Supporting Information

Salomon et al. 10.1073/pnas.1406110111

SI Materials and Methods

Bacterial Strains and Media. The *Vibrio parahaemolyticus* RIMD 2210633 derivative strain POR1 (RIMD 2210633 Δ*tdhAS*) (1) and its derivatives, as well as *Vibrio alginolyticus* 12G01 and its derivatives, were routinely cultured in marine LB (MLB) broth [LB broth containing 3% (wt/vol) sodium chloride] or on marine minimal media agar (2) at 30 °C. *Escherichia coli* DH5α and *E. coli* S17 (λ pir) were routinely cultured in 2× yeast-extract and tryptone broth at 37 °C. The medium was supplemented with kanamycin (30 μg/mL for *E. coli* and 250 μg/mL for *Vibrios* species) or chloramphenicol (25 μg/mL), when necessary, to maintain a plasmid.

Plasmids. For arabinose-inducible expression of *vp1389*, *vp1416*, and *vti2*, the genes coding the sequences were amplified and cloned into the pBAD/*Myc*-His vector (Invitrogen), in which the antibiotic resistance was changed from ampicillin to kanamycin, or into the pBAD33 vector (Addgene). Empty pBAD/*Myc*-His and pBAD33 vectors were used to provide resistance to kanamycin and chloramphenicol, respectively.

Construction of Deletion Strains. Gene deletions were performed as previously described using the pDM4 vector (3, 4) for *vp1388*, *vp1388–vp1389*, *vp1415*, *vp1415–vp1416*, *vpa1263*, *vpa1263–vti2*, *v12G01_02265*, *v12G01_02265–v12G01_02260*, and *v12G01_01540* (*V. alginolyticus hcp1*). Construction of POR1Δ*hcp1* was described previously (4).

Bacterial Killing Assays. Bacterial killing assays were performed as previously described (4). Cells were cocultured at an OD_{600} ratio of 1:4 (prey/attacker) on MLB (for *V. parahaemolyticus* attacker) or LB (for *V. alginolyticus* attacker) agar plates for 4 h at 30 °C in triplicate. Where appropriate, plates were supplemented with 0.1% arabinose to induce expression from plasmids. Each experiment was done in triplicate and repeated at least twice, with similar results. A representative experiment is presented. A two-tailed Student *t* test was used to determine significance between indicated sample groups unless stated otherwise.

Proteomics and MS. *V. parahaemolyticus* cultures of POR1Δ*opaR* [type 6 secretion system (T6SS1)$^+$] and POR1Δ*opaR*Δ*hcp1* (T6SS1$^-$) were grown in triplicate in 50 mL of M9 minimal media with 3% (wt/vol) NaCl media containing 20 μM phenamil for 5 h at 30 °C. Media were collected, and protein was precipitated as previously described (4). Protein samples were run 10 mm into the top of an SDS/PAGE gel, stained with Coomassie Blue, and excised. Overnight digestion with trypsin (Promega) was performed after reduction and alkylation with DTT and iodoacetamide (Sigma–Aldrich). The resulting samples were analyzed by tandem MS using a QExactive mass spectrometer (Thermo Electron) coupled to an Ultimate 3000 RSLC-Nano liquid chromatography system (Dionex). Peptides were loaded onto a 180–μm i.d., 15-cm long, self-packed column containing 1.9 μm C18 resin (Dr. Maisch, Ammerbuch, Germany) and eluted with a gradient of 0–40% buffer B for 60 min. Buffer A consisted of 2% (vol/vol) acetonitrile (ACN) and 0.1% formic acid in water. Buffer B consisted of 80% (vol/vol) ACN, 10% (vol/vol) trifluoroethanol, and 0.08% formic acid in water. The QExactive mass spectrometer acquired up to 10 high-energy, collision-induced dissociation fragment spectra for each full spectrum acquired.

Raw MS data files were converted to peak list format using ProteoWizard msconvert (version 3.0.3535) (5). The resulting files were analyzed using the central proteomics facilities pipeline (CPFP), version 2.0.3 (6, 7). Peptide identification was performed using the X!Tandem (8) and open MS search algorithm (OMSSA) (9) search engines against a database consisting of *V. parahaemolyticus* RIMD 2210633 sequences from the National Center for Biotechnology Information, with common contaminants and reversed decoy sequences appended (10). Fragment and precursor tolerances of 20 ppm and 0.1 Da were specified, and three missed cleavages were allowed. Carbamidomethylation of Cys was specified as a fixed modification, and oxidation of Met was specified as a variable modification. Label-free quantitation of proteins across samples was performed using SINQ normalized spectral index software (11).

