Genomic overview of closely related fungi with different *Protea* host ranges

Janneke Aylward[a,b,*], Brenda D. Wingfield[b], Léanne L. Dreyer[a], Francois Roets[c], Michael J. Wingfield[b] & Emma T. Steenkamp[b]

[a] Department of Botany and Zoology, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa

[b] Department of Biochemistry, Genetics and Microbiology, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria 0002, South Africa

[c] Department of Conservation Ecology and Entomology, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa

[*]Corresponding author: Janneke Aylward

Postal address: Department of Conservation Ecology and Entomology, Stellenbosch University, Victoria Rd, Stellenbosch, 7600, South Africa

Email: janneke@sun.ac.za

Tel: +27 76 948 2033

Fax: +27 21 808 2405

Highlights

• We compared the genomes of two *Knoxdaviesia* species with different *Protea* hosts.

• Similarity was apparent at a structural and gene content level.

• Few secreted proteins and secondary metabolite gene clusters were identified.

• A degraded secondary metabolite gene suggests limited competition in *K. proteae*.

• The specialist *Protea* niche has likely facilitated a smaller genetic complement.

Abstract

Genome comparisons of species with distinctive ecological traits can elucidate genetic divergence that influenced their differentiation. The interaction of a microorganism with its biotic environment is largely regulated by secreted compounds, and these can be predicted from genome sequences. In this study, we considered *Knoxdaviesia capensis* and *K. proteae*, two closely related saprotrophic fungi found exclusively *Protea* plants. We investigated their genome structure to compare their potential inter-specific interactions based on gene content. Their genomes displayed macrosynteny and were approximately 10 % repetitive. Both species had fewer secreted proteins than pathogens and other saprotrophs, reflecting their specialized habitat. The bulk of the predicted species-specific and secreted proteins coded for carbohydrate metabolism, with a slightly higher number of unique carbohydrate-degrading proteins in the broad host-range *K. capensis*. These fungi have few secondary metabolite gene clusters, suggesting minimal competition with other microbes and symbiosis with antibiotic-producing bacteria common in this niche. Secreted proteins associated with detoxification and iron sequestration likely enable these *Knoxdaviesia* species to tolerate antifungal compounds and compete for resources, facilitating their unusual dominance. This study confirms the genetic cohesion between *Protea*-associated *Knoxdaviesia* species and reveals aspects of their ecology that have likely evolved in response to their specialist niche.

Keywords: competition; genomes; interactions; *Knoxdaviesia*; secretome

1. Introduction

The increasing availability of whole genome sequences has enabled species comparisons to be extended beyond ecology, morphology and single gene sequences. Species can now be investigated in the context of their genome structure and the presence or absence of genes and other genomic features (Liti and Louis, 2005). Divergent genomic characteristics between closely related species are specifically intriguing, as they may elucidate the genetic basis of ecological differentiation (e.g. Rocap et al., 2003).

Every living organism interacts with its surroundings and, for microorganisms, their secreted proteins provide a picture of the potential range of environmental interactions. These proteins mediate external metabolism of complex substrates, competition with other organisms and establishment of microbe-host symbioses (Girard et al., 2013). For example, most plant pathogenic fungi secrete an array of enzymes to degrade plant cell walls (Kubicek et al., 2014), as well as numerous effector proteins that modulate the host's immune system and defence responses (Schneider and Collmer, 2010). Secreted low molecular weight proteins, rich in the amino acid cysteine and often uncharacterized, have been particularly implicated in both pathogenic and symbiotic fungal-host interactions (Martin et al., 2008; 2004; Rep, 2005b; Templeton et al., 1994). Indeed, the complement of secreted proteins tends to be strikingly different between microorganisms with different lifestyles (Girard et al., 2013; Kohler et al., 2015).

The ecological success of an organism's biotic interactions is related to its ability to adapt to a changing environment. This "evolutionary potential" is influenced by mutability (McDonald and Linde, 2002) and the proportion of repetitive sequences in a genome is one of the factors that may drive such mutability (Wöstemeyer and Kreibich, 2002). Repetitive regions, specifically transposable elements (TEs), may create sequence duplications and/or rearrangements via transposition or by mediating homologous recombination (Gray, 2000). In contrast to these long-term mutagenic effects, TEs may also cause more immediate phenotypic changes, such as repressing the expression of nearby genes in some fungi (Castanera et al., 2016). TEs, therefore, have the potential to profoundly affect overall genome structure, diversification and functionality.

In this study, we considered whether the genome structure and gene content, including secreted proteins and repetitive sequences, of two closely related fungal species reflect the differences in their ecologies. For this purpose, we investigated *Knoxdaviesia capensis* M.J. Wingf. & P.S. van Wyk and *K. proteae* M.J. Wingf., P.S. van Wyk & Marasas,

ophiostomatoid fungi (Wingfield et al., 1993) that dominate within the flower heads of *Protea* plants in the Cape Floristic Region, South Africa (Lee et al., 2005; Marais and Wingfield, 1994; Seifert et al., 2013; Wingfield and Van Wyk, 1993). *Knoxdaviesia proteae* occurs on a single *Protea* L. host species, whereas *K. capensis* occurs in multiple *Protea* hosts (Roets et al., 2009; Wingfield and Van Wyk, 1993) and, therefore, encounters a variety of host chemistries (Roets et al., 2011). Both *K. capensis* and *K. proteae* are, however, specialists in that they occupy a niche outside of which they have never been observed and most likely do not occur (Roets et al., 2005; 2009). Therefore, their biomass-degrading enzymes (Aylward et al., 2017b), and likely other aspects of their lifestyle, have been reduced to the extent that only *Protea* species are suitable hosts. It is consequently plausible that their different host ranges and their specialization on a single host genus will be reflected in their genomes.

Once serotinous *Protea* flower heads are fertilized, they mature and develop into seed storage organs known as infructescences (Fig. 1A, B). The closed infructescence is moist and protected from external factors such as wind and rain, providing an ideal micro-habitat for the *Knoxdaviesia* species (Fig. 1C), as well as other microorganisms, arthropods and nematodes (Coetzee and Giliomee, 1985; Human et al., 2016; Lee et al., 2003; 2005; Roets et al., 2005; 2006b; Theron et al., 2012; Zwölfer, 1979). Within this environment, ophiostomatoid fungi such as *Knoxdaviesia* represent a food source for the arthropods and, potentially, nematodes (Roets et al., 2013; Ruess and Lussenhop, 2005). Additionally, at least 22 species of saprotrophic fungi other than ophiostomatoid symbionts occupy *Protea* infructescences (Lee et al., 2005), generating further inter-organismal competition.

Here, we compare the publically available genomes of *K. capensis* and *K. proteae* (Aylward et al., 2016a) to (i) provide an overview of their genome structure and (ii) compare their potential inter-specific interactions based on gene content, with a specific focus on secreted proteins and secondary metabolite biosynthesis clusters. Because of their close phylogenetic relationship (Wingfield et al., 1999) and the few differences revealed by previous genetic studies (Aylward et al., 2016b; 2017b), we expected a high level of similarity between these genomes. The expected similarity could, however, be impeded if the *Knoxdaviesia* genomes are highly repetitive or repeats have proliferated in one or both of these genomes since speciation. The confined niche of these species may have promoted small secretomes and few secreted protein effectors. It is plausible that this reduction would be more pronounced in *K. proteae*, since it has presumably diverged from the broad-host range *K. capensis* due to restricted host association (Aylward et al., 2017b). With its broad host-range, *K. capensis* must interact with multiple hosts and, therefore, may also encounter a greater range of

competitive organisms. We, therefore, hypothesized that the secondary metabolite biosynthesis capability of these *Protea*-associated *Knoxdaviesia* species would be adapted to contending with microbial and animal competitors and that *K. capensis* requires a larger arsenal than *K. proteae*.

Fig. 1. The unique habitat of *Protea*-associated *Knoxdaviesia*. Flower heads of *Protea repens* (A) and *P. neriifolia* (B). Perithecia of *Knoxdaviesia* on the style and tepal of a *Protea* flower (C). Scale bar = 1 mm.

2. Materials and methods

2.1. *Knoxdaviesia* genomes and whole genome alignment

The genomes of *K. capensis* CBS139037 (LNGK00000000.1) and *K. proteae* CBS140089 (LNGL00000000.1), as well as their predicted proteins were utilized in this study. Both genomes were sequenced on the Illumina HiSeq platform, were previously annotated with MAKER (Cantarel et al., 2008; Holt and Yandell, 2011) and are estimated to be > 98% complete (Aylward et al., 2016a). The 29 scaffolds of the *K. capensis* genome and the 133 scaffolds of the *K. proteae* genome were masked with RepeatScout 1.0.5 (Price et al., 2005) and aligned with the MUMmer 3.23 package (Kurtz et al., 2004). Alignment statistics were computed with DNAdiff 1.3, distributed with MUMmer.

2.2. Transposable elements and repeat-induced point mutation (RIP)

To investigate whether TEs could have contributed to genome evolution and plasticity (Castanera et al., 2016) in *Knoxdaviesia*, families of these elements were identified with the REPET v2.2 package (Gilgado et al., 2005). The annotation files were filtered in Microsoft Office Excel 2010 (Microsoft Corp., Redmond, WA, USA) to remove duplicated regions where TE annotations overlap. Because short-read sequencing technologies often underestimate repeat content (Alkan et al., 2011), we determined whether repeated regions of the *K. capensis* and *K. proteae* genomes have higher-than-average sequence depth (indicative of sequence reads that collapse onto repeats). This was done by mapping the trimmed raw sequence reads (GenBank Accession *K. capensis*: SRX1453186, SRX1453795, SRX1453796; *K. proteae*: SRX1453891, SRX1453905, SRX1453906) back to the genomes with bowtie2 version 2.2.6 (Langmead and Salzberg, 2012). Sequence depth was calculated with Samtools 1.5 (Li et al., 2009). For each genome, the mean depth across the ten largest scaffolds was compared to the mean depth across TEs (identified by REPET) on those scaffolds, using R 3.2.5 (R Core Team, 2016). Low complexity (tandem) repeats were identified with RepeatScout 1.0.5 (Price et al., 2005) and mapped to the genome using RepeatMasker version open-4.0.5 (Smit et al., 2013-2015). Telomere repeats were identified following the methods of Fulnečková et al. (2013).

The genome sequences of the two *Knoxdaviesia* species were investigated for evidence of repeat-induced point mutation (RIP), a mechanism that limits the spread and diversification of TEs (Slotkin and Martienssen, 2007). A Perl script "RIP_index_calculation.pl" (available from https://github.com/hyphaltip/fungaltools/tree/master/scripts) was used to calculate the

composite RIP index (substrate index (TpA / ApT) subtracted from the product index (CpA + TpG / ApC + GpT)) in 500 base windows, sliding the window 100 bases at a time. A positive composite RIP index indicates a RIP-affected region (Lewis et al., 2009).

2.3. *Knoxdaviesia* orthologous groups and secretome

The total predicted *K. capensis* and *K. proteae* proteomes were clustered into orthologous groups (orthogroups) using OrthoMCL v2.0.9 (Li et al., 2003). Populous orthogroups (comprising ≥ 5 proteins in the two species combined) and species-specific proteins were annotated using BLASTp searches (Camacho et al., 2009) against the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov) non-redundant protein database and by identifying protein family (Pfam) domains with Reversed Position Specific (RPS)-BLAST (Finn et al., 2016). Blast2GO 3.3.5 (Götz et al., 2008) was used to parse the BLAST results. Species-specific proteins were subsequently grouped into categories according to the Functional Catalogue (FunCat) annotation scheme (Ruepp et al., 2004).

The putative secreted proteins of *K. capensis* and *K. proteae* were predicted following the protocol of Brown et al. (2012). Proteins with a secretion signal were identified with SignalP 4.1 (Petersen et al., 2011) and TargetP 1.1b (Emanuelsson et al., 2000). These were filtered by discarding proteins with transmembrane domains identified by TMHMM 2.0 (Krogh et al., 2001). All proteins with a secretion signal were interrogated for a glycosylphosphatidylinositol (GPI)-anchor site with PredGPI (http://gpcr2.biocomp.unibo.it/predgpi/, accessed 14 July 2016) (Pierleoni et al., 2008). The secretome was subsequently refined by excluding sequences lacking an initial methionine residue (once verified that the annotation was not erroneous) and GPI-anchor proteins. Protein location was predicted with ProtComp 9.0 (http://www.softberry.com), and only those with extracellular and unknown localizations were retained for further refinement. The localization of these remaining sequences was finally predicted with WoLFPSort 0.2 (https://github.com/fmaguire/WoLFPSort; Horton et al., 2007). Sequences with an extracellular score above 17 were labelled as proteins with a high likelihood of being secreted (Brown et al., 2012), while those that scored ≤ 17, but for which the extracellular environment remained the most probable localization, were labelled as proteins with a lower likelihood of secretion.

