BIOINFORMATICS ORIGINAL PAPER

Vol. 22 no. 9 2006, pages 1072–1079
doi:10.1093/bioinformatics/bti793

Phylogenetics

Describing ancient horizontal gene transfers at the nucleotide and gene levels by comparative pathogenicity island genometrics

F. Collyn[1,2], L. Guy[2], M. Marceau[1], M. Simonet[1] and C.-A. H. Roten[2,*]

[1]Inserm E0364—Université de Lille II, Faculté de Médecine Henri Warembourg, Institut Pasteur de Lille, 1 rue du Pr Calmette, F-59021 Lille, France and [2]Département de Microbiologie Fondamentale, Bâtiment de Biologie, Faculté de Biologie et de Médecine, Université de Lausanne, CH-1015 Dorigny, Switzerland

Received on July 27, 2005; revised on October 18, 2005; accepted on November 15, 2005
Advance Access publication November 22, 2005
Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Lateral gene transfer is a major mechanism contributing to bacterial genome dynamics and pathovar emergence via pathogenicity island (PAI) spreading. However, since few of these genomic exchanges are experimentally reproducible, it is difficult to establish evolutionary scenarios for the successive PAI transmissions between bacterial genera. Methods initially developed at the gene and/or nucleotide level for genomics, i.e. comparisons of concatenated sequences, ortholog frequency, gene order or dinucleotide usage, were combined and applied here to homologous PAIs: we call this approach comparative PAI genometrics.

Results: YAPI, a *Yersinia* PAI, and related islands were compared with measure evolutionary relationships between related modules. Through use of our genometric approach designed for tracking codon usage adaptation and gene phylogeny, an ancient inter-genus PAI transfer was oriented for the first time by characterizing the genomic environment in which the ancestral island emerged and its subsequent transfers to other bacterial genera.

Contact: claude-alain.roten@unil.ch

Supplementary information: http://www.unil.ch/comparativegenometrics/collyn_et_al_2005/collyn.htm

INTRODUCTION

In pathogenic bacteria, virulence genes are often clustered into pathogenicity islands (PAIs) (Hacker and Kaper, 2000). These laterally acquired genetic elements (GEs) are present on the chromosomes of bacterial pathogens but not on their counterparts in related but harmless bacteria. The foreign origin of PAIs is evidenced by the presence of various mobility genes, as well as by G + C content and codon usage that generally differ from those of the core genome. Since gene acquisition is an important evolutionary process by which microorganisms obtain novel phenotypes (Dobrindt and Hacker, 2001), PAIs play an essential role in virulence gene spreading and contribute to the emergence of new pathogens (Hacker and Kaper, 2000; Ochman *et al*., 2000). However, even though homologous PAIs have already been described in distant bacterial species, no clear scenario describing ancient inter-genus transmission of these mobile units has yet been demonstrated—partly because attempts to experimentally observe PAI transfer have rarely been successful. Thus, bioinformatics appears to be the method of choice for establishing accurate descriptions of lateral genetic exchanges in prokaryotes.

Bioinformatics tools have already been developed for comparative genomics. Two different strategies have been used successfully: the first, based on comparisons of homologous DNA sequences or gene units, measures either (1) gene concatenate similarities (Wolf *et al*., 2001), (2) ortholog frequencies (Snel *et al*., 1999) or (3) gene order conservation (Blanchette *et al*., 1999). In contrast to the first-listed technique, the latter two, which use genes as basic units, are insensitive to homing, i.e. the tendency of the codon usage of mobile elements to become similar to that of the cell counterpart. Both techniques were efficiently developed for viral or mitochondrial DNA sequences affected by high mutation rates (Wolfe *et al*., 1987; Blanchette *et al*., 1999; Montague and Hutchison, 2000; Herniou *et al*., 2003) and have subsequently been applied to bacterial genome phylogeny (Fitz-Gibbon and House, 1999; Snel *et al*., 1999; Tekaia *et al*., 1999; Snel *et al*., 2002). Since homing of transmissible GEs affects phylogenetic analyses based on the similarity of laterally acquired genes, these methods can also be used for comparative PAI analysis.

The second strategy is based on nucleotide signatures and enables comparison of non-homologous DNA sequences. The genome signature reflects codon usage by measuring dinucleotide biases (Karlin and Cardon, 1994; Karlin *et al*., 1998) and is able to (1) specify evolutionary relationships (Karlin *et al*., 1994), (2) measure taxonomical distances between bacteria, plasmids and eukaryotic organelles (Campbell *et al*., 1999) and (3) detect bacterial, chromosomal PAIs (Karlin, 2001). However, the genome signature must be carefully interpreted, since horizontally acquired genes gradually adapt their codon usage and thus their nucleotide composition to those of the core genome: the more ancient the genetic acquisition, the less conserved the PAI signature.

When applied to whole PAIs or some of their gene subsets, a combination of these comparative methods referred to here as comparative pathogenicity island genometrics appears capable of documenting lateral transfer at the gene and nucleotide levels. Homologous PAIs shared by non-closely-related species constitute suitable models for testing this strategy.

We recently identified a 11 kb type IV pilus gene cluster (*pil*) on the chromosome of the enteropathogenic bacteria *Yersinia pseudotuberculosis* (Collyn *et al*., 2002). This virulence factor is encoded by a 98 kb PAI called YAPI (Collyn *et al*., 2004a). A DNA

[*]To whom correspondence should be addressed.

