# Bastion6: a bioinformatics approach for accurate prediction of type VI secreted effectors

## Supplementary Material

### SUPPLEMENTAL INFORMATION

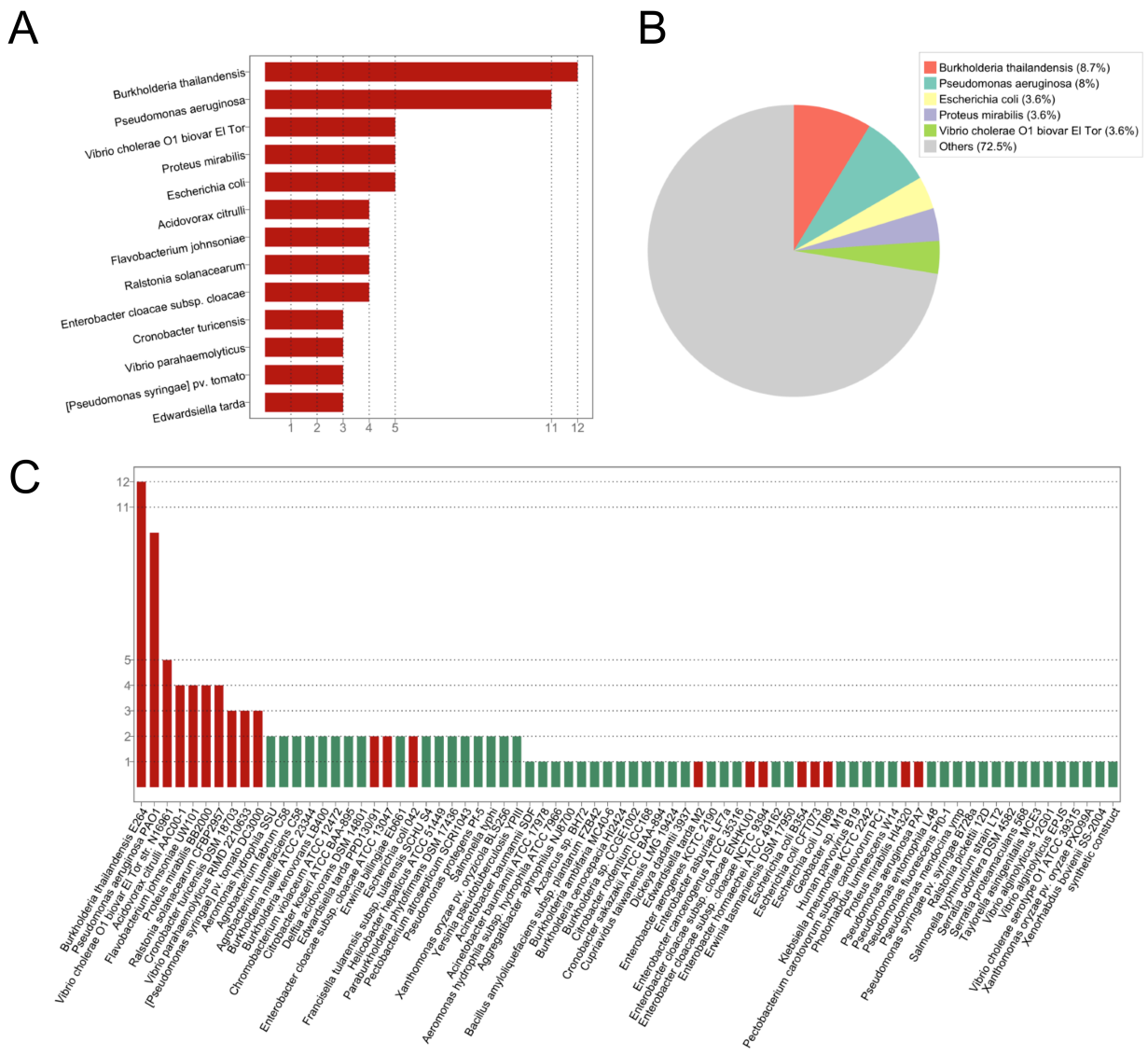


**Fig. 1**. Distribution of proteins in the training dataset by organism. (A) Bar chart listing the numbers of proteins in the training dataset for the top 13 species; (B) Pie chart showing the percentages of major protein associated species; and (C) Bar chart detailing the numbers of proteins per organism.

