Supporting Information

SI Materials and Methods

RNA isolation for 454 pyrosequencing, Northern blot and RACE analysis.

RNA was extracted for 454 pyrosequencing as follows: *Xcv* strains 85-10 and 85* were grown in NYG medium to exponential growth phase (OD_{600} = 0.6). Then, 10 ml stop-solution (95% ethanol, 5% phenol) was added to 40 ml bacterial culture which was snap-frozen in liquid nitrogen, thawn on ice and centrifuged. Cells were resuspended in 6 ml buffer (0.02 M sodium acetate pH 5.5, 0.5% SDS, 1 mM EDTA). RNA was isolated by addition of 6 ml phenol, preheated to 60°C, followed by two chloroform extractions. The RNA was precipitated at -80°C overnight with 2.1 volumes of an ethanol/0.15 M sodium acetate solution. After centrifugation, the RNA was washed with 70% ethanol, dried, resuspended in water and treated with DNase I (Roche) followed by phenol-chloroform extraction.

For RACE and Northern blot analyses, RNA was isolated from NYG-grown *Xcv* strains at exponential and both exponential and stationary (OD_{600} = ~1.5) growth phase, respectively, and treated with DNase I (Roche) as described (1).

5' and 3' RACE analyses.

RACE analyses (see Table 1) were carried out as described (2) with the following modifications: Reverse transcription was performed with 2 µg RNA, the Thermoscript RT system (Invitrogen) and a gene-specific primer for 5' RACE and a primer complementary to the 3' RNA adapter for 3' RACE analysis. Oligonucleotides used for RACE analyses are listed in Table S1. RACE-PCR was performed with Hotstar *Taq*-Polymerase (Qiagen). Cycling conditions: 95°C/15 min; 35 cycles of 95°C/40ss, 58°C/40 s, 72°C/40 s; 72°C/7 min. PCR products were cloned into pCR2.1-TOPO and transformed into *E*. coli TOP10F' (Invitrogen). Bacterial colonies were screened by colony PCR with vector-specific primers (see Table S1). Plasmid DNA was sequenced with an ABI PRISM 377 "Genetic Analyzer" DNA sequencer (Applied Biosystems).

Construction of cDNA libraries for dRNA-seq and 454 pyrosequencing.

Equal amounts of RNA from *Xcv* strains 85-10 and 85* were mixed. Next, we constructed dRNA-seq libraries as described (3). Briefly, primary transcripts of total RNA were enriched by a selective degradation of RNAs containing a 5' mono-phosphate (5'P) by treatment with Terminator^{TM} 5'P-dependent exonuclease (Epicentre). Prior to cDNA library construction, equal amounts of *Xcv* RNA were incubated 60 min at 30°C with terminator exonuclease (for generation of cDNA-library 2) or in buffer (for generation of cDNA-library 1). We used 1 unit terminator exonuclease per µg total RNA. Following organic extraction (25:24:1 v/v phenol/chloroform/isoamyalcohol), RNA was precipitated overnight with 2.5 volumes of an ethanol/0.1 M sodium acetate (pH 6.5) solution, and treated with 1 unit TAP (tobacco acid pyrophosphatase) (Epicentre) for 1 hour at 37°C to generate 5' mono-phosphates for linker ligation, and again purified by organic extraction and precipitation as above.

cDNA libraries for 454 pyrosequencing were constructed by *vertis* Biotechnology AG, Germany (http://www.vertisbiotech.com/) as described for eukaryotic microRNA (4) but omitting RNA size-fractionation prior to cDNA synthesis. Briefly, equal amounts of RNA treated with terminator exonuclease and untreated RNA, respectively, were poly(A)-tailed using poly(A) polymerase, followed by ligation of an RNA adapter to the 5´ P-RNA fragments. First-strand cDNA synthesis was performed using an oligo(dT)-adapter primer and M-MLV RNase H^- reverse transcriptase. Incubation temperatures were 42°C for 20 min, ramp to 55°C, followed by 55°C for 5 min. The cDNAs were PCR-amplified to yield a concentration of 20-30 ng/µl using a high fidelity DNA polymerase. Libraries were generated for the 454 FLX and Titanium kits. Each library contains a specific barcode sequence, which is attached to the 5' end of the cDNAs during PCR amplification: For FLX libraries, CCGA and CGCA were used as barcode tags for library 1 and 2, respectively. For Titanium libraries, ACGTGC and AGCGTA were used as barcode tags for library 1 and 2, respectively. 454 pyrosequencing was performed on a Roche 454 sequencer using FLX and Titanium chemistry at the Max Planck Institute for Molecular Genetics (Berlin, Germany). For library 1, a total of 62,056 and 98,293 reads was

sequenced using the FLX and Titanium kits, respectively. For library 2, a total of 51,091 and 98,505 reads was sequenced using the FLX and Titanium kits, respectively.

Sequence mapping.

For mapping of 454 reads, 5'-end-linker sequences were clipped, and reads with a poly(A) content of > 70% were discarded to prevent mapping errors. The remaining reads, including poly(A) tails and the 3' adapter sequence, were aligned to the genome sequence of *Xcv* strain 85-10 using the segemehl program (parameter settings E 10 A 65 D 1 H 2) (5). Mapped reads were post-processed by clipping of poorly aligned 3' ends as follows: For each alignment generated by segemehl (scoring scheme: match = 2, substitution = -2 and insertion/deletion = -3) the alignment score from the start of the read to each downstream nucleotide was calculated and stored in an array. All elements stored at a position greater than the maximum score at index i_{max} presumably correspond to the poly(A) tail and the 3'-linker sequence. Hence, the mapped read was clipped at position i_{max}. Reads that mapped with an identity of ≥ 85% and a minimum length of 12 nt were analyzed further whereas reads mapping to rRNA or tRNA genes were excluded.

