Quentin *et al. BMC Genomics* (2018) 19:475
https://doi.org/10.1186/s12864-018-4836-1

BMC Genomics

RESEARCH ARTICLE Open Access

Single-strand DNA processing: phylogenomics and sequence diversity of a superfamily of potential prokaryotic HuH endonucleases

Yves Quentin*, Patricia Siguier, Mick Chandler* and Gwennaele Fichant

Abstract

Background: Some mobile genetic elements target the lagging strand template during DNA replication. Bacterial examples are insertion sequences IS*608* and IS*Dra2* (IS*200*/IS*605* family members). They use obligatory single-stranded circular DNA intermediates for excision and insertion and encode a transposase, $\mathrm{TnpA}_{\mathrm{IS}\mathit{200}}$, which recognizes subterminal secondary structures at the insertion sequence ends. Similar secondary structures, Repeated Extragenic Palindromes (REP), are present in many bacterial genomes. $\mathrm{TnpA}_{\mathrm{IS}\mathit{200}}$-related proteins, $\mathrm{TnpA}_{\mathrm{REP}}$, have been identified and could be responsible for REP sequence proliferation. These proteins share a conserved HuH/Tyrosine core domain responsible for catalysis and are involved in processes of ssDNA cleavage and ligation. Our goal is to characterize the diversity of these proteins collectively referred as the $\mathrm{TnpA}_{\mathrm{Y1}}$ family.

Results: A genome-wide analysis of sequences similar to $\mathrm{TnpA}_{\mathrm{IS}\mathit{200}}$ and $\mathrm{TnpA}_{\mathrm{REP}}$ in prokaryotes revealed a large number of family members with a wide taxonomic distribution. These can be arranged into three distinct classes and 12 subclasses based on sequence similarity. One subclass includes sequences similar to $\mathrm{TnpA}_{\mathrm{IS}\mathit{200}}$. Proteins from other subclasses are not associated with typical insertion sequence features. These are characterized by specific additional domains possibly involved in protein/DNA or protein/protein interactions. Their genes are found in more than 25% of species analyzed. They exhibit a patchy taxonomic distribution consistent with dissemination by horizontal gene transfers followed by loss. The $\mathit{tnpA}_{\mathrm{REP}}$ genes of five subclasses are flanked by typical REP sequences in a REPtron-like arrangement. Four distinct REP types were characterized with a subclass specific distribution. Other subclasses are not associated with REP sequences but have a large conserved domain located in C-terminal end of their sequence. This unexpected diversity suggests that, while most likely involved in processing single-strand DNA, proteins from different subfamilies may play a number of different roles.

Conclusions: We established a detailed classification of $\mathrm{TnpA}_{\mathrm{Y1}}$ proteins, consolidated by the analysis of the conserved core domains and the characterization of additional domains. The data obtained illustrate the unexpected diversity of the $\mathrm{TnpA}_{\mathrm{Y1}}$ family and provide a strong framework for future evolutionary and functional studies. By their potential function in ssDNA editing, they may confer adaptive responses to host cell physiology and metabolism.

Keywords: HuH superfamily, Transposition, REP sequences, Replication fork, Insertion sequence, RAYT/$\mathrm{TnpA}_{\mathrm{REP}}$

* Correspondence: Yves.Quentin@ibcg.biotoul.fr; Mike.Chandler@ibcg.biotoul.fr
Laboratoire de Microbiologie et Génétique Moléculaires, UMR5100, Centre de Biologie Intégrative (CBI), Centre National de la Recherche Scientifique (CNRS), Université de Toulouse, UPS, F-31062 Toulouse, France

Background

HuH enzymes are dedicated to processing single-strand DNA (ssDNA) and use particular DNA recognition and reaction mechanisms for site-specific ssDNA cleavage and ligation. Members of this protein family are numerous and widespread in all three domains of life. There are two major classes within the HuH superfamily [1]: the Rep (replication) proteins and the relaxase or Mob (mobilization) proteins which process DNA during plasmid replication and conjugation, respectively. However, HuH endonucleases have also been identified in other processes involving ssDNA, such as replication of certain phages [2] and eukaryotic viruses [3], and in different types of transposon. These proteins have also been appropriated for cellular processes such as intron homing [4] and processing of bacterial Repetitive Extragenic Palindromic (REP) sequences [5]. The family relationship is based on several conserved amino acid motifs. These include the HuH motif [6, 7], composed of two His residues (H) separated by a bulky hydrophobic residue (u), and the Y motif, containing either one or two Tyr (Y) residues separated by several amino acids. Together, these constitute the core catalytic domain.