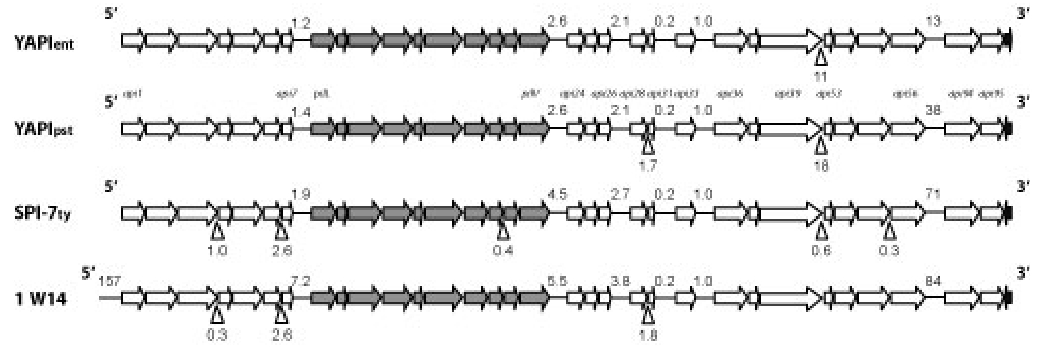

Fig. 1. Genetic map of YAPI-related PAIs. Only homologous CDSs present on four PAIs (YAPI$_{ent}$ from *Y. enterocolitica* 8081, YAPI$_{pst}$ from *Yersinia pseudotuberculosis* 32777, SPI-7$_{ty}$ from *Salmonella enterica* Typhi CT18 and 1 W14 from *Photorhabdus luminescens* W14) are shown. Gray and black arrows represent *pil* and *phe-tRNA* genes, respectively. The size in kb of non-homologous regions is indicated on the map. YAPI$_{pst}$ gene designations are those from Genbank. Supplementary Table 1 displays correspondences between orthologs. SPI-7 described in serovar Typhi (SPI-7$_{ty}$) is also harbored by serovars Dublin and Paratyphi. Despite the fact that SPI-7 gene composition varies in different serovars, the 32 YAPI orthologs are present in all known variants (Pickard *et al*., 2003). We used SPI-7$_{ty}$ in our study, since only the latter was fully sequenced.

segment homologous to *Y.pseudotuberculosis* YAPI (YAPI$_{pst}$) was found in the chromosome of *Yersinia enterocolitica* (Collyn *et al*., 2004b), the other enteropathogenic species in the *Yersinia* genus. Designated as YAPI$_{ent}$, this 66 kb PAI comprises 61 ORFs, of which 41 are homologous to YAPI$_{pst}$ genes. Furthermore, 37 of the latter (including the *pil* genes) have counterparts on the 134 kb SPI-7 PAI from *Salmonella enterica* serovar Typhi (SPI-7$_{ty}$) (Fig. 1), an Enterobacteriaceae found in the host digestive tract along with enteropathogenic Yersiniae. Of the 37 orthologs, 32 shared by YAPIs and SPI-7 were also detected in 1 W14, a 297 kb PAI from the entomopathogenic bacterium *Photorhabdus luminescens* (Duchaud *et al*., 2003). Furthermore, sequence analyses of SPI-7, YAPIs and 1 W14 showed that all contain coding sequences (CDS) related to type IV pilus-encoding genes and other transfer genes harbored on the *Salmonella* conjugative plasmids from the same family, R64 and ColIbP9. All these data suggest a common origin for R64-like plasmids, SPI-7, 1 W14 and YAPIs (Pickard *et al*., 2003; Collyn *et al*., 2004a). Since YAPIs, SPI-7 and 1 W14 are (1) fully sequenced and well documented, (2) harbored by distinct bacterial genera and (3) related to a conjugative plasmid likely to have contributed to their emergence (Pickard *et al*., 2003), they constitute a paradigmatic system for challenging PAI-scale phylogenetic scenarios.

MATERIALS AND METHODS

Sequence sources

The database accession numbers of PAI and plasmid sequences are, respectively, SPI-7$_{ty}$, NC_003198; 1W14, NC_005126, YAPI$_{pst}$, AJ627388; pSLT, NC_003277; R721, NC_002525; R64, NC_005014; ColIb-P9, NC_002122; pCD1, NC_003131; pCP1, NC_003132; pMT1, NC_003134 and pYV, NC_005017. The YAPI$_{ent}$ sequence was downloaded from the Sanger website (www.sanger.ac.uk/Projects/Y_enterocolitica/).

Ortholog identification

Amino acid sequences of all CDSs from the four PAIs and two plasmids were compared with those of related GEs using BLASTP 2.2.10 (Altschul *et al*., 1997) with the BLOSUM62 matrix and gap penalties of 11 and 1. The threshold of *E*-values was 0.01 and non-reciprocal matches were removed. When a protein sequence presented several orthologs, only the best *E*-value comparison was considered.

Similarity analysis of concatenated *pil* genes

All the *pil* gene products housed by the YAPIs, SPI-7, 1 W14, R64 and ColIbP9 were concatenated as previously described (McGeoch *et al*., 2000; Wolf *et al*., 2001; Ling *et al*., 2002; Sharp *et al*., 2005) and compared using ClustalW (Thompson *et al*., 1994). The Neighbor-Joining (NJ) tree was drawn with Treeview v.1.1.6 (Page, 1996).

Evolutionary distance estimated by ortholog frequencies

Evolutionary relationships between PAIs and plasmids were measured by the proportion of orthologous genes shared by pairs of GEs. The method differs slightly from that previously reported (Snel *et al*., 1999). For instance, when selecting GE*a* as reference, the distance between GE*a* and GE*b* corresponds to the proportion of GE*b* genes having a GE*a* homolog. The resulting distance matrix was analyzed using the Phylip package (Felsenstein, 2004) via the Unweighted Pair Group Method with Arithmetic Means (UPGMA), Neighbor-Joining (NJ) and Fitch Margoliash methods. Resulting dendrograms were drawn with Treeview v.1.1.6. Calculated for the UPGMA tree, the root of these ultrametric dendrograms is posted on the figures at the midpoint of the longest pathway between taxa.

Gene order comparison

In order to quantify the inversion and transposition events leading to the current organization of PAIs and plasmids, we measured the proportion of conserved ortholog pairs in these GEs, as previously described (Blanchette *et al*., 1999; Greub *et al*., 2004). Dendrograms and related roots were calculated as for the comparison of ortholog frequencies.