**Table 1.** Detailed information regarding positive samples in the independent dataset.

| Effector ID | Effector name | Species | Reference | Comments |
|---|---|---|---|---|
| 1 | Hcp-effector | *Desulfobacterium autotrophicum* | (Wang, et al., 2015) | |
| 2 | Hcp3 | *Pseudomonas fluorescens* | (Brunet, et al., 2015) | |
| 3 | EvpP | *Edwardsiella tarda* | (Durand, et al., 2014) | |
| 4 | VgrG3 | *Pseudomonas fluorescens* | (Durand, et al., 2014) | |
| 5 | Tse1 | *Pseudomonas aeruginosa* | (Durand, et al., 2014) | |
| 6 | Tae4 | *Enterobacter cloacae* | (Durand, et al., 2014) | |
| 7 | TecA | *Burkholderia cenocepacia* | (Aubert, et al., 2016) | |
| 8 | VgrG2b | *Pseudomonas aeruginosa* | (Sana, et al., 2015) | |
| 9 | KatN | *Pseudomonas aeruginosa* | (Wan, et al., 2017) | |
| 10 | Tle1 | *Escherichia coli* | (Flaugnatti, et al., 2016) | |
| - | RhsA | *Escherichia coli* | (Koskiniemi, et al., 2013) | Removed due to high similarity with RhsB |
| 11 | RhsB | *Escherichia coli* | (Koskiniemi, et al., 2013) | |
| 12 | Hcp-ET1 | *Escherichia coli* | (Ma, et al., 2017) | |
| 13 | Hcp-ET2 | *Escherichia coli* | (Ma, et al., 2017) | |
| 14 | Hcp-ET3 (1) | *Salmonella bongori* | (Ma, et al., 2017) | |
| - | Hcp-ET3 (2) | *Escherichia coli* | (Ma, et al., 2017) | Removed due to high similarity with Hcp-ET3+4 |
| - | Hcp-ET3 (3) | *Escherichia coli* | (Ma, et al., 2017) | Removed due to high similarity with Hcp-ET3+4 |
| 15 | Hcp-ET3 (4) | *Escherichia coli* | (Ma, et al., 2017) | |
| 16 | Hcp-ET3+4 | *Escherichia coli* | (Ma, et al., 2017) | |
| 17 | Hcp-ET5 | *Salmonella enterica* | (Ma, et al., 2017) | |
| 18 | Unclear | *Escherichia coli* | (Ma, et al., 2017) | |
| 19 | MIX-effector1 | *Vibrio proteolyticus* | (Salomon, 2016) | |

| | | | |
|---|---|---|---|
| 20 | MIX-effector2 | *Vibrio proteolyticus* | (Salomon, 2016) |

