



## EFFECTORANNOTATOR, A SCRIPT FOR AN EASY AND AUTOMATED IDENTIFICATION AND COMPARISON OF THE TYPE III SECRETION SYSTEM EFFECTORS OF PSEUDOMONAS SYRINGAE

Simone Marcelletti<sup>1\*</sup>, Marco Scortichini<sup>1,2</sup>, Giuseppe Firrao<sup>3</sup>

<sup>1</sup> C.R.E.A.- Research Centre for Fruit Tree, Via di Fioranello, 52, 00134 Rome, Italy

<sup>2</sup> C.R.E.A.- Research Centre for Fruit Tree, Via Torrino, 3, 81100 Caserta, Italy

<sup>3</sup> Department of Agricultural Sciences and Environment, University of Udine, Via delle Scienze, 208, 33100, Udine, Italy

### ABSTRACT

The type III secretion system of *Pseudomonas syringae* is a biological machinery that plays a fundamental role in causing diseases to both wild and cultivated plants through the releasing of effector proteins into the plant cell. Problems with the correct identification and comparison of effectors are usually encountered when a putative effector is annotated, particularly during the automated annotation in genome sequencing project. To improve the reliable identification of effector proteins and the address of their identity by homology comparison with the type III secretion system effectors of *Pseudomonas syringae*, we have developed and set up a program that enables fast and automated protein effectors search and comparison. EffectorAnnotator is free and open source software, available under a GPLv2 licence. EffectorAnnotator can be downloaded and installed from <http://www.atlasplantpathogenicbacteria.it/> written in Perl and tested on Linux.

**Keywords:** Genomic annotation; *Pseudomonas syringae*; Effectors; Bioinformatics

### INTRODUCTION

*Pseudomonas syringae* is a polyphagous bacterial phytopathogen able to infect both monocot and dicot plant species, including many cultivated crops. Within the species, more than 50 pathovars (i.e., pathogenic varieties, specifically causing infection to one or few host plants in most cases) have been described [1]. The type III secretion enables to inject into the plant cell a vast arrays of proteins that, in the compatible interaction with the host, subvert the plant defence mechanisms and greatly contribute to start the infection process. Such proteins are known as effector proteins or effectors [2]. Each strain possess a unique set of effectors, also called repertoire, that may include between 20 to 33 different effectors, whereas the *P. syringae* complex as a whole shows a number of validated effector families of about 50 [3]. Each effector, in addition to being horizontally transferred, can mutate to adapt

to a new host plant and/or cultivar by a genetic mechanism called pathoadaptation [4].

For the relevance in causing plant disease, the establishment of the correct effector repertoire of each single strain can greatly contribute to elucidate their similarities/differences in the interaction with the host. However, the great number of effector families and the abundance of mutations and recombinations makes the reliable comparison of effector repertoires a hard task. Given the large diversity of effectors and their involvement in pathoadaptation, a correct identification requires the comparison with a large panel of genomes, but this introduces the noise associated with poor or obsolete annotation of past released genomes. As a matter of fact, the identification and assignment to families of the effectors presently requires specific manual curation and cannot be satisfactorily carried out by automated annotation procedures.

As a recent example, Templeton et al. [5] observed that for a reliable effector proteins identification both the widely used NCBI Prokaryotic Genome Annotation Pipeline (PGAP) (<http://www.ncbi.nlm.nih.gov/genome/annotation>

\*Corresponding author:

Email: [simone.marcelletti@virgilio.it](mailto:simone.marcelletti@virgilio.it)

[http://dx.doi.org/10.20530/EJB\\_4\\_1-5](http://dx.doi.org/10.20530/EJB_4_1-5)

ISSN 2056-9912 © 2016

**Table1. Results of comparison between the automatic annotation using NCBI database and the EffectorAnnotator software.**

Orf number	Result software	Result annotation
orf00062FRU11contig41	hopAX1.689	FullMatch hypothetical protein
orf00063FRU11contig52	hopI1.243	FullMatch molecular chaperone DnaJ
orf00087FRU11contig53	hopAG1.577	FullMatch type III effector
orf00083FRU11contig53	hopAI1.609	FullMatch type III effector
orf00007FRU11contig71	hopAH1.597	FullMatch type III effector
orf00045FRU11contig87	hopBF1.734	FullMatch type III effector HopBF1
orf00134FRU11contig116	hopAA1.492	FullMatch type III effector
orf00131FRU11contig116	hopM1.250	FullMatch restriction endonuclease Eco57I
orf00126FRU11contig116	avrE1.65	FullMatch type III effector
orf00078FRU11contig116	hopZ3.466	FullMatch avirulence protein
orf00073FRU11contig67	hopI1.239	putative molecular chaperone DnaJ
orf00051FRU11contig120	hopI1.239	putative MULTISPECIES: molecular chaperone DnaJ
orf00028FRU11contig138	hopI1.229	putative molecular chaperone DnaJ
orf00029FRU11contig36	hopAZ1.717	truncated type III effector HopAZ1
orf00117FRU11contig44	hopB1.139	truncated type III effector
orf00085FRU11contig53	hopAH1.583	truncated type III effector
orf00079FRU11contig116	hopAX1.689	truncated type III effector HrpK
orf00022FRU11contig44		*HopAN1 protein
orf00017FRU11contig91		*type III helper protein HopAK1
orf00020FRU11contig166		*MULTISPECIES: type III effector HopJ1