The final dataset of secreted proteins was annotated with BLASTp and RPS-BLAST, as described above, and the Pathogen Host Interactions (PHI) database version 4.1. Blast2GO was used to investigate potential functional enrichment in the secretomes of either *K. capensis*

or *K. proteae*. Since many plant pathogen effectors have been described as cysteine-rich proteins (Lu and Edwards, 2016; Rep, 2005a), the cysteine content of the secreted proteins was determined from the amino acid composition statistics of each protein as calculated by the EMBOSS PEPSTATS tool (Rice et al., 2000). The *K. capensis* and *K. proteae* secreted proteins were compared according to their OrthoMCL orthogroups as an indication of which secreted proteins are likely to be functionally similar or unique. As with the species-specific proteins, secreted orthogroups and unique secreted proteins were categorised according to the FunCat database.

2.4. Secondary metabolite gene clusters

Secondary metabolite clusters were predicted with the online tool antiSMASH 3.0.5 (Blin et al., 2013; Medema et al., 2011; Weber et al., 2015) using the whole genome sequences as input. Genomic regions in which secondary metabolite clusters had been predicted were investigated by extracting the protein predictions of the putative biosynthetic genes, as well as approximately five upstream and five downstream flanking genes. These were subsequently annotated with BLASTp and Blast2GO as described above. The presence of the necessary catalytic domains in the backbone non-ribosomal peptide synthase (NRPS) and polyketide synthase (PKS) proteins (Keller et al., 2005) were verified by searching for protein family (Pfam; Finn et al., 2016) domains with HMMER 2.3.2 (Eddy, 2011).

One of the PKS genes in *K. proteae* appeared to be non-functional due to deletions and frameshifts causing nonsense mutations (see results below). To test whether these mutations were isolate-specific or whether the mutated PKS gene is characteristic of *K. proteae*, primers SM9-1F (5'-SGATACGTTGGTGTTCATGG-3') and SM9-1R (5'-ATCTTGGGAGCCTACTGCAA-3') were designed to amplify a diagnostic region in *K. proteae* and *K. capensis*. These primers were applied to a set of 11 *K. capensis* and 11 *K. proteae* isolates, representing all populations of these species that had previously been collected from the CFR (Aylward et al., 2014, 2015; 2016a; 2017a). Each 20 µl PCR reaction consisted of 10 µl Ampliqon Taq DNA Polymerase 2x Master Mix (Ampliqon, Denmark), 0.5 µM of each primer and a final $MgCl_2$ concentration of 3.4 mM. The cycling conditions included an initial denaturation at 95°C for 3 min, followed by 35 cycles of 94°C for 30 s, annealing at 60°C for 30 s and extension at 72°C for 45 s. Final extension was 72°C for 10 min.

3. Results

3.1. *Knoxdaviesia* genome overview

3.1.1. Genome structure

More than 85 % of bases in the two *Knoxdaviesia* genomes were aligned to each other with an average nucleotide identity of 89.5 % (Table 1). Ninety-eight of the *K. proteae* scaffolds (containing 99.3 % of the *K. proteae* genome) were aligned to 24 *K. capensis* scaffolds (99.9 % of the *K. capensis* genome), with a mean alignment length of approximately 3702 bases (Fig. 2). Several large inversions (100 – 700 kb) were apparent in the alignments. The unaligned scaffolds were small (< 26 kb) and had a low GC percentage (< 40 %). Uninterrupted diagonals of ≥ 20 kb in MUMmer genome dotplots have been proposed to constitute macrosynteny (Hane et al., 2011). Considering that all scaffolds in Fig 2(B) were > 500 kb in length, these lengthy diagonals are indicative of macrosynteny between the two *Knoxdaviesia* genomes.

Table 1 Whole genome nucleotide alignment statistics for *Knoxdaviesia capensis* and *K. proteae*.[a]

	K. capensis	***K. proteae***
Total scaffolds	29	133
Aligned scaffolds[b]	24 (82.76%)	98 (73.68%)
Unaligned scaffolds	5 (17.24%)	35 (26.32%)
Total Bases	35537816	35489142
Aligned Bases[b]	30327269 (85.34%)	30341562 (85.50%)
Unaligned Bases	5210547 (14.66%)	5147580 (14.50%)
1-to-1 alignments	8198	8198
Total Length (bases)	30342580	30352057
Mean Length (bases)	3701.22	3702.37
Mean Identity	89.45%	89.45%

[a] Statistics were computed with DNAdiff (Kurtz *et al.* 2004)

[b] Nucleotide alignments were computed with nucmer (Kurtz, et al. 2004)

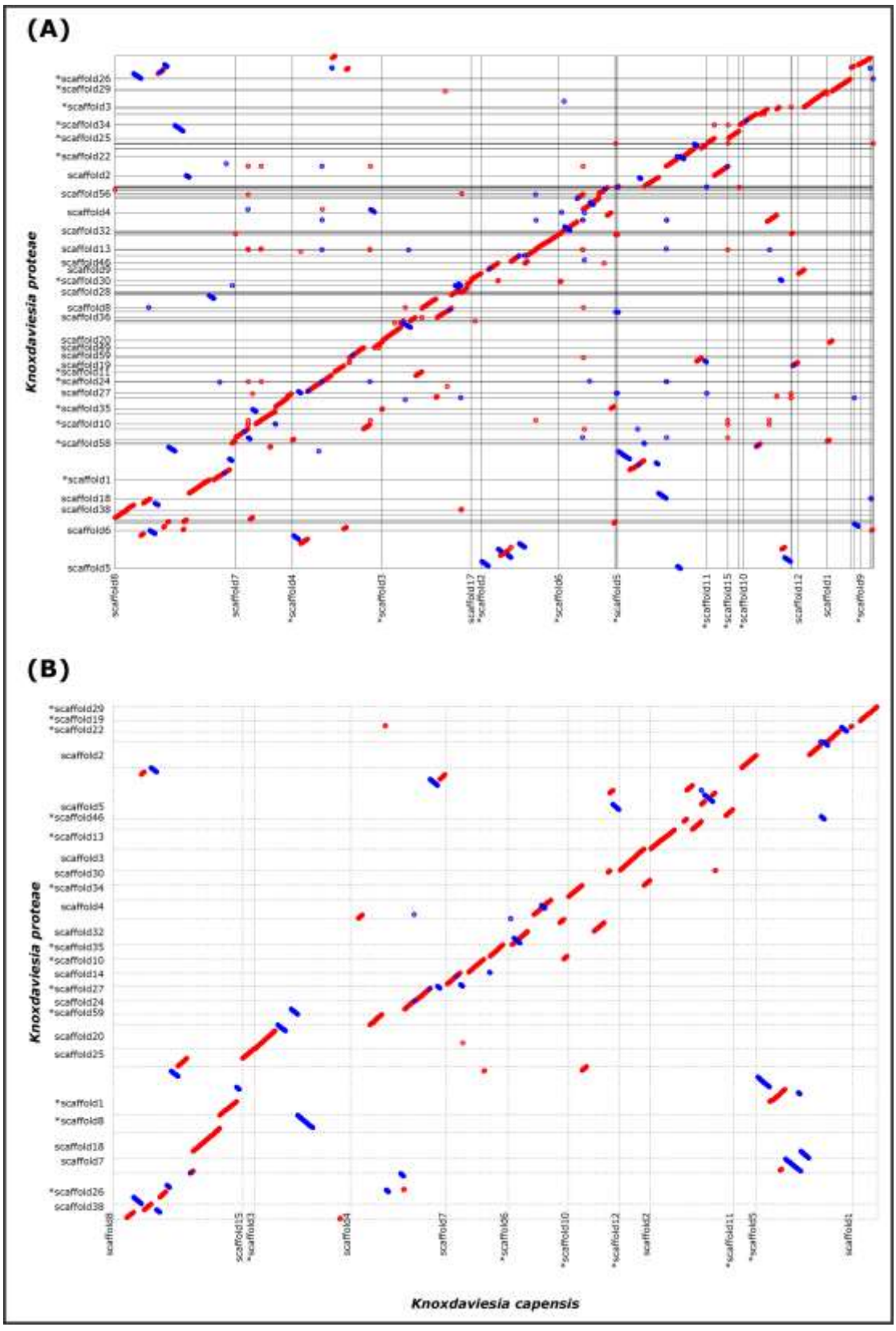

Fig. 2. Dotplots representing nucleotide alignments of the *Knoxdaviesia capensis* and *K. proteae* genomes. The alignment of all scaffolds (A) as well as scaffolds > 500 kb (B) are shown. Vertical grey bars indicate the start of *K. capensis* scaffolds and horizontal bars the start of *K. proteae* scaffolds. Only the names of scaffolds > 400 kb are annotated on the axes. Red diagonals represent alignments in the same direction, whereas blue diagonals indicate a reverse orientation in one of the species.

3.1.2. Repeats, transposable elements and repeat-induced point mutation (RIP)

REPET and RepeatScout identified 9.9 % of the *K. capensis* genome and 10.8 % of the *K. proteae* genome as repetitive (Table 2). Class I retrotransposons, specifically *Gypsy* and *Copia* Long Terminal Repeat (LTR) transposons, constituted more than half of the repetitive content in both genomes. Low complexity (short tandem) repeats occupied approximately 3 % of each genome, whereas Class II DNA transposons were rare, comprising only up to 0.2 %. *Knoxdaviesia proteae* had a slightly greater diversity of repeat families and 1066 additional individual TEs than *K. capensis*. Although the additional TEs could be partly explained by the fragmented nature of the *K. proteae* genome assembly, increased numbers of the non-autonomous Class I elements, large retrotransposon derivatives (LARD) and terminal repeat retrotransposons in miniature (TRIM), were especially apparent in this species (Table 2). The increase in LARD and TRIM elements was reflected in that they occupied more than double the number of sequence bases in *K. proteae* compared to *K. capensis*. The Class I order DIRS that made up almost 0.2 % of the *K. capensis* genome were, however, not identified in *K. proteae*.

The proportion of each genome affected by RIP was summarized by considering the number of sliding windows with a positive RIP composite index and dividing them by the total number of sliding windows. Following this approach, 44.5 % of the entire *K. capensis* genome and 42.7 % of the entire *K. proteae* genome was affected by RIP. Many of the smaller sized scaffolds, had positive RIP scores over their entire length (Supplementary File 1), alluding to their repetitive nature. All the scaffolds of *K. capensis* that could not be aligned to *K. proteae* were affected by RIP across their entire length, whereas 34 of the 35 unaligned *K. proteae* scaffolds had > 80 % RIP. Across the largest scaffolds (> 100 kb), considerably fewer regions were affected by RIP; 16.6 % in *K. capensis* and 13.9 % in *K. proteae*, and thus more than the estimated repetitive proportion of each genome.

A longer variant of the TTAGGG "vertebrate" telomere repeat (Fulnečková et al., 2013; Traut et al., 2007) was identified at one terminal of seven *K. capensis* scaffolds and 10 of *K. proteae* (Table S1). This TTAGGG telomeric unit is also found in some Pezizomycotina fungi (Podlevsky et al., 2008) and *Aspergillus oryzae* is known to have a 12 bp (TTAGGGTCAACA) variant (Kusumoto et al., 2003). The putative *Knoxdaviesia* telomere repeat is composed of 10 nucleotides with the sequence TTAGGGTTAC (Kusumoto et al., 2003). In *K. capensis*, TTAGGGTTAC telomere repeats were also identified within a TE on scaffold 3 and approximately 3 kb upstream of the 3' terminal in scaffold 20 (Table S1). The latter may suggest that the 3' end of scaffold 20 was assembled incorrectly.

Table 2 Repetitive elements identified in the genomes of *Knoxdaviesia capensis* and *K. proteae*.