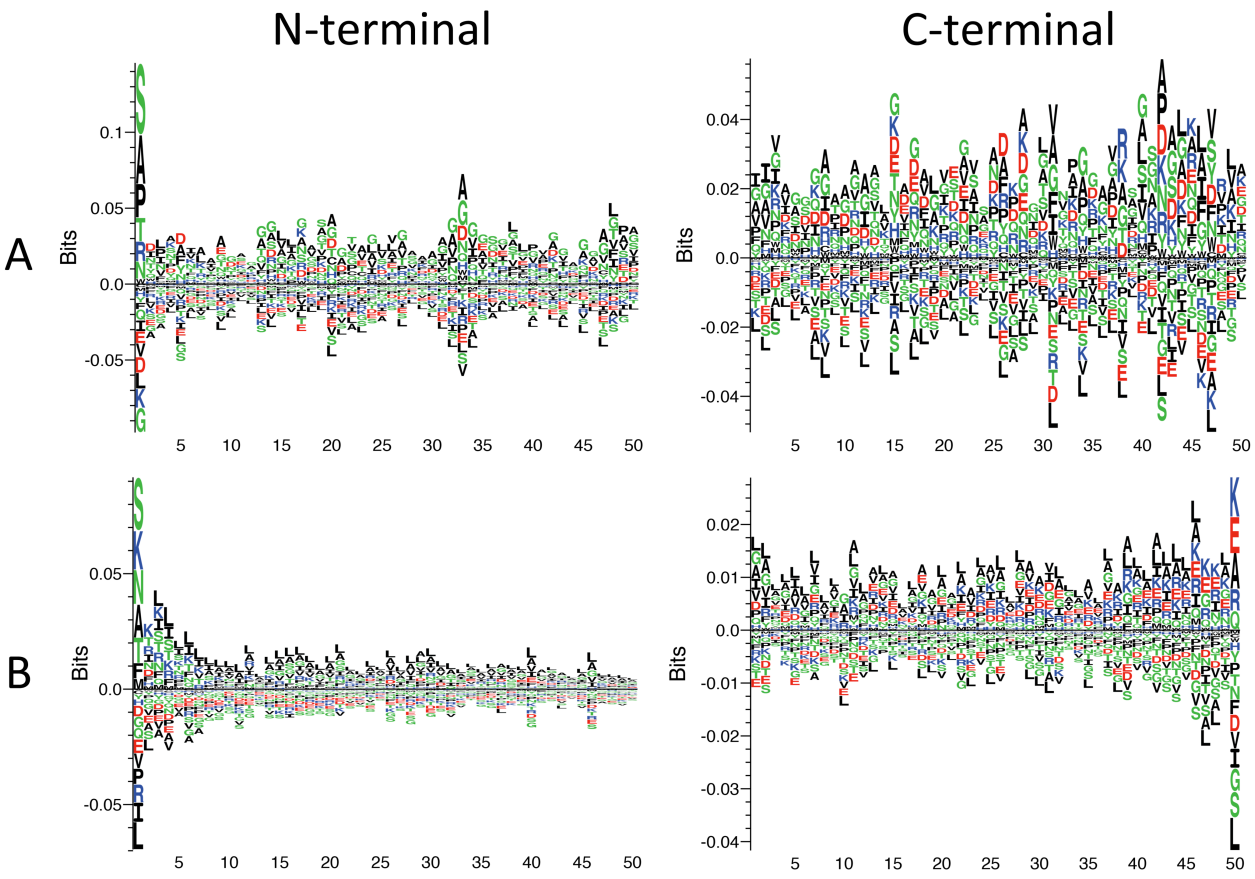


**Fig. 2.** Position-specific amino acid sequence profiles of 138 T6SEs and 1112 non-effectors, for 50 different N- and C-terminal positions. Images were generated with Seq2Logo (Thomsen and Nielsen, 2012) using the default settings. The positive y-axis depicts enriched amino acids in terms of amount of information in bits, while negative y-axis depicts corresponding depleted amino acids. The horizontal axis represents the N-/C-terminal position number. For the N terminal sequences, the methionine (M) at position 1 of each sequence is removed to improve readability. Here, the height of the stack represents the conservation level at each position, while the size of the letters depicts the relative frequency of each amino acid. (A) and (B) illustrate sequence logo representations for T6SEs and non-effectors, respectively.

**Table 2.** Details of distributions of T6SEs and non-T6SEs in each cluster.

| Encoding | Cluster1 | | | Cluster2 | | |
|---|---|---|---|---|---|---|
| | Total | T6SEs | non-T6SEs | Total | T6SEs | non-T6SEs |
| AAC | 115 | 96 (83.5%) | 19 (16.5) | 161 | 42 (26.1%) | 119 (73.9%) |
| DPC | 105 | 92 (87.6%) | 13 (12.4%) | 171 | 46 (26.9%) | 125 (73.1%) |
| QSO | 68 | 49 (72.1%) | 19 (27.9%) | 208 | 89 (42.8%) | 119 (57.2%) |
| BLOSUM | 244 | 107 (43.9%) | 137 (56.1%) | 32 | 31 (96.9%) | 1 (3.1%) |
| DPC-PSSM | 114 | 1 (0.9%) | 113 (99.1%) | 162 | 137 (84.6%) | 25 (15.4%) |
| S-FPSSM | 79 | 54 (68.4%) | 25 (31.6%) | 197 | 84 (42.6%) | 113 (57.4%) |
| Pse-PSSM | 118 | 91 (77.1%) | 27 (22.9%) | 158 | 47 (29.7%) | 111 (70.3%) |
| CTDC | 132 | 78 (59.1%) | 54 (40.9%) | 144 | 60 (41.7%) | 84 (58.3%) |
| CTDT | 238 | 124 (52.1%) | 114 (47.9%) | 38 | 14 (36.8%) | 24 (63.2%) |