Dinucleotide signature analysis

Dinucleotide signature analysis was performed according to the method of Campbell and co-workers (Karlin and Cardon, 1994; Karlin *et al*., 1998; Campbell *et al*., 1999). The distance between pairs of GEs was estimated by summing up all the absolute differences in dinucleotide bias (Manhattan distances). Resulting matrices were represented as UPGMA, NJ and Fitch Margoliash dendrograms. A dissimilarity matrix of Euclidean distances generated for all PAIs and plasmids was also represented by a principal coordinates (PCO) analysis (Gower, 1966).

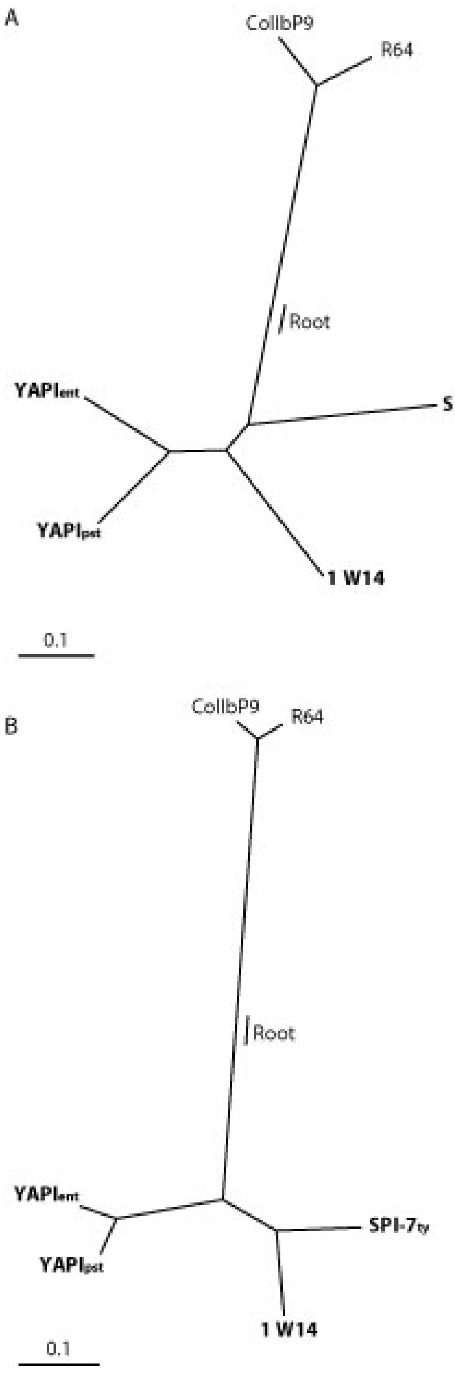

Fig. 2. Gene map comparisons of *pil*-harboring plasmids and PAIs: UPGMA trees of ortholog content (**A**) and gene order conservation (**B**). PAIs are in bold type. The root is indicated on the UPGMA dendrogram by a short segment running parallel to the deepest branch. The bar represents estimated evolutionary distance scale, based on dissimilarity frequencies. Dendrogram topologies were assessed using omit tests.

Dendrogram robustness

The robustness of the analyses shown in Figures 2 and 4 was evaluated by performing an omit test on the UPGMA, NJ and Fitch Margoliash dendrograms: for each final tree of *n* GEs, *n* subtrees are produced with $n - 1$ GEs, where each GE is omitted in turn. The conservation of subtree topologies indicates the robustness of the final dendrogram.

In NJ and Fitch Margoliash distance analysis, the dinucleotide usage distance biases were bootstrapped: 1000 distance matrices were built by random resampling with R 2.0.1 (Ihaka and Gentleman, 1996) and Fitch Margoliash trees were calculated using the Phylip package. A consensus dendrogram obtained with the extended Majority Rule method was used to obtain a bootstrapped evaluation of the original analysis.

RESULTS

YAPI, SPI-7 and 1 W14 were generated by horizontal transfer of an ancestral pathogenicity island

Since the *pil* operon was the only identified functional unit shared by the four PAIs and the two R64-related plasmids, we first measured sequence similarity between Pil proteins encoded by these GEs using ClustalW (Supplementary Fig. 1). Very similar trees were drawn when comparing individual genes (data not shown). Unfortunately, this investigation failed to shed light on PAI emergence: no unambiguous conclusion about PAI transfer could be drawn from the dendrogram, which was characterized by poor resolution of major branches. The divergence of type IV pilus function in the different GEs (plasmids, bacterial conjugation; PAIs, bacterial adhesion, respectively) might also be responsible for *pil* operon sequence variations.

To bypass this difficulty, the frequency of orthologs present on PAIs and plasmids was used to identify those GEs exhibiting the most similar organization (Snel *et al*., 1999). The UPGMA representation of the resulting dissimilarity matrix provided more information than ClustalW analyses: R64-like plasmids were clearly discriminated from YAPIs, SPI-7 and 1 W14 (Fig. 2A), suggesting that a unique, ancestral PAI emerged in a bacterium after chromosomal integration of a R64-like plasmid and was then transferred to other bacterial genera. NJ representations of this distance matrix provided tree topologies similar to those generated by UPGMA (Supplementary Fig. 2). To confirm this scenario, PAIs and plasmids were compared by using gene order breakpoint analysis (Blanchette *et al*., 1999) measuring local gene order and rearrangement events (insertion, deletion, transposition or inversion) in conserved gene sets: we obtained similar results to those obtained by ortholog frequency analysis (Fig. 2B).

