EffectorO: motif-independent prediction of effectors in oomycete genomes using machine learning and lineage specificity

Munir Nur[1#], Kelsey Wood[1,2#], and Richard Michelmore[1,3*]

[1]The Genome Center, University of California, Davis, USA

[2]Integrative Genetics & Genomics Graduate Group, University of California, Davis, USA

[3]Departments of Plant Sciences, Molecular & Cellular Biology, Medical Microbiology & Immunology, University of California, Davis, USA.

Co-first authors

* Corresponding author e-mail: rwmichelmore@ucdavis.edu

Abstract

Oomycete plant pathogens cause a wide variety of diseases, including late blight of potato, sudden oak death, and downy mildew of many plants. These pathogens are major contributors to losses in many food crops. Oomycetes secrete "effector" proteins to manipulate their hosts to the advantage of the pathogen. Plants have evolved to recognize effectors, resulting in an evolutionary cycle of defense and counter-defense in plant–microbe interactions. This selective pressure results in highly diverse effector sequences that can be difficult to computationally identify using sequence similarity. We developed a pipeline, EffectorO, that uses two complementary approaches to predict effectors in oomycete pathogen genomes: (1) a machine learning-based pipeline that predicts effector probability based on the biochemical properties of the N-terminal amino acid sequence of a protein and

24 is trained on experimentally verified oomycete effectors and (2) a pipeline based on lineage-
25 specificity to find proteins that are unique to one species or genus, a sign of evolutionary
26 divergence due to adaptation to the host. We tested EffectorO on *Bremia lactucae*, which
27 causes lettuce downy mildew, and *Phytophthora infestans*, which causes late blight of potato
28 and tomato, and predicted many novel effector candidates, while still recovering the
29 majority of known effector candidates. EffectorO will be useful for discovering novel families
30 of oomycete effectors without relying on sequence similarity to known effectors.
31

32 **Introduction**

33 Oomycetes are filamentous eukaryotic organisms that resemble fungi but are more
34 closely related to brown algae (Baldauf 2003). Pathogenic oomycetes cause huge losses in
35 agriculture and aquaculture worldwide (Derevnina et al. 2016; Haverkort et al. 2008).
36 Oomycete pathogens and other parasites must remain undetected by their hosts in order to
37 successfully infect and complete their life cycle. This is accomplished through the action of
38 effectors, secreted proteins that interfere with host immune signaling and physiology to
39 promote pathogen growth (Wang et al. 2019). To thwart the action of effectors and infection
40 by the pathogen, plants have evolved receptors encoded by resistance genes (R genes) to
41 detect effectors and trigger a defense response resulting in disease resistance (Wang et al.
42 2019). Therefore, identification of effectors is important for both understanding how
43 pathogens infect their hosts and for understanding and breeding for disease resistance.

44 For some oomycetes, such as *Phytophthora* species and downy mildew pathogens, the
45 discovery of conserved amino acid motifs and protein domains in effector proteins has
46 assisted in the prediction of effectors from genomes. One such motif is the RXLR motif, which

47 is located in the N-terminus of the protein just following the secretion-related signal peptide
48 (Rehmany et al. 2005). The RXLR motif is often associated with a downstream EER motif,
49 and both motifs have been implicated in secretion and/or translocation of effectors into host
50 cells (Birch et al. 2009; Whisson et al. 2007). Although their role in effector movement
51 remains unresolved (Ellis and Dodds 2011), the predictive value of the RXLR and EER motifs
52 has been demonstrated by the discovery of numerous effectors predicted in pathogen
53 genomes using the motif and validated by testing the candidate proteins for recognition by
54 resistant genotypes (avirulence activity) or immune suppression (virulence activity) *in*
55 *planta* (Deb et al. 2018; Fabro et al. 2011; Pecrix et al. 2019; Pel et al. 2009; Stassen et al.
56 2013; Zheng et al. 2014). While the RXLR motif is clearly a conserved effector motif in
57 *Phytophthora* species, effectors in the related downy mildew pathogens show divergence
58 and diversification of this motif (Bailey et al. 2011; Sharma et al. 2015).

59 Due to the sequence divergence of the RXLR motif in downy mildew species, both
60 string- and Hidden Markov Model (HMM)-based searches recover fewer effector candidates
61 in these pathogens compared to *Phytophthora* spp. (Baxter et al. 2010; McGowan and
62 Fitzpatrick 2017). Another approach to predict effectors in downy mildew species (as well
63 as in *Phytophthora*) is to search for the structurally conserved WY domain, which is present
64 in ~25–50% of RXLR effectors in *Phytophthora* and downy mildew pathogens (Bos et al.
65 2010; Dou et al. 2008; He et al. 2019; King et al. 2014; Win et al. 2012) and has been
66 implicated in effector function (Bos et al. 2010; Dou et al. 2008; He et al. 2019; King et al.
67 2014). While HMM searches for the WY domain have a low false positive rate for predicting
68 candidate effectors (Wood et al. 2020), they likely have a high false negative rate, as not all
69 RXLR effectors have WY domains and there may be *bona fide* effectors with novel N-terminal

70 motifs and uncharacterized effector domains. Unlike in oomycetes, there are no widely
71 conserved effector motifs across all fungi. Degenerate RXLR-like motifs have been found in
72 some fungal pathogens (Gu et al. 2011; Kale et al. 2010) and others have an N-terminal
73 Y/F/WxC motif (Godfrey et al. 2010).

74 Recently, EffectorP, a machine learning based approach to identify effectors in fungal
75 genomes using biochemical characteristics, successfully expanded the pool of candidate
76 effectors in fungal species (Sperschneider et al. 2016, 2018). Machine learning-based
77 effector prediction pipelines, such as EffectorP, work by building effector classification
78 models trained on known effectors and secreted non-effectors from the target organisms.
79 Once a suitable model is determined, the classifier calculates the biochemical features from
80 the amino acid sequence of the protein and outputs a probability that the protein is an
81 effector based on those features. To date, machine learning has not been applied to effector
82 prediction in oomycetes.

83 Due to the constant evolutionary cycle between pathogen and host, pathogenicity-
84 related proteins, such as effectors (as well as their corresponding R genes), often show
85 increased rates of sequence divergence (Dong et al. 2015). This is especially true for
86 organisms that have narrow host ranges, such as downy mildew pathogens. For example,
87 over 6,000 genes in the downy mildew pathogen *Hyaloperonospora arabidopsidis* had no
88 identified ortholog in *Phytophthora* species (Baxter et al. 2010) and over 10,000 genes were
89 lineage-specific in the animal pathogen *Saprolegnia parasitica* (Jiang et al. 2013). This lineage
90 specificity can be accounted for in part by the lack of genomic resources in closely related
91 organisms, but for many of these proteins, it is likely due to co-adaptation to specific hosts
92 of these pathogens. We reasoned that by using lineage specificity in combination with a

93 machine learning-based effector classifier, we would be able to expand the candidate effector
94 repertoires of oomycete pathogens, especially those with narrow host ranges.

95 Here we present EffectorO: a pipeline to predict novel effectors in oomycete genomes
96 that avoids the limitations of motif/domain-based searches. EffectorO consists of a machine
97 learning based effector classifier trained on a diverse set of oomycete effectors (EffectorO-
98 ML) combined with a lineage-specific protein sequence identification pipeline (EffectorO-
99 LSP). We used two economically important oomycete pathogens as model organisms:
100 *Bremia lactucae*, an obligate biotrophic downy mildew pathogen with complete host-
101 specificity to *Lactuca* spp., and *Phytophthora infestans*, which causes late blight of potato and
102 tomato. EffectorO had high cross-validation accuracy for classifying protein sequences as
103 effectors or non-effectors and predicted thousands of novel effectors in these two oomycete
104 genomes, while recovering known effectors. This analysis was expanded to 27 additional
105 oomycete genomes. One EffectorO prediction was validated by demonstrating the elicitation
106 of cell death on downy mildew resistant genotypes by an effector of *B. lactucae* with no
107 obvious canonical sequence domains. Thus, a machine learning approach, especially when
108 combined with lineage-specificity analysis, can be useful in gene annotation and effector
109 prediction in oomycete genomes.

110

111 **Materials and Methods**

112 *Training sets: Positive (effectors) and negative (non-effectors)*

113 Our positive training set consisted of oomycete effectors experimentally shown to be
114 either recognized by R genes (avirulence activity) or to have an immune-suppressing effect
115 on the host (virulence activity) (Supplemental Table 1). Seventy-seven of these sequences

116 were obtained from the Pathogen-Host Interaction database (PHI-base) (Urban et al. 2019)
117 and 15 effector candidates were added from *B. lactucae* that had avirulence (Giesbers et al.
118 2017; Pelgrom et al. 2019; Stassen et al. 2013; Wood et al. 2020) or virulence activity (Wood
119 et al. 2020) resulting in an initial set of 92 sequences. We then used CD-Hit at a sequence
120 identity threshold of 70% and word length of five amino acids in order to identify and
121 remove closely related sequences (Fu et al. 2012; Li and Godzik 2006), resulting in a final
122 positive training set of 88 unique oomycete effectors.

123 Our negative training set consisted of representative oomycete sequences that were
124 predicted to be secreted and highly conserved throughout the genomes of 28 oomycete
125 species (Supplemental Table 2). To find these sequences, we first performed an Orthofinder
126 search on the oomycete genomes at an e-value cutoff of 1e-10 and a query coverage cutoff of
127 90% (Emms and Kelly 2015, 2019). We then retrieved ortholog centroids, with the species
128 chosen at random, that were present in at least 26 genomes. We predicted signal peptides
129 using SignalP v4.1 (sensitive mode) (Petersen et al. 2011). To further reduce potential model
130 bias and redundancy in the set of non-effectors, we ran CD-hit and removed sequences with
131 greater than 40% identity. This identified 320 sequences, from which a training set of 88
132 sequences was chosen at random to obtain a balanced 1:1 positive to negative training set
133 ratio.

