illumina Short Reads

A *de novo* genome assembly of *S. sclerotiorum* strain 1980 was generated using MHAP version 1.5b1 (Berlin et al. 2014) with default settings. To assess contiguity of the assembled sequences, they were aligned to the previously generated optical map with MapSolver version 3.2 (OpGen, Gaithersburg, MD). Overlapping or adjacent sequence contigs were manually joined and gap-filled with PBJelly2 version 15.2.20

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

(English et al. 2012) with default settings, except for the assembly stage where "-maxTrim" and "-maxWiggle" were set to 15 kb. A single region on chromosome 9 was not completely assembled *de novo* using MHAP, and this region was inferred from the previous *S. sclerotiorum* genome assembly (Amselem et al. 2011). The final assembly was error corrected after manual gap filling using Quiver (Chin et al. 2013).

Illumina short reads were first trimmed using Trimmomatic version 0.36 (Bolger et al. 2014) then mapped to the *de novo* assembly using Bowtie version 2.2.6 (Langmead et al. 2009). Modules from the Genome Analysis Toolkit version 3.5-0-g36282e4 (McKenna et al. 2010) and Picard Tools version 2.1.0 (available at http://broadinstitute.github.io/picard/) were then used to call variants between the *de novo* assembly and the Illumina reads.

After mapping to the version two genome sequence, GATK was used to create a consensus sequence from the Illumina reads and PacBio *de novo* assembly. Substitutions, of which there were very few, were ignored and only insertions and deletions (InDels) were considered. Polymorphisms that were identified based on a depth of less than 30× or a mapping quality of less than 30 were also excluded.

Alignment of RNA Sequencing Data for Use in Gene Predictions

RNASeq data obtained from the experimental conditions enumerated in the previous section were used for gene prediction. Reads were trimmed using Trimmomatic version 0.36 (Bolger et al. 2014). Reads from *in planta* samples were first binned based on whether they mapped better to the *S. sclerotiorum* genome or the host genome using BBSplit version 34.86 (available at https://sourceforge.net/projects/bbmap/files/). For this purpose, the *B. napus* genome was downloaded from http://www.genoscope.cns.fr/brassicanapus/ and the *A. thaliana* genome was downloaded from https://www.arabidopsis.org/ (Chalhoub et al. 2014; 2000).

Reads were then mapped to the *S. sclerotiorum* genome using Top Hat version 2.1.0 (Trapnell et al. 2012) with the following nondefault settings: "–min-intron-length 10" and "–max-intron-length 5000." Cufflinks version 2.2.1 (Trapnell et al. 2012) was used with default settings to assemble spliced transcripts from read mappings. Transcripts from all experimental conditions were merged for use in gene prediction using the cuffmerge module of Cufflinks.

Gene Prediction Using *Ab Initio* and RNASeq-Guided Softwares, Protein Homology, Transcript, and Expressed Sequence Tag Evidence

The following gene prediction software packages were used to generate predictions for the version two *S. sclerotiorum* genome: Augustus version 2.5.5 (Stanke and Morgenstern, 2005), Coding Quarry version 1.2 (Testa et al. 2015), and GeneMark-ES version 3.1 (Borodovsky and Lomsadze, 2011). These softwares were used with default settings and gene predictions were supplied to Evidence Modeler (Haas et al. 2008) alongside assembled transcripts, available expressed sequence tags (ESTs) and protein homology data, and weighted accordingly. A detailed description of methods used in gene calling is supplied in the supplementary file 2, Supplementary Material online. Genomes used for protein homology information are detailed in the supplementary table 1, Supplementary Material online.

Comparison of Version One Gene Predictions with Version Two Gene Predictions

To determine what genomic loci had changed or remained the same, two approaches were used: 1) Bedtools version 2.17.0 (Quinlan and Hall, 2010) was used to determine which new gene predictions overlapped annotations that had been transferred from the previous assembly to infer fusions or splits and 2) nucleotide sequence predictions from both genome assemblies were subjected to a BLAST analysis in both directions using BLASTn version 2.2.31 to determine differences in gene sequence length and content. Approach two was used to generate a table of version two loci and corresponding version one loci. The script used for the reciprocal BLAST analysis is available at: https://github.com/markcharder/GeneralBioinformaticsScripts/blob/master/reciprocalBestHitsBlast.pl.

Validation of Version Two Gene Annotations

To determine overall concordance with RNASeq data, CDS sequences from both genome versions' gene predictions were compared with Cufflinks transcripts (derived from the previously mentioned read mapping) using BLASTn. Overall concordance with RNASeq data was inferred from coverage of query sequence with the best BLAST hit from this analysis. Sequencing depth for each site in all exonic regions for both genome assemblies was calculated using Bedtools version 2.17.0. Both sets of gene predictions were also tested against the NCBI nonredundant (nr) database using BLASTp, excluding previous *S. sclerotiorum* gene models. Percent identity and query coverage of subject sequence based on best BLAST hits were used to infer quality of gene models.

Finally, a random set of 30 sequences selected from the 300 loci that differed the most between the two genome versions were manually reinspected to check which version's prediction most closely matched supporting evidence and why there might have been a discrepancy.

Protein Domain Prediction in the Version Two Annotations

Protein domains in amino acid sequences were predicted using Interpro Scan version 5.17-56.0 (Jones et al. 2014), which was also used to retrieve Gene Ontology (GO) (Ashburner et al. 2000) terms. GO terms were mapped to "GO-slim" terms using Goatools version 0.5.9 (available at https://libraries.io/github/tanghair/goatools).

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

Effector Prediction in the Version Two Annotations

Putative effector prediction was carried out using the following pipeline: 1) SignalP version 4.1 (Petersen et al. 2011) was used to identify proteins with secretion signals, 2) TMHMM version 2.0 (Krogh et al. 2001) was used to filter out those that contained transmembrane domains, 3) GPIsom (available at http://gpi.unibe.ch/) (Fankhauser and Mäser, 2005) was used to filter out proteins that harbored a putative glycophosphatidylinositol membrane-anchoring domain, and 4) EffectorP version 1.0 (Sperschneider et al. 2016) was used to predict potential effector sequences among those remaining after the previous filtering steps.

Analysis of Differential Expression of Putative Effectors

A subset of the RNASeq data, data set 1, that were used for gene prediction were aligned to the version two assembly using the previously described methods. This subset included samples taken from the following conditions: *In vitro*, 1, 3, 6, 12, 24, and 48 hpi of *B. napus* with *S. sclerotiorum* strain 1980. Samples were replicated three times as per previously mentioned. Aligned reads were then used in conjunction with the version two gene annotations to determine differential expression using the cuffdiff module of Cufflinks (Trapnell et al. 2012). All sequences that exhibited homology to ribosome-related sequences were specified to cuffdiff at run-time. Fold-change in expression was considered significantly different if it was at a *P*-adjusted value of below 0.05 indicating a false discovery rate of below 0.05.

Identification of Putative Effector Families and Their Association with Segmental Duplications and TEs

To identify repetitive sequences in the version two *S. sclerotiorum* genome, the REPET pipeline (Quesneville et al. 2005) was run with default settings. *De novo* repeat annotations produced by REPET were used in all downstream analyses. TEs were divided into nine subclasses (automatically by REPET) based on the nomenclature system devised by (Wicker et al. 2007). These subclasses included noCat, DXX, RIX, RXX, RLX DTX, RPX, XXX and RYX, corresponding to no category, unclassified DNA element, unclassified long interspersed nuclear element, unclassified retroelement, long terminal repeat (LTR) element, DNA terminal inverted repeat (TIR) element, Penelope retroelement, unclassified TE, and DIRS-like element, respectively.

To identify paralogous effectors in *S. sclerotiorum*, OrthoMCL version 2.0.9 (Li et al. 2003) was used with default settings. The *Botrytis cinerea* reference genome version 3.0 (van Kan et al. 2016) was downloaded from the ENSEMBL Fungi database (http://fungi.ensembl.org/Botrytis_cinerea/Info/Index) for this purpose (Pedro et al. 2016).

For effector sequences that were determined to be recent paralogs by OrthoMCL, regions of 4–5 kb encompassing the gene models were extracted and aligned using ClustalW version 2.1 (Thompson et al. 1994). Up to 20-kb regions surrounding paralogous genes were manually inspected for the presence of LTRs using LTRharvest (Ellinghaus et al. 2008), TEs predicted by REPET, and overlapping gene models containing typical retro-element Pfam domains (based on the previously mentioned InterProScan analysis). Consensus TE sequences were used to search the Dfam database version 2.0 (Wheeler et al. 2013) to identify homologous TE families.

Assessment of RIP in the Most Abundant Repeats and at the Whole-Genome Scale in *S. sclerotiorum*

Before identifying RIP in repeat families, the whole genomes of *S. sclerotiorum* and the fungal and oomycete species *Blumeria graminis* f. sp. *hordei* (herein referred to as *B. graminis*), *Leptosphaeria maculans*, and *Phytophthora infestans* were scanned to determine whether there was a bimodal GC content. This was done using OcculterCut version 1.1 with default settings. The additional genome sequences were used to provide a context for genome bimodality against which *S. sclerotiorum* was compared. These additional genomes were downloaded from GenBank.

After identifying repeat families *de novo* using REPET, the five TE sequences with the highest copy number were subjected to analyses to identify RIP. For each repeat, only the 50 longest sequences were used. First, the most numerous repeat was subjected to analysis using RipCal version 1.0 (Hane and Oliver, 2008) through the alignment method to show dominant forms of RIP. Second, the ApT/TpA RIP index was calculated for each repeat family member and a random set of sequences from the *S. sclerotiorum* genome of the same size and lengths. Student's *t*-tests were used to determine whether the repeat sequences had a significantly higher ApT/TpA index than an equivalent random set of sequences. Random sequences were produced and RIP indices were calculated using a custom Perl script available at: https://github.com/markcharder/GeneralBioinformaticsScripts/blob/master/getRandom.pl.

Comparative Genomic Analysis of Association of Secreted and Effector Proteins with Repeat-Rich, Gene Sparse Genomic Regions and RIP-Affected Sequence

To identify secreted and effector-like proteins, the same secreted protein and effector discovery pipelines were run on the genomes of the three fungi *S. sclerotiorum, B. graminis* and *L. maculans*, and the oomycete *P. infestans*. The three additional species were chosen as their genomes have been well characterized and exhibit typical features described in the "two-speed" genome hypothesis (Rouxel et al. 2011; Haas et al. 2009; Amselem, Lebrun, et al. 2015). Genomes used in this analysis, apart from *S. sclerotiorum*, were downloaded from GenBank.

Before predicting secreted proteins, all gene sequences were filtered based on homology to TEs through blasting

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

against the RepBase database version 20.05 (Bao et al. 2015). Those sequences with more than 30% amino acid identity and alignment coverage and an *e*-value of $<1e^{-10}$ with subjects from RepBase were discarded. This was to ensure that TE genes were not included in downstream analysis, as this would affect CDS content and proportion of secreted proteins in repeat-rich regions.

The secreted protein prediction pipeline followed the same procedure as detailed in the "Effector prediction in the version two annotations" above but used Pfam instead of GPI-som to predict GPI-anchoring domains. To identify a subset of the secreted proteins that were potential effectors, EffectorP was run. The whole pipeline for secreted protein discovery is available at: https://github.com/markcharder/GeneralBioinformaticsScripts/blob/master/effector_pipeline.sh. Repeats in all of these genomes were identified for the purpose of this analysis using RepeatMasker; simple sequence repeats were discarded. To identify potentially RIP-affected sequence, RipCal version 1.0 was run on whole genomes with default settings to identify regions with a high RIP index.

To determine whether there was an association between the presence of secreted and effector-like proteins and repeats, RIP-affected sequence and gene-sparse sequence in the four genomes, two approaches were taken. First, for each set of secreted proteins, each set of effector-like proteins and each total gene set, the "closest" module of Bedtools was used to determine the distance from nearest repeats and RIP-affected sequence. Then, from the distances generated for the total set of genes, random sets the same size as the secreted protein sets and the effector-like sets were generated. The secreted protein sets and effector-like sets were compared against the random sets using Wilcoxon's test in R version 3.3.0 beta. The script used for this analysis is available at: https://github.com/markcharder/GeneralBioinformaticsScripts/blob/master/testingAssociations.r.

Second, percentage of bases covered by CDSs, GC content, and number of genes and number of secreted proteins were determined across a sliding window of 100 kb for each genome, incrementing by 1 kb. The R scripts used for this analysis are available at: Https://Github.Com/Markcharder/GeneralBioinformaticsScripts/Blob/Master/Test_Two-Speed.r.

Spearman's rank was used to assess correlation between proportion of secreted protein content (defined as number of secreted proteins/total number of genes) per sliding window with CDS and GC content. GC content was used to infer the likelihood of extensive RIP in sliding windows.

Results

The *S. sclerotiorum* Version Two Assembly Is Near Complete, with a Single Gap in the Nucleolus Organizer Region

To produce a more complete and accurate genome for *S. sclerotiorum*, PacBio reads were assembled *de novo* and scaffolded using the previously generated optical map and a portion of the Sanger assembly (Amselem et al. 2011). To further improve accuracy, Illumina reads that mapped to the new assembly were used to create a consensus sequence. The final assembly was based on 36× coverage of PacBio reads and consisted of 38,806,497 bp distributed across 16 scaffolds, representing the 16 chromosomes predicted for the *S. sclerotiorum* strain 1980 (Amselem et al. 2011). This is an improvement in both sequence content—an additional 805,146 bp have been sequenced—and contiguity, as the version one assembly contained 38,001,451 bp (excluding Ns) distributed across 36 scaffolds that contained numerous gaps (fig. 1; supplementary fig. 1*a* and table 2, Supplementary Material online). In the version two assembly, a single gap of unknown size was present within the nucleolus organizing region (identified through homology of repeating sequence to ribosomal DNA of other fungi; data not shown). This gap was given an arbitrary size of 100 Ns; this is contrasted to the 328,761 N-bases of the version one assembly (supplementary table 3 and fig. 1*a*, Supplementary Material online; fig. 1). The new version of the S. sclerotiorum genome and its annotations are deposited in GenBank under bioproject number PRJNA348385.

