

1 **Searching Algorithm for Type IV Effector proteins (S4TE) 2.0: improved**
2 **tools for type IV effector prediction, analysis and comparison**

3

4 Christophe Noroy^{1,2,3}, Thierry Lefrançois² and Damien F. Meyer^{1,2*}

5

6 ¹ CIRAD, UMR ASTRE, F-97170 Petit-Bourg, Guadeloupe, France

7 ² ASTRE, Univ Montpellier, CIRAD, INRA, Montpellier, France

8 ³ Université des Antilles, 97159 Pointe-à-Pitre, Guadeloupe, France

9 * To whom correspondence should be addressed. Tel: +590 (0)590 25 59 47; Email:

10 damien.meyer@cirad.fr

11 **ABSTRACT**

12 Bacterial pathogens have evolved numerous strategies to corrupt, hijack or mimic
13 cellular processes in order to survive and proliferate. Among those strategies, Type IV
14 effectors (T4Es) are proteins secreted by pathogenic bacteria to manipulate host cell
15 processes during infection. They are delivered into eukaryotic cells in an ATP-dependent
16 manner via the type IV secretion system, a specialized multiprotein complex. T4Es contain
17 a wide spectrum of features including eukaryotic-like domains, localization signals or a C-
18 terminal translocation signal. A combination of these features enables prediction of T4Es
19 in a given bacterial genome. In this study, we developed a web-based comprehensive
20 suite of tools with a user-friendly graphical interface. This version 2.0 of S4TE (Searching
21 Algorithm for Type IV Effector Proteins; <http://sate.cirad.fr>) enables accurate prediction and
22 comparison of T4Es. Search parameters and threshold can be customized by the user to
23 work with any genome sequence, whether publicly available or not. Applications range
24 from characterizing effector features and identifying potential T4Es to analyzing the
25 effectors based on the genome G+C composition and local gene density. S4TE 2.0 allows
26 the comparison of putative T4E repertoires of up to four bacterial strains at the same time.
27 The software identifies T4E orthologs among strains and provides a Venn diagram and
28 lists of genes for each intersection. New interactive features offer the best visualization of
29 the location of candidate T4Es and hyperlinks to NCBI and Pfam databases. S4TE 2.0 is
30 designed to evolve rapidly with the publication of new experimentally validated T4Es,
31 which will reinforce the predictive power of the algorithm. The computational methodology
32 can be used to identify a wide spectrum of candidate bacterial effectors that lack sequence
33 conservation but have similar amino acid characteristics. This approach will provide very
34 valuable information about bacterial host-specificity and virulence factors, and help identify
35 host targets for the development of new anti-bacterial molecules.

36

37

38 INTRODUCTION

39 Proteobacteria have evolved specific effector proteins to manipulate host cell gene
40 expression and processes, hijack immune responses and exploit host cell machinery
41 during infection. These proteins are secreted by ATP-dependent protein complexes named
42 type IV secretion systems (T4SS). Some T4Es have been identified and shown to be
43 crucial for pathogenicity. To facilitate the identification of putative T4Es, we previously
44 developed a bioinformatics tool called S4TE 1.0 (Searching Algorithm for Type IV
45 secretion system effector proteins) [1].

46 In the present article, we present the second version of 'S4TE'. S4TE 2.0 is a tool
47 for *in silico* screening of proteobacteria genomes and T4E prediction based on the
48 combined use of 14 distinctive features. In this updated version, modules searching for
49 promoter motifs, homology, NLS, MLS and E-block are more efficient. A new module has
50 been added in the workflow to locate phosphorylation (EPIYA-like) domains.

51 S4TE 2.0 consists of the S4TE 1.4 tool and a web interface available to non-
52 commercial users at <http://sate.cirad.fr>. The web interface is designed to make S4TE 2.0
53 easy to use for biologists and more time efficient. Most of the genomes and plasmids
54 available in the NCBI database of pathogenic bacteria that have a type IV secretion
55 system have been loaded into the S4TE 2.0 database so effectors can be predicted in only
56 a few clicks.

57 S4TE 2.0 offers advanced users an expert mode (S4TE-EM) they can use to
58 customize S4TE 2.0 search parameters (e.g. exclude modules, modify module weightings).
59 In this mode, S4TE 2.0 can be used as 14 independent programs to search for particular
60 features in a given bacterial genome (e.g. NLS, C-ter charges).

61 A new function for comparative genomics (S4TE-CG) has been added to compare
62 up to four predicted effectomes in just a few seconds.

63 All S4TE 2.0 results are interactive and linked to NCBI and Pfam databases.

64

65 SOFTWARE AND ALGORITHM

66

67 Programming

68 S4TE 2.0 software consists in a graphical interface (website) to use the S4TE 1.4
69 algorithm for genome analysis, Type IV effectors (T4Es) prediction and comparison of
70 effectomes. S4TE 1.4 is an update of S4TE 1.0[1]. It is written in Perl programming
71 language and uses NCBI, Pfam, EMBOSS, BioPerl and MitoFates libraries and its own
72 proper programs and database. It was developed to improve the prediction performances

73 of S4TE 1.0 and to provide new functionalities to search for new features, enable
74 interactivity and comparative genomics. The 10 S4TE search modules in S4TE 1.0 were
75 kept in S4TE 1.4. However, some modules have been modified (promoter motif search,
76 homology, MLS, NLS, E-block and Pfam database) to improve their predictive power. A
77 supplementary module (EPIYA search) has been added to the workflow. In this paper, only
78 the EPIYA search module and the revised modules are described.

