Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predector: an automated and combinative method for the predictive ranking of candidate effector proteins of fungal plant-pathogens

Darcy A. B. Jones
Curtin University
Lina Rozano
Curtin University
Johannes Debler
Curtin University
Ricardo L. Mancera
Curtin University
Paula Moolhuijzen
Curtin University
James K. Hane (James.Hane@curtin.edu.au)
Curtin University

Research Article

Keywords: Effector, virulence factor, Fungi, plant pathogen, host-pathogen interactions

Posted Date: April 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-379941/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published at Scientific Reports on October 5th, 2021. See the published version at https://doi.org/10.1038/s41598-021-99363-0.

Abstract

‘Effectors’ are a broad class of cytotoxic or virulence-promoting molecules that are released from plant-pathogen cells to cause disease in their host. Fungal effectors are a core research area for improving host disease resistance; however, because they generally lack common distinguishing features or obvious sequence similarity, discovery of effectors remains a major challenge. This study presents a novel tool and pipeline for effector prediction - Predector - which interfaces with multiple software tools and methods, aggregates disparate features that are relevant to fungal effector proteins, and ranks effector candidate proteins using a pairwise learning to rank approach. Predector outperformed alternative effector prediction methods that were applied to a curated set of confirmed effectors derived from multiple species. We present Predector as a useful tool for the prediction and ranking of effector candidates, which aggregates and reports additional supporting information relevant to effector and secretome prediction in a simple, efficient, and reproducible manner. Predector is available from https://github.com/ccdmb/predector and associated data from https://github.com/ccdmb/predector-data.

Introduction

‘Effectors’ are a broad class of cytotoxic or virulence-promoting molecules that are released from pathogen cells to cause disease in their host. Fungal effectors are a core research area for improving host disease resistance; however, because they generally lack common features or obvious sequence similarity, discovery of effectors is non-trivial[1–3]. Secreted effector proteins of plant pathogens have been studied more comprehensively in the Oomycetes (a separate lineage of filamentous microbes), in which *in silico* identification of effectors is more feasible compared to fungi as they exhibit highly conserved sequence motifs (e.g. RXLR, LXLFLAK)[4,5]. Effector prediction in fungal genomes may be more challenging as they are highly plastic, commonly exhibiting accelerated mutation rates, fungal-specific genome-wide mutagenesis mechanisms e.g. repeat-induced point mutation (RIP)[6,7], as well as increased rates of chromosome structure rearrangement[8,9] and lateral gene transfer[10]. Consequently, fungal effectors are highly diverse in sequence and function, and much effector candidate discovery is performed using experimental techniques such as phenotype association and comparative genomics[11–14], transcriptomics[15–17], proteomics[18,19] and GWAS[20,21]. There are, however, some protein characteristics - i.e. structural features (e.g. functional domains), signal peptides, amino-acid frequencies - that can be used as an alternative to simple homology searches. Several methods using these characteristics have been developed to prioritise effector candidates for experimental validation[2].

In-silico effector prediction of small-secreted proteins (SSPs) has typically involved ad hoc hard-set criteria such as a signal peptide, no transmembrane domains outside the signal peptide, small overall size (often < 300AA), and a high number of cysteine amino-acids. These thresholds were based on the properties of early discovered effectors; however, numerous known effectors do not conform to this profile (Supplementary Table 1). The use of simple hard filters risks excluding these proteins from

candidacy. Signal peptide prediction is the most common *in-silico* technique used to refine effector candidates from proteomes[22], with SignalP the most common prediction tool[23–25] although other tools are frequently used in combination[26,27], and different tools can perform better or worse with different protein groups or organisms[22]. Subcellular localisation prediction tools such as TargetP[23] or DeepLoc[28] are also frequently used to predict the location of proteins. Their reliability for predicting protein secretion is questionable[22], but proteins predicted to be localised in organelles might reasonably be excluded. Because most effectors are expected to be free in the extracellular space or host cells, transmembrane domains (TM) are also an important feature for excluding candidates, commonly predicted using TMHMM[29] or Phobius[26].

Recently developed machine learning tools tailored to predicting effector-like properties have presented new opportunities for improving effector prediction pipelines. EffectorP[30,31] and FunEffector-Pred[32] use amino acid frequencies, molecular weight, charge, AA k-mers, and other protein characteristics to predict effector-like proteins directly. In combination with secretion prediction, tools like EffectorP and FunEffector-Pred may be a more robust alternative to simple hard filters. LOCALIZER[33] and ApoplastP[31], which predict host subcellular or apoplastic localisations, are useful for evaluating candidates but are not necessarily predictive of effector candidature themselves.

While many fungal effectors have previously not had similar sequences in public databases, a small but increasing number of families based on conserved domains or structure are becoming known[2], including the ToxA-like[34], MAX[35], RALPH[36], and RXLR-like[37] families. Homology to effector-like conserved domains (i.e. selected Pfam domains) or effector-like sequences within databases such as the Plant-Host Interactions database (PHI-base)[38] and the Database of Fungal Virulence Factors (DFVF)[39], are growing in their relevance. Secondary and tertiary structural modelling and similarity searches against known effectors are not commonly used for high-throughput effector discovery, but this could yet become an important component of future effector prediction pipelines[2].

Current effector prediction pipelines face two major challenges: 1) the necessity of reducing 10–20 thousand proteins per genome down to a set of effector candidates that is both reliable and within a number that is feasible for experimental validation, and 2) the amalgamation of outputs from a large and diverse range of bioinformatics tools and methods, for both prediction and informative purposes. Fungal genome datasets typically contain thousands of secreted proteins, of which hundreds of SSPs may be predicted by standard methods[2]. Further filtering or ranking based on supporting data from GWAS, RNAseq, positive selection, or comparative genomics can still generate hundreds of candidates[40–43]. The prioritisation of effector candidates based on simple biochemical properties is, therefore, still relevant to effector prediction. Furthermore, there is little consensus on how to combine multiple analyses[22], and the common use of multiple successive hard filters risks increasing the error with each step, causing good candidates to be excluded.

Rank-based methods are a simple way to avoid exclusion of candidates lacking clearly discriminative features, via assigning weighted scores to features that are presumed to be important in determining effector-likelihood, and summing these into a single score that is used to rank candidates[43]. However, these simple combinations of manually assigned feature weights may still fail to place proteins with uncommon characteristics near the top of the list. More sophisticated ranking decisions may come from a group of machine learning techniques called “learn to rank”. Rather than offering a binary classification (i.e. effector or non-effector), these methods attempt to order elements optimally so that relevant elements are nearer the beginning of the list. Although these algorithms are most often employed in search engine and e-commerce websites, they have been used successfully to combine diverse sources of information and rank protein structure predictions[44], remote homology predictions[45], gene ontology term assignments[46], and predicting protein-phenotype associations in human disease[47].

In this study, we present a novel tool and pipeline for effector prediction - Predector - which interfaces with multiple software tools and methods, aggregates disparate features that are relevant to fungal effector proteins, and ranks effector candidate proteins using a pairwise learning to rank approach. Predector simplifies effector prediction workflows by providing simplified software dependency installation, a standardised pipeline that can be run efficiently on both commodity hardware and supercomputers, and user friendly tabular formatted results. In this study, we compare the performance of Predector against a typical effector prediction method (i.e. signal peptide prediction, transmembrane domain prediction, and EffectorP), on a curated set of confirmed effectors derived from multiple species. While the small number of currently known effectors and relatively loose definition of the group precludes the possibility of perfectly precise effector prediction tools, we present Predector as a tool enabling useful effector candidate ranks alongside supporting information for effector and secretome prediction in a simple, efficient, and reproducible manner.

