MICROBIAL GENOMICS

RESEARCH ARTICLE

Jones *et al.*, *Microbial Genomics* 2021;7:000637

DOI 10.1099/mgen.0.000637

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Remote homology clustering identifies lowly conserved families of effector proteins in plant-pathogenic fungi

Darcy A. B. Jones[1], Paula M. Moolhuijzen[1] and James K. Hane[1,2,*]

Abstract

Plant diseases caused by fungal pathogens are typically initiated by molecular interactions between 'effector' molecules released by a pathogen and receptor molecules on or within the plant host cell. In many cases these effector-receptor interactions directly determine host resistance or susceptibility. The search for fungal effector proteins is a developing area in fungal-plant pathology, with more than 165 distinct confirmed fungal effector proteins in the public domain. For a small number of these, novel effectors can be rapidly discovered across multiple fungal species through the identification of known effector homologues. However, many have no detectable homology by standard sequence-based search methods. This study employs a novel comparison method (RemEff) that is capable of identifying protein families with greater sensitivity than traditional homology-inference methods, leveraging a growing pool of confirmed fungal effector data to enable the prediction of novel fungal effector candidates by protein family association. Resources relating to the RemEff method and data used in this study are available from https://figshare.com/projects/Effector_protein_remote_homology/87965.

DATA SUMMARY

Supplementary Material: https://doi.org/10.6084/m9.figshare.13285733.v1. Additional data: https://figshare.com/projects/Effector_protein_remote_homology/87965

INTRODUCTION

Fungal-plant pathogens expose or secrete molecules called 'effectors' into the extracellular environment, which may interact with or be internalized by their host, to promote infection. Hosts in turn may recognize pathogen-associated molecular patterns (PAMPs) and initiate defence responses, which confers innate immunity against the majority of pathogens, termed pattern triggered immunity (PTI) [1]. However, some pathogens employ a specialized infection strategy on a narrow range of hosts by secreting proteinaceous or secondary metabolite effectors, which either bypass host defences or cause host-cell death. Necrotic effector activity has been observed to rely on the presence of a cognate sensitivity (S) gene in the host genome, whereupon recognition of the NE by the S-protein will initiate host-cell death and promote necrotrophic infection [2]. A second layer of defence, termed effector-triggered immunity (ETI) may also be employed by the host, where the presence of a cognate resistance (R) gene will confer the ability to activate host defences in the presence of a recognized effector. Effectors are sometimes divided into subclasses based on their known interactions with host S and R genes, with necrotrophic effectors (NEs) interacting with S genes but having no known R genes, and avirulence effectors (AVRs) interacting with known R genes [2, 3]. Crop disease resistance breeding is usually conducted on the basis of introducing beneficial R genes and removing deleterious S genes. The study and discovery of fungal effectors among the growing pool of fungal genome data is vital for ongoing resistance breeding efforts [4], however there are a number of challenges that need to be overcome.

Proteinaceous fungal effectors have long been considered to lack sequence conservation, and in many cases have been

Received 25 November 2020; Accepted 22 June 2021; Published 01 September 2021

Author affiliations: [1]Centre for Crop & Disease Management, School of Molecular & Life Sciences, Curtin University, Perth, Australia; [2]Curtin Institute for Computation, Curtin University, Perth, Australia.

***Correspondence:** James K. Hane, james.hane@curtin.edu.au

Keywords: effector; fungi; pathogen; protein family; protein function.

Abbreviations: AVR, avirulence effector; CRN, Crinkler; CSEP, candidate secreted effector protein; ETI, effector-triggered immunity; HMM, Hidden Markov Model; MAX, Magnaporthe Avrs and ToxB-like; MSA, multiple sequence alignment; NBS-LRR, nucleotide binding site - leucine-rich repeat; NE, necrotrophic effector; NEP, necrosis and ethylene induced protein; NMR, nuclear magnetic resonance; PAMP, pathogen-associated molecular pattern; PSSM, position-specific scoring matrix; PTI, pattern-triggered immunity; RALPH, RNase-Like Protein associated with Haustoria; R gene, resistance gene; RIP, repeat-induced point mutation; S gene, sensitivity gene; SIX, secreted in xylem; SP, signal peptide; SRL, sarcin-ricin loop.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Supplementary material is available with the online version of this article.

presumed to have arisen independently. The collective term 'effector' is most frequently used to describe a highly diverse group of proteins with a common but broadly defined role in virulence on a narrow range of hosts, but is sometimes also used to describe highly conserved families of pathogenicity proteins with broad host specificity, such as the NEPs [5], cerato-platanins [6], and ribotoxins [7]. Little sequence similarity has been observed between known host-specific fungal effectors, potentially due to relatively high levels of genome plasticity in some fungi [8–13]. This is in direct contrast to effectors of a separate microbial lineage – the oomycetes – which are comparatively less plastic due to a lack of repeat-induced point mutation (RIP), fewer rearrangements to chromosome structure between genera [14, 15]. Unlike fungal effectors, oomycete effectors frequently have conserved motifs including RxLR-dEER [16] and LXLFLAK (Crinklers/CRN) [17]. Traditional biochemical and structural analyses are the gold standard for the functional characterization of effector candidates [18] but are unsuitable for high-throughput analyses. Moreover, existing high-throughput experimental methods, such as proteomics and genome-wide association studies, routinely return numerous genes or proteins that may be associated with the phenotype of interest, necessitating some additional information to prioritize future experimental validation.

High-throughput bioinformatic identification of fungal effector candidates remains a significant challenge due to the apparent lack of homology among most fungal effectors [19]. The vast majority of fungal proteins have no experimentally determined function and the accurate annotation of fungal genes is impeded by the narrow taxonomic range of fungal species with high-quality gene annotation and by the relatively high gene density observed in fungal genomes [20]. Nevertheless, a small but growing number of fungal effector families have been described with members in taxonomically distinct pathogens including: ToxA-like [21, 22] MAX [23], RALPH [24] and RXLR-like [25]. In line with elevated fungal genome plasticity, these effector 'families' share conserved structures but lack significant primary sequence similarity. This raises the possibility that at least some effectors – rather than arising independently or via lateral transfer – may have been vertically inherited from ancestor effector genes that were subsequently heavily mutated by fungal-specific genome mutagenesis mechanisms such as RIP [26]. Among the currently identified effector families, conserved structural folds with similar functions can be observed, which are typically missed by simple sequence alignments. Effector family relationships with high sequence divergence are difficult to predict with traditional methods (e.g. BLASTP), but more sophisticated structural prediction and comparison methods (e.g. protein threading and structural alignment) are not yet computationally feasible to include in a high-throughput analysis of a whole fungal proteome. Suitable alternatives come in the form of search methods that incorporate protein redundancy, such as profile-hidden Markov models (HMMs) or position site-specific models (PSSMs), which offer viable methods for finding remote homologues of confirmed effector proteins. Also of note are the cysteine-spacing classification systems that have been successfully applied to non-fungal cytotoxic venoms, which appear to have similar basic protein properties to fungal effectors [27, 28]. As our understanding of fungal effector biology improves, it may also become possible to apply similarly simple pattern-based heuristics for fungal effector classification.

Impact Statement

Effector proteins of fungal-plant pathogens may be the key to understanding and developing new methods of controlling important agricultural crop diseases. Progress towards identification of new effector proteins has been slowed by challenges related to mutations in fungal genomes. It is typically ineffective to search for new effectors by looking for similar sequences to those of known effectors. This study describes a new, more sensitive method of searching for remote effector homologs (RemEff), broadly applying pattern-based searches and hierarchical networks of protein similarity relationships across multiple fungal species. This allowed prediction of many new effector protein candidates, which are relevant across multiple pathosystems. This study also highlights previously observed and newly predicted effector family groups among known and predicted effectors, which are predominantly unrelated by sequence. The RemEff dataset is publicly available and can benefit ongoing research across the molecular plant pathology community, through fast-tracking effector discovery efforts.

Fungal effector protein families

Fungal pathogenicity effector proteins can be divided into those which (1) form family groupings using simple bioinformatics methods, i.e. conserved motifs/patterns identified via sequence-based alignment, and (2) those which cannot be grouped by the above methods. In the case of the latter, there have been several studies to date piecing together a growing set of small cysteine-rich, secreted, low molecular weight, protein families with at least some members having effector-like phenotypes. There is remarkable diversity across these families, both between families and within them, yet common themes are emerging. Structural homology and in some cases similar modes of action [21] are observed between proteins with very low sequence identity, and some conserved or functional motifs appear to comprise surface-exposed, positively charged residues. A selected set of these emerging protein families are introduced below.

ToxA

The ToxA-like family is named after the ToxA effector originally characterized in the wheat pathogen *Pyrenophora tritici-repentis* [29, 30], and for which putatively horizontally transferred loci were later identified by varying degrees of sequence similarity of the locus and a~14 Kbp flanking region [31–34], to genomes of other cereal-pathogenic fungi *Parastagonospora nodorum* [31], *Bipolaris maydis* [21] and *B. sorokiniana* [32, 35]. The full PtrToxA pre-pro-protein is 178 aa in length, with a signal peptide (SP) domain at position 22–23, and an N-terminal pro-peptide with a conserved 'LXXR' motif [21], which is cleaved during secretion at positions 60–61, producing the mature ToxA protein that corresponds to position 61 to 178 [29, 36, 37]. PtrToxA (and the identical PnToxA) interact with a NBS-LRR domain membrane protein Tsn1, which confers host sensitivity to ToxA [38]. Within wheat host cells, mature PtrToxA is reported to bind two chloroplast-localized proteins: ToxA binding protein 1 (ToxABP1, syn. *Triticum aestivum* thylakoid formation protein TaThf1), plastocyanin protein TaPCN [39] and TaPR-1–5 PR-1–5 [40]. ToxA-mediated disruption of chloroplast function leads to host cell necrosis, which requires light [41] and conservation of a structural loop possessing an 'RGD' motif [42].

The ToxA homologue of *B. maydis* (syn. *Cochliobolus heterostrophus*), ChToxA, has poor sequence similarity (64%) with Ptr/PnToxA, but has highly conserved structural similarity [21] and a similar light-dependent necrosis phenotype on maize. Despite the similar structure, the 'RGD' motif required by Ptr/PnToxA for necrosis of wheat is substituted with a 'SGN' motif [21]. Broadened similarity searches using hidden-Markov model (HMM)-based methods have predicted many other ToxA-like proteins across the classes Dothideomycetes and Sordariomycetes [21], including Avr2 of *Fusarium oxysporum*. Like ChToxA, Avr2 has a virulence-promoting phenotype, poor sequence identity with Ptr/PnToxA (~5%), and high structural similarity [43]. There are however, a few motifs that are conserved across the currently predicted members of the ToxA-like effector family, including the 'LXXR' motif within the pro-domain, three motifs located in beta sheets 4, 5 and 8 (LXVXIXN, LILTXY, WXXQ respectively), and an asparagine-rich WXXN(S)NXIXVXI motif [21].

MAX

The Magnaporthe Avrs and ToxB-like (MAX) effector family comprises another set of fungal proteins that are structurally conserved but divergent at the sequence level. The MAX family was originally derived from effectors of *Magnaporthe oryzae* [23]. Similarity of NMR structures containing two anti-parallel three-stranded beta sheets with a single disulfide bond has been demonstrated between *M. oryzae* AVR-Pia, AVR1-CO39, AvrPiz-t and *Pyrenophora tritici-repentis* ToxB [23, 44]. Sequence alignment, position-specific score matrix (PSSM) and profile-HMM searches against these structural homologues had subsequently revealed numerous homologues in other species, including *P. bromi* [45], *Bipolaris oryzae*, *Colletotrichum* spp., *Zymoseptoria tritici*, *Leptosphaeria maculans* and even low but significant similarity a protein in plant-associated bacteria *Pseudomonas* sp. *StFLB209* [23]. Multiple paralogues of members of this family have also been reported for some species, including *Pyrenophora* spp. [34], *C. fioriniae*, *C. orbiculare* and *C. gloeosporioides* [23], suggesting the potential for duplication and diversification of the relatively broadly-conserved MAX effector family.

AvrLm6

AvrLm6 is a well characterized AVR effector of the brassica pathogen *Leptosphaeria maculans*, which causes necrosis but has an avirulent phenotype in *Brassica napus* and *B. juncea* hosts [46] possessing the resistance (R) locus *Rlm6* [47]. Several AvrLm6-like proteins have been reported in other fungal pathogen species, including: *Colletotrichum* spp., *Fusarium oxysporum*, *L. biglobosa* and *Venturia* spp. [48, 49]. Notably in *V. inaequalis* and *V. pirina*, this family has undergone extensive clonal expansion [49].

Ribotoxins and RALPHs

Fungi secrete a broad variety of toxic and non-toxic RNases into the extracellular space and host [50]. One set of cytotoxic RNAses, the ribotoxins, are a group of fungal proteins that target the sarcin-ricin loop (SRL) of the host ribosome. This cleaves a single phosphodiester bond of the ribosomal RNA, rendering it catalytically inactive and ultimately causing cell death [7, 51]. Fungal secreted RNases tend to share a common *α*-helix *β*-sheet fold topology, but differ in their terminal and loop domains [7, 50]. Ribotoxins possess an extended positively charged loop that is absent in non-cytotoxic secreted RNases, which is presumed to be important for interacting with the host-SRL [50, 51]. Ribotoxins are well documented in entomopathogens of the Ascomycetes (e.g. *Aspergillus giganteus* α-sarcin and *Aspergillus restrictus* restrictocin [52] and are also found in Basidiomycetes (e.g. white-rot *Agrocybe aegerita* [53, 54]). Aspergillus ribotoxins are relatively well conserved [52], however members from other genera (e.g. *Hirsutella thompsonii* - Hirsutellin and *Agrocybe aegerita* - Ageritin) share low sequence identity with the Aspergillus varieties but retain similar structures and activities [54, 55]. Recently a cytotoxic secreted RNase protein, Zt6, was reported in the wheat pathogen *Zymoseptoria tritici* [56]. Although SRL binding has not yet been demonstrated for Zt6, it structurally resembles canonical ribotoxins and has RNase catalytic activity and exhibits toxicity to plants, some fungi and bacteria, but not to *Z. tritici* [56].