The ancestral YAPI emerged from a *Salmonella* GE

The next step was to specify the bacterial environment in which the ancestral PAI emerged. Dinucleotide usage biases, constant along ~50 kb sequences, are efficient tools for comparing small GEs such as plasmids or mitochondria (Campbell *et al*., 1999) and detecting PAIs along prokaryotic chromosomes (Karlin, 2001). Hence, this genometric method appears to be appropriate for PAI comparison but, until now, had never been used for this purpose—most probably because PAIs often display chimeric genetic structures and thus exhibit heterogeneous nucleotide patterns. However, determination of dinucleotide frequencies on related modules (i.e. gene subsets from a common source and no smaller than animal mitochondrial genomes) might be useful for PAI comparison.

Before measuring inter-species/genus relationships, we first challenged the signal constancy of genome signatures of sets of PAI orthologs. Dinucleotide analyses of (1) *pil* operons, (2) genes located upstream or (3) downstream of the *pil* operon and (4) the entire

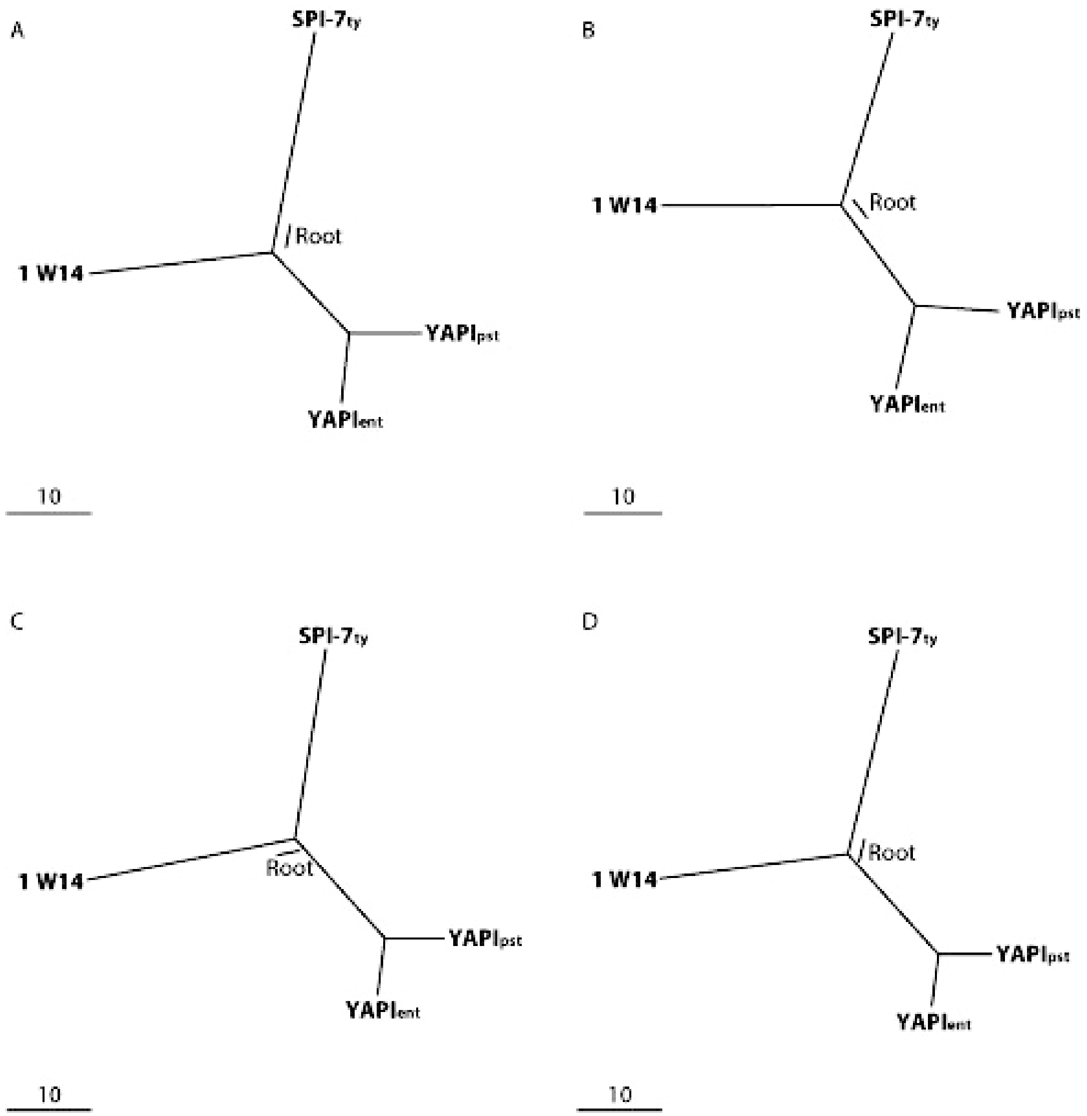

Fig. 3. Dinucleotide signature comparisons for the four PAIs: UPGMA dendrograms of dinucleotide signature comparisons with orthologs (Fig. 1) located upstream (**A**), within (**B**) and downstream (**C**) of the *pil* operon and for the whole set of orthologs (**D**). The similar topology of all four trees indicates that dinucleotide signature is homogeneous within ortholog sets. Roots on UPGMA trees are posted as in Figure 2. The bar represents the distance scale in thousandths of the average dinucleotide usage bias, normalized to nucleotide content (Campbell *et al*., 1999). Dendrogram topologies were assessed using omit tests.

set of orthologs for all PAIs indicated that this gene module constitutes an homogeneous core of an ancestral PAI, since the four dendrogram topologies were conserved (Fig. 3). Figure 3 also shows that *pil* operon dinucleotide usage was similar to those for the orthologs common to the four PAIs: we therefore compared the genome signatures of (1) PAI- and plasmid-borne *pil* operons and (2) all PAI orthologs and the R64-like plasmid sequences, regardless of whether they included *pil* genes or not. All the resulting trees were very similar (Fig. 4). As a control, our analysis also included the R721 self-transmissible plasmid—distantly related to R64 but also bearing *pil* CDSs (Kim and Komano, 1992): its dinucleotide usage diverged strongly from those of the R64-related plasmids, demonstrating that the genome signature method can indeed be used for phylogenetic analysis. NJ and Fitch Margoliash representations provided topologies similar to those generated by UPGMA (Supplementary Figs 3 and 4). All the results presented in Figures 3 and 4 clearly reveal that dinucleotide signatures have been conserved along the sequences of PAIs subsets and plasmids and demonstrate that PAI gene subsets can be reliably compared with whole plasmid sequences, including non-homologous genes.