Many family members include additional functional domains such as helicases, primases or zinc fingers. The simplest examples are transposases of the IS*200*/IS*605* insertion sequence (IS) family (TnpA$_{IS\mathit{200}}$) [8]. They are approximately 150 amino acids long, possess only the HuH-Y core domain of about 113 amino acids (Fig. 1) and function as dimers [9, 10] (Fig. 1a). The two H residues provide two of the three ligands involved in coordinating an essential Mg^{2+} ion. The third amino acid involved in this is located approximately four residues downstream from the catalytic Y residue (Q131 in the *Helicobacter pylori* IS*608* transposase and N119 in TnpA$_{REP}$ from *Escherichia coli* MG1655, Fig. 1d). The Y residues act as nucleophiles in the cleavage reactions generating covalent phosphotyrosine intermediates. They use ssDNA IS substrates and catalyze IS insertion into and excision from the lagging strand template at replication forks [11] via a circular ssDNA intermediate. The HuH transposases, TnpA$_{IS\mathit{608}}$ and TnpA$_{IS\mathit{Dra2}}$ (from IS*608* and IS*Dra2* respectively) have been extensively studied at the functional level using a combination of genetics, biochemistry and structural biology [9, 10, 12–21] and the transposition pathway has been described in detail [11]. These enzymes do not directly recognize the DNA sequences which they cleave at each IS end during transposition. Instead, they bind small subterminal DNA hairpin structures of about 20–25 nt at both the left (LE) and right (RE) ends. The cleavage position is determined by a complex series of base interactions between

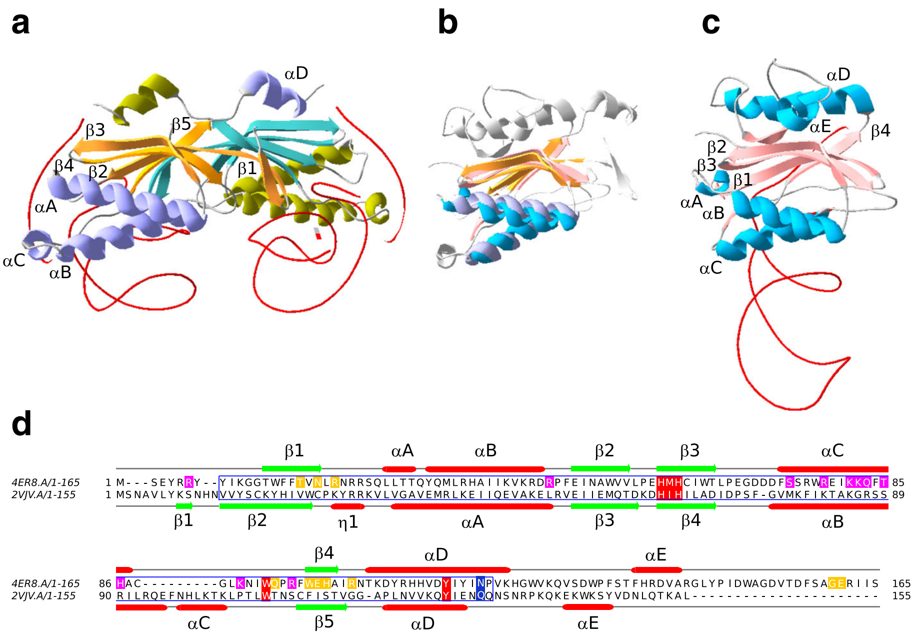
a
αD
β3
β5
β4
β2
β1
αA
αC
αB
b
c
αD
αE
β4
β2
β3
β1
αA
αB
αC
d
β1
αA
αB
β2
β3
αC
4ER8.A/1-165
1 M---SEYRRY--YIKGGTWFFTVNLRNRRSQLLTTQYQMLRHAIIKVKRDRPFEINAWVVLPEHMHCIWTLPEGDDDFSSRWREIKKQFT 85
2VJV.A/1-155
1 MSNAVLYKSNHNVVYSCKYHIVWCPKYRRKVLVGAVEMRLKEIIQEVAKELRVEIIEMQTDKDHIHILADIDPSF-GVMKFIKTAKGRSS 89
β1
β2
η1
αA
β3
β4
αB
β4
αD
αE
4ER8.A/1-165
86 HAC---------GLKNIWQPRFWEHAIRNTKDYRHHVDYIYINPVKHGWVKQVSDWPFSTFHRDVARGLYPIDWAGDVTDFSAGERIIS 165
2VJV.A/1-155
90 RILRQEFNHLKTKLPTLWTNSCFISTVGG-APLNVVKQYIENQQNSNRPKQKEKWKSYVDNLQTKAL---------------------- 155
αC
β5
αD
αE

