INFECTION AND IMMUNITY, Apr. 2005, p. 1927–1938 Vol. 73, No. 4
0019-9567/05/$08.00+0 doi:10.1128/IAI.73.4.1927–1938.2005

Cloning and Sequencing of a Genomic Island Found in the Brazilian Purpuric Fever Clone of *Haemophilus influenzae* Biogroup Aegyptius

Glen McGillivary,† Andrew P. Tomaras, Eric R. Rhodes, and Luis A. Actis*

Department of Microbiology, Miami University, Oxford, Ohio

Received 1 July 2004/Returned for modification 21 September 2004/Accepted 22 November 2004

A genomic island was identified in the *Haemophilus influenzae* biogroup aegyptius Brazilian purpuric fever (BPF) strain F3031. This island, which was also found in other BPF isolates, could not be detected in non-BPF biogroup aegyptius strains or in nontypeable or typeable *H. influenzae* strains, with the exception of a region present in the type b Eagan strain. This 34,378-bp island is inserted, in reference to *H. influenzae* Rd KW20, within a choline transport gene and contains a mosaic structure of Mu-like prophage genes, several hypothetical genes, and genes potentially encoding an *Erwinia carotovora* carotovoricin Er-like bacteriocin. The product of the tail fiber ORF in the bacteriocin-like region shows a hybrid structure where the C terminus is similar to an *H. influenzae* phage HP1 tail protein implicating this open reading frame in altering host specificity for a putative bacteriocin. Significant synteny is seen in the entire genomic island with genomic regions from *Salmonella enterica* subsp. *enterica* serovar Typhi CT18, *Photorhabdus luminescens* subsp. *laumondii* TT01, *Chromobacterium violaceum*, and to a lesser extent *Haemophilus ducreyi* 35000HP. In a previous work, we isolated several BPF-specific DNA fragments through a genome subtraction procedure, and we have found that a majority of these fragments map to this locus. In addition, several subtracted fragments generated from an independent laboratory by using different but related strains also map to this island. These findings underscore the importance of this BPF-specific chromosomal region in explaining some of the genomic differences between highly invasive BPF strains and non-BPF isolates of biogroup aegyptius.

Classically, *Haemophilus influenzae* biogroup aegyptius has been characterized as a pathogen causing only self-limiting conjunctivitis in young patients. These infections are generally confined to treatable, non-life-threatening cases, and the infection usually resolves after treatment with appropriate antibiotics (7). However, this definition was dramatically altered in the mid-1980s (5, 6) when an outbreak of a lethal disease known as Brazilian purpuric fever (BPF) emerged in several small towns in close proximity to the large industrial cities of Sao Paulo and Rio de Janeiro, Brazil (7). The symptoms associated with the severe disease caused by this strain differed markedly from known patterns of biogroup aegyptius infections and included vascular involvement shown as extensive hemorrhaging, purpura, vomiting, coma, and eventually death. Two years after this initial outbreak, it was postulated that all of the cases from the original outbreak were caused by a single clone called the BPF clone (7). Since the initial description of this outbreak of BPF, this disease has been found in other towns in Brazil (18, 47), in two towns in Australia (25, 51), and more recently in the United States (48), although it seems that the same clone is not responsible for all new cases (25, 47).

About 20 years after this initial outbreak, the molecular mechanisms responsible for the massive vascular damage seen in infected patients have not been elucidated. Progress in this field has been slow, mainly due to the lack of and difficulties in developing appropriate genetic systems that could allow the identification and characterization of potential bacterial virulence factors involved in the pathogenesis of this disease. With this in mind, the first approach used in our laboratory to identify and study genes unique to the BPF strain F3031 has been a genome subtraction hybridization procedure utilizing total DNA from the invasive F3031 strain and the noninvasive biogroup aegyptius strain F1947 (40). These are two related strains isolated in Brazil that belong to the same genus and species but differ in their rRNA gene restriction and sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) protein patterns, as well as their restriction pattern of the 24-MDa plasmid that is found in both of them (7). The genome subtraction hybridization approach yielded small DNA fragments that are specific to the BPF prototype strain F3031, several of which appear to have been acquired from unrelated microbial sources and encode unknown functions. Similar results were obtained by Li et al. (23) when they compared the BPF strain F3028 with the conjunctivitis strain F3043.

Complete and partial phages have been described for different strains of *H. influenzae* (2, 10, 19, 38, 52). It is known that two Mu-like phages are found in the sequenced strain of *H. influenzae* Rd, which are referred to as the prophage FluMu (29) and the cryptic prophage ϕ*flu* (19). The prophage FluMu contains most of the genes present in Mu and is found integrated within a gene encoding molybdenum transport functions. It also contains genes encoding products involved in transposition, has only one proposed tail fiber structural protein, and contains no invertible gene segments (29). In addition to FluMu, *H. influenzae* has two other temperate phages called HP1 (10) and HP2 (52), of which HP1 was the first phage identified for *H. influenzae* Rd (10, 17). The HP1 genome has

* Corresponding author. Mailing address: Department of Microbiology, Miami University, 40 Pearson Hall, Oxford, OH 45056. Phone: (513) 529-5424. Fax: (513) 529-2431. E-mail: actisla@muohio.edu.

† Present address: Columbus Children's Research Institute, Department of Pediatrics, Ohio State University College of Medicine and Public Health, Columbus, Ohio.

41 genes that encode typical structural proteins such as a tail sheath, tail fibers, a tail collar, and it also encodes proteins involved in integration and excision of the viral genome from that of the bacterial host (10). The recently described HP2 genome is related to HP1 with the important caveat that it infects nontypeable *Haemophilus* strains. Importantly, comparisons between *Haemophilus* phages suggest that extensive recombination events have occurred within this group of phages and also between the phages and resident host chromosomes (52). Some of these events within genes encoding tail fiber proteins have been proposed to explain differences in host specificities (39, 52).

Recently, Chang et al. (8) have described a genetic island that is specific to the *H. influenzae* type b serotype and contains a mosaic collection of genes, some of which appear to be of phage origin, whereas others have homologs to genes in other bacteria. Although the contribution of this genetic island to virulence is not known, the propensity of finding this 16-kb region in type b strains, coupled with the finding that numerous characteristics for pathogenicity islands are contained in this region, suggests that it may play a role in virulence (8). In addition to this island, other type b-specific genetic islands, in relationship to *H. influenzae* Rd KW20, have been described (3). Some of these islands contain genes whose products play a known role in virulence, capsular and fimbrial gene clusters, whereas the role in virulence of other gene products contained in these islands remains cryptic, such as tail fiber proteins and outer membrane proteins of undefined function (3). We describe here a genomic region in the BPF clone F3031 of biogroup aegyptius that is not found in its noninvasive counterpart F1947. This 34-kb region, which was found in other BPF isolates but not in non-BPF biogroup aegyptius strains and nontypeable or typeable *H. influenzae* strains, with the exception of a region present in the type b Eagan strain, is similar to the genetic island 1 of the type b strains of *H. influenzae* in that it contains numerous genes encoding either putative phage or bacteriocin functions and hypothetical proteins with unknown functions.

MATERIALS AND METHODS

Bacterial strains and growth conditions. The *Haemophilus* strains used in the present study are listed in Table 1. These strains were routinely grown on chocolate agar (PML Microbiologicals, Warwick, R.I.) or brain heart infusion agar supplemented with 2 μg of nicotinamide-adenine dinucleotide and 2 μg of hemin/ml (sBHI) at 37°C with 5% CO_2. *Escherichia coli* DH5α (Gibco-BRL, Gaithersburg, Md.) and EPI300 (Epicentre, Madison, Wis.), which were used as hosts in routine molecular biology experiments, were cultured under standard conditions in Luria-Bertani medium (37) containing the appropriate antibiotics.

TABLE 1. Bacterial strains and plasmids used in this study

H. influenzae strain	Relevant characteristics	Source or reference[b]
Biogroup aegyptius strains		
F3031	Brazilian invasive BPF isolate, type strain	7
F3028	Brazilian invasive BPF isolate	7
Valparaiso	Brazilian invasive BPF isolate	CDC via A. Lesse
Connecticut	USA invasive BPF isolate	48
F4380	Australian invasive BPF	51
F1947	Brazilian noninvasive non-BPF isolate	7
F3043	Brazilian noninvasive non-BPF isolate	A. Lesse
Other *H. influenzae* strains		
AMC 36-A-3	Type a	ATCC
Eagan	Type b	S. Goodgal
AMC 36-A-5	Type c	ATCC
AMC 36-A-6	Type d	ATCC
AMC 36-A-7	Type e	ATCC
AMC 36-A-8	Type f	ATCC
86-028NP	Non-typeable, otitis isolate	L. Bakaletz

[a] With the exception of F4380, Connecticut, and Valparaiso, all of the *H. aegyptius* strains contained the pF3031 plasmid.
[b] CDC, Centers for Disease Control; ATCC, American Type Culture Collection

General DNA procedures. Total DNA was isolated by ultracentrifugation in CsCl density gradients (26). Plasmid DNA was isolated by ultracentrifugation in CsCl-ethidium bromide density gradients (37) or with a commercial kit (Qiagen, Valencia, Calif.). DNA was digested with restriction enzymes as indicated by the supplier (New England Biolabs, Beverly, Mass.) and size fractionated by agarose gel electrophoresis (37). Lambda DNA digested with HindIII (New England Biolabs) was used as a molecular weight marker. Both strands of cloned DNA fragments were sequenced with BigDye (Applied Biosystems, Foster City, Calif.) or DYEnamic ET (Amersham Pharmacia Biotech, Piscataway, N.J.) chemistries on ABI Prism 310 or 3100 instruments using M13 forward and reverse primers (54), pCC1FOS-based primers (forward 5′-GGATGTGCTGCAAGGCGATTAAGTTGG-3′ and reverse 5′-CTCGTATGTTGTGTGGAATTGTGAGC-3′), or custom-designed primers. Sequences were examined and assembled with SeqMan II (DNASTAR, Madison, Wis.). Nucleotide and amino acid sequences were analyzed with GeneQuest and MapDraw (DNASTAR), BLAST (http://www.ncbi.nlm.nih.gov), Artemis (http://www.sanger.ac.uk/Software/Artemis/), and the software available through the ExPASy Molecular Biology Server (http://www.expasy.ch). Southern blot analyses were conducted by using standard methods under high-stringency conditions (16, 37). Probes PCR#1 and PCR#2 (Fig. 1), which

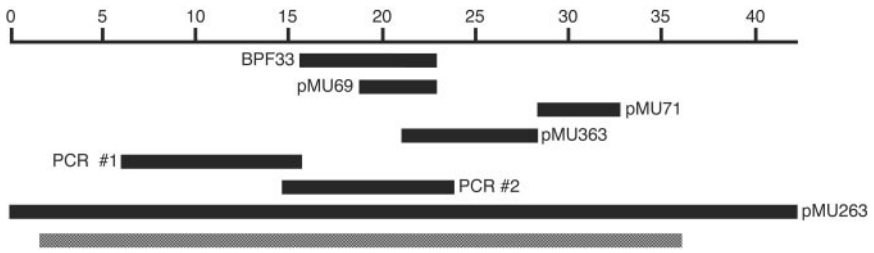

FIG. 1. Diagram of DNA fragments used in cloning and analysis of the genomic island. Fragments of F3031 DNA that were cloned in vectors are indicated by pMU numbers. PCR#1 (9,519 bp) and PCR#2 (8,885 bp) were not cloned but were used as probes in Southern blot hybridization analyses. BPF33 is a BPF locus that has been described (40). The length of the thick black lines show the sizes of the F3031 fragments used in experiments as indicated by the size scale, in kilobases, located on top. The stippled line at the bottom shows the location of the genomic island.