To identify statistically significant differences in protein amount between T6SS1$^+$ and T6SS1$^-$ strains, SINQ quantitation results for three biological replicates per strain were processed using the power law global error model (PLGEM) package in R (12, 13). Protein identifications were filtered to an estimated 1% protein false discovery rate (FDR) using the concatenated target-decoy method (10). An additional requirement of two unique peptide sequences per protein was imposed, resulting in a final protein FDR <1%. Spectral index quantitation was performed using peptide-to-spectrum matches (PSMs) with a q-value ≤0.01, corresponding to a 1% FDR rate for PSMs.

The dataset of tandem MS results was uploaded to the PeptideAtlas repository (www.peptideatlas.org/PASS/PASS00442; accession no. PASS00442).

Bioinformatics. The N-terminal sequence from VP1388 (GI|28898162, range 144–297) identifies conserved sequence motifs (PhhPhR and GhhYhhh) near the N terminus of numerous bacterial sequences. One of these sequences (GI|520945140, range 1–249) was queried against the nonredundant (NR) database using the position-specific iterative (PSI)-BLAST application of BLAST+ (14) [10 iterations, E-value cutoff of 0.005, collecting all subject GIs (NCBI protein sequence identifier)] to collect all potential MIX (marker for type six effectors) sequences. Collected sequences were filtered to include only those from complete genomes in the Reference Sequences (RefSeq) database (15) and were clustered using CLANS (16). Sequences from five resulting broad groups were independently aligned using the MAFFT server (17), and were colored according to conservation using Jalview (18) to highlight hydrophobicity patterns surrounding the conserved motifs. The corresponding domain sequences from each of the five multiple sequence alignments were aligned manually using the motifs and surrounding hydrophobicity patterns and PROMALS3D (19) alignments of representatives as a guide. Motifs were highlighted with WebLogo (20).

The domain organization of representative sequences was defined using Batch Entrez and conserved domains (CD)-search (21) with default cutoffs. Identified domains were sorted according to E-value, and incomplete domains with E-values >1.6e-4 were excluded from counts (excluded partial domains include mainly low-complexity sequence, such as coiled coils and transmembrane helices). Two low-complexity coil domains with E-values better than the cutoff were also excluded. Domains identified within this lower E-value range ($0.01–1.6e^{-4}$) that were also identified in other sequences with higher confidence were included in the counts (i.e., 38 LysM domains were identified with an E-value range from 0.00015 to $6.8e^{-14}$). Unknown

sequences represent 855 sequences that identified no additional domains and 58 sequences that identified domains with low-complexity regions that were excluded. Given the presence of a number of excluded transmembrane helix (TMH)-containing domain predictions, Phobius (22) was used to predict TMHs in the C-terminal sequence of all unknown MIX-containing proteins. Several of the sequences with TMH-containing domains predicted by CD-search that were excluded due to low confidence were submitted to the HHpred server (23) for domain validation. All of the submitted sequences identified various colicin pore-forming toxins (ranging from 9–93.3% probability) in their predicted TMH-containing regions.

T6SS components VrgG (GI|28898168, range 1–525) and haemolysin coregulated proteins (Hcp) (GI|256599595, range 1–163) were queried against the NR database (two iterations with an E-value cutoff of 0.0001 and 10 iterations with an E-value cutoff of 0.001, respectively, collecting all subject GIs) using PSI-BLAST (14). Subject sequences were filtered to include only those from complete genomes in the RefSeq database (15), and VrgG sequences were clustered using CLANS (16) to purge false-positive hits. Species containing the resulting VrgG and Hcp sequence representatives were sorted and compared with those containing MIX using GeneVenn (24).