Another group of non-toxic RNases have been reported in the *Blumeria* genus of biotrophic plant pathogens. *Blumeria* possesses several large families of effector candidates, with one of the largest groups containing RNase-associated domains with predicted structural similarity to RNase proteins – the RNase-Like Proteins associated with Haustoria (RALPHs) [57]. RALPH effectors include *AvrPm2* (*BgtE-5845*) in *B. graminis* f. sp. *tritici* [24, 58] and AVRa13 (aka CSEP0372), BEC1011 (aka CSEP0264) and BEC1054 (aka CSEP0064). Like many other mildew effectors the RALPHs possess a conserved Y(x)xC motif after the signal peptide [57] as well as a RALPH-specific RxFP motif, which may suggest roles in protein localization or virulence [58]. Like ribotoxins, some RALPHs appear to bind the ribosomal SRL but lack a catalytic site for mRNA cleavage. They have been proposed to have a protective function against host Ribosome Inactivating Ribonucleases (RIPs), which may be induced as part of a resistance response [59].

Prior efforts in remote homology

Fungal effectors do not frequently exhibit detectable sequence similarity with other known sequences, thus finding novel effector candidates in the form of distant homologues is challenging, and may involve relaxing significance thresholds of BLAST *e*-values beyond recommended limits [23, 60]. However, a range of more sensitive sequence-based search techniques are available, which can exploit sequence features that may indicate conserved tertiary structures. Classification systems using cysteine spacing are well established for antimicrobial peptides and some venoms/toxins, where the number of and distance between cysteine residues indicates a possible shared topology of disulphide bonds [61–63]. However, conserved cysteine patterns are not guaranteed to indicate common structure or function [28], and known functional domains or discriminative motif analysis may also be necessary to separate active from non-functional forms [64, 65]. Although they have been a useful heuristic in other applications, cysteine spacing classification generally requires prior knowledge of a well-defined family, which would limit their application to effector family discovery.

Generally, *de novo* remote homology detection falls into two camps: iterative searches and alignments utilizing sequence information from similar proteins (profile search methods) [66], and machine-learning methods, which map the sequence into a multidimensional space (called an embedding, sequence space or feature space) and perform a classification or ranking task. The latter form may use relatively simple sequence features such as k-mers and sequence autocorrelation/covariance features [67], or may themselves use profile search results to construct a redundant representation of the sequence [68, 69]. Although these methods can achieve excellent results, they lack some of the interpretability of classical sequence search methods and are still not in general use. Sequence-based searches are much more commonly used, and the profile sequence-based methods like position-specific scoring matrices (PSSMs; e.g. PSI-BLAST [70]) or profile hidden Markov models (profile-HMMs; e.g. HMMER [71]) can find protein homologues with less than 30% sequence identity. Even more divergent homologues can be found using profile HMM-HMM comparisons [72] or Markov random fields (MRF) [73]. These more sensitive methods can be relatively computationally intensive and some pipelines for remote homology detection will first run PSSM-based methods to reduce run-time [74, 75].

Many of these remote homology detection methods are designed to find homologues of a single protein and are not always applicable to the task of protein family identification. Although iterative PSSM or profile-HMM methods are capable of detecting similarity between very distantly related proteins, extensive post-processing of search results is necessary to remove spurious matches, identify families within search results, and recover potential matches missed by search heuristics or filtering thresholds. Identification of protein families typically involves an all-vs-all comparison between proteins and the construction of a graph (aka network) from significant alignments, from which families can be identified as subgraphs [76]. The best known and still most commonly used algorithm for finding subgraphs corresponding to protein families is by Markov clustering (MCL) originally used in TRIBE-MCL [77]. More recent heuristic algorithms that do not require all-vs-all comparisons have been investigated [78], but are yet to gain widespread use or a stable toolset.

In this study we apply a combination of sensitive sequence comparison methods and protein clustering methods to investigate the possibility of extending fungal effector protein families from the currently known set of fungal effectors. We use an agglomerative clustering approach with iteratively increasing sensitivity to find clusters of protein groups that show differing levels of similarity, which we have termed RemEff. These groups highlight previously unreported relationships between several known effectors, the presence of large effector families, and will support future studies of fungal effector function and evolution. RemEff and the resulting data from this study will also serve as an important resource in the field of molecular plant pathology for effector candidate prediction and study, with relevance to multiple fungal-plant pathogen species.

METHODS

Data sets

Non-redundant fungal protein datasets (Table S1) were downloaded from the UniParc database (https://www.uniprot.org/uniparc/, filter: 'taxonomy:'Fungi (9FUNG) [4751]', downloaded 24 January 2020) and the NCBI Identical Protein Groups database (https://www.ncbi.nlm.nih.gov/ipg/, filter: 'Fungi'[Organism] OR fungi[All Fields]', downloaded 28 January 2020) totalling 10946400 and 11351342 proteins, respectively. Data was supplemented using published genomes from JGI MycoCosm (https://genome.jgi.doe.gov/mycocosm/home), an Endophyte genome database (http://csbio-l.csr.uky.edu/endophyte/) [79–83], the *Alternaria* genome database (https://mycocosm.jgi.doe.gov/Alternaria/Alternaria.info.html) [84], and the 'Gemo' database (http://genome.jouy.inra.fr/gemo/) [85].

Additional genomes, proteomes and effector sequences collected from selected papers were included if they were not represented in the databases (Tables S2 and S3).

Datasets were combined to give a single non-redundant dataset using 'seguid' checksums [86] implemented in BioPython [87]. Proteins were filtered by length, including only proteins longer than 30AA and shorter than 6000AA. Unique sequences corresponding to published effectors and PHI-base entries were identified by searching the initial dataset using MMSeqs2 (version 10-6d92c) [88], requiring a minimum sequence identity of 90% and at least 90% reciprocal coverage, selecting the match with the highest bit-score.

Clustering

The non-redundant fungal protein set was clustered in multiple stages using 'MMSeqs2' (version 10-6d92c) [88]. Protein sequences were initially clustered using the 'cascade' clustering pipeline in three steps to a minimum of 30% sequence identity and 80% coverage of all members. To group more distant sequences, a second stage of clustering was performed using sequence profiles. Clusters were converted to sequence profiles and the profiles were enriched using the original input dataset of fungal proteins (including those sequences ≤ 30 aa or ≥ 6000 aa) to include information from sequences that did not pass the coverage threshold. The enriched profiles were searched against consensus sequences from the cluster profiles and were clustered to have a minimum of 10% identical AAs and 70% reciprocal coverage. In further analyses in this study, these resulting clusters are referred to as 'cluster level 1'.

MSAs for each cluster's sequences were constructed using DECIPHER version 2.10 [89] using the PFASUM15 substitution matrix [90], two iterations, two refinement iterations and alignment adjustments with staggering. A consensus sequence was added to the MSAs using DECIPHER, where columns with more than 50% gaps were considered gaps in the consensus. Code used for clustering sequences and constructing MSAs is available at https://githubcom/darcyabjones/pclust.

Remote homology comparison

To find 'low-level' sequence similarities between level 1 clusters, profile HMM-HMM searches were performed (Supplementary Data Sheet 1). MSAs with consensus sequences were first converted to MMSeqs2 profiles (`--match-mode 1 --match-ratio 1`) and enriched by searching against a database consisting of all fungal sequences of UniRef-90 (downloaded 2020-01-30. Query: 'taxonomy:'Fungi [4751]' AND identity:0.9') and the entire UniRef-50 database (downloaded 30 January 2020), selecting matches with a maximum *e*-value of 10^{-5}. Cluster MSAs and MSAs constructed from the profile matches were combined and converted into an HH-Suite database (version 3.2.0) [72]. 'Match' states in the HMMs and A3M alignments were determined by the consensus sequences of the cluster MSAs prior to enrichment (`--match-ratio=first`), where gaps in the consensus represent an insertion in the model.

To reduce computational requirements and to focus on fungal effectors, a subset of clusters were found by searching selected sequences of known effectors and virulence factors from numerous pathogens included in PHI-base version 4.8 [91], and a custom database of known effector sequences and homologues (Table S3). PHI-base entries to use for subsetting were selected based on annotated phenotypes, functional descriptions and secretion prediction by SignalP versions 3, 4.1 g, and 5.0b [92–94], DeepSig version 1 [95], Phobius version 1.01 [96], and TMHMM version 2.0 c [97]. The sequences were first enriched into MSAs using the cluster HMMs, using two HHBlits search iterations. The enriched sequence MSAs were then searched against the cluster HMMs allowing a maximum *e*-value of 0.01, minimum probability of 0.20 and realigning up to 20000 matches (`-n 1 -e 0.01p 20 -Z 20000 -z 0 -B 20000 -b 0`). Code used for constructing HMMs, subsetting the database, and performing all-vs-all comparison is available at https://github.com/darcyabjones/pclust.

To identify remotely homologous clusters (referred to here as level 2 clusters or superclusters), the subset of HMMs matching selected PHI-base or effector sequences were searched against themselves (all-vs-all) using HHblits (`-n 1 -e 0.01 -E 0.01 -z 0 -Z 20000 -b 0 -B 20000 -pre_evalue_thresh 10min_prefilter_hits 10 -realign_max 20000`). Pairwise matches were considered significant if they had an *e*-value $\leq 10^{-5}$, probability ≥0.9, alignment length ≥>=30 AAs, and where the alignment covered at least 70% of at least one HMM in the pair. Where there were multiple alignments between the same pair of proteins, the alignment with the highest score was selected. Alignments were then filtered so that only reciprocally significant matches were retained. To reduce any score bias in alignments caused by HMM lengths, we adopt the normalization approach used by OrthoFinder [98] with modifications. Briefly, HMM search self matches were selected from the alignments, the HMM length was squared, the selected alignments were sorted by the squared HMM length, and the top 5% of alignments (by score) were selected from non-overlapping 1000 element sized bins in the sorted list. The $\log_2$ transformed alignment scores were regressed on the $\log_2$ transformed squared HMM lengths, and the slope and intercept were taken to transform scores using the same formula described in Emms and Kelly [98]. Conceptually, this transformation normalizes the scores by the average maximum possible score for an alignment of two proteins with those lengths. Alignments were then further filtered to require alignments between both HMMs to be covered at least 70% of their respective lengths, and reciprocal matches were selected again. Each pair of alignments were grouped and the arithmetic mean of the two normalized scores was used as a single score for each pair, and the scores were converted to a value between 0 and 1 by dividing by the highest score in the full set of pairwise matches.

The filtered, score-normalized alignments were used to construct a weighted, undirected graph (AKA network) using the Python libraries, networkx [99], Pandas [100] and SciPy [101]. Clusters (superclusters) in the graph were found using a reimplementation of the greedy set cover algorithm [102] and with Markov clustering [77] (https://github.com/GuyAllard/markov_clustering), which due to their relatively similar stringencies were designated 'cluster level 2A' and 'cluster level 2B', respectively. Connected components were also found to summarize higher-order relationships, which were also designated as 'cluster level 3'. The MCL inflation parameter was selected by running MCL on ten randomly selected connected components containing 300–600 nodes,

Table 1. Summary of the number of unique sequences in the input dataset (A) and the number of clusters obtained using various methods for remote homology clustering (B)

(A) Initial clustering of input data for removal of sequence redundancies				
	Total	**Uniparc**	**NCBI IPG**	**Custom**
Unique Sequences	14425844	11987341	12293758	3130080
Unique sequences per taxid	23351787	19081482	14297803	3302707

(B) Remote homology clustering		
Level	**Total dataset**	**Containing known effectors**
1 (profile)	286512	200
2A (greedy)	45363	103
2B (Markov)	27851	104
3 (connected components)	6538	80

for a range of inflation parameters between 1.1 and 2.0. The inflation parameter that gave consistently higher modularity scores [103] was selected for overall clustering. Graphs and subgraphs were visualized using the Graph Tool Python library [104].

Supercluster comparison

To interrogate sequence conservation within and between cluster levels 2 and 3, composite multiple sequence alignments were constructed and visualised as sequence logos [105].

Enriched multiple sequence alignments from level 1 clusters (used to form HMMs) were combined using a progressive algorithm, guided by the maximum spanning tree of the subgraph containing the clusters of interest, where MSAs were pairwise aligned using hhalign [72]. For each pairwise alignment, two alignments were computed using each MSA as the template. A pairwise alignment was considered to have succeeded if the resulting MSA contained sequences from both input MSAs, and if both pairwise MSAs succeeded the result containing more sequences was selected for further iterations. If an alignment failed to merge two MSAs, the alignment was scheduled to be retried after all other alignments had been completed, stopping if no more MSAs could be merged. Un-connected components from this progressive method were then pairwise aligned in random order, shuffling the list if no MSAs had been merged in a full pass through the list, and stopping the process if no MSAs had been merged in ten passes through the list, resulting in one or more MSAs. Each combined MSA was converted into an HMMER profile HMM [71], and used as a template profile to align all sequences from the level 1 clusters (not including the enrichment sequences) using Clustal omega [106].

Sequence logos were computed by filtering out sequences from the final MSA with more than 90% pairwise identity using `hhfiltr`; [72], and the resulting MSAs were plotted using Logomaker [107] and matplotlib [108] using the information content as the logo heights.

Comparison to iterative search results JackHMMER

We compared the output clusters of the RemEff pipeline to an output from a standard progressive HMM search using JackHMMER [71] for benchmarking purposes. The same sequences used to initial clusters of interest before remote homology inference (PHI-base version 4.8, and effector sequences and homologues in Table S3) were searched against the entire non-redundant fungal protein dataset using JackHMMER with a maximum of five iterations and a maximum *e*-value of 1 (`-N 5 -E 1 -domE 1`). The resulting matches were then filtered to have a full sequence *e*-value of ≤ 0.01, and summarized to show the number of significant matches, and matches to known effectors or published effector homologues.