Results

To develop and evaluate the predector pipeline, a dataset of unprocessed fungal proteins was collected and split into train and test datasets (Supplementary Table 2). The datasets included redundancy reduced proteins of known fungal effectors (train: 125, test: 28), fungal proteins in the SwissProt database annotated as secreted (train: 256, test: 64) and non-secreted (train: 8676, test: 2169), and the whole proteomes from 10 well studied fungal genomes (train: 52224, test: 13056). The predector pipeline runs numerous tools related to effector and secretome prediction (Table 3). Benchmarking those tools against the set of confirmed effector proteins in the train dataset, it was observed that the secretion prediction tools were frequently correct with a small number of exceptions (Fig. 1). Signal peptide prediction recall in the training dataset of known effectors ranged from 84% (DeepSig) to 92% (TargetP 2). SignalP 3, 4, 5, and Phobius generally predicted about 90% of effectors to have signal peptides (Fig. 1). Transmembrane (TM) predictors were, as expected, generally not able to predict TM domains in confirmed effectors, with the few single TM predictions by TMHMM or Phobius likely to be mis-predictions within N-terminal signal peptides. In the case of TMHMM, all effectors with at least one TM

domain had more than ten AAs predicted to be TM associated in the first 60 residues by TMHMM (Supplementary File 1:39). Effector prediction tools (EffectorP 1 and 2) were also able to predict most, but not all, of the confirmed effector set. EffectorP correctly predicted 85.6% and 76.8% of effectors in the training dataset for versions 1 and 2, respectively. Evaluation of protein features that might allow for distinction between the different protein classes considered in this study (effectors, effector homologues, secreted proteins, non-secreted proteins, and unlabelled proteomes) identified twelve features that could be used effectively. These included: the proportion of cysteines, small, non-polar, charged, acidic, and basic amino acids; ApoplastP prediction; DeepLoc extracellular or membrane predicted localisations; molecular weight; EffectorP scores, and signal peptide raw scores (see Supplementary File 1:3–40).

Table 3
Bioinformatics tools and methods integrated into the Predector pipeline. *Non-default parameters are indicated where applicable.

Software	Description	References
A) Localisation		
SignalP v3.0, 4.1g, 5.0b	Extracellular secretion via signal peptide. Both NN and HMM methods are run for v3.0. Eukaryotic types specified.	(José Juan Almagro Armenteros et al., 2019; Dyrløv Bendtsen et al., 2004; Petersen et al., 2011)
Deepsig commit 69e01cb	Extracellular secretion. *-k euk	(Savojardo et al., 2018)
Phobius 1.01	Extracellular secretion	(Käll et al., 2004)
LOCALIZER v1.0.4	Host sub-cellular localisation. Using predicted mature proteins from SignalP 5.0b. *-e -M	(Sperschneider et al., 2017)
ApoplastP v1.0.1	Apoplast-specific localisation	(Sperschneider, Dodds, Singh, et al., 2018)
DeepLoc v1.0	Sub-cellular localisation	(Almagro Armenteros et al., 2017)
TargetP v2.0	Sub-cellular localisation. *-org non-pl	(Jose Juan Almagro Armenteros et al., 2019)
TMHMM v2.0c	Membrane localisation via transmembrane domains. *-d	(Krogh et al., 2001)
B) Effector-like properties		
EffectorP v1.0, 2.0	Probabilistic prediction of effector likelihood	(Sperschneider, Dodds, Gardiner, et al., 2018; Sperschneider et al., 2016)
EMBOSS: pepstats v6.5.7	Amino acid properties and frequencies	(Rice et al., 2000)
C) Functional annotation		
HMMER (vs dbCAN v8) v3.2.1	Used to search dbcan	(Eddy, 2011; Zhang et al., 2018)
MMSeqs2 v10-6d92c (vs PHIBase v4.9)	Used to search phibase. *–max-seqs 300 -e 0.01 -s 7 –num-iterations 3 -a	(Steinegger & Söding, 2017; Urban et al., 2020)
MMSeqs2 v10-6d92c (vs known effectors in Supplementary Table 2)	*–max-seqs 300 -e 0.01 -s 7 –num-iterations 3 -a	(Steinegger & Söding, 2017; Urban et al., 2020)
PfamScan (vs Pfam v33.1)	With active site prediction. *-as	(Finn et al., 2014)

To incorporate information from the selected features related to effector and secretion prediction, a pairwise learning to rank model was trained. The mean cross validated normalised discount cumulative gain (NDCG) in the top 500 ranked predictions (NDCG@500) for the hyper-parameter optimised model was 0.925942 with standard deviation 0.009421, indicating high performance and little effect of substructure within the dataset. The mean NDCG@500 for the train sets within the cross validation was 0.886542 (0.015099), indicating that the model was not overfitting.

Benchmarked against a test set of confirmed effectors (Fig. 2), the Predector model consistently gave higher scores to effector proteins, and also to homologues of confirmed effectors (those on which the model was not trained). Secreted proteins from SwissProt tended to have intermediate scores centred around 0. Non-secreted and the unlabelled effectors were heavily skewed towards more negative scores, with a long tail that included some proteins with high scores (which in the case of proteomes was expected as this dataset was unlabelled). The test and train sets showed similar distributions of scores, though there tended to be slightly lower scores for known effectors in the test set.

The main features used for sorting effectors from non-effectors in the Predector model were TargetP secretion prediction, SignalP 3-HMM S-scores, SignalP4 D-scores, DeepLoc extracellular and membrane predictions, and EffectorP 1 and 2. TargetP secretion was overwhelmingly the most important feature according to the gain metric (the average increase in predictive score when the feature is used), which was consistent with the observation that it was the most sensitive of the signal peptide prediction methods for effectors (Fig. 1). The most commonly used predictors were EffectorP 2 pseudo-probabilities, molecular weight, and the proportions of cysteines, basic AAs, non-polar AAs and tiny AAs. Feature importance and boosted trees indicated overall that the Predector model first coarsely sorts proteins into the predicted secretome and non-secreted proteins, then proceeds to separate proteins with effector-like properties from the remainder of the secretome using more decision nodes each with smaller overall gain (Supplementary File 1:43).

Predector separated some proteins predicted to be secreted (i.e. with a signal peptide and fewer than two TM domains), from those that are not (Fig. 3). Most “non-secreted” proteins have a score < 0, while a tri-modal distribution of “secreted proteins” was observed, which spanned the full range of scores and roughly coincided with the distributions of effectors/homologues, SwissProt secreted and the non-secreted/proteome datasets (Fig. 2). This contrasted with EffectorP predictions (which was trained and is intended to be used on secretomes only), which gave poor separation of non-secreted and secreted proteins. EffectorP 1 showed a high bias to predicting proteins as either 0 or 1, indicating that it may be unsuited for ranking and should only be used as a decision classifier with a score threshold of 0.5. EffectorP 2 showed a more continuous separation of known effectors, and was moderately correlated with Predector scores for secreted proteins.

Predector consistently outperformed EffectorP 1 and 2 (restricted to the predicted secretome, as per intended usage) in classification recall and Matthews correlation coefficient, and in metrics assessing the ranked order of effector candidates (Table 1, Supplementary Table 4). While EffectorP was not optimised

for effector candidate ranking or intended to be used this way, we note that its probability score can often be mis-used for this purpose. Conversely, although Predector was not intended to be used for effector classification, we also compared its predictive performance with EffectorP 1 and 2 on the secreted subset, and on the full dataset using the joint estimator of secretion and EffectorP score > 0.5. For the purpose of this comparison, a minimum Predector score of 0 was selected as a classification threshold based on the observation that the model assigns positive scores to effector associated branches in the trees (and negative scores to non-effector associated branches). EffectorP 1 and 2 performed identically in terms of effector classification on our test dataset, and gave highly similar results on the training dataset (Supplementary Table 4, Supplementary File 1:50), although fewer false positives were reported by EffectorP 2. Predector correctly predicted all but two effectors in the full test set, and all but one in the secreted test subset. In contrast, EffectorP 1 and 2 both mis-classified six effectors in the secreted subset, and two known effectors in the test dataset were not predicted to be secreted thus would have been excluded from prediction by an EffectorP pipeline. Predector also correctly predicted a confirmed effector (AvrSr35) that was not predicted to be secreted as an effector. Although Predector, not being optimised for classification, had a higher false positive rate than EffectorP 1 and 2, it compared favourably for the MCC metric which is considered more reliable for unbalanced datasets[48]. It is worth noting that in this study secretion prediction incorporates multiple methods, whereas many studies rely on a single prediction tool, thus the proportion of potentially missed effector candidates may be higher than we report here.

Table 1
Effector prediction and ranking statistics for Predector and EffectorP on the test dataset.