Mapping of Illumina reads to the version two assembly scaffolds revealed 653 insertions (In), 222 deletions (Del), and 8 substitutions (supplementary table 3, Supplementary Material online); all 875 InDels were subsequently corrected. The version two assembly was congruent with the version one assembly (fig. 1; supplementary fig. 1*b*, Supplementary Material online); although upon mapping Illumina reads to the version one assembly, it became apparent that there were more discrepancies between the Illumina reads and this assembly than the version two assembly, totaling 751 insertions, 1,828 deletions, and 1,764 substitutions (supplementary table 3, Supplementary Material online).

Gene Predictions in the *S. sclerotiorum* Version Two Genome Assembly Exhibit Significant Divergence from the Version One Predictions

To determine how individual gene models had changed from version one to version two, a reciprocal best hits BLAST analysis in conjunction with Bedtools-based analyses were conducted. The Bedtools analyses were used to determine fusion and splitting events between the version two predictions and the transferred version one predictions. The BLAST analysis was used to identify version one loci corresponding to version two loci.

The total number of gene models in the version two assembly is 11,130. This is 3,392 less than the previous total of 14,522, and 730 less than the previous "nondubious" total of 11,860. Furthermore, a total of 435 predictions in the version two assembly did not have a one-to-one BLAST correspondence with any version one loci. From version one to version two, 1,301 loci were fused to neighboring loci. A total of 279 version one loci were split to form separate loci in the version two assembly. The mean length of version two models was

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

Fig. 1.—Circos plot depicting features of the new *S. sclerotiorum* assembly and positions of corresponding version one contigs. From outer to inner circular tracks: The 16 version two assembly (Ss1980 version 2) contigs colored to distinguish neighboring contigs; the contigs from the version one assembly (Ss1980 version 1) colored to distinguish neighboring contigs; circles depicting positions of secreted proteins and SsPEs—clear circles = secreted proteins, colored circles = SsPEs; log fold change in expression of secreted proteins from *in vitro* to each of the *B. napus* time points tested, from blue (downregulated) to red (upregulated); regions with a high RIP index as predicted by RipCal; repeat sequences from REPET *de novo* annotation of more than 5 kb. The lines crossing the center of the Circos plot join sequences of the same repeat element.

1,439.88 bp, whereas the mean length of version one models was 1,088.47 bp. Based on the reciprocal BLAST analysis, 6,891 sequences were identical at the nucleotide level (supplementary table 4, Supplementary Material online).

Gene Predictions in the Version Two Genome Assembly Are More Accurate than Version One

To assess the accuracy of the new gene models, both the version one predictions and the version two predictions

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

were 1) tested against the NCBI nr database, 2) scanned for protein domains using InterProScan, 3) assessed for congruence with RNASeq data by comparison to Cufflinks transcripts and determining RNASeq depth for all CDS regions, and 4) manually reinspected randomly to determine which version's models were best supported by accompanying evidence.

Based on the BLAST analysis against the NCBI nr database, the version two predictions exhibited a higher mean percent identity (86.7%) with their best BLAST hits than the version one predictions (81.78%). Furthermore, the version two sequences exhibited a higher mean subject coverage per alignment with their best BLAST hits (90.39%) than the version one sequences (84.9%) (fig. 2*a*). The InterProScan analysis showed that the version two sequences contained more predicted protein domains than the version one sequences (fig. 2*b*; supplementary table 5, Supplementary Material online).

Comparing CDS sequences from both sets of gene predictions with assembled Cufflinks transcripts demonstrated that 10,523 version two sequences had RNASeq-based transcript support. A total of 7,506 of these were fully supported, that is, gene predictions were fully covered by an assembled Cufflinks transcript which was 100% identical. Furthermore, the version two sequences were more concordant with RNASeq data than the version one sequences. Mean query coverage per high scoring segment pair (HSP) for best BLAST hits was 91.88% for the version two sequences and 83.53% for the version one sequences, whereas median coverage per HSP was moderately increased to 100% in the version two sequences from 99% in the version one sequences. However, interquartile range for the version two sequences was 1, whereas interquartile range for the version one sequences was 32. Additionally, mean RNASeq coverage for all CDS regions of the version two predictions was 535.187×, whereas for the version one predictions it was 456.359×, median values were 132 and 124, respectively (fig. 2*c* and *d*).

Thirty of the 300 most divergent sequences from version one to two were randomly selected. Reinspection of these genes (after initial manual curation in a previous step) confirmed that the version two models were more in accordance with supporting evidence than version one models (supplementary fig. 2, Supplementary Material online).

TE Content Is Increased in the *S. sclerotiorum* Version Two Genome Assembly

To assess the amount of repetitive sequence in the new *S. sclerotiorum* genome the REPET pipeline (Quesneville et al. 2005) was run. This demonstrated that the version one assembly contained 4,329,439 bp (11.39%) of repetitive sequence, whereas the version two assembly contained 5,041,697 bp (12.96%); this is an increase of 712,258 bp (1.57%). As a proportion of the total additional sequence content (905,046 bp), this is 78.7%; thus, most of the additional sequence was repetitive.

Both class I (retro) elements and class II (DNA) elements were identified in the version two genome. The first and second-most abundant TEs were unclassified retroelements and DNA TIR elements, respectively. Overall proportions of TE subclasses did not markedly differ between the two genome versions, though the total number of bases within TEs did (fig. 3).

Seventy Putative Effector Sequences, 61 of Which Have Not Been Reported Before, Were Predicted for *S. sclerotiorum*

To determine whether the new genome contained previously undiscovered effector-like sequences, version two-predicted proteins were filtered based on various criteria, including identification using the EffectorP software (Sperschneider et al. 2016). Based on these analyses, the version two *S. sclerotiorum* genome contains 900 predicted proteins with a predicted secretion signal; 695 of these lack a predicted transmembrane domain, of which 523 lack both a predicted transmembrane domain and a predicted GPI-anchoring domain. Of these 523 amino acid sequences, 70 were determined to be likely effectors based on EffectorP analysis; the mean amino acid length of these sequences was 176.543. These will henceforth be referred to as "*S. sclerotiorum* putative effectors" (SsPEs) (fig. 1; table 1).

Of the 70 SsPEs, 13 are different in length to their corresponding loci in the version one assembly. The most divergent sequence was 4.95 times longer than the previous model (table 1). An additional three did not correspond to genes predicted in the version one assembly. The remaining 54 SsPEs were identical to their corresponding loci in the version one assembly. Of the 70 SsPE sequences, 61 were not identified as likely effectors by Guyon et al. (2014) who used a different effector-prediction pipeline that did not include EffectorP analysis.

Of the 70 SsPEs, 63 had homologs (*e*-value $< e^{-10}$) in other species of fungi, 22 of which contained predicted protein domains based on an InterProScan analysis. A total of 48 SsPEs had no predicted protein domains, seven of which were unique to *S. sclerotiorum* (table 1).

RNASeq Analysis Demonstrates Significant Differential Expression of SsPEs during Infection of *B. napus* Relative to during Growth *In Vitro*

To determine whether SsPEs were significantly upregulated *in planta* relative to during growth *in vitro*, an RNASeq time course consisting of samples encompassing *in vitro* growth, 1, 3, 6, 12, 24, and 48 hpi of *B. napus* was analyzed.

In total, 66 SsPEs were expressed under at least one condition tested (figs. 1 and 4*a*). Of these 66, 19 were significantly differentially expressed at at least one *in planta* time point relative to during growth *in vitro*. Of these 19 gene models, nine were significantly upregulated during infection relative to

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

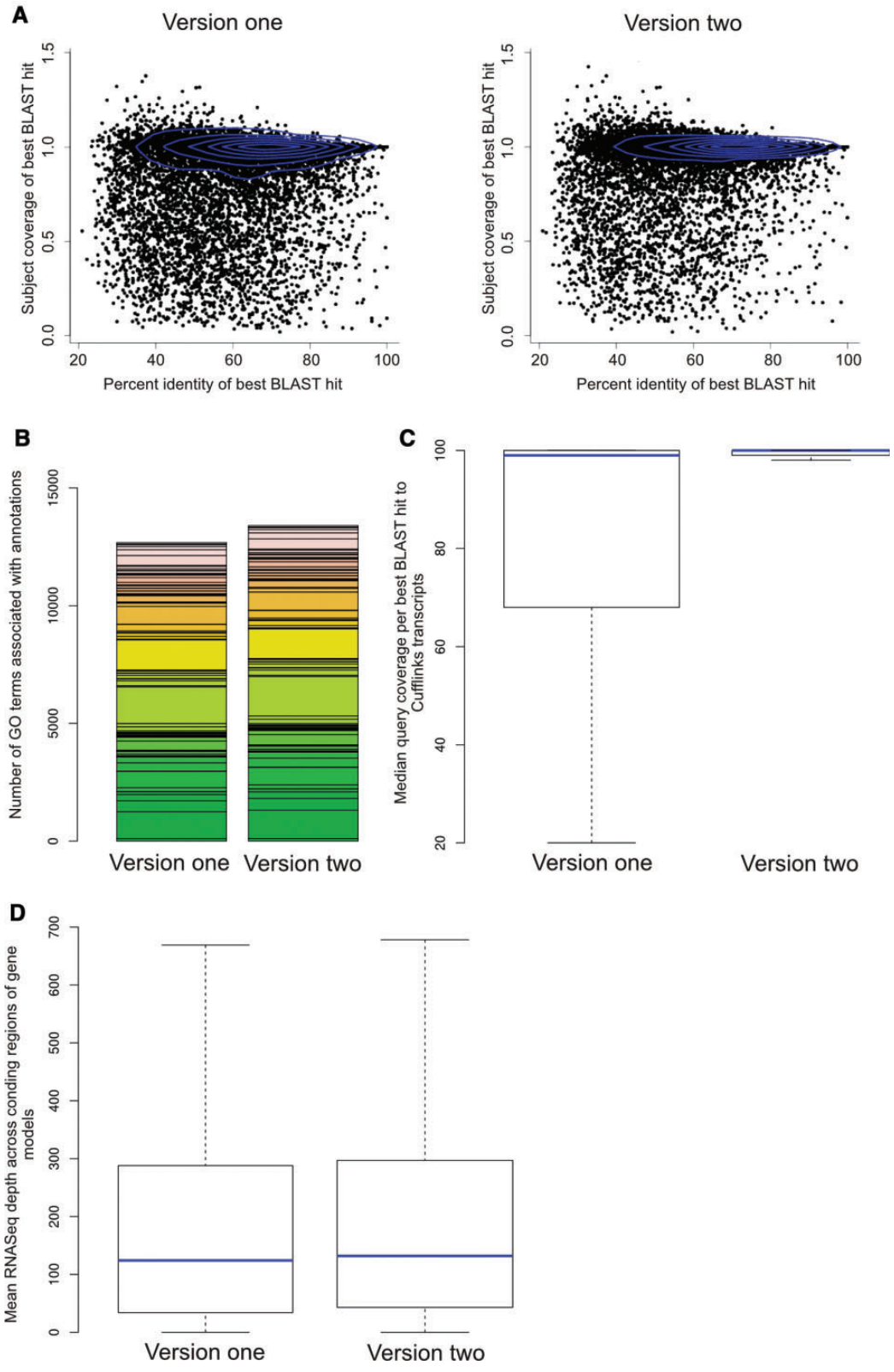

Fig. 2.—BLAST and RNA sequencing validation of version two annotations. (*a*) Alignments of best BLASTp hits of gene annotations were higher quality for version two annotations than version one. On the *x* axis is the percent identity of the best BLAST hits and on the *y* axis is the coverage of the subject sequence by the query sequence. Blue contour lines represent kernel density. (*b*) More GO slim terms were present in version two gene predictions than

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

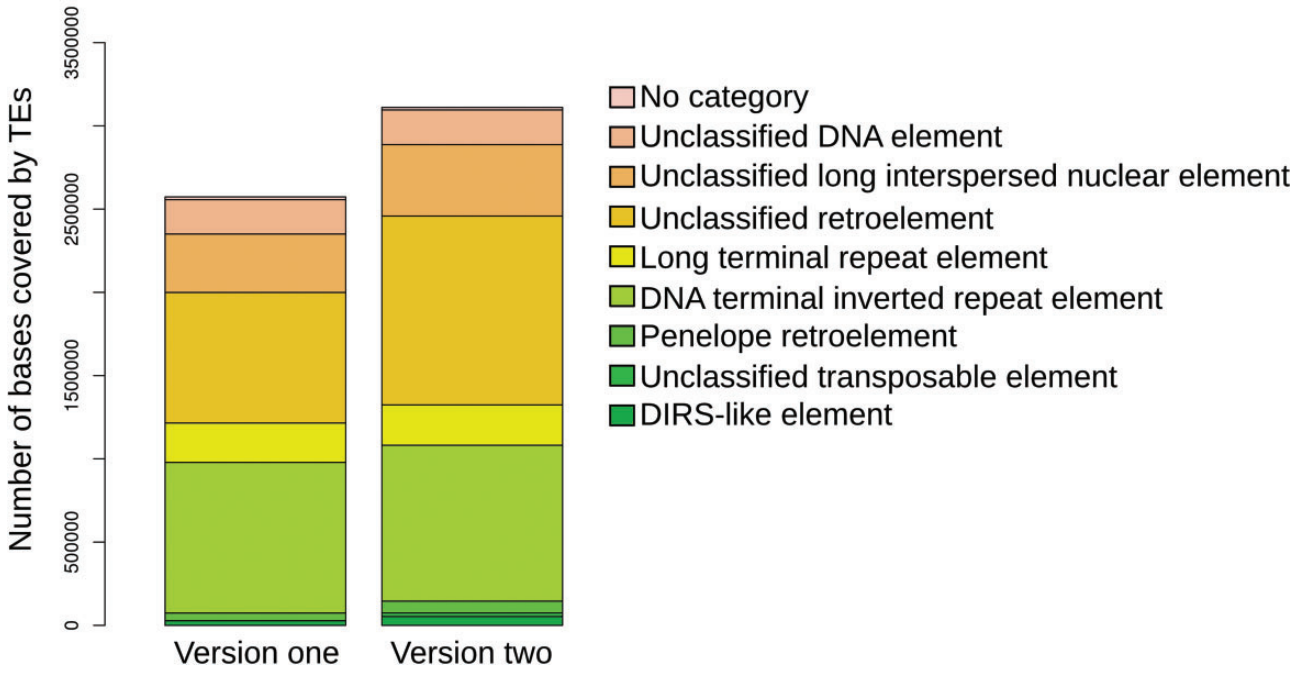

Fig. 3.—Transposable element content of the version one and version two assemblies. On the *y* axis is the number of bases covered by TEs in each family, represented by different colored blocks in the stacked histogram for both genome versions. The figure is based on *de novo* prediction by REPET.

during growth *in vitro* at at least one time point and exhibited a log fold change of at least 1 (fig. 3*b* and *c*). Of these gene models, five were significantly upregulated at 1, 3, 6, and/or 12 hpi relative to during growth *in vitro*; three of which were also upregulated at 24 and/or 48 hpi (fig. 3*b* and *c*). A further four gene models were exclusively upregulated at later time points, 24 and/or 48 hpi (fig. 3*c*).