RESULTS

Protein dataset and initial sequence clustering

Nearly fourteen and a half million unique sequences spanning 69724 distinct NCBI taxonomic ids were collected from public databases for clustering and remote homology comparison (Table 1). A first pass clustering of these proteins with MMSeqs2 [88] yielded 3111468 clusters, which were designated as ‘level 1’ clusters (Fig. 1). Within this first pass, position-specific scoring matrix (PSSM) profile clustering did not merge any clusters from the standard MMSeqs2 ‘cascaded clustering’ pipeline, but was observed to merge clusters in datasets with fewer sequences and when the coverage criteria were relaxed from 80% reciprocal as required here. The majority (1918741) of level 1 clusters consisted of one or two (the median) unique sequences (Fig. 2f). A smaller number of large clusters were observed, with the largest 10% of clusters containing eight or more unique sequences and a maximum cluster size of 10244. From these level 1 clusters, enriched profile-hidden markov models (HMMs) were constructed for use with HH-Suite version 3.2.0 [72]. To focus on finding potentially grouped families of effectors, HMMs

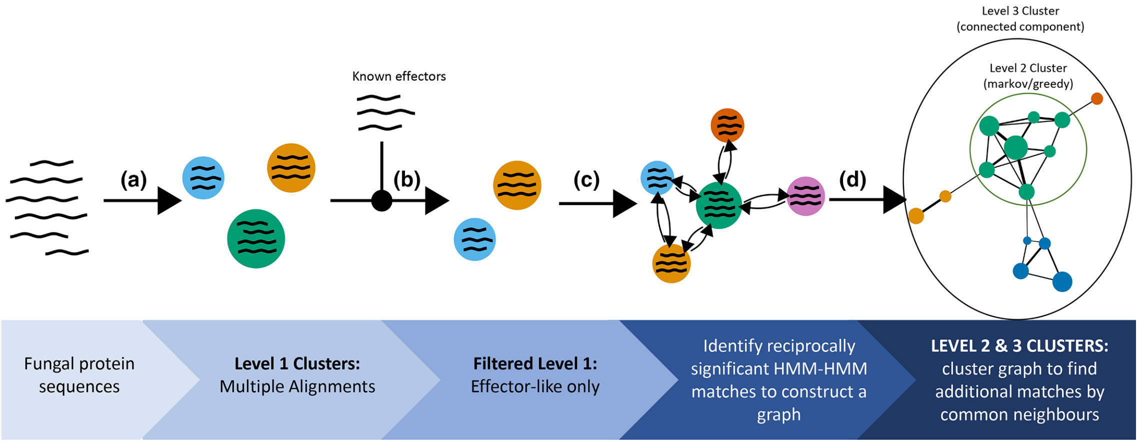
Level 3 Cluster
(connected component)
Level 2 Cluster
(markov/greedy)
Known effectors
(a)
(b)
(c)
(d)
Fungal protein sequences
Level 1 Clusters: Multiple Alignments
Filtered Level 1: Effector-like only
Identify reciprocally significant HMM-HMM matches to construct a graph
LEVEL 2 & 3 CLUSTERS: cluster graph to find additional matches by common neighbours

Fig. 1. The clustering workflow employed in this study. (a) Sequences are initially clustered using MMSeqs2 resulting in 3111468 level 1 clusters. (b) A subset of 286512 of these clusters with any similarity to known effectors are found using HHBlits. (c) All clusters from this subset are searched against each other and reciprocally significant alignments are selected to form a graph. (d) Clusters of the initial clusters are found within the resulting graph to form more distant sequence families. In the final graph, each point represents a level 1 cluster resulting from step A, the colours indicate level 2 clusters (Markov or greedy clustering), and the whole graph forms a single connected component (level 3 cluster).

for remote homology comparison were selected based on HHBlits matches to 6598 selected PHI-base [91] and effector sequences (Table S3). Of these sequences, 1078 sequences matched 286512 level 1 cluster HMMs with a maximum *e*-value of 0.01 and minimum probability of 20%, which were selected for HMM-HMM comparisons. Of the subset of clusters selected for remote homology comparison, 2856 clusters contained unique sequences corresponding to 310

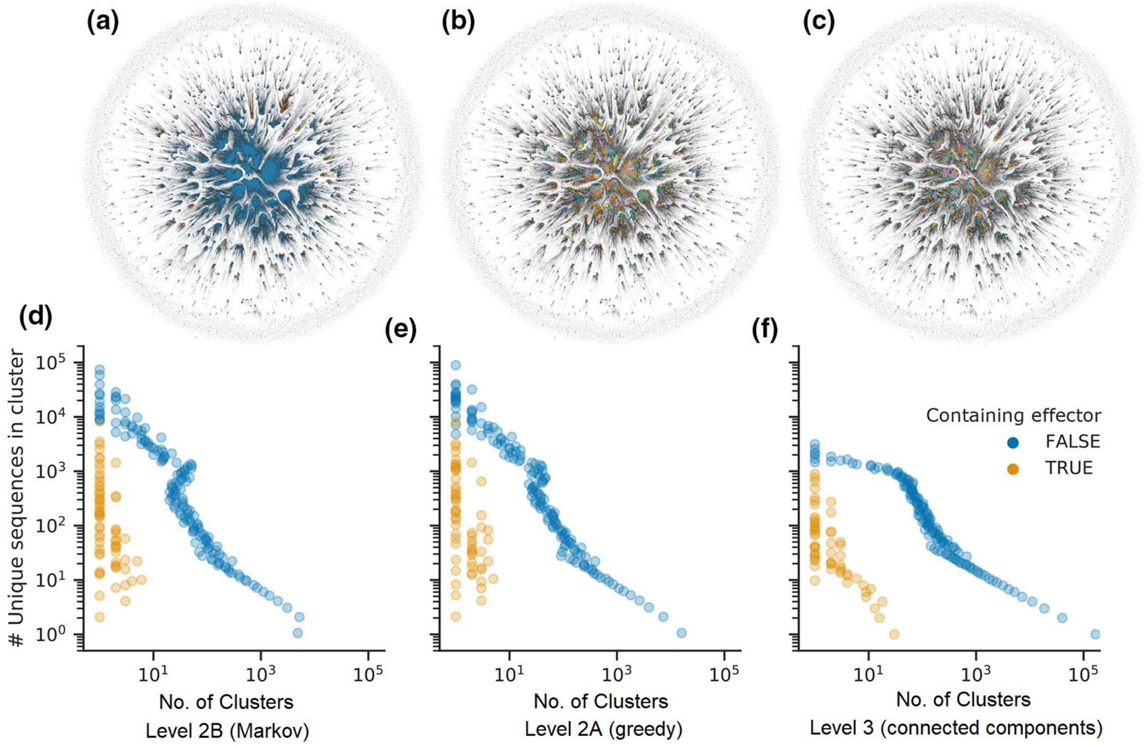

Fig. 2. Top row: plot of graph coloured by connected components (a), and Markov (b) and greedy (c) clusters. Bottom row: the number of unique sequences compared with the number of clusters with that size, within Markov (d), greedy (e) and profile clusters (f). For the bottom row, Y-axis values are in binned into 100 evenly sized ranges taken from a 10-based exponential space ($10^{0 \cdot\cdot \max(\#seqs)}$).

sequences from the effector and effector homologue dataset, and 3571 PHI-base entries.

Clustering of profile HMM-HMM matches to identify remote homology relationships between effector-like sequence clusters

To identify more distant relationships, all-vs-all profile HMM-HMM comparisons were performed on the 286512 selected (effector-like) level 1 clusters (Fig. 1). A total of 224230 level 1 clusters were connected by 30472762 reciprocally significant alignments (*e*-value <=10^{-5}, probability >90%, reciprocal coverage >70%). The remaining clusters had no matches at this significance threshold, and so could not be grouped into more remote clusters. Of these clusters without significant matches, 71 clusters contained unique sequences corresponding to 104 effectors and effector homologues. A strong correlation between alignment scores and sequence lengths was observed, which was effectively removed by normalization (Fig. S1). A graph was constructed of the level 1 effector-like clusters and their connecting alignments, using the mean of the pair of normalized scores as edge weights. The graph consisted of 6538 connected components (sub-graphs of directly or indirectly connected level 1 clusters), which were designated 'level 3' clusters. A single large component containing 171346 nodes/level 1 clusters (representing 1.9 million unique proteins) was observed, with numerous small components typically with fewer than 1000 nodes also present (Table 1b, Fig. 2a). Despite the presence of one large connected component, most level 1 clusters, which corresponded to known effectors were found in smaller components, with only 19 out of 310 known effectors and effector homologues found in the largest component. These were typically highly conserved protein families including LysM-domain containing proteins, CRN, Tom1, Ave1 (expansins/PNPs), MoMSP (a cerato-platanin), though the NEPs and ribotoxins each formed a separate component. In order to sub-divide larger connected components into more stringent remote-homology groupings, 'communities' or 'superclusters' of level 1 clusters were found within connected components using the greedy set cover [102] and Markov clustering [77] algorithms. A Markov clustering inflation value of 1.35 was selected for clustering, which gave the highest average modularity scores [103] for a random selection of smaller connected components. Greedy and Markov clustering (referred to as 'level 2A' and 'level 2B' clusters, respectively) generally yielded comparable groupings, but greedy clustering tended to produce more clusters with only one member (Fig. 2d, e, Tables S4 and S5). Although each clustering method is different, conceptually the cluster levels from 1 to 3 represent progressively more distant homology relationships.

Level 2 and 3 clusters grouped multiple known effectors and predicted an expanded set of effector candidates across multiple pathogen species

We found 80 clusters at level 3, 103 and 104 clusters at level 2A and 2B, respectively, and 200 clusters at level 1, which contained known effectors and published effector homologues (Table 1). Of these, 20, 24 and 26, and 10, respectively, contained two or more known effectors (Table S4, Fig. 2d, e). To demonstrate how known effectors have been grouped into novel 'families' in this study, we present three examples in detail. The first examples consist of a level 3 cluster (connected component) that contains *Leptosphaeria maculans* AvrLm6 [47], *Magnaporthe oryzae* BAS4 and SPD5 [109, 110], *Fusarium oxysporum* f. sp. *lycopersici* SIX5 [111], and *Cercospora beticola* NIP1 [112] (Fig. 3a, b). At cluster level 2, this group is further divided into sub-groups, with AvrLm6 and some published *Venturia inaequalis* AvrLm6 homologues [49] forming a distinct sub-group, *M. oryzae* SPD5 and BAS4 forming another sub-group, and *C. beticola* NIP1 and *F. oxysporum* SIX5 both forming their own distinct sub-groups, with additional sub-groups that did not match a known effector. Sequence logos generated from a multiple sequence alignment of all level 1 clusters contained in these subgroups (Figs 3 and S2, Supplementary Data Sheets 2–5) indicated conservation of specific cysteine, threonine, and glycine residues, as well as distinct motifs that were specific to each sub-group. Level 2A clusters PC_02VR38 (containing AvrLm6), PC_01204B (containing CbNip1), and PC_03MDGJ (containing SPD5 and BAS4) are found in numerous species from the Leotiomyceta clade including *Bipolaris* spp., *Colletotrichum* spp., *Leptosphaeria* spp., *Venturia* spp., and *Fusarium* spp. (Tables S4 and S5). The cluster containing *Fol* SIX5 (PC_05PCSX) was found in a broader range of taxa including the basidiomycetes *Jaapia argillacea* and *Plicaturopsis crispa*, but most observed sequences were from species in the Pezizomycotina, including other significant plant pathogens such as *Zymoseptoria tritici* and *Pyrenophora teres* f. sp. *teres*.

The second example consists of two separate connected components (level 3) clusters that correspond to the conserved ToxA effectors of *Pyrenophora tritici-repentis*, *Parastagonospora nodorum* and *Bipolaris* spp. [21, 35], and a set of loosely conserved 'ToxA-like' proteins, which had been previously identified in other studies using PSI-BLAST searches, including ChEC13 [21] and AvrFOM2 [22] (Fig. 4a, b). Our method did not link these two reportedly related groups within a single level 3 cluster. Alignments between the two connected components were observed, but failed the *e*-value significance threshold (data not shown). Multiple sequence alignment combining all sequences from both level 3 clusters showed only low-level similarity between sequences of these two level 3 clusters (Figs 4 and S3, Supplementary Data Sheet 6). Between the ToxA and ToxA-like/AvrFOM2 clusters, there are several broadly conserved residues, most notably two cysteines, two aromatic [W|Y] residues, several aliphatic [L,I] residues, and an LxxRQ…C motif. The AvrFOM2 cluster(s) are more diverse than ToxA, with conserved residues separated by hypervariable regions, and also possess a phenylalanine rich region in the signal peptide. The AvrFOM2 cluster also lacks a recognizable 'RGD' [42] or 'SGN' [21] motif (positions 138–140 in Fig. 4c), which is absent in the level 2B (Markov) cluster PC_08EP4N and replaced in this alignment by a poorly conserved 'TTP' consensus in level 2B cluster PC_07OLPP. The component containing ToxA sequences