TABLE 2. Oligonucleotide primers used in RT-PCR experiments

Oligonucleotide	Sequence
ORF 21-5′	5′-CGGGTTGGGATTATAACG-3′
ORF 21-3′	5′-CTTCATCAATCACGCTTGC-3′
ORF 24-5′	5′-TGGCACTCGTAAAGATGC-3′
ORF 24-3′	5′-ATGCGGTGACATAAGACAC-3′
ORF 25-3′	5′-GTAACCAGCAGCTTTACC-3′
ORF 26-5′	5′-GTTATATGCCGTCGTTGG-3′
ORF 26-3′	5′-CTGCGCTTTTTCAGCAAC-3′
ORF 29-5′	5′-CATCATGGCACAGAAACC-3′
ORF 29-3′	5′-TCATTACTGACGTGTTGG-3′
ORF 30-5′	5′-CTCATTACTGGGCAAAGC-3′
ORF 30-3′	5′-TTACCGCCAACGGTTTGC-3′
ORF 36-5′	5′-CCCATAATTTCACTGCTACC-3′
ORF 36-3′	5′-CGCCTTGCTCAGTATGTTTG-3′
ORF 37-5′	5′-CACATACTGACTTAACGACAC-3′
ORF 37-3′	5′-CCTTGACCTCTGTTGAATAG-3′
ORF 38-5′	5′-GTGTCTGAAGATGTGAAAC-3′
ORF 38-3′	5′-CTCTCGCTTTCGACTTCTAC-3′
ORF 39-5′	5′-CGTTACCCCGATATTATCGC-3′
ORF 39-3′	5′-TGTAGAGTACTGGTGCAGAC-3′
ORF 40-3′	5′-TGCCGTTTCTATCAACAGTC-3′
ORF 40-5′H	5′-TCCTCGCGGTTATCTAAGAG-3′
ORF 40-5′S	5′-CGTATAGAGGAAAACGATCC-3′

were used to screen genomic libraries or to detect the presence of particular F3031 genomic island-like regions in different subclones, were prepared by PCR amplification with appropriate sets of primers and *Pfu* DNA polymerase (Stratagene, La Jolla, Calif.). PCR#1 was amplified by using the primers F18 (5′-TGCATATTGCAAGGAGCC-3′) and 33.5′0.1 (5′-CGCCAATTTCTTCCTCCG-3′) and PCR#2 used primers p26.F (5′-GGAAAGATTACGATATGG-3′) and p27.R (5′-TTGAACAGTTAGAAATG-3′). The pCC1FOS-based construct pMU263 (Fig. 1) was used as the template for these amplification reactions. Cycle conditions used an initial denaturation at 94°C for 5 min, followed by 30 cycles of denaturation at 94°C for 30 s, annealing at 55°C for 2 min, and extension at 72°C for 22 min. A final extension at 72°C for 7 min was also used. All amplicons used were purified by using the GeneClean II kit (Q-Biogene, Carlsbad, Calif.) and labeled with [α-^{32}P]dCTP (Perkin-Elmer, Boston, Mass.) (11). The radioactive bands were detected either with X-ray film or a Storm 860 scanner (Molecular Dynamics, Piscataway, N.J.).

Cloning of the BPF genomic island. The PCR products from F3031 total DNA representing the subtracted clones E3 and F20 (40) were used to screen a genomic library, which was made by cloning 4- to 6-kb Sau3AI partially digested F3031 DNA fragments into the BamHI site of calf intestinal alkaline phosphatase-treated pUC18 (Pharmacia, Piscataway, N.J.). This approach yielded the plasmids pMU69 and pMU71 (Fig. 1), but further attempts to isolate other overlapping clones from this high-copy-number library failed. In order to clone larger DNA fragments at lower copy numbers, a library was constructed with the CopyControl fosmid library construction kit (Epicentre). Briefly, after randomly shearing F3031 DNA by passage through a small-bore pipette tip 100 times, the ends of the DNA fragments were repaired with T4 DNA polymerase and T4 polynucleotide kinase. DNA fragments of between 40 and 50 kb were isolated after size fractionation in a 1% low-melting-point agarose gel, purified by incubation with GELase, and concentrated by ethanol precipitation. The DNA fragments were subsequently ligated to pCC1FOS, packaged in vitro with MaxPlax lambda packaging extracts, and transduced into *E. coli* EPI300 cells. Transductants were selected on Luria-Bertani agar plates containing 12.5 μg of chloramphenicol/ml. Screening of the fosmid library with a PCR product representing the genomic copy of the fragment cloned in pMU69 resulted in the isolation of the fosmid clone pMU263 (Fig. 1), which contains an insert of ~45-kb that encompasses the genomic island found in the F3031 BPF strain. The construct pMU363 was derived from pMU263 by digesting this plasmid with BamHI, gel isolating an ~7-kb fragment, and subsequently ligating the purified fragment into BamHI-digested pCC1FOS.

Transcriptional analysis of gene expression. Expression of some genes located within the F3031 genomic island was evaluated by reverse transcription-PCR (RT-PCR) analysis with total RNA isolated as described previously (53) from bacteria grown in 35 ml of sBHI to an optical density at 600 nm of 0.7. The RNA samples were treated with RNase-free DNase I (Roche, Indianapolis, Ind.), and the presence of selected transcripts was detected with a RT-PCR commercial kit (Qiagen) under the conditions suggested by the supplier. The primers used in the amplification reactions are listed in Table 2. The cDNA products were analyzed by agarose gel electrophoresis. PCR of total RNA with *Pfu* DNA polymerase without reverse transcription was used to test for DNA contamination of RNA samples.

Detection of the F3031 genomic island in other *Haemophilus* strains. The presence of the three regions of the genomic island in other *Haemophilus* strains was tested either by Southern blotting as described before (40) or colony PCR. For the latter method, colonies were picked and suspended in 50 μl of Tris-EDTA buffer and then incubated for 10 min at 98°C. The supernatant was collected after centrifugation for 10 min at full speed, and 1 μl was used in PCR amplification reactions with the *Taq* PCR Master Mix kit from Qiagen. Agarose gel electrophoresis was used to detect the production of the appropriate amplicons. The primers 2338 (5′-TAACGGAAAGTAAAAAGC-3′) and 2481 (5′-CCAGTGATGCCTCCCAC-3′) were used to amplify a 1.1-kb fragment that encompasses open reading frames (ORFs) 11 to 13 of region I. The primers 2334 (5′-ACTCAACGGAAACCTTAC-3′) and 2141-B (5′-CTGAACCAGTCTCATTAA-3′) were used to amplify a 980-bp fragment that encompasses ORFs 15 to 18 of region II. A 1-kb fragment that maps within ORF 32 of region III was amplified with the primers 2148 (5′-GGGTGCCAAAATGGCG-3′) and 2334 (5′-GGTCGCCAAATTTTGCTCGG-3′). These primers were used either in colony PCR assays or to prepare the cognate probes with pMU263 as a template. The primers D1F (5′-AGAGTTTGATCCTGGCTCAG-3′) and 1540R (5′-AAGGAGGTGATCCAGCC-3′), which amplify 16S rRNA gene (55), were used as PCR positive controls.

Nucleotide sequence accession number. The nucleotide sequence of the biogroup aegyptius F3031 genomic island described in the present study was deposited in GenBank under accession number AY647244.

RESULTS AND DISCUSSION

Discovery of the F3031 genomic island. In an early attempt at defining genetic loci specific to the F3031 BPF prototype strain of biogroup aegyptius, we isolated 13 DNA fragments that were found in this strain but not in the non-BPF biogroup aegyptius isolate F1947 (40). At further glance it was noticed that several of the subtracted clones, such as E3, F7, and F20, contained genes similar to those encoding proteins responsible for the production of the phage-like bacteriocin carotovoricin Er produced by the plant pathogen *Erwinia carotovora* subsp. *carotovora* (also called *Pectobacterium* spp.) (33). This evidence suggested that these F3031 DNA subtracted fragments were contiguous and could be used as probes to isolate genomic clones harboring all of them in a single insert. Screening of a genomic library made in pUC18 with the subtracted fragments E3 and F20 resulted in the isolation of the genomic clones pMU69 and pMU71 (Fig. 1). PCR analysis with F3031 total DNA with primers annealing to the ends of the cloned fragments in pMU69 and pMU71 confirmed the orientation shown in this figure and indicated that an ~5-kb gap existed between the fragments cloned in these two constructs. In addition, the DNA sequence of the insert in pMU69 could be assembled with BPF33 (Fig. 1), a BPF-specific locus we described before (40).