Finally, in order to identify the bacterial environment in which the ancestral PAI emerged, our study also included plasmids not harboring the *pil* operon: (1) pSLT1, isolated from *Salmonella* and not involved in YAPIs and SPI-7 emergence (McClelland *et al*., 2001); (2) virulence-associated plasmids from *Yersinia pestis* (pCD) and *Y.enterocolitica* (pYV) and (3) *Y.pestis* plasmids encoding a murine toxin (pMT) and a plasminogen activator (pPCP). Unfortunately, no *Photorhabdus* plasmid sequences were available in databases for comparative purposes. Figure 5A shows that the dinucleotide frequencies of two R64-family plasmids (R64 and ColIb-P9) were

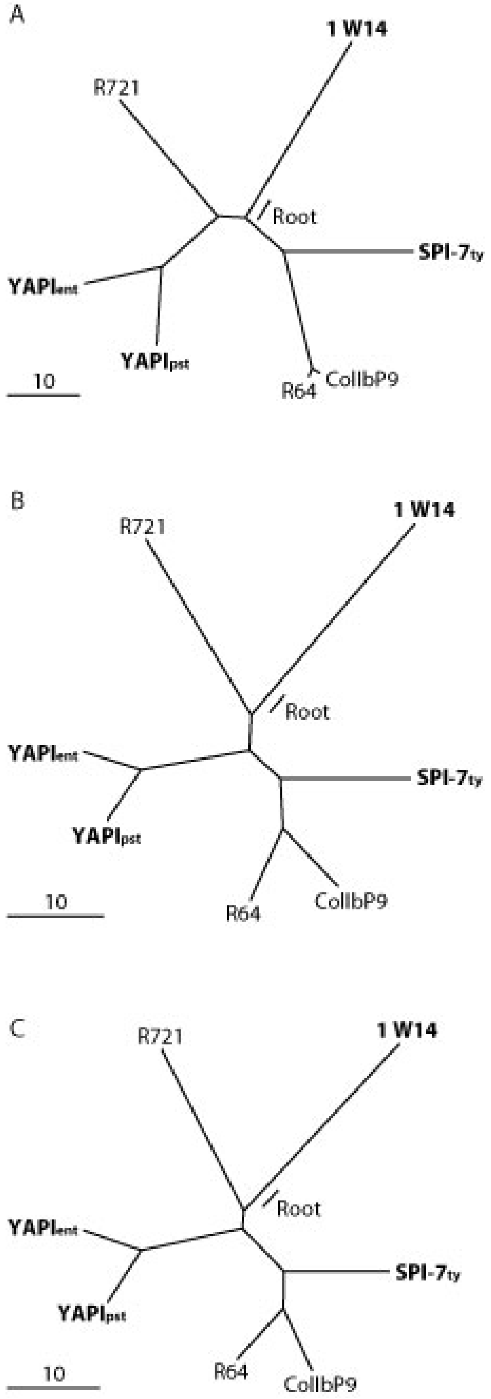

Fig. 4. Dinucleotide signature comparison of PAIs and *pil*-encoding plasmids. UPGMA dendrograms for comparisons of *pil* operons (**A**) and whole plasmid sequences associated with PAI ortholog subsets without (**B**) or with *pil* genes (**C**). PAIs are in bold type. Scale units and roots (if any) are described in Figure 3.