Genomic neighborhoods encompassing the T6SS machinery were explored using the Microbial Genome Database for Comparative Analysis (25) and the Search Tool for the Retrieval of Interacting Genes (STRING) (26). The STRING database v9.1 describing protein association networks was used to generate genome neighborhood information for (*i*) core T6SS components and (*ii*) MIX-containing groups. For the core T6SS components, a network was generated in the clusters of orthologous groups (COG) mode from Hcp (COG3157). Only neighborhood and gene fusion scores with a high-confidence cutoff (>0.7) were considered for generating the T6SS network. MIX sequences for each group were submitted to the STRING database using "multiple sequences" submission in the COG mode to identify relevant groups for neighborhood analysis. The nonorthologous group populated with the most MIX sequences was selected for neighborhood analysis, which was limited to neighborhood and gene fusion scores with a medium-confidence cutoff for network generation. For each of the five MIX groups, combined STRING scores were ranked, and all links with high confidence scores are shown in Table 1. For links of lower confidence, only those scores up to and including the top core T6SS component were reported. For the MIX II group, a low-confidence link to a T6SS component was also reported.

1. Park KS, et al. (2004) Functional characterization of two type III secretion systems of Vibrio parahaemolyticus. *Infect Immun* 72(11):6659–6665.
2. Eagon RG (1962) Pseudomonas natriegens, a marine bacterium with a generation time of less than 10 minutes. *Journal of Bacteriology* 83:736–737.
3. Salomon D, et al. (2013) Effectors of animal and plant pathogens use a common domain to bind host phosphoinositides. *Nat Commun* 4:2973.
4. Salomon D, Gonzalez H, Updegraff BL, Orth K (2013) Vibrio parahaemolyticus type VI secretion system 1 is activated in marine conditions to target bacteria, and is differentially regulated from system 2. *PLoS ONE* 8(4):e61086.
5. Kessner D, Chambers M, Burke R, Agus D, Mallick P (2008) ProteoWizard: Open source software for rapid proteomics tools development. *Bioinformatics* 24(21):2534–2536.
6. Trudgian DC, Mirzaei H (2012) Cloud CPFP: A shotgun proteomics data analysis pipeline using cloud and high performance computing. *J Proteome Res* 11(12): 6282–6290.
7. Trudgian DC, et al. (2010) CPFP: A central proteomics facilities pipeline. *Bioinformatics* 26(8):1131–1132.
8. Craig R, Beavis RC (2004) TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics* 20(9):1466–1467.
9. Geer LY, et al. (2004) Open mass spectrometry search algorithm. *J Proteome Res* 3(5): 958–964.
10. Elias JE, Gygi SP (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* 4(3):207–214.
11. Trudgian DC, et al. (2011) Comparative evaluation of label-free SINQ normalized spectral index quantitation in the central proteomics facilities pipeline. *Proteomics* 11(14):2790–2797.
12. Pavelka N, et al. (2008) Statistical similarities between transcriptomics and quantitative shotgun proteomics data. *Mol Cell Proteomics* 7(4):631–644.
13. Pavelka N, et al. (2004) A power law global error model for the identification of differentially expressed genes in microarray data. *BMC Bioinformatics* 5:203.
14. Camacho C, et al. (2009) BLAST+: Architecture and applications. *BMC Bioinformatics* 10:421.
15. Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy. *Nucleic Acids Res* 40(Database issue):D130–D135.
16. Frickey T, Lupas A (2004) CLANS: A Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* 20(18):3702–3704.
17. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* 30(4):772–780.
18. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview Version 2—A multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25(9): 1189–1191.
19. Pei J, Tang M, Grishin NV (2008) PROMALS3D web server for accurate multiple protein sequence and structure alignments. *Nucleic Acids Res* 36(Web Server issue):W30–W34.
20. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: A sequence logo generator. *Genome Res* 14(6):1188–1190.
21. Marchler-Bauer A, et al. (2011) CDD: A Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* 39(Database issue):D225–D229.
22. Käll L, Krogh A, Sonnhammer EL (2004) A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* 338(5):1027–1036.
23. Hildebrand A, Remmert M, Biegert A, Söding J (2009) Fast and accurate automatic structure prediction with HHpred. *Proteins* 77(Suppl 9):128–132.
24. Pirooznia M, Nagarajan V, Deng Y (2007) GeneVenn—A web application for comparing gene lists using Venn diagrams. *Bioinformation* 1(10):420–422.
25. Uchiyama I, Mihara M, Nishide H, Chiba H (2013) MBGD update 2013: The microbial genome database for exploring the diversity of microbial world. *Nucleic Acids Res* 41(Database issue):D631–D635.
26. Franceschini A, et al. (2013) STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 41(Database issue):D808–D815.