\*Not validated as true effector in the most recent release of the effector database.

[prok](http://prok)) and the webserver RAST (<http://rast.nmpdr.org>) yielded inconsistent results, so that a manual search of such proteins must be performed for each comparison.

To improve the reliability of identification, assignment and further comparison of the type III secretion system effectors of *P. syringae*, we have developed and set up a program that enables fast and automated protein effectors search and comparison by using the *P. syringae* effectors databank (that can be found at [www.pseudomonas-syringae.org/pst\\_func\\_gen2.htm](http://www.pseudomonas-syringae.org/pst_func_gen2.htm).) The program was validated in Linux operating system (Fedora and Ubuntu) and has been already applied to a case study[6].

## RESULTS

In order to verify the efficiency of effectorAnnotator, we compared the results obtained by applying this script to the annotation of *Pseudomonas syringae* pv. *syringae* CRAFRU11 (ATSU00000000) genome obtained by using the database of protein retrieved from

<ftp://ftp.ncbi.nlm.nih.gov/refseq/release/bacteria> (last update 15/09/2015).

The annotation using NCBI database identified 13 effectors, while the effectorAnnotator found 17 effector proteins. Moreover, within the effectors found through the annotation using NCBI database, there were a Hop helper and two effectors defined by the website [http://pseudomonas-syringae.org/pst\\_func\\_gen2.htm](http://pseudomonas-syringae.org/pst_func_gen2.htm) as discontinued. Therefore effectorAnnotator found seven additional effectors than the annotation: four belonged to the category full match and three were referred as putative (Table 1).

The maximum number of proteomes analyzed on a PC Intel® Core™ i7-2600 CPU @ 3.40GHz × 8 with 16GiB of Ram is 30. The time spent by the script was about 793m 48.899s; during this period it identifies 841 effector proteins (Figure 1).

EffectorAnnotator can be downloaded and installed from <http://www.atlasplantpathogenicbacteria.it/> written in Perl and tested on Linux.

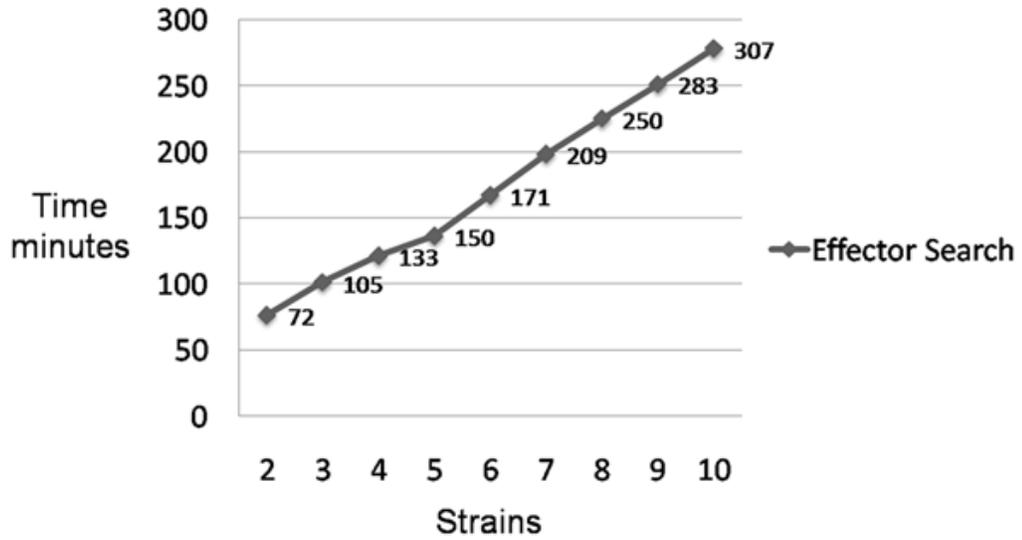


Fig. 1. Execution time (min) of EffectorAnnotator software for analyzing *Pseudomonas syringae* strains.