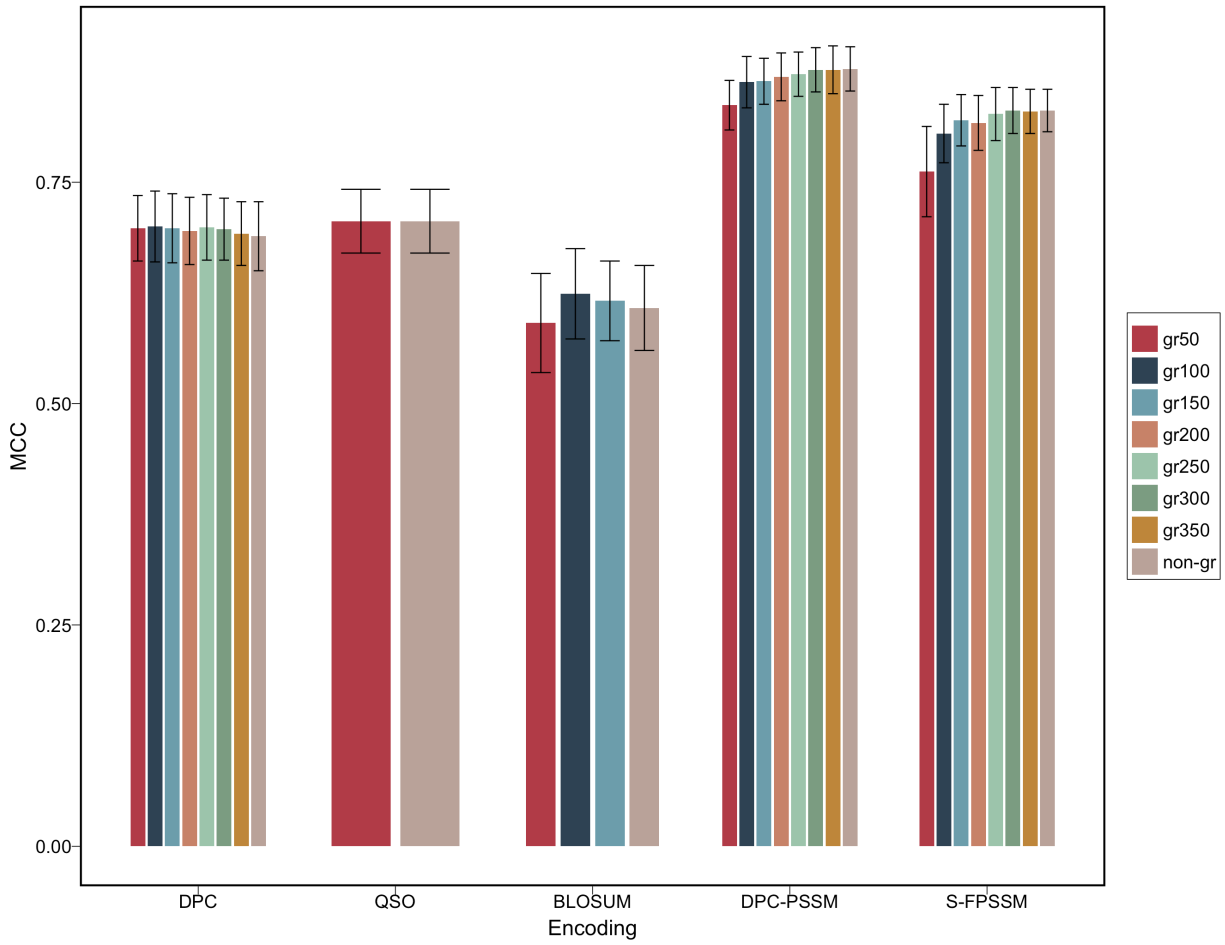


**Fig. 3**. Comparisons of the performance of various feature encoding methods with different numbers of top features, selected by GainRatio based on 5-fold cross-validation tests. grX (X=50,100,150,200,250,300,350) means top X features as ranked by GainRatio, while non-gr means full features without feature selection by GainRatio.

**Table 3.** The detailed prediction performance of various models in the independent test.

| | Model | SN | SP | ACC | F-value | MCC |
|---|---|---|---|---|---|---|
| **Single feature-based models** | AAC | 0.900±0.000 | 0.875±0.075 | 0.887±0.038 | 0.890±0.033 | 0.777±0.074 |
| | DPC | 0.800±0.000 | 0.865±0.097 | 0.833±0.049 | 0.829±0.041 | 0.670±0.101 |
| | QSO | 0.850±0.000 | 0.875±0.072 | 0.863±0.036 | 0.862±0.031 | 0.727±0.074 |
| | BLOSUM | 0.800±0.000 | 0.830±0.101 | 0.815±0.050 | 0.814±0.042 | 0.634±0.105 |
| | DPC-PSSM | 0.950±0.000 | 0.745±0.101 | 0.848±0.051 | 0.863±0.039 | 0.712±0.088 |
| | S-FPSSM | 0.750±0.000 | 0.770±0.116 | 0.760±0.058 | 0.760±0.042 | 0.523±0.115 |
| | Pse-PSSM | 0.950±0.000 | 0.780±0.111 | 0.865±0.056 | 0.878±0.044 | 0.743±0.098 |
| | CTDC | 0.900±0.000 | 0.850±0.094 | 0.875±0.047 | 0.880±0.040 | 0.753±0.090 |
| | CTDT | 0.850±0.000 | 0.795±0.064 | 0.823±0.032 | 0.828±0.026 | 0.647±0.063 |
| **Ensemble model** | Group 1 | 0.850±0.000 | 0.880±0.079 | 0.865±0.039 | 0.864±0.034 | 0.733±0.081 |
| | Group 2 | **1.000±0.000** | 0.825±0.072 | 0.912±0.036 | 0.920±0.030 | 0.839±0.062 |
| | Group 3 | 0.950±0.000 | 0.840±0.088 | 0.895±0.044 | 0.902±0.038 | 0.797±0.082 |
| **Final ensemble model** | Bastion6 | **1.000±0.000** | **0.885±0.053** | **0.943±0.026** | **0.946±0.024** | **0.892±0.049** |