134

135 *Feature selection and encoding for classical machine learning models*

136 To calculate biochemical characteristics of effectors and non-effectors, we used
137 published amino acid scales from ExPASyProtScale (ExPaSy; Expert Protein Analysis System
138 2004) and evaluated six features that could be informative for effector classification. We

139 focused on the first 100 amino acids of the N-terminus of protein sequences, with the
140 rationale that (1) secretion and translocation signals are typically encoded in the N-terminus
141 and (2) effectors may differ in their C-terminal domains depending on their specific functions
142 (for example, not all validated oomycete effectors have WY domains). To ensure the feature
143 scales were equivalent to each other, we transformed the calculated feature distributions to
144 zero mean and unit variance (Z score). We calculated the average values for six N-terminal
145 features: the grand average of hydropathy (GRAVY), hydrophobicity, surface exposure,
146 disorder propensity, bulkiness, and residue interface propensity of the first 100 amino acids
147 using data from ExPASyProtScale (ExPaSy; Expert Protein Analysis System 2004).
148 PONDR VSL2 (Peng et al. 2006) was used as an additional method to visualize levels
149 of intrinsic disorder for effectors and secreted non-effectors across the N-terminal sequence.
150 The average sequence disorder was calculated at each amino acid position for the first 160
151 amino acids including the signal peptide in order to visualize the differences between
152 candidate effectors and non-effectors.
153
154 *Feature encoding for Convolutional Neural Network models*
155 In order to account for signals that may be concealed through net average
156 calculations, we also tried using a convolutional neural network (CNN) to extract positional-
157 dependent differences in amino acid sequence between effectors and non-effectors. We used
158 the first 100 amino acid values for each feature and padded the ends with zeros for
159 sequences shorter than 100 aa. For each protein in our positive and negative training sets,
160 this feature encoding resulted in a 6 x 100 matrix to include every positional encoding for all
161 six biochemical features in the first 100 amino acids.

162

163 *Machine learning architectures*

164 For the classical machine learning pipeline, we tested various parametric and non-
165 parametric models and evaluated what worked best for effector classification. We tested and
166 compared the following models using multiple cross-validation trials (resampling
167 procedures that rotate training and testing splits): Random Forest, Naïve Bayes (Bernoulli
168 and Gaussian), Logistic Regression, and K-Nearest Neighbors. To reduce overfitting, we
169 monitored training and testing scores (accuracy, sensitivity, specificity, false positive rate,
170 and area under the precision-recall curve). Models were used from the scikit-learn machine
171 learning library in Python (Pedregosa et al. 2011). We tested the Logistic Regression, and
172 Gaussian/Bernoulli Naive-Bayes models with default parameters, as well as the Random
173 Forest classifier with 400 estimators and a max depth of 4, and the K-Nearest Neighbors
174 classifier with the neighbors parameter set to 5.

175

176 *Lineage Specificity*

177 Proteins encoded by lineage specific genes were obtained using the predicted
178 secretome of a species as a BLASTp query against the translated open reading frames (ORFs)
179 of other oomycete species (Supplemental Table 2). Conserved proteins were defined as
180 proteins with a BLASTp hit in another species with an e-value cutoff of <10e-7, percent
181 identity of >30%, and a query coverage of > 30%. Protein sequences that were not conserved
182 in another oomycete species were considered to be lineage-specific. Similar BLASTp cutoffs
183 are used in other pipelines to obtain species-specific genes for numerous species, including
184 other oomycetes (Rujirawat et al. 2019; Zhou et al. 2015). ORFs were used instead of gene

185 models because most oomycete effectors are encoded by a single exon and because gene
186 prediction algorithms often misannotate effector genes. Orthofinder (Emms and Kelly 2015)
187 was tested as well but was found to predict a much higher number of lineage specific
188 proteins (and thus likely contained more false positives) so we used our more conservative
189 BLASTp-based pipeline for EffectorO-LSP.
190
191 *Prediction of RXLR-EER motifs and WY domains*
192 To predict the RXLR-EER motif, we used a string search for [RQGH]XLR and
193 [DE][DE][KR] for EER as well as the HMM for RXLR-EER developed by Whisson et al. (2007)
194 for *Phytophthora* species. To predict WY domains, we used the HMM training set containing
195 three *Phytophthora* species by Boutemy et al. (2011), with proteins with a HMMer bitscore
196 >0 classified as having a WY domain. HMMer v3.1 was used for all HMM searches (Eddy
197 2011).
198
199 *Scripts*
200 All scripts for running EffectorO-ML and EffectorO-LSG can be found on Github at
201 https://github.com/mjnur/oomycete-effector-prediction. A webapp for EffectorO-ML is
202 available at https://bremia.ucdavis.edu/effectorO.php.
203
204 *Candidate gene cloning and agroinfiltration*
205 The *Avr6* candidate *BLE01* was cloned from *B. lactucae* genomic DNA using Gateway
206 cloning into pEG100 (35S promoter) as described previously (Wood et al. 2020). The signal

207 peptide was excluded from the coding sequence to ensure expression inside the plant cell.
208 Two distinct alleles were obtained. *Agrobacterium tumefaciens* strain C58Rif+ was
209 transformed with pEG100-BLE01 and infiltrations were performed on three- to five-week-
210 old lettuce plants as described previously (Wood et al. 2020). Leaves were scored for
211 necrosis 4-5 days after infiltration. *A. tumefaciens* containing pEG100 (empty vector) and
212 pEG100-GFP were used as negative controls, and *A. tumefaciens* containing pBAV139-
213 35S:HopM1 and pBAV139-35S:AvrPto were used as positive controls for cell death
214 (Wroblewski et al. 2009).
215

216 **Results**

217 *Effector and non-effector proteins differ in their biochemical characteristics*

218 To determine whether various N-terminal amino acid characteristics would be
219 sufficient to distinguish effectors from non-effector proteins, we compared predicted
220 biochemical parameters between validated oomycete effectors and secreted, conserved
221 proteins (presumed non-effectors). We found that validated effectors had higher intrinsic
222 disorder and had lower scores for GRAVY, hydrophobicity, exposed residues, bulkiness, and
223 interface, compared with secreted, conserved proteins (Fig. 1A). The intrinsic disorder score
224 was significantly higher in validated effectors, starting from just after the signal peptide (first
225 ~20 aa) until approximately 80 aa after the start codon (Fig. 1B).
226

227 *Evaluation of machine learning effector classifier models*

228 To find the effector classifier with the best performance, we evaluated several
229 machine learning models (Random Forest, Linear SVCs, Bernoulli and Gaussian Naïve Bayes,

230 Logistic Regression, and K-Nearest Neighbors) as well as a convolutional neural network
231 model that takes into account positional information for classification. The Random Forest
232 classifier initially outperformed other models, so we attempted several combinations of
233 parameters, adjusting the number of estimators (n_estimators) and maximum tree depth
234 (max_depth), while monitoring the cross-validation accuracy. The best receiver-operator
235 curve (ROC) was achieved by the Random Forest classifier (with parameters n_estimators =
236 400 and max_depth = 4) with an area under the curve (AUC) of 0.89 (Figure 2) and an overall
237 5-fold cross-validation accuracy of 84.26% (Table 1). We also tested a convolutional neural
238 network (CNN), which takes position into account, to predict effectors; however, it was found
239 to perform similarly but not better than the most accurate Random Forest model. Therefore,
240 we used the Random Forest model for our EffectorO-ML pipeline to classify effectors based
241 on the biochemical characteristics of the first 100 amino acids of the N-terminus for all
242 subsequent analyses.
243