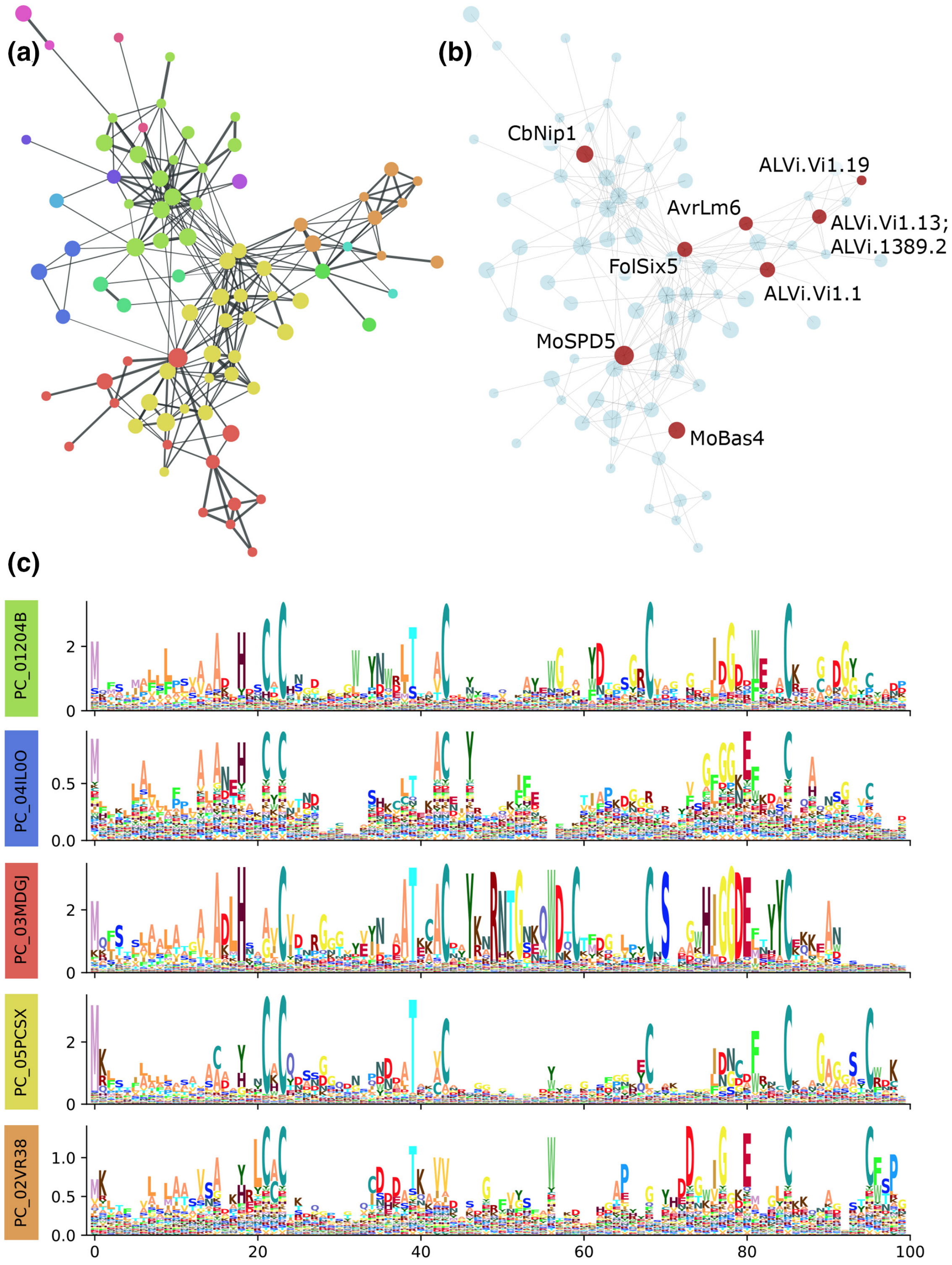

Fig. 3. A family of SIX5-like effector sequences. (a) The connected component containing the effectors AvrLm6, Bas4, SPD5, CbNIP1 and SIX5, coloured by Markov cluster membership (level 2B). Nodes in the graph each represent a single HMM (level 1 cluster) with node size indicating the relative number of unique sequences contained in the HMM, and edges represent where a significant sequence alignment was found between the two HMMs. (b) The same graph, but highlighting the level 1 clusters containing effector sequences and published effector homologues (ALVI*). (c) Sequence logos resulting from multiple sequence alignment of all sequences in the connected component (level 3 clusters). Logos for Markov clusters with more than 10 members are shown separately. Columns in the multiple sequence alignment with more than 50% gaps are excluded.

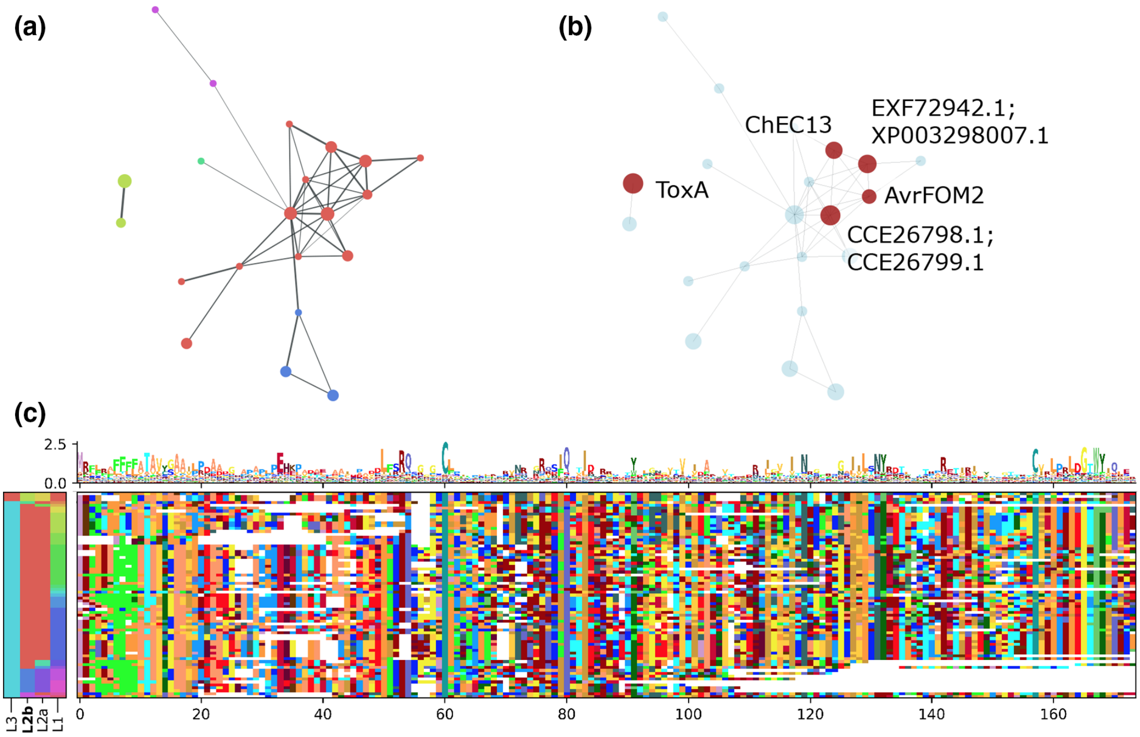

Fig. 4. ToxA-like fungal effector groups. (a) The connected components (level 3 clusters) containing ToxA-like and AvrFOM2-like sequences, coloured by Markov cluster membership (level 2B). Nodes in the graph each represent a single HMM (level 1 cluster) with node size indicating the relative number of unique sequences contained in the HMM, and edges represent where a significant sequence alignment was found between the two HMMs. (b) The same graph shown in (a), but highlighting level 1 clusters containing known effectors and published effector homologues. (c) A multiple sequence alignment constructed from all sequences in the ToxA-like and AvrFOM2-like connected components. Columns in the multiple sequence alignment with more than 50% gaps are excluded. Colours on the y-axis indicate the level 1, 2 and 3 clusters that members belong to, with level 2B (Markov) cluster colours matching those in (a).

(PC_03B2DN) was observed to have members in several species that have been previously described: *Pyrenophora teres formae speciales*, *Pyrenophora tritici-repentis*, *Parastagonospora nodorum*, *Parastagonospora avenae*, *Bipolaris maydis* and *Bipolaris sorokiniana* (Tables S4 and S5). The level 2B cluster containing AvrFOM2 and ChEC13 was observed in numerous species within the leotiomyceta clade, including *Epichloe* spp., *Fusarium* spp., *Pyrenophora* spp., *Colletotrichum* spp. and *Bipolaris* spp. Other level 2 clusters within the component containing AvrFOM2 and ChEC13 also contain members from the leotiomyceta, but are specific to genus (*Epichloe*), or a strain (*Zymoseptoria ardabiliae* STIR04_1.1.1, *Balansia obtecta* B249). In a similar manner to the ToxA and ToxA-like clusters, members of the MAX effector family and the homologues published by de Guillen *et al.* [23] were found in 12 separate connected components in this study. None of these components contained more than one of the experimentally validated MAX members (ToxB, Avr1_CO39, AVR_Pik, AvrPiz_t and AvrPib). (Fig. S4, Supplementary Data Sheet 7) (4).

The third example is a level 3 cluster of RNase-like effectors that grouped level 1 clusters, which were sufficiently divergent that profile alignments between all sequences in the MSA was not possible. Consequently, further presentation of this example focuses on a sub-graph, which includes level 2A and 2B clusters containing known RALPH [24, 58, 59] and ribotoxin [56] effectors (Fig. 5a). The ribotoxins (including Zt6) formed a large and densely connected cluster, which was distinct from all RALPH effectors (Fig. 5b, c). The RALPH effectors consist of three main groups: AvrPm3$^{a2/f2}$, AvrPm2/BEC1054/AVR$_{a13}$ and SvrPm3$^{a1/f1}$, and are sparsely connected. Multiple sequence alignment of all sequences in the selected clusters indicate two or four conserved cysteine positions in the RALPH and ribotoxin logos, respectively (Figs 5 and S5, Supplementary Data Sheet 8). Additional conserved proline, aromatic [Y|F], and aliphatic [V|I] residues were observed. The clusters containing AvrPm2-like RALPH proteins (level 2B cluster PC_04SK9M) were more similar to the ribotoxin/Zt6-like cluster (level 2B cluster PC_032CKH), than the clusters containing AvrPm3$^{a2/f2}$/SvrPm3$^{a1/f1}$ sequences (level 2B clusters PC_01D3OM and PC_0278ZT, respectively). The Y(x)xC motif commonly found in after the signal peptide in *Blumeria* effectors [57, 58] appeared to be enriched in AvrPm2-like and AvrPm3$^{a2/f2}$-like RALPH sequences, but may be replaced by an F(x)xC motif in SvrPm3$^{a1/f1}$-like sequences. The level 2B cluster containing the known ribotoxin effector Zt6 (PC_032CKH) was broadly conserved in the Fungal kingdom (Tables S4 and

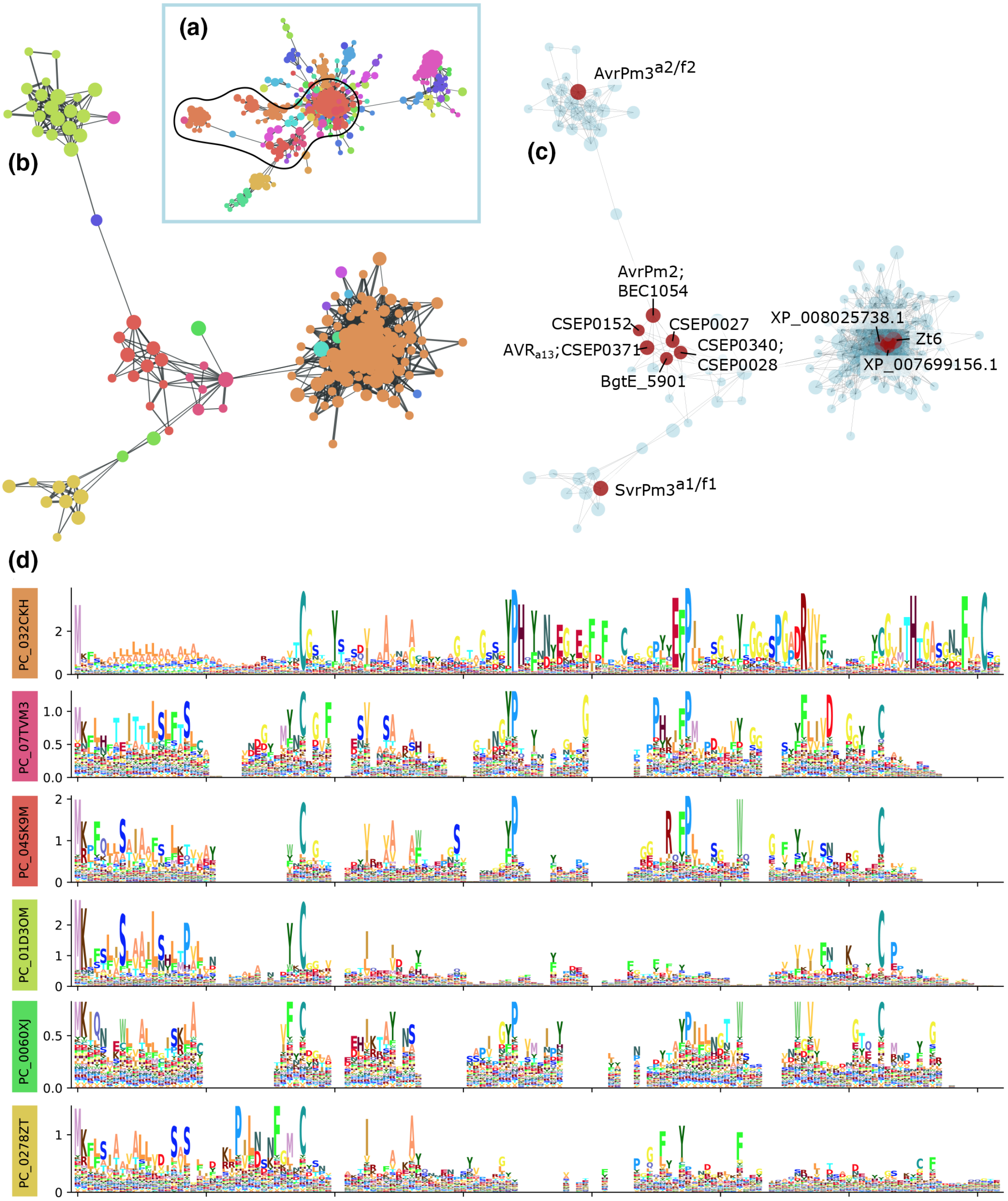

Fig. 5. A connected component containing RNase-like effectors. A single connected component containing the ribotoxins and RALPH effectors was observed (a). Nodes in the graph each represent a single HMM (level 1 cluster) with node size indicating the relative number of unique sequences contained in the HMM, and edges represent where a significant sequence alignment was found between the two HMMs. (b) Shows a subset of the connected component containing all level 2 clusters containing effector sequences (c). (d) Sequence logos for selected level 2B (Markov) cluster from a multiple sequence alignment of all sequences in (b). Colours in the left boxes corresponding to colours in (b). Logos with fewer than 10 members are not shown. Columns in the MSA with greater than 50% gaps are excluded from the visualisation.

S5). Sequences belonging to level 2B clusters corresponding to RALPH effectors (AvrPm2/PC_04SK9M, SvrPm3$^{a1/f1}$/PC_0278ZT, and AvrPm3$^{a2/f2}$/PC_01D3OM) were only found in *Blumeria graminis formae speciales*. However, several other lineage specific level 2 clusters were observed within the same connected component, which were most often associated with the Pezizomycotina.

Ten other level 2 clusters that grouped two or more known effectors were identified (Table S4), grouping: *Leptosphaeria maculans* AvrLm2 and *Fusarium oxysporum f.* sp. *lycopersici* SIX1 [113, 114]; *Zymoseptoria tritici* NIP1 and *Passalora fulva* Ecp2 [115, 116]; *Blumeria graminis f.* sp. *hordei* BEC2 and *Golovinomyces orontii* GoEC2 [117]; *Passalora fulva* Ecp6 and *Zymoseptoria tritici* Mg3LysM [118, 119]; NIS1 effectors [120, 121]; *Zymoseptoria tritici* MgxLysM and Mg1LysM [119]; *Magnaporthe oryzae* AVR-Pita and AVR-Pita2 [122, 123]; *Puccinia striiformis* Shr4 and Shr6 [124]; Pit2 effectors [125, 126]; and the NEP virulence factors [127–129, 130].