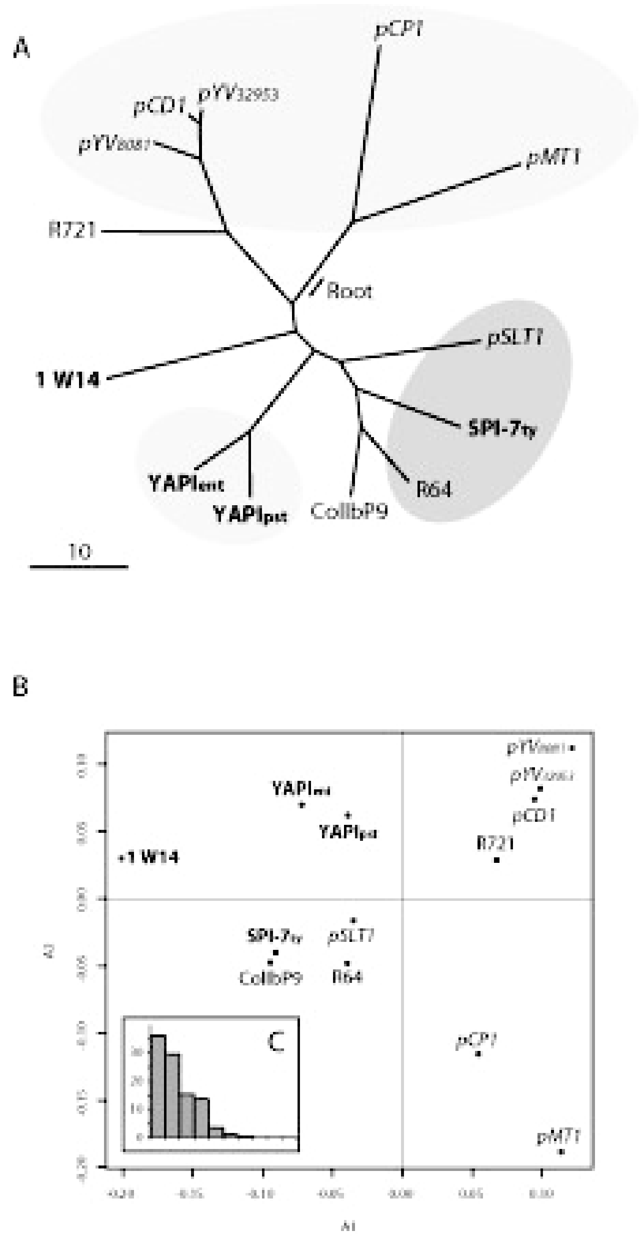

Fig. 5. Dinucleotide signature comparisons for SPI-7, YAPI, 1W14 and various plasmids from *Yersinia*, *Salmonella* and *Escherichia*. (**A**) UPGMA dendrogram of these comparisons, supported by a bootstrap analysis of NJ and Fitch dendrograms (Supplementary Fig. 6). (**B**) PCO analysis of the same comparisons and (**C**) its percentage variation expressed by the 10 major axes, sorted by decreasing order of magnitude along the horizontal axis. *Yersinia* and *Salmonella* GEs are shown on light gray and dark gray backgrounds, respectively. GEs not bearing a type IV pilus gene cluster are in italics and PAIs are in bold type. Scale units and roots (if any) are described in Figure 3. The distance matrix used for these representations is posted as Supplementary Fig. 5.

similar and resembled that of *Salmonella* plasmid pLST1, thus validating the specificity of the *Salmonella* plasmid signature. Since R721 houses *tra* and *pil* genes, it could theoretically be the source of the ancestral island formed by plasmid integration into a bacterial chromosome. Nevertheless, dinucleotide frequency comparisons clearly showed that the R721 plasmid did not contribute to the emergence of YAPIs, 1 W14 or SPI-7, since its dinucleotide signature diverged strongly from those of all PAIs. Our

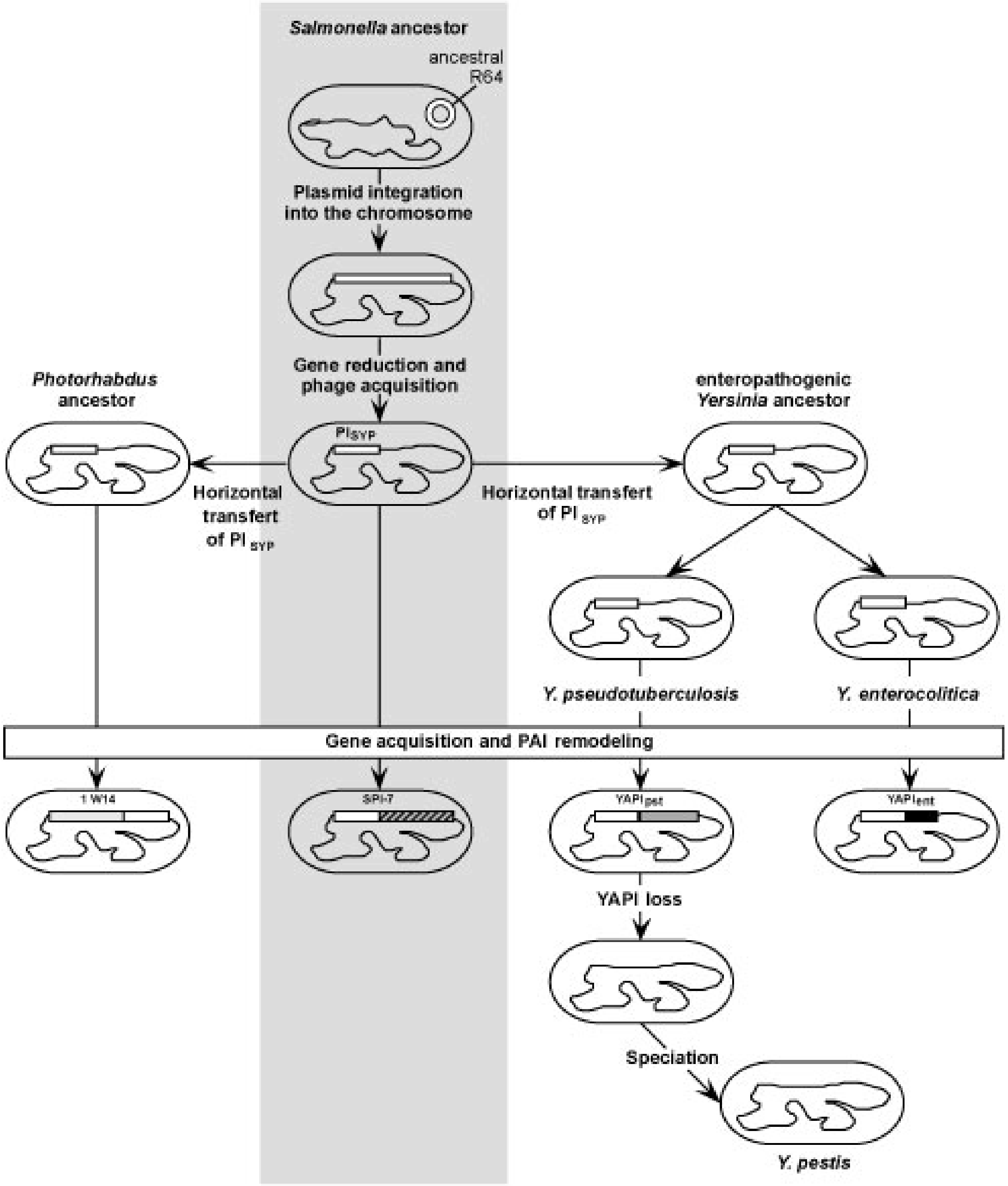
Salmonella ancestor
ancestral R64
Plasmid integration into the chromosome
Gene reduction and phage acquisition
PI SYP
Photorhabdus ancestor
Horizontal transfert of PI SYP
Horizontal transfert of PI SYP
enteropathogenic Yersinia ancestor
Y. pseudotuberculosis
Y. enterocolitica
Gene acquisition and PAI remodeling
1 W14
SPI-7
YAPI pst
YAPI ent
YAPI loss
Speciation
Y. pestis