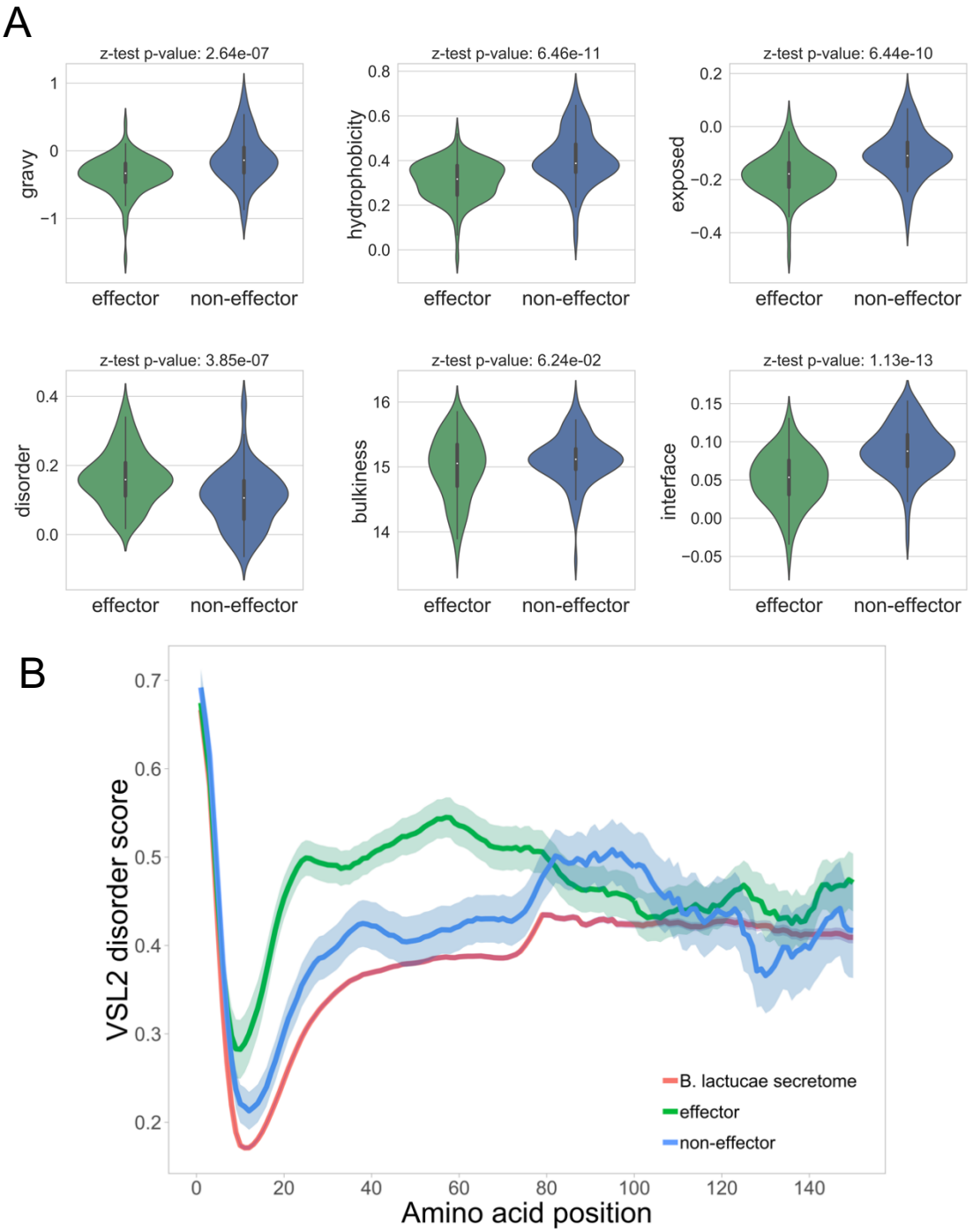


Figure 1. (A) Amino acid biochemical characteristics for validated oomycete effectors and secreted non-effectors as calculated by the ExPASyProtScale and used as features for the machine learning classifier. Violin plot shows the distribution of the N-terminal Z scores for each group of proteins, with the p-value shown above plot. **(B)** Intrinsic disorder as calculated by PONDR VSL2 for validated oomycete effectors and secreted non-effectors with the *B. lactucae* secretome for reference. The average and standard error are shown for each group of proteins at each amino acid position.

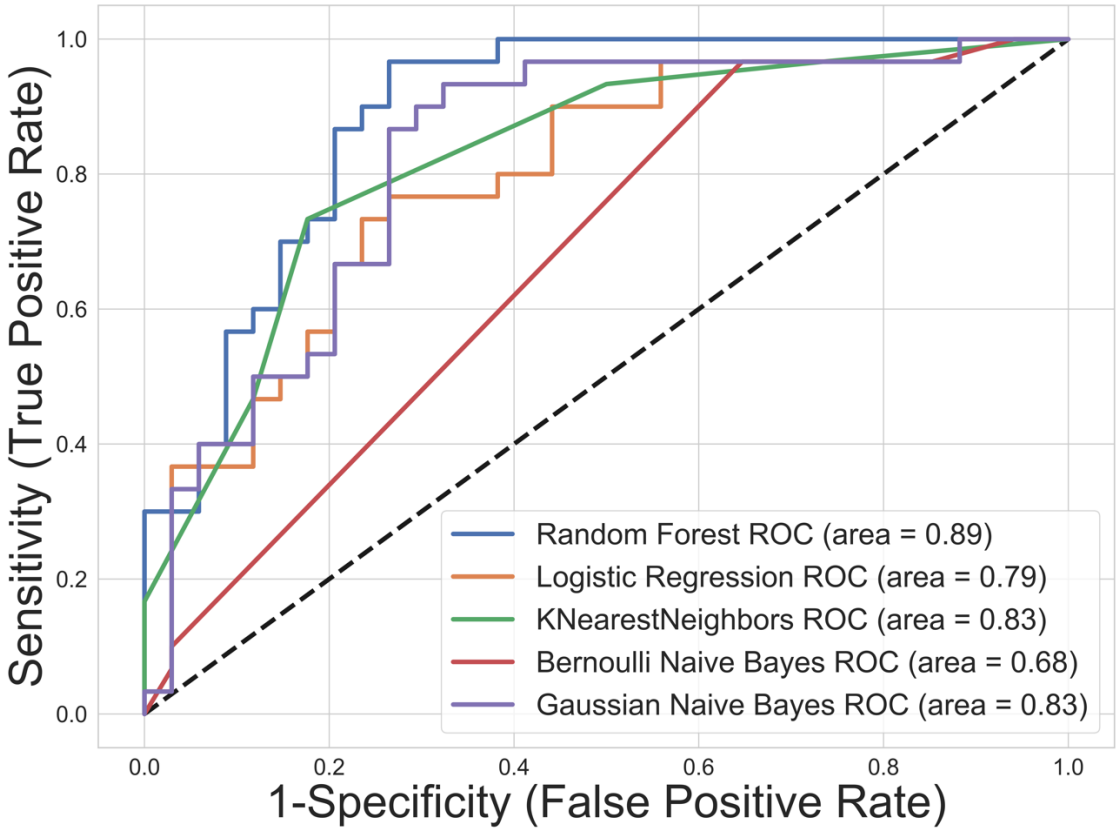
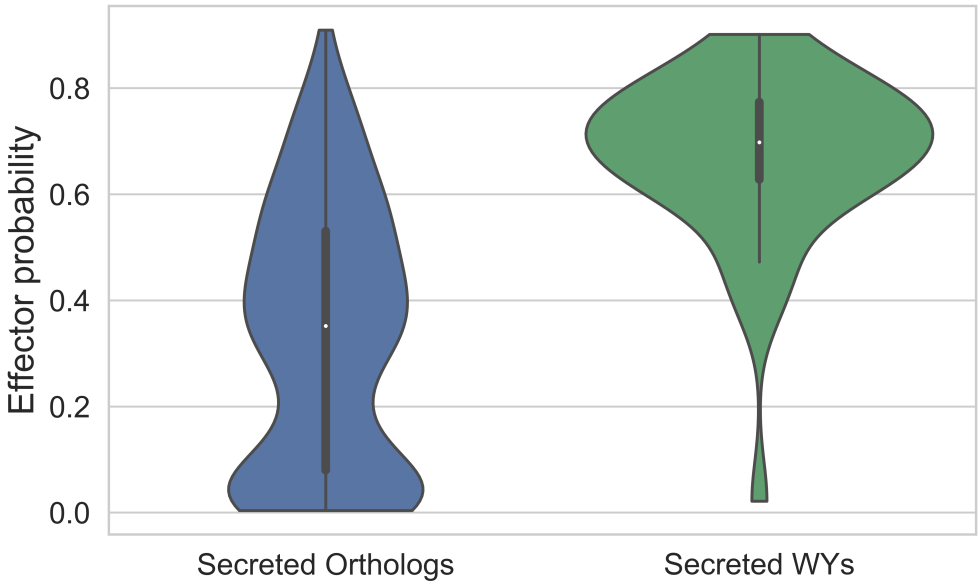


Figure 2. Receiver-operator curves (ROCs) for the various machine learning models tested on a cross validation fold, using 1:1 effector to non-effector testing and training sets.

The Random Forest effector classifier uses a default probability threshold of 0.5 to classify a protein as an effector; this probability threshold can be modified depending on preference for either inclusivity/sensitivity or specificity. For the top performing Random Forest model, we calculated and averaged accuracy, specificity, sensitivity (true positive rate), and false positive rate along the five cross-validation folds for different probability thresholds, illustrating the tradeoffs between rates of false positives and true positives when varying probability thresholds (Table 2). To obtain another estimate of the false positive and false negative rate, we ran EffectorO-ML on a set of 634 predicted secreted proteins that were conserved in at least 24 of the 28 sequenced oomycete species used for EffectorO-LSP

268 (presumed non-effectors) as well as on the 41 secreted *B. lactucae* WY effectors (presumed
269 true effectors). Secreted conserved orthologs and secreted WY proteins had significantly
270 different distributions (Figure 3). We found that a probability cutoff of 0.5 resulted in
271 approximately 28% of these conserved proteins being classified as effectors while capturing
272 90% of the WY effectors (Table 3). At a probability cutoff of 0.6, only 16% of the conserved
273 proteins were classified as effectors, with 85% of the WY proteins classified as effectors
274 (Table 3). Since we presume most of the conserved secreted proteins are not effectors due
275 to lineage specificity, 16% would be a good estimate of the false positive rate at the 0.6
276 threshold, which is similar to the false positive rate for string searches for the RXLR motif
277 (Wood et al. 2020). The false positive rate can be decreased further if desired by increasing
278 the probability threshold, but at the loss of true effectors. For inclusive searches (especially
279 if combined with other lines of evidence), we recommend between 0.5 and 0.6 as the
280 threshold for analyses of oomycete genomes.



281

282 **Figure 3.** Violin plots of effector probabilities for conserved secreted oomycete orthologs
283 and secreted *B. lactucae* WY domain containing proteins, generated by the top-performing
284 Random Forest model.