Seven Putative *S. sclerotiorum* Effectors Are Paralogous and Associated with Potentially Recent Duplication Events

OrthoMCL was used to determine which SsPEs were likely to be recent paralogs (i.e., generated after the divergence of *B. cinerea* and *S. sclerotiorum*). This identified seven SsPEs grouped into a single family, with a single homolog in *B. cinerea*. None of these sequences contained predicted functional domains based on InterProScan analysis, though they were broadly conserved throughout fungi, matching numerous models without functional domain predictions (supplementary table 6, Supplementary Material online).

Six of these proteins were previously identified by Guyon et al. (Guyon et al. 2014), who demonstrated a close proximity (1.1 kb) to a Tad1 retroelement of one of them. To expand on these analyses, and to determine whether these sequences were embedded within regions that might have resulted from relatively recent segmental duplications, 4–5 kb regions spanning SsPEs were extracted and aligned; this created an alignment of 4,869 bp. These regions were homologous and exhibited a higher overall percentage identity across the multiple alignment in the region immediately surrounding and spanning the SsPEs. Four of the SsPE-containing loci (those containing SsPE models Sscle13g097000, Sscle16g111300, Sscle06g055280, and Sscle07g061960) were more than 90% identical across an approximately 3,000-bp region immediately surrounding and spanning the aligned SsPEs. Additionally, similarly sized regions surrounding SsPE models Sscle09g074030 and Sscle16g107730 exhibited a pairwise identity of more than 90% but were divergent from the previously mentioned SsPE regions (fig. 5*a*).

Fig. 2.—Continued

version one. Each GO slim term identified is plotted as a separate colored block in the stacked histogram for each version's annotations. On the *y* axis is the number of GO slim numbers associated with each annotation set for each GO slim term. (*c*) Version two annotations were more concordant with RNASeq data, based on comparison to assembled Cufflinks transcripts. The *y* axis represents percent coverage of the gene annotation CDS with best BLAST hit. Blue lines represent median values, boxes and whiskers represent second and third quartiles, and interquartile range, respectively. (*d*) Version two gene annotations showed a greater depth of RNASeq coverage than version one annotations. Depth of RNASeq alignment across exonic regions is plotted on the *y* axis. Blue lines represent median values and boxes and whiskers represent second and third quartiles, and interquartile range, respectively.

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

Table 1

Predicted Effectors in the Version Two *Sclerotinia sclerotiorum* Genome (SsPEs) and Their Corresponding Version One Loci

Version One ID	Version Two ID	Percent Identity	Version One Length	Version Two Length	Guyon et al. Prediction?	Version Two Length/Version One Length	Domains Predicted by InterProScan (prediction database|domain identifier|description|)	% AA Identity of Best Blast Hit	*E*-Value of Best Blast Hit
SS1G_09693	Sscle01g000660	100	459	459		1	|||	57.534	$2.34e^{-050}$
SS1G_02068	Sscle01g003850	100	565	565		1	|||	55.621	$2.7e^{-056}$
SS1G_01974	Sscle01g004620	100	465	558		1.2	|||	86.607	$5.13e^{-052}$
SS1G_01867	Sscle01g005390	100	490	490		1	|||	90.741	$3.46e^{-059}$
SS1G_01754	Sscle01g006330	100	2,266	532		0.234774934	|||	65.035	$9.79e^{-055}$
SS1G_01427	Sscle01g008940	100	1,512	1,512		1	Pfam|PF00024|PAN domain|	69.421	$3.37e^{-168}$
SS1G_01426	Sscle01g008950	100	1,079	1,079		1	|||	82.158	$2.12e^{-119}$
SS1G_01287	Sscle01g009960	100	624	624		1	|||	29.353	$3.44e^{-011}$
SS1G_01214	Sscle01g010490	100	404	404		1	Pfam;SUPERFAMILY;Gene3D|PF06766;SSF101751; G3DSA:3.20.120.10|Fungal hydrophobin;;|	63.265	$9.41e^{-039}$
SS1G_12778	Sscle02g012940	100	282	282		1	|||	—	—
SS1G_04766	Sscle02g014280	100	881	348		0.395005675	Pfam;SUPERFAMILY;Gene3D|PF00190;SSF51182; G3DSA:2.60.120.10|Cupin;;|	84.404	$2.38e^{-059}$
SS1G_04618	Sscle02g015390	100	468	468		1	|||	72.611	$1.13e^{-079}$
SS1G_04155	Sscle02g019000	100	1,139	1,139		1	|||	77.551	$1.96e^{-149}$
SS1G_13012	Sscle02g021780	100	755	838		1.109933775	ProSiteProfiles;Gene3D;SUPERFAMILY; Pfam|PS50059;G3DSA:3.10.50.40;SSF54534; PF00254|FKBP-type peptidyl-prolyl *cis-trans* isomerase domain profile.;;;FKBP-type peptidyl-prolyl *cis-trans* isomerase|	90.909	$3.64e^{-112}$
SS1G_01032	Sscle03g022550	100	1,178	1,178		1	|||	79.894	0
SS1G_01003	Sscle03g022790	100	450	450		1	Gene3D;Pfam|G3DSA:3.20.120.10;PF06766|;Fungal hydrophobin|	34.653	$1.63e^{-012}$
SS1G_13371	Sscle03g031910	100	880	880	x	1	|||	61.404	$7.12e^{-083}$
SS1G_02522	Sscle04g035160	100	897	897		1	|||	87.9	$5.62e^{-169}$
SS1G_02700	Sscle04g036550	100	276	610		2.210144928	|||	70.27	$7.9e^{-081}$
SS1G_03057	Sscle04g039210	100	484	484		1	Pfam;Gene3D|PF03966;G3DSA:2.20.25.10| Trm112p-like protein;|	96.85	$8.66e^{-083}$
SS1G_03080	Sscle04g039420	100	852	852	x	1	Pfam;PIRSF;Coils|PF05630;PIRSF029958;Coil|Necrosis-inducing protein (NPP1);;|	82.114	$5.47e^{-147}$
SS1G_14237	Sscle04g040080	100	1,186	1,186		1	Pfam|PF11327|Protein of unknown function (DUF3129)|	82.4	$9e^{-144}$
SS1G_12123	Sscle05g041050	100	1,365	620		0.454212454	|||	69.565	$4.06e^{-075}$
SS1G_06068	Sscle05g045060	100	352	352		1	ProSiteProfiles|PS51257|Prokaryotic membrane lipoprotein lipid attachment site profile.|	67.742	$2.5e^{-035}$
SS1G_05939	Sscle05g046060	100	264	264		1	ProSiteProfiles|PS51257|Prokaryotic membrane lipoprotein lipid attachment site profile.|	74.713	$4.11e^{-041}$
SS1G_05938	Sscle05g046070	100	639	639	x	1	|||	58.772	$1.7e^{-045}$

(continued)

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

Table 1 Continued

Version One ID	Version Two ID	Percent Identity	Version One Length	Version Two Length	Guyon et al. Prediction?	Version Two Length/Version One Length	Domains Predicted by InterProScan (prediction database|domain identifier|description|)	% AA Identity of Best Blast Hit	*E*-Value of Best Blast Hit
SS1G_07491	Sscle06g048920	100	709	526		0.741889986	|||	75.51	$5.86e^{-078}$
SS1G_07320	Sscle06g050100	100	930	930		1	SUPERFAMILY;TIGRFAM;PRINTS;PRINTS;PRINTS;PRINTS; PRINTS;Gene3D;SMART;ProSiteProfiles;Pfam; SMART|SSF52540;TIGR00231;PR00328;PR00328; PR00328;PR00328;PR00328;G3DSA:3.40.50.300;SM00178; PS51422;PF00025;SM00177|;small_GTP: small GTP-binding protein domain;GTP-binding SAR1 protein signature;GTP-binding SAR1 protein signature; GTP-binding SAR1 protein signature;GTP-binding SAR1 protein signature;GTP-binding SAR1 protein signature;;Sar1p-like members of the Ras-family of small GTPases;small GTPase SAR1 family profile.;ADP-ribosylation factor family;ARF-like small GTPases; ARF, ADP-ribosylation factor|	99.471	$8.88e^{-135}$
SS1G_07230	Sscle06g050820	100	298	298		1	|||	36.364	2.1
SS1G_07027	Sscle06g052360	100	710	710		1	|||	87.931	$1.17e^{-109}$
SS1G_12482	Sscle06g054400	100	797	797		1	|||	73.82	$1.5e^{-114}$
SS1G_12431	Sscle06g054810	100	939	616		0.656017039	|||	45.37	$3.47e^{-022}$
SS1G_12365	Sscle06g055280	100	843	843	+	1	|||	67.606	$1.84e^{-096}$
SS1G_03381	Sscle07g057000	100	1,841	2,124		1.153720804	|||	61.029	$1.69e^{-043}$
SS1G_11673	Sscle07g061770	100	506	506		1	SUPERFAMILY;ProSiteProfiles;Gene3D|SSF57414;PS50948; G3DSA:3.50.4.10|;PAN/Apple domain profile.;|	65.625	$3.07e^{-057}$
SS1G_11693	Sscle07g061960	100	850	850	+	1	|||	67.606	$1.84e^{-096}$
SS1G_11706	Sscle07g062060	100	297	614		2.067340067	|||	82.639	$4.75e^{-082}$
SS1G_05103	Sscle08g064180	100	520	520		1	Coils|Coil|	66.99	$1.51e^{-042}$
SS1G_05152	Sscle08g064590	100	529	529		1	|||	74.051	$2.25e^{-055}$
SS1G_05569	Sscle08g067710	100	498	498		1	|||	46.914	$4e^{-035}$
SS1G_14184	Sscle08g068200	100	1,560	1,350	x	0.865384615	ProSitePatterns;SMART;ProSiteProfiles;ProSitePatterns; SUPERFAMILY;SUPERFAMILY;ProSitePatterns;ProSiteProfiles; SUPERFAMILY;ProSiteProfiles;Gene3D;Gene3D;SMART; SMART;SMART;Pfam;Gene3D;Gene3D;ProSitePatterns; ProDom;Pfam;SUPERFAMILY|PS00026;SM00236; PS51164;PS00026;SSF57016;SSF57016;PS00562;PS50941; SSF57016;PS50941;G3DSA:3.30.60.10;G3DSA:3.30.60.10; SM00270;SM00270;SM00270;PF00187;G3DSA:3.30.60.10; G3DSA:3.30.60.10;PS00026;PD001821;PF00734; SSF57180|Chitin recognition or binding domain signature.;Fungal-type cellulose-binding domain;CBM1	34.759	$4.64e^{-014}$

(continued)

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

Table 1 Continued

Version One ID	Version Two ID	Percent Identity	Version One Length	Version Two Length	Guyon et al. Prediction?	Version Two Length/Version One Length	Domains Predicted by InterProScan (prediction database|domain identifier|description|)	% AA Identity of Best Blast Hit	*E*-Value of Best Blast Hit
							(carbohydrate binding type-1) domain profile.;Chitin recognition or binding domain signature.;;;CBM1 (carbohydrate binding type-1) domain signature.;Chitin-binding type-1 domain profile.;;Chitin-binding type-1 domain profile.;;;Chitin binding domain;Chitin binding domain;Chitin binding domain;Chitin recognition protein;;;Chitin recognition or binding domain signature.;DEGRADATION HYDROLASE GLYCOSIDASE CELLULOSE METABOLISM CARBOHYDRATE POLYSACCHARIDE PRECURSOR SIGNAL I;Fungal cellulose binding domain;|		
SS1G_10892	Sscle09g069090	100	576	576		1	Pfam|PF12296|Hydrophobic surface binding protein A|	38.15	$1.68e^{-028}$
SS1G_03897	Sscle09g072630	100	481	481		1	|||	64.384	$9.4e^{-025}$
SS1G_03721	Sscle09g074030	100	869	869	+	1	|||	61.404	$3.39e^{-084}$
SS1G_08128	Sscle10g074920	100	444	444		1	|||	83.81	$3.41e^{-047}$
SS1G_08163	Sscle10g075140	99.49	393	395		1.005089059	|||	66.667	$1.75e^{-033}$
—	Sscle10g076600	—	—	—		—	Pfam;SUPERFAMILY|PF07249;SSF50685|Cerato-platanin;|	87.417	$3.34e^{-085}$
SS1G_13851	Sscle10g080580	100	441	275		0.623582766	|||	49.231	$4.2e^{-010}$
SS1G_08088	Sscle11g081020	100	159	159		1	|||	—	—
SS1G_07837	Sscle11g082970	100	1,012	1,012		1	|||	50.968	$2.64e^{-051}$
SS1G_07613	Sscle11g084720	100	628	628		1	Pfam;Gene3D;SUPERFAMILY;SMART|PF02221; G3DSA:2.70.220.10;SSF81296;SM00737|ML domain;;;Domain involved in innate immunity and lipid metabolism.|	86.628	$3.87e^{-107}$
SS1G_11151	Sscle12g087960	100	471	471		1	|||	33.758	$5.21e^{-017}$
SS1G_11085	Sscle12g088530	100	772	772		1	|||	47.573	$1.37e^{-055}$
SS1G_11065	Sscle12g088660	100	636	636		1	|||	40.385	3.4
SS1G_11928	Sscle12g090380	100	159	159		1	|||	—	—
—	Sscle13g094760	—	—	—		—	Pfam|PF10270|Membrane magnesium transporter|	89.041	$1.21e^{-090}$
SS1G_06729	Sscle13g094920	100	802	802		1	Pfam;ProSiteProfiles;Coils;SUPERFAMILY|PF01105;PS50866; Coil;SSF101576|emp24/gp25L/p24 family/GOLD;GOLD domain profile.;;|	96.552	$5.94e^{-145}$
SS1G_06763	Sscle13g095230	100	709	709		1	|||	44.624	$4.16e^{-039}$
SS1G_14379	Sscle13g097000	100	843	843	+	1	|||	68.075	$3.96e^{-097}$
SS1G_08669	Sscle14g098710	100	213	1,055		4.953051643	Gene3D;SUPERFAMILY;Pfam|G3DSA:1.10.1280.10;SSF48056; PF00264|;;Common central domain of tyrosinase|	68.705	$7.45e^{-099}$
SS1G_08706	Sscle14g098920	100	405	405		1	SUPERFAMILY;Gene3D;Pfam|SSF50685;G3DSA:2.40.40.10; PF03330|;;Rare lipoprotein A (RlpA)-like double-psi beta-barrel|	88.806	$6.08e^{-081}$
SS1G_08892	Sscle14g100310	100	908	908		1	|||	65.845	$8.4e^{-123}$