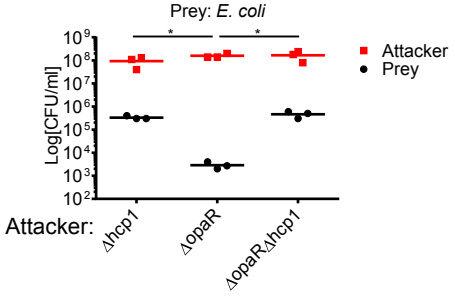

Fig. S1. *V. parahaemolyticus* Δ*opaR* strain kills *E. coli* in a T6SS1-dependent manner. Viability counts of *E. coli* prey 4 h after coculturing with *V. parahaemolyticus* POR1Δ*hcp1*, POR1Δ*opaR*, or POR1Δ*hcp1*Δ*opaR* attacker strains. Mixed cultures at an OD_{600} ratio of 1:4 (prey/attacker) were spotted on MLB plates at 30 °C for 4 h. Asterisks mark statistical significance between prey sample groups at $t = 4$ h ($P < 0.05$).

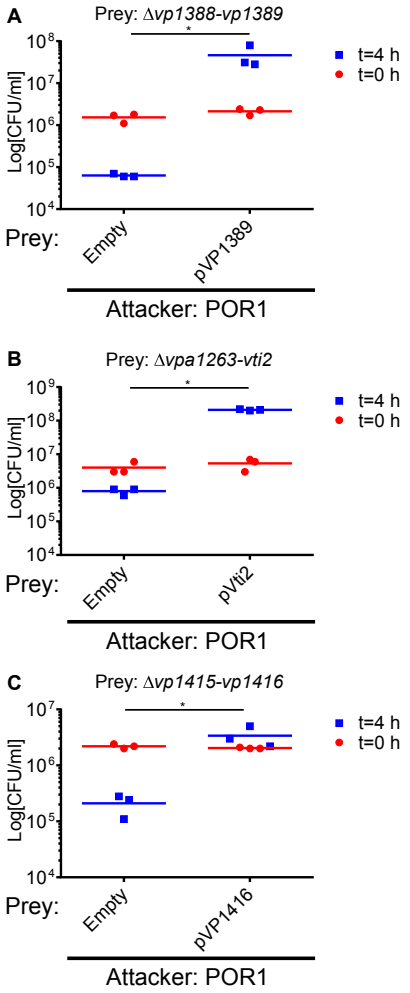

Fig. S2. *vp1389*, *vti2*, and *vp1416* encode T6SS immunity proteins. Viability counts of prey strains before (0 h) and after (4 h) coculture with the attacker strain POR1 mixed at an OD_{600} ratio of 1:4 (prey/attacker) on MLB plates containing 0.1% arabinose at 30 °C are shown. Prey strains used were as follows: POR1Δ*vp1388-vp1389* (*A*), POR1Δ*vpa1263-vti2* (*B*), and POR1Δ*vp1415-vp1416* (C); they contained an arabinose-inducible expression vector that was either empty or encoding the putative immunity protein. Asterisks mark statistical significance between sample groups at $t = 4$ h based on a one-tailed Student *t* test ($P < 0.05$).

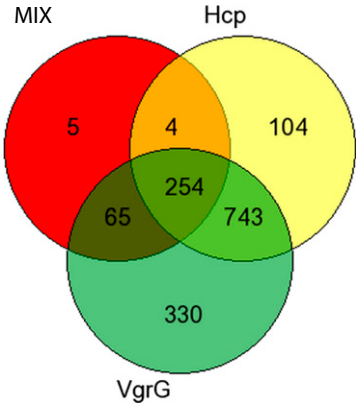

Fig. S3. MIX-containing proteins are found in bacterial species that encode additional T6SS components. A Venn diagram comparing species with MIX-containing proteins (red) identified in the RefSeq database of complete genomes with those containing valine-glycine repeat protein G (VgrG) domains (green) and Hcp domains (yellow) is shown.