	Knoxdaviesia capensis				*Knoxdaviesia proteae*			
Repeats[a]	**Repeat families**[b]	**Copies in genome**[c]	**Length occupied (bp)**[d]	**% of genome**	**Repeat families**[b]	**Copies in genome**[c]	**Length occupied (bp)**[d]	**% of genome**
Class I elements								
DIRS	1	27	64667	0.18	0	-	-	-
LTR *(Gypsy & Copia)*	57	506	2048344	5.76	72	761	2205424	6.21
LTR *(LARD)**	3	75	160706	0.45	2	759	354216	1.00
LTR *(TRIM)**	5	64	35330	0.10	12	283	136517	0.38
unknown	1	46	100804	0.28	1	1	1192	0.00
Class I Total	*67*	*718*	*2409851*	6.78	*87*	*1804*	*2697349*	7.60
Class II elements								
TIR (*MITE*)*	2	23	8417	0.02	1	12	4150	0.01
TIR (*Tc1-Mariner*)	3	69	53180	0.15	6	55	48147	0.14
TIR (*unknown*)	1	10	8164	0.02	4	15	6880	0.02
Class II Total	*6*	*102*	*69761*	0.20	*11*	*82*	*59177*	0.17
Low complexity repeats[d]	-	24894	1042985	2.93	-	25704	1089080	3.07
Total repeats (TEs and low complexity)	***73***	***25714***	***3522597***	***9.91***	***98***	***27590***	***3845606***	***10.84***

[a] DIRS = *Dictyostelium* intermediate repeat sequence; LTR = long terminal repeat; LARD = large retrotransposon derivative; TRIM = terminal repeat retrotransposons in miniature; TIR = terminal inverted repeat; MITE = miniature inverted-repeat transposable element; TEs = transposable elements; asterisks (*) indicate non-autonomous elements

[b] Identified by TEdenovo package in REPET and clustered with NCBI Blastclust (Gilgado *et al.* 2005)

[c] Identified by TEannot package in REPET (Gilgado, et al. 2005)

[d] Identified with RepeatScout (Price *et al.* 2005) and annotated with RepeatMasker (Smit *et al.* 2013-2015)

The TE-containing regions had a greater than average sequence depth, which confirmed that the repeat content in both the *K. capensis* and *K. proteae* genomes is underestimated. The average sequencing depth was > 200 in both genomes (*K. capensis* 213.3; *K. proteae* 285.5), but with high standard deviations (224.5 and 261.8, respectively). By repeating the calculation using only the ten largest scaffolds in each genome, and thereby excluding smaller scaffolds that may not have been assembled further due to high repeat content, the standard deviations decreased drastically (*K. capensis* 206.5 x ± 6.8; *K. proteae* 268.4 ± 18.3). The mean sequence depth across the TEs identified on the ten largest scaffolds was also determined (*K. capensis* 303.2 ± 48.9; *K. proteae* 478.4 ± 64.3) and a Welch t-test, performed in R, confirmed that the depth across these regions differed significantly ($P < 0.001$) from the overall sequencing depth (*K. capensis*: $t = -6.1917$, $df = 9.3487$; *K. proteae*: $t = -9.9306$, $df = 10.448$). Additionally, visualisation of the read mappings in the Integrative Genomics Viewer 2.3.97 (Robinson et al., 2011; Thorvaldsdóttir et al., 2013), indicated more mismatches in regions containing TEs compared to non-repetitive regions (see Supplementary File 2), providing additional evidence that some repeats may have been overlooked. Inferences from the repeat content remain valid, however, since the actual number of repeats in these genomes would be more, and not less, than our estimation. Similarly, the RIP indexes would be calculated across a smaller number of repeated elements and are, therefore, also likely to underestimate, rather than overestimate, the extent of RIP in these genomes.

3.1.3. Knoxdaviesia *orthogroups*

OrthoMCL clustered 7116 (89.6 %) of the *K. capensis* and 7098 (86.6 %) of the *K. proteae* proteome into 6982 shared orthogroups (Table 3). Eighteen shared orthogroups comprised ≥ 5 proteins (Table S2). The majority (8/18) of these populous orthogroups were associated with membrane transport, including two groups of ABC transporters, two groups that transport small proteins (oligopeptides), a general major facilitator superfamily (MFS) transporter group and three groups involved in the transport of ions or inorganic molecules. Notably, three other orthogroups contained Glycoside Hydrolase (GH) family 3, GH55 and glycosyl transferase (GT) 2 carbohydrate-active enzymes (CAZys).

Knoxdaviesia capensis and *K. proteae* encoded 825 and 1075 species-specific proteins, respectively. In both species, OrthoMCL identified only two orthogroups within the species-specific set. Of the species-specific proteins, only 258 (31.3 %) in *K. capensis* and 228 (21.2 %) in *K. proteae* could be sorted into meaningful FunCat categories, since a large proportion of these proteins did not have significant BLASTp hits or were hypothetical / uncharacterised (Table S3).

Table 3 Number of predicted shared orthologous groups and secreted proteins in *Knoxdaviesia capensis* and *K. proteae*.[a]

	K. capensis	*K. proteae*
Total predicted proteins	7940	8173
Shared orthogroups	6982	
Proteins in shared orthogroups	7116 (89.6 %)	7098 (86.6 %)
Species-specific proteins	825 (10.4 %)	1075 (13.2 %)
Proteins with putative secretion signals	1163	1166
Excluded from secretome		
GPI-anchor, TM region, no Met[b]	593	598
Localization not extracellular[c]	204	214
Annotation not extracellular[d]	27	28
Refined secretome		
Shared orthogroups[e]	246	244
Unique secreted orthogroups [f]	65	49
Species-specific proteins[g]	28	33
Proportion of total predicted proteins	*4.27%*	*3.99%*

[a] Prediction methods are outlined in the text
[b] Proteins containing a glycosylphosphatidylinositol (GPI) anchor and/or transmembrane (TM) region or lacking an initial Methionine residue
[c] Proteins whose most likely final destination is not extracellular
[d] Proteins with functional annotations inconsistent with secreted proteins
[e] Secreted proteins with a putative secreted ortholog in the other species
[f] Secreted proteins with an ortholog in the other species that is not predicted as secreted
[g] Secreted proteins without an ortholog in the other species

The singleton proteins represent the divergence that has taken place between the *K. capensis* and *K. proteae* proteomes. Although species-specific, these proteins could be divided into 17 FunCat categories, with roughly equal proportions of *K. capensis* and *K. proteae* proteins in each category (Table 4; Supplementary File 3 and 4). The majority are putatively involved in metabolism (FunCat 01; 27.5 % and 26.3 % in *K. capensis* and *K. proteae*, respectively) and binding to substrates or cofactors (FunCat 16; 11.2 % and 12.3 %, respectively). Regarding metabolism, the carbohydrate and secondary metabolism sub-categories were the largest, with the latter containing several enzymes potentially able to protect the two species by degrading harmful substrates (Supplementary File 5). In addition to the populous orthogroups with transporter annotations (Table S2), a cellular transport role (FunCat 20) was also predicted for 9.7 % and 8.8 % of *K. capensis* and *K. proteae* species-specific proteins. Three other categories that comprise similar, but species-specific, proteins suggest some divergence in signalling pathways and reproductive cues since speciation; namely, FunCat 30.01 (“cellular

Table 4 Functional categories present in the species-specific proteins and predicted secretome of *Knoxdaviesia capensis* and *K. proteae*

	Species-specific proteins		Putatively secreted proteins			
			K. capensis		*K. proteae*	
FunCat Category[a]	***K. capensis***	***K. proteae***	**Total (%)**	**Species-specific (%)**	**Total (%)**	**Species-specific (%)**
01 METABOLISM	72 (27.9%)	60 (26.3%)	184 (54.3%)	8 (28.6%)	165 (73%)	3 (9.1%)
02 ENERGY	18 (7.0%)	6 (2.6%)	-	-	-	-
10 CELL CYCLE AND DNA PROCESSING	10 (3.9%)	15 (6.6%)	-	-	-	-
11 TRANSCRIPTION	20 (7.8%)	21 (9.2%)	-	-	-	-
12 PROTEIN SYNTHESIS	6 (2.3%)	11 (4.8%)	-	-	-	-
14 PROTEIN FATE	15 (5.8%)	20 (8.8%)	43 (12.7%)	2 (7.1%)	36 (15.9%)	0
16 PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT	29 (11.2%)	28 (12.3%)	21 (6.2%)	3 (10.7%)	17 (7.5%)	1 (3.0%)
18 REGULATION OF METABOLISM AND PROTEIN FUNCTION	-	-	1 (0.3%)	0	1 (0.4%)	0
20 CELLULAR TRANSPORT, TRANSPORT FACILITIES AND TRANSPORT ROUTES	25 (9.7%)	20 (8.8%)	-	-	-	-
30 CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM	13 (5.0%)	18 (7.9%)	10 (2.9%)	0	12 (5.3%)	1 (3.0%)
32 CELL RESCUE, DEFENCE AND VIRULENCE	7 (2.7%)	5 (2.2%)	27 (8%)	1 (3.6%)	23 (10.2%)	0
34 INTERACTION WITH THE ENVIRONMENT	5 (1.9%)	1 (0.4%)	8 (2.4%)	1 (3.6%)	8 (3.5%)	0
38 TRANSPOSABLE ELEMENTS, VIRAL AND PLASMID PROTEINS	1 (0.4%)	3 (1.3%)	-	-	-	-
40 CELL FATE	12 (4.7%)	5 (2.2%)	4 (1.2%)	1 (3.6%)	4 (1.8%)	0
41 DEVELOPMENT	3 (1.2%)	2 (0.9%)	-	-	-	-
42 BIOGENESIS OF CELLULAR COMPONENTS	1 (0.4%)	0	-	-	-	-
70 SUBCELLULAR LOCALIZATION	17 (6.6%)	10 (4.4%)	-	-	-	-
98 CLASSIFICATION NOT YET CLEAR-CUT	4 (1.6%)	3 (1.3%)	25 (7.4%)	2 (7.1%)	28 (12.4%)	3 (9.1%)
99 UNCLASSIFIED PROTEINS	excluded[b]		16 (4.7%)	10 (35.7%)	32 (14.2%)	24 (72.7%)
TOTAL	*258*	*228*	*339*	*28*	*326*	*33*

[a] FunCat = Functional Catalogue (Ruepp et al. 2004)

[b] The majority of genome-wide species-specific proteins could not be annotated (>50%) or were annotated as hypothetical / predicted proteins (>20%); see Table S3. These are excluded for clarity in the proportions of other categories.

signalling") containing various protein kinases, FunCat 40 ("cell fate") containing heterokaryon incompatibility proteins that mediate programme cell death and FunCat 41.01 (fungal "mating") comprised of a few proteins that influence reproduction.

The most apparent differences between *K. capensis* and *K. proteae* were in the number, rather than the identity, of proteins in the different categories (Table 4; Supplementary File 3). *Knoxdaviesia capensis* had several additional electron transport chain components (FunCat 02), five proteins that mediate cell adhesion compared to one in *K. proteae* (FunCat 34) and six more species-specific proteins classified as extracellular or secreted (FunCat 70). In contrast, *K. proteae* had additional enzymes for DNA processing (FunCat 10) and protein degradation (FunCat 14).

3.2. The *Knoxdaviesia* secretome

Approximately 4 % of the total putative proteins of *K. capensis* and *K. proteae*, were predicted to be secreted (Table 3). Initial identification of protein signal peptides by SignalP and TargetP identified more than 1 000 potentially secreted proteins in each of the *Knoxdaviesia* species. By excluding proteins that contain transmembrane domains, GPI-anchor sites and proteins that do not start with a methionine residue, this number was reduced to less than 600 for each species (Table 3; Supplementary File 6). Analysis of protein localization indicated < 400 proteins from each species that were likely secreted to the external environment. For *K. capensis*, WolfPSort predicted 248 proteins that have a high likelihood of being secreted (score > 17) and 108 proteins with lower likelihoods of being secreted. Slightly fewer secreted proteins were included in the *K. proteae* "high likelihood" dataset (229), whereas 125 *K. proteae* proteins have lower likelihoods of secretion. Annotation of these putative secreted proteins revealed classifications in both the high and low likelihood datasets that are not secreted, such as protein translation machinery and transmembrane transport proteins. Based on annotation, a further 27 proteins in *K. capensis* and 28 in *K. proteae* were excluded from the secretome (Table 3; Supplementary File 6).

3.2.1. Functional annotation of secretome proteins

More than 90 % of the predicted secreted proteins in the high likelihood datasets could be annotated, whereas the lower likelihood datasets contained > 10 % unannotated proteins. No significant enrichment of functional annotations between the *K. capensis* and *K. proteae* datasets could be detected in Blast2GO at a critical value of 0.05. More than 70 % of the secreted proteins in each species formed part of the shared set consisting of 231 orthogroups.

Species-specific proteins comprised 8.5 % in *K. capensis* and 10.4 % in *K. proteae*, of which the majority (32.1 % and 70.6 %, respectively) were not annotated. The remaining 19.7 % *K. capensis* and 18.0 % in *K. proteae* were unique within the secretome dataset (unique secreted), but had an ortholog in the other species that was not predicted as secreted. Further investigation revealed that most of these orthologs (61.5 % in *K. capensis* and 39.0 % in *K. proteae*) lacked an identifiable signal peptide, while the remainder were excluded due to a transmembrane domain, GPI anchor site or location prediction (Supplementary File 6). Ten of the *K. capensis* orthologs were excluded due to a single transmembrane helix at the N-terminal. Since TMHMM may occasionally recognize signal peptides at the N-terminal as a transmembrane domain (Krogh et al., 2001) and WoLFPSort predicted the localization of these orthologs as extracellular, they were included in the secretome. The shared final secretome dataset, therefore, contained 241 orthogroups (72.4 % and 74.4 % of the *K. capensis* and *K. proteae* secretome, respectively; Supplementary File 7).