Fig. 6. The most parsimonious model of YAPI formation. An ancestral R64 plasmid is incorporated into the ancestral *Salmonella* chromosome. Loss of R64 genes probably results from imperfect excision of the conjugative plasmid. The ancestral PAI (PI_{SYP}) was transmitted to *Yersinia* prior to *Yersinia enterocolitica* and *Y.pseudotuberculosis* speciation. In each bacterium, various DNA rearrangements (e.g. gene deletions and additions) result in a segment which is common to the four islands (ancestral PAI) and others which are specific to each bacterium. YAPI is absent from the genome of *Yersinia pestis* [the latter having recently emerged from *Y.pseudotuberculosis* (Achtman *et al*., 1999)]. Since YAPI can spontaneously excise from *Y.pseudotuberculosis* (Collyn *et al*., 2004a), *Y.pestis* derives from a YAPI-deleted clone. None of our data indicates an intermediary role of *Photorhabdus* 1 W14 in the ancestral PAI transmission from *Salmonella* to *Yersinia* or that YAPI was involved in the transfer from *Salmonella* to *Photorhabdus*. Indeed, since no PAIs or plasmids share more CDSs with 1 W14, we therefore prefer to consider a more parsimonious scenario based on the independent acquisition of an ancestral island from *Salmonella*.

conclusion is also supported by the lower degree of sequence conservation between R721-encoded Pil proteins and their counterparts on PAIs and the R64-related plasmids. In contrast, the genome signatures of the R64-plasmid family and SPI-7 were found to be very similar, indicating that a plasmid from the R64 family was the source of the *Salmonella* PAI. Since representations of phylogenetic trees introduce some bias in relative distances and imply that all sequences have a common ancestor, we also performed a PCO analysis (Fig. 5B) providing no dendrogram, but calculating distance estimations of GE relationships more accurately. The close clustering of SPI-7 and the R64-related and pSLT1 plasmids demonstrates that the SPI-7 module presents a *Salmonella*-plasmid signature. In conclusion, our dinucleotide analyses revealed that a R64-related plasmid was the source of an ancestral PAI in a *Salmonella* environment. These results consequently enabled us to orient an ancient PAI transfer for the first time: a pristine island referred as $\mathrm{PI}_{\mathrm{SYP}}$ (i.e. PAI module shared by *Salmonella*, *Yersinia* and *Photorhabdus*) first emerged in a *Salmonella* environment and then spread into the *Yersinia* and *Photorhabdus* genera.

Y.pseudotuberculosis and *Y.enterocolitica* acquired ancestral YAPI prior to speciation

The close relationship between $\mathrm{YAPI}_{\mathrm{ent}}$ and $\mathrm{YAPI}_{\mathrm{pst}}$ raises the question of how this PAI was acquired by both species and whether this DNA transfer occurred before or after enteropathogenic *Yersinia* speciation. In all the above analyses, the distances separating either $\mathrm{YAPI}_{\mathrm{pst}}$ or $\mathrm{YAPI}_{\mathrm{ent}}$ from the other GEs were found to be almost identical (Figs 2–5 and Supplementary Fig. 1). This observation supports a parsimonious scenario proposing a contemporary acquisition of $\mathrm{PI}_{\mathrm{SYP}}$ predating the enteropathogenic Yersiniae speciation. Furthermore, the similar degree of identity between (1) $\mathrm{YAPI}_{\mathrm{ent}}$ and $\mathrm{YAPI}_{\mathrm{pst}}$ gene products (64–97% identity; average 82%) (Collyn *et al*., 2004b) and (2) homologous proteins encoded by the chromosomal cores of both *Yersinia* species (70–97% identity; average 86%) (Supplementary Table 2) reinforce our evolutionary scenario.

DISCUSSION

This study is the first to measure evolutionary relationships between PAIs and/or PAI modules using comparative PAI genometrics, i.e. the simultaneous application to mobile GEs of tools developed for prokaryotic chromosome comparisons at nucleotide or gene levels. Since large PAIs usually result from successive additions of heterologous DNA modules, the identification of a common gene set in PAIs was essential for evaluating this overall strategy. We showed that dinucleotide signatures can be used as phylogenetic tools for PAI modules containing genes from the same origin, when applied to sequences not smaller than 30 kb. Consequently, tools have to be developed for highlighting segmentation in PAI modules that lack homology to identified sequences. Moreover, mobile GEs responsible for PAI emergence (including the *Salmonella* and *Yersinia* plasmids used here) can display distinct genome signatures when compared with their recipient chromosome. The latter GEs are not directly relevant to the understanding of ancestral PAI emergence, since their horizontal transfer can be accurately oriented by comparative genometrics of PAIs and mobile GEs.

Comparative PAI genometrics enabled us to orient $\mathrm{PI}_{\mathrm{SYP}}$ horizontal transfers. Since conjugative plasmids containing type IV pilus genes have not been isolated from *Yersinia* or *Photorhabdus*, $\mathrm{PI}_{\mathrm{SYP}}$ would have most probably emerged in bacteria such as *Salmonella* harboring conjugative plasmids and PAIs. This proposal is supported by dinucleotide frequency analyses: the YAPI- and *Yersinia* plasmids signatures diverge significantly, confirming the various origins of these GEs. Moreover, YAPI signatures are more closely related to those of *Salmonella* plasmids than those of *Yersinia* plasmids. All these data enable us to propose a parsimonious evolutionary scenario for YAPI emergence (Fig. 6). Since YAPI acquisition by Yersiniae is probably an ancient transfer event, this PAI would have progressively adopted the codon and nucleotide usage of the host (Hacker and Kaper, 2000), explaining the divergence of the YAPI and SPI-7 dinucleotide signatures. This nucleotide adaptation is a strong argument in favor of genome signatures as phylogenetic tools, when they are challenged using several representations of distance matrix comparisons. However, our scenario does not provide a time frame for the ancestral R64 plasmid integration into the *Salmonella* chromosome.