79

80 **Promoter motif search**

81 As several T4Es in a given bacterium can be subjected to coordinated regulation with
82 the same protein, *e.g.* PmrA[2], we used S4TE 2.0 to conduct a search for conserved
83 motifs (potential regulatory motifs) in the short promoter regions of the genes. The aim was
84 to improve S4TE 2.0 prediction of possible regulons of T4Es. Enriched DNA motifs were
85 searched in a window of 100 nucleotides (nt) placed upstream of the start codon, using
86 MEME[3]. Eight consensus motifs were identified in different bacteria (table 1). The
87 corresponding motif search module of S4TE 2.0 extracts the 5' Flanking intergenic regions
88 (5' FIRs) and searches for all these motifs thanks to a position-specific scoring matrix
89 generated from multiple sequence alignments with the promoters of known T4Es. Only
90 alignments with a score above the chosen threshold are selected. The threshold that
91 yielded the highest sensitivity and specificity for each motif in the corresponding bacterium
92 was chosen (Table 1).

93

94 **Homology**

95 BLAST 2.2 was used to compare proteins to search for homologies with known T4Es [4].
96 The cut-off of the S4TE 1.0 homology module was changed. S4TE 2.0 compares the
97 database containing all known T4Es with the query proteome and returns all homologs
98 with a cut-off of the expected value (E) $<10^{-4}$. This E-value cut-off was selected to find real
99 homologs between phylogenetically distant bacterial species. Databases containing
100 proven effectors have also been updated (Table S1).

101

102 **Nuclear localization signals (NLS)**

103 NLS are protein sequences that target proteins in the nucleus of eukaryotic cells[5]. We
104 assume that the occurrence of NLS in a bacterial protein sequence would be a good
105 indicator of secretion. There are two classes of NLS, monopartite and Bipartite. In S4TE
106 2.0, the search for monopartite NLS has been improved according to Ruhanen *et al.* [6].
107 We rewrote this module to add more known NLS motifs in the search. Monopartite NLS

108 consist of [KR]-[KR]-[KR]-[KR]-[KR], X-K-[KR]-[KRP]-[KR]-X, X-R-K-[KRP]-[KR]-X, X-R-K-
109 X-[KR]-[KRP], X-K-[KR]-[KR]-X-[KRP], X-R-K-[KR]-X-[KRP], X-K-[KR]-X-[KR]-X-X, X-R-K-
110 X-[KR]-X-X, X-K-[KR]-[KR]-X-X-X and X-R-K-[KR]-X-X-X motifs. Bipartite NLS were also
111 searched with S4TE 1.0 motif (K-[KR]-X(6,20)-[KR]-[KR]-X-[KR]). The new module was
112 tested with a dataset of 32 NLS and 32 no-NLS containing proteins (dataset 1). The
113 module selected 24 true positives (TP) and only three false positives (FP). This represents
114 a sensitivity (Se) of 75% and a specificity (Sp) of 91%.

115

116 **Mitochondrial Localization Signals (MLS)**

117 MLS are signal sequences located in the N-terminus of proteins that are targeted to
118 mitochondria. This sequence is cleaved after translocation of the protein inside the
119 mitochondria[5,7]. To predict MLS in S4TE 2.0, we used the MitoFates tool[8]. MitoFates
120 predicts mitochondrial presequences, a cleavable localization signal located in the N-
121 terminal, and its cleaved position.

122

123 **E-block**

124 The E-block domain consists of a glutamate sequence rich in C-terminal 30 amino acids
125 and is associated with T4Es translocation in *L. pneumophila*. Huang *et al.* showed that an
126 E-block motif is also important for the translocation of T4SS substrates[9]. In S4TE 2.0, the
127 E-block module was modified according to Lifshitz *et al.* [10]. The E-block was searched in
128 a window of 22 amino acids between position -4 C-terminal and -26 C-terminal. The motif
129 that is searched for is a motif of 10 amino acids containing three or more glutamate (E)
130 residues. The module was tested on 98 E-block and 98 no-E-block containing proteins
131 (dataset 2). This module selected 60 TP and only 6 FP (Sensitivity of 61%, Specificity of
132 94%).

133

134 **Pfam database**

135 The local Pfam database has been updated to find more eukaryotic domains of known
136 effectors of *Legionella pneumophila*[10]. Eukaryotic domains were extracted from the
137 whole Pfam database and added to the S4TE 2.0 workflow. All eukaryotic domains used
138 for this search are listed in Table S2.

139

140 **EPIYA search**

141 EPIYA search is a new module implemented in S4TE 2.0. The EPIYA domain is an
142 eukaryotic phosphorylation motif[11]. In *H. pylori*, EPIYA has been shown to contribute to

143 the secretion of a CagA effector[12]. We searched for conserved EPIYA motifs (EPIYA,
144 ENIYE, NPLYE, EHLYA, TPLYA, EPLYA, ESIYE, EDLYA, EPIYG, EPVYA, VPNYA,
145 EHIYD) in different bacteria that have a type IV secretion system and we searched for
146 hypothetical EPIYA motifs using the motif E-X-X-Y-X.

