Supplementary Methods

The *Photorhabdus luminescens* genome reveals a biotechnological weapon to fight microbes and insect pests.

Methods

Cloning, sequencing, assembly and annotation

Genome sequencing was performed using the whole genome shotgun strategy[1] as described by Frangeul[2]. Two libraries (1–2 kb and 2–3 kb inserts) were generated by random mechanical shearing of genomic DNA and cloning into pcDNA-2.1 (Invitrogen) and a medium size insert library (5-10 kb) was generated in the low copy number vector pSYX34[3]. The BAC library was constructed as previously described[4]. Briefly, *P. luminescens* genomic DNA, embedded in agarose plugs, was partially digested with *Hind*III and 50–90 kb fragments were cloned into pBeloBAC11[5]. Recombinant plasmids were used as templates for cycle sequencing reactions of 35 cycles (96°C for 30 s; 50°C for 15 s; 60°C for 4 min). Samples were precipitated and loaded onto a 96-lane capillary automatic 3700 DNA sequencer or a conventional 377 DNA sequencer (Applied Biosystems). In an initial step 63,475 sequences from the four libraries were assembled into 472 contigs using the Phred/Phrap/Consed software[6,7] (sequence coverage 7-fold). CAAT-Box[8] was used to predict links between contigs. Walks on individual clones and polymerase chain reaction products amplified from a TT01 chromosomal DNA template were used to fill gaps, to resolve assembly ambiguities and to re-sequence low quality regions. Primers were designed using the software Consed. The correctness of the assembly was confirmed by analysing the scaffold obtained by end-sequencing of 1200 clones from the BAC library (7-fold coverage).

Coding sequences (CDS) were defined by combining Genemark predictions[9] with visual inspection of the open reading frames (ORF) for the presence of start codons with an upstream ribosome binding site and BLASTP similarity searches using the NCBI Nrprot database[10]. The Genemark predictions were trained on a set of ORFs longer than 300 codons. Initially, all predicted CDSs longer than 80 codons were retained. Subsequently, all CDSs between 40 and 80 codons were searched using the same matrix, but only those with a high coding probability were retained. All predicted CDSs were examined visually. Function predictions were based on BLASTP similarity searches and on the analysis of motifs using the PFAM databases[11]. Toppred2 was used to identify transmembrane domains[12]; SignalP vs2.0 was used to predict signal peptide regions[13] and the Petrin algorithm was used to predict transcriptional terminators[14]. Lipoproteins were defined as proteins containing a lipoprotein modification/processing motif[15] and a signal sequence, identified by SignalP vs2.0. Secreted proteins were defined as proteins containing a signal peptide (predicted by SignalP vs2.0) but no other transmembrane domain. Functional annotation of each predicted protein was manually inspected. Repeats in protein sequences were identied using multiple sequence alignment with hierarchical clustering[16]. The genome sequence and the annotation are accessible via a Sybase relational database constructed according to the SubtiList model[17] at http://genolist.pasteur.fr/PhotoList.

Plasmid and cloning strategies.

BAC1A02 encompassing the JHE-like toxin locus *plu4093-plu4092* extends from position 4 729 455 to position 4 789 585 of the *P. luminescens* genome; BAC8C11 encompassing the JHE-like toxins locus *plu4437-plu4436* extends from position 5 179 747 to position 5 205 606.

To construct the pDIA700 plasmid, a DNA fragment containing both genes *plu4093* and *plu4092* was generated by PCR with primers JHE2 (5'-AACTGCAGCATTGAAGCAGAGCGTTGACAT-3') and JHE3 (5'-CGGGATCCGACGTCGGCAAGTGCATCAAAT-3'). The amplified DNA-fragment of 2060-bp was purified and cloned into the pBluescript SK vector (Stratagene) into the *EcoR*V site.

In order to inactivate *plu4092*, the pDIA700 was restricted by *EcoR*I leading to a deletion of the DNA region located between an *EcoR*I site located at the 270th codon of *plu4092* and the pBluescript *EcoR*I site and self-ligated to yield pDIA701 which contains a 3' truncated *plu4092* gene.

Insecticidal assays.