Fig. 1 Structural homology between TnpA$_{IS\mathit{200}}$ and TnpA$_{REP}$. Panel **a** TnpA$_{IS\mathit{200}}$ from *H. pylori* TnpA$_{IS\mathit{608}}$ dimer (PDB 2VJV) as a ribbon diagram with one molecule colored cyan (β sheets) and purple (α helix) the other molecule in orange (β sheets) and green (α helix), with the DNA in red. Panel **c** *E. coli* TnpA$_{REP}$ monomer (PDB 4ER8). Panel **b** TnpA$_{IS\mathit{608}}$ and TnpA$_{REP}$ monomer structure superposition, with un-conserved structural elements in grey. The superpositioning was accomplished using SWISS-MODEL and the visualization with Swiss-PdbViewer [86] (https://spdbv.vital-it.ch/). Panel **d** TnpA$_{REP}$ (4ER8) and TnpA$_{IS\mathit{608}}$ (2VJV) sequence alignment with the secondary structures. HuH, W and Y conserved residues in red boxes, the residues essential for transposition in vivo (Q131 in *H. pylori* and N119 in TnpA$_{REP}$ from *E. coli* MG1655) in blue, residues involved in 5′ GTAG interaction in yellow and residues REP sequence hairpin interaction in magenta, the core conserved domain in blue frame

a tetranucleotide (guide sequence) just 5′ to the hairpin foot and a conserved target tetranucleotide which flanks LE. At the right end, similar types of interaction occur between the final tetranucleotide of RE and an equivalent tetranucleotide guide sequence. The interactions can involve bases which form canonical (Watson and Crick) and non-canonical interactions including base triples [10, 14].

Closely related members of this group are the REP-associated tyrosine transposases, RAYTs [22] also referred as $\mathrm{TnpA_{REP}}$ to highlight their functional proximity with $\mathrm{TnpA}_{\mathrm{IS}\mathit{200}}$ [5, 22]. These also include the HuH-Y functional core [23]. In contrast to $\mathrm{TnpA}_{\mathrm{IS}\mathit{200}}$ (Fig. 1a), $\mathrm{TnpA_{REP}}$ from *E. coli* appears as a monomer (Fig. 1c). The proteins share significant structural similarities (Fig. 1b).

REP sequences provide genomic binding sites for proteins such as Integration Host Factor, DNA polymerase I, and DNA gyrase [24–26], can increase mRNA stability [27] and cause transcription termination [28, 29]. They have also been implicated in regulating translation [30]. In many ways REP sequences resemble the subterminal secondary structures at the ends of IS*200*/IS*605* family members and are of similar size (21–65 nt). Like the IS LE and RE, they also carry a tetranucleotide guide sequence 5′ to the foot of the REP sequence hairpin necessary for cleavage by $\mathrm{TnpA_{REP}}$ [5]. As their name suggests, they are found largely between genes and are often present in many copies in a given host genome (up to 1% of *E. coli* chromosomes). They can be present as isolated copies or in clusters and often form particular structures, BIMES (bacterial interspersed mosaic elements [31]), whose basic unit includes two REP sequence copies in an inverted orientation. These were later renamed REPINs (REP doublets forming hairpINs; [32, 33]). There are a number of different REP sequence families or groups and several different groups can sometimes be found within a single genome [22, 34–37]. In general, there appears to be only a single full length $\mathit{tnpA}_{\mathrm{REP}}$ gene for each REP sequence type and this is often flanked by a number of REP sequences forming a structure called a REPtron [5]. REPtrons do not appear to transpose as a unit but the REPs/BIMEs are likely to be mobilized by $\mathrm{TnpA_{REP}}$ activity [5, 22, 32, 33]. Indeed $\mathrm{TnpA_{REP}}$ is able to cleave and rejoin REP sequences [5].

The apparent stability of $\mathit{tnpA}_{\mathrm{REP}}$ genes raises the question of how and why they are maintained in their host bacterial genomes. This was recently addressed through an analysis of $\mathit{tnpA}_{\mathrm{REP}}$ properties and a comparison to transposases of IS families (IS*200*/IS*605* and the entirely unrelated IS*110*) and two housekeeping genes [38]. The observations that $\mathit{tnpA}_{\mathrm{REP}}$ share more characteristics with housekeeping genes than with insertion sequences and that $\mathit{tnpA}_{\mathrm{REP}}$ are predominantly vertically transmitted, at the subgenus level, suggest that they have a yet uncharacterized function(s) that benefit their host cell and hence insure their maintenance [38].