Attempts to isolate high-copy-number genomic clones harboring the ~5-kb gap between the inserts of pMU69 and pMU71 (Fig. 1) or to clone it by PCR amplification in vectors such as pCR-BluntII-TOPO (Invitrogen) failed to produce the appropriate derivatives. However, the screening of a fosmid library made by cloning large DNA fragments in pCC1FOS yielded the derivative pMU263, which includes the fragments cloned in both pUC18-based genomic clones and the previously described BPF33 locus (Fig. 1). Sequencing and nucleotide analysis of the 42-kb insert showed that carotovoricin Er-related genes are part of a 34,252-bp fragment that disrupts

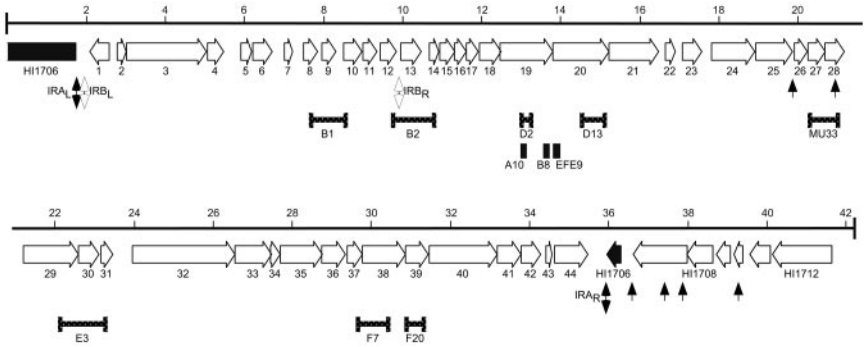

FIG. 2. Graphical representation of the F3031 genomic island and surrounding regions. Horizontal arrows represent ORFs with the direction of transcription shown by the arrows orientation. The black rectangle and arrow represent the 5′ and 3′ fragments of HI1706, respectively, that were produced by the insertion of a genomic island. ORFs contained within the island are numbered, and those outside the island are indicated with HI locus numbers according to their homologs in *H. influenzae* Rd KW20. The stippled bars show the subtracted fragments that were previously isolated (40), and the black boxes show subtracted fragments isolated by Li et al. (23). Vertical arrows designate uptake signal sequence sites. The solid and open double-headed arrows highlight the inverted repeats IRA and IRB, respectively. Half-site copies of the inverted repeats are designated as left (L) or right (R). The numbers in the size scales are kilobases.

a homolog of the *H. influenzae* Rd HI1706 gene (Fig. 2), which is predicted to encode a high-affinity choline transport protein (13). The last 1,720 bp and the first 240 bp of the F3031 HI1706 homolog flank the 5′ and 3′ ends of the disrupting fragment, respectively. Based on similarity, the entire HI1706 homolog is not represented in pMU263. The sequence 5′-TTTAAAATG AGAA-3′ (IRA_L and IRA_R) is found repeated twice in inverted orientation, with one copy located 17 nucleotides downstream of the interruption site of the 5′ fragment of the HI1706 homolog and the other copy found 321 nucleotides upstream of the interruption site of the 3′ fragment of this homolog (Fig. 2).

Taken together, the data suggest that this region contains features that resemble those of genomic islands described in other bacterial genomes, which have inserted within a gene encoding putative transport functions (29). Figure 2 shows that this F3031 genomic region contains 44 predicted ORFs, with all but the first transcribed in the same direction and contains only two copies of the 5′-AAGTGCGGT-3′ DNA uptake signal sequence (15). The overall G+C content of this region is 39.9%, a value that is close to the 38% reported for the genome of the *H. influenzae* Rd strain (13).

Based on the analysis of the putative products encoded by these genes (Table 3), this seems to be a genomic island that can be functionally divided into three regions in which the first initiates immediately after the 3′ end of the 5′ fragment of HI1706 and continues through ORF 15. The second region encompasses ORF 16 through ORF 28, and the third region extends from ORF 29 to the 5′ end of the 3′ fragment of the interrupted HI1706 homolog.

Analysis of region I of the genomic island. This region is 9,604-bp in length and encompasses 15 ORFs, with the first 12 of them flanked by the sequence 5′-GTTTAAATTGATTTT AA-3′ found in inverted orientation (Fig. 2, IRB_L and IRB_R). Although the overall G+C content of this region is 39.1%, a value close to that determined for the *H. influenzae* Rd chromosome (13), a wide variation was observed when individual ORFs were analyzed (Table 3). This observation suggests that some components of this region were acquired from unrelated sources. The first ORF, the only one transcribed in a direction opposite to that of the remaining predicted genes, codes for a product similar in sequence and length to the *Yersinia pestis* YP03514 conserved hypothetical protein, which is similar to the bacteriophage D108 repressor protein CI (36). This hypothetical protein also showed similarity to the N terminus of transposases, particularly that found in the phage Mu, which has been described previously (28). The product of the second gene is similar to the Mu *ner* gene that codes for a DNA-binding protein that plays a role in regulating bacteriophage lysogeny (46). In addition, in the entomopathogenic bacterium *Photorhabdus luminescens* K122, *ner* has recently been described to affect expression of several virulence genes which are important in a phenotypic switch allowing the bacterium to survive and grow in a nematode and an insect (34). Interestingly, the *ner* homolog in strain W14 (accession no. AY144119) is part of a candidate pathogenicity island containing several toxin-encoding genes (49). This F3031 *ner*-like gene is followed by two ORFs whose translation products are related to the A and B subunits, respectively, of the transposase found in the Mu-like *Haemophilus* phage FluMu (29). Although the F3031 MuA homolog is similar in size to that found in the Rd genome, the MuB-like protein encoded by ORF 4 is about half of that encoded by the HD0095 and HI1481 genes found in *H. ducreyi* 35000HP (genome accession reference NC_002940) and *H. influenzae* Rd (13), respectively. Although the products of ORFs 5 and 7 are related to the hypothetical *H. ducreyi* HD0096 and *H. somnus* 2336 (accession reference ZP_00132800) proteins, respectively, the product of ORF 6 is similar to the *H. influenzae* Rd HI1483 protein, a putative Mu-like prophage FluMu host-nuclease inhibitor protein Gam (13).

The product of ORF 8 is related to a hypothetical protein of the *Y. enterocolitica* temperate bacteriophage PY54 (20), which is significantly related to the putative products of genes found in the lambdoid phage 434 (1) and the enterobacteria phage HK022 (21). The E16 Mu-like protein encoded by the HD0110 gene of *H. ducreyi* showed the highest similarity to the predicted product of ORF 10, which is also related to the FluMu protein gp16 (HI1488) of *H. influenzae* Rd (13). Between ORF 8 and ORF 10 (Fig. 2), there is a 118-amino-acid coding sequence that is related to the *H. ducreyi* proteins HD0495 and

TABLE 3. Description of ORFs contained in the BPF genomic island

ORF	Size (bp)		% G + C	Homolog(s)[c]	Score[d]	E-value[e]
	A[a]	H[b]				
1	501	399	43.1	Phage D108 repressor protein CI	59	3e-08
2	228	219	33.8	Phage D108 Ner DNA-binding protein	80	7e-15
3	2,010	2,061	42.0	HI1478, phage MuA transposase	379	1e-104
4	417	915	42.2	HD0095, phage MuB transposase	156	1e-37
5	204	231	39.7	HD0096, hypothetical protein	47	7e-05
6	507	507	47.1	HI1483, Mu host nuclease inhibitor Gam	186	3e-47
7	159	165	41.5	Haso109501, hypothetical protein	53	7e-07
8	375	465	37.6	Phage PY54 ORF 50	70	4e-12
9	354	276	40.1	HD0495, hypothetical protein	44	0.0005
10	489	462	35.7	HD0110, Mu-like E16 protein	120	1e-26
11	375	351	36.4	CV2165, phage transcriptional regulator	119	1e-26
12	420	NA[f]	32.6	No significant homology	NA	NA
13	537	519	44.7	HD0112, *N*-acetylmuramoyl-L-alanine amidase	305	3e-82
14	231	228	39.8	HD0114, conserved hypothetical protein	113	7e-25
15	384	252	38.0	HI1496, hypothetical protein	74	3e-13
16	309	330	37.9	Plu3445, unnamed protein	87	5e-17
17	303	309	41.2	Plu3444, unnamed protein	140	6e-33
18	549	543	40.3	Plu3443, unnamed protein	165	4e-40
19	1,323	1,374	44.6	DRA0094, conserved hypothetical	207	6e-52
20	1,416	1,464	41.2	CV2139, conserved hypothetical	432	1e-120
21	1,278	1,248	38.0	HD1557, bacteriophage Mu GP30-like	211	4e-53
22	189	NA	33.9	No significant homology	NA	NA
23	501	489	41.1	HD1558, bacteriophage Mu G-like	127	8e-29
24	1,104	1,092	43.0	Plu3438, hypothetical protein	289	7e-77
25	927	924	38.5	CV2134, hypothetical protein	209	9e-53
26	318	1,014	38.1	Erp22, *B. burgdorferi* OM protein	55	2e-07
27	435	447	41.9	Plu3435, hypothetical protein	99	2e-20
28	498	462	42.0	CV2130, hypothetical protein	97	1e-19
29	1,386	1,425	41.2	ORF 1, *E. carotovora* tail sheath protein	374	1e-102
30	519	522	41.7	ORF 2, *E. carotovora* tail core protein	147	2e-34
31	285	342	39.3	PA0624, hypothetical protein	43	0.001
32	2,592	2,550	41.3	Plu3429, hypothetical protein	271	7e-71
33	933	882	39.8	STY1630, hypothetical protein	113	5e-24
34	231	213	40.1	STY1631, bacteriophage tail fiber protein	80	1e-14
35	1,065	1,170	39.6	CV2118, bacteriophage regulatory protein	279	7e-74
36	609	402	42.0	Plu3426, phage baseplate component	121	1e-26
37	366	357	39.7	STM4203, phage baseplate protein	74	4e-13
38	1,107	1,101	42.8	STY1635, bacteriophage baseplate protein	256	9e-67
39	570	576	41.8	STY1636, bacteriophage tail fiber protein	145	5e-34
40	1,710	2,775	39.2	HP1p38, phage tail fiber protein	608	1e-172
41	603	630	36.2	HP2p34, phage tail collar	361	5e-99
42	501	456	37.9	Hflu2020135, enoyl-coenzyme A hydratase/carnithine racemase	251	6e-66
43	120	117	34.2	HI1522.1, Mu-like prophage protein Com	74	6e-13
44	846	888	38.5	HI1523, hypothetical protein	412	1e-114

[a] Actual size of ORF in the F3031 genomic island.
[b] Size of the homolog.
[c] Letters followed by locus numbers are as dictated by finished sequences. HI represents locus numbers from *H. influenzae* Rd KW20, Haso from *H. somnus* 2336, Hflu from *H. influenzae* R2866, HD from *H. ducreyi* 35000HP, CV from *C. violaceum* ATCC 12472, Plu from *P. luminescens* subsp. *laumondii* TT01, DR from *D. radiodurans* R1, PAO from *P. aeruginosa* PAO1, STM from *S. enterica* serovar Typhimurium LT2, and STY *S. enterica* subsp. *enterica* serovar Typhi CT18.
[d] A score of <40 was not considered significant.
[e] Blastx analysis using the NCBI website was used to get the expectation value (E-value).
[f] NA, not applicable.