In other cases, clusters containing a single known effector were assigned functional annotations of high relevance to potential effector functions. For example, a large cluster (level 3: PC_07OBLJ, level 2: PC_058FSP, Table S4) corresponding to known effector BAS3 (biotrophy-associated secreted protein 3) of *Magnaporthe oryzae* [109], was functionally annotated as similar to scorpion knottin toxins [InterPro: IPR036574]. This group contained unconfirmed candidates from other *Pyricularia* spp., as well as *Colletotrichum* spp., *Macrophomina phaseolina*, *Neofusicoccum parvum* and *Monosporascus* spp.

Some known effectors were not able to be grouped beyond cluster level 1 (Table S4). Of these, 21 were within clusters that contained a single unique sequence, including AvrLm11 and AvrLmJ1 of *Leptosphaeria maculans* [131, 132], Avr5 of *Passalora fulva* [133], AVR$_{a10}$ of *Blumeria graminis f.* sp. *tritici* [134], PIIN_08944 and FGB1 of *Piriformospora indica* [135, 136], CDIP3 and Slp1 of *Magnaporthe oryzae* [137, 138], lsc1 of *Verticillium dahliae* [139], Zt80707 of *Zymoseptoria tritici* [140], and the putative effector CSEP-07 of *Phakopsora pachyrhizi* [141]. Another 50 level 1 clusters containing a single effector possessed two or more unique sequences, for which most were restricted to isolates of the same species or genus. These included Tox1 of *Parastagonospora* [142]; SIX2, SIX4, and SIX8 of *Fusarium oxysporum* [111, 143]; DN3 and EP1 of *Colletotrichum* [144, 145]; AvrP4 and AvrL567 of *Melampsora* [146, 147]; SSVP1 of Sclerotiniaceae [148]; Shr5 and Shr7 of *Puccinia* [124]; NIP3 of *Rhynchosporium* [149]; and SCP7 of *Verticillium* [150]. Notable exceptions of level 1 clusters which spanned genera were Ecp1 of *Passalora fulva* [116] which had a homologue in *Pseudocercospora eumusae*, AVR4 of *P. fulva* [151] which had a homologue in *Dothistroma septosporum*, AvrLm3 of *L. maculans* [152] which had homologues in *P. fulva* and *Fusarium oxysporum f.* sp. *narcissi*, and SSVP1 of *Sclerotinia sclerotiorum* [148] which had homologs in *Botrytis* spp. and *Monilinia laxa*.

Benchmarking RemEff clusters against progressive profile-HMMs (JackHMMER) cluster

To determine whether the observed relationships reported by RemEff are comparable to outputs from progressive profile-HMM methods that have previously been used to identify several effector families [21–23, 48, 49], the 591 effector and effector-like sequences (Table S3) were searched against the complete non-redundant protein dataset using JackHMMER [71] (five search iterations, full length *e*-value ≤ 0.01). Of the 591 query sequences, 433 had significant matches to 942978 distinct proteins in this dataset. Of those, 640727 were present in the filtered clustering dataset used for HMM-HMM comparison in the RemEff pipeline, corresponding to 97887 level 1, 15986 level 2A (greedy), 10935 level 2B (Markov), and 2566 level 3 (connected component) clusters.

JackHMMER recovered many of the putative relationships between effectors identified by RemEff as well as some additional matches, but also missed relationships inferred by graph clustering (Table S6). In the SIX5-like group presented in Fig. 3, each of SIX5, CbNIP1, BAS4 and SPD5 were reciprocally significant. However, using AvrLm6 as the query only identified SIX5 and SPD5 as candidate homologues, and only the match with SIX5 was reciprocally significant. As previously described by [22] (with corresponding RemEff clusters presented in Fig. 4), Pn/Ptr-ToxA, ChToxA, ChEC13 and AvrFOM2 were all identified as being reciprocally significant matches using JackHMMER. The RALPH effectors AvrPm2, AvrA13 and BEC1054, were all reciprocally significant, and all matched the ribotoxin Zt6, but only the match to BEC1054 was reciprocally significant. Neither AvrPm3a2/f2 nor SvrPm3a2/f2 were identified as potential homologues of other RALPH members using JackHMMER. SvrPm3a2/f2 matched Zt6 but was not reciprocally significant. Similar to RemEff, none of the experimentally validated MAX effectors (ToxB, Avr1_CO39, AVR_Pik, AvrPiz_t, and AvrPib) were identified as matches to each other using JackHMMER.

Some additional potential effector similarities not reported by RemEff were observed in the JackHMMER results. Searching using AvrLm3 as the query identified AvrLmJ1 as a potential match, but this was not reciprocally significant. Local alignment of the AvrLm3 and AvrLmJ1 amino acid sequences using the EMBOSS water aligner with a BLOSUM45 substitution matrix [153] did show some similarity between the two sequences and potentially a shared cysteine spacing pattern. SIX6, SIX1 and SIX13 were identified as possible matches to SIX4 using JackHMMER, and SIX13 also matched SIX6. However, none of these alignments were reciprocally significant. Multiple sequence alignment of SIX1, SIX6, SIX13, SIX4 and AvrLm2 (which is a reciprocally significant match to SIX1, also identified using the clustering method) amino acid sequences using MAFFT [154] showed only a small central region of similarity containing four conserved cysteine positions. Using CfECP2 as a query JackHMMER identified PgShr8 as a potential candidate match, but this match was not reciprocally significant. Multiple sequence alignment of CfECP2, ZtNIP1 (which was a reciprocally significant match to CfECP2), and PgShr8 showed some conserved glycine, threonine, cysteine and valine residues between the three sequences in the central and C-terminal regions. However, the alignment indicated the presence of an additional N-terminal region in Shr8, and Shr8 was clearly distinct from ZtNIP1 and CfEcp2 (Supplementary Data Sheet 9), Finally, BEC2, GoEC2, BEC4 and MoCDIP2 were all identified as reciprocally significant matches using JackHMMER, though some alignments did have a high 'bias' score indicating that they may be spurious matches. Multiple sequence alignment of these four sequences identified a conserved N-terminal region after the signal peptide with several conserved cysteine, proline, aspartic acid and glycine residues. BEC2 and GoEC2 were more similar than BEC4 and MoCDIP2.

DISCUSSION

With a growing number of experimentally confirmed fungal 'effector' proteins in the public domain (Table S3) [91], there are emerging opportunities to mine this data and develop improved methods for effector and virulence factor discovery. However, basic homology-inference methods cannot necessarily be applied, as many known effector proteins are either sufficiently divergent or of independent origin to prevent their grouping into larger 'effector families'. Comparisons between effector proteins and candidates at the structural level have indicated recognizable structural similarity between many emerging groupings, including the ToxA-like [21, 22], MAX [23], RALPH [24, 58, 59] and Hce2 [155] families. Tertiary structural homology may become the basis for reliable effector prediction in future studies; however, the application of protein structure prediction to large sets of effector candidates is not currently computationally feasible. This study applied a highly sensitive sequence clustering approach - termed 'RemEff' - to a large protein dataset to form novel protein clusters, leveraging known effectors to identify effector 'family' clusters and predict homologous effector candidates within them by association.

While the RemEff method has taken a 'top-down' approach that has identified a large number of 'effector families', we focus here on selected examples. In our first detailed example, we presented a previously undescribed expanded family of effectors containing the effectors *Leptosphaeria maculans* AvrLm6 [47], *Magnaporthe oryzae* BAS4 and SPD5 [109, 110], *Fusarium oxysporum* f. sp. *lycopersici* SIX5 [111] and *Cercospora beticola* NIP1 [112]. Each study describing the effectors has noted the presence of homologues of these effectors in multiple species. Numerous homologues of AvrLm6 have been previously observed in *Venturia*, *Colletotrichum* and *Fusarium* species [48, 49]. However, not all *Venturia* AvrLm6 homologues published by Shiller *et al.* [49] were identified as members of this superfamily. In that study the only restriction on matches was a maximum PSI-BLAST *e*-value of 10^{-2}, so it is likely that the focus here on finding full-length homologues might have excluded these potential matches. Each of the five level 2 clusters within the level 3 cluster had a different cysteine spacing pattern, with four or six conserved cysteine positions each. Some cysteine residues were conserved across multiple groups, and two positions were conserved in all subgroups suggesting their functional relevance.

None of these SIX5-like effectors have yet been structurally determined. SIX5 appears to interact with plasmodesmata and mediates the intercellular translocation of another effector Avr2 (where it can then exert virulence promoting and avirulence function) [156]. Intriguingly, another pair of *Leptosphaeria maculans* effectors not present in this study, one of which is a SIX5 homologue, appear to show a similar interaction in that pathosystem [157]. *Magnaporthe* biotrophy-associated secreted protein 4 (BAS4) elicits a host defence response late in the biotrophic phase, which promotes cell death during the necrotrophic phase [158, 109]. Cytoplasmic effectors PWL2 and BAS1, but not BAS4, move from cell to cell preceding the invasive hyphae (IH), possibly through plasmodesmata [159]. Suppressor of cell death 5 (SPD5) is a known homologue of BAS4, which suppresses BAX- and NEP1-induced cell death [110]. CbNIP1 induces light-independent necrosis [112], but its specific activity and cell localization is unknown. In contrast to SIX5, which internalizes into the host cytoplasm [156], BAS4 is accumulated in the apoplast [158]. The host internalization mechanisms of most effector proteins are not well understood, but some may require short, conserved, amino acid motifs [25, 160]. If this is also the case for SIX5, these internalization motifs are not likely to be the conserved pattern or structure being detected by RemEff across the SIX5-like cluster. In a similar vein, previous tertiary structure comparisons (although not RemEff) had defined the MAX-effector group [23] which also grouped several cytoplasmic *Magnaporthe* effectors with the apoplastic *Pyrenophora* ToxB [161]. *Magnaporthe oryzae* appears to have a distinct mechanism for effector host-internalization involving a specialized infection structure – the biotrophic interfacial complex (BIC) – and via extracellular vesicles [162]. The necrotrophic *Pyrenophora* spp. lack such structures. We surmise that for both the SIX5-like cluster and the MAX family, any common functions related to detectable conserved structures are unrelated to cell internalization.

In our second detailed example, we compared two other clusters containing the effectors ToxA and AvrFOM2, which were previously reported as similar [22]. The cluster containing AvrFOM2 is much larger and more sequence diverse compared to the one containing ToxA. Within the ToxA level 3 cluster are only the canonical ToxA-like effectors of *Parastagonospora* spp., *Pyrenophora* spp., *Bipolaris sorokiniana* and *B. maydis*, many of which are identical and are thought to have arisen by a complex horizontal transfer event [32]. The level 3 cluster containing AvrFOM2 and ChEC13 overlaps considerably with the candidate homologues identified in Lu *et al.* [21]. Although a *Fusarium* *oxysporum f.* sp. *melonis* homologue was described in that paper, it does not appear to have been AvrFOM2. The multiple sequence alignment does show the conservation of some of the motifs previously described [21], including the LXXR pro-peptide cleavage site, and the three motifs found in beta sheets 4 (LXVXIXN, here replaced by IXVXIXN in PC_07OLPP containing AvrFOM2), 5 (LILTXY, replaced by I[VI]LSNY in PC_07OLPP) and 8 (WXXQ). However, neither the asparagine rich motif (WXXN(S)NXIXVXI) nor the RGD/SGN motif were observed. The level 2 cluster containing AvrFOM2 and ChEC13 (PC_07OLPP) exhibits a number of phenylalanine residues in the signal peptide (SP) at the junction of the N-region and the hydrophobic core. Hydrophobic amino acids near the N-region tend to decrease secretion efficiency [163] and although phenylalanine residues are found in the hydrophobic core regions of human signal peptides, it is not generally known in yeasts [164]. However, the amino acid composition of efficient signal peptides can vary between species, and the hydrophobic and N-terminal regions of the SP may be involved in directing proteins through different secretion pathways [163].

In our last detailed example including ribotoxins and RALPH effectors, the clusters containing ribotoxins (Zt6) and RNase-like proteins associated with haustoria (RALPH) effectors formed distinct clusters, but a clear similarity existed at specific regions between clusters PC_032CHK containing Zt6 and PC_04SK9M containing AvrPm2 and BEC1054. In the ribotoxin sequence α-sarcin the active sites are Histidine in the YPH motif, Glutamine in EFP motif, and a Histidine between the last two cysteine residues, all of which are missing in RALPHs though they possess the conserved surrounding sequence of the former two [165–167]. Additionally, all RALPH sequences lacked the extended N-terminal loop that has previously been thought to be necessary for ribotoxin activity, though it was also poorly conserved in the cluster containing Zt6 [7]. Overall, the profiles of RALPH effectors, with only two conserved cysteine positions, is more like RNase T1 than the ribotoxins, are missing many of the previously described active sites, and have shorter loop sequences than the canonical ribotoxins. This is consistent with previous structural prediction analysis [58], and makes sense given that *Blumeria graminis*, to which this group appears to be restricted, are obligate biotrophs which would not benefit from effectors with cytotoxic activity. This also supports speculation that BEC1014 acts as a pseudoenzyme, binding host ribosomes but not cleaving the SRL [59]. Both AvrPm3$^{a2/f2}$ and the suppressor SvrPm3$^{a1/f1}$ form distinct level 2 clusters branching from the main group of RALPH effectors (PC_01D3OM and PC_0278ZT, respectively). SvrPm3$^{a1/f1}$ was originally described as being a member of the RALPH group, but AvrPm3$^{a2/f2}$ was not [24]. It has previously been demonstrated that high expression of SvrPm3$^{a1/f1}$ suppresses the recognition of AvrPm3$^{a2/f2}$ by Pm3 receptors [168], and that positive selection in the *avrpm3*$^{a2/f2}$ gene does not appear to be related to evasion of recognition by Pm3 [169]. Although the clusters are quite distinct, their association may suggest a possible mode of SvrPm3$^{a1/f1}$ suppression, where it may act as a 'bodyguard decoy' to AvrPm3$^{a2/f2}$ [170]. However, we note that the level 2 clusters containing AvrPm3$^{a2/f2}$ and SvrPm3$^{a1/f1}$ may be poorly aligned here, and AvrPm3$^{a2/f2}$ shares little conserved sequence similarity with the other RALPH effectors beyond the signal peptide and the cysteine positions.