Our goal in the work presented here was to characterize the diversity of this large protein family referred to collectively as the $\mathrm{TnpA_{Y1}}$ family, since all members share the conserved HuH/Y core domain responsible for ssDNA cleavage and ligation [11]. We provide a detailed classification of family members which was consolidated by the analysis of the conserved core domains and the characterization of additional domains. Only one subclass has the characteristic of insertion sequence transposases while five subclasses are more closely related to $\mathrm{TnpA_{REP}}$ proteins and their genes are flanked by typical REP sequences in a REPtron-like arrangement. Other subclasses are not associated with REP sequences and illustrate the unexpected diversity of proteins belonging to the $\mathrm{TnpA_{Y1}}$ family. The data obtained provide a strong framework for future evolutionary study and the characteristics describing each subclass may help to further experimentally unravel their functions which are still largely unknown.

Results

Overall class and subclass organization of the $\mathrm{TnpA_{Y1}}$ family

Our initial $\mathrm{TnpA_{Y1}}$ library was composed of 924 proteins (Methods). To explore the family organization, the graph of the large connected component used in constructing the library (Methods) was reprocessed with MCL. The inflate factor (IF) value is an important parameter of MCL as it regulates the cluster granularity. We tested different IF values and analyzed their impact on graph partitioning. An IF of 1.2 generated three distinct classes (Fig. 2a-b). Class 1 includes the $\mathrm{TnpA}_{\mathrm{IS}\mathit{200}}$; class 2, the $\mathrm{TnpA_{REP;}}$ while the smaller class 3 contains a previously unknown family of HuH-Y proteins. Increasing IF values increased the number of sub-groups (Fig. 2a) from five (IF values of 1.4) to thirteen (IF values > 3.8). This increase resulted from hierarchical decomposition of class 1 and class 2 into subclasses. Class 3 remained coherent and did not decompose into subclasses. We chose an intermediate IF value of 3.0 since the number of partitions remained stable at 12 subclasses over the range 3.0 ≤ IF ≤3.6. Here, class 1 forms two subclasses, 1.1 and 1.2, class 2 forms nine subclasses, 2.1 to 2.9, (Fig. 2c), while class 3 remains unified. This classification into 12 subclasses was subsequently propagated to all proteins of the sample to obtain a complete overview (i.e. sequences of each cluster received the subclass of its medoid; Methods). 4043 of the 4081 initial proteins fell into the 12 $\mathrm{TnpA_{Y1}}$ subclasses. The remaining 38 proteins were false positives and were discarded.