Based on their Blast2GO annotations and orthologous groups, the secreted proteins were divided into 10 of the "top level" FunCat hierarchical categories (Table 4; Supplementary File 7 and 8), including unclassified proteins and those with an unclear function. Most *K. capensis* and *K. proteae* proteins in both the shared (56.5 % and 56.6 %, respectively) and unique secreted (55.4 % and 49.0 %, respectively) datasets were associated with metabolism, including amino acid, carbohydrate and lipid metabolism (FunCat 01). Of these, metabolism of carbon compounds (FunCat 01.05) was by far the largest subcategory. The second largest category comprised proteolytic enzymes (FunCat 14.13 under "Protein Fate"). Together, these categories (01 and 14.13) indicate that the majority of all *K. capensis* (65.2 %) and *K. proteae* (60.4 %) secreted proteins are dedicated to degrading organic substrates, such as plant cell wall and plasma membrane components.

The largest category of secreted proteins unique to one of the *Knoxdaviesia* species was metabolism, specifically carbohydrate metabolism (Supplementary File 7). Nineteen unique secreted and six species-specific carbohydrate metabolism proteins could be identified from *K. capensis*, whereas *K. proteae* had only 11 and two, respectively. *Knoxdaviesia capensis* also had three additional unique secreted and two species-specific proteolytic enzymes. Other categories with notable differences between the two species include proteins related to disease, virulence and defence (FunCat 32.05) and secondary metabolism (FunCat 1.20) (discussed below).

3.2.2. *Potential secreted virulence / defence proteins in* Knoxdaviesia

In *K. capensis* and *K. proteae,* 107 and 101 secreted proteins, respectively, composed of 84 shared secreted orthogroups had significant hits to the Pathogen Host Interactions (PHI) database (Table S4). These orthogroups were primarily related to metabolic and proteolytic activity. In 74/84 orthogroups, the *K. capensis* and *K. proteae* orthologs had the same PHI-BLAST classification, comprised primarily of reduced virulence (35/84) or unaffected pathogenicity (33/83) mutant phenotypes. Most of the differences between orthologs were due to only one ortholog having a match to the PHI database. Two orthogroups consisted of putative necrosis inducing (NPP1) family proteins, where the best *K. capensis* hit was a protein that does not affect pathogenicity and the best *K. proteae* hit is an effector or a protein with a reduced virulence phenotype. Such a difference in the best hits was also true for two other orthogroups, serine endopeptidases where one member did not have a PHI match and the two others had reduced virulence and unaffected mutant phenotypes, and two putative cerato-platanin phytotoxic proteins of which one had a reduced virulence and the other an unaffected mutant phenotype. Only two secreted proteins in *K. capensis* and three in *K. proteae* were homologous to proteins required for pathogenicity (loss of pathogenicity mutant phenotype).

Less than 2 % of the secreted proteins in each *Knoxdaviesia* species were homologous to known effector proteins and approximately a third of the secreted proteins could be classified as "cysteine-rich". Five potential effector proteins were present in *K. capensis* and three in *K. proteae*. The species-specific effector identified in *K. capensis* was annotated as a lytic polysaccharide monooxygenase, which belongs to a group of enzymes that introduce internal breaks in polysaccharides, such as cellulose, by oxidation (Hemsworth et al., 2015). This protein was, however, not identified as an AA9 family enzyme in the CAZy database (Supplementary File 8). For *K. capensis* and *K. proteae*, 33 and 32 shared secreted proteins, respectively, had a cysteine content of ≥ 2 % and consisted of < 200 amino acids and may, therefore, be classified as "small secreted cysteine-rich proteins" (SSCPs; Lu and Edwards, 2016). A further eight *K. capensis* and 13 *K. proteae* SSCPs were identified in the unique secreted and singleton datasets (Table S4). Most of these SSCPs are involved in metabolism (FunCat 01) and cell defence (FunCat 32), some in proteolysis (FunCat 14.13) or cell communication (FunCat 30 and 34) and a few have a predicted binding function (FunCat 16). Hypothetical / unclassified proteins (FunCat 98 and 99) comprise 24.4 % of *K. capensis* and 27.3% of *K. proteae* SSCPs (Supplementary File 9).

Table 5 Secreted proteins potentially involved in cell defence against microbial competitors[a].

Putative defence protein[b]	Description	*Kc*[c]	*Kp*[c]
Production of reactive oxygen species			
Glucose oxidase	Hydrogen peroxide production	5**ę	2
Secondary metabolite biosynthesis			
Snoal-like polyketide cyclase family	Nogalamycin biosynthesis	1	1
Tyrosinase	Pigment production	4***	2*
2OG-Fe(II)oxygenase superfamily	Antibiotic biosynthesis	2*	1
Other			
Peptidoglycan-binding lysin domain		1*	0
Detoxification			
Aflatoxin b1 aldehyde reductase member 2	Detoxifies ketones and aldehydes	1*	0
Carboxylesterase	Xenobiotic metabolism	1	1
Chlorocatechol 1,2-dioxygenase	Cleaves catechol (aromatic compound)	1	1
FAD-containing monooxygenase	Oxidises xenobiotics	0	1*
Fumonisin B1 esterase	Degrades fumonisin B1	0	1*
Haloacid dehalogenase / Epoxide hydrolase family	Hydrolyses halogenated aromatic compounds	2	2
Intradiol ring-cleavage dioxygenase	Cleaves aromatic rings	1	1
Nitroreductase	Reduction of nitrosubstituted compounds	2*	1
Superoxide dismutase	Dismutates superoxide	2	2
TOTAL		**23**	**16**

[a] Proteins were defined as potentially involved in cell defence against microbes if they belonged to FunCat 01.20 ("secondary metabolism") or 32.07 ("detoxification"). FunCat 32.05 ("disease, virulence and defense") was not included as the annotations of these proteins suggested interactions with the plant host only, although the putative peptidoglycan-binding protein of *K. capensis* was the single exception.

[b] FAD = flavin adenine dinucleotide; OG-Fe(II) = oxoglutarate/iron-dependent

[c] *Kc* = *Knoxdaviesia capensis*; *Kp* = *K. proteae*; each asterisk (*) denotes a protein from the unique secreted dataset, whereas ę denotes a species-specific protein.

Enzymes potentially involved in virulence and cell defence were identified in three FunCat subcategories: 01.20 secondary metabolism, 32.05 disease, virulence and defence and 32.07 detoxification. FunCat 32.05 comprised 13 shared orthogroups with putative roles in fungal pathogenesis against a plant host. *Knoxdaviesia capensis* had a further four of these proteins unique to the secreted dataset of which only one was not related to the plant host, but had a peptidoglycan-binding domain. The other two subcategories contained 13 types of enzymes that potentially confront microbial competition in the environment (Table 5). These were glucose oxidase enzymes that produce hydrogen peroxide, nine different enzymes that detoxify antimicrobials or reactive compounds, two that potentially produce antimicrobial

compounds and one involved in pigment production. Of these defence-related proteins, 23 were identified in *K. capensis* and 16 in *K. proteae*. The Aflatoxin B1 aldehyde reductase, Fumonisin B esterase and FAD-containing monooxygenase proteins are part of the unique secereted dataset and, therefore, have orthologs in the other species, although these are apparently not secreted. *Knoxdaviesia capensis* had a species-specific glucose oxidase as well as additional glucose oxidase, tyrosinase and 2OG-Fe(II)oxygenase superfamily proteins in the unique secreted dataset.

3.3. Secondary metabolite gene clusters

AntiSMASH identified 11 secondary metabolite biosynthesis clusters in *K. capensis* and 10 in *K. proteae*. Three NRPS clusters (Fig. 3), two terpene clusters and one unclassified cluster (Fig. 4) were found in both species. In two *K. proteae* NRPS clusters, however, the adenylation (A) domain and / or peptidyl carrier protein (P) domain of the synthase gene could not be detected (Fig. 3), implying that these genes may not be functional (Keller et al., 2005). Of the five identified PKS clusters (Fig. 5), the PKS gene in one *K. proteae* cluster (T1PKS-4) was split into multiple ORFs and, therefore, not identified by AntiSMASH. Only one of these truncated ORFs in *K. proteae* had a detectable β-ketoacyl synthase, C-terminal domain (Fig. 5) and two large deletions (303 and 530 bp) were apparent at the 3' end of the gene. The larger deletion was targeted for a diagnostic PCR, yielding a 280 bp product in *K. proteae*, but an 820 bp product in *K. capensis* (Fig. 5). Amplification of this PKS gene region confirmed a deletion in all *K. proteae* isolates (Table S5).

The homologous clusters between *K. capensis* and *K. proteae* were identified by considering the upstream and downstream flanking regions. In four of the clusters (NRPS-1, T1PKS-2, T1PKS-3 and T1PKS-4), retrotransposons likely affected the gene content of their flanking regions. In clusters T1PKS-2 and T1PKS-3 the synteny of genes beyond the TE insertion appeared to be maintained. In cluster T1PKS-4, the synteny of the genes immediately upstream was minimally affected, however, synteny was lost downstream of the truncated *K. proteae* PKS gene. Further investigation indicated an inversion, with the homologous region of the downstream genes in *K. capensis* found further upstream in *K. proteae* (Fig. 5). In cluster NRPS-1, the genes further downstream of the *K. proteae* retrotransposons did not coincide with the remainder of the *K. capensis* biosynthetic genes (results not shown). A tBLASTn search against the *K. proteae* genome revealed that these biosynthesis genes were

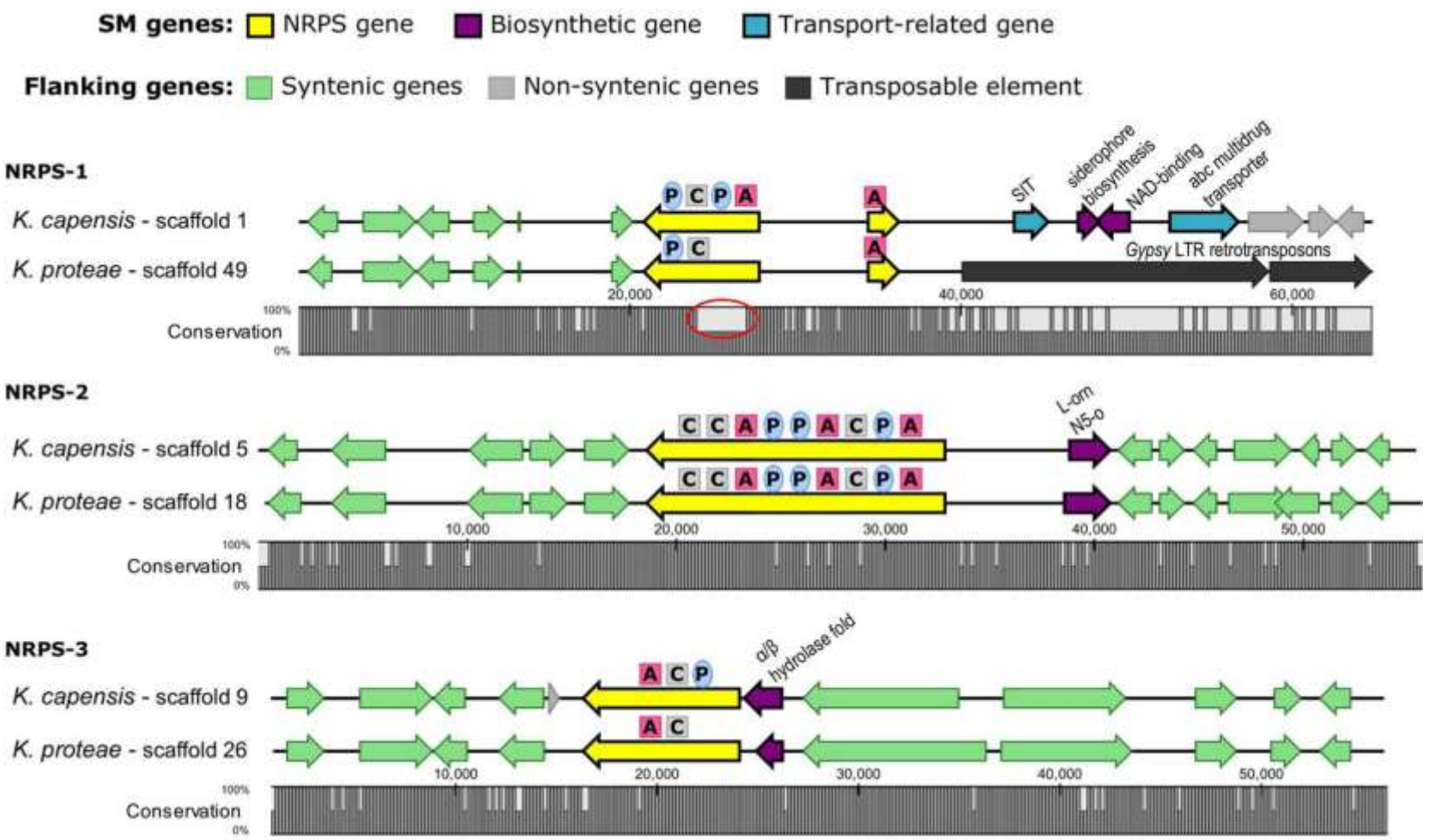

Fig. 3. Alignment between the homologous non-ribosomal peptide synthase (NRPS) secondary metabolite (SM) biosynthesis clusters in *Knoxdaviesia capensis* and *K. proteae*. The identified NRPS domains are indicated above the genes: A = adenylation (PF00501), C = condensation (PF00668), P = peptidyl carrier protein (PF00550). Other putative SM genes and long terminal repeat (LTR) retrotransposons have been annotated: L-orn N5-o = L-ornithine-N5-oxygenase; SIT = siderophore iron transporter. The red circle indicates missing sequence data in *K. proteae*.