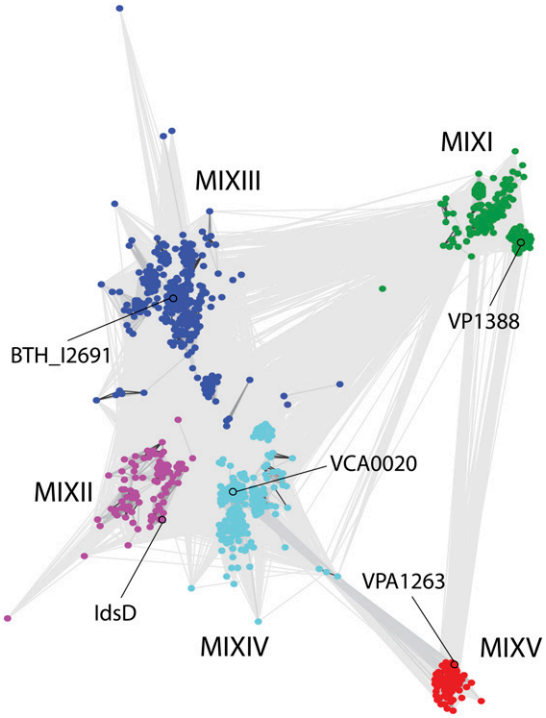

Fig. S4. MIX regions are grouped into five clans. Pairwise BLAST scores of all identified MIX sequences were used to cluster similar sequences into groups depicted in a 2D representation. Nodes are colored and labeled according to visual grouping: MIX I (green), MIX II (magenta), MIX III (blue), MIX IV (cyan), and MIX V (red), with connections between nodes representing pairwise BLAST scores better than E = 0.0001. MIX sequences of known T6SS effectors are labeled.