147

148 **Validation**

149 S4TE 2.0 is a software program with 14 independent modules. We tested all the
150 modules independently. The 14 modules were weighted to make S4TE 2.0 efficient. The
151 weighting of each module was calculated according to its performance in finding effectors
152 in *L. pneumophila* Philadelphia I which has been shown to have the most extensive
153 repertoire of T4Es ever identified, with 286 confirmed effectors [10].

154 Each module has its own weighting in S4TE 2.0 searches. The weightings were
155 calculated for each module based on their Positive Predictive Value (PPV
156 [PPV=TP/(TP+FP)]) for *L. pneumophila* (Table 2).

157 The S4TE 2.0 prediction threshold was then defined to enable the best prediction by
158 disregarding homology with known effectors. The threshold was chosen by examining the
159 Sensitivity (Se), Specificity (Sp), Positive Predictive Value (PPV), Negative Predictive
160 Value (NPV) and Accuracy (Acc) for thresholds ranging from 40 to 120 on the test dataset
161 (Figure 2). The threshold was set at a score of 72 to obtain the global PPV possible with
162 the least possible impact on sensitivity.

163 This threshold combined with weightings led to the correct prediction (true positives) of
164 282 of the 286 effectors of *L. pneumophila* (Se=98%, PPV=60%) and 96 incorrect
165 predictions (false positives) (Sp = 96%, NPV = 99%).

166 With this update, S4TE 2.0 prediction is more powerful than that of S4TE 1.0 whose
167 sensitivity was 14% lower. Without homology, sensitivity increased by 25% (data not
168 shown). Other characteristics including specificity, accuracy and negative predictive value
169 did not change significantly (table 3). S4TE 2.0 allows flexible, highly sensitive and specific
170 detection of new putative T4SS effectors.

171

172 **SATE-CG**

173 S4TE-CG is a new tool designed to compare different repertoires of putative T4Es
174 identified by S4TE 2.0. The corresponding S4TE-CG algorithm is described in Figure 3.
175 The user can compare up to four effectomes simultaneously. S4TE 2.0 results from
176 selected genomes (effectomes) are compared with Blastp 2.2 with an expected value (E)
177 cut-off of $<10^{-4}$ to find homologous proteins in each effectome. S4TE-CG successively

178 compares all effectomes in a pairwise manner, the overlaps between the effectomes of
179 each genome are calculated and the final results are plotted on a Venn diagram and listed
180 in an interactive table. All effectors are clickable and the user is redirected to the S4TE 2.0
181 results on the effector concerned. The table can be easily copied and pasted for export.

182

183 **Software availability**

184 S4TE 2.0 is a web interface and the S4TE 1.4 package is freely available to non-
185 commercial users at <http://sate.cirad.fr/S4TE-Doc.php>. All programming was done using
186 Perl 5.18 and BioPerl 1.6.1. The software runs on Linux platforms (Ubuntu 14.04 and Mac
187 OS X). All required packages and the installation process are described in the user guide
188 included in the package. The user guide also details S4TE options for running S4TE. By
189 default, the command line to launch S4TE is `./S4TE.pl -f "Genbank_file"` in the
190 S4TE folder (`cd way_to_S4TE/S4TE/`). Some options are available for the user to
191 launch S4TE: `-c`, suppression of a module in the pipeline; `-w`, modification of the weight of
192 each module in the pipeline; `-t`, imposition of a threshold for effector selection. Each
193 S4TE module creates a `.txt` file in the folder `way_to_S4TE/S4TE/Jobs/`
194 `job<Name_of_genome_folder><year><month><day><hour><min>`

195 All the results are compiled in the *CompilationFile.txt* and *Results.txt* in the same folder.

196

197 **WEB INTERFACE**

198

199 **Design and general features**

200 The S4TE 2.0 website is powered from scratch on the 'CIRAD web server'. All the features
201 of the web site were tested on common web browsers. S4TE 2.0 found T4Es in large
202 genome databases (Table S3) available to all users. A user account is available and
203 necessary to keep your jobs up to three months, to import your own genome in a S4TE 2.0
204 temporary database and to ask to add a new proved effector in the database. The addition
205 of an effector to the database must be accompanied by a reference (scientific article) and
206 will be checked manually before the effector is added to the database. Those who
207 subscribe to the newsletter will be notified by email about the addition of new effectors to
208 the database and the effector will be visible in the S4TE 2.0 tab strip. This free account
209 allows users to search for proteins in the S4TE 2.0 database using the name, the locus tag
210 or NCBI number of a protein in the search bar. The account also allows the user to
211 subscribe to the S4TE newsletter that summarizes any changes made to the software, and
212 provide updates on the latest research on Type IV Effectors.

213

214 **S4TE 2.0 is a simple and user-friendly tool**

215 S4TE 2.0 is a web-based user-friendly tool that gets results in only a few clicks. The user
216 can locate a chromosome in more than 340 bacterial genomes and plasmids available in
217 the database and the results can be viewed by clicking on run S4TE 2.0 (Table S3).

218 If the desired genome is not available in the databases, the user can import it with a
219 GenBank file (.gbk). S4TE 2.0 will import the file to a temporary database for three months.
220 All S4TE tools (S4TE-EM and S4TE-CG) can then be used on the genome by the owner.

221 The S4TE 2.0 web page allows users to read some of the news published in the
222 newsletter. Five news items are visible on the S4TE2.0 web page, but all the news can be
223 found by clicking on the bottom right link.