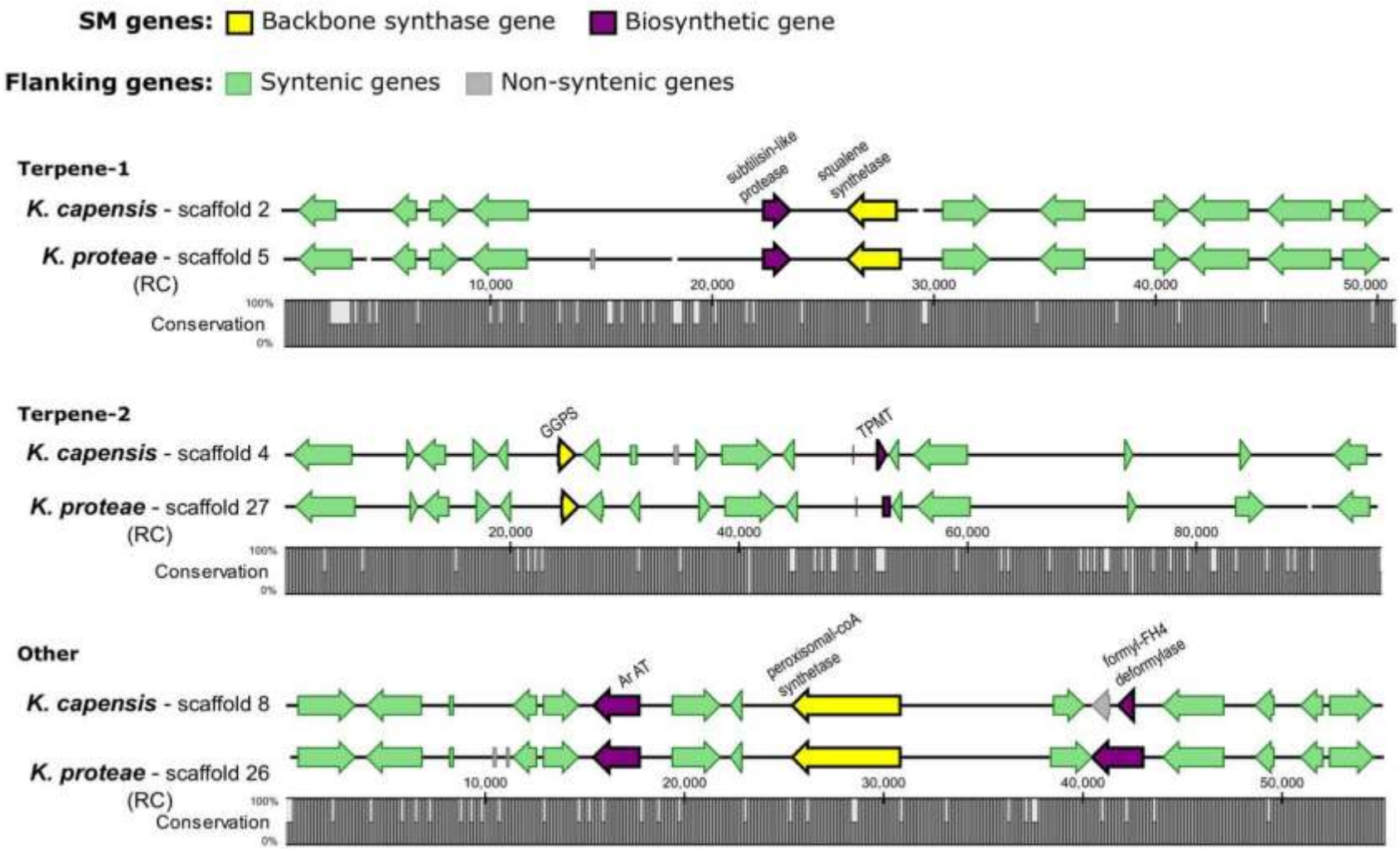

Fig. 4. Alignment of the two homologous terpene biosynthesis clusters and one peroxisomal-coenzyme A synthetase cluster of *Knoxdaviesia capensis* and *K. proteae*. Other putative SM genes have been annotated: ArAT = aromatic amino acid aminotransferase; formyl-FH4 = formyltetrahydrofolate; GGPS = geranylgeranyl pyrophosphate synthetase; TPMT = thiopurine S-methyltransferase.

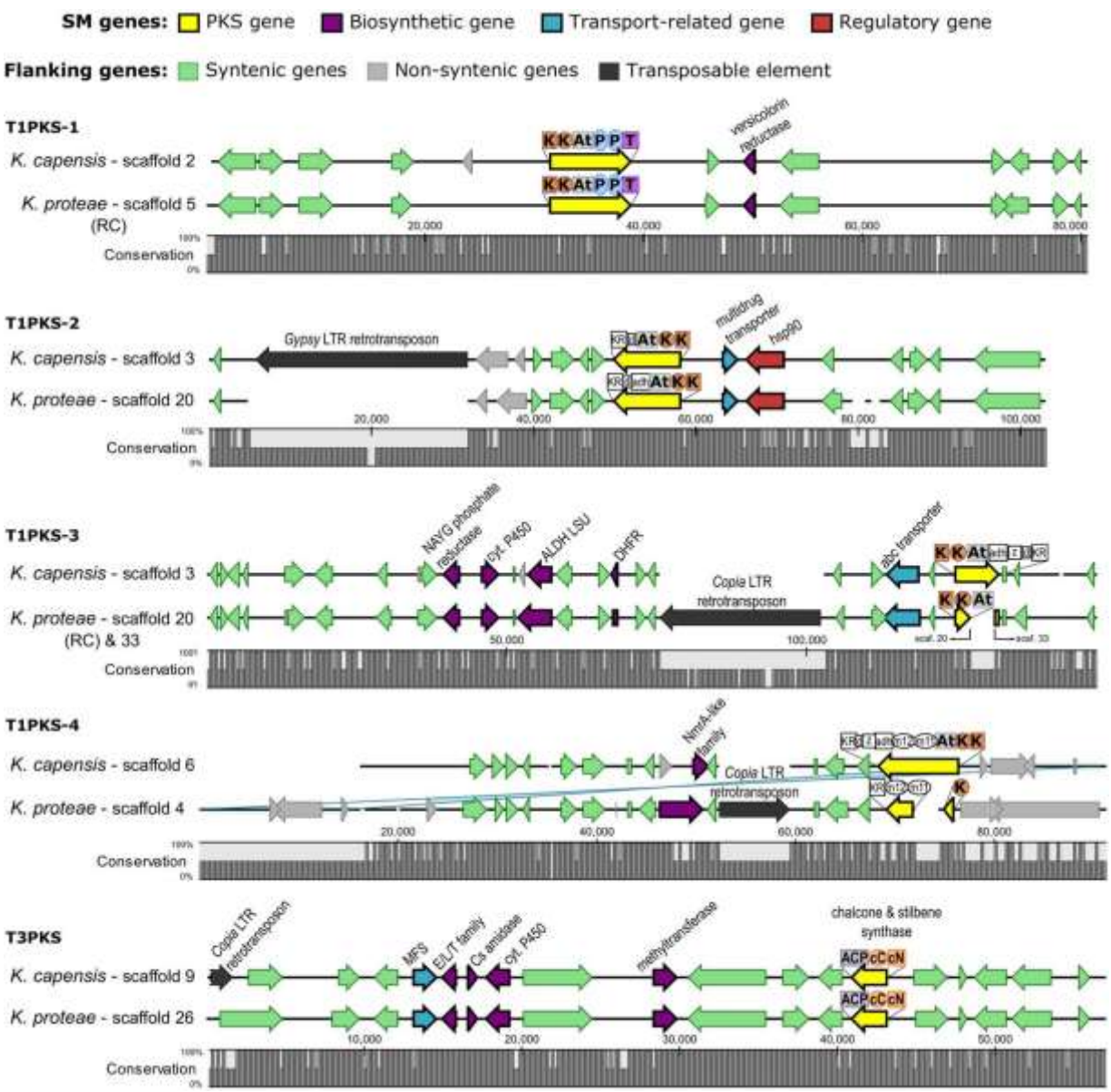

Fig. 5. Alignment between the homologous polyketide synthase (PKS) secondary metabolite (SM) biosynthesis clusters of *Knoxdaviesia capensis* and *K. proteae*. Essential PKS domains are indicated in colour above the genes, whereas accessory domains are white: ACP = acyl carrier protein (PF08541); adh = alcohol dehydrogenase (PF08240); At = acyltransferase (PF00698); cC = chalcone and stilbene synthase, C-terminal (PF02797); cN = chalcone and stilbene synthase, N-terminal (PF00195); d = short-chain dehydrogenases/reductase (PF00106); K, boxed = β-ketoacyl synthase, N-terminal (PF00109); K, circled = β-ketoacyl synthase, C-terminal (PF02801); KR = β-ketoacyl-ACP reductase (PF08659); m11 = methyltransferase (PF08241); m12 = methyltransferase (PF08242); P = peptidyl carrier protein (PF00550); T = thioesterase (PF00975); z = zinc finger (PF00172). Other putative SM genes and long terminal repeat (LTR) retrotransposons have been annotated: ALDH LSU = L-aminoadipate-semialdehyde dehydrogenase large subunit; Cyt. = cytochrome; DHFR = dihydrofolate reductase; E/L/T = esterase lipase thioesterase; Cs amindase = N-carbamoylsarcosine amidase; hsp90 = heat shock protein 90; MFS = major facilitator superfamily; NAYG = N-acetyl-gamma-glutamyl; NmrA = nitrogen metabolite repressor A. The inversion in the T1PKS-4 cluster is shaded blue.

located on scaffold 62. The retrotransposons had, therefore, interrupted this cluster and could have caused displacement or inversion.

4. Discussion

Knoxdaviesia proteae and *K. capensis* are saprotrophic, yet highly niche-specific fungi (Roets et al., 2013). They have evolved in concert with the iconic *Protea* plants of the Cape Floristic Region biodiversity hotspot, together with arthropods, mites, microbes and most likely other organisms. The results of this study show a high level of structural genome conservation between these two closely related species. The similarity extended to a high proportion of orthologous proteins, small secretomes, few effectors and few proteins to mediate inter-organismal competition. However, proteins and other compounds secreted to the outside govern the interactions of fungi with their biotic environment (Girard et al., 2013) and our findings suggest some inter-specific divergence within them.

4.1. Knoxdaviesia genome similarity

Knoxdaviesia capensis and *K. proteae* showed large-scale genome synteny and aligned to each other with an average of 89.5 % nucleotide identity. The few identified telomeres and the lack of telomeres at both scaffold ends in *K. capensis* and *K. proteae* implies that neither genome assembly contains a complete chromosome. Although several inversions and rearrangements were detected, the lengthy aligned portions indicated macrosynteny between the two genomes (Hane et al., 2011). Both genomes also had similar sizes (35.54 vs. 35.49 Mb) and a comparable number of predicted genes (7940 vs. 8173; Aylward et al., 2016a). Such large scale genome conservation was expected, since macrosynteny is often observed between taxa that are phylogenetically closely related (De Vos et al., 2014; Ohm et al., 2012). The average nucleotide identity between them was less than that observed in comparisons of other closely related *Grosmannia* Goid. (98 %; Alamouti et al., 2014) and dermatophyte species (94.8 %; Burmester et al., 2011), but similar to the 85 - 91 % identity between three species of *Fusarium* Link (Ma et al., 2010).