285 **Table 1. Average cross validation metrics for various machine learning models for**
286 **classifying proteins as effectors or non-effectors**
287
288

	Random Forest	Logistic Regression	K-Nearest Neighbors	Bernoulli Naïve Bayes	Gaussian Naïve Bayes
Accuracy	84.26%	65.90%	70.16%	63.61%	67.87%
Specificity	82.97%	63.31%	69.12%	30.95%	61.14%
Sensitivity	86.95%	68.73%	71.10%	96.09%	74.59%
False Positive Rate	17.03%	36.69%	30.88%	69.05%	38.86%

289

290 **Table 2. Metrics at different probability thresholds for the most accurate Random**
291 **Forest model**
292

	Random Forest Model Effector Probability Threshold =		
	0.7	0.6	0.5
Accuracy	76.23%	79.02%	84.26%
Specificity	89.84%	85.09%	82.97%
Sensitivity	62.74%	73.00%	86.95%
False Positive Rate	10.16%	14.91%	17.03%

293

294 **Table 3. Number and percent of oomycete conserved secreted orthologous proteins**
295 **(634 total) and *B. lactucae* secreted WY proteins (41 total) classified as an effector by**
296 **the EffectorO-ML Random Forest model at different effector probability thresholds**
297

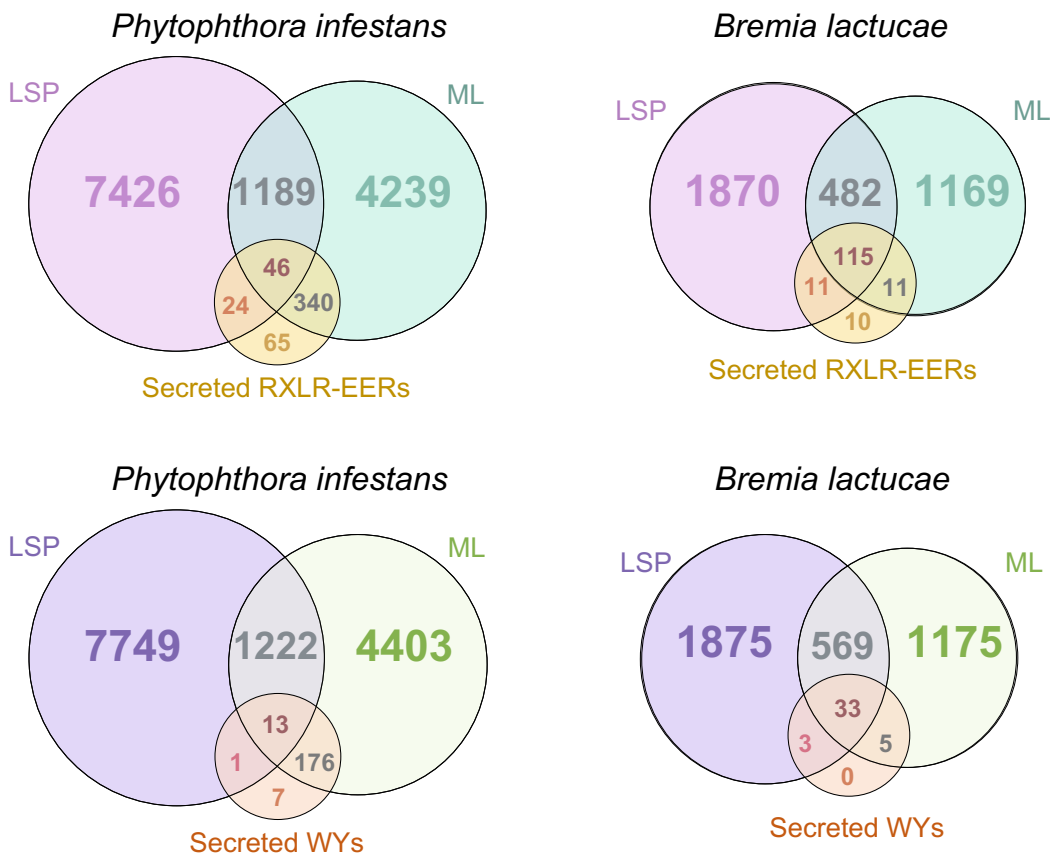
	Random Forest Model Effector Probability Threshold =				
	0.9	0.8	0.7	0.6	0.5
Secreted Orthologs	2 (0.3%)	12 (1.9%)	54 (8.5%)	103 (16%)	180 (28%)
Secreted WY Proteins	1 (2.5%)	5 (12.5%)	20 (50%)	34 (85%)	36 (90%)

298

299

300 *EffectorO predicts novel effectors in oomycete species and recovers known effectors*

301 To find novel effectors from *P. infestans* and *B. lactucae*, we used our EffectorO-ML
302 and LSP pipelines on the predicted secreted ORFs of these genomes. For *P. infestans*, we
303 found 8,685 proteins (41% of the secretome) to be lineage-specific and 5,814 proteins (30%)
304 to be classified as candidate effectors by machine learning, with 1,235 proteins (6%) present
305 in both categories (Figure 4). For *B. lactucae,* we found 2,478 proteins (34% of the
306 secretome) to be lineage-specific and 1,777 proteins (24%) to be classified as candidate
307 effectors by machine learning, with 597 proteins (8%) present in both categories (Figure 4).
308 EffectorO was largely able to recover proteins containing effector signatures, such as the
309 RXLR-EER motif and the WY domain (Figure 4 and Supplemental Table 3), which were not
310 used as classifiers in our analysis but were contained in proteins in the positive training set.

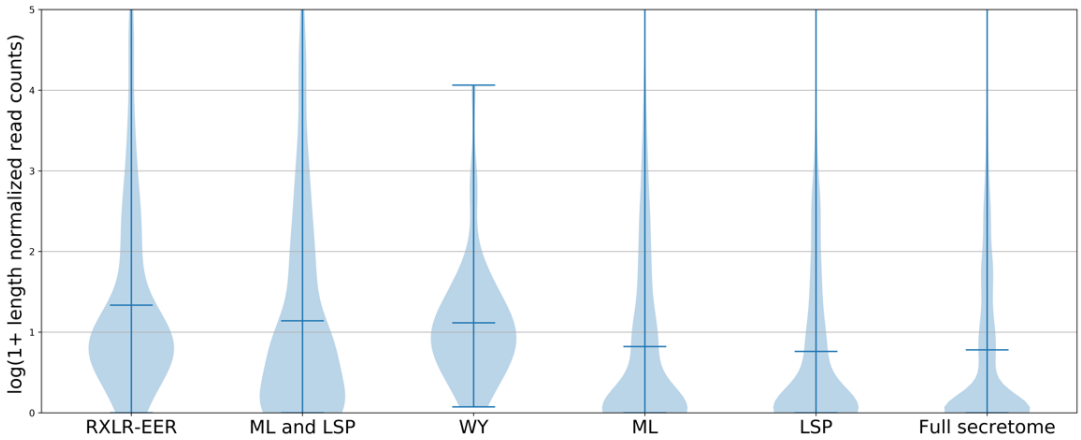


311

312 **Figure 4.** Venn diagrams showing the number of open reading frames (ORFs) identified in
313 the predicted secretomes of *P. infestans* and *B. lactucae* that are lineage specific proteins
314 (LSPs) and/or predicted to be an effector by the machine learning classifier (ML). The
315 number of RXLR-EER and WY proteins that are included or excluded from these datasets are
316 also shown.
317

318 *The majority of EffectorO candidates are expressed in B. lactucae*

319 To investigate whether EffectorO candidates are expressed at the transcript level, we
320 used an available RNA-seq dataset of *B. lactucae* grown on lettuce seedlings (NCBI BioProject
321 PRJNA523226). We found that transcripts that encoded proteins predicted to be effectors by
322 EffectorO-ML and EffectorO-LSP had similar mean expression levels as transcripts encoding
323 proteins containing WY domains and slightly lower expression than those containing RXLR-
324 EER motifs (Figure 5). Transcripts that encoded proteins predicted to be effectors by either
325 the ML or LSP pipeline, but not both, had lower mean expression levels and more proteins
326 with zero reads present in the RNA-seq dataset (Figure 5). The full secretome had the lowest
327 mean expression and also had 1,716 predicted transcripts with zero reads, suggesting that
328 many of the predicted ORFs may not represent real genes (Figure 5).

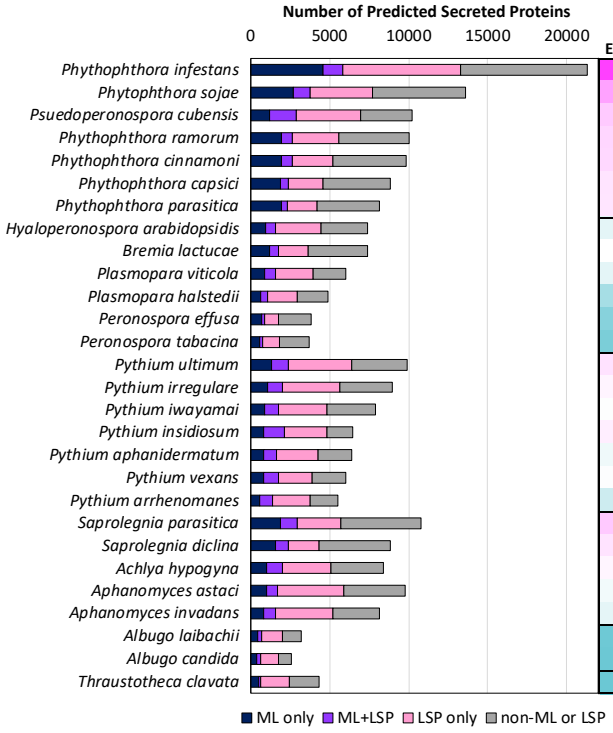


329

330 **Figure 5.** RNA-seq expression levels (log of the length-normalized read counts) for
331 candidate effector transcripts encoding proteins with the RXLR-EER motif (RXLR-EER), WY
332 domain (WY), EffectorO-ML -LSP (ML and LSP), EffectorO-ML only (ML), EffectorO-LSP only
333 (LSP), or the full secretome.