HD0960. ORF 11 codes for a product that is related to several putative bacteriophage transcriptional regulators, with the highest match to the β-proteobacterium *Chromobacterium violaceum* CV2165 protein (4). Significant homology was also detected with the DNA-binding protein RdgB of the plant pathogen *Erwinia carotovora* subsp. *carotovora* Er, which together with RdgA regulates the production of pectin lyase in response to DNA-damaging agents (24). This regulator is also known to affect expression of the phage tail-like bacteriocin carotovoricin Er in a temperature- and *recA*-dependent manner (32). The predicted product of ORF 12 did not show any significant match in the GenBank database. The ORF 13 codes for a predicted protein that is highly related to the products of the HD0112 and HI1494 genes found in the genomes of *H. ducreyi* and *H. influenzae*, respectively. These putative *Haemophilus* proteins show strong similarity to the phage T3 and T7 *N*-acetylmuramoyl-L-alanine amidases that are involved in hydrolysis of *N*-acetylmuramoyl residues and certain L-amino acid residues in the cell wall of bacteria (29). ORFs 14 and 15 code for predicted proteins with the highest similarity to the *H. ducreyi* HD0114 conserved hypothetical protein (genome accession reference NC_002940) and the *H. influenzae* HI1496 hypothetical protein (13), respectively.

Analysis of region II. After the region containing the mosaic FluMu-like genes comes a series of 13 ORFs where the similarity shifts from predominantly *Haemophilus*-like sequences to genes whose predicted encoded products are similar to proteins found in *Salmonella enterica* subsp. *enterica* serovar Typhi CT18 (35), the soil- and waterborne bacterium *C. violaceum* ATCC 12472 (4), *P. luminescens* subsp. *laumondii* TT01 (*P. luminescens* TT01) (9), and *E. carotovora* subsp. *carotovora* (accession reference AB017338) (Fig. 2 and 3). Starting in this region, the genes found in F3031 and the abovementioned strains are highly similar both in sequence and organization. The ORF 16 product is similar to a membrane protein in serovar Typhi CT18 (STY1612), but it also has significant similarity to the product encoded by a *P. luminescens* hypothetical gene (Plu3445), and both hits occur in genomic regions containing multiple hypothetical and Mu-like genes (Fig. 3). Proteins encoded by ORFs 17 and 18 have similarities to the products of two genes in the same region of the serovar Typhi CT18 (STY1613 and STY1614) and *Photorhabdus* (Plu3444 and Plu3443) genomes. Although the function of these ORFs is not known, Plu3443 does have similarity to gp27 of Mu, the small subunit of a putative terminase (29). ORF 19 has the highest G+C content in this region and the highest similarity to the product of the *Deinococcus radiodurans* R1 ORF DRA0094 (50), which has recently been described as a component of the Mu-like RadMu prophage most likely encoding the large subunit of the terminase (29).

ORFs 19 and 20 are unique in the genomic island in that not only do they contain sequences identical to our published subtracted fragments, D2 and D13 (40), but they also encompass similar sequences from subtracted fragments A10, B8, and EFE9 (Fig. 2), which were generated from an independent laboratory (23) utilizing a different BPF isolate, F3028, and a separate biogroup aegyptius conjunctivitis isolate, F3043. The product of ORF 20 is similar to conserved hypothetical proteins found in *C. violaceum* (CV2139), *S. enterica* serovar Typhi CT18 (STY1616), and *P. luminescens* TT01 (Plu3441). The remaining ORFs (21–28) belong to a F3031 genomic region we have called the BPF33 locus (40). Even though the annotations for these ORFs have been published, we present updated information here because of the relevance of newly published sequences. ORF 21 is most similar to the product encoded by the *H. ducreyi* Mu-like gp30 ORF HD1557, and ORF 22 has no homolog in GenBank. The products of ORFs 23 to 25 now show matches to the products of loci in *H. ducreyi* 35000HP (HD1558, HD1559, and HD1561), *C. violaceum* (CV 2137, CV2135, and CV2134), *P. luminescens* TT01 (Plu3439, Plu3438, and Plu3437), and *S. enterica* serovar Typhi CT18 (STY1618, STY1619, and STY1620). The products of the STY1618 and Plu3439 ORFs have similarity to the G protein of Mu (29). ORF 26 shows the most similarity at the protein level to outer membrane lipoproteins Erp22 and ErpD from *Borrelia burgdorferi*, which are identical outer membrane proteins that are expressed in a temperature-dependent manner from one of the many resident plasmids named cp32s (27, 41). These proteins are highly expressed within mammalian hosts and other Erp family members are known to provide resistance to killing by complement (27, 41). ORF 26 is much shorter in length than these predicted proteins and, as such, is likely not a homolog but shares a similar domain, one close to the center of the Erp22 and D proteins (42). Interestingly, the family of Erps has recently been shown to be horizontally transferred, and the cp32 family of plasmids are now considered prophages (42). ORFs 27 and 28 are similar to the encoded products from the *P. luminescens* TT01 ORF Plu3435 and the *C. violaceum* CV2130 locus, respectively, both of which have no described function.

Analysis of region III in the genomic island. This region consists of ORFs 29 to 44, and almost all of them encode functions putatively related to the production of a bacteriocin (Table 3). Quite surprisingly, proteins with similarity to all of the region III products are highly conserved in serovar Typhi CT18 and in *P. luminescens* with the exception of ORF 34. ORFs 31 and 34 are not annotated in the *E. carotovora* Er sequence either. In *C. violaceum*, the order of genes is conserved from ORF 31 to ORF 36 and then the sequence diverges (Fig. 3).

The products of ORFs 29 and 30 have significant similarity to the tail sheath and core structural components, respectively, of a phage tail-like bacteriocin called carotovoricin Er from *E. carotovora* subsp. *carotovora* Er, which kills other *Erwinia* strains through an undefined mechanism (31, 33). The carotovoricin sheath is a contractile structure that surrounds the stable tail core and, interestingly, these two ORFs are represented partially in the subtracted fragment E3 (Fig. 2) (40). The product of ORF 31 is the only one in the genomic island that has similarity to the product of a *P. aeruginosa* gene, PAO624 (44), although ORF 31 is 57-bp shorter than PAO624 and the similarity is almost below the cutoff for significance, as is ORF 31's similarity to the ATP synthase alpha subunit. ORF 32 is the largest in the genomic island at 2,592 bp, and the putative protein product is similar to the largest component in the *E. carotovora* bacteriocin, ORF 5 (accession reference AB017338), and the *P. luminescens* ORF Plu3429 (Fig. 3). Despite having such a large coding region and two predicted transmembrane domains, a function could not be attributed to this ORF even with extensive in silico analyses. The product of ORF 33 is similar to protein products that are described as hypothetical, such as serovar Typhi CT18 STY1630. ORF 34 is the first of three ORFs in region III that have similarity at the protein level to bacteriophage tail fiber products. The translated products of both ORF 34 and ORF 39 have the most similarity to serovar Typhi CT18 encoded tail fibers, whereas the putative protein encoded by ORF 35 is related to predicted bacteriophage regulatory proteins, with the highest similarity to the product of CV2118 of *C. violaceum* ATCC 12472. Interestingly, the second highest match is the product of the BcepMu47 found in the *Burkholderia cenocepacia* phage BcepMu, which appears to contain genes encoding potential virulence factors (GenBank accession number YP_024720).

The first analysis showed that the protein encoded by ORF 40 has the most similarity to the resident *H. influenzae* Rd lysogenic phage HP1 tail fiber ORF 31 (10). However, upon closer examination it was noticed that the putative tail fiber encoded by ORF 40 was hybrid in nature. The N terminus has highest similarity to the protein product of the serovar Typhi CT18 gene STY1637, and the C terminus shows similarity to the encoded product from ORF 31 of the bacteriophage HP1 (Fig. 4A). Even though the average G+C content of this coding region is 39.2%, its 5′ and 3′ regions have values of 41 and