The estimated 10 % repetitive proportion of the *Knoxdaviesia* genomes is congruent with an intermediate repeat content in fungi (Amselem et al., 2015; Castanera et al., 2016). Although the genomes of obligate biotrophic fungi are often rich (> 45 % of the genome) in transposable elements (Amselem et al., 2015), repeats in Sordariomycete genomes typically

range between 0.8 – 28 % (Castanera et al., 2016; Ma et al., 2010). The repeat content of *Knoxdaviesia* is therefore comparable to those of Sordariomycetes such as *Neurospora crassa* Shear & B.O. Dodge (10 %; Galagan et al., 2003), *Grosmannia clavigera* (Rob.-Jeffr. & R.W. Davidson) Zipfel, Z.W. de Beer & M.J. Wingf. (10.4 %; DiGuistini et al., 2011) and *Magnaporthe oryzae* B.C. Couch (11.1 %; Amselem et al., 2015). Like other fungi (Castanera et al., 2016), the *Knoxdaviesia* genomes contained a large number of Class I elements, specifically *Gypsy* and *Copia* LTR retrotransposons, that use an RNA intermediate to "copy and paste" themselves. However, our data suggest that RIP has curtailed extended proliferation of repeats and TEs. In both genomes, repetitive sequences occupied a smaller proportion than those that have been affected by RIP, which implies that certain repeats could have been mutated to the extent that they are no longer detected (Hane and Oliver, 2010).

The slightly greater proportion of repetitive elements in the *K. proteae* genome was not indicative of significant differences between *K. capensis* and *K. proteae*, since TE content may vary even between different strains of a fungal species (Castanera et al., 2016). A difference in the repeat identity of retrotransposons was, however, apparent and may correlate with differences in the selective forces shaping their genomes. Both *K. capensis* and *K. proteae* have massive gene and genotypic diversity, disperse over large geographic distances (Aylward et al., 2014, 2015; 2017a) and continually generate novel genetic combinations through an outcrossing reproductive strategy (Aylward et al., 2016b). These properties all point toward these fungi's high "evolutionary potential" or ability to adapt to changing environments (McDonald and Linde, 2002). Our results suggest that TEs have likely contributed to the genetic diversity of these fungi and could have played an important role in their divergence. This is because TEs are known to increase the plasticity of genomes through transposition to different locations and by providing sites for homologous recombination (Castanera et al., 2016).

The 18 multi-gene orthogroups shared by *Knoxdaviesia* were dominated by membrane transporter proteins (see Table S2). This suggests that similar substances are moved across the membranes of *K. capensis* and *K. proteae*, as would be expected in their specialist niche. The shared family of heat shock proteins is likely significant to the survival of these species, since temperatures in *Protea* infructescences may exceed 40°C (Roets et al., 2012).

Less than 15 % of the proteins in each *Knoxdaviesia* species were classified as species-specific. This proportion is similar to that reported between closely related Dothideomycetes (Ohm et al., 2012) and two *Ustilago* (Pers.) Roussel species in the Basidiomycetes (Laurie et al., 2012) and exceeds the proportion between conspecific Dothideomycetes (Ohm et al.,

2012). As expected, protein categories such as metabolism, protein degradation, and substrate-binding that potentially influence environmental interactions, comprised the majority of annotated species-specific proteins. Many similar species-specific annotations were noted for proteins in the two *Knoxdaviesia* species, but these were not identified as orthologs or recent paralogs, suggesting functional divergence (Li et al., 2003). Considering the host range disparity of *K. capensis* and *K. proteae*, it is possible that divergence in these proteins may be linked to the adaptation of each species to host-specific substrates. This claim, however, requires further investigation.

4.2. Small secretomes and few effectors

In comparison to plant pathogens and other saprotrophs, few secreted proteins were identified in *K. capensis* and *K. proteae* (Brown et al., 2012; Kim et al., 2016). Previously, Aylward et al. (2017b) characterized polysaccharide-degrading enzymes in *Knoxdaviesia* and noted the prevalence of plant cell wall-degrading enzymes, although the specialist niche of *Knoxdaviesia* likely facilitated a reduction in the number of these enzymes compared to saprotrophs. Even so, carbohydrate-degrading proteins were the largest category of secreted proteins identified in the current study, which is congruent with the idea that *Protea* cell walls provide the primary source of nutrition for these fungi. The secreted peptidases and lipases in both species may indicate that they have some capacity to degrade components beyond the plant cell wall, perhaps also including some protein and lipid remains of other inhabitants in the infructescence.

A small proportion of the *Knoxdaviesia* secretome was made up of proteins associated with plant pathogenesis, such as small secreted cysteine-rich proteins (SSCPs; Rep, 2005a), proteins with homology to pathogen effector proteins and members of the necrosis and ethylene inducing-like protein family (Nep1-like proteins; Gijzen and Nürnberger, 2006). The presence of pathogenesis-related proteins is not uncommon in saprotrophic species (Seidl et al., 2015), although they are usually distinguished from pathogens by their number. For example, 190 SSCPs were recently identified from the genome of the cereal head blight pathogen, *F. graminearum* Schwabe (Lu and Edwards, 2016), in comparison with the < 50 potential SSCP candidates of *Knoxdaviesia*. The *Knoxdaviesia* species, therefore, have a very limited arsenal of SSCPs that are potentially involved in establishing and maintaining a symbiotic interaction with their hosts, rather than mediating pathogenesis.

4.3. Compounds mediating competition

Several secreted defence enzymes were detected in *K. capensis* and *K. proteae*, potentially allowing them to compete with arthropods, nematodes and microbes in infructescences. The glucose oxidase enzymes are intriguing since they convert glucose to hydrogen peroxide (Bankar et al., 2009) and may enable these species to establish their dominance early during the colonization of young flower heads. Glucose is abundant in *Protea* nectar (Nicolson and Van Wyk, 1998) and the production of hydrogen peroxide may prevent establishment of other microbes during the flowering season. Once nectar is depleted and the infructescence develops, the amount of free glucose would depend on the degradation of cellulose components. Numerous organisms colonize the enclosed infructescence, suggesting that the *Knoxdaviesia* species require additional measures to compete successfully during this stage of niche capture.

Besides the secreted defence proteins, a small arsenal of potential secondary metabolite biosynthesis clusters was found in both *Knoxdaviesia* species. Whereas most Ascomycetes have upwards of 20 NRPS and PKS backbone genes (Kubicek et al., 2011), only eight of these genes (three NRPS and five PKS) were identified in *Knoxdaviesia*. Additionally, at least one of the PKS genes in *K. proteae* is mutated to the extent that it cannot be functional in this species. Members in the closely related Ceratocystidaceae family have similarly low numbers of PKS clusters, between three and six per species (Sayari et al., 2018), suggesting that the few secondary metabolite biosynthesis clusters in *Knoxdaviesia* could be a common trait in Microascalean fungi. These numbers are even lower than the 10 NRPS and PKS genes identified from *N. crassa*, the model for RIP activity (Galagan et al., 2003). Low repeat content and low gene duplication levels have been identified as the reason for the lack of secondary metabolite cluster diversity in *N. crassa* (Galagan et al., 2003). Similarly, the large sections of RIP-affected sequences in the *Knoxdaviesia* genomes could imply that RIP has played a role in keeping the evolution of secondary metabolite clusters in check.

The highly specialized niche occupied by *Protea*-associated *Knoxdaviesia* species may have made large scale diversification of secondary metabolites unnecessary. The abundance of ophiostomatoid fungi in infructescences was previously thought to be due to their fitness and ability to resist colonization by other contaminating fungi (Lee et al., 2005; Marais and Wingfield, 1994). However, three novel clades of antifungal-producing *Streptomyces* Waksman & Henrici bacteria have recently been isolated from *Protea* infructescences, to which *Knoxdaviesia* species have been shown to be at least partially resistant (Human, 2013; Human et al., 2016). Although *Streptomyces* species may lessen inter-specific fungal

competition to a certain extent, *Protea*-associated ophiostomatoid fungi would have to tolerate the antifungal compounds produced by these species. Our results suggest that *Knoxdaviesia* secrete various detoxification enzymes that may enable these species to tolerate some antifungal compounds produced by both the actinomycete bacteria and its plant host. Interestingly, a protein with a peptidoglycan-binding domain is present in both *Knoxdaviesia* species, although only predicted to be secreted in *K. capensis*, and may mediate an interaction with bacteria present in the niche. It is also possible that additional detoxifying processes are present in these *Knoxdaviesia* species, but could not be detected with the methods used in this study.

In both *Knoxdaviesia* species, two of the secondary metabolite clusters identified (NRPS-1 and 2) were predicted to be involved in synthesising siderophores, whereas eight different types of secreted proteins potentially detoxify harmful compounds (see Table 5). Siderophores sequestrate iron from the environment (Haas, 2014). As a scarce resource essential in many catalytic reactions, effective iron uptake plays an important role in determining whether a species will be a successful competitor (Loper and Buyer, 1991). The products synthesized by the remaining nine secondary metabolite clusters in *K. capensis* and eight in *K. proteae* are unknown, but may offer protection from predators and other competitors (Brakhage and Schroeckh, 2011). The predicted chalcone and stilbene synthase PKS type-III is also present in the closely related Ceratocystidaceae family (Sayari et al., 2018) and theoretically produces a flavonoid with a role in pigmentation or defence (Austin and Noel, 2003), while the putative geranylgeranyl pyrophosphate synthetase may produce an antimicrobial diterpene or carotenoid (Keller et al., 2005). These secondary metabolite clusters may, therefore, contribute toxic substances that enable *Knoxdaviesia* to compete in infructescences.

5. Conclusions

A high level of genetic similarity has been observed between *K. capensis* and *K. proteae* in previous studies (Aylward et al., 2016b; 2017b). The present comparative genomics study has affirmed the conservation between these species at the genome level and revealed subtle differences in their species-specific and secreted proteins and defence compounds. The macrosynteny between the *Knoxdaviesia* genomes, the similarities in their secreted proteins and their syntenic secondary metabolite cluster localization, reflect the phylogenetic and ecological relatedness of these species. Their different host associations were not clearly

defined by gene content, although more secreted and species-specific metabolic proteins were identified in *K. capensis*, which may enable it to colonize multiple *Protea* hosts having different chemistries (Valente et al., 2010). This is in contrast to the host-specific *K. proteae* from which less species-specific and secreted proteins were identified.

The greater number of secreted enzymes associated with cell defence and the maintenance of the secondary metabolite clusters in *K. capensis* likely reflects its wider host range. Due to the occurrence of this species in multiple hosts, it will likely also encounter greater numbers of competitors than the host-specific *K. proteae*. For example, two *Streptomyces* groups additional to those already known from *P. repens* L. have been isolated from a *K. capensis* host, *P. neriifolia* R. Br. (Human et al., 2016). Conversely, the fewer defence proteins and degraded T1PKS-4 secondary metabolite gene cluster in *K. proteae* implies that the selective pressures experienced within the *P. repens* host environment are not sufficient to maintain these genes.

This study aimed to determine how genome structure and gene content differentiate these *Knoxdaviesia* genomes. The next step will be to employ transcriptomics to answer questions concerning the expression of certain genes in these species. Additionally, a second lineage of *Sporothrix* Hektoen & C.F. Perkins ophiostomatoid fungi, although phylogenetically distant to *Knoxdaviesia* (De Beer et al., 2016; Wingfield et al., 1999), have adapted to the same *Protea* niche (Marais and Wingfield, 2001). It would, therefore, be interesting to investigate whether *Sporothrix* has followed the same path of adaptation as hypothesized for *Protea*-associated *Knoxdaviesia*. Specifically, the set of secondary metabolite biosynthesis clusters and predicted secretomes of these species would indicate whether the *Protea* environment promotes a smaller genetic complement. Further, some *Sporothrix* species have multiple *Protea* hosts (Marais and Wingfield, 2001; Roets et al., 2006a), while others are apparently host-specific (Roets et al., 2008; 2010), enabling further investigation into how host association has potentially resulted in speciation.

Acknowledgements

This work was supported by the National Research Foundation (NRF) and the Department of Science and Technology (DST)-NRF Centre of Excellence in Tree Health Biotechnology (CTHB) and the SARChI chair in Fungal Genomics.