334 *EffectorO-ML predicts candidate effectors in diverse oomycete pathogens*

335 To determine whether EffectorO could predict effectors in other pathogens across the
336 phylum Oomycota, we tested EffectorO-ML on an additional 27 published oomycete
337 genomes. We found that even though EffectorO was trained primarily on effectors from
338 downy mildew and *Phytophthora* pathogens, the algorithm predicted hundreds to thousands
339 of candidate effectors in distantly related oomycetes infecting animals (e.g., *Achlya hypogyna*,
340 *Aphanomyces* spp., *Saprolegnia* spp.), as well as oomycetes from the genera *Albugo* and
341 *Pythium* (Figure 6). Critically, EffectorO-ML predicted a similar number of effectors in the
342 free-living oomycete species *Thraustotheca clavata* (14.6% of the genome) as the false
343 positive rate for EffectorO (17.0%, Table 1).



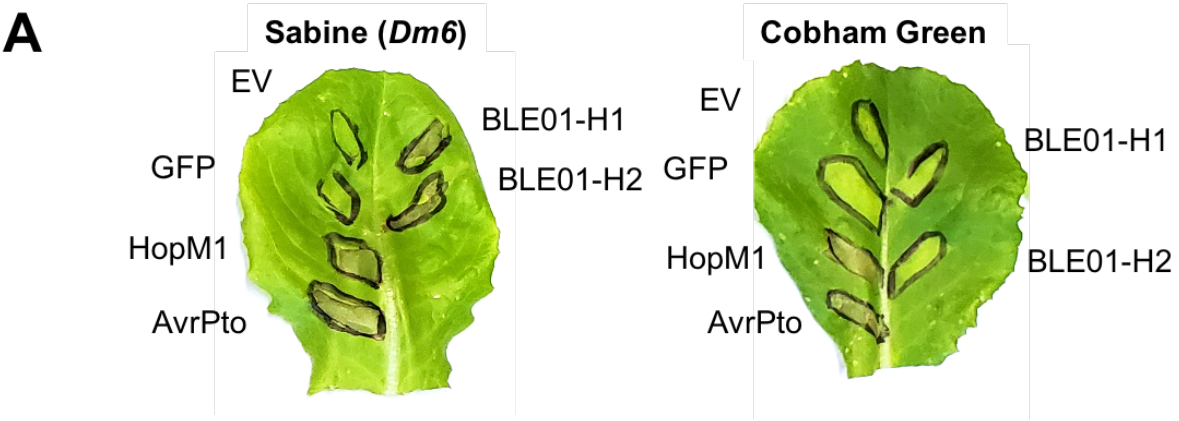
	EffectorO-ML	ML+LSP	LSP	RXLR-EER	WY	Total Secreted
Phythophthora infestans	5814	1235	8685	475	197	21320
Phytophthora sojae	3749	1078	5065	343	118	13605
Psuedoperonospora cubensis	2899	1684	5727	131	48	10192
Phythophthora ramorum	2654	702	3612	270	131	10043
Phythophthora cinnamoni	2601	631	3228	238	96	9847
Phythophthora capsici	2348	494	2698	226	76	8820
Phythophthora parasitica	2328	372	2244	336	144	8160
Hyaloperonospora arabidopsidis	1563	654	3515	98	45	7383
Bremia lactucae	1777	597	2478	147	40	7376
Plasmopara viticola	1562	664	3074	110	214	5997
Plasmopara halstedii	1087	433	2271	58	34	4863
Peronospora effusa	856	146	1025	99	67	3840
Peronospora tabacina	761	206	1247	73	37	3671
Pythium ultimum	2380	1052	5048	87	0	9928
Pythium irregulare	2013	949	4598	71	0	8953
Pythium iwayamai	1751	860	3934	80	0	7907
Pythium insidiosum	2132	1345	4018	72	0	6467
Pythium aphanidermatum	1646	808	3427	50	0	6407
Pythium vexans	1743	907	3072	63	0	6019
Pythium arrhenomanes	1388	810	3201	34	0	5523
Saprolegnia parasitica	2922	1057	3819	112	0	10767
Saprolegnia diclina	2365	793	2767	63	0	8835
Achlya hypogyna	1984	1006	4118	52	0	8370
Aphanomyces astaci	1667	673	4887	44	0	9747
Aphanomyces invadans	1569	781	4381	33	0	8175
Albugo laibachii	664	235	1600	20	0	3200
Albugo candida	643	295	1399	16	0	2572
Thraustotheca clavata	635	127	1907	11	0	4352

344

345 **Figure 6.** Number of predicted secreted proteins (translated ORFs) classified as effectors
346 using EffectorO-ML only (cutoff = 0.5) (dark blue bars), EffectorO-LSP only (pink bars),
347 EffectorO-ML and LSP (purple bars), and those classified as non-effectors (grey bars) in 28
348 species of oomycetes, including Peronosporales (Phytophthora and downy mildew species),
349 *Pythium* species, animal pathogens (*Achlya, Saprolegnia*, and *Aphanomyces* spp.), *Albugo*

350 spp., and a free-living oomycete (*Thraustotheca clavata*). Numbers are estimated by
351 subtracting the number of EffectorO-ML hits from the false positive rate for each genome
352 (14.6%, estimated using *T. clavata*, which had fewer predicted effectors than the estimated
353 false positive rate of 17.0%).
354
355
356 *An EffectorO candidate causes cell death when expressed in lettuce*

357 To see whether any EffectorO candidates are recognized effectors (i.e., have
358 avirulence activity), we cloned a gene from *B. lactucae* that encoded a protein that does not
359 have RXLR or EER motifs nor predicted WY domains, but was classified as a probable effector
360 by both EffectorO-ML and Effector-LSP. The gene encoding this protein was not found in the
361 gene models (i.e. was not predicted to be a gene by the annotation software), but did encode
362 a predicted ORF. The candidate effector, named BLE01 for (*Bremia lactucae* EffectorO
363 candidate 1), was within the mapping region for *Avr6,* as determined by sequencing of
364 segregating sexual progeny of *B. lactucae* isolates (R.J. Gil, unpublished data). We found that
365 transient expression of *BLE01* caused cell death on the lettuce cultivar Sabine, which
366 contains *Dm6*, but not Cobham Green, which does not contain any known *Dm* genes (Fig 7A).
367 When we screened a larger panel of lettuce cultivars with various *Dm* genes, we found that
368 it also elicited cell death in RYZ2164, but not in the other seven cultivars with other *Dm* genes
369 or in the non-host plant *Nicotiana benthamiana* (Fig 7B).



B

Cultivar	*Dm* genes	N plants	EV	GFP	BLE01-H1	BLE01-H2	HopM1	AvrPto
N. benthamiana	n/a	3	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	4.0 ± 0.0	0.0 ± 0.0
Cobham Green	n/a	4	0.5 ± 0.6	0.75 ± 0.5	0.5 ± 0.6	0.5 ± 0.6	4.0 ± 0.0	3.8 ± 0.5
UCDM2	*Dm2*	4	0.5 ± 0.6	0.3 ± 0.5	0.0 ± 0.0	0.3 ± 0.5	3.8 ± 0.5	3.5 ± 0.6
Dandie	*Dm3*	4	0.5 ± 0.6	0.5 ± 0.6	0.3 ± 0.5	0.3 ± 0.5	4.0 ± 0.0	3.3 ± 0.6
Sabine	*Dm6*	6	0.8 ± 0.4	0.8 ± 0.4	2.7 ± 0.8	3.5 ± 0.6	3.7 ± 0.5	3.0 ± 0.0
LSE57/15	*Dm7*	4	0.5 ± 0.6	0.5 ± 0.6	0.3 ± 0.5	0.3 ± 0.5	3.3 ± 1.5	3.8 ± 0.5
Capitan	*Dm11*	4	1.0 ± 0.8	1.0 ± 0.8	0.8 ± 1.0	0.8 ± 0.5	3.3 ± 0.5	2.8 ± 0.5
Ninja	*Dm3,4,11,36*	4	0.3 ± 0.5	0.3 ± 0.5	0.3 ± 0.5	0.3 ± 0.5	4.0 ± 0.0	4.0 ± 0.0
Discovery	*Dm7, 37*	4	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	3.8 ± 0.5	3.5 ± 0.6
RYZ2164	*Dm25*	4	0.8 ± 1.0	0.5 ± 0.6	3.0 ± 0.0	3.3 ± 0.5	4.0 ± 0.0	2.5 ± 1.7
ViAE	R20	4	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	3.5 ± 0.6	3.3 ± 0.5

Figure 7. (A) Example transient expression results from lettuce infiltrated with *A. tumefaciens* containing two candidate effector alleles from *B. lactucae* (BLE01-H1 and BLE01-H2) as well as with the negative controls empty vector (EV) and green fluorescent protein (GFP) and the positive controls for cell death, *Pseudomonas syringae* pv. *tomato* effectors HopM1 and AvrPto. Leaves are representative of four replicates and photos were taken 4 days after infiltration. **(B)** Necrosis scores four days post-agroinfiltration of the candidate effectors and controls in a panel of lettuce cultivars and the non-host *Nicotiana benthamiana*. Necrosis was scored as follows: 0 = no necrosis or chlorosis, 1 = mild chlorosis, 2 = chlorosis, 3 = chlorosis and necrosis, 4 = full necrosis. Number of replicates (N), averages, and standard deviations are shown for each line and effector combination.