(continued)

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

Table 1 Continued

Version One ID	Version Two ID	Percent Identity	Version One Length	Version Two Length	Guyon et al. Prediction?	Version Two Length/Version One Length		Domains Predicted by InterProScan (prediction database|domain identifier|description|)	% AA Identity of Best Blast Hit	*E*-Value of Best Blast Hit
SS1G_13394	Sscle15g102390	100	587	587		1	=		58.757	$2.72e^{-059}$
SS1G_09420	Sscle15g105160	100	528	528		1	=		77.586	$1.6e^{-095}$
SS1G_09288	Sscle15g106080	100	594	594		1	=		33.333	$2.32e^{-019}$
SS1G_09150	Sscle15g107190	100	648	648		1	=		36.782	0.001
SS1G_10104	Sscle16g107730	100	880	880	+	1	=		61.842	$3.08e^{-084}$
SS1G_10129	Sscle16g107890	100	333	333		1	=		46.032	0.026
SS1G_14320	Sscle16g111080	100	483	483		1	=	Pfam;ProSiteProfiles;Gene3D;SMART;SUPERFAMILY|PF01817; PS51168;G3DSA:1.20.59.10;SM00830;SSF48600|Chorismate mutase type II;Chorismate mutase domain profile.;Chorismate mutase type II;|	58.904	$4.27e^{-048}$
—	Sscle16g111300	—	—	—		—	=		68.075	$3.96e^{-097}$

NOTE.—In column six, a "x" indicates an "effector-like" prediction in Guyon et al., (2014), and a "+" indicates a sequence in the six-gene family paralogous to Sscle03g031910 (previously SS1G_13371) identified in Guyon et al.

To determine whether homologous SsPE-containing loci may have arisen through activity of TEs, genomic regions containing these seven SsPEs were inspected for the presence of TE sequences predicted by REPET and assessed for the presence of LTRs and transposition-associated Pfam domains. This showed that all SsPEs were embedded in either a complete or fragmented predicted chimeric LTR retroelement (fig. 5*b*). Blasting of the consensus retroelement to the Dfam database showed that it belongs to the *Gypsy* family of TEs. The model Sscle09g074030 was embedded within the complete retroelement, which contained four upstream open reading frames (ORFs) (within the gene models Sscle09g073990, Sscle09g074000, Sscle09g07410 and Sscle09g07420) that harbored unknown function, zinc knuckle, reverse transcriptase, integrase core, and chromatin organization modifier domains.Additionally, a 5′-LTR was identified 269 bp upstream of the first ORF (Sscle09g073990) and a 3′-LTR was identified 2,170 bp downstream of the SsPE. The SsPE model Sscle07g061960 was situated between 5′- and 3′-LTRs but the region lacked ORFs with predicted transposition-associated domains. The 5′-LTR was situated 5,376 bp from the 5′-end of the SsPE and the 3′-LTR was situated 238 bp from the 3′-end of the SsPE (fig. 5*c*). The 5′- and 3′-LTRs surrounding these two SsPEs were adjacent to target site duplications. Other SsPE models were not situated between 5′- and 3′-LTRs but were flanked by predicted retroelements and DNA TEs containing LTRs and inverted terminal repeats, respectively (data not shown).

To determine whether members of this SsPE family were active during infection, data from the previously mentioned RNASeq analysis were analyzed for evidence of expression. This showed that six of the seven paralogous SsPEs were expressed at at least one time point during infection of *B. napus*. A single paralogous SsPE was not expressed under any conditions, including both *in planta* and *in vitro*. The most highly and consistently (across all conditions tested) expressed SsPE was the model Sscle07g061960. None of the paralogous SsPEs were significantly differentially expressed between conditions. Overall, expression levels of these gene models were low, with mean FPKM values of below 10 (figs. 1, 4*a*, and 5*d*).

Identification of RIP in the Five Most Frequently Occurring Repeats in the *S. sclerotiorum* Genome

To determine whether *S. sclerotiorum* exhibits bimodal GC content, OcculterCut was run on the version two genome sequence and the genome sequences of *L. maculans, B. graminis*, and *P. infestans*. The only genome with a bimodal GC content was *L. maculans*, which has been shown previously to harbor alternating RIP-affected and non-RIP-affected regions (Rouxel et al. 2011). The genomes of the other three species tested, including *S. sclerotiorum*, exhibited unimodal GC content (fig. 6*a*).

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

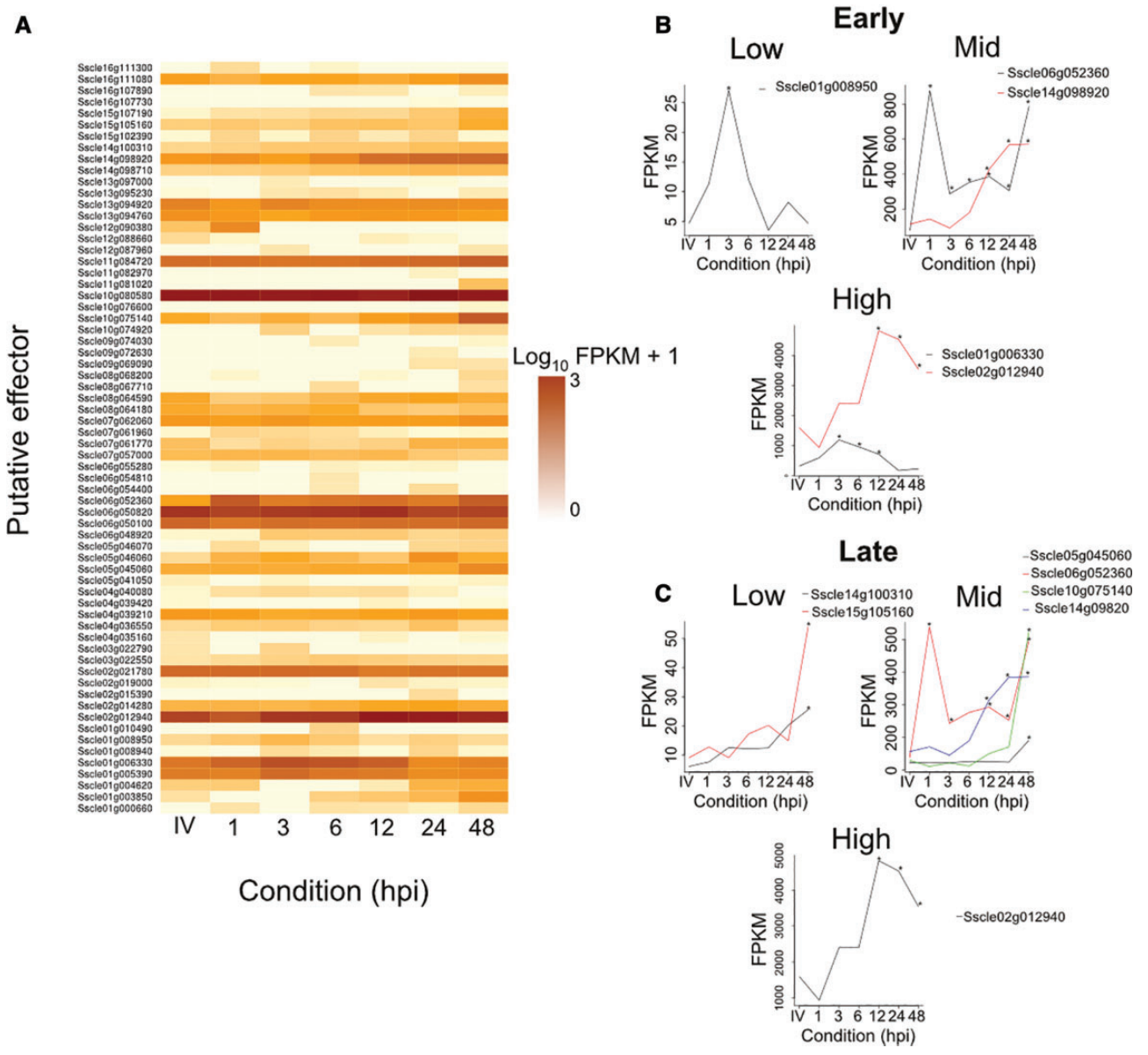

Fig. 4.—Expression analysis of SsPEs. (*a*) Heatmap showing expression ($\log_{10}$(FPKM + 1)) of all expressed SsPEs *in vitro* (IV) and at each *B. napus* time point tested (X1h, X3h, X6h, X12h, X24h, X48h); from light to dark amber represent low to high expression (0 to 3+). (*b*) Expression of SsPEs significantly upregulated *in planta* at early time points (1, 3, 6, or 12 hpi) relative to during growth *in vitro*; asterisks represent a *P*-adjusted value of < 0.05 based on cuffdiff analysis. The *y* axis scale is FPKM and the *x* axis is all conditions tested from *in vitro* (left) to 48 hpi (right). (*c*) Same as (*b*) but for SsPEs that were significantly upregulated at later time points (24 or 48 hpi) relative to during growth *in vitro*. There are some overlapping SsPEs between (*b*) and (*c*).

To determine whether particular repeats in *S. sclerotiorum* were affected by RIP, regardless of whether they were compartmentalized into particular low-GC content regions or not, the five highest copy number repeats in the *S. sclerotiorum* genome were subjected to RIP-based analyses. First, the 50 longest copies of the most abundant repeat, flagged as "RXX-chim_Blc59_repet-L-B64-Map1_reversed" by REPET, were subjected to an alignment-based RipCal version 1.0 (Hane and Oliver, 2008) analysis to identify dominant forms of RIP. This showed that these 50 repeat elements are likely to have undergone both CpA => TpA and CpT => TpT RIPs (fig. 6*b*). Second, the RIP index TpA/ApT was calculated for the 50 longest sequences of all five repeats and compared against a random sequence set of equivalent size and number from the *S. sclerotiorum* genome. This showed that for each repeat family, the TpA/ApT index was significantly higher than found in a random set of sequences (Student's *t*: $P < 0.001$) (fig. 6*c*).

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

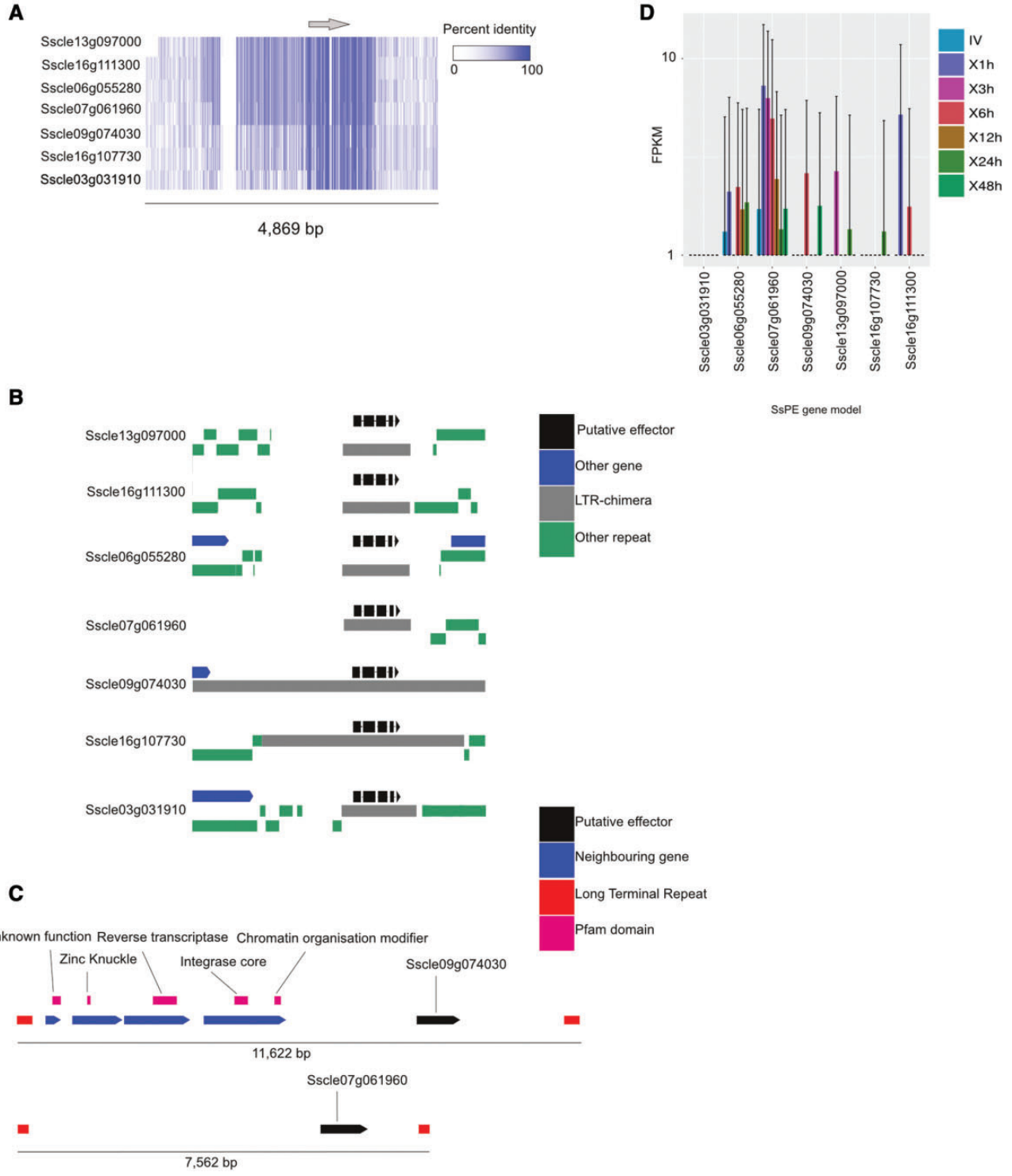

Fig. 5.—A seven member SsPE family associated with transposition of a retro element. (*a*) Alignment of the seven paralogous SsPEs (six of which were identified by Guyon et al.), including flanking genomic regions. Coloring from white (0%) to blue (100%) represents identity across the multiple alignment. The gray arrow above represents the position of the SsPE loci in relation to the alignment of the flanking regions. (*b*) The genomic context of each of the seven SsPEs. Black gene diagrams represent SsPEs, gray bars represent fragments or the complete *Gypsy* retro element associated with this family, green bars represent other adjacent repeats, and blue gene diagrams represent other gene models. (*c*) Genomic context of two SsPEs flanked by corresponding LTR sequences and target site duplications. Red bars represent LTRs, black gene diagrams represent positions of SsPEs, magenta bars represent Pfam domains of genes in this region, blue gene diagrams represent other genes. (*d*) Expression of the seven member SsPE family *in vitro* and across the *B. napus* time points tested. Expression (FPKM + 1) is plotted on the *y* axis. Bars represent standard deviation. Different colored bars represent different sample conditions for each SsPE.