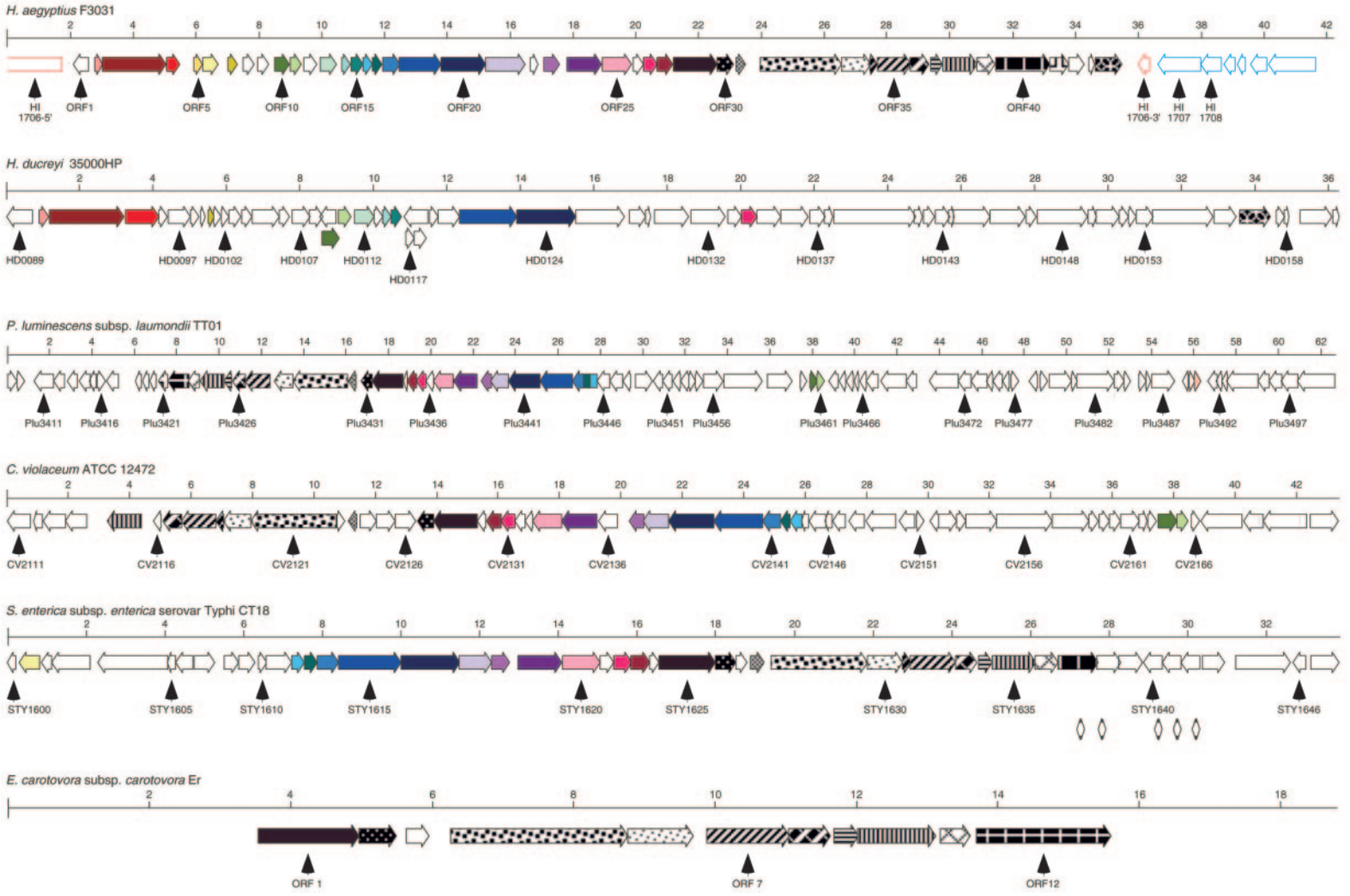

FIG. 3. Graphical representation of loci with synteny to the F3031 genomic island. Loci from several organisms are depicted with individual arrows representing ORFs with direction of transcription shown. ORFs are drawn to scale within organism sections but are not drawn to scale between organisms. ORFs contained within the genomic island in F3031 are numbered, and those outside the island are shown as an open red arrow and rectangle (split HI1706 ORF) or open light blue arrows. ORFs in other bacteria are numbered according to conventions set by finished sequences and are only numbered every five ORFs for clarity. ORFs are not always numbered serially in finished sequences as illustrated by the absence of HD0125 in the *H. ducreyi* annotation. Diamonds in the *S. enterica* subsp. *enterica* Typhi CT18 locus show a series of invertible sequence repeats. Colors and shades of ORFs designate similar sequences among organisms. Numbers in the size scales are kilobases.

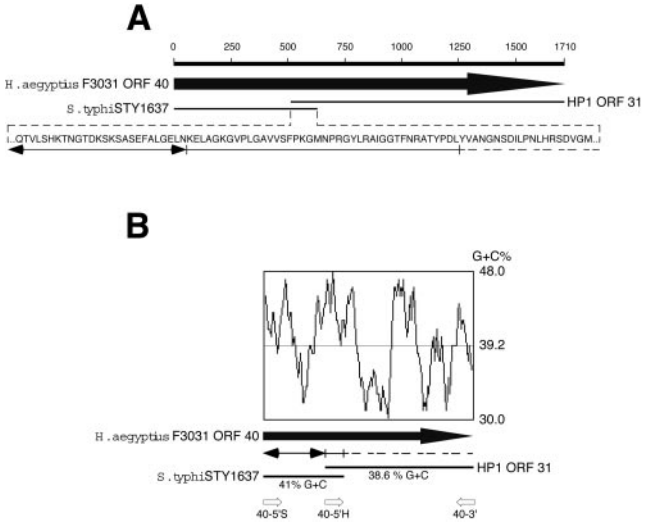

FIG. 4. Illustration of the chimeric structure of the genomic island ORF 40. (A) ORF 40 is graphically depicted by the large arrow with direction of transcription shown. Portions of the encoded product from this ORF that have similarity to proteins from serovar Typhi CT18 and *H. influenzae* HP1 phage are shown as lines. The region in ORF 40 that shares similarity to both proteins is enlarged to show the amino acid sequence for the ORF 40 product. Amino acids underscored with a double-headed arrow have similar sequences in the serovar Typhi CT18 ORF STY1637, portions that are similar to HP1 ORF 31 are shown as underscored hashmarks, and the region in ORF40 that is common to both proteins is underlined with a solid line. (B) The G+C content from the F3031 ORF 40 is shown along the entire length of the gene with a window of 100 nucleotides. The minimum, maximum, and mean of the G+C content are shown on the right side. The bars on the bottom illustrates, to scale, the ORF 40 with the separate regions of homology with the encoded products of serovar Typhi CT18 STY1637 and HP1 ORF 31 depicted as in panel A. The individual G+C percent values for the separate domains are also indicated. The open arrows indicate the location of the primers 40-5′S, 40-5′H, and 40-3′ used in the RT-PCRs.

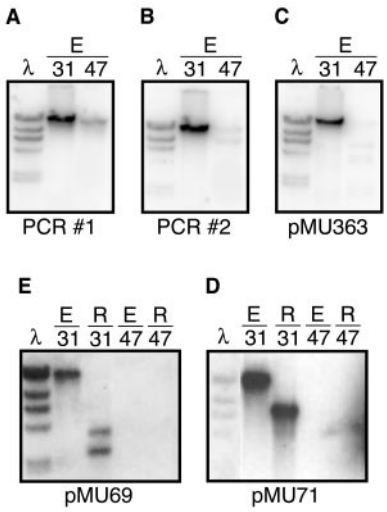

FIG. 5. Prevalence of genomic island homologous sequences in the invasive and noninvasive isolates. Total DNA from F3031 (lanes labeled 31) and F1947 (lanes labeled 47) was digested with either EcoRI (E) or RsaI (R) and blotted to nitrocellulose after agarose gel electrophoresis. Lambda DNA (λ) cut with HindIII was used as a size marker. The probe used to test each blot is indicated at the bottom of each panel. All blots were also probed with radiolabeled λ DNA.

38.6% (Fig. 4B), respectively, giving evidence for the potential chimeric structure of this predicted gene. In between the genes encoding tail fibers are three ORFs (ORF 36 to ORF 38) that putatively encode phage baseplate proteins. ORF 38 has the highest G+C content in this region III of the genomic island, suggesting that it was acquired horizontally from an unrelated source. It should be noted that all three ORFs code for products with similarity to products encoded by *P. luminescens* TT01, serovar Typhi CT18, and *E. carotovora* (Fig. 3). It would seem from the sequence similarity that all of the structural components needed for the bacteriocin could be encoded from the already presented ORFs. Based on the organization in the *E. carotovora* carotovoricin Er locus (accession reference AB017338), expression of ORFs 11, 29, 30, 32, 33, 35, 36, 37, 38, 39, and 40 in the genomic island would be sufficient for expression of a carotovoricin-like bacteriocin. This conclusion is also supported by the drastic drop-off in synteny seen between the remaining ORFs in the genomic island and the serovar Typhi CT18 (35), *C. violaceum* (4), and *E. carotovora* loci (accession reference AB017338) (Fig. 3). Still, ORF 41 has the potential to encode another virus-like component since it has similarity to the product of ORF 32 from phage HP2 (52), which is thought to encode a tail collar protein. This HP2 phage is known to infect nontypeable *H. influenzae* strains, unlike HP1 that has only been described to infect *H. influenzae* Rd strains (52). This ORF also has similarity to the encoded product Plu3421 from *P. luminescens*. ORF42 differs from the other components in this region since it codes for a protein product with similarity to the *H. influenzae* R2866 enoyl-coenzyme A hydratase/carnithine racemase (accession reference ZP_00156541) and also has similarity, although to a lesser extent, to *H. somnus* ORF113401 (accession reference ZP_00132835), for which no function has been defined. The Mu-like prophage protein Com has similarity to the product of ORF 43, which has the lowest G+C content in region III at 34.2%. Com has been described previously as a regulator to the Mu gene *mom*, which is involved in DNA modification of adenine bases (29), although no such *mom* homolog is present in the F3031 genomic island. The final ORF in the genomic island resumes similarity to *Haemophilus* genes with matches to the *H. influenzae* Rd KW20 protein from HI1523 (13) and the protein produced from the *H. ducreyi* ORF HD0156 (accession reference NC_002940), whose functions have not been described.

Analyses of the presence of F3031 genomic island regions in F1947 and other *Haemophilus* strains. The fact that eight subtracted fragments from our previously published genome subtraction hybridization (40) and three fragments from another laboratory's subtraction procedure (23) localize to the cloned DNA fragment in pMU263 (Fig. 1 and 2) suggests that most of the DNA in this construct is unique to F3031 and is important in differentiating BPF strains from noninvasive biogroup aegyptius strains. Southern analysis was completed to compare the presence of genomic island sequences between F3031 and the conjunctivitis-producing F1947 strain (Fig. 5). Figure 5A shows that when total DNA from the F1947 strain is hybridized with a probe covering from the beginning of ORF 5 to the 5′

region of ORF 21, a faint band can be seen (lane 47), suggesting that some similar sequences exist in the noninvasive strain but that the entire region is most likely not represented in F1947. This is supported by the difference in signal intensity between the F3031 and F1947 samples even though DNA was loaded from the F1947 strain at a higher concentration. This assertion is also supported by the fact that the subtracted fragments B1, B2, D2, and D13, which proved to be unique to the BPF strain F3031 (40), mapped to this location. Taken together, these observations indicate that genes similar to ORF 9 and ORF 13 and sequences related to portions of ORFs 8, 10, 12, 14, 19, and 20 (Fig. 2) are not found in F1947. These results limit the area that gave hybridization seen in Fig. 5A to regions encompassing ORFs 5, 6, 7, 11, 15, 16, 17, and/or 18. We suggest that the region from ORF 5 to ORF 7 is responsible for the hybridization seen, given both the conservation of sequences with other *Haemophilus* sequences in this region and the lack of *Haemophilus*-like sequences between ORF 15 and ORF 18.