224 Figure 4 presents some results obtained with S4TE 2.0. All the proteins in the selected
225 genome are represented on the S4TE 2.0 web results page. A score was calculated for
226 each protein based on the weighting of each module. Proteins were ranked according to
227 the same score. All proteins whose scores are above the threshold are considered as
228 belonging to the S4TE 2.0 effectome. An iconography was created to help read the list
229 (Figure 4A). Users can find all the details concerning each characteristic of a given protein
230 by clicking on the protein concerned on the web results page.

231 When a user runs S4TE 2.0, in addition to the results page, two graphs are automatically
232 drawn. The first shows the distribution of predicted effectors according to local gene
233 density (Figure 4B). The second one displays the distribution of predicted T4Es according
234 to the G+C content along the genome (Figure 4C).

235

236 **S4TE-EM Expert mode for accurate searching**

237 S4TE-EM is the expert mode of S4TE 2.0. S4TE-EM allows the user to modify the weights
238 of each module and to deactivate one or more modules in the search (Figure 1). The
239 weight of a module can be changed by moving the position of the cursor next to the name
240 of each module. Weightings can be changed between the lowest weighting available for
241 the module and the threshold of S4TE 2.0 ($\tau=72$). The lowest weight is calculated
242 independently for each module as a function of the positive predictive value and
243 corresponds to a value equal to 0.5. Users can also cancel one or more modules in the
244 pipeline by unchecking the box next to the name of the module (Figure 1).

245 All the modules are independent and users can use S4TE-EM to locate the same
246 characteristic throughout the genome. For example, if the user disables all the modules

247 except NLS, S4TE-EM will find all proteins with an NLS in the genome, meaning users can
248 use S4TE-EM as a new genome analysis tool.

249

250 **S4TE-CG Comparative genomics to compare effectomes**

251 S4TE-CG is a new tool designed to compare different effectomes predicted by S4TE 2.0.
252 Users can choose up to four effectomes in S4TE 2.0 databases or upload a genome
253 present in the temporary database. S4TE-CG displays results in a Venn diagram and in an
254 interactive table. Users can easily find different subsets of information in the appropriate
255 table by referring to the different colors in the Venn diagram (Figure 3). Information about
256 each effector can easily be found by clicking on the name of the effector in the table. Or
257 users can simply copy and paste the table in a .csv file.

258

259 **CONCLUSION**

260 This paper presents updated S4TE software. The computational tool is designed to predict
261 the presence of T4SS effector proteins in bacteria. The identification of T4Es and some
262 characteristics are improved in this update. Compared with a machine learning approach,
263 using S4TE 2.0 to predict T4Es in *Legionella* and *Coxiella* species[10,13,14] improved
264 sensitivity (98% for S4TE 2.0 and 89% for Wang *et al.*) and equivalent specificity (97% for
265 Wang *et al.* and 93% for S4TE 2.0). S4TE 2.0 is easy to use. Only an internet connection
266 and a few clicks are needed to search for T4Es in more than 340 bacterial genomes and
267 plasmids. The results are displayed instantaneously for easy reading. An automated
268 pipeline is also provided to analyze and visualize effector distribution in the genome
269 according to G+C content and local gene density. S4TE 2.0 results are linked to
270 bioinformatics databases like NCBI and Pfam. The S4TE 2.0 database is designed to
271 evolve and will be updated by adding new proven effectors and new bacterial genomes.
272 S4TE 2.0 not only predicts the T4Es but also their subcellular localization (NLS, MLS,
273 prenylation) and the function of these proteins (Coiled coils, EPIYA, Euk-like, etc.). All
274 these features make S4TE 2.0 a powerful software for studies of T4Es.

275 S4TE 2.0 also offers an expert mode, which allows users to make manual adjustments to
276 the weight of the modules. Each module that searches for a feature or a characteristic can
277 be used independently. S4TE EM can be viewed and use as 14 independent programs.
278 This could facilitate the annotation of new genomes by looking for specific features such
279 as NLS, prenylation domains, etc.

280 Finally, S4TE-CG makes it possible for users to compare effectomes to highlight core T4
281 effectomes and/or accessory T4 effectomes to understand how effectomes evolved, and
282 may provide clues to the specificity of different strains.