Prediction of regulatory motifs and small ORFs.

Promoter regions, 50 nt upstream of the annotated TSS, and 5' UTRs were scanned with MEME (6) for regulatory motifs. To identify short conserved protein coding genes in *Xcv*, a multiple sequence alignment of 19 bacterial genomes (see Table S8) was calculated with the Multiz package (7). The alignments were analyzed for potential coding segments using RNAcode (8) and a p-value cutoff of 0.05. High scoring segments were combined if they were ≤ 15 nt apart and in the same reading frame. Regions that overlapped with annotated genes were discarded. The remaining 265 regions were inspected for potential open reading frames starting with an ATG and ending with a canonical stop codon. If no complete ORF was detected, the RNAcode high scoring segment was extended by 51 nt up- and downstream followed by repeated analysis. The RNAcode prediction resulted in 24 potential short ORFs in *Xcv* (Table S7; annotation files are available at www.bioinf.uni-leipzig.de/publications/supplements/10-035).

Rfam scan.

The *Rfam* database version 10.0 was downloaded from ftp://ftp.sanger.ac.uk/pub/databases/Rfam/10.0/. To scan the *Xcv* genome for known noncoding RNAs the *Rfam* provided Perl script rfam_scan.pl with an e-value cutoff of 100 was used. Eight riboswitches (FMN, SAH, Glycine, SAM, Cobalamin, TPP, yybP-ykoY) and five RNAs (RNase P, SRP, tmRNA, 6S-RNA, RrT) were identified (see Table S6). Annotation files are available at http://www.bioinf.uni-leipzig.de/publications/supplements/10-035.

Homology analysis.

Homology searches were based on scans of the bacterial NCBI genome database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/; downloaded 08/02/2010). To identify homologs of *Xcv* sRNA genes, Gotohscan (9) was used. Results were aligned with RNAclust, which is based on the LocARNA algorithm (10), and visualized with the SoupViewer (www.bioinf.uni-leipzig.de/software.html). Alignments of the analyzed *Xcv* sRNAs are available at http://www.bioinf.uni-leipzig.de/publications/supplements/10-035.

Protein detection.

The analysis of type III secretion was performed with *Xcv* strains incubated in minimal medium A as described (11). Total cell extracts and culture supernatants were concentrated 10 and 100 times, respectively, and were analyzed by sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) and immunoblotting. For protein detection, specific polyclonal antibodies directed against HrpF (11), HrcJ (12) and GroEL (Stressgen) were used. A horseradish peroxidase-labeled anti-rabbit antibody (Amersham Pharmacia Biotech) was used as secondary antibody. The antibody reactions were visualized by enhanced chemiluminescence (Amersham Pharmacia Biotech).

For detection of the sX6-c-Myc protein, total cell extracts of NYG-grown bacteria (harvested at OD_{600} = 0.7) were concentrated 10-fold and analyzed by SDS-PAGE and immunoblotting using PVDF membranes. sX6-c-Myc was visualized with a monoclonal anti-c-Myc antibody (Roche) and a horseradish peroxidase-labeled anti-mouse secondary antibody (Amersham Pharmacia Biotech) by enhanced chemiluminescence (Amersham Pharmacia Biotech).

References

1. Hartmann, R.K., Bindereif, A., Schön, A. and Westhof, E. (2005) Handbook of RNA biochemistry. *Wiley-VCH, Weinheim, Germany*, **2**, 636-637.
2. Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E.G., Margalit, H. and Altuvia, S. (2001) Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol.*, **11**, 941-950.
3. Sharma, C.M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R. *et al.* (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature*, **464**, 250-255.
4. Berezikov, E., Thuemmler, F., van Laake, L.W., Kondova, I., Bontrop, R., Cuppen, E. and Plasterk, R.H. (2006) Diversity of microRNAs in human and chimpanzee brain. *Nat. Genet.*, **38**, 1375-1377.
5. Hoffmann, S., Otto, C., Kurtz, S., Sharma, C.M., Khaitovich, P., Vogel, J., Stadler, P.F. and Hackermüller, J. (2009) Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol*, **5**, 10.1371/journal.pcbi.1000502.
6. Bailey, T.L. and Elkan, C. (1995) The value of prior knowledge in discovering motifs with MEME. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **3**, 21-29.
7. Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708-715.
8. Washietl, S., Findeiß, S., Müller, S.A., Kalkhof, S., von Bergen, M., Hofacker, I.L., Stadler, P.F. and Goldman, N. (2011) RNAcode: Robust discrimination of coding and noncoding regions in comparative sequence data. *RNA*, **17**, 578-594.
9. Hertel, J., de Jong, D., Marz, M., Rose, D., Tafer, H., Tanzer, A., Schierwater, B. and Stadler, P.F. (2009) Non-coding RNA annotation of the genome of *Trichoplax adhaerens*. *Nucleic Acids Res.*, **37**, 1602-1615.
10. Will, S., Reiche, K., Hofacker, I.L., Stadler, P.F. and Backofen, R. (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, **3**, e65.
11. Büttner, D., Nennstiel, D., Klüsener, B. and Bonas, U. (2002) Functional analysis of HrpF, a putative type III translocon protein from *Xanthomonas campestris* pv. vesicatoria. *J. Bacteriol.*, **184**, 2389-2398.
12. Rossier, O., Van den Ackerveken, G. and Bonas, U. (2000) HrpB2 and HrpF from *Xanthomonas* are type III-secreted proteins and essential for pathogenicity and recognition by the host plant. *Mol. Microbiol.*, **38**, 828-838.