**Table 4.** Prediction results of positive samples from the independent dataset using single encoding method-based models, group based ensemble models and Bastion6. Here, samples with a prediction score larger than 0.5 are recognized as T6SS effectors, and otherwise as non-T6SS effectors (marked in grey).

| Effector ID | Effector Name | Single encoding method based model | | | | | | | | | Ensemble Model | | | Bastion6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AAC | DPC | QSO | BLOSUM | DPC-PSSM | S-FPSSM | Pse-PSSM | CTDC | CTDT | Group 1 | Group 2 | Group 3 | |
| 1 | Hcp | 0.960 | 0.919 | 0.938 | 0.970 | 0.950 | 0.939 | 0.976 | 0.913 | 0.815 | 0.939 | 0.959 | 0.864 | 0.921 |
| 2 | Hcp3 | 0.658 | 0.854 | 0.819 | 0.482 | 0.980 | 0.841 | 0.961 | 0.440 | 0.715 | 0.777 | 0.816 | 0.577 | 0.723 |
| 3 | EvpP | 0.752 | 0.850 | 0.838 | 0.787 | 0.966 | 0.864 | 0.905 | 0.568 | 0.585 | 0.813 | 0.880 | 0.576 | 0.757 |
| 4 | VgrG3 | 0.921 | 0.940 | 0.977 | 0.931 | 0.987 | 0.961 | 0.948 | 0.719 | 0.750 | 0.946 | 0.957 | 0.735 | 0.879 |
| 5 | Tse1 | 0.831 | 0.708 | 0.843 | 0.710 | 0.963 | 0.939 | 0.981 | 0.905 | 0.781 | 0.794 | 0.898 | 0.843 | 0.845 |
| 6 | Tae4 | 0.842 | 0.923 | 0.848 | 0.948 | 0.992 | 0.968 | 0.968 | 0.897 | 0.798 | 0.871 | 0.969 | 0.848 | 0.896 |
| 7 | TecA | 0.575 | 0.423 | 0.313 | 0.299 | 0.987 | 0.495 | 0.819 | 0.693 | 0.768 | 0.437 | 0.650 | 0.730 | 0.606 |
| 8 | VgrG2b | 0.934 | 0.936 | 0.941 | 0.845 | 0.969 | 0.993 | 0.959 | 0.838 | 0.810 | 0.937 | 0.941 | 0.824 | 0.901 |
| 9 | KatN | 0.684 | 0.838 | 0.746 | 0.643 | 0.830 | 0.186 | 0.361 | 0.803 | 0.929 | 0.756 | 0.505 | 0.866 | 0.709 |
| 10 | Tle1 | 0.406 | 0.572 | 0.536 | 0.482 | 0.336 | 0.450 | 0.824 | 0.746 | 0.460 | 0.505 | 0.523 | 0.603 | 0.543 |
| 11 | RhsB | 0.987 | 0.979 | 0.987 | 0.935 | 0.957 | 0.707 | 0.950 | 0.961 | 0.929 | 0.985 | 0.888 | 0.945 | 0.939 |
| 12 | Hcp-ET1 | 0.584 | 0.375 | 0.210 | 0.491 | 0.956 | 0.909 | 0.848 | 0.632 | 0.413 | 0.390 | 0.801 | 0.522 | 0.571 |
| 13 | Hcp-ET2 | 0.507 | 0.475 | 0.572 | 0.829 | 0.767 | 0.391 | 0.570 | 0.588 | 0.386 | 0.518 | 0.639 | 0.487 | 0.548 |
| 14 | Hcp-ET3 (1) | 0.878 | 0.909 | 0.939 | 0.956 | 0.928 | 0.725 | 0.754 | 0.842 | 0.675 | 0.909 | 0.841 | 0.758 | 0.836 |
| 15 | Hcp-ET3 (4) | 0.926 | 0.932 | 0.971 | 0.940 | 0.968 | 0.967 | 0.923 | 0.903 | 0.792 | 0.943 | 0.949 | 0.848 | 0.913 |
| 16 | Hcp-ET3+4 | 0.707 | 0.741 | 0.889 | 0.793 | 0.969 | 0.777 | 0.812 | 0.484 | 0.647 | 0.779 | 0.838 | 0.565 | 0.728 |
| 17 | Hcp-ET5 | 0.679 | 0.501 | 0.514 | 0.893 | 0.973 | 0.242 | 0.745 | 0.662 | 0.713 | 0.565 | 0.713 | 0.688 | 0.655 |
| 18 | Unclear | 0.432 | 0.446 | 0.374 | 0.850 | 0.980 | 0.847 | 0.873 | 0.851 | 0.641 | 0.417 | 0.887 | 0.746 | 0.684 |
| 19 | MIX-effector1 | 0.989 | 0.965 | 0.976 | 0.707 | 0.950 | 0.982 | 0.993 | 0.953 | 0.906 | 0.976 | 0.908 | 0.930 | 0.938 |
| 20 | MIX-effector2 | 0.987 | 0.920 | 0.953 | 0.783 | 0.945 | 0.976 | 0.983 | 0.888 | 0.689 | 0.953 | 0.922 | 0.789 | 0.888 |