		Full test dataset			Secreted test subset		
		EP1 & Sec	EP2 & Sec	Predector	EP1	EP2	Predector
RANK	Coverage error*	-	-	8054	2275	1593	1115
	NDCG@50	-	-	0.640	0.615	0.629	0.652
	NDCG@500	-	-	0.928	0.916	0.926	0.933
	NDCG	-	-	0.447	0.365	0.402	0.448
CLASSIFICATION	TP	20	20	26	20	20	25
	TN	14450	14609	14317	1410	1569	1323
	FP	839	680	972	839	680	926
	FN	8	8	2	6	6	1
	Precision	0.023	0.028	0.026	0.023	0.028	0.026
	Recall	0.714	0.714	0.928	0.769	0.769	0.961
	Accuracy	0.944	0.955	0.936	0.628	0.698	0.592
	Balanced accuracy	0.829	0.834	0.932	0.698	0.733	0.774
	MCC	0.122	0.137	0.149	0.086	0.107	0.118

EP1 = EffectorP v1, EP2 = EffectorP v2, Sec = Secreted, TP = true positives, TN = true negatives, FP – false positives, FN = false negatives (for classification), NDGC = measure of how often effectors are placed ahead of unlabelled samples in the list sorted by score, penalising incorrect orderings more highly near the top of the list, NDGC@N = NDGC for only the top N items in the sorted list

* index of the last known effector in the test dataset

For a set of 14 fungal proteomes retained separately for evaluation (Table 2, Supplementary Table 4), Predector predicted on average 7.2% of proteins to have a score > 0, with an average of 6.4 effector homologues in the 50 highest scoring predictions. Predector processed whole fungal proteome datasets with an average rate of 1,814.67 proteins per hour on four CPUs, and 3,922.15 proteins per hour on 16

CPUs. DeepLoc was overwhelmingly the longest running task (~ 75 mins for 5000 proteins), with most of the 16 CPUs to idle while waiting for DeepLoc to finish.

Table 2

Predector effector predictions and run time evaluation for results on proteomes held out of the training set. Runs with 4 CPUs were performed on a cloud instance running ubuntu 20.04 (4 AMD EPYC vcpus, 16Gb RAM), runs with 16 CPUs are performed on a partially occupied single HPC node (16 Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz, 48 Gb RAM). Both were carried out with default configuration (maximum within task parallelisation is 4 CPUs). The number of protein sequence similarity matches to known effectors and matches to Pfam domains with putative virulence functions are noted for the top 50 candidates.

Organism	CPUs	# proteins	Run time (h:m:s)	# > 0[a]	#homologues in top 50	#Pfam domain in top 50
Blumeria graminis f. sp. *hordei* RACE1	4	5317	3:12:54	366	12	0
Blumeria graminis f. sp. *tritici* 96224	4	8347	4:33:55	696	20	0
Dothistroma septosporum NZE10	4	12415	6:53:22	610	2	4
Fusarium oxysporum f. sp. *lycopersici* MN25	4	24733	12:41:24	1313	8	8
Fusarium oxysporum f. sp. *melonis* 26406	4	26719	13:47:00	1464	7	8
Leptosphaeria maculans G12-14	4	12678	6:56:57	821	5	10
Leptosphaeria maculans NzT4	4	14026	7:26:47	868	9	9
Melampsora larici-populina 98AG31	4	16372	8:39:07	1282	1	0
Puccinia graminis f. sp. *tritici* 21 – 0	4	37843	20:08:17	4169	6	0
Pyrenophora teres f. sp. *maculata* SG1	4	10571	6:50:25	981	3	2
Pyrenophora tritici-repentis M4	4	13795	7:49:00	850	6	8
Zymoseptoria tritici 1A5	16	12072	2:55:22	970	2	1
Zymoseptoria tritici 1E4	16	12023	3:33:47	981	5	1
Zymoseptoria tritici 3D1	16	11991	2:48:49	971	4	1

[a] Number of proteins with a Predector score greater than 0.

Discussion

The Predector pipeline unites, for the first time, numerous computational tasks commonly involved in effector and secretion prediction to determine a ranked set of candidate effectors from unprocessed (immature) proteins, simplifying complex data gathering steps. The effector ranking model run as part of Predector provides additional benefits over the standalone use of its composite tools, in combining their individual strengths while being less prone to their weaknesses. It was observed that while the most recently updated effector prediction tool available – EffectorP 2[30] - performed well as a very specific classifier, it still missed several confirmed effectors. The preliminary step of secretion prediction can also be error prone, and the combined false positives from both effector and secretion prediction methods, coupled with their common implementation as hard filters, may result in many genuine candidate effectors being discarded. For this reason, we propose that ranking and clustering methods should be preferred over hard filters for prioritising effector candidates.

In terms of effector candidate ranking, EffectorP 2 performed reasonably well for ordering confirmed effectors based on probability score, but was not designed to be used in this way. Predector maintained higher recall with higher scores (Supplementary File 1:46 & 47, Table 1) and achieved comparable or better precision than EffectorP 2 alone for higher effector scores. Thus, while Predector is not intended to be used as a classifier, we demonstrate its utility as a highly sensitive method for combined secretion and effector prediction, and suggest a decision threshold (score) of 0 for summarisation purposes alongside standard EffectorP and secretion classifiers (which can be obtained from Predector output). However, the appropriate threshold may change with future versions. Although the recall scores for Predector were very high, Predector also predicted 292 more false positives in the test dataset than the commonly used method of combining a predicted secretion hard filter with EffectorP 2 (Table 1). We argue that recall should be prioritised for effector prediction, as the unlabelled proteome datasets used here may contain genuine novel effectors, and the focus of Predector on ranking rather than classification mitigates some of the issues associated with lower precision. Encouragingly, we observed that Predector was capable of giving positive scores to known effectors which were not predicted to have a signal peptide (in both the train and test datasets) and thus would have failed to be predicted by alternate methods with a secretion prediction hard filter.

The predictive rankings provided by Predector are complemented with additional information that can be used to manually evaluate groups of effector candidates, and represents a comprehensive summary of various predicted types of proteins within a fungal proteome dataset, including candidate pathogenicity effectors, effector homologues, predicted secreted proteins, and carbohydrate-active enzymes (CAZymes)[49]. Predector reports the results of database searches against PHI-base, a curated set of known fungal effectors, Pfam domains, and dbCAN HMMs. We recommend that users examine the functionally annotated candidates closely, particularly with respect to homologues of confirmed effectors, prior to consideration of candidates ranked by Predector scores. Similarly, supplementation with experimental evidence or information derived from external tools and pipelines will further improve the utility of the Predector outputs, e.g. selection profiles derived from pan-genome comparisons[43,50],

presence-absence profiles in comparative genomics, genome wide association studies, differential gene expression, or pathogenicity-relevant information relating to the genomic landscape: the distance to a DNA repeats, telomeres or distal regions of assembled sequences[8,51]; or codon adaptation. By selecting indicators of general effector properties or molecular interactions of interest, and sorting these lists first by those functionally-guided features and then by Predector score(s), users gain a rich and clear guide for prioritising candidates before proceeding to more resource-expensive experiments (e.g. cloning or structure modelling).

Among known effectors there is considerable diversity of their molecular roles and functions. The modern plant pathology community has yet to come to firm agreement on the broad definition of an effector beyond the gene-for-gene and inverse gene-for-gene models, or to refine a broader definition with effector sub-types. Effectors may promote virulence through directly targeting and disrupting host cell biological processes, including ribogenesis, photosynthesis or mitochondrial activity. In contrast various extracellular chitin-binding proteins have also long been described as effectors, yet promote virulence through passively protecting the pathogen cell from host PAMP and DAMP recognition. CAZymes are not typically considered to act as effectors, yet there are several examples of secreted CAZymes that are reported as virulence factors or may be recognised by host major resistance genes[38]. Furthermore, the focus of many effector prediction methods (including Predector) on biochemical or functional aspects of effector proteins also neglects the crucial contribution of host R- and S-proteins in gene-for-gene interactions, which must be determined experimentally. An inclusive predictive model spanning diverse effector types may not offer a reliable pathway to rapid effector identification, rather they are likely to focus on general biochemical properties unrelated to necrotrophic or avirulence activities, e.g. that would enable the majority to interact with membranes and translocate into a host cell or to function in the apoplast. We present Predector as a reasonable compromise between functional diversity and common purpose, accounting for this inherent diversity through incorporation of multiple predictive methods. Additionally, with rapidly decreasing costs of genome sequencing and improvements to the automation of genome analysis and gene feature annotation, the availability and utility of fungal pathogen genomes is steadily increasing[52]. There is a growing need for tools which will minimise the effects of poor data quality control and ensure reproducibility and comparability across multiple genome resources. The Predector pipeline is an important time-saving tool which applies a standardised and reproducible set of tests for effector prediction

Methods

Pipeline implementation

The Predector pipeline runs a range of commonly used effector and secretome prediction bioinformatics tools for complete predicted proteome, accepted as input in FASTA formatted files (Table 3), and combines all raw and summarised outputs into newline-delimited JSON, tab-delimited text and GFF3 formats. The pipeline is implemented in Nextflow (version >20)[53], and a conda environment and Docker container are available for easy installation of dependencies, with scripts to integrate user-downloaded

proprietary software into these environments. Predector is available from https://github.com/ccdmb/predector.