283

284 REFERENCES

- 285 1. Meyer DF, Noroy C, Moumène A, Raffaele S, Albina E, Vachiéry N: **Searching algorithm for type IV**
286 **secretion system effectors 1.0: a tool for predicting type IV effectors and exploring their genomic context.**
287 *Nucleic acids research* 2013, **41**:9218–2910.1093/nar/gkt718.
- 288 2. Zusman T, Aloni G, Halperin E, Kotzer H, Degtyar E, Feldman M, Segal G: **The response regulator PmrA is**
289 **a major regulator of the icm/dot type IV secretion system in Legionella pneumophila and Coxiella**
290 **burnetii.** *Molecular microbiology* 2007, **63**:1508–2310.1111/j.1365-2958.2007.05604.x.
- 291 3. Bailey T, Boden M, Buske F, Frith M, Grant C, Clementi L, Ren J, Li W, Noble W: **MEME SUITE: tools for**
292 **motif discovery and searching.** *Nucleic acids research* 2009, **37**:W202–810.1093/nar/gkp335.
- 293 4. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden T: **NCBI BLAST: a better web**
294 **interface.** *Nucleic acids research* 2008, **36**:W5–910.1093/nar/gkn201.
- 295 5. Hicks SW, Galán JE: **Exploitation of eukaryotic subcellular targeting mechanisms by bacterial effectors.**
296 *Nature Reviews Microbiology* 2013, **11**:316–32610.1038/nrmicro3009.
- 297 6. Ruhanen H, Hurley D, Ghosh A, O'Brien KT, Johnston CRR, Shields DC: **Potential of known and short**
298 **prokaryotic protein motifs as a basis for novel peptide-based antibacterial therapeutics: a computational**
299 **survey.** *Front Microbiol* 2014, **5**:410.3389/fmicb.2014.00004.
- 300 7. Niu H, Kozjak-Pavlovic V, Rudel T, Rikihisa Y: **Anaplasma phagocytophilum Ats-1 Is Imported into Host**
301 **Cell Mitochondria and Interferes with Apoptosis Induction.** *PLoS Pathogens* 2010,
302 **6**:e100077410.1371/journal.ppat.1000774.
- 303 8. Fukasawa Y, Tsuji J, Fu S-CC, Tomii K, Horton P, Imai K: **MitoFates: improved prediction of mitochondrial**
304 **targeting sequences and their cleavage sites.** *Mol. Cell Proteomics* 2015, **14**:1113–
305 2610.1074/mcp.M114.043083.
- 306 9. Huang L, Boyd D, Amyot W, Hempstead A, Luo Z-Q, Connor T, Chen C, Machner M, Montminy T,
307 Isberg R: **The E Block motif is associated with Legionella pneumophila translocated substrates.** *Cellular*
308 *Microbiology*, **13**:227–24510.1111/j.1462-5822.2010.01531.x.
- 309 10. Lifshitz Z, Burstein D, Peeri M, Zusman T, Schwartz K, Shuman H, Pupko T, Segal G: **Computational**
310 **modeling and experimental validation of the Legionella and Coxiella virulence-related type-IVB secretion**
311 **signal.** *Proc National Acad Sci* 2013, **110**:E707–E71510.1073/pnas.1215278110.
- 312 11. Safari F, Murata-Kamiya N, Saito Y, Hatakeyama M: **Mammalian Pragmin regulates Src family kinases**
313 **via the Glu-Pro-Ile-Tyr-Ala (EPIYA) motif that is exploited by bacterial effectors.** *Proc. Natl. Acad. Sci.*
314 *U.S.A.* 2011, **108**:14938–4310.1073/pnas.1107740108.
- 315 12. Papadakos KS, Sougleri IS, Mentis AF, Hatziloukas E, Sgouras DN: **Presence of terminal EPIYA**
316 **phosphorylation motifs in Helicobacter pylori CagA contributes to IL-8 secretion, irrespective of the**
317 **number of repeats.** *PLoS ONE* 2013, **8**:e5629110.1371/journal.pone.0056291.
- 318 13. Lifshitz Z, Burstein D, Schwartz K, Shuman H, Pupko T, Segal G: **Identification of Novel Coxiella burnetii**
319 **Icm/Dot Effectors and Genetic Analysis of Their Involvement in Modulating a Mitogen-Activated Protein**
320 **Kinase Pathway.** *Infect Immun* 2014, **82**:3740–375210.1128/IAI.01729-14.
- 321 14. Wang Y, Wei X, Bao H, Liu S-L: **Prediction of bacterial type IV secreted effectors by C-terminal features.**
322 *Bmc Genomics* 2014, **15**:1–1410.1186/1471-2164-15-50.

323









324

325

326 **TABLE AND FIGURE LEGENDS**

327

328 **Table 1. Enriched DNA motifs found in several bacteria in the 100 nucleotides**
 329 **upstream of known type IV effectors and implemented in S4TE 2.0 searches**

Name	Organism	Length	Threshold	Effector ¹	Non-effector ²	Logo ³
PmrA	<i>Legionella</i>	20	0.748	18.6	4.1	
Cpm	<i>Coxiella</i>	20	0.87	18.2	0.02	
Cpm2	<i>Coxiella</i>	7	0.875	13.8	3.9	
Apm	<i>Anaplasma</i>	14	0.7	77.8	9.9	
Apm2	<i>Anaplasma</i>	15	0.86	66.7	0.41	
Bapm	<i>Bartonella</i>	19	0.75	62.5	2.2	
Hpm	<i>Helicobacter</i>	20	0.68	1	0.39	
Bopm	<i>Bordetella</i>	20	0.8	52.6	0.51	

330 ¹Frequency of motif in effector promoters

331 ²Frequency of motif in non-effector promoters

332 ³Logo and motif were established using MEME software (Bailey TL *et al.*, 2009).

333

334

335

336 **Table 2. Calculation of S4TE 2.0 weighting according to *Legionella pneumophila***
 337 **Philadelphia 1 Positive Predictive Values (PPV) of each module**

338

S4TE 2.0 Features	1	2	3	4	5	6	7	8	9	10	11	12	13	14
True Positives	108	285	13	2	30	105	6	1	100	262	62	41	114	98
False Positives	434	34	27	106	101	783	79	6	231	2376	863	339	156	232
PPV(%)	20	89	32	2	23	12	7	14	30	10	7	10	42	30

339

340

341

342

343

344

345

346 **Table 3. Comparison between S4TE 1.0 and S4TE 2.0**

Software	S4TE 1.0		S4TE 2.0	
	With	Without	With	Without
Homology				
Sensitivity	0.86	0.16	1	0.41
Specificity	0.97	0.97	0.93	0.93
Positive Predictive Value	0.74	0.44	0.60	0.43
Negative Predictive Value	0.98	0.91	1	0.94

347

348 **FIGURE CAPTIONS**

349 **Figure 1. The new front page of the S4TE-EM tool.** The right side provides some
350 information about the page. The right side matches the user account. The user account shows all
351 the jobs previously ran in S4TE 2.0 and S4TE-CG. This account makes it possible to search a
352 protein with the search bar and to ask to add a proven T4 effector in the database. In the central
353 part of the work space, the user can select a genome in the drop-down menu. In S4TE-EM, the
354 user can change the weighting or disable one or more modules (on the left) shown in the S4TE
355 diagram (on the right), and run S4TE-EM by clicking on the 'Run S4TE-EM' button.