In conclusion, this contribution reveals how comparative genometrics enables characterization of GEs responsible for the formation of PAIs or related modules (prophages, plasmids, transposons, etc.). Our multiscale, comparative approach opens up new horizons in the understanding of microbial genome evolution due to horizontal transfer in general and the emergence of pathogenic bacterial species by virulence gene transfer in particular.

ACKNOWLEDGEMENTS

We warmly thank Philippe Marti and Ruedi Füllemann for providing DinucOne (preliminary dinucleotide analyses) and Philippe Moreillon, Christophe Carnoy and Jean-Claude Sirard for stimulating discussions. F.C. received a doctoral studentship from the French Ministère de l’Enseignement Supérieur, de la Recherche et de la Technologie and the Fondation pour la Recherche Médicale.

Conflict of Interest: none declared.

REFERENCES

Achtman,M. *et al*. (1999) *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proc. Natl Acad. Sci. USA*, **96**, 14043–14048.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Blanchette,M. *et al*. (1999) Gene order breakpoint evidence in animal mitochondrial phylogeny. *J. Mol. Evol*., **49**, 193–203.

Campbell,A. *et al*. (1999) Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc. Natl Acad. Sci. USA*, **96**, 9184–9189.

Collyn,F. *et al*. (2002) *Yersinia pseudotuberculosis* harbors a type IV pilus gene cluster that contributes to pathogenicity. *Infect Immun*., **70**, 6196–6205.

Collyn,F. *et al*. (2004a) YAPI, a new *Yersinia pseudotuberculosis* pathogenicity island. *Infect. Immun*., **72**, 4784–4790.

Collyn,F. *et al*. (2004b) YAPI, a new pathogenicity island in Enteropathogenic Yersiniae. In Carniel,E. and Hinnebusch,B.J. (eds), *Yersinia Molecular and Cellular Biology*. Horizon Biosciences, Wynmondham, Norfolk, UK, pp. 307–317.

Dobrindt,U. and Hacker,J. (2001) Whole genome plasticity in pathogenic bacteria. *Curr. Opin. Microbiol*., **4**, 550–557.

Duchaud,E. *et al*. (2003) The genome sequence of the entomopathogenic bacterium *Photorhabdus luminescens*. *Nat. Biotechnol*., **21**, 1307–1313.

Felsenstein,J. (2004) PHYLIP (Phylogeny Inference Package). (2004), Distributed by the author, Department of Genetics, University of Washington, Seattle.

Fitz-Gibbon,S.T. and House,C.H. (1999) Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res*., **27**, 4218–4222.

Gower,J.C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325–338.

Greub,G. *et al*. (2004) A genomic island present along the bacterial chromosome of the *Parachlamydiaceae* UWE25, an obligate amoebal endosymbiont, encodes a potentially functional F-like conjugative DNA transfer system. *BMC Microbiol*., **4**, 48.

Hacker,J. and Kaper,J.B. (2000) Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbiol*., **54**, 641–679.

Herniou,E.A. *et al*. (2003) The genome sequence and evolution of baculoviruses. *Annu. Rev. Entomol*., **48**, 211–234.

Ihaka,R. and Gentleman,R. (1996) R: a language for data analysis and graphics. *Comp. Graphical Stat.*, **5**, 299–314.

Karlin,S. (2001) Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol.*, **9**, 335–343.

Karlin,S. and Cardon,L.R. (1994) Computational DNA sequence analysis. *Annu. Rev. Microbiol.*, **48**, 619–654.

Karlin,S. *et al.* (1994) Molecular evolution of herpesviruses: genomic and protein sequence comparisons. *J. Virol.*, **68**, 1886–1902.

Karlin,S. *et al.* (1998) Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.*, **32**, 185–225.

Kim,S.R. and Komano,T. (1992) Nucleotide sequence of the R721 shufflon. *J. Bacteriol.*, **174**, 7053–7058.

Ling,L. *et al.* (2002) Proteome-wide analysis of protein function composition reveals the clustering and phylogenetic properties of organisms. *Mol. Phylogenet. Evol.*, **25**, 101–111.

McClelland,M. *et al.* (2001) Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. *Nature*, **413**, 852–856.

McGeoch,D.J. *et al.* (2000) Toward a comprehensive phylogeny for mammalian and avian herpesviruses. *J. Virol.*, **74**, 10401–10406.

Montague,M.G. and Hutchison,C.A.,III (2000) Gene content phylogeny of herpesviruses. *Proc. Natl Acad. Sci. USA*, **97**, 5334–5339.

Ochman,H. *et al.* (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.

Page,R.D. (1996) TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.*, **12**, 357–358.

Pickard,D. *et al.* (2003) Composition, acquisition, and distribution of the Vi exopolysaccharide-encoding *Salmonella enterica* pathogenicity island SPI-7. *J. Bacteriol.*, **185**, 5055–5065.

Sharp,P.M. *et al.* (2005) Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.*, **33**, 1141–1153.

Snel,B. *et al.* (1999) Genome phylogeny based on gene content. *Nat. Genet.*, **21**, 108–110.

Snel,B. *et al.* (2002) The identification of functional modules from the genomic association of genes. *Proc. Natl Acad. Sci. USA*, **99**, 5890–5895.

Tekaia,F. *et al.* (1999) The genomic tree as revealed from whole proteome comparisons. *Genome Res.*, **9**, 550–557.

Thompson,J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Wolf,Y.I. *et al.* (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.*, **1**, 8.

Wolfe,K.H. *et al.* (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Natl Acad. Sci. USA*, **84**, 9054–9058.