Datasets

The training and evaluation datasets consisted of: confirmed fungal effectors, fungal proteins with confirmed subcellular localisation, and an 'unlabelled' fungal protein set derived from whole proteomes of well-annotated, model fungal species. The experimentally-confirmed effector protein dataset was curated from literature, PHI-base[38], and EffectorP[30,54] training datasets (Supplementary Table 2). Effector homologues were also identified from literature (Supplementary Table 2) and by searching the UniRef-90 fungal proteins (UniProtKB query: taxonomy:"Fungi [4751]" AND identity:0.9, UniProt version 2020_01, Downloaded 2020-06-01) using MMSeqs2 version 11-e1a1c[55] requiring a minimum reciprocal coverage of 70% and a maximum e-value of 10^{-5} (-e 0.00001 --start-sens 3 -s 7.0 --sens-steps 3 --cov-mode 0 -c 0.7). Fungal proteins with experimentally annotated subcellular localisation were downloaded from UniProtKB/SwissProt (version 2020_01, Downloaded 2020-06-01), and were labelled "secreted" (non-transmembrane) or "non-secreted" (membrane associated, endoplasmic reticulum localised, golgi localised, and Glycosylphosphatidylinositol (GPI) anchored). UniProtKB download queries are provided in Supplementary Table 2. The 'unlabelled' whole proteome dataset was derived from well-studied pathogens, with at least one representative chosen from a range of trophic phenotypes [56]: monomertrophs/biotrophs: *Blumeria graminis* f. sp. *hordei*

[57], *Blumeria graminis* f. sp. *tritici* [58], *Melampsora lini* [59], *Melampsora larici-populina* [60], *Puccinia graminis* f. sp. *tritici* [61]; polymertrophs/necrotrophs - *Parastagonospora nodorum* [43,62], *Pyrenophora tritici-repentis* [63], *Pyrenophora teres* f. *teres* [64], and *Pyrenophora teres* f. *maculata* [64]; mesotrophs/hemibiotrophs - *Leptosphaeria maculans* [41], *Zymoseptoria tritici* [65,66], *Passalora fulva*[67], *Dothistroma septosporum*[67]; wilts/vasculartrophs - *Fusarium oxysporum* f. sp. *lycopersici* [68,69], *Fusarium oxysporum f. sp. melonis* [70]; and saprotroph (or opportunistic monomertroph/biotroph) *Neurospora crassa* [71] (Supplementary Table 2). Fourteen of the 24 proteomes above were retained as a separate dataset for final evaluation (Supplementary Table 2). The remainder of the datasets were combined, and redundant sequences were removed to prevent the undue influence of conserved or well studied sequences with multiple records. Redundancy was reduced by clustering proteins with MMSeqs2 version 11-e1a1c [55] requiring a minimum reciprocal coverage of 70% and minimum sequence identity of 30% (--min-seq-id 0.3 --cov-mode 0 -c 0.7 --cluster-mode 0). A single sequence was chosen to represent a set of clustered, redundant sequences, which was prioritised based on supporting information (in order of preference): known effector, SwissProt secreted, SwissProt non-secreted, proteome/effector homologue, longest member of cluster. Clusters that corresponded to the known effectors from the EffectorP 2 [30] training and test data sets were automatically assigned to training and test data sets in this study. A randomly selected subset of 20% of the remaining representative members of clusters were also assigned to the test dataset. Data and scripts for generating the datasets are available at https://github.com/ccdmb/predector-data.

Manual effector and secretion prediction scoring

Predicted proteins were ranked using the sum of several weight-adjusted scores derived from a range of software and methods (Table 3, Supplementary Table 3). Proteins were annotated as “multiple_transmembrane” if it was assigned more than one transmembrane (TM) domains by either TMHMM or Phobius, and “single_transmembrane” if it was assigned one TM domain by TMHMM or Phobius (but neither had more than one). For TMHMM “single_transmembrane” we add the additional constraint that if there is a signal peptide prediction (by any method), the number of expected TM AAs in the first 60 residues is less than ten. A protein was annotated as “secreted” if it was predicted to have a signal peptide by any method and was not annotated as a multiple transmembrane protein.

Protein matches to PHI-base were summarised based on the experimental phenotypes of the matched proteins. Proteins were marked as a “phibase_effector_match” if they had any matches with the “Loss of pathogenicity”, “Increased virulence (Hypervirulence)”, or “Effector (plant avirulence determinant)” phenotypes; as a “phibase_virulence_match” if they had any matches with the “Reduced virulence” phenotype and not any of the effector phenotypes; and as a “phibase_lethal_match” if they had any matches with the “Lethal” phenotype. Proteins were also labelled as “effector_match”, “pfam_match”, or “dbcan_match” if they had a significant match to a custom database of known effectors, selected virulence associated Pfam HMMs, or selected virulence associated dbCAN HMMs, respectively (Supplementary Table 2).

Each protein was given two manually designed scores to evaluate effector or secreted protein candidates based on the values and weights in Supplementary Table 3. The secretion score is the sum of the products of value and weight for transmembrane, secreted, signalp3_hmm, signalp3_nn, phobius, signalp4, deepsig, targetp, and deeploc parameters. The effector score is the sum of the secretion score and the sum of the products of EffectorP, and the homology parameters (effector match, virulence match, and lethal match) values and weights.

Learning to rank model training

A “learning to rank” pairwise machine learning method based on LambdaMart [72] was developed using XGBoost [73] to prioritise effectors. Effector homologues in the training data set were held out as an informal validation set, known effector proteins were considered relevant (priority 2), and all other proteins in the train dataset were considered irrelevant (priority 1). To mitigate issues caused by unbalanced class sizes, training data were weighted for effectors as #irrelevant / #relevant and unlabelled proteins were given weight #relevant / #irrelevant. A subset of features output by the Predector pipeline and model constraints for the direction of effect (indicated in brackets as + or - when a constraint was applied; + indicating that increasing values of the feature can only contribute positively towards effector prediction) were selected based on the distributions of parameters in Supplementary File 1:3-40): molecular weight, proportion of cysteines, proportion of tiny AAs (Gly, Ala, Ser and Pro), proportion of small AAs (Thr, Asp and Asn), proportion of non-polar AAs, proportion of basic AAs, EffectorP 1

probability (+), EffectorP 2 probability (+), ApoplastP probability (+), TMHMM TM count (-), TMHMM expected TM residues in first 60 AAs, Phobius TM count (-), DeepLoc membrane probability (-), DeepLoc extracellular probability (+), DeepSig signal peptide (SP) prediction (+), Phobius SP prediction (+), SignalP 3 neural network D-score (+), SignalP 3 HMM S-score (+), SignalP 4 D-score (+), SignalP 5 SP probability (+), and TargetP secreted probability (+). The hyperparameters max_depth, min_child_weight, gamma, lambda (L2 regularisation), subsample (dropout), colsample_bytree, eta (learning rate), and num_boost_round (number of boosted trees) were optimised by maximising the normalised discounted cumulative gain (NDCG) [74] for the highest 500 ranked proteins (NDCG@500) in 5-fold cross validated training. The final model was trained using the optimised hyper-parameters.