Despite the slight hybridization seen in Fig. 5A, it is clear from other Southern hybridization experiments that sequences similar to the entire remaining portion of the genomic island are not found in the F1947 strain (Fig. 5B to E). Using probes that span from the 3′ region of ORF 20 to the end of ORF 40, no hybridization can be seen. In our previous study (40), we found that MU33-related sequences encompassing ORFs 27 and 28 could be found in three other BPF strains, two of four typeable strains tested, and none of eight *H. influenzae* typeable strains, including all serotypes except d. Taken together, the Southern hybridization data show that, at least, 18,649 bp of contiguous sequence, which is conserved in other BPF isolates, is not found in the genome of F1947, some of which are conserved in other BPF isolates. Although not shown, the results obtained with DNA hybridization and colony PCR support the latter possibility. The Brazilian BPF strains F3028 and Valparaiso were positive when tested for the presence of sequences related to all three regions of the F3031 genomic island. In contrast, the F3043 non-BPF strain was negative for regions II and III and showed the same faint product detected when the conjunctivitis isolate F1947 was tested for the presence of region I. None of the three regions could be detected in the total DNA isolated from the nontypeable otitis strain 86-028NP and the type a to f strains, with the exception of type b, which tested positive for the region III by colony PCR. However, the amplicon obtained with this strain was different in size and fainter than that obtained when F3031 DNA was used as a template. This result is consistent with a report describing the presence of genomic islands in type b strains that could code for phage tail fiber and outer membrane proteins of undefined functions (3). Interestingly, none of the three regions of the F3031 genomic island could be detected in the Australian BPF strain F4380, a result that is consistent with our previous report showing that this strain does not have sequences related to the BPF33 locus (40). Taken together, these results indicate that the sequences of this F3031 genomic island, particularly those located in regions I and II, are unique to the invasive BPF strains of biogroup aegyptius isolated in Brazil. The fact that the DNA isolated from the biogroup aegyptius Connecticut strain was positive when tested with a BPF33 probe (40) but negative when tested for the presence of sequences related to ORFs 11 to 13, ORFs 15 to 18, and ORF 32, which represent regions I, II, and III (Fig. 2), respectively, is in accordance with the observation that this isolate does not display all of the genetic and phenotypic markers described for the BPF type strain F3031 (48). These observations indicate that this isolate may contain some but not all of the elements of the F3031 genomic island.

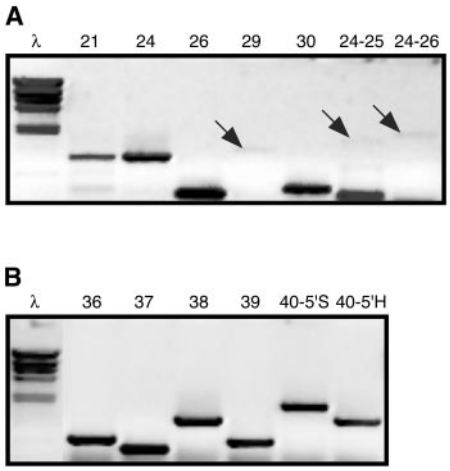

FIG. 6. RT-PCR of selected ORFs from the genomic island in F3031. Lambda DNA digested with HindIII (λ) was used as a size marker. Transcripts from each ORF, identified by the numbers on top of each lane, were reverse transcribed and PCR amplified with the primers listed in Table 1. The amplicons shown in the lanes labeled 40-5′S and 40-5′H were obtained by using the primer 40-3′ in combination with 40-5′S and 40-5′H (Table 1 and Fig. 4B), respectively. The arrows identify the weak bands corresponding to the predicted sizes according to the sequence analysis of each ORF. The images of the ethidium bromide-stained agarose gels were inverted with Photoshop for more clarity of the images.

Expression analysis of selected ORFs within the F3031 genomic island. The transcriptional expression was verified for some of the ORFs in the genomic island by using RT-PCR. Figure 6 shows that all tested ORFs were expressed in F3031 cells cultured under routine conditions, although not all of them to the same levels. The transcription products of ORF 21, ORF 24, ORF 26, ORF 30, and ORF 36 to ORF 40 are readily detectable. In contrast, the transcription product of ORF 29, which is predicted to code for a protein similar to the *E. carotovora* tail sheath protein (Table 3), is the weakest. This is an interesting observation because Nguyen and others have recently reported that expression of the *E. carotovora* gene encoding the tail sheath is minimal at 37°C when treated with mitomycin C but increases in expression at lower temperatures (32). This expression is regulated, in part, by the homolog of *rdgB*, which has a counterpart in the F3031 genomic island in ORF 11. We have attempted purification of these phage-tail-like particles from F3031 cultures grown at 37°C after UV treatment, and we have failed to see these carotovoricin-like particles in electron microscope preparations. Also, we have never detected zones of killing when testing for bacteriocin activity in the usual agar plate diffusion assays when incubated overnight at 37°C (33). The lack of activity at higher temperatures has also been seen with the *Yersinia enterocolitica* bacteriocin enterocoliticin and, therefore, may be a general fea-

ture of this class of bacteriocins (45). We are currently investigating expression of the F3031 carotovoricin Er-like bacteriocin at lower temperatures.

This RT-PCR analysis also showed that ORF 24 to ORF 26 (Fig. 6) are transcribed as a polycistronic message, although at low levels, suggesting that they may have similar or coordinated biological functions. The encoded product of ORF 26 has similarity to Erp22 and ErpD, and the *erpD* gene is in a bicistronic message with *erpC* in *B. burgdorferi* N40 (43). No ErpC homolog is found in the F3031 genomic island but the functions of ORF 24 and ORF 25 are poorly described (Table 3). This analysis also showed the presence of mRNA that was produced from the hybrid ORF 40 in both the region that has similarity to the HP1 phage ORF 31 (Fig. 6B, 40-5′H) and the serovar Typhi CT18 ORF STY1637 (Fig. 6B, 40-5′S). This is an important result because it supports the hypothesis that this ORF encodes a product with two separate domains that are contained within a single coding region. Taken together, the RT-PCR data show that the tested ORFs are expressed and suggest that the in silico analysis presented here has biological relevance.

Further comparison of the F3031 genomic island with other bacterial loci. The conservation between this F3031 region and the regions in serovar Typhi CT18, *P. luminescens* TT01, *H. ducreyi* 35000HP, *E. carotovora*, and *C. violaceum* ATCC 12472 is striking (Fig. 3), considering that these bacteria are found in different ecological niches. The shared region between F3031 and *S. enterica* serovar Typhi CT18 is the most impressive in regards to conservation (Fig. 3). The F3031 genomic island-like sequence in serovar Typhi CT18 covers the region from STY1601 to STY1637 (35). The serovar Typhi CT18 region from STY1638 to STY1640, which flanks the genomic island in this bacterium, contains numerous invertible sequences, suggesting that the entire or part of the island could be inverted. This possibility is supported by the observation that the predicted product from STY1604 has similarity to an integrase, whereas the predicted protein produced from STY1643 is most similar to an invertase/recombinase. In contrast, no such invertible sequences and invertase/recombinase genes were found in the regions flanking either ends of the F3031 genomic island, suggesting that this genetic element could be more stable compared to that of serovar Typhi CT18. When we compared the *S. enterica* subsp. *enterica* serovar Typhi Ty2 (*S. enterica* Ty2) and CT18 genomes, in relation to the described STY locus, we found that the strains only differed in a single ORF, STY1641. However, comparison of the CT18 and serovar Typhimurium LT2 genomes (24) showed that the entire locus from STY1600 to STY1643 was unique to the enteropathogens serovar Typhi CT18 and *S. enterica* Ty2, an observation that suggests that functions encoded within this genomic region may play a role in disease (36). All of these facts suggest that some of the genes located within the F3031 genomic island may play a role in the unusual virulence attributes of the F3031 BPF isolate of biogroup aegyptius.

Another interesting observation is that a gene with similarity, at the protein level, to arginine/ornithine antiporters is located at STY1645 and a two-component response regulator is present at STY1646 and STY1647 (35). A similar arrangement is found in the area adjacent to the 3′ end of the F3031 genomic island, where two ORFs (HI1707 and HI1708) encoding potential two-component regulators were mapped after the 3′ fragment of the interrupted choline transport homolog HI1706 (Fig. 3). This would suggest that transporter-encoding genes, particularly those next to two-component regulatory genes, are targets for the insertion of genomic island elements. This possibility is supported by the observation that the *C. violaceum* genomic region from CV2115 to CV2165, which has sequences similar to F3031 genomic island sequences (Fig. 3), is next to an ORF (CV2167) with an encoded product similar to a sodium/malate symporter (4).

The production of a bacteriocin in *Haemophilus* is not novel (30). It has been known for some time that only type b strains can express haemocin, although the importance of this strict relationship is not known. However, no phage tail-like bacteriocin similar to that found in the plant pathogen *E. carotovora* has been described previously for a *Haemophilus* species, and it is interesting that the genome of biogroup aegyptius F3031 contains BPF-specific genes predicted to encode this type of bacteriocin. Equally interesting is the observation that a noncontiguous region similar to the F3031 genomic island, which covers from Plu3421 to Plu3490, is found in *P. luminescens* TT01. This is a symbiont of nematodes that acts as a vector for transmission to certain insects, which has the highest number of proposed toxin genes of any sequenced bacterium to date (9, 12). Upon infection of an insect, bacteria quickly overwhelm and kill the insect, presumably through the expression of a combination of virulence factors, including the production of a R-type bacteriocin, photorhabdicin, which was seen when cells were grown at 28°C (14). Analogous structures in *P. luminescens* W14 has shown killing of other *Photorhabdus* strains, as well as of *E. coli* (12). Plu3461 encodes a pectin lyase-like regulator and Plu3414 contains an endolysin- and lysozyme-like ORF, which is analogous to the predicted *N*-acetylmuramoyl-L-alanine amidase encoded by the ORF 13 of F3031 genomic island. The protein product of this ORF in *Photorhabdus* is presumably involved in release of the bacteriocin into the external environment (12). Based on these observations, it is tempting to speculate that F3031 may encode a phage-like bacteriocin that remains to be detected and characterized. The finding that ORF 40, predicted to encode a phage tail protein, has a hybrid structure, including a domain with similarity to a tail gene from HP1, helps to account for the evolutionary significance of the F3031 region harboring bacteriocin-like genes. Bacteriocins are usually specific for killing strains that are closely related to the producing bacterium, and the importance of this BPF region would be diminished if we could not explain how this bacteriocin could affect more closely related strains other than *Erwinia*- or *Photorhabdus*-like bacteria. Instead, this fusion could ensure that killing activity can be directed against other *Haemophilus* species via the HP1-like tail gene.