**Table 5.** Detailed prediction results of Bastion 6 and of two motif-based methods for positive samples in the independent dataset. Misclassified proteins are marked in grey.

| Effector ID | Effector Name | Bastion6 | | Motif Methods | |
|---|---|---|---|---|---|
| | | | | MIX | SAVC |
| 1 | Hcp | 0.921 | √ | × | × |
| 2 | Hcp3 | 0.723 | √ | × | × |
| 3 | EvpP | 0.757 | √ | × | × |
| 4 | VgrG3 | 0.879 | √ | × | × |
| 5 | Tse1 | 0.845 | √ | × | × |
| 6 | Tae4 | 0.896 | √ | × | × |
| 7 | TecA | 0.606 | √ | × | × |
| 8 | VgrG2b | 0.901 | √ | × | × |
| 9 | KatN | 0.709 | √ | × | × |
| 10 | Tle1 | 0.543 | √ | × | × |
| 11 | RhsB | 0.939 | √ | × | √ |
| 12 | Hcp-ET1 | 0.571 | √ | × | × |
| 13 | Hcp-ET2 | 0.548 | √ | × | √ |
| 14 | Hcp-ET3 (1) | 0.836 | √ | × | × |
| 15 | Hcp-ET3 (4) | 0.913 | √ | × | × |
| 16 | Hcp-ET3+4 | 0.728 | √ | × | × |
| 17 | Hcp-ET5 | 0.655 | √ | × | × |
| 18 | Unclear | 0.684 | √ | × | × |
| 19 | MIX-effector1 | 0.938 | √ | × | × |
| 20 | MIX-effector2 | 0.888 | √ | × | × |