Model and score evaluation

The learning to rank model, manually designed scores, and EffectorP pseudo-probabilities were evaluated using rank summarisation statistics using the scikit-learn library [75], which included the coverage error (the rank of the lowest scoring effector), label ranking average precision (LRAP; average proportion of correctly labelled samples with a lower score than each position in the sorted results), the label ranking loss (the average number of results that are incorrectly ordered), and the normalised discount cumulative gain (NDCG; the sum of all ranking priorities divided by the $\log_2$ of the rank position in the sorted list (DCG), normalised by the best theoretically possible DCG score) [74]. NDCG, LRAP, and label ranking loss were also evaluated for the top 50, 500, and 5000 proteins (indicated with the suffix @50, @500, or @5000). Additionally, to compare classification performance of the learn to rank model with the combined EffectorP and secretion prediction decisions, a decision threshold of 0 was set for the learn to rank model (with > 0 indicating an effector prediction), and the classification metrics precision (the proportion of predicted effectors that are labelled as true effectors), recall (the proportion of known effectors that are predicted to be effectors), accuracy (the fraction of correct predictions), balanced accuracy (the arithmetic mean of precision and recall for binary cases like this, and is less affected by unbalanced data-sets than accuracy), F1 score (the harmonic mean of precision and recall), and matthews correlation coefficient (MCC). For unbalanced datasets like the training set of effectors and non-effectors, MCC is considered a more reliable indicator of model performance than the other methods mentioned above [48]. Additionally, to evaluate the performance at different decision thresholds, the precision, recall, and MCC were calculated for 100 score thresholds along the range of each score, and the receiver operating characteristic (ROC) curves were plotted.

For the effector ranking scores, only known effectors were used as the relevant (positive) set with the irrelevant (negative or unlabelled) set consisting of secreted, non-secreted, and proteomes. Because EffectorP is intended to be run on secreted datasets, ranking statistics were only calculated for the subset of proteins that were predicted to have a signal peptide (by any method) and with fewer than two predicted TM domains (by either Phobius or TMHMM), and classification statistics were considered on both this secreted subset, and as a combined classifier (secretion and EffectorP prediction) on the whole

datasets. For the secretion ranking score the positive set consisted of the known effectors and SwissProt secreted set, and the negative set was made of the SwissProt non-secreted proteins.

Declarations

Acknowledgements

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

Author Contributions

Authors and Contributors: Conceptualisation, Methodology: DABJ, LR, JD, JKH; Software, Formal Analysis, Visualisation: DABJ; Writing – Original Draft Preparation: DABJ, JKH; Writing – Review and Editing: DABJ, LR, JD, RLM, PM, JKH; Supervision: RLM, PM, JKH.

Additional Information

Conflicts of Interest: The authors declare that there are no conflicts of interest.

Funding Information: JKH and PM were funded via Grains Research & Development Corporation project CUR00023.

Ethical Approval: n/a

Consent for Publication: n/a

References

1. He, Q., McLellan, H., Boevink, P. C. & Birch, P. R. J. All roads lead to susceptibility: the many modes-of-action of fungal and oomycete intracellular effectors. *Plant Communications.* **100050**, https://doi.org/10.1016/j.xplc.2020.100050 (2020).
2. Jones, D. A. B., Bertazzoni, S., Turo, C. J., Syme, R. A. & Hane, J. K. Bioinformatic prediction of plant–pathogenicity effector proteins of fungi. *Current Opinion in Microbiology.* **46**, 43–49 https://doi.org/10.1016/j.mib.2018.01.017 (2018).
3. Liu, L. *et al.* Arms race: diverse effector proteins with conserved motifs. *Plant Signal. Behav.* **14**, 1557008 https://doi.org/10.1080/15592324.2018.1557008 (2019).
4. Boutemy, L. S. *et al.* Structures of Phytophthora RXLR Effector Proteins: A CONSERVED BUT ADAPTABLE FOLD UNDERPINS FUNCTIONAL DIVERSITY. *J. Biol. Chem.* **286**, 35834–35842 https://doi.org/10.1074/jbc.M111.262303 (2011).

5. Jiang, R. H. Y., Tripathy, S., Govers, F. & Tyler, B. M. RXLR effector reservoir in two *Phytophthora* species is dominated by a single rapidly evolving superfamily with more than 700 members. *Proceedings of the National Academy of Sciences* **105**, 4874–4879, doi:10.1073/pnas.0709303105 (2008).
6. Galagan, J. E. & Selker, E. U. RIP: the evolutionary cost of genome defense. *Trends in Genetics.* **20**, 417–423 https://doi.org/10.1016/j.tig.2004.07.007 (2004).
7. Ohm, R. A. *et al.* Diverse Lifestyles and Strategies of Plant Pathogenesis Encoded in the Genomes of Eighteen Dothideomycetes Fungi. *PLOS Pathogens.* **8**, e1003037 https://doi.org/10.1371/journal.ppat.1003037 (2012).
8. Bertazzoni, S. *et al.* Accessories Make the Outfit: Accessory Chromosomes and Other Dispensable DNA Regions in Plant-Pathogenic Fungi. *MPMI.* **31**, 779–788 https://doi.org/10.1094/MPMI-06-17-0135-FI (2018).
9. Hane, J. K. *et al.* A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. *Genome Biology* **12**, R45, doi:10.1186/gb-2011-12-5-r45 (2011).
10. Schmidt, S. M. & Panstruga, R. Pathogenomics of fungal plant parasites: what have we learnt about pathogenesis? *Current Opinion in Plant Biology.* **14**, 392–399 https://doi.org/10.1016/j.pbi.2011.03.006 (2011).
11. Beckerson, W. C. *et al.* Cause and Effectors: Whole-Genome Comparisons Reveal Shared but Rapidly Evolving Effector Sets among Host-Specific Plant-Castrating Fungi. *mBio* **10**, doi:10.1128/mBio.02391-19 (2019).
12. Mousavi-Derazmahalleh, M. *et al.* Prediction of pathogenicity genes involved in adaptation to a lupin host in the fungal pathogens Botrytis cinerea and Sclerotinia sclerotiorum via comparative genomics. *BMC Genomics.* **20**, 385 https://doi.org/10.1186/s12864-019-5774-2 (2019).
13. Plissonneau, C. *et al.* Using Population and Comparative Genomics to Understand the Genetic Basis of Effector-Driven Fungal Pathogen Evolution. *Front. Plant Sci.* **8**, https://doi.org/10.3389/fpls.2017.00119 (2017).
14. Williams, A. H. *et al.* Comparative genomics and prediction of conditionally dispensable sequences in legume–infecting Fusarium oxysporum formae speciales facilitates identification of candidate effectors. *BMC Genomics.* **17**, 191 https://doi.org/10.1186/s12864-016-2486-8 (2016).
15. Gervais, J. *et al.* Different waves of effector genes with contrasted genomic location are expressed by Leptosphaeria maculans during cotyledon and stem colonization of oilseed rape. *Mol. Plant Pathol.* **18**, 1113–1126 https://doi.org/10.1111/mpp.12464 (2017).
16. Human, M. P., Berger, D. K. & Crampton, B. G. Time-Course RNAseq Reveals Exserohilum turcicum Effectors and Pathogenicity Determinants. *Front. Microbiol.* **11**, https://doi.org/10.3389/fmicb.2020.00360 (2020).
17. Jones, D. A. B. *et al.* A specific fungal transcription factor controls effector gene expression and orchestrates the establishment of the necrotrophic pathogen lifestyle on wheat. *Sci. Rep.* **9**, 1–13 https://doi.org/10.1038/s41598-019-52444-7 (2019).