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

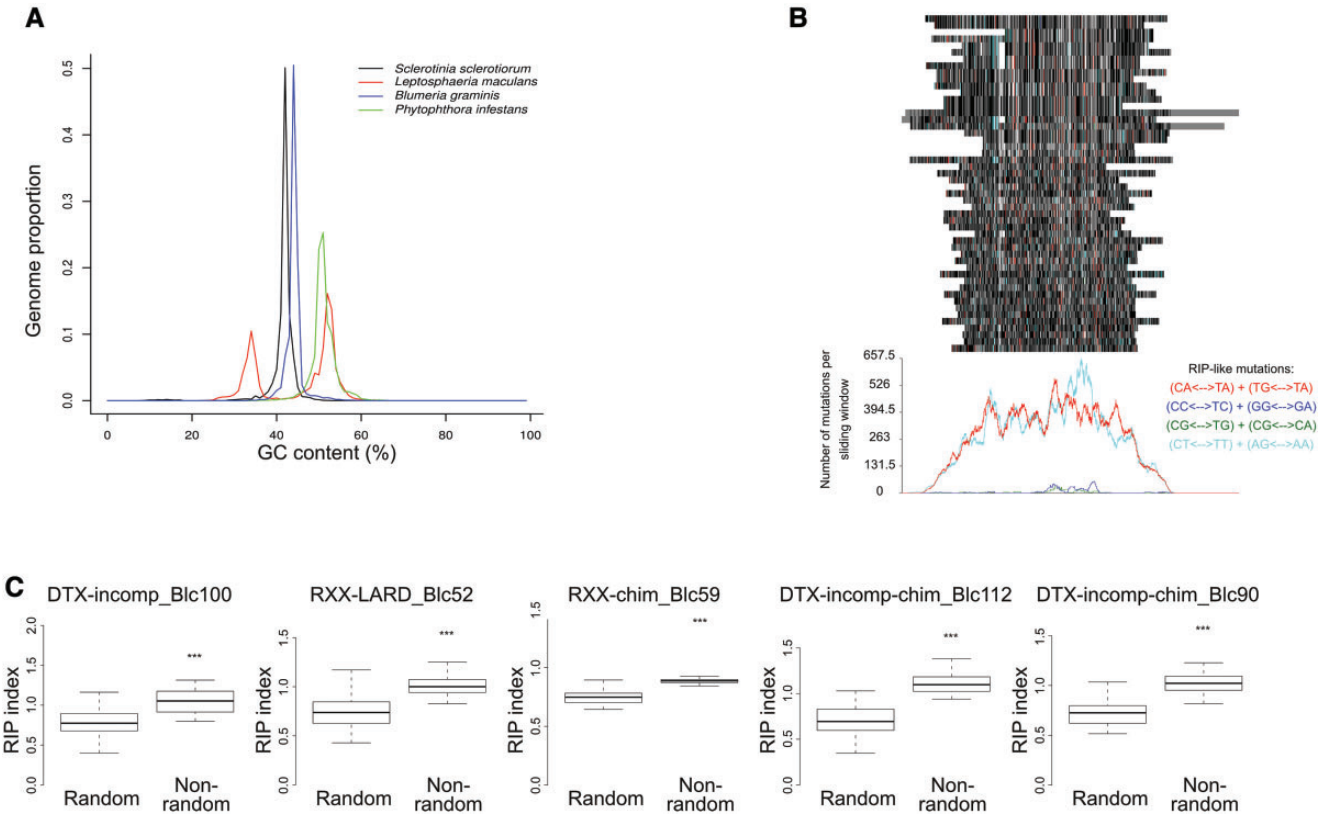

Fig. 6.—Analysis of RIP in the version two *S. sclerotiorum* genome. (*a*) Plot from OcculterCut analysis showing GC content for *S. sclerotiorum, P. infestans, L. maculans*, and *B. graminis* f. sp. *Hordei.* The *x* axis scale is GC content (%) and the *y* axis scale is proportion of the genome that shows a GC content of *x* percentage. (*b*) RipCal alignment based analysis of the most abundant repeat in the *S. sclerotiorum* genome "RXX-chim_Blc59_repet-L-B64-Map1_reversed." The colored bars at the top of the figure represent an alignment of the 50 longest copies of this repeat. Red bars represent likely CpA = > TpA mutation, blue bars represent likely CpC = > TpC mutation, green bars represent likely CpG = > TpG mutation, and turquoise bars represent CpT = > TpT mutation; gray bars, black bars and white bars represent mismatches, matches, and gaps relative to the consensus sequence, respectively. Graph below shows total number of mutations for each potential type of RIP (i.e., CpA = > TpA, CpC = > TpC, CpG = > TpG, CpT = > TpT) at each site in the alignment. (*c*) Comparison of the TpA/ApT index for the five most numerous repeats in the version two *S. sclerotiorum* genome against a random set of equivalent sequences from the same genome (same number and sizes). Only the 50 longest sequences were considered for each repeat family. The *y* axis scale represents the TpA/ApT index for each set of sequences. Black bars represent median values and boxes and whiskers represent second and third quartiles, and interquartile range, respectively. Asterisks represent statistical significance (***$P < 0.001$; Student's *t*-test).

Comparison of the *S. sclerotiorum* Genome Sequence with Genome Sequences of Three Filamentous Fungal Pathogens Known to Harbor Two-Speed Genome Properties

Based on our observations of apparent TE-mediated expansion of a seven-member effector family in *S. sclerotiorum* and potential RIP activity, we speculated that effectors within *S. sclerotiorum* could be associated with rapidly evolving, repeat-rich and potentially RIP-affected genomic regions.

To test whether both secreted proteins and effector-like proteins in *S. sclerotiorum* were significantly associated with RIP-affected sequence, repeats and low gene content, we performed two different analyses. Both of these analyses were also carried out on the two plant pathogenic fungi *B. graminis* and *L. maculans*, and the oomycete *P. infestans.*

The first analysis compared positions of secreted proteins and effector-like proteins with random sets of gene sequences. This analysis showed that in *S. sclerotiorum* there was no significant difference between the distance of secreted or effector proteins from nearest repeat elements than random sets. This was different to all three additional genomes tested. In *L. maculans*, secreted proteins were significantly closer to repeats than a random set (Wilcoxon's test: $P < 0.05$), though there was no significant difference between effector proteins and the random set. However, adjusting α to 0.1 indicated a potential association between effector proteins and repeat regions in *L. maculans* (Wilcoxon's test: $P < 0.1$). In both *B. graminis* and *P. infestans*, both secreted and effector-like proteins were significantly closer to repeats than random sets of proteins (Wilcoxon's test: $P < 0.001$ for *B. graminis* and $P < 0.01$ for

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

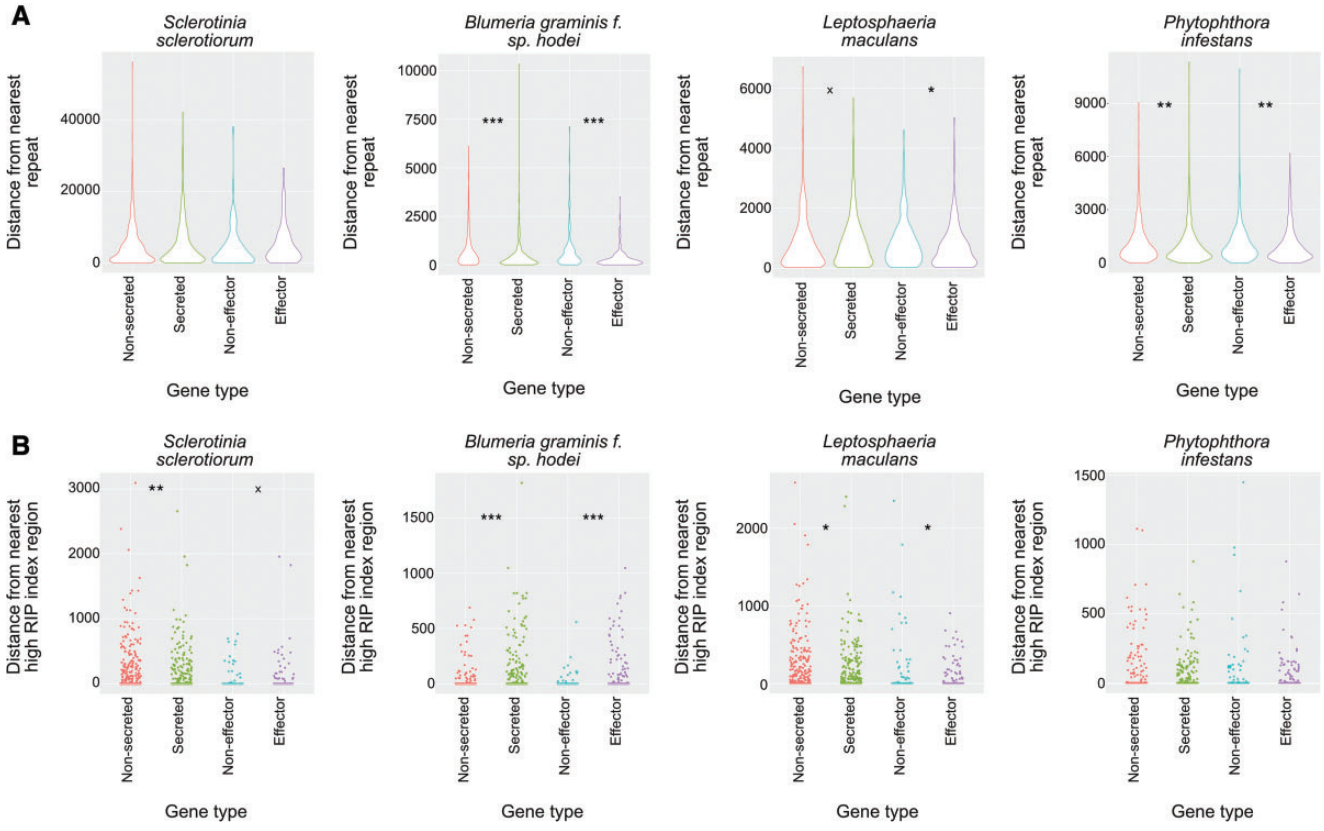

Fig. 7.—Association of repeat secreted protein and effector-like protein encoding genes with repeat regions and regions with a high RIP index. (*a*) Violin plots showing distances of secreted and effector-like protein encoding genes compared with random gene sets from each of the genomes tested. Violin plots represent kernel density of the data points. Asterisks and "x"s represent different *P* values—x = $P < 0.1$; * = $P < 0.05$; ** = $P < 0.01$; and *** = $P < 0.001$ (Wilcoxon's test). The *y* axis scale is distance in bp from the nearest repeat element. Each type of sequence (from left to right: randomized set equivalent in size to secreted protein encoding gene set; secreted protein encoding genes; randomized set equivalent in size to effector-like gene set; effector-like gene set) is colored differently for clarity. Organism names are displayed above plots. (*b*) Jitter plots showing the same as (*a*) but for regions with a high RIP index (as predicted by RipCal). Points are staggered horizontally for each gene set so that differences, as statistically evaluated, are easier to interpret.

P. infestans) (fig. 7*a*). In *S. sclerotiorum* secreted proteins were significantly closer than the random set to regions with a high RIP index (Wilcoxon's test: $P < 0.01$), whereas effector proteins were not. There was some indication that effector proteins in *S. sclerotiorum* may be further away from repeats if α was set to 0.1 (Wilcoxon's test: $P < 0.1$) (fig. 7*b*).

In *B. graminis*, both secreted and effector-like proteins were significantly further away from regions with a high RIP index than the random sets (Wilcoxon's test: $P < 0.001$). In *L. maculans*, both secreted proteins and effector-like proteins were closer to regions with a high RIP index than the random sets (Wilcoxon's test: $P < 0.05$). In *P. infestans*, no significant differences were observed for either secreted or effector proteins (fig. 7*b*).

The second analysis tested correlation between proportion of genes encoding secreted proteins and proportion of CDS sequence and GC content in a 100-kb sliding window, incrementing by 100 kb (end-to-end). This showed that in *S. sclerotiorum* there was no correlation between proportion of secreted proteins and CDS content or GC content. In *L. maculans*, there was a significant negative correlation between proportion of secreted proteins and CDS content (Spearman's test: rho = −0.31; $P < 0.001$), and a significant negative correlation between proportion of secreted proteins and GC content (Spearman's test: rho = −0.36; $P < 0.001$). In *B. graminis*, there was a significant negative correlation between proportion of secreted proteins and CDS content (Spearman's test: rho = −0.80; $P < 0.001$), and no correlation between secreted protein proportion and GC content. In *P. infestans*, there was a significant negative correlation between secreted protein proportion and CDS content (Spearman's test: rho = −0.72; $P < 0.001$), and no correlation between secreted protein proportion and GC content (fig. 8).