Another startling finding in this F3031-specific genomic island is the number of subtracted fragments that map to this location. In our previous work, we isolated 13 fragments specific to the BPF strain F3031 (40), 8 of which are located in this genomic island. In addition, another subtraction procedure done by Li and others (23), who performed a subtraction procedure similar to ours but with different strains, has three of their fragments map to this location in close proximity to each other. It is thus evident from two studies that these subtraction

procedures did not merely produce short pieces of DNA found scattered in the genome with no relationship to each other but rather signify a region to be investigated more closely. Many of the ORFs in this region have no previously described function and therefore could be important in explaining what makes the F3031 strain so virulent compared to the F1947 strain. The fact that all of the genes tested in this region have an mRNA product substantiates that these genes need to be investigated further, especially now that it is known that they are genetically linked.

In summary, the present study describes the largest genomic island found to date in a *Haemophilus* strain, a finding that indicates that members of this genus can acquire large DNA fragments from unrelated sources as it has been described for other human pathogens. Our data also support the hypothesis that the integration of FluMu- and HP1-like phages played a role in the integration of foreign DNA within a gene coding for transport functions. However, it seems that at least one additional process reshaped this genomic region of the BPF strains. The conservation of sequences from region II and III in the genomic island to similar regions in serovar Typhi and *P. luminescens* supports this hypothesis. Based on the composition of region III, it is also possible to speculate that a phage-like DNA-containing bacteriocin, similar to that produced by *P. aeruginosa* (22), could have been involved in the reshaping process. The fact that this island also has extensive conservation with regions found in the genome of the soil bacterium *C. violaceum* and the insect pathogen *P. luminescens* suggests that the Brazilian F3031 isolate has acquired DNA material from unrelated sources in the environment, a process that has not been described for any *Haemophilus* species. Clearly, the role that horizontal gene transfer has played in the evolution *Haemophilus* is substantial and is only starting to be appreciated.

ACKNOWLEDGMENTS

Miami University research funds and Public Health Service grants R15AI37781-01 and R15AI44776-01A1 from the National Institutes of Health funded this work.

We thank D. J. Brenner (CDC) and A. Lesse (VAMC, Buffalo, N.Y.) for providing the biogroup aegyptius isolates. We thank L. Bakaletz (Ohio State University) for the nontypeable *H. influenzae* otitis isolates and S. H. Goodgal (University of Pennsylvania) for providing *H. influenzae* strain Eagan. We thank M. Fields (Miami University) for providing the primers D1F and 1540R. We are grateful to C. Wood, Miami University Center of Bioinformatics and Functional Genomics, for support and assistance with automated DNA sequencing and nucleotide sequence analysis.

REFERENCES

1. **Baker, J., R. Limberger, S. J. Schneider, and A. Campbell.** 1991. Recombination and modular exchange in the genesis of new lambdoid phages. New Biol. **3:**297–308.
2. **Bendler, J. W., and S. H. Goodgal.** 1968. Prophage S2 mutants in *Haemophilus influenzae*: a technique for their production and isolation. Science **162:**464–465.
3. **Bergman, N. H., and B. J. Akerley.** 2003. Position-based scanning for comparative genomics and identification of genetic islands in *Haemophilus influenzae* type b. Infect. Immun. **71:**1098–1108.
4. **Brazilian National Genome Project and Consortium.** 2003. The complete genome sequence of *Chromobacterium violaceum* reveals remarkable and exploitable bacterial adaptability. Proc. Natl. Acad. Sci. USA **100:**11660–11665.
5. **Brazilian Purpuric Fever Study Group.** 1987. Brazilian Purpuric fever: epidemic purpura fulminans associated with antecedent purulent conjunctivitis. Lancet **ii:**761–763.
6. **Brazilian Purpuric Fever Study Group.** 1987. *Haemophilus aegyptius* bacteraemia in Brazilian purpuric fever. Lancet **ii:**757–761.
7. **Brenner, D. J., L. W. Mayer, G. M. Carlone, L. H. Harrison, W. F. Bibb, M. C. de Cunto Brandileone, F. O. Sottnek, K. Irino, M. W. Reeves, J. M. Swenson, K. A. Birkness, R. S. Weyant, S. F. Berkley, T. C. Woods, A. G. Steigerwalt, P. A. D. Grimont, R. C. Cooksey, R. J. Arko, C. V. Broome, and The Brazilian Purpuric Fever Study Group.** 1988. Biochemical, genetic, and epidemiological characterization of *Haemophilus influenzae* biogroup aegyptius. J. Clin. Microbiol. **26:**1524–1534.
8. **Chang, C. C., J. R. Gilsdorf, V. J. DiRita, and C. F. Marrs.** 2000. Identification and genetic characterization of *Haemophilus influenzae* genetic island 1. Infect. Immun. **68:**2630–2637.
9. **Duchaud, E., C. Rusniok, L. Frangeul, C. Buchrieser, A. Givaudan, S. Taourit, S. Bocs, C. Boursaux-Eude, M. Chandler, J. F. Charles, E. Dassa, R. Derose, S. Derzelle, G. Freyssinet, S. Gaudriault, C. Medigue, A. Lanois, K. Powell, P. Siguier, R. Vincent, V. Wingate, M. Zouine, P. Glaser, N. Boemare, A. Danchin, and F. Kunst.** 2003. The genome sequence of the entomopathogenic bacterium *Photorhabdus luminescens*. Nat. Biotechnol. **21:**1307–1313.
10. **Esposito, D., W. P. Fitzmaurice, R. C. Benjamin, S. D. Goodman, A. S. Waldman, and J. J. Scocca.** 1996. The complete nucleotide sequence of bacteriophage HP1 DNA. Nucleic Acids Res. **24:**2360–2368.
11. **Feinberg, A. P., and B. Vogelstein.** 1983. A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal. Biochem. **132:**6–13.
12. **Ffrench-Constant, R., N. Waterfield, P. Daborn, S. Joyce, H. Bennett, C. Au, A. Dowling, S. Boundy, S. Reynolds, and D. Clarke.** 2003. *Photorhabdus*: toward a functional genomic analysis of a symbiont and pathogen. FEMS Microbiol. Rev. **26:**433–456.
13. **Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J.-F. Tomb, B. A. Dougherty, J. M. Merrick, K. McKenney, G. Sutton, W. FitzHugh, C. Fields, J. D. Gocayne, J. Scott, R. Shirley, L.-I. Liu, A. Glodek, J. M. Kelley, J. F. Weidman, C. A. Phillips, T. Spriggs, E. Hedblom, M. D. Cotton, T. R. Utterback, M. C. Hanna, T. Nguyen, D. M. Saudek, R. C. Brandon, L. D. Fine, J. L. Fritchman, J. L. Fuhrmann, N. S. M. Geoghagen, C. L. Gnehm, L. A. McDonald, K. V. Small, C. M. Fraser, H. O. Smith, and J. C. Venter.** 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science **269:**496–512.
14. **Gaudriault, S., J. O. Thaler, E. Duchaud, F. Kunst, N. Boemare, and A. Givaudan.** 2004. Identification of a P2-related prophage remnant locus of *Photorhabdus luminescens* encoding an R-type phage tail-like particle. FEMS Microbiol. Lett. **233:**223–231.
15. **Goodgal, S. H., and M. A. Mitchell.** 1990. Sequence and uptake specificity of cloned sonicated fragments of *Haemophilus influenzae* DNA. J. Bacteriol. **172:**5924–5928.
16. **Graber, K., L. M. Smoot, and L. A. Actis.** 1998. Expression of iron binding proteins and hemin binding activity in the dental pathogen *Actinobacillus actinomycetemcomitans*. FEMS Microbiol. Lett. **163:**135–142.
17. **Harm, W., and C. S. Rupert.** 1963. Infection of transformable cells of *Haemophilus influenzae* by bacteriophage and bacteriophage DNA. Zentbl. Vererbungsl. **94:**336–348.
18. **Harrison, L. H., G. A. da Silva, M. Pittman, D. W. Fleming, A. Vranjac, and C. V. Broome.** 1989. Epidemiology and clinical spectrum of Brazilian purpuric fever. Brazilian Purpuric Fever Study Group. J. Clin. Microbiol. **27:**599–604.
19. **Hendrix, R. W., M. C. Smith, R. N. Burns, M. E. Ford, and G. F. Hatfull.** 1999. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proc. Natl. Acad. Sci. USA **96:**2192–2197.
20. **Hertwig, S., I. Klein, V. Schmidt, S. Beck, J. A. Hammerl, and B. Appel.** 2003. Sequence analysis of the genome of the temperate *Yersinia enterocolitica* phage PY54. J. Mol. Biol. **331:**605–622.
21. **Juhala, R. J., M. E. Ford, R. L. Duda, A. Youlton, G. F. Hatfull, and R. W. Hendrix.** 2000. Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. J. Mol. Biol. **299:**27–51.
22. **Lee, F. K., K. C. Dudas, J. A. Hanson, M. B. Nelson, P. T. LoVerde, and M. A. Apicella.** 1999. The R-type pyocin of *Pseudomonas aeruginosa* C is a bacteriophage tail-like particle that contains single-stranded DNA. Infect. Immun. **67:**717–725.
23. **Li, M. S., J. L. Farrant, P. R. Langford, and J. S. Kroll.** 2003. Identification and characterization of genomic loci unique to the Brazilian purpuric fever clonal group of *H. influenzae* biogroup aegyptius: functionality explored using meningococcal homology. Mol. Microbiol. **47:**1101–1111.
24. **Liu, Y., A. Chatterjee, and A. K. Chatterjee.** 1994. Nucleotide sequence, organization and expression of *rdgA* and *rdgB* genes that regulate pectin lyase production in the plant pathogenic bacterium Erwinia carotovora subsp. carotovora in response to DNA-damaging agents. Mol. Microbiol. **14:**999–1010.
25. **McIntyre, P., G. Wheaton, J. Erlich, and D. Hansman.** 1987. Brazilian purpuric fever in Central Australia. Lancet **ii:**112.
26. **Meade, H. M., S. R. Long, S. E. Ruvkum, S. E. Brown, and F. M. Ausubel.** 1982. Physical and genetic characterization of symbiotic and auxotrophic