18. Gawehns, F. *et al.* The effector repertoire of Fusarium oxysporum determines the tomato xylem proteome composition following infection. *Front. Plant Sci.* **6**, https://doi.org/10.3389/fpls.2015.00967 (2015).
19. Mesarich, C. H. *et al.* Specific Hypersensitive Response–Associated Recognition of New Apoplastic Effectors from Cladosporium fulvum in Wild Tomato. *MPMI.* **31**, 145–162 https://doi.org/10.1094/MPMI-05-17-0114-FI (2017).
20. Richards, J. K. *et al.* Local adaptation drives the diversification of effectors in the fungal wheat pathogen Parastagonospora nodorum in the United States. *PLOS Genetics.* **15**, e1008223 https://doi.org/10.1371/journal.pgen.1008223 (2019).
21. Sánchez-Vallet, A., Hartmann, F. E., Marcel, T. C. & Croll, D. Nature's genetic screens: using genome-wide association studies for effector discovery. *Mol. Plant Pathol.* **19**, 3–6 https://doi.org/10.1111/mpp.12592 (2018).
22. Sperschneider, J., Williams, A. H., Hane, J. K., Singh, K. B. & Taylor, J. M. Evaluation of Secretion Prediction Highlights Differing Approaches Needed for Oomycete and Fungal Effectors. *Front. Plant Sci.* **6**, https://doi.org/10.3389/fpls.2015.01168 (2015).
23. Armenteros, J. J. A. *et al.* Detecting sequence signals in targeting peptides using deep learning. *Life Science Alliance.* **2**, https://doi.org/10.26508/lsa.201900429 (2019).
24. Dyrløv Bendtsen, J., Nielsen, H., von Heijne, G. & Brunak, S. Improved Prediction of Signal Peptides: SignalP 3.0. *Journal of Molecular Biology.* **340**, 783–795 https://doi.org/10.1016/j.jmb.2004.05.028 (2004).
25. Petersen, T. N., Brunak, S., Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods.* **8**, 785–786 https://doi.org/10.1038/nmeth.1701 (2011).
26. Käll, L., Krogh, A. & Sonnhammer, E. L. L. A Combined Transmembrane Topology and Signal Peptide Prediction Method. *Journal of Molecular Biology.* **338**, 1027–1036 https://doi.org/10.1016/j.jmb.2004.03.016 (2004).
27. Savojardo, C., Martelli, P. L., Fariselli, P. & Casadio, R. DeepSig: deep learning improves signal peptide detection in proteins. *Bioinformatics.* **34**, 1690–1696 https://doi.org/10.1093/bioinformatics/btx818 (2018).
28. Armenteros, J. J. A., Sønderby, C. K., Sønderby, S. K., Nielsen, H. & Winther, O. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics.* **33**, 3387–3395 https://doi.org/10.1093/bioinformatics/btx431 (2017).
29. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology.* **305**, 567–580 https://doi.org/10.1006/jmbi.2000.4315 (2001).
30. Sperschneider, J., Dodds, P. N., Gardiner, D. M., Singh, K. B. & Taylor, J. M. Improved prediction of fungal effector proteins from secretomes with EffectorP 2.0. *Mol. Plant Pathol.* **19**, 2094–2110 https://doi.org/10.1111/mpp.12682 (2018).

31. Sperschneider, J., Dodds, P. N., Singh, K. B. & Taylor, J. M. ApoplastP: prediction of effectors and plant proteins in the apoplast using machine learning. *New Phytol.* **217**, 1764–1778 https://doi.org/10.1111/nph.14946 (2018).
32. Wang, C. *et al.* FunEffector-Pred: Identification of Fungi Effector by Activate Learning and Genetic Algorithm Sampling of Imbalanced Data. *IEEE Access.* **8**, 57674–57683 https://doi.org/10.1109/ACCESS.2020.2982410 (2020).
33. Sperschneider, J. *et al.* LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell. *Sci. Rep.* **7**, 1–14 https://doi.org/10.1038/srep44598 (2017).
34. Lu, S., Turgeon, B. G. & Edwards, M. C. A ToxA-like protein from Cochliobolus heterostrophus induces light-dependent leaf necrosis and acts as a virulence factor with host selectivity on maize. *Fungal Genetics and Biology.* **81**, 12–24 (2015).
35. de Guillen, K. *et al.* Structure analysis uncovers a highly diverse but structurally conserved effector family in phytopathogenic fungi. *PLoS Pathog.* **11**, e1005228 (2015).
36. Spanu, P. D. Cereal immunity against powdery mildews targets RNase-Like Proteins associated with Haustoria (RALPH) effectors evolved from a common ancestral gene(2017).
37. Kale, S. D. *et al.* External lipid PI3P mediates entry of eukaryotic pathogen effectors into plant and animal host cells. *Cell.* **142**, 284–295 (2010).
38. Urban, M. *et al.* PHI-base: the pathogen–host interactions database. *Nucleic Acids Res.* **48**, D613–D620 https://doi.org/10.1093/nar/gkz904 (2020).
39. Lu, T., Yao, B. & Zhang, C. DFVF: database of fungal virulence factors. Database 2012, bas032-bas032, doi:10.1093/database/bas032 (2012).
40. Anderson, J. P. *et al.* Comparative secretome analysis of Rhizoctonia solani isolates with different host ranges reveals unique secretomes and cell death inducing effectors. *Sci. Rep.* **7**, 10410 https://doi.org/10.1038/s41598-017-10405-y (2017).
41. Dutreux, F. *et al.* De novo assembly and annotation of three Leptosphaeria genomes using Oxford Nanopore MinION sequencing. *Scientific Data.* **5**, 180235 https://doi.org/10.1038/sdata.2018.235 (2018).
42. Sonah, H. *et al.* Comparative Transcriptomic Analysis of Virulence Factors in Leptosphaeria maculans during Compatible and Incompatible Interactions with Canola. *Front. Plant Sci.* **7**, https://doi.org/10.3389/fpls.2016.01784 (2016).
43. Syme, R. A. *et al.* Pan-Parastagonospora Comparative Genome Analysis—Effector Prediction and Genome Evolution. *Genome Biol Evol.* **10**, 2443–2457 https://doi.org/10.1093/gbe/evy192 (2018).
44. Qiu, J., Sheffler, W., Baker, D. & Noble, W. S. Ranking predicted protein structures with support vector regression. *Proteins: Structure, Function, and Bioinformatics.* **71**, 1175–1182 https://doi.org/10.1002/prot.21809 (2008).
45. Liu, B., Chen, J. & Wang, X. Application of learning to rank to protein remote homology detection. *Bioinformatics.* **31**, 3492–3498 https://doi.org/10.1093/bioinformatics/btv413 (2015).

46. You, R. *et al.* GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics.* **34**, 2465–2473 https://doi.org/10.1093/bioinformatics/bty130 (2018).
47. Liu, L., Huang, X., Mamitsuka, H. & Zhu, S. HPOLabeler: improving prediction of human protein–phenotype associations by learning to rank. *Bioinformatics.* https://doi.org/10.1093/bioinformatics/btaa284 (2020).
48. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics.* **21**, 6 https://doi.org/10.1186/s12864-019-6413-7 (2020).
49. Zhang, H. *et al.* dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **46**, W95–W101 https://doi.org/10.1093/nar/gky418 (2018).
50. Schweizer, G. *et al.* Positively Selected Effector Genes and Their Contribution to Virulence in the Smut Fungus Sporisorium reilianum. *Genome Biol Evol.* **10**, 629–645 https://doi.org/10.1093/gbe/evy023 (2018).
51. Testa, A. C., Oliver, R. P. & Hane, J. K. OcculterCut: a comprehensive survey of AT-rich regions in fungal genomes. *Genome Biol Evol.* **8**, 2044–2064 (2016).
52. Aylward, J. *et al.* A plant pathology perspective of fungal genome sequencing. *IMA Fungus.* **8**, 1–15 https://doi.org/10.5598/imafungus.2017.08.01.01 (2017).
53. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nature Biotechnology.* **35**, 316–319 https://doi.org/10.1038/nbt.3820 (2017).
54. Sperschneider, J. *et al.* EffectorP: predicting fungal effector proteins from secretomes using machine learning. *New Phytol.* **210**, 743–761 https://doi.org/10.1111/nph.13794 (2016).
55. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology.* **35**, 1026–1028 https://doi.org/10.1038/nbt.3988 (2017).
56. Hane, J. K., Paxman, J., Jones, D. A. B., Oliver, R. P. & 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.* **10**, 3088 https://doi.org/10.3389/fmicb.2019.03088 (2020).
57. Frantzeskakis, L. *et al.* Signatures of host specialization and a recent transposable element burst in the dynamic one-speed genome of the fungal barley powdery mildew pathogen. *BMC Genomics.* **19**, 381 https://doi.org/10.1186/s12864-018-4750-6 (2018).
58. Müller, M. C. *et al.* A chromosome-scale genome assembly reveals a highly dynamic effector repertoire of wheat powdery mildew. *New Phytol.* **221**, 2176–2189 https://doi.org/10.1111/nph.15529 (2019).
59. Nemri, A. *et al.* The genome sequence and effector complement of the flax rust pathogen Melampsora lini. *Front. Plant Sci.* **5**, https://doi.org/10.3389/fpls.2014.00098 (2014).