Several other effectors formed groups of more than one effector, including two that have previously been unreported: AvrLm2 and SIX1 [113, 114], and *Puccinia striiformis* Shr4 and Shr6 [124]. However, in addition to the groupings that the RemEff method has formed between known and candidate effector proteins, the absence of predicted groupings may also offer biological insights. The presence of effectors in 'orphan' clusters might be an indicator of their evolutionary histories involving either high sequence divergence or independent origin.

There were some notable cases where RemEff reported clusters that conflicted with past reports of fungal effector families. The AvrFOM2 level 3 cluster described above, which contained sequences that were previously reported to be ToxA-like [21, 22], failed to group with the ToxA (level 3) cluster despite weak overall sequence similarity (Fig. 4). Similarly, RemEff also failed to group the previously described MAX family proteins [23] into a single component. Our analysis of JackHMMER results did identify a potential relationship between ToxA-like proteins and AvrFOM2 and also recovered many of the relationships identified by our clustering method, including some members of the SIX5-like family but failed to identify similarity between any canonical MAX family members. An important distinction is that progressive search methods (e.g. JackHMMER or HHBlits) search for sequence homologues from a single query sequence. Whereas our clustering method extends the results of these search methods to identify families of reciprocally matching, full length homologues at a higher level of stringency, with the network of clusters representing the structure within and between these families. Additionally, detecting low sequence identity matches to a single query sequence may require manually evaluating matches (as was done with the JackHMMER results in this study), whereas clustering and HMM-HMM comparisons used by RemEff are an automated and unbiased procedure. Where past studies have used progressive searches to focus on a single or reduced set of effectors (e.g. AvrLm6, Zt6, MAX), RemEff mines a larger dataset spanning multiple fungal species which greatly increases its predictive potential.

Some RemEff clusters also contained proteins from non-pathogen species – including saprotrophs – which may initially appear to be in error. However, these results may be valid and worthy of deeper investigation on a case-by-case basis. The presence of false positive predictions of effector-like small secreted proteins (SSPs) in saprotrophs poses a challenge to several effector prediction methods, and it has been reported that all fungal species may have ~40–60 % effector-like SSP predictions within their secretomes: [171]. Nevertheless, some species that have been conventionally regarded as unequivocally saptrotrophic have recently had new evidence supporting rarely observed pathogenic modes [172]. For example, *Neurospora crassa* had long been regarded as the model saprotroph, yet we observed its proteins in ten level 3 clusters that contained known effectors (Table S4). *N. crassa* has been recently isolated as an pathogen of Scots pine [173], and its CAZyme content indicates this may resemble a biotrophic interaction [172]. Additionally, a proportion of fungal effectors are assumed to have broadly cytotoxic functions (albeit with a diverse range of mechanisms). While this study focusses on effectors that target host plant cells, there may be some functional overlap with cytotoxic peptides targeting other types of cells. Saprotrophic yeast species (also present in several RemEff clusters, Table S4) may still require mechanisms to compete with foreign microbes in a resource-scarce environment, e.g. via secretion of yeast killer toxins. It is not unreasonable to allow for the possibility that some plant pathogen effectors may share distant, underlying structural homology to certain saprotroph proteins and cytotoxic peptides from other organisms with vastly different lifestyles, including but not limited to: antimicrobial peptides (AMPs), defensins, conotoxins and venoms. A RemEff cluster matching any of these classes of non-plant pathogen proteins would be considered worthy of experimental follow-up.

CONCLUSION

The RemEff method predicts remote homology relationships between known effectors and candidate effector proteins, allowing for the prediction of distantly related effector 'families' in plant-pathogenic fungi. HMM-HMM clustering based on pattern and/or structural similarities can be useful in organizing known effectors and predicting novel ones, however much more work will be needed to draw links between any structural and functional similarities that have been highlighted – which is far beyond the scope of this current study. We cannot demonstrate that pattern-based/structural homology detected by RemEff is consistently grouping clusters based on common internalisation, pathogenicity or other functions. We assume that there is a reasonable possibility of a common function within each cluster based on their similar sequence or secondary structure patterns, and that by using known effectors as initial 'seeds', most of the resulting clusters will be relevant to pathogenicity and useful in an effector discovery context. We have made RemEff's underlying profile data available for further analysis (https://figshare.com/projects/Effector_protein_remote_homology/87965), which can serve as a useful resource for future plant pathology studies.

We have presented case studies of novel effector family groupings that both demonstrate the utility of this method to enhance effector discovery research, and highlight important similarities and differences between effector family sub-groups. This is illustrated well by sequence logos generated from the AvrLm6-like (Fig. 3) and ribonuclease (Fig. 5) level 2 clusters, which show a combination of overall conservation and motif diversity. We observe cysteine spacing to be a major conserved feature, sometimes in the absence of other defining sequence features. Given the potential overlap in modes of action of some fungal effectors and other non-fungal cytotoxic peptides, we speculate that it may be useful to further explore conservation of cysteine-spacing as a heuristic classification system for some groups of fungal effectors, similar to those that have been established for arthropod venoms [28, 174] and for snail conotoxins [63]. Alternatively, as protein structure prediction methods become more feasible to apply at large scale, it may become possible to predict effector candidates on the basis of structure modelling and molecular docking simulation. However, given the number and diversity of 'effector' proteins and their functions, we anticipate that neither method would be broadly applicable, and maintaining an ensemble of profile HMMs will be preferable for the foreseeable future.

Funding information

An Australian Government Research Training Program Scholarship funded the studentship of DABJ. This study was conducted within the Centre for Crop and Disease Management, a cooperative research centre funded by the Grains and Research Development Corporation (GRDC) and Curtin University (CUR00023).

Acknowledgements

This research was undertaken with the assistance of resources from the Pawsey Supercomputer Centre in Western Australia, and the National Computational Infrastructure (NCI Australia) - an NCRIS enabled capability supported by the Australian Government.

Conflicts of interest

The authors declare that there are no conflicts of interest.

References

1. Thomma B, Nürnberger T, Joosten M. Of pamps and effectors: The blurred PTI-ETI dichotomy. *Plant Cell* 2011;23:4–15.
2. De Wit PJGM, Mehrabi R, Van den Burg HA, Stergiopoulos I. Fungal effector proteins: Past, present and future. *Mol Plant Pathol* 2009;10:735–747.
3. Thrall PH, Barrett LG, Dodds PN, Burdon JJ. Epidemiological and evolutionary outcomes in gene-for-gene and matching allele models. *Front Plant Sci* 2016;6:1084.
4. Vleeshouwers V, Oliver RP. Effectors as tools in disease resistance breeding against biotrophic, hemibiotrophic, and necrotrophic plant pathogens. *Mol Plant Microbe Interact* 2014;27:196–206.
5. Oome S, Van den Ackerveken G. Comparative and functional analysis of the widely occurring family of Nep1-like proteins. *Mol Plant Microbe Interact* 2014;27:1081–1094.
6. Chen H, Kovalchuk A, Keriö S, Asiegbu FO. Distribution and bioinformatic analysis of the cerato-platanin protein family in *Dikarya*. *Mycologia* 2013;105:1479–1488.
7. Olombrada M, Lázaro-Gorines R, López-Rodríguez JC, Martínez-Del-Pozo Á, Oñaderra M, *et al*. Fungal ribotoxins: A review of potential biotechnological applications. *Toxins (Basel)* 2017;9.
8. Bertazzoni S, Williams AH, Jones DA, Syme RA, Tan KC, *et al*. Accessories make the outfit: accessory chromosomes and other dispensable DNA regions in plant-pathogenic fungi. *Mol Plant Microbe Interact* 2018;31:779–788.

9. Hane JK, Rouxel T, Howlett BJ, Kema GH, Goodwin SB, *et al*. A novel mode of chromosomal evolution peculiar to filamentous *Ascomycete fungi*. *Genome Biol* 2011;12:R45.

10. Rouxel T, Grandaubert J, Hane JK, Hoede C, van de Wouw AP, *et al*. Effector diversification within compartments of the Leptosphaeria maculans genome affected by Repeat-Induced Point mutations. *Nat Commun* 2011;2:202.

11. Sánchez-Vallet A, Fouché S, Fudal I, Hartmann FE, Soyer JL, *et al*. The genome biology of effector gene evolution in filamentous plant pathogens. *Annu Rev Phytopathol* 2018;56:21–40.

12. Testa AC, Oliver RP, Hane JK. OcculterCut: A comprehensive survey of AT-rich regions in fungal genomes. *Genome Biol Evol* 2016;8:2044–2064.

13. Torres DE, Oggenfuss U, Croll D, Seidl MF. Genome evolution in fungal plant pathogens: Looking beyond the two-speed genome model. *Fungal Biology Reviews* 2020;34:136–143.

14. Adhikari BN, Hamilton JP, Zerillo MM, Tisserat N, Lévesque CA, *et al*. Comparative genomics reveals insight into virulence strategies of plant pathogenic Oomycetes. *PLoS One* 2013;8:e75072.

15. Fang Y, Coelho MA, Shu H, Schotanus K, Thimmappa BC, *et al*. Long transposon-rich centromeres in an oomycete reveal divergence of centromere features in *Stramenopila-Alveolata-Rhizaria* lineages. *PLoS Genet* 2020;16:e1008646.

16. Anderson RG, Deb D, Fedkenheuer K, McDowell JM. Recent Progress in RXLR Effector Research. *Mol Plant Microbe Interact* 2015;28:1063–1072.

17. Amaro T, Thilliez GJA, Motion GB, Huitema E. A Perspective on CRN proteins in the genomicsage: evolution, classification, delivery and function revisited. *Front Plant Sci* 2017;8:99.

18. de Guillen K, Lorrain C, Tsan P, Petre B, Saveleva N, *et al*. Structural genomics applied to the rust fungus Melampsora laricipopulina reveals two candidate effector proteins adopting cystine knot and NTF2-like protein folds. *Sci Rep* 2019;9:18084.

19. Jones DA, Bertazzoni S, Turo CJ, Syme RA, Hane JK. Bioinformatic prediction of plant–pathogenicity effector proteins of fungi. *Curr Opin Microbiol* 2018;46:43–49.

20. Testa AC, Hane JK, Ellwood SR, Oliver RP. CodingQuarry: Highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. *BMC Genomics* 2015;16:170.

21. Lu S, Gillian Turgeon B, Edwards MC. 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* 2015;81:12–24.

22. Schmidt SM, Lukasiewicz J, Farrer R, Dam P van, Bertoldo C, *et al*. Comparative genomics of *Fusarium oxysporum* f. sp. Melonis reveals the secreted protein recognized by the Fom-2 resistance gene in melon. *New Phytol* 2016;209:307–318.

23. de Guillen K, Ortiz-Vallejo D, Gracy J, Fournier E, Kroj T, *et al*. Structure analysis uncovers a highly diverse but structurally conserved effector family in phytopathogenic fungi. *PLoS Pathog* 2015;11:e1005228.

24. Spanu PD. Cereal immunity against powdery mildews targets RNase-like proteins associated with haustoria (RALPH) effectors evolved from a common ancestral gene. *New Phytol* 2017;213:969–971.

25. Kale SD, Gu B, Capelluto DGS, Dou D, Feldman E, *et al*. External lipid PI3P mediates entry of eukaryotic pathogen effectors into plant and animal host cells. *Cell* 2010;142:284–295.

26. Galagan JE, Selker EU. RIP: The evolutionary cost of genome defense. *Trends Genet* 2004;20:417–423.

27. Kaas Q, Yu R, Jin A-H, Dutertre S, Craik DJ. ConoServer: Updated content, knowledge, and discovery tools in the conopeptide database. *Nucleic Acids Res* 2012;40:D325–D330.

28. Saucedo AL, Flores-Solis D, Vega RCR, la de, Ramírez-Cordero B, *et al*. New tricks of an old pattern structural versatility of scorpion toxins with common cysteine spacing. *J Biol Chem* 2012;287:12321–12330.

29. Ballance GM, Lamari L, Bernier CC. Purification and characterization of a host-selective necrosis toxin from *Pyrenophora tritici-repentis*. *Physiol Mol Plant Path* 1989;35:203–213.

30. Tuori RP, Wolpert TJ, Ciuffetti LM. Purification and immunological characterization of toxic components from cultures of *Pyrenophora tritici*-repentis. *Mol Plant Microbe Interact* 1995;8:41–48.

31. Friesen TL, Stukenbrock EH, Liu Z, Meinhardt S, Ling H, *et al*. Emergence of a new disease as a result of interspecific virulence gene transfer. *Nat Genet* 2006;38:953–956.

32. McDonald MC, Ahren D, Simpfendorfer S, Milgate A, Solomon PS. The discovery of the virulence gene ToxA in the wheat and barley pathogen *Bipolaris sorokiniana*. *Mol Plant Pathol* 2018;19:432–439.

33. McDonald MC, Taranto AP, Hill E, Schwessinger B, Liu Z, *et al*. Transposon-mediated horizontal transfer of the host-specific virulence protein toxa between three fungal wheat pathogens. *mBio* 2019;10.

34. Moolhuijzen P, See PT, Hane JK, Shi G, Liu Z, *et al*. Comparative genomics of the wheat fungal pathogen Pyrenophora triticirepentis reveals chromosomal variations and genome plasticity. *BMC Genomics* 2018;19:279.

35. Friesen TL, Holmes DJ, Bowden RL, Faris JD. ToxA is present in the US bipolaris sorokiniana population and is a snificant virulence factor on wheat harboring Tsn1. *Plant Disease* 2018;102:2446–2452.

36. Ciuffetti LM, Tuori RP, Gaventa JM. A single gene encodes a selective toxin causal to the development of tan spot of wheat. *Plant Cell* 1997;9:135–144.

37. Sarma GN, Manning VA, Ciuffetti LM, Karplus PA. Structure of Ptr ToxA: an RGD-containing host-selective toxin from *Pyrenophora tritici-repentis*. *Plant Cell* 2005;17:3190–3202.

38. Liu Z, Friesen TL, Ling H, Meinhardt SW, Oliver RP, *et al*. The *Tsn1* –ToxA interaction in the wheat–*Stagonospora nodorum* pathosystem parallels that of the wheat–tan spot system. *Genome* 2006;49:1265–1273.

39. Tai YS, Bragg J, Meinhardt SW. Functional characterization of ToxA and molecular identification of its intracellular targeting protein in wheat. *American J of Plant Physiology* 2007;2:76–89.