Discussion

We demonstrated that our EffectorO pipeline can predict novel candidate effectors using a machine learning-based approach combined with a lineage-specificity criterion for oomycete effector classification. Therefore, EffectorO can be used to mine oomycete

386 genomes to expand the set of predicted effectors without relying on canonical N-terminal
387 motifs (such as RXLR and EER) or C-terminal domains (such as the WY domain); thus, it may
388 reveal effectors that were previously missed by only searching for these features. EffectorO
389 predicts an additional ~600 ML+LSP candidate effectors from *B. lactucae* and ~1,200
390 additional ML+LSP candidate effectors from *P. infestans*. The percent of effectors predicted
391 in the secretome by EffectorO-ML was approximately 25%, which is similar to what was
392 predicted for fungal secretomes with EffectorP (Sperschneider et al. 2018). The number of
393 potentially novel effectors in *B. lactucae* and *P. infestans* that were identified is much larger
394 than previous reports of effector repertoires from these two species (Fletcher et al. 2019;
395 Haas et al. 2009). The large size of these candidate sets should not preclude functional
396 analysis when combined with additional data from the pathogen of interest such as
397 avirulence gene mapping, population genomics, and gene expression data.

398 EffectorO-ML was trained on the N-terminal sequences of validated effectors
399 compared to conserved secreted proteins (presumed non-effectors), which we show differ
400 in several biochemical characteristics and especially in the degree of N-terminal disorder.
401 Sequences in the N-terminus of effector proteins are known to be important for secretion
402 and translocation of effectors; here we demonstrate that they have distinct biochemical
403 signatures that can be used to predict the probability that a given sequence is an effector.
404 EffectorO-ML was able to recover the majority of predicted RXLR-EER and WY domain
405 containing proteins from *B. lactucae* and *P. infestans*, without relying on using those motifs
406 for prediction. By focusing on the N-terminus for effector classification, our pipeline allows
407 for a diversity of C-terminal domains, which are typically considered to be the functional
408 region of effector proteins. EffectorO may help to uncover new effector families, which can

409 then be bioinformatically characterized using motif/domain-based analysis, as well as
410 structurally characterized using cryo-EM or x-ray crystallography of effector proteins, as has
411 been done for WY-containing effectors (Boutemy et al. 2011). The predicted effector
412 repertoires should also be analyzed with the latest ML structural prediction algorithms such
413 as AlphaFold (Senior et al. 2020) and trRosetta (Yang et al. 2020). Recently, trRosetta was
414 used to predict effectors based on structure for the fungal species *Magnaporthe oryzae*
415 (Seong and Krasileva 2021) and a similar analysis should be done with EffectorO candidates.
416 The most accurate model for EffectorO-ML was the Random Forest classifier. Random
417 Forest classifiers work by building many decision trees and then taking the consensus of all
418 of the decision trees to calculate an overall probability for a given classification (Pavlov
419 2000). Because the final output is the average of many independent models, Random Forest
420 classifiers are robust to individual errors in any given single model. The fungal effector
421 classifier, EffectorP 2.0, uses a similar approach by utilizing an ensemble classifier that takes
422 the consensus predictions of multiple Naive-Bayes and decision trees to obtain an accuracy
423 of 89% for predicting fungal effectors. While EffectorP is trained on a larger set of effectors
424 than EffectorO and relies on additional biochemical characteristics (16 features vs. our six),
425 we achieved a similar accuracy to EffectorP 2.0 in predicting oomycete effectors. Our
426 Random Forest model had an estimated 84% accuracy and 82% specificity at a probability
427 cutoff of 0.5. The false positive rate can be decreased by increasing the probability cutoff for
428 classification of an effector; however, this will also result in more false negatives (missed
429 true effectors). We recommend 0.5 as an inclusive threshold for casting a wide net to catch
430 effectors; however, depending on the application the researcher may want to use a more
431 stringent threshold to find higher confidence effectors. It may also be possible to further

432 improve the accuracy and specificity of EffectorO by adding additional proteins to the
433 training set or by adding more biochemical and structural characteristics. Gene models may
434 be used instead of ORFs to reduce false positives, although gene prediction algorithms may
435 miss true effectors (as was the case for the *Avr6* candidate *BLE01*).

436 Similar to EffectorO-ML, EffectorO-LSP will predict additional proteins besides
437 effectors, not only due to false positives but also due to the biological reality that there may
438 be lineage specific proteins that fulfill other functions. Thus, EffectorO-LSP predictions are
439 most useful in combination with other lines of evidence of effector activity (such as
440 EffectorO-ML, expression during infection, RXLR motifs and/or WY domains).
441 Approximately 41% of the *P. infestans* secretome and 34% of the *B. lactucae* secretome were
442 predicted to be lineage specific. Some of these predicted proteins may not actually be
443 lineage-specific; some may be encoded by pseudogenes or result from failure to detect
444 homology (Weisman et al. 2020). We found that many were expressed in *B. lactucae* during
445 infection, suggesting that some of these represent true genes. In addition to effectors, these
446 lineage-specific secreted proteins could also be related to the lifestyles of these organisms,
447 their sexual reproduction, or may be products of co-evolution with other microbes rather
448 than co-evolution with the host. The amount of lineage specificity expected for a given
449 organism will depend on whether or not closely related organisms are used in the analysis
450 and how closely pathogens are adapted to their hosts. Depending on the relatedness of a
451 given species to those in the database, we recommend that BLAST results from very closely
452 related species be filtered out in order to find proteins that are truly lineage specific.

453 EffectorO predictions were validated for an effector from *B. lactucae* that was within
454 the region co-segregating with *Avr6* but did not have RXLR or EER motifs nor any WY

455 domains. Here we have shown that this effector, *BLE01*, causes cell death when expressed in
456 lettuce with the *Dm6* gene but not on cultivars containing other known *Dm* genes, making it
457 a good candidate for the protein encoded by *Avr6.* Expression of *BLE01* also caused cell death
458 on RYZ2164. The background of RYZ2164 is proprietary, so it may have *Dm6* in its lineage
459 or it may contain another *Dm* gene that recognizes the effector.

460 EffectorO-ML and -LSP can expand the set of predicted effectors for identification of
461 genes for resistance to agriculturally important oomycetes. These predictions can be used to
462 choose candidates for "effectoromics" (Vleeshouwers and Oliver 2015) screens to identify
463 effectors with avirulence and virulence activities. EffectorO predictions will be especially
464 useful when combined with mapping of avirulence phenotypes for the cloning of specific
465 avirulence genes enabling an effector-driven approach to R gene identification.

466

467 **Acknowledgements**

468 The authors would like to thank John Emerson for greenhouse support, Elizabeth
469 Georgian for manuscript editing and formatting, Kyle Fletcher for very helpful comments and
470 suggestions, and Mary Corrigan for additional minor editing.

471

472 **Literature Cited**

473 Bailey, K., Cevik, V., Holton, N., Byrne-Richardson, J., Sohn, K. H., Coates, M., Woods-Tör, A.,
474 Aksoy, H. M., Hughes, L., Baxter, L., Jones, J. D. G., Beynon, J., Holub, E. B., and Tör, M.
475 2011. Molecular cloning of ATR5(Emoy2) from *Hyaloperonospora arabidopsidis*, an
476 avirulence determinant that triggers RPP5-mediated defense in Arabidopsis.
477 Molecular Plant-Microbe Interactions. 24:827–838

478 Baldauf, S. L. 2003. The Deep Roots of Eukaryotes. Science. 300:1703–1706

479 Baxter, L., Tripathy, S., Ishaque, N., Boot, N., Cabral, A., Kemen, E., Thines, M., Ah-Fong, A.,
480 Anderson, R., Badejoko, W., Bittner-Eddy, P., Boore, J. L., Chibucos, M. C., Coates, M.,
481 Dehal, P., Delehaunty, K., Dong, S., Downton, P., Dumas, B., Fabro, G., Fronick, C.,
482 Fuerstenberg, S. I., Fulton, L., Gaulin, E., Govers, F., Hughes, L., Humphray, S., Jiang, R.
483 H. Y., Judelson, H., Kamoun, S., Kyung, K., Meijer, H., Minx, P., Morris, P., Nelson, J.,
484 Phuntumart, V., Qutob, D., Rehmany, A., Rougon-Cardoso, A., Ryden, P., Torto-
485 Alalibo, T., Studholme, D., Wang, Y., Win, J., Wood, J., Clifton, S. W., Rogers, J., Van den
486 Ackerveken, G., Jones, J. D. G., McDowell, J. M., Beynon, J., and Tyler, B. M. 2010.
487 Signatures of adaptation to obligate biotrophy in the *Hyaloperonospora*
488 *arabidopsidis* genome. Science. 330:1549–1551

489 Birch, P. R. J., Armstrong, M., Bos, J., Boevink, P., Gilroy, E. M., Taylor, R. M., Wawra, S.,
490 Pritchard, L., Conti, L., Ewan, R., Whisson, S. C., van West, P., Sadanandom, A., and
491 Kamoun, S. 2009. Towards understanding the virulence functions of RXLR effectors
492 of the oomycete plant pathogen *Phytophthora infestans*. Journal of Experimental
493 Botany. 60:1133–1140