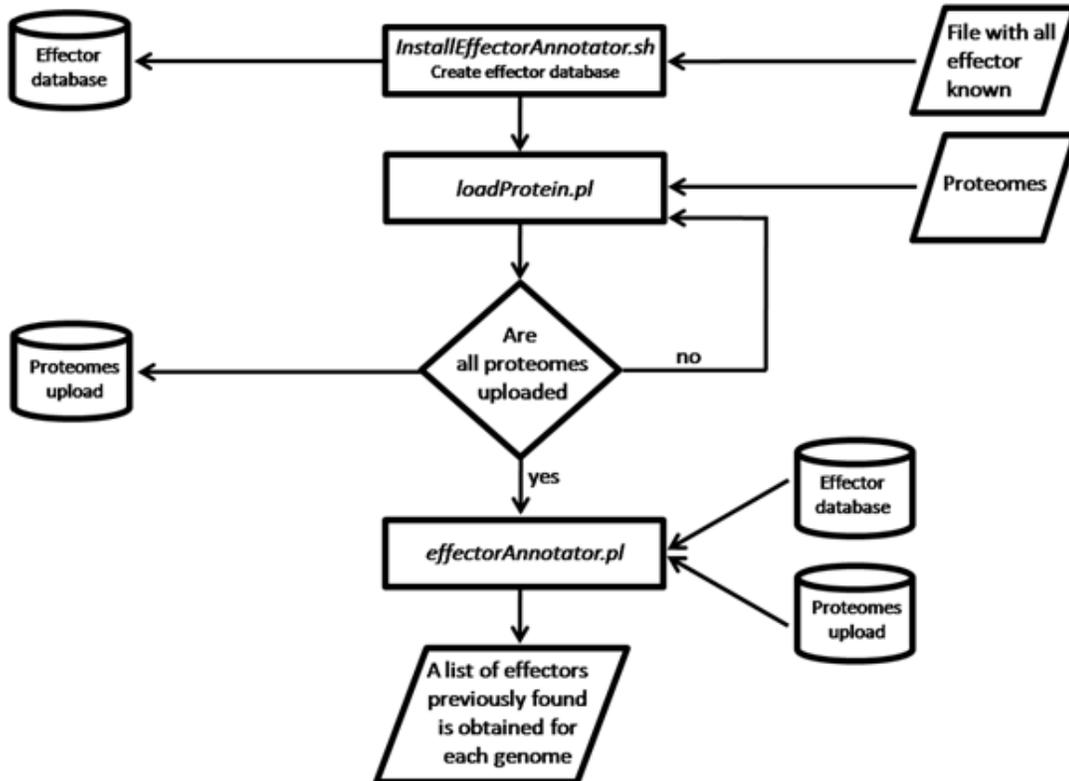


Fig. 2. Complete flow chart of EffectorAnnotator software installation and run.

### FUTURE DEVELOPMENTS

In order to identify the effectors, the software exploits the bacterial genome database as an input.

Regarding to effectorAnnotator, the database includes effectors of *Pseudomonas*. However, by substituting or increasing the database, it would be

also possible to use it as a tool for the identification of other effector proteins for which the automatic annotations fail. For example, it could be useful to find effectors of *Xanthomonas* strains by adding their protein sequences within the software.

## METHODS

The software uses three external softwares. Each of them is contained in the folder “support software”, BLAST v. 2.2.29 [7]; clustalW v2.0.10 [8] and PhyML v3.1 [9].

The program is composed by five scripts: four in perl and one in bash. The script bash, `installeffectorAnnotator.sh`, verifies if Perl is already installed and through the script Perl, `buildEffector.pl`, installs the software. The script Perl `loadProtein.pl` uploads proteomes of bacterial genomes to be analyzed.

The script Perl `effectorAnnotator.pl` takes as input the effector database and the proteomes upload identifies a list of effector for each proteomes in input. Finally, Perl `uninstalleffectorAnnotator.pl` uninstalls the software (Figure 2).

### Input files

The `effectorAnnotator` software requires two input files: a database file containing the protein effectors (downloadable at [http://pseudomonas-syringae.org/pst\\_func\\_gen2.htm](http://pseudomonas-syringae.org/pst_func_gen2.htm)) and a file containing one or more (currently up to 30) proteomes of the bacterial strain(s) to be analyzed.