Discussion

A Complete Assembly of the *S. sclerotiorum* Genome

In this article, we present a complete and accurately annotated genome of the destructive plant pathogenic fungus

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

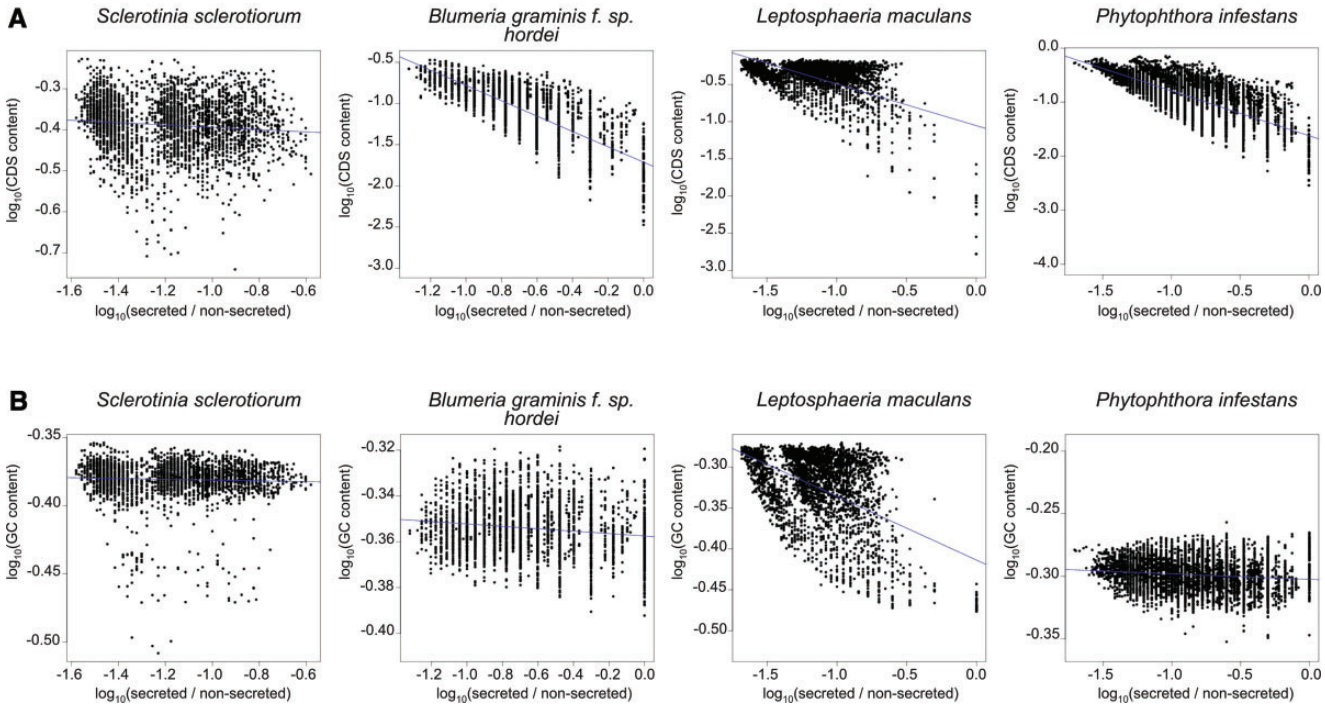

Fig. 8.—Correlation of GC content and CDS content with secreted protein proportion across a 100-kb end-to-end sliding window for all species tested. (*a*) The *x* axis scale is proportion of secreted protein encoding genes (defined as secreted/nonsecreted) for each sliding window, and the *y* axis scale is GC content. Both axes are plotted on a log scale. Blue lines represent least squares regression. (*b*) Same as for (*a*) but for CDS content instead of GC content .

S. sclerotiorum. The previous genome assembly was based on Sanger sequencing at a depth of 9.1× combined with optical mapping. Though this assembly was relatively good quality, it was estimated to have been missing approximately 1.6 Mb of sequence. The new assembly of *S. sclerotiorum* contains an additional 805,146 bp. Though this is only half the estimated missing sequence, we conclude that the new *S. sclerotiorum* genome is practically complete, because 14 of the 16 chromosomes have been sequenced from telomere to telomere with virtually no gaps (fig. 1; supplementary fig. 1*a* and table 2, Supplementary Material online). The only gap present is a region of the nucleolar organizing complex (containing rDNA repeats). A similar result was found in the recently published PacBio-sequenced genome of the plant-pathogenic fungus *V. dahliae*. In this study, which resulted in what was determined to be a finished genome, the authors noted read-stacking in the rDNA repeat region (Faino et al. 2015).

The only other two complete fungal plant pathogen genomes (apart from *V. dahliae*) are those of the wheat head blight fungus *Fusarium graminearum* (King et al. 2015), and the broad host-range necrotroph *B. cinerea* (van Kan et al. 2016); the first of these was assembled with Illumina mate pair technology and the second was assembled using PacBio sequencing and optical mapping. Thus, the new *S. sclerotiorum* assembly represents an addition to a thus far relatively small pool of complete genomes of fungal plant pathogens, and should be of use in future comparative genomics studies.

Most of the additional bases assembled (78.7%) fall within what were predicted as repetitive regions. The total proportion of the genome predicted to be composed of TEs in this study is 12.96%, which is higher than the previously predicted 7.7–9.5% (Amselem et al. 2011; Amselem, Lebrun, et al. 2015). However, as TEs were not manually curated following automated detection, several of these sequences could, in the future, be marked as dubious. Though they have not been thoroughly investigated in this study, more complete analysis of TEs in *S. sclerotiorum*, expanding on findings by Santana et al. (2014) and Amselem, Lebrun, et al. (2015) may now be possible with the additional repetitive sequence.

A More Accurate Set of Gene Annotations in the New *S. sclerotiorum* Genome

Though moderate improvements were made in the genome assembly of *S. sclerotiorum*, the main improvements were in new gene annotations. In the previous genome assembly, 14,522 genes were predicted using *ab initio* gene finders trained on a manually curated gene set of 542 *S. sclerotiorum* genes. These 14,522 gene predictions were further evaluated based on predictions from additional softwares, EST and BLAST evidence, homology to TEs, and length. This resulted

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

in a total of 11,860 nondubious gene calls (Amselem et al. 2011).

In the new gene annotation set of *S. sclerotiorum*, even when considering only the 11,860 "nondubious" genes of the previous genome version, there are still 730 fewer, as the new version contains 11,130 predictions. Furthermore, only 10,528 of these sequences had reciprocal best BLAST hits with version one sequences (supplementary table 4, Supplementary Material online), which suggests that a further 602 sequences were either not present in the previous genome or had changed substantially enough in the version two prediction set for them to be unable to retrieve their corresponding loci as hits through BLAST analysis.

Further inspection revealed that 1,301 version two loci resulted from fusion of previous loci. This is in contrast to the 279 version two gene predictions that resulted from splitting of version one predictions. Thus, it can be surmised that the decrease in filtered gene predictions from the version one assembly to the version two is largely a result of the joining of previous separate loci.

Further analyses illustrated that the version two gene predictions were more accurate than the version one predictions. Blasting of amino acid sequences against the NCBI nr database indicated that version two sequences on average covered more of and had a higher identity to their best BLAST hits. This implies that the new gene predictions were able to obtain more BLAST hits from accurate protein predictions that represent truly conserved sequences (fig. 2*a*). A similar analysis was performed to compare the recent *Parastagonospora nodorum* annotation set with a previous version (Syme et al. 2016). Accuracy at the protein level can also be inferred by the increased number of predicted functional domains and GO terms (fig. 2*b*; supplementary table 5, Supplementary Material online).

Additionally, version two predictions were more similar to their corresponding Cufflinks transcripts based on BLAST analysis. Though median coverage was only improved from 99% to 100% in the version two predictions relative to version one, a high interquartile range of 32% for the version one sequences, as opposed to 1% for the version two sequences, would indicate a decreased degree of variability in correspondence with Cufflinks transcripts in the version two predictions (fig. 2*b*). That the version two predictions are more supported by the RNASeq data is also evident in the observation that the median coverage of version two CDSs is higher than those of version one (fig. 2*c*).

Prediction of Effector Sequences Using an Updated Approach and the New, Improved Gene Models

A major theme of plant pathological research in recent decades has been the identification and analysis of so-called "effector" proteins. These are small, secreted proteins produced by pathogenic fungi whose function is to manipulate host physiology to promote infection (Lo Presti et al. 2015). To date, several effector-like sequences have been functionally characterized in *S. sclerotiorum* (Zhang et al. 2014; Zhu et al. 2013; Wang et al. 2009; Lyu et al. 2016; Dallal Bashi et al. 2010). Two *in silico* analyses using the previous genome version attempted to define the *S. sclerotiorum* secretome and in one of these 79 effector-candidates were predicted (Heard et al. 2015; Guyon et al. 2014). In this study, a new list of 70 effector candidates was identified that was markedly different from the previous list. Only nine genes in the new effector-candidate list were identified by Guyon et al. which highlights the differences not only between the annotation sets but also between the outcomes of effector-prediction pipelines that use different criteria (fig. 1; table 1).

Of the sequences in the new set, only 22 had predicted functional domains. Four of these functional domain predictions have been associated with effector-like activity in other fungi. A cerato-platanin domain was identified in the SsPE Sscle10g076600. Although the exact role of cerato-platanins remains elusive, it has been proposed that they may be involved in causing plant cell-wall instability or could possibly act as pathogen-associated molecular patterns (Baccelli, 2015). Indeed, a cerato-platanin protein deleted in *B. cinerea* was shown to be essential for full virulence (Frías et al. 2011). This protein was also shown to induce a hypersensitive response in tobacco and *A. thaliana* leaves and induce systemic acquired resistance in tobacco (Frías et al. 2011; Frías et al. 2013). The low level of expression of this gene *in vitro* and throughout infection (fig. 4*a*), however, would indicate that abundance of its protein product during *S. sclerotiorum* infection is not necessary for full virulence on *B. napus*. However, further wet-lab studies are needed to test this hypothesis, perhaps including deletion experiments and heterologous expression or infiltration of the protein in *B. napus*.

A necrosis-inducing protein (NPP) domain was identified in the effector-candidate Sscle04g039420. The previous locus corresponding to this gene model was already characterized as being only weakly expressed but nonetheless able to cause necrosis when infiltrated into tobacco leaves (Dallal Bashi et al. 2010). In our study, expression was not detected *in vitro* or at any time points during infection (fig. 4*a*), supporting the observation that this gene is only weakly expressed.

A chitin-binding domain was identified in the gene model Sscle08g068200. The previous locus corresponding to this model was also identified by Guyon et al. Expanding on this discovery, our RNASeq analysis demonstrated that this gene was weakly expressed *in vitro* and throughout infection (fig. 4*a*).

Finally, a chorismate mutase domain was identified in the effector candidate Sscle16g111080. A chorismate mutase with an effector-like function was first identified in the maize pathogen *Ustilago maydis*, where it was shown to be secreted into host cells where it catalyses conversion of chorismate to prephenate. This reaction diverts chorismate away

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

from the salycilic acid biosynthesis pathway, leading to a dampened immune response in infected plants. This dampening occurs because SA is a key signaling molecule in plant defense (Djamei et al. 2011). RNASeq analysis showed that Sscle16g111080 was expressed *in vitro* and throughout infection. This would indicate that the gene is active and may also have an important function in *S. sclerotiorum* infection of *B. napus*.

SsPEs that lacked predicted functional domains were generally homologous to sequences from other fungi (mean amino acid identity of best hit = 66.425%) (supplementary table 6, Supplementary Material online); however, seven were not. These seven sequences were all expressed *in planta*. Intriguingly, one of these genes, Sscle02g012940, was significantly upregulated in *B. napus* from 12 to 48 hpi relative to during growth *in vitro* (fig. 4*b* and *c*). At these time points during *S. sclerotiorum* infection, the plant becomes increasingly necrotized. It is possible that this gene model encodes a necrosis-inducing effector in *S. sclerotiorum*. The fact that this protein was not conserved in any other sequenced fungus may indicate that it is under positive selection pressure, and has become specialized to hosts of *S. sclerotiorum*. To determine this, further studies involving resequencing of *S. sclerotiorum* isolates from diverse regions and hosts, infiltration of this protein onto various host plants and cultivars, and deletion or knockdown experiments could be performed. An alternative hypothesis is that this gene sequence has no homologs simply because no fungal species with a homologous gene sequence has yet been sequenced. As the repertoire of fungal genomes increases, this may be elucidated.

Another putative nonconserved gene model that was highly expressed was Sscle06g050820, though differentially expressed *in planta* relative to during growth *in vitro*, this gene was downregulated. However, as expression levels were high *in vitro* and throughout infection (fig. 4*a*), it still offers an attractive candidate for further study.

The rest of the SsPEs that were not conserved in other fungi were not significantly differentially expressed *in planta* relative to during growth *in vitro*. Overall, they exhibited relatively low levels of expression. It is possible that these sequences are specific to hosts other than *B. napus*, and require differential signals for induction *in planta*. An alternative hypothesis is that low transcript abundance is all that is needed for the activity of these proteins. It is also possible that they are only highly expressed at a very specific time point, which was missed in this infection assay. Such an expression pattern has been shown for effector-like genes in other fungi, for example, *Zymoseptoria tritici* (Rudd et al. 2015).

A total of nine of SsPEs were significantly upregulated either early (defined as 1, 3, 6, and 12 hpi) or late (defined as 24 and 48 hpi) or both early and late during infection (fig. 4*b* and *c*). Apart from the already mentioned model Sscle02g012940, all of these SsPEs were conserved in other fungi (supplementary table 6, Supplementary Material online). Of particular note was the SsPE Sscle01g008950, which was only significantly upregulated at 3 hpi (fig. 4*b*). This would suggest that this SsPE is required before the onset of host necrosis, perhaps as a suppressor of immune responses as has been demonstrated for the *S. sclerotiorum*-secreted integrin like protein and numerous effector proteins in other fungal plant pathogens (Wang et al. 2014; Zhu et al. 2013).

Identification of a Seven-Member Effector Gene Family in *S. sclerotiorum* That May Have Arisen through Recent Transposition

In several plant pathogenic fungal and oomycete species, a clear link between the genomic positions of effector-like genes and TEs has been identified (Rouxel et al. 2011; Grandaubert et al. 2014; Amselem, Lebrun, et al. 2015; Selin et al. 2016; Åsman et al. 2016; Faino et al. 2016). It has been hypothesized that such positioning allows for expansion of effector gene families and subsequent diversification through mutation (Dong et al. 2015). This, in theory, could enhance the capacity of a fungus to rapidly adapt to newly evolved R genes or lost susceptibility loci in host plants. Additionally, RIP, a fungal defense mechanism against TEs, is thought to enhance mutation rates of effectors embedded within TE-rich regions. This process involves mutation of C to T bases in duplicated sequences. Repeat-rich regions containing effector-like proteins are usually gene-poor compared with the rest of the genome. This allows for housekeeping genes to remain unaffected by TE-associated evolutionary processes.