40. Lu S, Faris JD, Sherwood R, Friesen TL, Edwards MC. A dimeric PR-1-type pathogenesis-related protein interacts with ToxA and potentially mediates ToxA-induced necrosis in sensitive wheat: PR-1 potentially mediates ToxA-induced necrosis. *Mol Plant Pathol* 2014;15:650–663.

41. Manning VA, Chu AL, Steeves JE, Wolpert TJ, Ciuffetti LM. A host-selective toxin of Pyrenophora tritici-repentis, Ptr ToxA, induces photosystem changes and reactive oxygen species accumulation in sensitive wheat. *Mol Plant Microbe Interact* 2009;22:665–676.

42. Meinhardt SW, Cheng W, Kwon CY, Donohue CM, Rasmussen JB. Role of the Arginyl-Glycyl-Aspartic motif in the action of Ptr ToxA produced by Pyrenophora tritici-repentis. *Plant Physiol* 2002;130:1545–1551.

43. Di X, Cao L, Hughes RK, Tintor N, Banfield MJ, *et al*. Structure–function analysis of the *Fusarium oxysporum* Avr2 effector allows uncoupling of its immune-suppressing activity from recognition. *New Phytol* 2017;216:897–914.

44. Nyarko A, Singarapu KK, Figueroa M, Manning VA, Pandelova I, *et al*. Solution NMR structures of pyrenophora Tritici-repentis ToxB and its inactive homolog reveal potential determinants of toxin activity. *J Biol Chem* 2014;289:25946–25956.

45. Andrie RM, Schoch CL, Hedges R, Spatafora JW, Ciuffetti LM. Homologs of ToxB, a host-selective toxin gene from Pyrenophora tritici-repentis, are present in the genome of sister-species *Pyrenophora bromi* and other members of the *Ascomycota*. *Fungal Genet Biol* 2008;45:363–377.

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

47. Fudal I, Ross S, Gout L, Blaise F, Kuhn ML, *et al*. Heterochromatin-like regions as ecological niches for avirulence genes in the

Leptosphaeria maculans genome: map-based cloning of AvrLm6. *Mol Plant Microbe Interact* 2007;20:459–470.

48. Grandaubert J, Lowe RG, Soyer JL, Schoch CL, Van de Wouw AP, *et al*. Transposable element-assisted evolution and adaptation to host plant within the *Leptosphaeria maculans-Leptosphaeria biglobosa* species complex of fungal pathogens. *BMC Genomics* 2014;15:891.
49. Shiller J, Van de Wouw AP, Taranto AP, Bowen JK, Dubois D, *et al*. A large family of AvrLm6-like genes in the apple and pear scab pathogens, *Venturia inaequalis* and *Venturia pirina*. *Front Plant Sci* 2015;6:980.
50. Lacadena J, Álvarez-García E, Carreras-Sangrà N, Herrero-Galán E, Alegre-Cebollada J, *et al*. Fungal ribotoxins: Molecular dissection of a family of natural killers. *FEMS Microbiol Rev* 2007;31:212–237.
51. Glück A, Wool IG. Determination of the 28 S Ribosomal RNA identity element (G4319) for Alpha-sarcin and the relationship of recognition to the selection of the catalytic site. *J Mol Biol* 1996;256:838–848.
52. Martínez-Ruiz A, Kao R, Davies J, Martínez del Pozo Á. Ribotoxins are a more widespread group of proteins within the filamentous fungi than previously believed. *Toxicon* 1999;37:1549–1563.
53. Citores L, Ragucci S, Ferreras JM, Di Maro A, Iglesias R. Ageritin, a ribotoxin from poplar mushroom (*Agrocybe aegerita*) with defensive and antiproliferative activities. *ACS Chem Biol* 2019;14:1319–1327.
54. Landi N, Pacifico S, Ragucci S, Iglesias R, Piccolella S, *et al*. Purification, characterization and cytotoxicity assessment of Ageritin: The first ribotoxin from the basidiomycete mushroom *Agrocybe aegerita*. *Biochimica et Biophysica Acta (BBA) - General Subjects* 2017;5:1113–1121.
55. Herrero-Galán E, Lacadena J, Pozo Á, del M, Boucias DG, *et al*. The insecticidal protein hirsutellin A from the mite fungal pathogen *Hirsutella thompsonii* is a ribotoxin. *Proteins* 2008;72:217–228.
56. Kettles GJ, Bayon C, Sparks CA, Canning G, Kanyuka K, *et al*. Characterization of an antimicrobial and phytotoxic ribonuclease secreted by the fungal wheat pathogen *Zymoseptoria tritici*. *New Phytol* 2018;217:320–331.
57. Pedersen C, Ver Loren van Themaat E, McGuffin LJ, Abbott JC, Burgis TA, *et al*. Structure and evolution of Barley powdery mildew effector candidates. *BMC Genomics* 2012;13:694.
58. Praz CR, Bourras S, Zeng F, Sánchez-Martín J, Menardo F, *et al*. AvrPm2 encodes an RNase-like avirulence effector which is conserved in the two different specialized forms of wheat and rye powdery mildew fungus. *New Phytol* 2017;213:1301–1314.
59. Pennington HG, Jones R, Kwon S, Bonciani G, Thieron H, *et al*. The fungal ribonuclease-like effector protein CSEP0064/BEC1054 represses plant immunity and interferes with degradation of host ribosomal RNA. *PLoS Pathog* 2019;15:e1007620.
60. Deng CH, Plummer KM, Jones DAB, Mesarich CH, Shiller J, *et al*. Comparative analysis of the predicted secretomes of Rosaceae scab pathogens venturia inaequalis and V. Pirina reveals expanded effector families and putative determinants of host range. *BMC Genomics* 2017;18:339.
61. Islam SMA, Kearney CM, Baker E. Classes, databases, and prediction methods of pharmaceutically and commercially important cystine-stabilized peptides. *Toxins (Basel)* 2018;10:251.
62. Linial M, Rappoport N, Ofer D. Overlooked short toxin-like proteins: A shortcut to drug design. *Toxins (Basel)* 2017;9.
63. Robinson SD, Norton RS. Conotoxin gene superfamilies. *Mar Drugs* 2014;12:6058–6101.
64. Asgari E, McHardy AC, Mofrad MRK. Probabilistic variable-length segmentation of protein sequences for discriminative motif discovery (DiMotif) and sequence embedding (ProtVecX). *Sci Rep* 2019;9:3577.
65. Negi SS, Schein CH, Ladics GS, Mirsky H, Chang P, *et al*. Functional classification of protein toxins as a basis for bioinformatic screening. *Sci Rep* 2017;7:13940.
66. Park J, Teichmann SA, Hubbard T, Chothia C. Intermediate sequences increase the detection of homology between sequences11Edited by J. *J Mol Biol* 1997;273:349–354.
67. Chen J, Liu B, Huang D. Protein remote homology detection based on an ensemble learning approach [Research article]. *Biomed Res Int* 2016;2016:5813645.
68. Li S, Chen J, Liu B. Protein remote homology detection based on bidirectional long short-term memory. *BMC Bioinformatics* 2017;18:443.
69. Rangwala H, Karypis G. Profile-based direct kernels for remote homology detection and fold recognition. *Bioinformatics* 2005;21:4239–4247.
70. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, *et al*. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402.
71. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: Hmmer3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res* 2013;41:e121.
72. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, *et al*. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* 2019;20:473.
73. Ma J, Wang Z, Wang Z, Xu J. MRFalign: Protein homology detection through alignment of markov random fields. *PLoS Comput Biol* 2014;10:e1003500.
74. Szklarczyk R, Wanschers BF, Cuypers TD, Esseling JJ, Riemersma M, *et al*. Iterative orthology prediction uncovers new mitochondrial proteins and identifies C12orf62 as the human ortholog of COX14, a protein involved in the assembly of cytochrome oxidase. *Genome Biol* 2012;13:R12.
75. Wagner I, Volkmer M, Sharan M, Villaveces JM, Oswald F, *et al*. morFeus: A web-based program to detect remotely conserved orthologs using symmetrical best hits and orthology network scoring. *BMC Bioinformatics* 2014;15:263.
76. Tatusov RL, Koonin E, Lipman DJ. A genomic perspective on protein families. *Science* 1997;278:631–637.
77. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 2002;30:1575–1584.
78. Petegrosso R, Li Z, Srour MA, Saad Y, Zhang W, *et al*. Scalable remote homology detection and fold recognition in massive protein networks. *Proteins* 2019;87:478–491.
79. Chen L, Li X, Li C, Swoboda GA, Young CA, *et al*. Two distinct *Epichloë* species symbiotic with *Achnatherum inebrians*, drunken horse grass. *Mycologia* 2015;107:863–873.
80. Gao Q, Jin K, Ying SH, Zhang Y, Xiao G, *et al*. Genome sequencing and comparative transcriptomics of the model entomopathogenic fungi *Metarhizium anisopliae* and *M. acridum*. *PLoS Genet* 2011;7:e1001264.
81. Pan J. *Ether Bridge Formation and Chemical Diversification in Loline Alkaloid Biosynthesis*. University of Kentucky, 2014.
82. Schardl CL, Young CA, Hesse U, Amyotte SG, Andreeva K, *et al*. Plant-symbiotic fungi as chemical engineers: Multi-genome analysis of the *Clavicipitaceae* reveals dynamics of alkaloid loci. *PLoS Genet* 2013;9:e1003323.
83. Schardl CL, Young CA, Moore N, Krom N, Dupont PY, *et al*. Chapter ten genomes of plant-associated *Clavicipitaceae. In Fungi* 2014;Vol. 70:291–327.
84. Dang H, Pryor B, Peever T, Lawrence CB. The Alternaria genomes database: A comprehensive resource for a fungal genus comprised of saprophytes, plant pathogens, and allergenic species. *BMC Genomics* 2015;16:239.
85. Chiapello H, Mallet L, Guérin C, Aguileta G, Amselem J, *et al*. Deciphering genome content and evolutionary relationships of isolates from the fungus Magnaporthe oryzae attacking different host plants. *Genome Biol Evol* 2015;7:2896–2912.
86. Babnigg G, Giometti CS. A database of unique protein sequence identifiers for proteome studies. *Proteomics* 2006;6:4514–4522.

87. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, *et al*. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009;25:1422–1423.

88. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;35:1026–1028.

89. Wright ES. DECIPHER: Harnessing local sequence context to improve protein multiple sequence alignment. *BMC Bioinformatics* 2015;16:322.

90. Keul F, Hess M, Goesele M, Hamacher K. PFASUM: A substitution matrix from Pfam structural alignments. *BMC Bioinformatics* 2017;18:293.

91. Urban M, Cuzick A, Seager J, Wood V, Rutherford K, *et al*. PHI-base: The pathogen–host interactions database. *Nucleic Acids Res* 2020;48:D613–D620.

92. Armenteros JJA, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, *et al*. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol* 2019;37:420–423.

93. Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 2004;340:783–795.

94. Petersen TN, Brunak S, Heijne G von, Nielsen H. SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat Methods* 2011;8:785–786.

95. Savojardo C, Martelli PL, Fariselli P, Casadio R. DeepSig: Deep learning improves signal peptide detection in proteins. *Bioinformatics* 2018;34:1690–1696.

96. Käll L, Krogh A, Sonnhammer ELL. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* 2004;338:1027–1036.

97. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J Mol Biol* 2001;305:567–580.

98. Emms DM, Kelly S. OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 2015;16:157.

99. Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using NetworkX. Varoquaux G, Vaught T and Millman J (eds). In: *Proceedings of the 7th Python in Science Conference*. 2008. pp. 11–15.

100. McKinney W. Data structures for statistical computing in python. van der Walt S and Millman J (eds). In: *Proceedings of the 9th Python in Science Conference*. 2010. pp. 56–61.

101. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, *et al*. SCIPY 1.0: Fundamental algorithms for scientific computing in Python. *Nat Methods* 2020;17:261–272.

102. Hauser M, Steinegger M, Söding J. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics* 2016;32:1323–1330.

103. Malliaros FD, Vazirgiannis M. Clustering and community detection in directed networks: A survey. *Physics Reports* 2013;533:95–142.

104. Peixoto TP. *The Graph-Tool Python Library*. 2014.

105. Schneider TD, Stephens RM. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res* 1990;18:6097–6100.

106. Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci* 2018;27:135–145.

107. Tareen A, Kinney JB. Logomaker: Beautiful sequence logos in Python. *Bioinformatics* 2020;36:2272–2274.

108. Hunter JD. Matplotlib: A 2D graphics environment. *Comput Sci Eng* 2007;9:90–95.

109. Mosquera G, Giraldo MC, Khang CH, Coughlan S, Valent B. Interaction transcriptome analysis identifies magnaporthe oryzae bas1-4 as biotrophy-associated secreted proteins in rice blast disease. *Plant Cell* 2009;21:1273–1290.

110. Sharpee W, Oh Y, Yi M, Franck W, Eyre A, *et al*. Identification and characterization of suppressors of plant cell death (SPD) effectors from *Magnaporthe oryzae*. *Mol Plant Pathol* 2017;18:850–863.

111. Lievens B, Houterman PM, Rep M. Effector gene screening allows unambiguous identification of Fusarium oxysporum f. sp. *Lycopersici* races and discrimination from other formae speciales. *FEMS Microbiol Lett* 2009;300:201–215.

112. Ebert MK. *Effector Biology of the Sugar Beet Pathogen Cercospora Beticola*. Wageningen University, 2018.

113. Ghanbarnia K, Fudal I, Larkan NJ, Links MG, Balesdent M-H, *et al*. Rapid identification of the *Leptosphaeria maculans* avirulence gene AvrLm2 using an intraspecific comparative genomics approach. *Mol Plant Pathol* 2015;16:699–709.

114. Rep M, Does HCVD, Meijer M, Wijk R, Houterman PM, *et al*. 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* 2004;53:1373–1383.

115. Ben M'Barek S, Cordewener JHG, Tabib Ghaffary SM, van der Lee TAJ, Liu Z, *et al*. FPLC and liquid-chromatography mass spectrometry identify candidate necrosis-inducing proteins from culture filtrates of the fungal wheat pathogen *Zymoseptoria tritici*. *Fungal Genet Biol* 2015;79:54–62.

116. Laugé R, Joosten M, Van den Ackerveken G, Van den Broek HWJ, De Wit P. The in planta-produced extracellular proteins ECP1 and ECP2 of *Cladosporium fulvum* are virulence factors. *MPMI* 1997;10:725–734.