356 **Figure 2. Distribution of S4TE 2.0 performances according to the threshold.** Plot of
357 the sensitivity (Se), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV)
358 and accuracy (Acc) of S4TE 2.0 with no homology module on *L. pneumophila* genome as a
359 function of the S4TE 2.0 threshold. A threshold of 72 proved to be the best combination of these
360 characteristics.

361 **Figure 3. Flow chart of the comparison of 4 effectomes using S4TE-CG.** Users can
362 compare up to four genomes simultaneously. **1.** S4TE 2.0 results from selected genomes
363 (effectomes) are compared with Blastp 2.2 to find homologous proteins in each effectome.
364 **2.** S4TE-CG successively compares all effectomes in a pairwise manner, and calculates
365 any overlaps between the effectomes of each genome. **3.** The final results are plotted on a
366 Venn diagram and listed in an interactive table.

367

368 **Figure 4. Example of S4TE 2.0 results for *Anaplasma phagocytophilum* HZ. APH-**
369 **0740. A.** Schematic representations of proteins with different characteristics present in the
370 sequence are shown. Characteristics are easy to find by highlighting the corresponding
371 sequence in the effector sequence. These characteristics are detailed below the sequence.
372 **B.** Distribution of S4TE 2.0 predicted type IV effectors (T4Es) according to local gene
373 density. The predicted T4Es are plotted according to the length of their flanking intergenic
374 regions (FIRs). All *A. phagocytophilum* genes were sorted into 2-dimensional bins

375 according to the length of their 5' (y-axis) and 3' (x-axis) FIRs. The number of genes in the
376 bins is represented by a color-coded density graph. Genes whose FIRs are both longer
377 than the median FIR length were considered as gene-sparse region (GSR) genes. Genes
378 whose FIRs are both below the median value were considered as gene-dense region
379 (GDR) genes. In-between region (IBR) genes are genes with a long 5' FIR and short 3' FIR,
380 or inversely. Candidate T4Es predicted using the S4TE2.0 algorithm were plotted on this
381 distribution according to their own 3' and 5' FIRs. A color is assigned to each of the three
382 following groups: Red to GDRs, orange to IBRs, and blue to GSRs. **C.** Genome-wide
383 distribution of predicted effectome according to the G+C content. From outer track to inner
384 track, sense and antisense genes (black), S4TE 2.0 putative T4Es (pink), proved T4Es
385 (turquoise), S4TE 2.0 putative T4Es in genomic region with low G+C content (yellow),
386 S4TE 2.0 putative T4Es in genomic region with high G+C content (blue), $G+C \geq$ average
387 $G+C$ (red), $G+C <$ average $G+C$ (green).
388



Searching Algorithm for
Type IV Effector proteins 2.0

Home | [S4TE 2.0](#) | [S4TE-EM](#) | [S4TE-CG](#) | [S4TE-Doc](#)

S4TE-EM Select a Genome, choose your best settings and run the search for candidate effectors

S4TE-EM (Expert Mode)

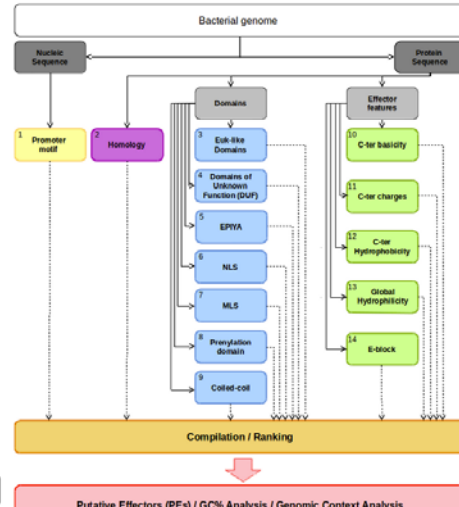
S4TE-EM is a modular software and the user keeps the possibility to adjust various parameters such as the selection of search modules and the weight of each module.