In the *B. napus* pathogen *L. maculans*, it has been shown that most TEs are inactive. This is thought to be a result of an ancestral TE invasion followed by extensive RIP. Approximately 20% of effector-like sequences in *L. maculans* are associated with RIP-affected TE containing regions, whereas only 4.2% are associated with gene-rich regions. Though the sequences in repetitive regions generally do not appear to be in families that have expanded through TE activity, it has been shown that they are likely to have undergone RIP-like mutation (Rouxel et al. 2011).

In the powdery mildew fungus, *B. graminis*, several hundred effector-like genes have been identified *in silico*. These genes occur in 72 families of up to 59 members. These sequences are generally associated with TE-rich regions, and their expansion is thought to be a result of transposition (Pedersen et al. 2012). Interestingly, it has been shown that the EKA (effectors homologous to *Avr* k *1* and *Avr* a *10*) family of effectors is likely to have originated from a truncated TE ORF, which the fungus co-opted as an effector (Amselem, Vigouroux, et al. 2015).

In the oomycete pathogen *P. infestans*, approximately 74% of the genome is repetitive. These repetitive regions harbor extensively expanded, rapidly evolving effector proteins (Haas et al. 2009). This is perhaps one of the most extensively

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

studied and clearest examples of a compartmentalized genome in a microbial plant pathogen.

Based on these observations in fungal and oomycete species, we hypothesized that a similar dynamic may be present in *S. sclerotiorum*. We carried out several analyses to test this hypothesis. To determine whether SsPEs occurred in families, we conducted OrthoMCL analysis using the gene predictions of *S. sclerotiorum* and *B. cinerea* to detect recent paralogs. Further, to determine whether SsPE families were likely formed through TE activity, we conducted multiple sequence alignment of regions containing effectors and inspected them for the presence of TE sequences as predicted by REPET.

By doing this, we identified a single SsPE family, which included seven proteins with more than 90% amino acid identity. The sequences in this family exhibited a single ortholog in *B. cinerea*, indicating that they possibly appeared subsequent to the divergence of *S. sclerotiorum* and *B. cinerea*. Six of the sequences in this family were already identified by Guyon et al. (2014), though a detailed investigation of their association with transposition activity has not been carried out.

None of these sequences contained any predicted functional domains, underlining their possible specialized functions as effectors. Furthermore, multiple alignment of regions surrounding these sequences showed that they were within a 4–5 kb genomic segment that was repeated across six different chromosomes. Intriguingly, each of these SsPEs was embedded in either a partial or complete *Gypsy* LTR-retroelement (fig. 5). This would indicate that these duplicated sequences arose as a result of TE activity. Thus, there is some evidence of effector evolution occurring through transposition in *S. sclerotiorum*, the first evidence of this phenomenon in this species thus far.

Based on this observation, we hypothesized that *S. sclerotiorum* may harbor TE-rich, gene sparse, and potentially RIP-affected regions that are enriched for effector proteins. To test this hypothesis, we performed a comparative analysis of the genome of *S. sclerotiorum* with the "two-speed" genomes of *B. graminis*, *L. maculans*, and *P. infestans*. We found that out of these four microbial genomes, only that of *L. maculans* showed a bimodal GC content (fig. 6*a*). This is consistent with results from the same analysis performed by Testa et al. (2016), who used the previous version of the *S. sclerotiorum* genome. This would indicate that GC content of *B. graminis, P. infestans*, and *S. sclerotiorum* is consistent across the whole genome.

There are two possible explanations for this: 1) The genomes analyzed do not exhibit or only exhibit a limited level of RIP and 2) the genomes exhibit a consistent level of RIP throughout, and do not harbor RIP-affected regions in distinct, GC-depleted compartments. It has been shown that *S. sclerotiorum* and *L. maculans* exhibit active RIP, whereas *B. graminis* does not; *P. infestans* is not thought to harbor RIP as RIP is a fungus-specific mechanism (Hane et al. 2015). *Sclerotinia sclerotiorum* displays an intermediate frequency of RIP and *L. maculans* displays a high frequency of RIP. *Blumeria graminis* shows an RIP-like signature in some TE copies but it is thought that this is the result of ancestral activity, as this species does not harbor the genes necessary for RIP. Unlike *L. maculans*, *S. sclerotiorum* has been shown to exhibit two kinds of RIP, both CpT=>TpT and CpA=>TpA transitions (Amselem, Lebrun, et al. 2015). Our results are consistent with this analysis as *S. sclerotiorum* was found to exhibit dominance of these two types of RIPs across copies of the most abundant TE (fig. 6*b*). Additionally, the TpA/ApT RIP indices of the five most numerous TEs were significantly higher than random sets of control sequences (fig. 6*b*).

As *B. graminis* has only an extremely weak signature of ancestral RIP, it is unlikely to exhibit a bimodal GC content at the whole-genome scale. However, as RIP appears to be at a fairly significant level in *S. sclerotiorum*, the lack of bimodality of its genome GC content could indicate a lack of specific RIP-affected genome compartments such as those observed in *L. maculans*.

To test whether secreted and effector-like proteins were associated with TEs and/or RIP in the four filamentous plant pathogens, both secreted proteins and effector-like proteins were compared with randomized control sets of proteins. This showed that in both *B. graminis* and *P. infestans* there was a significant association between repeats and both secreted and effector-like proteins. In *L. maculans* there was a significant association between secreted proteins and TE sequences, but no significant association between effector-like proteins and TE sequences. However, adjusting α to 0.1 showed that there was possible evidence of a general trend toward association of effector sequences with repeats in *L. maculans* (fig. 7*a*). These results are consistent with the previously proposed hypotheses that *B. graminis* and *P. infestans* do not exhibit active RIP but do compartmentalize effector sequences in TE-rich regions (Haas et al. 2009; Pedersen et al. 2012). It is also consistent with the hypothesis that *L. maculans* repeats have been subjected to extensive RIP, and were hence not as easily discovered by our standardized repeat calling pipeline (RepeatMasker). Despite this, there was still evidence of association of secreted and effector-like proteins with repeat sequences in *L. maculans*.

Conversely, in *S. sclerotiorum*, there was no significant association between secreted or effector-like proteins and TEs. This would suggest that TE activity is not a dominant mode of effector or secreted protein evolution in *S. sclerotiorum*. Indeed, the high degree of homology of SsPEs to proteins in other fungi (90% were homologous at e^{-10}) would indicate that they may be fulfilling conserved functions that are not under extensive selection pressure exerted by host species.

To test whether secreted and effector-like proteins are co-located with RIP-affected sequence in *S. sclerotiorum*, we performed the same analysis as for the repeat sequences for regions with a high RIP index as specified by RipCal. This showed that *L. maculans* exhibited a significant association

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

between both secreted proteins and effector-like proteins and RIP-affected sequences, whereas *B. graminis* exhibited a significant negative association between effector and secreted proteins and RIP-affected sequence. In *P. infestans*, there was no association between secreted or effector proteins and RIP-affected sequence. This is consistent with the previous observation that *L. maculans* exhibits significant RIP activity, which has a particular impact on the evolution of effector proteins compartmentalized into AT-rich regions (Rouxel et al. 2011). The data would also indicate that *B. graminis* requires active transposons for effector diversification and that effector proteins are significantly further from regions that may have undergone RIP ancestrally.

Intriguingly, in *S. sclerotiorum*, there was a significant association between secreted proteins and RIP-affected sequence. However, there was no association between effector-like sequences and RIP-affected sequence. This would indicate that there may be some hot spots of RIP associated with secreted protein diversification in the *S. sclerotiorum* genome. Indeed, several regions can be readily identified in figure 1. For example, at the distal ends of chromosomes 7, 15, and 16 there appear to be clusters of secreted and effector-like proteins that occur in proximity to regions with a high RIP index and several repeat sequences. However, the impact of these regions on evolution and host specificity remain to be elucidated.

Based on these observations, we tested the hypothesis that secreted proteins were associated with AT-rich regions in *S. sclerotiorum*. Furthermore, we tested whether they were associated with gene sparse regions, which would be indicative of a significant level of compartmentalization away from housekeeping genes. We found that in *S. sclerotiorum*, *B. graminis*, and *P. infestans*, there was no correlation between secreted protein proportion and GC content. However, in *L. maculans*, there was a significant negative correlation between GC content and secreted protein proportion (fig. 8a). This would indicate that in *S. sclerotiorum*, there is no significant compartmentalization of secreted proteins into AT-rich (i.e., RIP-affected) sequence regions. Though it is possible that there are hotspots of this kind of activity in the genome, a 100-kb sliding window was too large to detect them. This is a limitation of this study. However, as it was able to detect an association in *L. maculans*, for which RIP-affected genome compartments with effectors have been characterized, it would at least demonstrate that in *S. sclerotiorum* this kind of phenomenon is not as extensive as in *L. maculans.* In addition, we found that the genomes of *B. graminis*, *P. infestans*, and *L. maculans* exhibited a significant negative correlation between CDS content (used to infer gene richness) and secreted protein proportion. This supports previous observations that effector-like and secreted proteins are often positioned in gene sparse regions in these species. Conversely, there was no correlation between gene content and secreted protein proportion in *S. sclerotiorum*. Again, it is possible that there are small hotpots of secreted protein diversification that are gene poor; figure 1 illustrates such candidate regions (e.g., the distal end of chromosome 7). However, at a macroscale, this phenomenon was not identified in *S. sclerotiorum* as it was in organisms previously characterized (Raffaele and Kamoun, 2012). Therefore, we propose that compartmentalization of secreted proteins in *S. sclerotiorum*, if at all present, is far subtler than it is in the host-specific biotrophic and hemibiotrophic organisms included in this study.

Conclusion

In conclusion, we have produced a complete and accurate genome of the important broad host range necrotrophic fungus *S. sclerotiorum*. We have identified a number of novel effector candidates for future studies and elucidated their expression patterns *in planta*. Further, we show that the genome architecture of *S. sclerotiorum*, in terms of TEs, RIP, and secreted and effector-like proteins, is different to three of the most well-characterized filamentous fungi and oomycetes that conform to the two-speed genome hypothesis. Instead, we demonstrate subtle signatures of enhanced mutation of secreted proteins and effectors in *S. sclerotiorum* through RIP activity and transposition, which is not generally detectable at a whole-genome scale as in the other species tested. This has implications for effector discovery and comparative genomic analyses in the future.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgements

MFS is supported by a Veni grant of the Research Council for Earth and Life Sciences (ALW) of the Netherlands Organization for Scientific Research (NWO). KH-K is supported by the UK Biotechnology and Biological Sciences Research Council (BBSRC) through the Institute Strategic Programme Grant 20:20 Wheat (BB/J/00426x/1). SH was funded by a BBSRC Industrial Collaborative Award in Science and Engineering (CASE) studentship supported by Syngenta entitled 'Early sensing of plant pathogenic fungi'. JR received support for this project from the USDA National Institute of Food and Agriculture, Hatch project 1005726. MM, ON and SR are funded by the European Research Council (ERC-StG 336808 project VariWhim) and the French Laboratory of Excellence project TULIP (ANR-10-LABX-41; ANR-11-IDEX-0002–02; New Frontiers grant 'ScleRNAi'). DH and SS are funded by SaskCanola and the Government of Canada through the Developing Innovative Agri-Products programme (project # J-000269: P032). RPO conducted part of this research in Wageningen Agricultural University Laboratory of Phytopathology as a KNAW research fellow. MCD, MD-G

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

and RPO are funded by the Grains Research and Development Corporation (GRDC) as part of a bilateral agreement between Curtin University under the grant CUR00023, within the Centre for Crop and Disease Management (CCDM). This work was in part supported by resources provided by The Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia. MCD would like to personally thank Robert King and Keywan Hassani-Pak (Rothamsted Research) for advice and instruction regarding genome assembly during the initial stages of the project.

Literature Cited

Amaral AMD, Antoniw J, Rudd JJ, Hammond-Kosack KE. 2012. Defining the predicted protein secretome of the fungal wheat leaf pathogen *Mycosphaerella graminicola*. PLoS ONE 7:e49904.

Amselem J, et al. 2011. Genomic analysis of the necrotrophic fungal pathogens *Sclerotinia sclerotiorum* and *Botrytis cinerea*. PLoS Genet. 7:e1002230.

Amselem J, Lebrun MH, Quesneville H. 2015. Whole genome comparative analysis of transposable elements provides new insight into mechanisms of their inactivation in fungal genomes. BMC Genomics 16.

Amselem J, Vigouroux M, et al. 2015. Evolution of the EKA family of powdery mildew avirulence-effector genes from the ORF 1 of a LINE retrotransposon. BMC Genomics 16:917.

Ashburner M, et al. 2000. Gene Ontology: tool for the unification of biology. Nat Genet. 25:25–29.

Åsman AKM, Fogelqvist J, Vetukuri RR, Dixelius C. 2016. *Phytophthora infestans* Argonaute 1 binds microRNA and small RNAs from effector genes and transposable elements. New Phytol. 211:993–1007.

Baccelli I. 2015. Cerato-platanin family proteins: one function for multiple biological roles? Front Plant Sci. 5.

Bae H, Kim MS, Sicher RC, Bae HJ, Bailey BA. 2006. Necrosis- and ethylene-inducing peptide from *Fusarium oxysporum* induces a complex cascade of transcripts associated with signal transduction and cell death in *Arabidopsis*. Plant Physiol. 141:1056–1067.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 6.

Berlin K, et al. 2014. Assembling large genomes with single-molecule sequencing and locality sensitive hashing. bioRxiv 8003.

Boland GJ, Hall R. 1994. Index of plant hosts of *Sclerotinia sclerotiorum*. Can J Plant Pathol. 16:93–108.

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

Borodovsky M, Lomsadze A. 2011. Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. In: Baxevanis AD, Petsko GA, Stein LD, Stormo GD, editors. Current protocols in bioinformatics. Hoboken (NJ): John Wiley & Sons, Inc.

Bourras S, et al. 2015. Multiple avirulence loci and allele-specific effector recognition control the Pm3 race-specific resistance of wheat to powdery mildew. Plant Cell 27:2991–3012.