***(i) Plutella xylostella* leaf bioassay**. The recombinant *E. coli* XL1-Blue strain containing pDIA700 or pDIA701 and *E.coli* DH10B containing BAC1A02 or BAC8C11 were grown for 20 h at 28°C in 50 ml of LB broth supplemented with 100 µg/ml ampicillin or 12,5 µg/ml chloramphenicol, respectively. For each clone, six cabbage leaves (3 cm diameter) were incubated in non-diluted bacterial cultures for 1 hour with 0.05% Tween 20 (Sigma). Treated leaves were put onto 6-well plates containing an agar-bed containing 15 g/l of agar (Difco) and 30 mg/l of fungicide nipagine (Sigma). Five second-instar larvae were placed into each well. All the assays were done at 28°C with a photoperiod of 11:13 h (night:day). After 2 days, the treated leafs were replaced by untreated leaves. The larval mortality was recorded at day 2 and day 3.

Recombinant *E. coli* strains containing the pBluescript SK vector without insert and *P. luminescens* TT01 cultures were used as negative and positive controls, respectively. The percentage of larval mortality was corrected with the negative control by using the Abbott equation[18]. The toxicity of each clone was tested on 30 larvae, and each experiment was done in triplicate.

(ii) ***Mosquitocidal*** **activity**. Recombinant *E. coli* clones containing plasmids pDIA700, pDIA701 or *E.coli* DH10B containing BAC1A02 or BAC8C11 were grown at 30°C in 50 ml LB broth supplemented with 100 µg/ml ampicillin or 12,5 µg/ml chloramphenicol, respectively, until the optical density at 600 nm (OD_{600}) reached a value of 2, then harvested by centrifugation and resuspended at the same optical density. Ten *Culex pipiens* second-instar larvae were placed in Petri dishes (2.5 cm diameter) containing 5 ml bacterial suspensions at $OD_{600} = 2$ (or water as a control). Yeast cells were given as food source in all dishes (to avoid mortality in the control), and larval mortality was recorded at days 1 and 2. Two independent experiments were conducted at 24 ± 2 °C, each one in duplicate. The same procedure was used for toxicity assays involving *Aedes aegypti* and *Anopheles gambiae.*

References

1. Fleischmann, R.D., *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496-512 (1995).
2. Frangeul, L. *et al.* Cloning and assembly strategies in microbial genome projects. *Microbiology* **145**, 2625-2634 (1999).
3. Xu, S.Y. & Fomenkov, A. Construction of pSC101 derivatives with Camr and Tetr for selection or LacZ' for blue/white screening. *Biotechniques* **17**, 57 (1994).
4. Buchrieser, C. *et al.* The 102-kilobase pgm locus of *Yersinia pestis*: sequence analysis and comparison of selected regions among different *Yersinia pestis* and *Yersinia pseudotuberculosis* strains. *Infect. Immun.* **67**, 4851-4861 (1999).
5. Woo, S. S., Jiang, J., Gill, B. S., Paterson, A. H. & Wing R. A. Construction and characterization of a bacterial artificial chromosome library of Sorghum bicolor. *Nucleic Acids Res.* **22**, 4922-4931 (1994).
6. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*. **8**, 186-194 (1998).
7. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8**, 195-202 (1998).
8. Frangeul, L. *et al*. CAAT-Box, Contigs-Assembly and Annotation tool-box for genome sequencing projects. *Bioinformatics*. in press.
9. Isono, K., McIninch, J.D. & Borodovsky, M. Characteristic features of the nucleotide sequences of yeast mitochondrial ribosomal protein genes as analyzed by computer program GeneMark. *DNA Res.* **1**, 263-269 (1994).
10. Altschul, S.F. *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997).

11. Bateman, A. *et al*. The Pfam protein families database. *Nucleic Acids Res.* **30**, 276-280 (2002)

12. Claros, M.G. & von Heijne, G. TopPred II: an improved software for membrane protein structure predictions. *Comput. Appl. Biosci.* **10**, 685-686 (1994).

13. Nielsen, H., Brunak, S. & von Heijne, G. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.* **12**, 3-9 (1999).

14. d'Aubenton Carafa, Y., Brody, E. & Thermes, C. Prediction of rho-independent *Escherichia coli* transcription terminators. A statistical analysis of their RNA stem-loop structures. *J. Mol. Biol.* **216**, 835-858 (1990).

15. Hayashi, S. & Wu, H.C. Lipoproteins in bacteria. *J. Bioenerg. Biomembr.* **22**, 451-471 (1990).

16. Corpet, F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* **25**, 10881-10890 (1988).

17. Moszer, I., Jones, L.M., Moreira, S., Fabry, C. & Danchin A. SubtiList: the reference database for the *Bacillus subtilis* genome. *Nucleic Acids Res.* **30,** 62-65 (2002).

18. Abbott, W. S. A method for computing the effectiveness of an insecticide. *J. Econ. Entomol.* **18**, 265-267 (1925).