The effectors are classified in full match, truncated and putative, on the basis of their evalue (modifiable by the users with `-e` option) and alignment length: an evalue  $< 1e-25$  and an alignment length  $> 60\%$  defines the full match effector; an evalue  $< 1e-25$  and an alignment length  $< 60\%$  a truncated effector; an evalue  $> 1e-25$  a putative effectors.

### Output files

`EffectorAnnotator` generates an output (Fig1.) that includes five output files and two folders.

The `histogram.jpeg` file contains three different bars: `FullMatchEffectors`; `TruncatedEffectors` and `PutativeEffectors`.

`Matrix.csv` file contains the number of effectors in each strain.

`Matrix2puMLST.csv` file contains the presence/absence of effector in each strain.

The matrix, without the first line, can be pasted a <http://pubmlst.org/cgi-bin/mlstanalyse/mlstanalyse.pl?site=pubmlst&page=treedraw&referer=pubmlst.org> to build the effector matrix tree.

The `stat.txt` file contains statistical information on the identified effectors.

`Cytoscape.sif` is a file used to create a Venn graph when used with the Venn and Euler App in Cytoscape (<http://www.cytoscape.org/>).

The program creates a folder containing a phylogenetic tree with effectors belonging to the same family for each bacterial strain and a folder containing phylogenetic trees comprehensive of all strain for each effector family. Files that contain phylogenetic tree can be opened with the free software `figtree` (<http://tree.bio.ed.ac.uk/>).

More extensive installation and user guides can be found in the README file.

## CONCLUSIONS

`EffectorAnnotator` software is a rapid, simple-to-use and reliable tool for the identification of effector proteins within the genome of *Pseudomonas syringae* strains. It allows a more accurate and reliable annotation compared to the current way of identification of such fundamental proteins involved in the pathogenicity of bacterial plant pathogens.

## COMPETING INTERESTS

The authors have declared no competing interests.

## REFERENCE

1. Bull, C.T., De Boer, S.H., Denny, T.P., Firrao, G., Fischer-Le Saux, M., Saddler, G.S., Scortichini, M., Stead, D.E. and Takikawa, Y. 2010 Comprehensive list of names of plant pathogenic bacteria, 1980-2007. *Journal of Plant Pathology*, 92 : 551-592.
2. Alfano JR, Collmer A. Type III secretion system effector proteins: Double Agents in Bacterial Disease and Plant Defense. *Annual Review of Phytopathology*. 2004 Sep;42(1):385–414. Available from: <http://dx.doi.org/10.1146/annurev.phyto.42.040103.110731>.
3. Baltrus DA, Nishimura MT, Romanchuk A, Chang JH, Mukhtar MS, Cherkis K, et al. Dynamic Evolution of Pathogenicity Revealed by Sequencing and Comparative Genomics of 19 *Pseudomonas syringae* Isolates. Guttman DS, editor. *PLoS Pathogens*. 2011 Jul

- 14;7(7):e1002132. Available from:  
<http://dx.doi.org/10.1371/journal.ppat.1002132>.
4. Ma W, Dong FFT, Stavrinides J, Guttman DS. Type III Effector Diversification via Both Pathoadaptation and Horizontal Transfer in Response to a Coevolutionary Arms Race. *PLoS Genet.* 2006;2(12):e209. Available from:  
<http://dx.doi.org/10.1371/journal.pgen.0020209>.
  5. Templeton MD, Warren BA, Andersen MT, Rikkerink EHA, Fineran PC. Complete DNA Sequence of *Pseudomonas syringae* pv. *actinidiae*, the Causal Agent of Kiwifruit Canker Disease. *Genome Announc* [Internet]. American Society for Microbiology; 2015 Sep 17;3(5):e01054–15. Available from:  
<http://dx.doi.org/10.1128/genomea.01054-15>.
  6. Marcelletti S, Scortichini M. Comparative Genomic Analyses of Multiple *Pseudomonas* Strains Infecting *Corylus avellana* Trees Reveal the Occurrence of Two Genetic Clusters with Both Common and Distinctive Virulence and Fitness Traits. Semsey S, editor. *PLoS ONE.* 2015 Jul 6;10(7):e0131112. Available from:  
<http://dx.doi.org/10.1371/journal.pone.0131112>.
  7. Altschul S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research.* 1997 Sep 1;25(17):3389–402. Available from:  
<http://dx.doi.org/10.1093/nar/25.17.3389>.
  8. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. *Bioinformatics.* 2007 Sep 10;23(21):2947–8. Available from:  
<http://dx.doi.org/10.1093/bioinformatics/btm404>.
  9. Guindon S, Gascuel O. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology.* 2003 Oct 1;52(5):696–704. Available from:  
<http://dx.doi.org/10.1080/10635150390235520>.