mutants *Rhizobium meliloti* induced by transposon Tn*5* mutagenesis. J. Bacteriol. **149:**114–122.
27. **Miller, J. C., K. von Lackum, K. Babb, J. D. McAlister, and B. Stevenson.** 2003. Temporal analysis of *Borrelia burgdorferi* Erp protein expression throughout the mammal-tick infectious cycle. Infect. Immun. **71:**6943–6952.
28. **Mizuuchi, M., R. A. Weisberg, and K. Mizuuchi.** 1986. DNA sequence of the control region of phage D108: the N-terminal amino acid sequences of repressor and transposase are similar both in phage D108 and in its relative, phage Mu. Nucleic Acids Res. **14:**3813–3825.
29. **Morgan, G. J., G. F. Hatfull, S. Casjens, and R. W. Hendrix.** 2002. Bacteriophage Mu genome sequence: analysis and comparison with mu-like prophages in *Haemophilus*, *Neisseria*, and *Deinococcus*. J. Mol. Biol. **317:** 337–359.
30. **Murley, Y. M., T. D. Edlind, P. A. Plett, and J. J. LiPuma.** 1998. Cloning of the haemocin locus of *Haemophilus influenzae* type b and assessment of the role of haemocin in virulence. Microbiology **144:**2531–2538.
31. **Nguyen, A. H., T. Tomita, M. Hirota, T. Sato, and Y. Kamio.** 1999. A simple purification method and morphology and component analyses for carotovoricin Er, a phage-tail-like bacteriocin from the plant pathogen *Erwinia carotovora* Er. Biosci. Biotechnol. Biochem. **63:**1360–1369.
32. **Nguyen, H. A., J. Kaneko, and Y. Kamio.** 2002. Temperature-dependent production of carotovoricin Er and pectin lyase in phytopathogenic *Erwinia carotovora* subsp. carotovora Er. Biosci. Biotechnol. Biochem. **66:**444–447.
33. **Nguyen, H. A., T. Tomita, M. Hirota, J. Kaneko, T. Hayashi, and Y. Kamio.** 2001. DNA inversion in the tail fiber gene alters the host range specificity of carotovoricin Er, a phage-tail-like bacteriocin of phytopathogenic *Erwinia carotovora* subsp. carotovora Er. J. Bacteriol. **183:**6274–6281.
34. **O'Neill, K. H., D. M. Roche, D. J. Clarke, and B. C. Dowds.** 2002. The *ner* gene of *Photorhabdus*: effects on primary-form-specific phenotypes and outer membrane protein composition. J. Bacteriol. **184:**3096–3105.
35. **Parkhill, J., G. Dougan, K. D. James, N. R. Thomson, D. Pickard, J. Wain, C. Churcher, K. L. Mungall, S. D. Bentley, M. T. Holden, M. Sebaihia, S. Baker, D. Basham, K. Brooks, T. Chillingworth, P. Connerton, A. Cronin, P. Davis, R. M. Davies, L. Dowd, N. White, J. Farrar, T. Feltwell, N. Hamlin, A. Haque, T. T. Hien, S. Holroyd, K. Jagels, A. Krogh, T. S. Larsen, S. Leather, S. Moule, P. O'Gaora, C. Parry, M. Quail, K. Rutherford, M. Simmonds, J. Skelton, K. Stevens, S. Whitehead, and B. G. Barrell.** 2001. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. Nature **413:**848–852.
36. **Parkhill, J., B. W. Wren, N. R. Thomson, R. W. Titball, M. T. Holden, M. B. Prentice, M. Sebaihia, K. D. James, C. Churcher, K. L. Mungall, S. Baker, D. Basham, S. D. Bentley, K. Brooks, A. M. Cerdeno-Tarraga, T. Chillingworth, A. Cronin, R. M. Davies, P. Davis, G. Dougan, T. Feltwell, N. Hamlin, S. Holroyd, K. Jagels, A. V. Karlyshev, S. Leather, S. Moule, P. C. Oyston, M. Quail, K. Rutherford, M. Simmonds, J. Skelton, K. Stevens, S. Whitehead, and B. G. Barrell.** 2001. Genome sequence of *Yersinia pestis*, the causative agent of plague. Nature **413:**523–527.
37. **Sambrook, J., E. F. Fritsch, and T. Maniatis.** 1989. Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
38. **Samuels, J., and J. K. Clarke.** 1969. New bacteriophage of *Haemophilus influenzae*. J. Virol. **4:**797–798.
39. **Sandmeier, H.** 1994. Acquisition and rearrangement of sequence motifs in the evolution of bacteriophage tail fibers. Mol. Microbiol. **12:**343–350.
40. **Smoot, L. M., D. D. Franke, G. McGillivary, and L. A. Actis.** 2002. Genomic analysis of the F3031 Brazilian purpuric fever clone of *Haemophilus influenzae* biogroup aegyptius by PCR-based subtractive hybridization. Infect. Immun. **70:**2694.
41. **Stevenson, B., N. El-Hage, M. A. Hines, J. C. Miller, and K. Babb.** 2002. Differential binding of host complement inhibitor factor H by *Borrelia burgdorferi* Erp surface proteins: a possible mechanism underlying the expansive host range of Lyme disease spirochetes. Infect. Immun. **70:**491–497.
42. **Stevenson, B., and J. C. Miller.** 2003. Intra- and interbacterial genetic exchange of Lyme disease spirochete *erp* genes generates sequence identity amidst diversity. J. Mol. Evol. **57:**309–324.
43. **Stevenson, B., K. Tilly, and P. A. Rosa.** 1996. A family of genes located on four separate 32-kilobase circular plasmids in *Borrelia burgdorferi* B31. J. Bacteriol. **178:**3508–3516.
44. **Stover, C. K., X. Q. Pham, A. L. Erwin, S. D. Mizoguchi, P. Warrener, M. J. Hickey, F. S. L. Brinkman, W. O. Hufnagle, D. J. Kowalik, M. Lagrou, R. L. Garber, L. Goltry, E. Tolentino, S. Westbrook-Wadman, Y. Yuan, L. L. Brody, S. N. Coulter, K. R. Folger, A. Kas, K. Larbig, R. M. Lim, K. A. Smith, D. H. Spencer, G. K. S. Wong, Z. Wu, I. T. Paulsen, J. Reizer, M. H. Saier, R. E. W. Hancock, S. Lory, and M. V. Olson.** 2000. Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. Nature **406:**959–964.
45. **Strauch, E., H. Kaspar, C. Schaudinn, P. Dersch, K. Madela, C. Gewinner, S. Hertwig, J. Wecke, and B. Appel.** 2001. Characterization of enterocoliticin, a phage tail-like bacteriocin, and its effect on pathogenic *Yersinia enterocolitica* strains. Appl. Environ. Microbiol. **67:**5634–5642.
46. **Tolias, P. P., and M. S. DuBow.** 1985. The cloning and characterization of the bacteriophage D108 regulatory DNA-binding protein *ner*. EMBO J. **4:**3031–3037.
47. **Tondella, M. L. C., F. D. Quin, and B. A. Perkins.** 1995. Brazilian purpuric fever caused by *Haemophilus influenzae* biogroup aegyptius strains lacking the 3031 plasmid. J. Infect. Dis. **171:**209–212.
48. **Virata, M., N. E. Rosenstein, J. L. Hadler, N. L. Barrett, M. L. Tondella, L. W. Mayer, R. S. Weyant, B. Hill, and B. A. Perkins.** 1998. Suspected Brazilian purpuric fever in a toddler with overwhelming Epstein-Barr virus infection. Clin. Infect. Dis. **27:**1238–1240.
49. **Waterfield, N. R., P. J. Daborn, and R. H. Ffrench-Constant.** 2002. Genomic islands in *Photorhabdus*. Trends Microbiol. **10:**541–545.
50. **White, O., J. A. Eisen, J. F. Heidelberg, E. K. Hickey, J. D. Peterson, R. J. Dodson, D. H. Haft, M. L. Gwinn, W. C. Nelson, D. L. Richardson, K. S. Moffat, H. Qin, L. Jiang, W. Pamphile, M. Crosby, M. Shen, J. J. Vamathevan, P. Lam, L. McDonald, T. Utterback, C. Zalewski, K. S. Makarova, L. Aravind, M. J. Daly, K. W. Minton, R. D. Fleischmann, K. A. Ketchum, K. E. Nelson, S. Salzberg, H. O. Smith, J. C. Venter, and C. M. Fraser.** 1999. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science **286:**1571–1577.
51. **Wilde, B. E., J. W. Pearman, P. B. Campbell, P. B. Swan, and D. L. Gurry.** 1989. Brazilian purpuric fever in Western Australia. Med. J. Aust. **150:**344–346.
52. **Williams, B. J., M. Golomb, T. Phillips, J. Brownlee, M. V. Olson, and A. L. Smith.** 2002. Bacteriophage HP2 of *Haemophilus influenzae*. J. Bacteriol. **184:**6893–6905.
53. **Wu, C.-J., and G. R. Janssen.** 1997. Expression of a streptomycete leaderless mRNA encoding chloramphenicol acetyltransferase in *Escherichia coli*. J. Bacteriol. **179:**6824–6830.
54. **Yanisch-Perron, C., J. Vieira, and J. Messing.** 1985. Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. Gene **33:**103–119.
55. **Zhou, J., B. Xia, D. S. Treves, L. Y. Wu, T. L. Marsh, R. V. O'Neill, A. V. Palumbo, and J. M. Tiedje.** 2002. Spatial and resource factors influencing high microbial diversity in soil. Appl. Environ. Microbiol. **68:**326–334.

Editor: V. J. DiRita