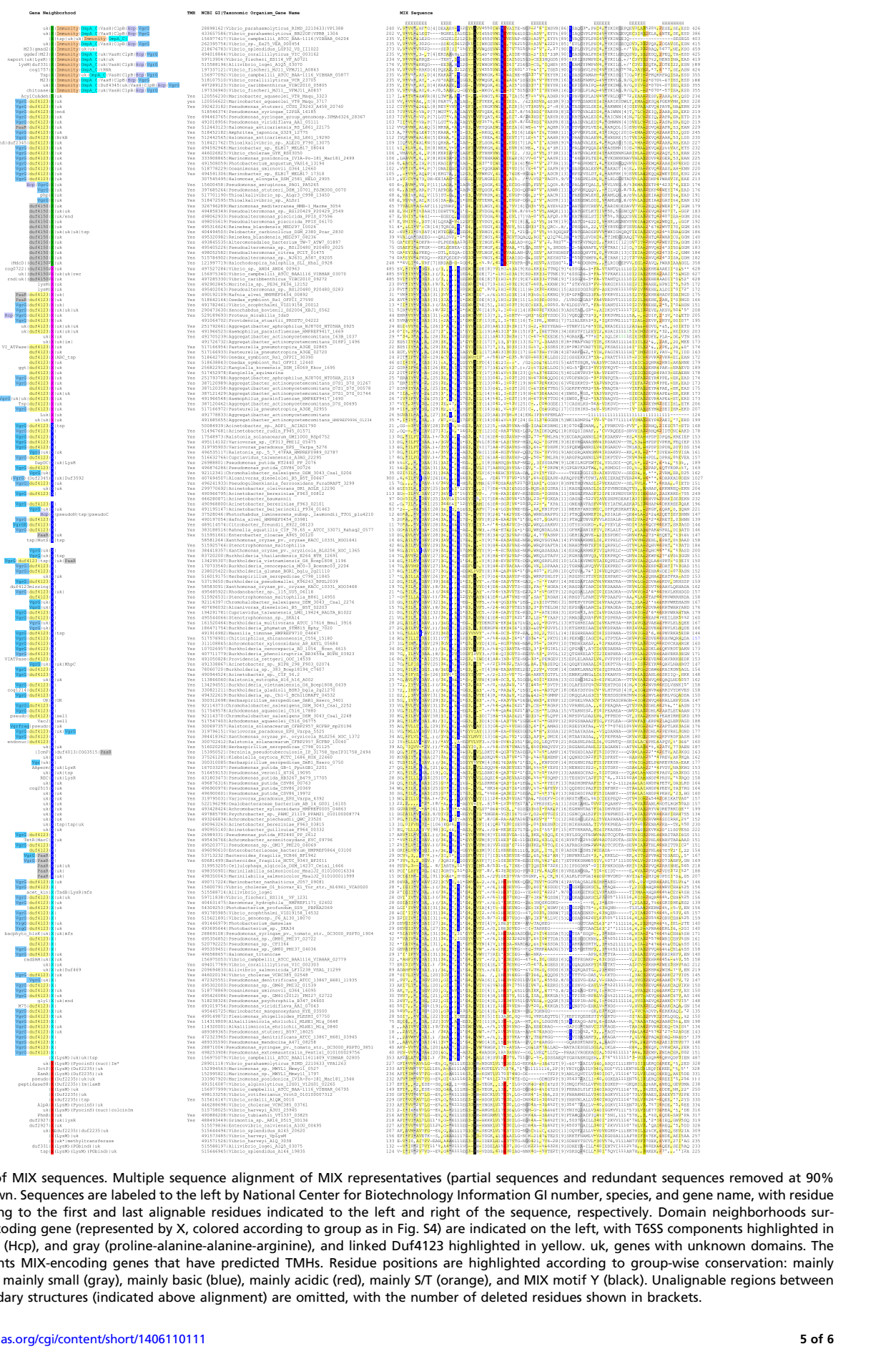

Fig. S5. Alignment of MIX sequences. Multiple sequence alignment of MIX representatives (partial sequences and redundant sequences removed at 90% identity cutoff) is shown. Sequences are labeled to the left by National Center for Biotechnology Information GI number, species, and gene name, with residue numbers corresponding to the first and last alignable residues indicated to the left and right of the sequence, respectively. Domain neighborhoods surrounding the MIX-encoding gene (represented by X, colored according to group as in Fig. S4) are indicated on the left, with T6SS components highlighted in blue (VgrG), lavender (Hcp), and gray (proline-alanine-alanine-arginine), and linked Duf4123 highlighted in yellow. uk, genes with unknown domains. The TMH column represents MIX-encoding genes that have predicted TMHs. Residue positions are highlighted according to group-wise conservation: mainly hydrophobic (yellow), mainly small (gray), mainly basic (blue), mainly acidic (red), mainly S/T (orange), and MIX motif Y (black). Unalignable regions between conserved core secondary structures (indicated above alignment) are omitted, with the number of deleted residues shown in brackets.

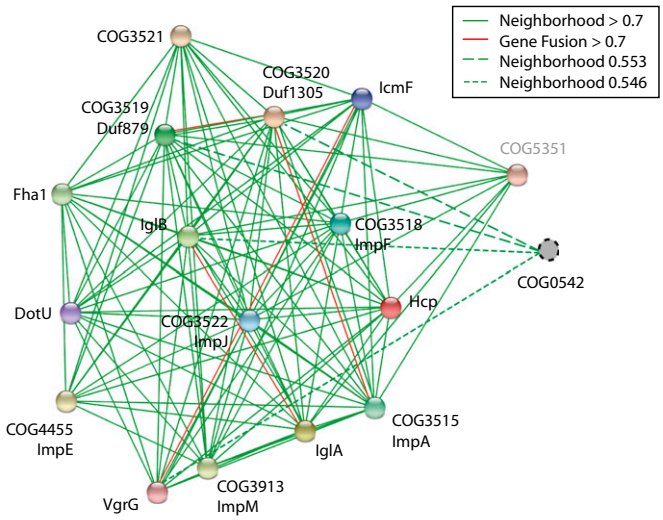

Fig. S6. T6SS functional network. A network of connections representing genomic neighborhood and gene fusion scores was derived by the STRING database. Nodes that represent conserved T6SS component COGs are labeled (black) and connected with solid lines for confident (>0.7) neighborhood (green) and gene fusion (red) scores, or with dashed lines for neighborhood scores (green) with medium confidence. One additional COG that was not defined as a conserved T6SS component yet retains a large number of confident connections is labeled in gray.

Other Supporting Information Files

Dataset S1 (XLSX)
Dataset S2 (XLSX)