117. Schmidt SM, Kuhn H, Micali C, Liller C, Kwaaitaal M, *et al*. Interaction of a *Blumeria graminis* f. sp. *hordei* effector candidate with a barley ARF-GAP suggests that host vesicle trafficking is a fungal pathogenicity target: *Blumeria graminis* effector candidates. *Mol Plant Pathol* 2014;15:535–549.

118. Bolton MD, Esse HP, Vossen JH, Jonge RD, Stergiopoulos I, *et al*. The novel *Cladosporium fulvum* lysin motif effector Ecp6 is a virulence factor with orthologues in other fungal species. *Molecular Microbiology* 2008;69:119–136.

119. Marshall R, Kombrink A, Motteram J, Loza-Reyes E, Lucas J, *et al*. 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* 2011;156:756–769.

120. Irieda H, Inoue Y, Mori M, Yamada K, Oshikawa Y, *et al*. Conserved fungal effector suppresses PAMP-triggered immunity by targeting plant immune kinases. *Proc Natl Acad Sci U S A* 2019;116:496–505.

121. Yoshino K, Irieda H, Sugimoto F, Yoshioka H, Okuno T, *et al*. Cell death of *Nicotiana benthamiana* is induced by secreted protein NIS1 of *Colletotrichum orbiculare* and is suppressed by a homologue of CgDN3. *Mol Plant Microbe Interact* 2012;25:625–636.

122. Chuma I, Isobe C, Hotta Y, Ibaragi K, Futamata N, *et al*. Multiple translocation of the AVR-Pita effector gene among chromosomes of the riceblast fungus *Magnaporthe oryzae* and related species. *PLoS Pathog* 2011;7:e1002147.

123. Dai Y, Jia Y, Correll J, Wang X, Wang Y. Diversification and evolution of the avirulence gene AVR-Pita1 in field isolates of *Magnaporthe oryzae*. *Fungal Genet Biol* 2010;47:973–980.

124. Ramachandran SR, Yin C, Kud J, Tanaka K, Mahoney AK, *et al*. Effectors from wheat rust fungi suppress multiple plant defense responses. *Phytopathology* 2017;107:75–83.

125. Mueller AN, Ziemann S, Treitschke S, Aßmann D, Doehlemann G. Compatibility in the Ustilago maydis-maize interaction requires inhibition of host cysteine proteases by the fungal effector Pit2. *PLoS Pathog* 2013;9:e1003177.

126. Schweizer G, Münch K, Mannhaupt G, Schirawski J, Kahmann R, *et al*. Positively selected effector genes and their contribution to virulence in the smut fungus *Sporisorium reilianum*. *Genome Biol Evol* 2018;10:629–645.

127. Bailey BA, Apel-Birkhold PC, Luster DG. Expression of NEP1 by *Fusarium oxysporum* f. Sp erythroxyli after gene replacement and overexpression using polyethylene glycol-mediated transformation. *Phytopathology* 2002;92:833–841.

128. Garcia O, Macedo JAN, Tibúrcio R, Zaparoli G, Rincones J, *et al*. Characterization of necrosis and ethylene-inducing proteins (NEP) in the basidiomycete *Moniliophthora perniciosa*, the causal agent of witches' broom in *Theobroma cacao*. *Mycol Res* 2007;111:443–455.
129. Staats M, van Baarlen P, Schouten A, van Kan JAL, Bakker FT. Positive selection in phytotoxic protein-encoding genes of Botrytis species. *Fungal Genet Biol* 2007;44:52–63.
130. Wang JY, Cai Y, Gou JY, Mao YB, Xu YH, *et al*. VdNEP, an Elicitor from *Verticillium dahliae*, Induces cotton plant wilting. *Appl Environ Microbiol* 2004;70:4989–4995.
131. Balesdent M-H, Fudal I, Ollivier B, Bally P, Grandaubert J, *et al*. The dispensable chromosome of *Leptosphaeria maculans* shelters an effector gene conferring avirulence towards *Brassica rapa*. *New Phytol* 2013;198:887–898.
132. Wouw APV de, Lowe RGT, Elliott CE, Dubois DJ, Howlett BJ. An avirulence gene, AvrLmJ1, from the blackleg fungus, *Leptosphaeria maculans,* confers avirulence to *Brassica juncea* cultivars. *Mol Plant Pathol* 2014;15:523–530.
133. Mesarich CH, Griffiths SA, van der Burgt A, Okmen B, Beenen HG, *et al*. Transcriptome sequencing uncovers the Avr5 avirulence gene of the tomato leaf mold pathogen *Cladosporium fulvum*. *Mol Plant Microbe Interact* 2014;27:846–857.
134. Ridout CJ, Skamnioti P, Porritt O, Sacristan S, Jones JDG, *et al*. Multiple avirulence paralogues in cereal powdery mildew fungi may contribute to parasite fitness and defeat of plant resistance. *Plant Cell* 2006;18:2402–2414.
135. Akum FN, Steinbrenner J, Biedenkopf D, Imani J, Kogel K-H. The *Piriformospora indica* effector PIIN_08944 promotes the mutualistic *Sebacinalean symbiosis*. *Front Plant Sci* 2015;6:906.
136. Wawra S, Fesel P, Widmer H, Timm M, Seibel J, *et al*. The fungal-specific β-glucan-binding lectin FGB1 alters cell-wall composition and suppresses glucan-triggered immunity in plants. *Nat Commun* 2016;7:13188.
137. Chen S, Songkumarn P, Venu RC, Gowda M, Bellizzi M, *et al*. Identification and characterization of In planta–expressed secreted effector proteins from *Magnaporthe oryzae* that induce cell death in rice. *MPMI* 2012;26:191–202.
138. Mentlak TA, Kombrink A, Shinya T, Ryder LS, Otomo I, *et al*. Effector-mediated suppression of chitin-triggered immunity by *Magnaporthe oryzae* is necessary for rice blast disease. *Plant Cell* 2012;24:322–335.
139. Liu T, Song T, Zhang X, Yuan H, Su L, *et al*. Unconventionally secreted effectors of two filamentous pathogens target plant salicylate biosynthesis. *Nat Commun* 2014;5:4686.
140. Poppe S, Dorsheimer L, Happel P, Stukenbrock EH. Rapidly evolving genes are key players in host specialization and virulence of the fungal Wheat pathogen *Zymoseptoria tritici* (*Mycosphaerella graminicola*). *PLoS Pathog* 2015;11:e1005055.
141. Kunjeti SG, Iyer G, Johnson E, Li E, Broglie KE, *et al*. Identification of *Phakopsora pachyrhizi* candidate effectors with virulence activity in a distantly related pathosystem. *Front Plant Sci* 2016;7:269.
142. Liu Z, Zhang Z, Faris JD, Oliver RP, Syme R, *et al*. The Cysteine rich Necrotrophic effector SnTox1 produced by *Stagonospora nodorum* triggers susceptibility of wheat lines harboring Snn1. *PLoS Pathog* 2012;8:e1002467.
143. Houterman PM, Cornelissen BJC, Rep M. Suppression of plant resistance gene-based immunity by a fungal effector. *PLoS Pathog* 2008;4:e1000061.
144. Stephenson SA, Hatfield J, Rusu AG, Maclean DJ, Manners JM. *CgDN3*: an essential pathogenicity gene of *Colletotrichum gloeosporioides* necessary to avert a hypersensitive-like response in the host *Stylosanthes guianensis*. *Mol Plant Microbe Interact* 2000;13:929–941.
145. Vargas WA, Sanz-Martín JM, Rech GE, Armijos-Jaramillo VD, Rivera LP, *et al*. A fungal effector with host nuclear localization and DNA-binding properties is required for maize anthracnose development. *MPMI* 2015;29:83–95.
146. Catanzariti A-M, Dodds PN, Lawrence GJ, Ayliffe MA, Ellis JG. Haustorially expressed secreted proteins from flax rust are highly enriched for avirulence elicitors. *Plant Cell* 2006;18:243–256.
147. Dodds PN, Lawrence GJ, Catanzariti AM, Ayliffe MA, Ellis JG. The Melampsora lini Avrl567 avirulence genes are expressed in *Haustoria* and their products are recognized inside plant cells. *Plant Cell* 2004;16:755–768.
148. Lyu X, Shen C, Fu Y, Xie J, Jiang D, *et al*. A small secreted virulence-related protein is essential for the necrotrophic interactions of *Sclerotinia sclerotiorum* with its host plants. *PLoS Pathog* 2016;12:e1005435.
149. Kirsten S, Navarro-Quezada A, Penselin D, Wenzel C, Matern A, *et al*. Necrosis-inducing proteins of *Rhynchosporium commune*, effectors in quantitative disease resistance. *Mol Plant Microbe Interact* 2012;25:1314–1325.
150. Zhang L, Ni H, Du X, Wang S, Ma X-W, *et al*. The Verticillium-specific protein VdSCP7 localizes to the plant nucleus and modulates immunity to fungal infections. *New Phytol* 2017;215:368–381.
151. Joosten MHAJ, Cozijnsen TJ, De Wit PJGM. Host resistance to a fungal tomato pathogen lost by a single base-pair change in an avirulence gene. *Nature* 1994;367:384–386.
152. Plissonneau C, Daverdin G, Ollivier B, Blaise F, Degrave A, *et al*. A game of hide and seek between avirulence genes AvrLm4-7 and AvrLm3 in *Leptosphaeria maculans*. *New Phytol* 2016;209:1613–1624.
153. Rice P, Longden I, Bleasby A. EMBOSS: the european molecular Biology open software suite. *Trends Genet* 2000;16:276–277.
154. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* 2013;30:772–780.
155. Stergiopoulos I, Kourmpetis YA, Slot JC, Bakker FT, De Wit P, *et al*. *In silico* characterization and molecular evolutionary analysis of a novel superfamily of fungal effector proteins. *Mol Biol Evol* 2012;29:3371–3384.
156. Cao L, Blekemolen MC, Tintor N, Cornelissen BJC, Takken FLW. The *Fusarium oxysporum* Avr2-Six5 effector pair alters plasmodesmatal exclusion selectivity to facilitate Cell-to-Cell movement of Avr2. *Mol Plant* 2018;11:691–705.
157. Petit-Houdenot Y, Degrave A, Meyer M, Blaise F, Ollivier B, *et al*. A two genes – for – one gene interaction between *Leptosphaeria maculans* and *Brassica napus*. *New Phytol* 2019;223:397–411.
158. Wang C, Liu Y, Liu L, Wang Y, Yan J, *et al*. The biotrophy-associated secreted protein 4 (BAS4) participates in the transition of *Magnaporthe oryzae* from the biotrophic to the necrotrophic phase. *Saudi J Biol Sci* 2019;26:795–807.
159. Khang CH, Berruyer R, Giraldo MC, Kankanala P, Park S-Y, *et al*. Translocation of *Magnaporthe oryzae* effectors into rice cells and their subsequent cell-to-cell movement. *Plant Cell* 2010;22:1388–1403.
160. Manning VA, Hamilton SM, Karplus PA, Ciuffetti LM. The Arg-Gly-Asp–containing, solvent-exposed loop of Ptr ToxA is required for internalization. *Mol Plant Microbe Interact* 2008;21:315–325.
161. Figueroa M, Manning VA, Pandelova I, Ciuffetti LM. Persistence of the host-selective toxin PTR ToxB in the apoplast. *Mol Plant Microbe Interact* 2015;28:1082–1090.
162. Giraldo MC, Dagdas YF, Gupta YK, Mentlak TA, Yi M, *et al*. Two distinct secretion systems facilitate tissue invasion by the rice blast fungus *Magnaporthe oryzae*. *Nat Commun* 2013;4:1996.
163. Owji H, Nezafat N, Negahdaripour M, Hajiebrahimi A, Ghasemi Y. A comprehensive review of signal peptides: Structure, roles, and applications. *Eur J Cell Biol* 2018;97:422–441.
164. Duffy J, Patham B, Mensa-Wilmot K. Discovery of functional motifs in h-regions of trypanosome signal sequences. *Biochemical Journal* 2010;426:135–145.
165. Pérez-Cañadillas JM, Santoro J, Campos-Olivas R, Lacadena J, Martínez del Pozo A, *et al*. The highly refined solution structure of the cytotoxic ribonuclease alpha-sarcin reveals the structural

requirements for substrate recognition and ribonucleolytic activity. *J Mol Biol* 2000;299:1061–1073.

166. Viegas A, Herrero-Galán E, Oñaderra M, Macedo AL, Bruix M. Solution structure of hirsutellin A-new insights into the active site and interacting interfaces of ribotoxins. *FEBS J* 2009;276:2381–2390.

167. Yang X, Moffat K. Insights into specificity of cleavage and mechanism of cell entry from the crystal structure of the highly specific *Aspergillus ribotoxin*, restrictocin. *Structure* 1996;4:837–852.

168. Bourras S, McNally KE, Ben-David R, Parlange F, Roffler S, *et al*. Multiple *Avirulence loci* and allele-specific effector recognition control the Pm3 race-specific resistance of wheat to powdery mildew. *Plant Cell* 2015;27:2991–3012.

169. McNally KE, Menardo F, Lüthi L, Praz CR, Müller MC, *et al*. Distinct domains of the AVRPM3A2/F2 avirulence protein from wheat powdery mildew are involved in immune receptor recognition and putative effector function. *New Phytol* 2018;218:681–695.

170. Paulus JK, van der Hoorn RAL. Tricked or trapped-Two decoy mechanisms in host-pathogen interactions. *PLoS Pathog* 2018;14:e1006761.

171. Feldman D, Yarden O, Hadar Y. Seeking the roles for fungal small-secreted proteins in affecting saprophytic lifestyles. *Front Microbiol* 2020;11:455.

172. Hane JK, Paxman J, Jones DAB, Oliver RP, de Wit P. "CATAStrophy", a genome-informed trophic classification of *Filamentous* plant pathogens – how many different types of *Filamentous* plant pathogens are there? *Front Microbiol* 2020;10:3088.

173. Kuo H-C, Hui S, Choi J, Asiegbu FO, Valkonen JPT, *et al*. Secret lifestyles of *Neurospora crassa*. *Sci Rep* 2015;4:5135.

174. Fry BG, Roelants K, Champagne DE, Scheib H, Tyndall JDA, *et al*. The toxicogenomic multiverse: Convergent recruitment of proteins into animal venoms. *Annu Rev Genomics Hum Genet* 2009;10:483–511.