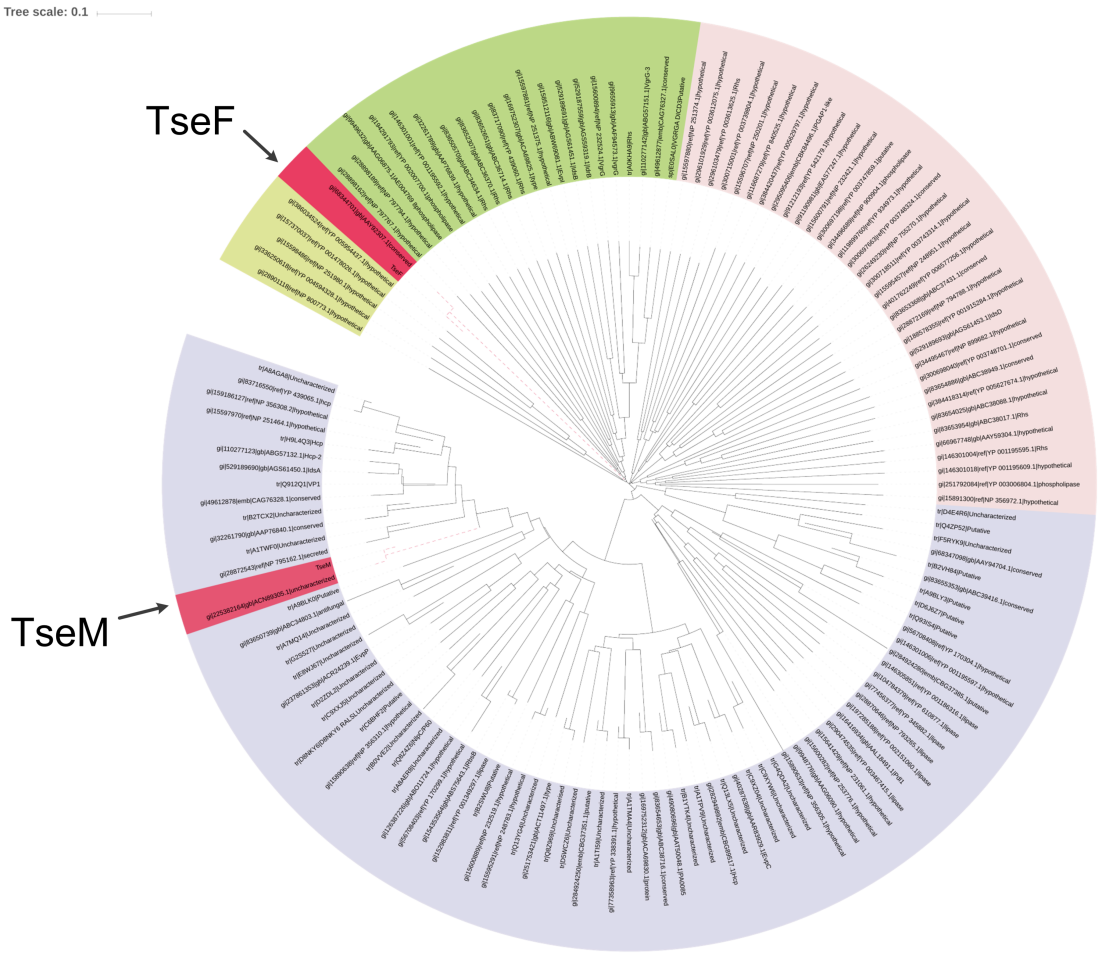


**Fig. 4**. Phylogenetic tree of all T6SEs in the training dataset and the two case study proteins TseM and TseF. Multiple sequence alignment was constructed for all the included proteins using Clustal Omega (Li, et al., 2015), with the phylogenetic tree generated using iTOL (Letunic and Bork, 2016).

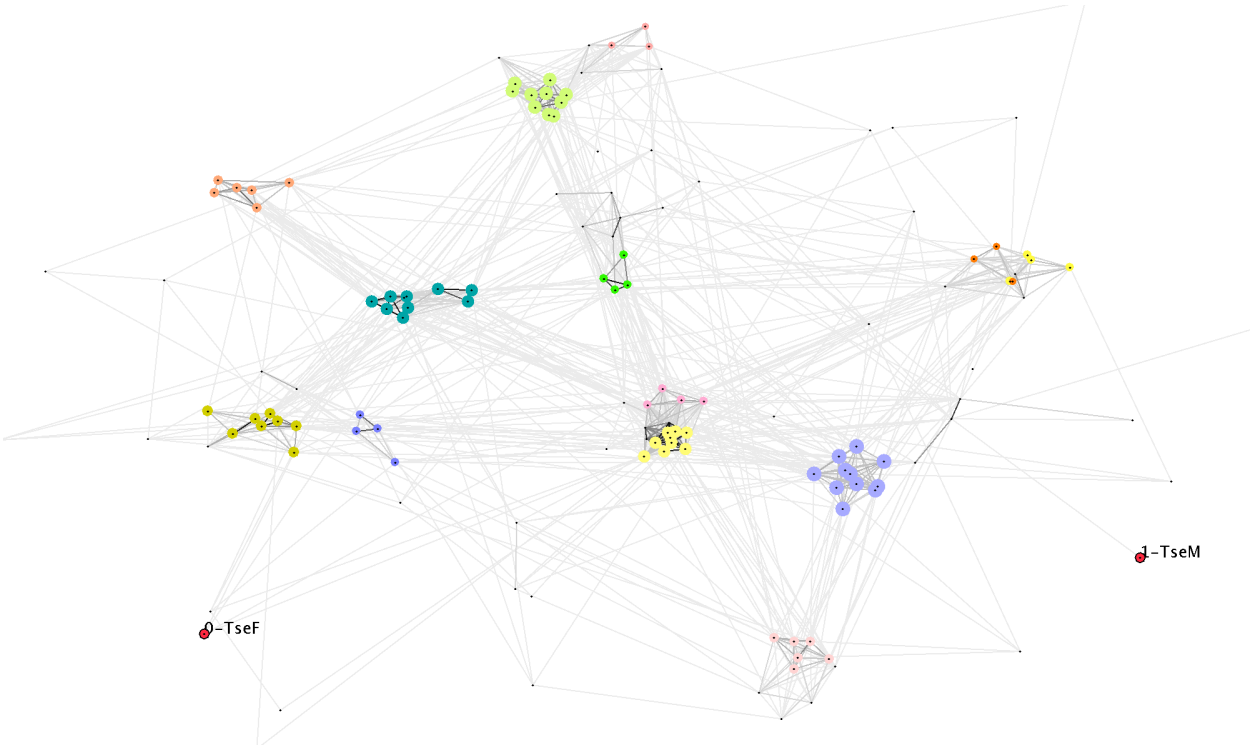


**Fig. 5.** Graphical two-dimensional representation of sequence similarities between the T6SS effectors of the training dataset and two case study effectors using the software CLANS. To draw a three-dimensional graph (projected here onto two dimensions), we performed all-against-all BLAST searches and used all significant high-scoring segment pairs (HSPs). In the graph, each node represents a T6SS effector protein and each edge (shaded according to p-value) represents a significant HSP with a p-value lower than 0.05 Each cluster is highlighted, while TseM and TseF are marked in the graph.