References

Alamouti, S.M., Haridas, S., Feau, N., Robertson, G., Bohlmann, J., Breuil, C., 2014. Comparative genomics of the pine pathogens and beetle symbionts in the genus *Grosmannia*. Molecular Biology and Evolution 31, 1454-1474.

Alkan, C., Sajjadian, S., Eichler, E.E., 2011. Limitations of next-generation genome sequence assembly. Nature Methods 8, 61-65.

Amselem, J., Lebrun, M.-H., Quesneville, H., 2015. Whole genome comparative analysis of transposable elements provides new insight into mechanisms of their inactivation in fungal genomes. BMC Genomics 16, 141.

Austin, M.B., Noel, J.P., 2003. The chalcone synthase superfamily of type III polyketide synthases. Natural Product Reports 20, 79-110.

Aylward, J., Dreyer, L.L., Laas, T., Smit, L., Roets, F., 2017a. *Knoxdaviesia capensis*: dispersal ecology and population genetics of a flower-associated fungus. Fungal Ecology 26, 28-36.

Aylward, J., Dreyer, L.L., Steenkamp, E.T., Wingfield, M.J., Roets, F., 2014. Panmixia defines the genetic diversity of a unique arthropod-dispersed fungus specific to *Protea* flowers. Ecology and Evolution 4, 3444-3455.

Aylward, J., Dreyer, L.L., Steenkamp, E.T., Wingfield, M.J., Roets, F., 2015. Long-distance dispersal and recolonization of a fire-destroyed niche by a mite-associated fungus. Fungal Biology 119, 245-256.

Aylward, J., Steenkamp, E.T., Dreyer, L.L., Roets, F., Wingfield, B.D., Wingfield, M.J., 2016a. Genome sequences of *Knoxdaviesia capensis* and *K. proteae* (Fungi: Ascomycota) from *Protea* trees in South Africa. Standards in Genomic Sciences 11, 1-7.

Aylward, J., Steenkamp, E.T., Dreyer, L.L., Roets, F., Wingfield, M.J., Wingfield, B.D., 2016b. Genetic basis for high population diversity in *Protea*-associated *Knoxdaviesia*. Fungal Genetics and Biology 96, 47-57.

Aylward, J., Wingfield, B.D., Dreyer, L.L., Roets, F., Wingfield, M.J., Steenkamp, E.T., 2017b. Contrasting carbon metabolism in saprotrophic and pathogenic Microascalean fungi from *Protea* trees. Fungal Ecology 30, 88-100.

Bankar, S.B., Bule, M.V., Singhal, R.S., Ananthanarayan, L., 2009. Glucose oxidase — An overview. Biotechnology Advances 27, 489-501.

Blin, K., Medema, M.H., Kazempour, D., Fischbach, M.A., Breitling, R., Takano, E., Weber, T., 2013. antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Research, gkt449.

Brakhage, A.A., Schroeckh, V., 2011. Fungal secondary metabolites – Strategies to activate silent gene clusters. Fungal Genetics and Biology 48, 15-22.

Brown, N.A., Antoniw, J., Hammond-Kosack, K.E., 2012. The predicted secretome of the plant pathogenic fungus *Fusarium graminearum*: a refined comparative analysis. PLOS ONE 7, e33731.

Burmester, A., Shelest, E., Glöckner, G., Heddergott, C., Schindler, S., Staib, P., Heidel, A., Felder, M., Petzold, A., Szafranski, K., 2011. Comparative and functional genomics provide insights into the pathogenicity of dermatophytic fungi. Genome Biology 12, R7.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421.

Cambareri, E.B., Jensen, B.C., Schabtach, E., Selker, E.U., 1989. Repeat-induced GC to AT mutations in *Neurospora*. Science 244, 1571-1575.
Cantarel, B.L., Korf, I., Robb, S.M., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A.S., Yandell, M., 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research 18, 188-196.
Castanera, R., López-Varas, L., Borgognone, A., LaButti, K., Lapidus, A., Schmutz, J., Grimwood, J., Pérez, G., Pisabarro, A.G., Grigoriev, I.V., Stajich, J.E., Ramírez, L., 2016. Transposable Elements *versus* the Fungal Genome: Impact on Whole-Genome Architecture and Transcriptional Profiles. PLOS Genetics 12, e1006108.
Coetzee, J.H., Giliomee, J.H., 1985. Insects in association with the inflorescence of *Protea repens* (L.) (Proteaceae) and their role in pollination. Journal of the Entomological Society of South Africa 48, 303-314.
De Beer, Z.W., Duong, T.A., Wingfield, M.J., 2016. The divorce of *Sporothrix* and *Ophiostoma*: solution to a problematic relationship. Studies in Mycology 83, 165-191.
De Vos, L., Steenkamp, E.T., Martin, S.H., Santana, Q.C., Fourie, G., van der Merwe, N.A., Wingfield, M.J., Wingfield, B.D., 2014. Genome-wide macrosynteny among *Fusarium* species in the *Gibberella fujikuroi* complex revealed by amplified fragment length polymorphisms. PLOS ONE 9, e114682.
DiGuistini, S., Wang, Y., Liao, N.Y., Taylor, G., Tanguay, P., Feau, N., Henrissat, B., Chan, S.K., Hesse-Orce, U., Alamouti, S.M., Tsui, C.K.M., Docking, R.T., Levasseur, A., Haridas, S., Robertson, G., Birol, I., Holt, R.A., Marra, M.A., Hamelin, R.C., Hirst, M., Jones, S.J.M., Bohlmann, J., Breuil, C., 2011. Genome and transcriptome analyses of the mountain pine beetle-fungal symbiont *Grosmannia clavigera*, a lodgepole pine pathogen. Proceedings of the National Academy of Sciences 108, 2504-2509.
Eddy, S.R., 2011. Accelerated profile HMM searches. PLOS Computational Biology 7, e1002195.
Emanuelsson, O., Nielsen, H., Brunak, S., Von Heijne, G., 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. Journal of Molecular Biology 300, 1005-1016.
Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., Bateman, A., 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research 44, D279-D285.
Fulnečková, J., Ševčíková, T., Fajkus, J., Lukešová, A., Lukeš, M., Vlček, Č., Lang, B.F., Kim, E., Eliáš, M., Sýkorová, E., 2013. A broad phylogenetic survey unveils the diversity and evolution of telomeres in eukaryotes. Genome Biology and Evolution 5, 468-483.
Galagan, J.E., Calvo, S.E., Borkovich, K.A., Selker, E.U., Read, N.D., Jaffe, D., FitzHugh, W., Ma, L.-J., Smirnov, S., Purcell, S., 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. Nature 422, 859-868.
Gijzen, M., Nürnberger, T., 2006. Nep1-like proteins from plant pathogens: recruitment and diversification of the NPP1 domain across taxa. Phytochemistry 67, 1800-1807.
Gilgado, F., Cano, J., Gené, J., Guarro, J., 2005. Molecular Phylogeny of the *Pseudallescheria boydii* Species Complex: Proposal of Two New Species. Journal of Clinical Microbiology 43, 4930-4942.
Girard, V., Dieryckx, C., Job, C., Job, D., 2013. Secretomes: the fungal strike force. Proteomics 13, 597-608.

Götz, S., García-Gómez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talón, M., Dopazo, J., Conesa, A., 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Research 36, 3420-3435.
Gray, Y.H.M., 2000. It takes two transposons to tango:transposable-element-mediated chromosomal rearrangements. Trends in Genetics 16, 461-468.
Haas, H., 2014. Fungal siderophore metabolism with a focus on *Aspergillus fumigatus*. Natural Product Reports 31, 1266-1276.
Hane, J.K., Oliver, R.P., 2010. *In silico* reversal of repeat-induced point mutation (RIP) identifies the origins of repeat families and uncovers obscured duplicated genes. BMC Genomics 11, 655.
Hane, J.K., Rouxel, T., Howlett, B.J., Kema, G.H., Goodwin, S.B., Oliver, R.P., 2011. A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. Genome Biology 12, 1.
Hemsworth, G.R., Johnston, E.M., Davies, G.J., Walton, P.H., 2015. Lytic polysaccharide monooxygenases in biomass conversion. Trends in Biotechnology 33, 747-761.
Holt, C., Yandell, M., 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12, 491.
Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C., Nakai, K., 2007. WoLF PSORT: protein localization predictor. Nucleic Acids Research 35, W585-W587.
Human, Z., Moon, K., Bae, M., de Beer, Z.W., Cha, S., Wingfield, M.J., Slippers, B., Oh, D.-C., Venter, S.N., 2016. Antifungal *Streptomyces* spp. associated with the infructescences of *Protea* spp. in South Africa. Frontiers in Microbiology 7, 1657.
Human, Z.R., 2013. The diversity and ecology of actinomycetes associated with environments dominated by ophiostomatoid fungi. MSc Thesis, University of Pretoria, Pretoria, South Africa.
Keller, N.P., Turner, G., Bennett, J.W., 2005. Fungal secondary metabolism—from biochemistry to genomics. Nature Reviews Microbiology 3, 937-947.
Kim, K.-T., Jeon, J., Choi, J., Cheong, K., Song, H., Choi, G., Kang, S., Lee, Y.-H., 2016. Kingdom-wide analysis of fungal small secreted proteins (SSPs) reveals their potential role in host association. Frontiers in Plant Science 7.
Kohler, A., Kuo, A., Nagy, L.G., Morin, E., Barry, K.W., Buscot, F., Canbäck, B., Choi, C., Cichocki, N., Clum, A., 2015. Convergent losses of decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal mutualists. Nature Genetics 47, 410-415.
Krogh, A., Larsson, B., Von Heijne, G., Sonnhammer, E.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of Molecular Biology 305, 567-580.
Kubicek, C.P., Herrera-Estrella, A., Seidl-Seiboth, V., Martinez, D.A., Druzhinina, I.S., Thon, M., Zeilinger, S., Casas-Flores, S., Horwitz, B.A., Mukherjee, P.K., Mukherjee, M., Kredics, L., Alcaraz, L.D., Aerts, A., Antal, Z., Atanasova, L., Cervantes-Badillo, M.G., Challacombe, J., Chertkov, O., McCluskey, K., Coulpier, F., Deshpande, N., von Döhren, H., Ebbole, D.J., Esquivel-Naranjo, E.U., Fekete, E., Flipphi, M., Glaser, F., Gómez-Rodríguez, E.Y., Gruber, S., Han, C., Henrissat, B., Hermosa, R., Hernández-Oñate, M., Karaffa, L., Kosti, I., Le Crom, S., Lindquist, E., Lucas, S., Lübeck, M., Lübeck, P.S., Margeot, A., Metz, B., Misra, M., Nevalainen, H., Omann, M., Packer, N., Perrone, G., Uresti-Rivera, E.E., Salamov, A., Schmoll, M., Seiboth, B., Shapiro, H., Sukno, S., Tamayo-Ramos, J.A., Tisch, D., Wiest, A., Wilkinson, H.H., Zhang, M., Coutinho, P.M., Kenerley, C.M., Monte, E., Baker, S.E.,

Grigoriev, I.V., 2011. Comparative genome sequence analysis underscores mycoparasitism as the ancestral life style of *Trichoderma*. Genome Biology 12, R40-R40.
Kubicek, C.P., Starr, T.L., Glass, N.L., 2014. Plant cell wall-degrading enzymes and their secretion in plant-pathogenic fungi. Annual Review of Phytopathology 52, 427-451.
Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S.L., 2004. Versatile and open software for comparing large genomes. Genome Biology 5, R12.
Kusumoto, K.I., Suzuki, S., Kashiwagi, Y., 2003. Telomeric repeat sequence of *Aspergillus oryzae* consists of dodeca-nucleotides. Appl Microbiol Biotechnol 61, 247-251.
Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9, 357-359.
Laurie, J.D., Ali, S., Linning, R., Mannhaupt, G., Wong, P., Güldener, U., Münsterkötter, M., Moore, R., Kahmann, R., Bakkeren, G., 2012. Genome comparison of barley and maize smut fungi reveals targeted loss of RNA silencing components and species-specific presence of transposable elements. The Plant Cell 24, 1733-1745.
Lee, S., Groenewald, J.Z., Taylor, J.E., Roets, F., Crous, P.W., 2003. Rhynchostomatoid fungi occurring on Proteaceae. Mycologia 95, 902-910.
Lee, S., Roets, F., Crous, P.W., 2005. Biodiversity of saprobic microfungi associated with the infructescences of *Protea* species in South Africa. Fungal Diversity 19, 69-78.
Lewis, Z.A., Honda, S., Khlafallah, T.K., Jeffress, J.K., Freitag, M., Mohn, F., Schübeler, D., Selker, E.U., 2009. Relics of repeat-induced point mutation direct heterochromatin formation in *Neurospora crassa*. Genome Research 19, 427-437.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078-2079.
Li, L., Stoeckert, C.J., Roos, D.S., 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Research 13, 2178-2189.
Liti, G., Louis, E.J., 2005. Yeast evolution and comparative genomics. Annual Review of Microbiology 59, 135-153.
Loper, J.E., Buyer, J.S., 1991. Siderophores in microbial interactions on plant surfaces. Molecular Plant-Microbe Interactions 4, 5-13.
Lu, S., Edwards, M.C., 2016. Genome-Wide Analysis of Small Secreted Cysteine-Rich Proteins Identifies Candidate Effector Proteins Potentially Involved in *Fusarium graminearum*− Wheat Interactions. Phytopathology 106, 166-176.
Ma, L.-J., Van Der Does, H.C., Borkovich, K.A., Coleman, J.J., Daboussi, M.-J., Di Pietro, A., Dufresne, M., Freitag, M., Grabherr, M., Henrissat, B., 2010. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature 464, 367-373.
Marais, G.J., Wingfield, M.J., 1994. Fungi associated with infructescences of *Protea* species in South Africa, including a new species of *Ophiostoma*. Mycological Research 98, 369-374.
Marais, G.J., Wingfield, M.J., 2001. *Ophiostoma africanum* sp. *nov*., and a key to ophiostomatoid species from *Protea* infructescences. Mycological Research 105, 240-246.
Margolin, B.S., Garrett-Engele, P.W., Stevens, J.N., Fritz, D.Y., Garrett-Engele, C., Metzenberg, R.L., Selker, E.U., 1998. A methylated *Neurospora* 5S rRNA pseudogene contains a transposable element inactivated by repeat-induced point mutation. Genetics 149, 1787-1797.
Martin, F., Aerts, A., Ahrén, D., Brun, A., Danchin, E., Duchaussoy, F., Gibon, J., Kohler, A., Lindquist, E., Pereda, V., 2008. The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. Nature 452, 88.