494 Bos, J. I. B., Armstrong, M. R., Gilroy, E. M., Boevink, P. C., Hein, I., Taylor, R. M., Zhendong, T.,
495 Engelhardt, S., Vetukuri, R. R., Harrower, B., Dixelius, C., Bryan, G., Sadanandom, A.,
496 Whisson, S. C., Kamoun, S., and Birch, P. R. J. 2010. *Phytophthora infestans* effector
497 AVR3a is essential for virulence and manipulates plant immunity by stabilizing host
498 E3 ligase CMPG1. Proceedings of the National Academy of Sciences. 107:9909–9914

499 Boutemy, L. S., King, S. R. F., Win, J., Hughes, R. K., Clarke, T. A., Blumenschein, T. M. A.,
500 Kamoun, S., and Banfield, M. J. 2011. Structures of *Phytophthora* RXLR effector

501 proteins: a conserved but adaptable fold underpins functional diversity. Journal of
502 Biological Chemistry. 286:35834–35842
503 Deb, D., Anderson, R. G., How-Yew-Kin, T., Tyler, B. M., and McDowell, J. M. 2018. Conserved
504 RxLR Effectors From Oomycetes *Hyaloperonospora arabidopsidis* and *Phytophthora*
505 *sojae* Suppress PAMP- and Effector-Triggered Immunity in Diverse Plants. Molecular
506 Plant-Microbe Interactions. 31:374–385
507 Derevnina, L., Petre, B., Kellner, R., Dagdas, Y. F., Sarowar, M. N., Giannakopoulou, A., De la
508 Concepcion, J. C., Chaparro-Garcia, A., Pennington, H. G., van West, P., and Kamoun, S.
509 2016. Emerging oomycete threats to plants and animals. Philos. Trans. R. Soc. Lond.
510 B Biol. Sci. 371
511 Dong, S., Raffaele, S., and Kamoun, S. 2015. The two-speed genomes of filamentous
512 pathogens: waltz with plants. Current Opinion in Genetics & Development. 35:57–65
513 Dou, D., Kale, S. D., Wang, X., Jiang, R. H. Y., Bruce, N. A., Arredondo, F. D., Zhang, X., and
514 Tyler, B. M. 2008. RXLR-mediated entry of *Phytophthora sojae* effector Avr1b into
515 soybean cells does not require pathogen-encoded machinery. Plant Cell. 20:1930–
516 1947
517 Eddy, S. R. 2011. Accelerated Profile HMM Searches. PLoS Comput. Biol. 7:e1002195
518 Ellis, J. G., and Dodds, P. N. 2011. Showdown at the RXLR motif: Serious differences of
519 opinion in how effector proteins from filamentous eukaryotic pathogens enter plant
520 cells. Proceedings of the National Academy of Sciences. 108:14381–14382
521 Emms, D. M., and Kelly, S. 2015. OrthoFinder: solving fundamental biases in whole genome
522 comparisons dramatically improves orthogroup inference accuracy. Genome
523 Biology. 16:157

524 Emms, D. M., and Kelly, S. 2019. OrthoFinder: phylogenetic orthology inference for
525 comparative genomics. Genome Biology. 20:238
526 ExPaSy (Expert Protein Analysis System). 2004. Encyclopedic Dictionary of Genetics,
527 Genomics and Proteomics.
528 Fabro, G., Steinbrenner, J., Coates, M., Ishaque, N., Baxter, L., Studholme, D. J., Körner, E.,
529 Allen, R. L., Piquerez, S. J. M., Rougon-Cardoso, A., Greenshields, D., Lei, R., Badel, J. L.,
530 Caillaud, M.-C., Sohn, K.-H., Van den Ackerveken, G., Parker, J. E., Beynon, J., and
531 Jones, J. D. G. 2011. Multiple Candidate Effectors from the Oomycete Pathogen
532 *Hyaloperonospora arabidopsidis* Suppress Host Plant Immunity. PLOS Pathogens.
533 7:e1002348
534 Fletcher, K., Gil, J., Bertier, L. D., Kenefick, A., Wood, K. J., Zhang, L., Reyes-Chin-Wo, S.,
535 Cavanaugh, K., Tsuchida, C., Wong, J., and Michelmore, R. 2019. Genomic signatures
536 of heterokaryosis in the oomycete pathogen *Bremia lactucae*. Nature
537 Communications. 10:2645
538 Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. 2012. CD-HIT: accelerated for clustering the next-
539 generation sequencing data. Bioinformatics. 28:3150–3152
540 Giesbers, A. K. J., Pelgrom, A. J. E., Visser, R. G. F., Niks, R. E., Van den Ackerveken, G., and
541 Jeuken, M. J. W. 2017. Effector-mediated discovery of a novel resistance gene against
542 *Bremia lactucae* in a nonhost lettuce species. New Phytologist. 216:915–926
543 Godfrey, D., Böhlenius, H., Pedersen, C., Zhang, Z., Emmersen, J., and Thordal-Christensen, H.
544 2010. Powdery mildew fungal effector candidates share N-terminal Y/F/WxC-motif.
545 BMC Genomics. 11:317

546 Gu, B., Kale, S. D., Wang, Q., Wang, D., Pan, Q., Cao, H., Meng, Y., Kang, Z., Tyler, B. M., and
547 Shan, W. 2011. Rust Secreted Protein Ps87 Is Conserved in Diverse Fungal
548 Pathogens and Contains a RXLR-like Motif Sufficient for Translocation into Plant
549 Cells. PLOS ONE. 6:e27217
550 Haas, B. J., Kamoun, S., Zody, M. C., Jiang, R. H. Y., Handsaker, R. E., Cano, L. M., Grabherr, M.,
551 Kodira, C. D., Raffaele, S., Torto-Alalibo, T., Bozkurt, T. O., Ah-Fong, A. M. V., Alvarado,
552 L., Anderson, V. L., Armstrong, M. R., Avrova, A., Baxter, L., Beynon, J., Boevink, P. C.,
553 Bollmann, S. R., Bos, J. I. B., Bulone, V., Cai, G., Cakir, C., Carrington, J. C., Chawner, M.,
554 Conti, L., Costanzo, S., Ewan, R., Fahlgren, N., Fischbach, M. A., Fugelstad, J., Gilroy, E.
555 M., Gnerre, S., Green, P. J., Grenville-Briggs, L. J., Griffith, J., Grünwald, N. J., Horn, K.,
556 Horner, N. R., Hu, C.-H., Huitema, E., Jeong, D.-H., Jones, A. M. E., Jones, J. D. G., Jones,
557 R. W., Karlsson, E. K., Kunjeti, S. G., Lamour, K., Liu, Z., Ma, L., Maclean, D., Chibucos,
558 M. C., McDonald, H., McWalters, J., Meijer, H. J. G., Morgan, W., Morris, P. F., Munro, C.
559 A., O'Neill, K., Ospina-Giraldo, M., Pinzón, A., Pritchard, L., Ramsahoye, B., Ren, Q.,
560 Restrepo, S., Roy, S., Sadanandom, A., Savidor, A., Schornack, S., Schwartz, D. C.,
561 Schumann, U. D., Schwessinger, B., Seyer, L., Sharpe, T., Silvar, C., Song, J., Studholme,
562 D. J., Sykes, S., Thines, M., van de Vondervoort, P. J. I., Phuntumart, V., Wawra, S.,
563 Weide, R., Win, J., Young, C., Zhou, S., Fry, W., Meyers, B. C., van West, P., Ristaino, J.,
564 Govers, F., Birch, P. R. J., Whisson, S. C., Judelson, H. S., and Nusbaum, C. 2009.
565 Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora*
566 *infestans*. Nature. 461:393–398
567 Haverkort, A. J., Boonekamp, P. M., Hutten, R., Jacobsen, E., Lotz, L. A. P., Kessel, G. J. T.,
568 Visser, R. G. F., and van der Vossen, E. A. G. 2008. Societal Costs of Late Blight in

569 Potato and Prospects of Durable Resistance Through Cisgenic Modification. Potato
570 Research. 51:47–57
571 He, J., Ye, W., Choi, D. S., Wu, B., Zhai, Y., Guo, B., Duan, S., Wang, Y., Gan, J., Ma, W., and Ma, J.
572 2019. Structural analysis of *Phytophthora* suppressor of RNA silencing 2 (PSR2)
573 reveals a conserved modular fold contributing to virulence. Proceedings of the
574 National Academy of Sciences. 116:8054–8059
575 Jiang, R. H. Y., de Bruijn, I., Haas, B. J., Belmonte, R., Löbach, L., Christie, J., van den
576 Ackerveken, G., Bottin, A., Bulone, V., Díaz-Moreno, S. M., Dumas, B., Fan, L., Gaulin,
577 E., Govers, F., Grenville-Briggs, L. J., Horner, N. R., Levin, J. Z., Mammella, M., Meijer,
578 H. J. G., Morris, P., Nusbaum, C., Oome, S., Phillips, A. J., van Rooyen, D., Rzeszutek, E.,
579 Saraiva, M., Secombes, C. J., Seidl, M. F., Snel, B., Stassen, J. H. M., Sykes, S., Tripathy,
580 S., van den Berg, H., Vega-Arreguin, J. C., Wawra, S., Young, S. K., Zeng, Q., Dieguez-
581 Uribeondo, J., Russ, C., Tyler, B. M., and van West, P. 2013. Distinctive expansion of
582 potential virulence genes in the genome of the oomycete fish pathogen *Saprolegnia*
583 *parasitica.* PLOS Genetics. 9:e1003272
584 Kale, S. D., Gu, B., Capelluto, D. G. S., Dou, D., Feldman, E., Rumore, A., Arredondo, F. D.,
585 Hanlon, R., Fudal, I., Rouxel, T., Lawrence, C. B., Shan, W., and Tyler, B. M. 2010.
586 External Lipid PI3P Mediates Entry of Eukaryotic Pathogen Effectors into Plant and
587 Animal Host Cells. Cell. 142:284–295
588 King, S. R. F., McLellan, H., Boevink, P. C., Armstrong, M. R., Bukharova, T., Sukarta, O., Win,
589 J., Kamoun, S., Birch, P. R. J., and Banfield, M. J. 2014. *Phytophthora infestans* RXLR
590 effector PexRD2 interacts with host MAPKKK ε to suppress plant immune signaling.
591 Plant Cell. 26:1345–1359