Genome Select a Genome

Settings

1- Promoter motif	<input checked="" type="checkbox"/>		6
2- Homology	<input checked="" type="checkbox"/>		6
3- Euk-like Domains	<input checked="" type="checkbox"/>		6
4- Domains of Unknown Function (DUF)	<input checked="" type="checkbox"/>		6
5- EPIYA	<input checked="" type="checkbox"/>		6
6- NLS	<input checked="" type="checkbox"/>		6
7- MLS	<input checked="" type="checkbox"/>		6
8- Prenylation Domain	<input checked="" type="checkbox"/>		6
9- Coiled-coils	<input checked="" type="checkbox"/>		6
10- C-ter basicity	<input checked="" type="checkbox"/>		6
11- C-ter charges	<input checked="" type="checkbox"/>		6
12- C-ter Hydrophobicity	<input checked="" type="checkbox"/>		6
13- Global Hydrophilicity	<input checked="" type="checkbox"/>		6
14- E-block	<input checked="" type="checkbox"/>		6

Reset **Run S4TE-EM**



User's account

Jobs will be deleted from database after three months

Recent Jobs

- CG05191614 ✖
- A.cen05232018 ✖
- CG05191646 ✖
- CG05191717 ✖

Search for proteins

Ask to add a new effector
Subscribe from newsletter

Sign out

[Contact us](#) | [Legal details](#)

Current version: 2.0 | Decembre 2017

The S4TE project was funded by EU research grant FP7-REGPOT (EPIGENESIS project), and by the Guadeloupe Region and the European Regional Development Fund (MALIN project).

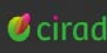


Figure 1

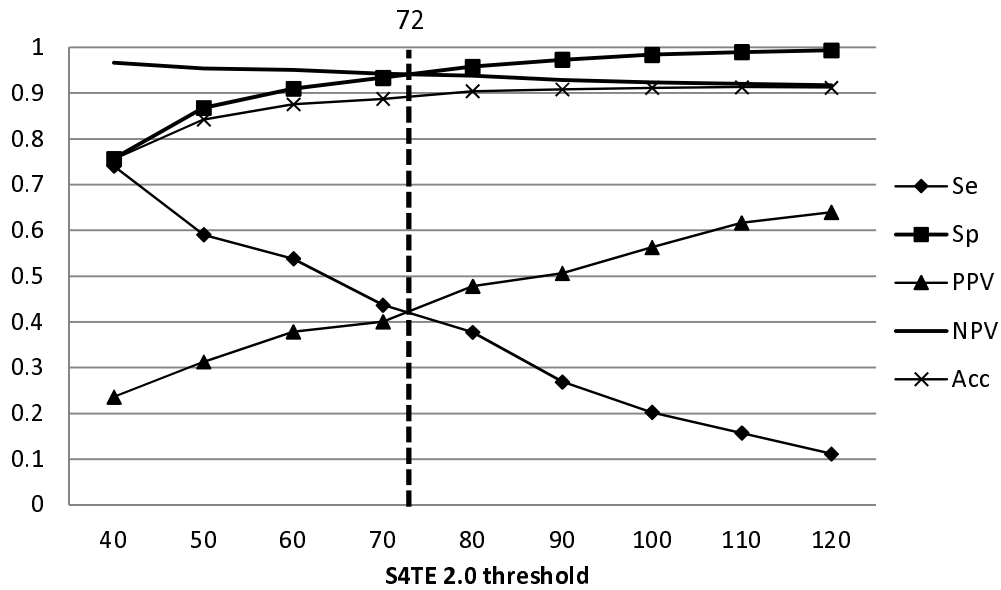


Figure 2.