McDonald, B.A., Linde, C., 2002. Pathogen population genetics, evolutionary potential, and durable resistance. Annual Review of Phytopathology 40, 349-379.
Medema, M.H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M.A., Weber, T., Takano, E., Breitling, R., 2011. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Research 39, W339-W346.
Nicolson, S.W., Van Wyk, B.-E., 1998. Nectar Sugars in Proteaceae: Patterns and Processes. Australian Journal of Botany 46, 489-504.
Ohm, R.A., Feau, N., Henrissat, B., Schoch, C.L., Horwitz, B.A., Barry, K.W., Condon, B.J., Copeland, A.C., Dhillon, B., Glaser, F., 2012. Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. PLOS Pathogens 8, e1003037.
Petersen, T.N., Brunak, S., von Heijne, G., Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods 8, 785-786.
Pierleoni, A., Martelli, P.L., Casadio, R., 2008. PredGPI: a GPI-anchor predictor. BMC Bioinformatics 9, 1.
Podlevsky, J.D., Bley, C.J., Omana, R.V., Qi, X., Chen, J.J., 2008. The telomerase database. Nucleic Acids Research 36, D339-343.
Price, A.L., Jones, N.C., Pevzner, P.A., 2005. *De novo* identification of repeat families in large genomes. Bioinformatics 21, i351-i358.
R Core Team, 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.
Rep, M., 2005a. Small proteins of plant-pathogenic fungi secreted during host colonization. FEMS Microbiology Letters 253, 19-27.
Rep, M., 2005b. Small proteins of plant-pathogenic fungi secreted during host colonization. FEMS Microbiology Letters 253, 19-27.
Rep, M., Van Der Does, H.C., Meijer, M., Van Wijk, R., Houterman, P.M., Dekker, H.L., De Koster, C.G., Cornelissen, B.J., 2004. A small, cysteine-rich protein secreted by *Fusarium oxysporum* during colonization of xylem vessels is required for I-3-mediated resistance in tomato. Molecular Microbiology 53, 1373-1383.
Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16, 276-277.
Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., Mesirov, J.P., 2011. Integrative genomics viewer. Nature Biotechnology 29, 24-26.
Rocap, G., Larimer, F.W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N.A., Arellano, A., Coleman, M., Hauser, L., Hess, W.R., 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424, 1042.
Roets, F., de Beer, Z.W., Dreyer, L.L., Zipfel, R., Crous, P.W., Wingfield, M.J., 2006a. Multi-gene phylogeny for *Ophiostoma* spp. reveals two new species from *Protea* infructescences. Studies in Mycology 55, 199-212.
Roets, F., de Beer, Z.W., Wingfield, M.J., Crous, P.W., Dreyer, L.L., 2008. *Ophiostoma gemellus* and *Sporothrix variecibatus* from Mites Infesting *Protea* Infructescences in South Africa. Mycologia 100, 496-510.

Roets, F., Dreyer, L.L., Crous, P.W., 2005. Seasonal trends in colonisation of *Protea* infructescences by *Gondwanamyces* and *Ophiostoma* spp. South African Journal of Botany 71, 307-311.
Roets, F., Dreyer, L.L., Geertsema, H., Crous, P.W., 2006b. Arthropod communities in Proteaceae infructescences: seasonal variation and the influence of infructescence phenology. African Entomology 14, 257-265.
Roets, F., Theron, N., Wingfield, M.J., Dreyer, L.L., 2011. Biotic and abiotic constraints that facilitate host exclusivity of *Gondwanamyces* and *Ophiostoma* on *Protea*. Fungal Biology 116, 49-61.
Roets, F., Theron, N., Wingfield, M.J., Dreyer, L.L., 2012. Biotic and abiotic constraints that facilitate host exclusivity of *Gondwanamyces* and *Ophiostoma* on *Protea*. Fungal Biology 116, 49-61.
Roets, F., Wingfield, B.D., de Beer, Z.W., Wingfield, M., Dreyer, L.L., 2010. Two new *Ophiostoma* species from *Protea caffra* in Zambia. Persoonia 24, 18-28.
Roets, F., Wingfield, M.J., Crous, P.W., Dreyer, L.L., 2009. Fungal radiation in the Cape Floristic Region: An analysis based on *Gondwanamyces* and *Ophiostoma*. Molecular Phylogenetics and Evolution 51, 111-119.
Roets, F., Wingfield, M.J., Crous, P.W., Dreyer, L.L., 2013. Taxonomy and ecology of ophiostomatoid fungi associated with *Protea* infructescences, in: Seifert, K.A., de Beer, Z.W., Wingfield, M.J. (Eds.), Ophiostomatoid fungi: expanding frontiers. CBS Biodiversity Series, Utrecht, The Netherlands, pp. 177-187.
Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Güldener, U., Mannhaupt, G., Münsterkötter, M., 2004. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Research 32, 5539-5545.
Ruess, L., Lussenhop, J., 2005. Trophic interactions of fungi and animals, in: Dighton, J., White, J.K., Oudemans, P. (Eds.), The fungal community—its organization and role in the ecosystem. CRC Press, Boca Raton, FL, pp. 581–598.
Sayari, M., Steenkamp, E.T., van der Nest, M.A., Wingfield, B.D., 2018. Diversity and evolution of polyketide biosynthesis gene clusters in the *Ceratocystidaceae*. Fungal Biology.
Schneider, D.J., Collmer, A., 2010. Studying plant-pathogen interactions in the genomics era: beyond molecular Koch's postulates to systems biology. Annual Review of Phytopathology 48, 457-479.
Seidl, M.F., Faino, L., Shi-Kunne, X., van den Berg, G.C., Bolton, M.D., Thomma, B.P., 2015. The genome of the saprophytic fungus *Verticillium tricorpu*s reveals a complex effector repertoire resembling that of its pathogenic relatives. Molecular Plant-Microbe Interactions 28, 362-373.
Seifert, K.A., De Beer, Z.W., Wingfield, M.J., 2013. The Ophiostomatoid Fungi: Expanding Frontiers. CBS Biodiversity Series, Utrecht, The Netherlands.
Selker, E.U., Tountas, N.A., Cross, S.H., Margolin, B.S., Murphy, J.G., Bird, A.P., Freitag, M., 2003. The methylated component of the *Neurospora crassa genome*. Nature 422, 893-897.
Slotkin, R.K., Martienssen, R., 2007. Transposable elements and the epigenetic regulation of the genome. Nature Reviews Genetics 8, 272-285.
Smit, A.F.A., Hubley, R., Green, P., 2013-2015. RepeatMasker Open-4.0. http://www.repeatmasker.org.

Templeton, M.D., Rikkerink, E.H., Beever, R.E., 1994. Small, cysteine-rich proteins and recognition in fungal-plant interactions. Molecular Plant-Microbe Interactions 7, 320-325.
Theron, N., Roets, F., Dreyer, L.L., Esler, K.J., Ueckermann, E.A., 2012. A new genus and eight new species of Tydeoidea (Acari: Trombidiformes) from *Protea* species in South Africa. International Journal of Acarology 38, 257-273.
Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 14, 178-192.
Traut, W., Szczepanowski, M., Vítková, M., Opitz, C., Marec, F., Zrzavý, J., 2007. The telomere repeat motif of basal Metazoa. Chromosome Research 15, 371-382.
Valente, L.M., Reeves, G., Schnitzler, J., Mason, I.P., Fay, M.F., Rebelo, T.G., Chase, M.W., Barraclough, T.G., 2010. Diversification of the African genus *Protea* (Proteaceae) in the Cape biodiversity hotspot and beyond: equal rates in different biomes. Evolution 64, 745-760.
Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H.U., Bruccoleri, R., Lee, S.Y., Fischbach, M.A., Müller, R., Wohlleben, W., 2015. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Research, gkv437.
Wingfield, B.D., Viljoen, C.D., Wingfield, M.J., 1999. Phylogenetic relationships of ophiostomatoid fungi associated with *Protea* infructescences in South Africa. Mycological Research 103, 1616-1620.
Wingfield, M.J., Seifert, K.A., Webber, J.F., 1993. *Ceratocystis* and *Ophiostoma*: taxonomy, ecology, and pathogenicity. American Phytopathological Society, St. Paul, Minnesota.
Wingfield, M.J., Van Wyk, P.S., 1993. A new species of *Ophiostoma* from *Protea* infructescences in South Africa. Mycological Research 97, 709-716.
Wöstemeyer, J., Kreibich, A., 2002. Repetitive DNA elements in fungi (Mycota): impact on genomic architecture and evolution. Current Genetics 41, 189-198.
Zwölfer, H., 1979. Strategies and counter strategies in insect population systems competing for space and food in flower heads and plant galls. Fortschritte der Zoologie 25, 331-353.

Supplementary information

The following supplementary tables, files and data are supplied with this submission.

Supplementary Table S1: Occurrence of the TTAGGGTTAC / GTAACCCTAA *Knoxdaviesia* telomere repeat in *K. capensis* and *K. proteae*.

Supplementary Table S2: Populous orthogroups in the *Knoxdaviesia* genomes.

Supplementary Table S3: Outcome of the protein BLAST for the *Knoxdaviesia capensis* and *K. proteae* species-specific proteins.

Supplementary Table S4: Cysteine-rich secreted proteins and proteins with hits to the Pathogen Host Interaction (PHI) database in *Knoxdaviesia capensis* and *K. proteae*.

Supplementary Table S5: Amplification of the T1PKS-4 cluster deletion in *Knoxdaviesia capensis* and *K. proteae*.

Supplementary File 1: Summary of repeat-induced-point mutation (RIP) per scaffold.

Supplementary File 2: Examples of the sequence depth and number of mismatches across repetitive regions in *Knoxdaviesia capensis* and *K. proteae*.

Supplementary File 3: Overview of the classification of *Knoxdaviesia* genome-wide species-specific proteins in Functional Catalogue categories.

Supplementary File 4: Classification and annotation of the genome-wide species-specific proteins of *Knoxdaviesia capensis* and *K. proteae*.

Supplementary File 5: *Knoxdaviesia* species-specific proteins putatively involved in secondary metabolism.

Supplementary File 6: Summary of proteins excluded from and included in the final secretome dataset.

Supplementary File 7: Overview of the classification of *Knoxdaviesia* secreted proteins in Functional Catalogue categories.

Supplementary File 8: Classification and annotation of the putative secreted proteins of *Knoxdaviesia capensis* and *K. proteae*.

S**upplementary File 9:** Classification and annotation of the small secreted cysteine-rich proteins (SSCPs) identified in the two *Knoxdaviesia* genomes.

Supplementary Data:

The predicted proteins of *Knoxdaviesia capensis* and *K. proteae* in FASTA format and the gff3 annotation files of the transposable elements identified by the REPET package have been made available on Mendeley Data (https://data.mendeley.com/), DOI:10.17632/rbx32w7crp.1 (this doi has been reserved, but not yet published).