60. Duplessis, S. *et al.* Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proceedings of the National Academy of Sciences* **108**, 9166, doi:10.1073/pnas.1019315108 (2011).
61. Li, F. *et al.* Emergence of the Ug99 lineage of the wheat stem rust pathogen through somatic hybridisation. *Nature Communications.* **10**, 5068 https://doi.org/10.1038/s41467-019-12927-7 (2019).
62. Syme, R. A. *et al.* Comprehensive Annotation of the Parastagonospora nodorum Reference Genome Using Next-Generation Genomics, Transcriptomics and Proteogenomics. *PLOS ONE.* **11**, e0147221 https://doi.org/10.1371/journal.pone.0147221 (2016).
63. Moolhuijzen, P. *et al.* Comparative genomics of the wheat fungal pathogen Pyrenophora tritici-repentis reveals chromosomal variations and genome plasticity. *BMC Genomics.* **19**, 279 https://doi.org/10.1186/s12864-018-4680-3 (2018).
64. Syme, R. A. *et al.* Transposable Element Genomic Fissuring in Pyrenophora teres Is Associated With Genome Expansion and Dynamics of Host–Pathogen Genetic Interactions. *Front. Genet.* **9**, https://doi.org/10.3389/fgene.2018.00130 (2018).
65. Goodwin, S. B. *et al.* Finished Genome of the Fungal Wheat Pathogen Mycosphaerella graminicola Reveals Dispensome Structure, Chromosome Plasticity, and Stealth Pathogenesis. *PLOS Genetics.* **7**, e1002070 https://doi.org/10.1371/journal.pgen.1002070 (2011).
66. Plissonneau, C., Hartmann, F. E. & Croll, D. Pangenome analyses of the wheat pathogen Zymoseptoria tritici reveal the structural basis of a highly plastic eukaryotic genome. *BMC Biology.* **16**, 5 https://doi.org/10.1186/s12915-017-0457-4 (2018).
67. de Wit, P. J. G. M. *et al.* The Genomes of the Fungal Plant Pathogens Cladosporium fulvum and Dothistroma septosporum Reveal Adaptation to Different Hosts and Lifestyles But Also Signatures of Common Ancestry. *PLOS Genetics.* **8**, e1003088 https://doi.org/10.1371/journal.pgen.1003088 (2012).
68. Delulio, G. A. *et al.* Kinome Expansion in the Fusarium oxysporum Species Complex Driven by Accessory Chromosomes. *mSphere* **3**, doi:10.1128/mSphere.00231-18 (2018).
69. Ma, L. J. *et al.* Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. *Nature.* **464**, 367–373 https://doi.org/10.1038/nature08850 (2010).
70. Ma, L. J., Shea, T., Young, S., Zeng, Q. & Kistler, H. C. Genome Sequence of Fusarium oxysporum f. sp. melonis Strain NRRL 26406, a Fungus Causing Wilt Disease on Melon. *Genome Announc.* **2**, https://doi.org/10.1128/genomeA.00730-14 (2014).
71. MacCallum, I. *et al.* ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. *Genome Biol.* **10**, R103 https://doi.org/10.1186/gb-2009-10-10-r103 (2009).
72. Wu, Q., Burges, C. J. C., Svore, K. M. & Gao, J. Adapting boosting for information retrieval measures. *Inf. Retr.* **13**, 254–270 https://doi.org/10.1007/s10791-009-9112-1 (2010).
73. Chen, T. & Guestrin, C. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* 785–794.
74. Wang, Y., Wang, L., Li, Y., He, D. & Liu, T. Y. in *Conference on Learning Theory.* 25–54 (PMLR).

75. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research.* **12**, 2825–2830 (2011).

Figures

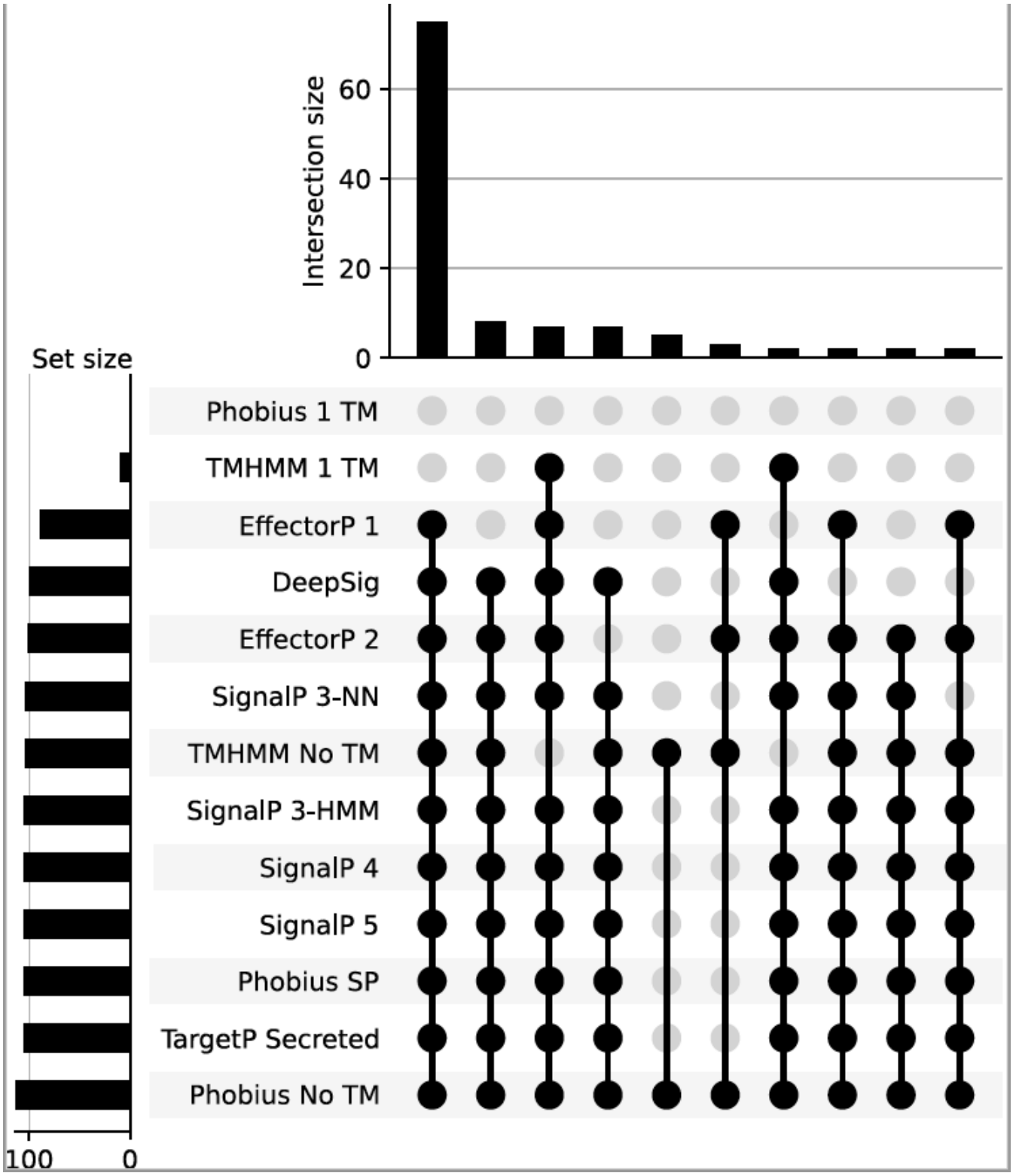

Figure 1

UpSet plot showing predictions of signal peptides, transmembrane domains, and effector-like properties for all known effectors in the training dataset (N=125). Rows indicate sets of proteins predicted to have a property related to effector prediction (e.g. a signal peptide), with the horizontal bar chart indicating set size. Columns indicate where the horizontal sets intersect with each other, where the vertical bar-chart indicates the number of proteins in that intersection. For clarity, intersections with only 1 member have been excluded, the full plot is presented in Supplementary File 1:1.