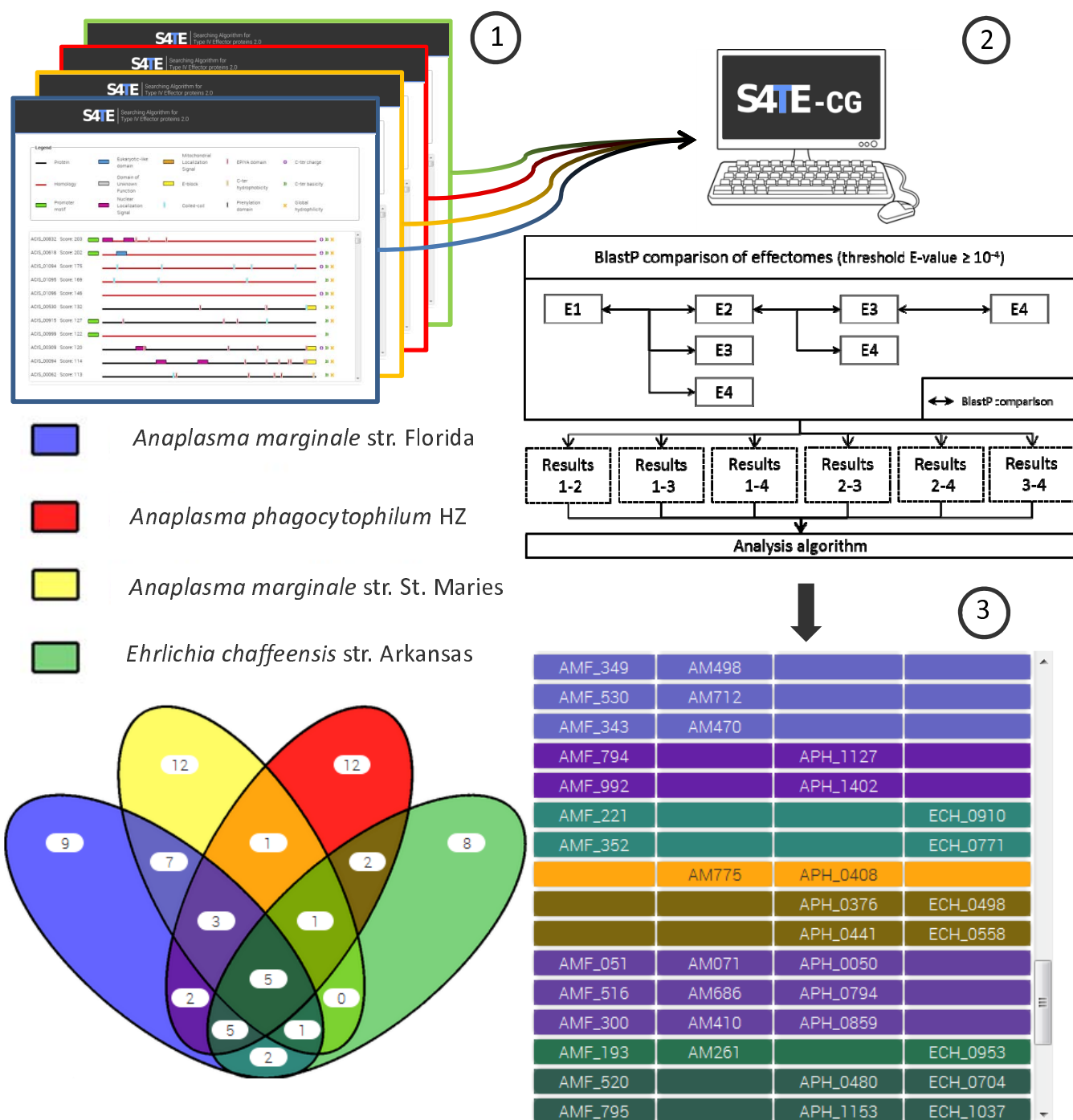


Figure 3.

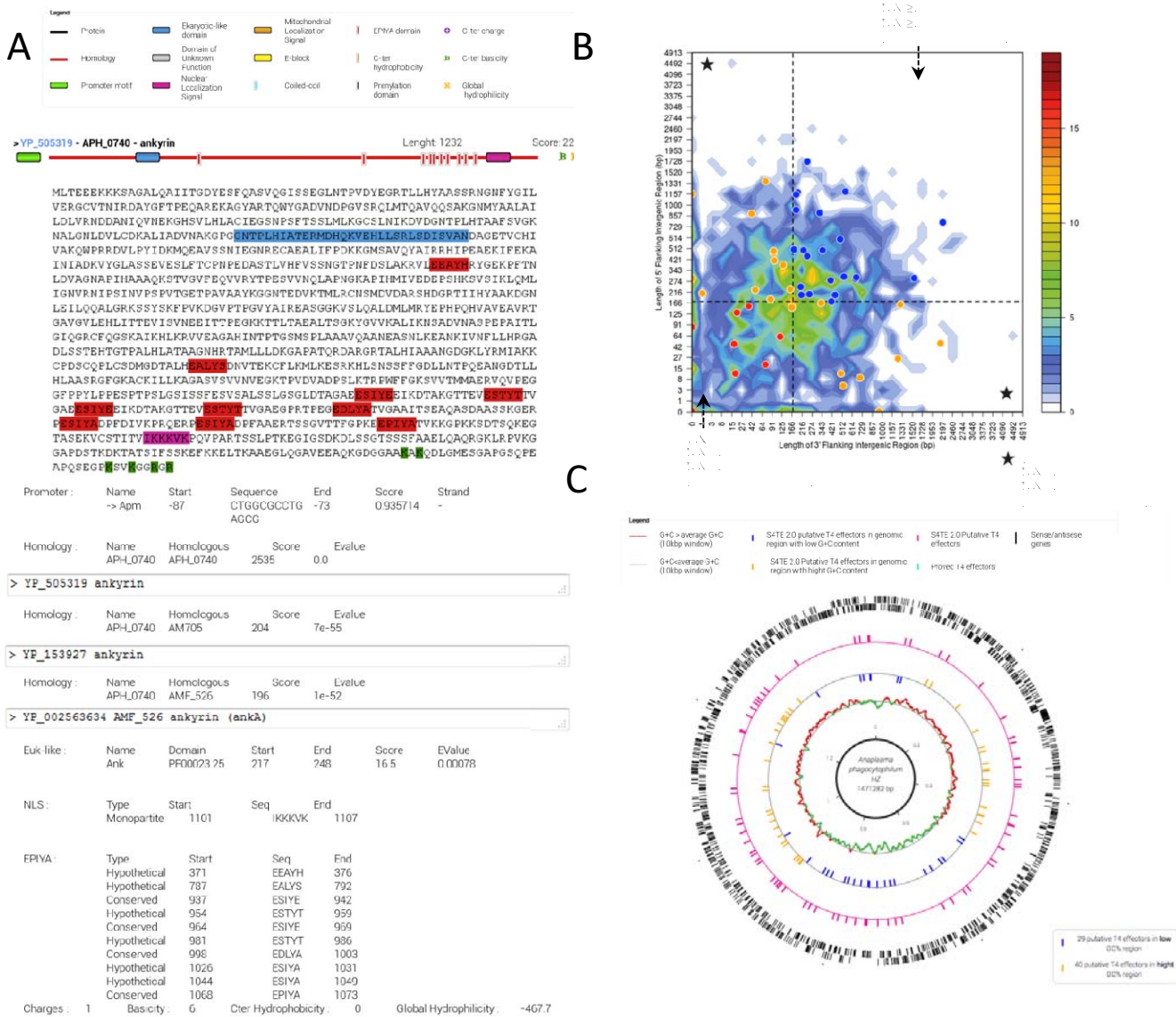


Figure 4.