Bourras S, McNally KE, Müller MC, Wicker T, Keller B. 2016. Avirulence genes in cereal powdery mildews: the gene-for-gene hypothesis 2.0. Plant Biot Interact. 241.

Chalhoub B, et al. 2014. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. Science 345:950–953.

Chin CS, et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 10:563–569.

Chitrampalam P, Figuli PJ, Matheron ME, Subbarao KV, Pryor BM. 2008. Biocontrol of lettuce drop caused by *Sclerotinia sclerotiorum* and *S. minor* in desert agroecosystems. Plant Dis. 92:1625–1634.

Ciuffetti LM, Manning VA, Pandelova I, Betts MF, Martinez JP. 2010. Host-selective toxins, Ptr ToxA and Ptr ToxB, as necrotrophic effectors in the *Pyrenophora tritici-repentis*–wheat interaction. New Phytol. 187:911–919.

Croll D, McDonald BA. 2012. The accessory genome as a cradle for adaptive evolution in pathogens. PLoS Pathog. 8:e1002608.

Dallal Bashi Z, Hegedus DD, Buchwaldt L, Rimmer SR, Borhan MH. 2010. Expression and regulation of *Sclerotinia sclerotiorum* necrosis and ethylene-inducing peptides (NEPs). Mol Plant Pathol. 11:43–53.

de Jonge R, et al. 2010. Conserved fungal LysM effector Ecp6 prevents chitin-triggered immunity in plants. Science 329:953–955.

Dean R, et al. 2012. The top 10 fungal pathogens in molecular plant pathology. Mol Plant Pathol. 13:414–430.

Derbyshire MC, Denton-Giles M. 2016. The control of sclerotinia stem rot on oilseed rape (*Brassica napus*): current practices and future opportunities. Plant Pathol. n/a-n/a.

Djamei A, et al. 2011. Metabolic priming by a secreted fungal effector. Nature 478:395–398.

Dong S, Raffaele S, Kamoun S. 2015. The two-speed genomes of filamentous pathogens: waltz with plants. Curr Opin Genet Dev. 35:57–65.

Ekblom R, Wolf JBW. 2014. A field guide to whole-genome sequencing, assembly and annotation. Evol Appl. 7:1026–1042.

Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9:18.

English AC, et al. 2012. Mind the gap: upgrading genomes with pacific biosciences RS long-read sequencing technology. PLoS ONE 7:e47768.

Faino L, et al. 2015. Single-molecule real-time sequencing combined with optical mapping yields completely finished fungal genome. mBio 6:e00936-15.

Faino L, et al. 2016. Transposons passively and actively contribute to evolution of the two-speed genome of a fungal pathogen. Genome Res. 26:1091–1100.

Fankhauser N, Mäser P. 2005. Identification of GPI anchor attachment signals by a Kohonen self-organizing map. Bioinformatics 21:1846–1852.

Flor HH, Comstock VE. 1972. Identification of rust-conditioning genes in flax cultivars. Crop Sci. 12:800.

Frías M, Brito N, González C. 2013. The *Botrytis cinerea* cerato-platanin BcSpl1 is a potent inducer of systemic acquired resistance (SAR) in tobacco and generates a wave of salicylic acid expanding from the site of application. Mol Plant Pathol. 14:191–196.

Frías M, González C, Brito N. 2011. BcSpl1, a cerato-platanin family protein, contributes to *Botrytis cinerea* virulence and elicits the hypersensitive response in the host. New Phytol. 192:483–495.

Friesen TL, et al. 2006. Emergence of a new disease as a result of interspecific virulence gene transfer. Nat Genet. 38:953–956.

Friesen TL, Faris JD, Solomon PS, Oliver RP. 2008. Host-specific toxins: effectors of necrotrophic pathogenicity. Cell Microbiol. 10:1421–1428.

Friesen TL, Zhang Z, Solomon PS, Oliver RP, Faris JD. 2008. Characterization of the interaction of a novel *Stagonospora nodorum* host-selective toxin with a wheat susceptibility gene. Plant Physiol. 146:682–693.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 17:333–351.

Gordon TR, Martyn RD. 1997. The evolutionary biology of *Fusarium oxysporum*. Annu Rev Phytopathol. 35:111–128.

Grandaubert J, et al. 2014. Transposable element-assisted evolution and adaptation to host plant within the *Leptosphaeria maculans–Leptosphaeria biglobosa* species complex of fungal pathogens. BMC Genomics 15:891.

Guyon K, Balagué C, Roby D, Raffaele S. 2014. Secretome analysis reveals effector candidates associated with broad host range necrotrophy in

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

the fungal plant pathogen *Sclerotinia sclerotiorum*. BMC Genomics 15:336.
Haas BJ, et al. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 9:R7.
Haas BJ, et al. 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. Nature 461:393–398.
Hane JK, Oliver RP. 2008. RIPCAL: a tool for alignment-based analysis of repeat-induced point mutations in fungal genomic sequences. BMC Bioinformatics 9:478.
Hane JK, Williams AH, Taranto AP, Solomon PS, Oliver RP. 2015. Repeat-induced point mutation: a fungal-specific, endogenous mutagenesis process. In: van den Berg MA, Maruthachalam K, editors. Genetic transformation systems in fungi. Vol. 2. Springer International Publishing. p. 55–68.
Heard S, Brown NA, Hammond-Kosack K. 2015. An interspecies comparative analysis of the predicted secretomes of the necrotrophic plant pathogens *Sclerotinia sclerotiorum* and *Botrytis cinerea*. PLoS ONE 10.
Heffer. 2007. White mold. Plant Health Instr.
Hemetsberger C, et al. 2015. The fungal core effector Pep1 is conserved across smuts of dicots and monocots. New Phytol. 206:1116–1126.
Jones P, et al. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics.
King R, Urban M, Hammond-Kosack MCU, Hassani-Pak K, Hammond-Kosack KE. 2015. The completed genome sequence of the pathogenic ascomycete fungus *Fusarium graminearum*. BMC Genomics 16:544.
Kleemann J, et al. 2012. Sequential delivery of host-induced virulence effectors by appressoria and intracellular hyphae of the phytopathogen *Colletotrichum higginsianum*. PLOS Pathog. 8:e1002643.
Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 305:567–580.
Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10:R25.
Lee WS, Rudd JJ, Hammond-Kosack KE, Kanyuka K. 2013. *Mycosphaerella graminicola* LysM effector-mediated stealth pathogenesis subverts recognition through both CERK1 and CEBiP homologues in wheat. Mol Plant Microbe Interact. 27:236–243.
Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.
Liang X, et al. 2015. Oxaloacetate acetylhydrolase gene mutants of *Sclerotinia sclerotiorum* do not accumulate oxalic acid, but do produce limited lesions on host plants. Mol Plant Pathol. 16:559–571.
Liang Y, Yajima W, Davis MR, Kav NNV, Strelkov SE. 2013. Disruption of a gene encoding a hypothetical secreted protein from *Sclerotinia sclerotiorum* reduces its virulence on canola (*Brassica napus*). Can J Plant Pathol. 35:46–55.
Liu Z, et al. 2006. The Tsn1–ToxA interaction in the wheat–*Stagonospora nodorum* pathosystem parallels that of the wheat–tan spot system. Genome 49:1265–1273.
Lo Presti L, et al. 2015. Fungal effectors and plant susceptibility. Annu Rev Plant Biol. 66:513–545.
Lu S, Gillian Turgeon B, Edwards MC. 2015. A ToxA-like protein from *Cochliobolus heterostrophus* induces light-dependent leaf necrosis and acts as a virulence factor with host selectivity on maize. Fungal Genet Biol. 81:12–24.
Lyu X, et al. 2015. Comparative genomic and transcriptional analyses of the carbohydrate-active enzymes and secretomes of phytopathogenic fungi reveal their significant roles during infection and development. Sci Rep. 5.
Lyu X, et al. 2016. A small secreted virulence-related protein is essential for the necrotrophic interactions of *Sclerotinia sclerotiorum* with its host plants. PLoS Pathog. 12:e1005435.
Ma LJ, et al. 2010. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature 464:367–373.
Marshall R, et al. 2011. Analysis of two in planta expressed LysM effector homologs from the fungus *Mycosphaerella graminicola* reveals novel functional properties and varying contributions to virulence on wheat. Plant Physiol. 156:756–769.
McKenna A, et al. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.
Naito S, Sugimoto T. 1986. *Sclerotinia* stalk rot of sugar beets. Jpn J Phytopathol. 52:217–224.
Nemri A, et al. 2014. The genome sequence and effector complement of the flax rust pathogen *Melampsora lini*. Front Plant Sci. 5:98.
Ohm RA, et al. 2012. Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. PLoS Pathog. 8:e1003037.
Pedersen C, et al. 2012. Structure and evolution of barley powdery mildew effector candidates. BMC Genomics 13:694.
Pedro H, et al. 2016. PhytoPath: an integrative resource for plant pathogen genomics. Nucleic Acids Res. 44:D688–D693.
Peltier AJ, et al. 2012. Biology, yield loss and control of *Sclerotinia* stem rot of soybean. J Integr Pest Manag. 3:1–7.
Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 8:785–786.
Quesneville H, et al. 2005. Combined evidence annotation of transposable elements in genome sequences. PLoS Comput Biol. 1:e22.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.
Raffaele S, Kamoun S. 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. Nat Rev Microbiol.
Raffaele S, et al. 2010. Genome evolution following host jumps in the Irish potato famine pathogen lineage. Science 330:1540–1543.
Redkar A, Bonequi MV, Doehlemann G. 2015. Conservation of the *Ustilago maydis* effector See1 in related smuts. Plant Signal Behav. 10:e1086855.
Rouxel T, et al. 2011. Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. Nat. Commun. 2:202.
Rowe RC. 1980. Comparative pathogenicity and host ranges of *Fusarium oxysporum* isolates causing crown and root rot of greenhouse and field-grown tomatoes in North America and Japan. Phytopathology 70:1143.
Rudd JJ, et al. 2015. Transcriptome and metabolite profiling of the infection cycle of *Zymoseptoria tritici* on wheat reveals a biphasic interaction with plant immunity involving differential pathogen chromosomal contributions and a variation on the hemibiotrophic lifestyle definition. Plant Physiol. 167:1158–1185.
Santana MF, Silva JCF, Mizubuti ESG, Arajo EF, Queiroz MV. 2014. Analysis of Tc1-Mariner elements in *Sclerotinia sclerotiorum* suggests recent activity and flexible transposases. BMC Microbiol. 14:256.
Sanz-Martín JM, et al. 2016. A highly conserved metalloprotease effector enhances virulence in the maize anthracnose fungus *Colletotrichum graminicola*. Mol Plant Pathol. n/a-n/a.
Selin C, de Kievit TR, Belmonte MF, Fernando WGD. 2016. Elucidating the role of effectors in plant-fungal interactions: progress and challenges. Front Microbiol. 7.
Selker EU, Stevens JN. 1985. DNA methylation at asymmetric sites is associated with numerous transition mutations. Proc Natl Acad Sci U S A. 82:8114–8118.
Smith KM, Galazka JM, Phatale PA, Connolly LR, Freitag M. 2012. Centromeres of filamentous fungi. Chromosome Res. 20:635–656.
Sperschneider J, et al. 2015. Advances and challenges in computational prediction of effectors from plant pathogenic fungi. PLoS Pathog. 11:e1004806.

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018

Sperschneider J, et al. 2016. EffectorP: predicting fungal effector proteins from secretomes using machine learning. New Phytol. 210:743–761.
Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 33:W465–W467.
Stergiopoulos I, de Wit PJGM. 2009. Fungal effector proteins. Annu Rev Phytopathol. 47:233–263.
Syme RA, et al. 2016. Comprehensive annotation of the *Parastagonospora nodorum* reference genome using next-generation genomics, transcriptomics and proteogenomics. PLoS ONE 11:e0147221.
Testa AC, Hane JK, Ellwood SR, Oliver RP. 2015. CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. BMC Genomics 16:170.
Testa AC, Oliver RP, Hane JK. 2016. OcculterCut: a comprehensive survey of AT-rich regions in fungal genomes. Genome Biol Evol. 8:2044–2064.
Thatcher LF, Gardiner DM, Kazan K, Manners JM. 2012. A highly conserved effector in *Fusarium oxysporum* is required for full virulence on *Arabidopsis*. Mol Plant Microbe Interact. 25:180–190.
Thomma BPHJ, et al. 2016. Mind the gap; seven reasons to close fragmented genome assemblies. Fungal Genet Biol. 90:24–30.
Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.
Trapnell C, et al. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 7:562–578.
van Kan JAL, et al. 2017. A gapless genome sequence of the fungus *Botrytis cinerea*. Mol Plant Pathol. 18:75–89.
Vargas WA, et al. 2016. A fungal effector with host nuclear localization and DNA-binding properties is required for maize anthracnose development. Mol Plant Microbe Interact. 29:83–95.
Wang X, et al. 2009. Characterization of a canola C2 domain gene that interacts with PG, an effector of the necrotrophic fungus *Sclerotinia sclerotiorum*. J Exp Bot. 60:2613–2620.
Wang X, Jiang N, Liu J, Liu W, Wang GL. 2014. The role of effectors and host immunity in plant–necrotrophic fungal interactions. Virulence 5:722–732.
Wheeler TJ, et al. 2013. Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucleic Acids Res. 41:D70–D82.
Whigham E, et al. 2015. Broadly conserved fungal effector BEC1019 suppresses host cell death and enhances pathogen virulence in powdery mildew of barley (*Hordeum vulgare* L.). Mol Plant Microbe Interact. 28:968–983.
Wicker T, et al. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 8:973–982.
Zhang H, et al. 2014. A novel protein elicitor (SsCut) from *Sclerotinia sclerotiorum* induces multiple defense responses in plants. Plant Mol Biol. 86:495–511.
Zhu W, et al. 2013. A secretory protein of necrotrophic fungus *Sclerotinia sclerotiorum* that suppresses host resistance. PLoS ONE 8:e53901.
Zuppini A, et al. 2005. An endopolygalacturonase from *Sclerotinia sclerotiorum* induces calcium-mediated signaling and programmed cell death in soybean cells. Mol Plant Microbe Interact. 18:849–855.
2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815.

Associate editor: Rebecca Zufall

Downloaded from https://academic.oup.com/gbe/article-abstract/9/3/593/2997436 by guest on 20 November 2018