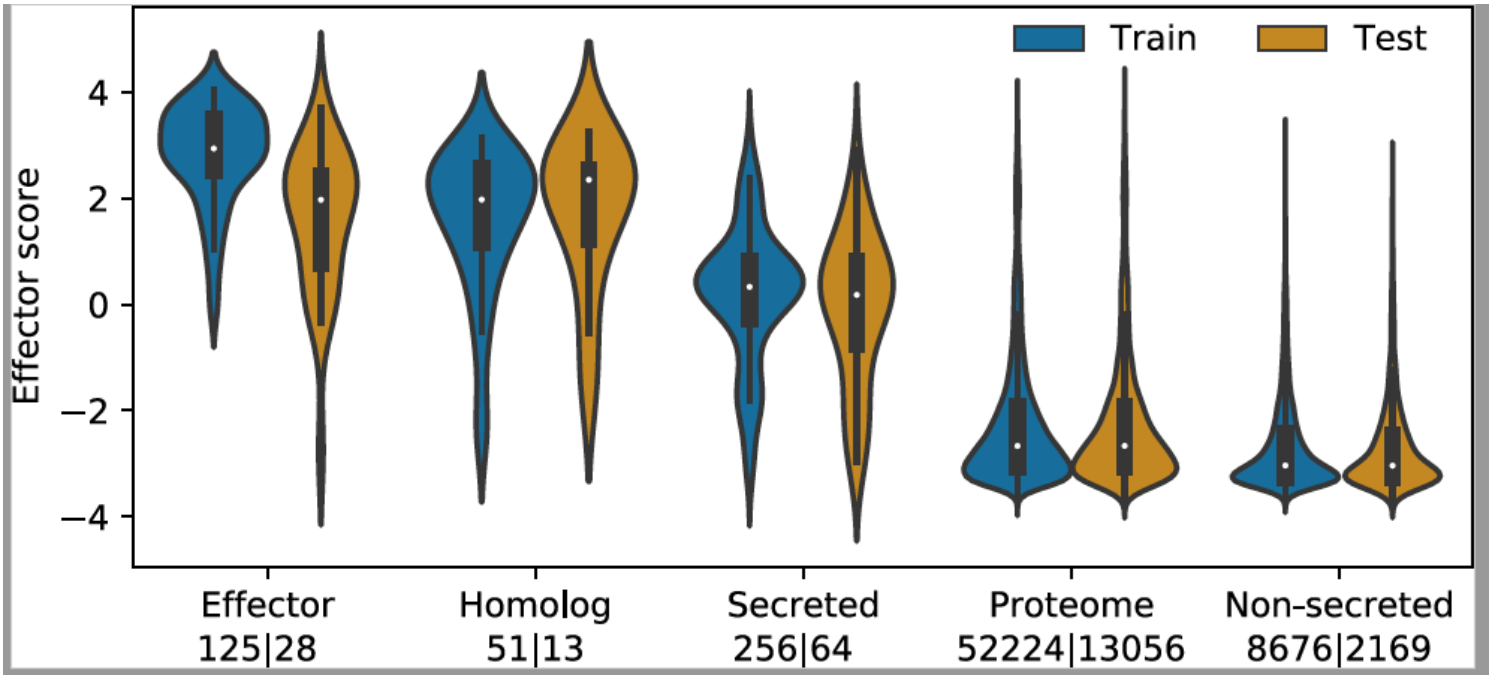

Figure 2

A violin plot showing the distributions of Predector effector ranking scores for each class in the test and training datasets. The effectors consist of experimentally validated fungal effector sequences. “Secreted” and “non-secreted” proteins are manually annotated proteins from the SwissProt database. Proteomes consist of the complete predicted proteomes from 10 well studied fungi (Supplementary Table 2). The number of proteins represented by each violin are indicated on the x-axis.

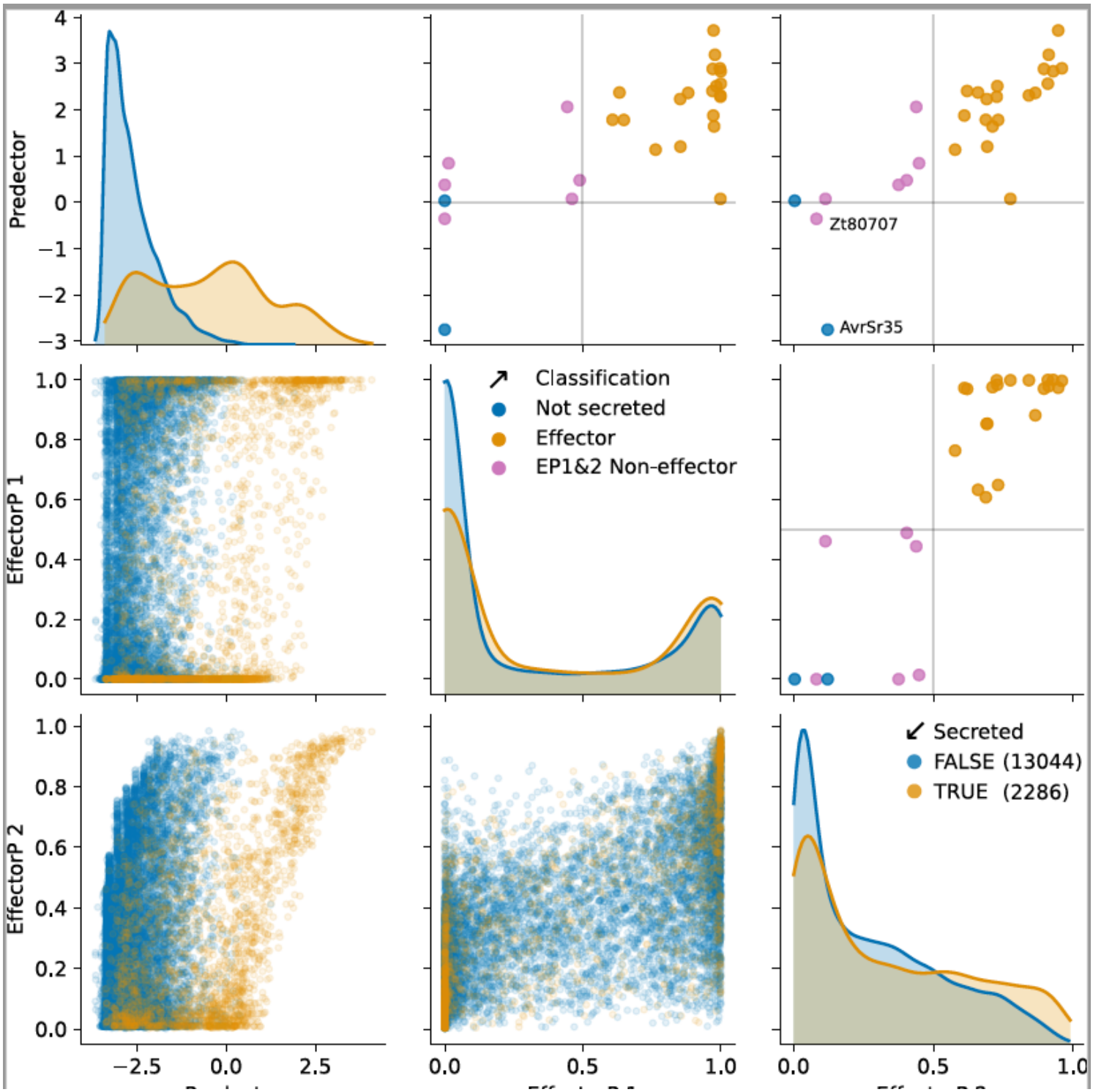

Figure 3

Comparing the scores of Predector with EffectorP versions 1 and 2 for proteins in the testing dataset. Scatter plots in the lower-left corner indicate comparisons of predictive scores between methods, with predicted secreted proteins (any signal peptide and fewer than two TM domains predicted) indicated in yellow, and non-secreted proteins indicated in blue. Density plots along the diagonal indicate distributions of the full test dataset versus predictive scores for each method (indicated along the x-axis), also coloured by secretion prediction as before (Note: there are far more non-secreted than secreted proteins in

the dataset). Scatter plots in the top-right corner indicate score comparisons between methods for confirmed effectors, coloured by whether they have been predicted as secreted (criteria as above), or additionally predicted by EffectorP versions 1 or 2. Two proteins that are misclassified by a Predector score > 0 are labelled in the top-right subplot.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FileS1.docx
- SupplementaryTable1.xlsx
- SupplementaryTable2.xlsx
- SupplementaryTable3.xlsx
- SupplementaryTable4.zip.001
- SupplementaryTable4.zip.002
- SupplementaryTable4.zip.003
- SupplementaryTable4.zip.004
- SupplementaryTable4.zip.005
- SupplementaryTable4.zip.006
- SupplementaryTable4.zip.007