**Table 6.** Detailed prediction results of single encoding method based models, group based ensemble models and Bastion6, for two case study T6SS effector sequences.

| Effector Name | Single encoding method based model | | | | | | | | | Ensemble Model | | | Bastion6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AAC | DPC | QSO | BLOSUM | DPC-PSSM | S-FPSSM | Pse-PSSM | CTDC | CTDT | Group 1 | Group 2 | Group 3 | |
| TseM | 0.351 | 0.497 | 0.390 | 0.267 | 0.728 | 0.445 | 0.734 | 0.585 | 0.763 | 0.413 | 0.544 | 0.674 | 0.544 |
| TseF | 0.809 | 0.808 | 0.606 | 0.396 | 0.168 | 0.902 | 0.500 | 0.815 | 0.805 | 0.741 | 0.491 | 0.810 | 0.681 |

**Table 7. Statistics of T6SE prediction results from 54,212 sequences of 12 bacterial species scanned by Bastion6.** We list results using different thresholds, noting all results were filtered by readily validated T6SS effectors.

| Species | Total number | >=0.5 | >=0.6 | >=0.7 | >=0.8 | >=0.9 |
|---|---|---|---|---|---|---|
| *Acidovorax citrulli strain* AAC00-1 | 4652 | 925 | 495 | 225 | 83 | 29 |
| *Klebsiella pneumoniae* AJ218 | 5108 | 524 | 299 | 142 | 47 | 4 |
| *Klebsiella pneumoniae* B5055 | 5198 | 552 | 308 | 131 | 34 | 3 |
| *Burkholderia thailandensis* E264 | 5763 | 954 | 497 | 240 | 89 | 14 |
| *Cronobacter turicensis* z3032 | 3987 | 556 | 303 | 173 | 68 | 11 |
| *Flavobacterium johnsoniae* UW101 | 5101 | 1009 | 594 | 284 | 70 | 6 |
| *Legionella pneumoniae* Phi1 | 2943 | 212 | 88 | 34 | 6 | 0 |
| *Klebsiella pneumoniae* MGH78578 | 4859 | 495 | 274 | 132 | 36 | 3 |
| *Proteus mirabilis* BB2000 | 3325 | 364 | 211 | 111 | 40 | 1 |
| *Pseudomonas aeruginosa* PAO1 | 5558 | 690 | 388 | 198 | 94 | 12 |
| *Ralstonia solanacearum* CFBP2957 | 3174 | 481 | 216 | 90 | 17 | 3 |
| *Vibrio parahaemolyticus* RIMD 2210633 | 4544 | 530 | 309 | 162 | 67 | 8 |

# Reference


Aubert, D.F.*, et al.* A Burkholderia Type VI Effector Deamidates Rho GTPases to Activate the Pyrin Inflammasome and Trigger Inflammation. *Cell host & microbe* 2016;19(5):664-674.

Brunet, Y.R.*, et al.* The Type VI secretion TssEFGK-VgrG phage-like baseplate is recruited to the TssJLM membrane complex via multiple contacts and serves as assembly platform for tail tube/sheath polymerization. *PLoS Genet* 2015;11(10):e1005545.

Durand, E.*, et al.* VgrG, Tae, Tle, and beyond: the versatile arsenal of Type VI secretion effectors. *Trends in microbiology* 2014;22(9):498-507.

Flaugnatti, N.*, et al.* A phospholipase A1 antibacterial Type VI secretion effector interacts directly with the C-terminal domain of the VgrG spike protein for delivery. *Molecular microbiology* 2016;99(6):1099-1118.

Koskiniemi, S.*, et al.* Rhs proteins from diverse bacteria mediate intercellular competition. *Proceedings of the National Academy of Sciences of the United States of America* 2013;110(17):7032-7037.

Letunic, I. and Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic acids research* 2016;44(W1):W242-245.

Li, W.*, et al.* The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic acids research* 2015;43(W1):W580-584.

Ma, J.*, et al.* The Hcp proteins fused with diverse extended-toxin domains represent a novel pattern of antibacterial effectors in type VI secretion systems. *Virulence* 2017:1-14.

Salomon, D. MIX and match: mobile T6SS MIX-effectors enhance bacterial fitness. *Mobile genetic elements* 2016;6(1):e1123796.

Sana, T.G.*, et al.* Internalization of Pseudomonas aeruginosa Strain PAO1 into Epithelial Cells Is Promoted by Interaction of a T6SS Effector with the Microtubule Network. *mBio* 2015;6(3):e00712.

Thomsen, M.C. and Nielsen, M. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic acids research* 2012;40(Web Server issue):W281-287.

Wan, B.*, et al.* Type VI secretion system contributes to Enterohemorrhagic Escherichia coli virulence by secreting catalase against host reactive oxygen species (ROS). *PLoS pathogens* 2017;13(3):e1006246.

Wang, N.*, et al.* Protective efficacy of recombinant hemolysin co-regulated protein (Hcp) of Aeromonas hydrophila in common carp (Cyprinus carpio). *Fish & shellfish immunology* 2015;46(2):297-304.