592 Li, W., and Godzik, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of
593 protein or nucleotide sequences. Bioinformatics. 22:1658–1659
594 McGowan, J., and Fitzpatrick, D. A. 2017. Genomic, Network, and Phylogenetic Analysis of
595 the Oomycete Effector Arsenal. mSphere. 2
596 Pavlov, Y. L. 2000. Random Forests. VSP Publishers, Utrecht, the Netherlands.
597 Pecrix, Y., Buendia, L., Penouilh-Suzette, C., Maréchaux, M., Legrand, L., Bouchez, O., Rengel,
598 D., Gouzy, J., Cottret, L., Vear, F., and Godiard, L. 2019. Sunflower resistance to
599 multiple downy mildew pathotypes revealed by recognition of conserved effectors
600 of the oomycete *Plasmopara halstedii*. Plant J. 97:730–748
601 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel, M.,
602 Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D.,
603 Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine Learning in Python. J
604 Mach Learn Res. 2011;12: 2825–2830.
605 Pelgrom, A. J. E., Eikelhof, J., Elberse, J., Meisrimler, C.-N., Raedts, R., Klein, J., and Van den
606 Ackerveken, G. 2019. Recognition of lettuce downy mildew effector BLR38 in
607 *Lactuca serriola* LS102 requires two unlinked loci. Molecular Plant Pathology.
608 20:240–253
609 Pel, M. A., Foster, S. J., Park, T.-H., Rietman, H., van Arkel, G., Jones, J. D. G., Van Eck, H. J.,
610 Jacobsen, E., Visser, R. G. F., and Van der Vossen, E. A. G. 2009. Mapping and cloning
611 of late blight resistance genes from *Solanum venturii* using an interspecific candidate
612 gene approach. Molecular Plant-Microbe Interactions. 22:601–615
613 Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K., and Obradovic, Z. 2006. Length-dependent
614 prediction of protein intrinsic disorder. BMC Bioinformatics. 7:208

615 Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. 2011. SignalP 4.0: discriminating
616 signal peptides from transmembrane regions. Nat. Methods. 8:785–786
617 Rehmany, A. P., Gordon, A., Rose, L. E., Allen, R. L., Armstrong, M. R., Whisson, S. C., Kamoun,
618 S., Tyler, B. M., Birch, P. R. J., and Beynon, J. L. 2005. Differential recognition of highly
619 divergent downy mildew avirulence gene alleles by RPP1 resistance genes from two
620 Arabidopsis lines. Plant Cell. 17:1839–1850
621 Rujirawat, T., Patumcharoenpol, P., Kittichotirat, W., and Krajaejun, T. 2019. Oomycete
622 Gene Table: an online database for comparative genomic analyses of the oomycete
623 microorganisms. Database. 2019
624 Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson,
625 A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P.,
626 Jones, D. T., Silver, D., Kavukcuoglu, K., and Hassabis, D. 2020. Improved protein
627 structure prediction using potentials from deep learning. Nature. 577:706–710
628 Seong, K., and Krasileva, K. V. 2021. Computational structural genomics unravels common
629 folds and predicted functions in the secretome of fungal phytopathogen
630 *Magnaporthe oryzae*. bioRxiv doi: 10.1101/2021.01.25.427855
631 Sharma, R., Xia, X., Cano, L. M., Evangelisti, E., Kemen, E., Judelson, H., Oome, S., Sambles, C.,
632 van den Hoogen, D. J., Kitner, M., Klein, J., Meijer, H. J. G., Spring, O., Win, J., Zipper, R.,
633 Bode, H. B., Govers, F., Kamoun, S., Schornack, S., Studholme, D. J., Van den
634 Ackerveken, G., and Thines, M. 2015. Genome analyses of the sunflower pathogen
635 *Plasmopara halstedii* provide insights into effector evolution in downy mildews and
636 *Phytophthora*. BMC Genomics. 16:741

637 Sperschneider, J., Dodds, P. N., Gardiner, D. M., Singh, K. B., and Taylor, J. M. 2018. Improved
638 prediction of fungal effector proteins from secretomes with EffectorP 2.0. Molecular
639 Plant Pathology. 19:2094–2110
640 Sperschneider, J., Gardiner, D. M., Dodds, P. N., Tini, F., Covarelli, L., Singh, K. B., Manners, J.
641 M., and Taylor, J. M. 2016. EffectorP: predicting fungal effector proteins from
642 secretomes using machine learning. New Phytol. 210:743–761
643 Stassen, J. H. M., den Boer, E., Vergeer, P. W. J., Andel, A., Ellendorff, U., Pelgrom, K., Pel, M.,
644 Schut, J., Zonneveld, O., Jeuken, M. J. W., and Van den Ackerveken, G. 2013. Specific In
645 Planta Recognition of Two GKLR Proteins of the Downy Mildew *Bremia lactucae*
646 Revealed in a Large Effector Screen in Lettuce. Molecular Plant-Microbe
647 Interactions. 26:1259–1270
648 Urban, M., Cuzick, A., Seager, J., Wood, V., Rutherford, K., Venkatesh, S. Y., De Silva, N.,
649 Martinez, M. C., Pedro, H., Yates, A. D., Hassani-Pak, K., and Hammond-Kosack, K. E.
650 2019. PHI-base: the pathogen–host interactions database. Nucleic Acids Research.
651 Vleeshouwers, V. G. A. A., and Oliver, R. P. 2015. Effectors as Tools in Disease Resistance
652 Breeding Against Biotrophic, Hemibiotrophic, and Necrotrophic Plant Pathogens.
653 Molecular Plant-Microbe Interactions. 2015:17–27
654 Wang, Y., Tyler, B. M., and Wang, Y. 2019. Defense and Counterdefense During Plant-
655 Pathogenic Oomycete Infection. Annu. Rev. Microbiol. 73:667–696
656 Weisman, C. M., Murray, A. W., and Eddy, S. R. 2020. Many but not all lineage-specific genes
657 can be explained by homology detection failure. PLOS Biology 18(11):e3000862.
658 Whisson, S. C., Boevink, P. C., Moleleki, L., Avrova, A. O., Morales, J. G., Gilroy, E. M.,
659 Armstrong, M. R., Grouffaud, S., van West, P., Chapman, S., Hein, I., Toth, I. K.,

660 Pritchard, L., and Birch, P. R. J. 2007. A translocation signal for delivery of oomycete
661 effector proteins into host plant cells. Nature. 450:115–118
662 Win, J., Krasileva, K. V., Kamoun, S., Shirasu, K., Staskawicz, B. J., and Banfield, M. J. 2012.
663 Sequence Divergent RXLR Effectors Share a Structural Fold Conserved across Plant
664 Pathogenic Oomycete Species. PLOS Pathogens. 8:e1002400
665 Wood, K., Nur, M., Gil, J., Fletcher, K., Lakeman, K., Gothberg, A., Khuu, T., Kopetzky, J.,
666 Pandya, A., Pel, M., and Michelmore, R. 2020. Effector prediction and
667 characterization in the oomycete pathogen *Bremia lactucae* reveal host-recognized
668 WY domain proteins that lack the canonical RXLR motif. PLOS Pathogens 10:
669 e1009012
670 Wroblewski T., Caldwell K .S., Piskurewicz U., Cavanaugh K. A., Xu H., Kozik A., Ochoa O.,
671 McHale L. K., Lahre K., Jelenska J., Castillo J. A., Blumenthal D., Vinatzer B. A.,
672 Greenberg J. T., and Michelmore R. W. 2009. Comparative large-scale analysis of
673 interactions between several crop species and the effector repertoires from multiple
674 pathovars of *Pseudomonas* and *Ralstonia*. Plant Physiology. 150: 1733–49.
675 doi:10.1104/pp.109.140251
676 Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. 2020. Improved
677 protein structure prediction using predicted interresidue orientations. Proceedings
678 of the National Academy of Sciences. 117:1496–1503
679 Zheng, X., McLellan, H., Fraiture, M., Liu, X., Boevink, P. C., Gilroy, E. M., Chen, Y., Kandel, K.,
680 Sessa, G., Birch, P. R. J., and Brunner, F. 2014. Functionally Redundant RXLR
681 Effectors from *Phytophthora infestans* Act at Different Steps to Suppress Early flg22-
682 Triggered Immunity. PLOS Pathogens. 10:e1004057

683 Zhou, K., Huang, B., Zou, M., Lu, D., He, S., and Wang, G. 2015. Data in support of genome-
684 wide identification of lineage-specific genes within *Caenorhabditis elegans*. Data in
685 Brief. 4:595–601