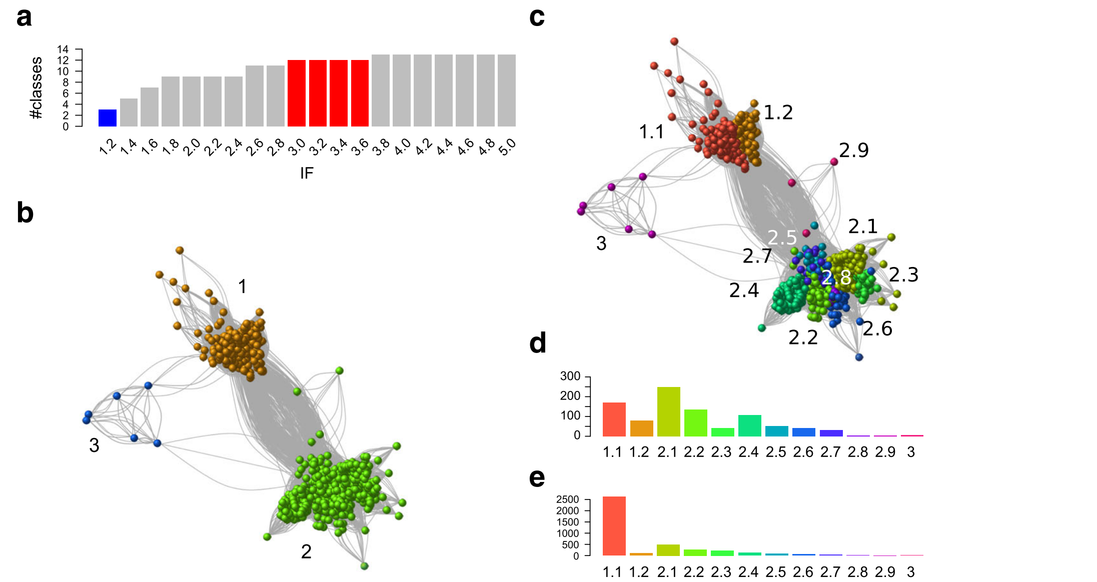

Fig. 2 Family and sub-family organization of the 931 TnpA$_{Y1}$ protein candidates explored with MCL. **a** Evolution of the number of classes as a function of IF value. Three classes (family level) are obtained with IF = 1.2 (blue bar) and rise to reach a maximum of 14 classes. Twelve classes (subclass level) were obtained with IF in the range [3.0, 3.6] (red bar). **b** Graph partitioning obtained with IF = 1.2. Edges are BlastP hits (e-value <1e-5) between proteins (vertices). The vertices are colored according to MCL classes. **c** Graph partitioning obtained with IF = 3.0. **d** Histograms of the subclass sizes obtained with IF = 3.0. **e** Histograms of the subclass sizes obtained after propagation of the annotation to the complete sample of protein candidates. A maximum of 2620 proteins is observed in subclass 1.1. The color code is the same in **c**, **d** and **e** panels

The distribution of sequences into subclasses was uneven both for the initial (Fig. 2d) and the full protein sets (Fig. 2e). In both, there is a preponderance of subclass 1.1 (TnpA$_{IS\mathit{200}}$) (171 in the restricted sample and 2620 in the complete set due to multiple genomic copies of IS*200*/IS*605* family members). The contribution of the second most abundant class, 2.1, decreased in the full protein set.

An additional filter was included based on protein length within each subclass. The median protein length varies from 148 to 325 AA between different subclasses (see Additional file 1: Figure S1). Subclass 2.9 included only three proteins and was eliminated from further analysis. A few unexpectedly long proteins, the result of gene fusions or annotation errors, observed in each subclass and a number of significantly shorter sequences (partial) observed in almost all subclasses were also discarded (see Additional file 1: Figure S1). Partial sequences were significantly more abundant in subclass 1.1 (TnpA$_{IS\mathit{200}}$) which show a tendency to decay [11].

Subclass characterization

Conserved core domain

Class 2 members, which include TnpA$_{REP}$ proteins, were initially retrieved with PF01797 (121 amino acids long, http://pfam.xfam.org/family/Y1_Tnp) which was built uniquely for the IS*200*/IS*605* family. However, the profile of most class 2 subfamilies aligns only partially with TnpA$_{REP}$ sequences (see Additional file 1: Figure S2A) and in some cases, two alignments are obtained for the N- and C-term regions respectively. This is also true if the length of the aligned part of the protein with the profile is analyzed (see Additional file 1: Figure S2B).

For better core domain coverage, we constructed a specific HMMER TnpA$_{REP}$-like family profile (Methods). The multiple alignments were manually edited to remove highly divergent sequences and to extract the conserved core domain as defined in PF01797. The TnpA$_{REP}$-like profile obtained with hmmbuild from 428 aligned sequences is 115 amino acids long. This new profile allows excellent coverage of class 2 sequences: a complete alignment of the profile with the protein sequences is observed for all the subfamilies (see Additional file 1: Figure S2C) despite the sequence length variability of class 2 proteins (see Additional file 1: Figure S2D).

Class 3 includes only 10 members, 5 of which do not have conserved HuH or Y motifs. To augment this number, we searched Genbank for other instances of this class using BlastP. A sample of 24 sequences was retrieved and used to construct a specific class 3 profile with HMMER. Compared to class 1 and 2, class 3 proteins have an extended sequence conservation in the N-term region leading to a profile length of 155 AA.

HuH and Y motifs

An *ab initio* search for conserved motifs with MEME confirmed the conservation of subclass-specific motifs overlapping the ubiquitous HuH, Y together with a W residue (Fig. 1d and see Additional file 1: Figure S3). The distance between HuH and Y motifs is relatively constant (median between 58 and 67 AA) in the majority of subclasses, but is larger in subclasses 2.3, 2.4 and 2.6 (median 110, 88 and 82 amino acids) while subclass 2.7 and class 3 have a more disperse distribution (see Additional file 1: Figure S4). MEME also revealed additional short conserved motifs upstream (subclasses 2.4, 2.5, 2.7 and 2.8) and/or downstream (subclasses 2.1, 2.2, 2.3, 2.6 and 2.8) of the core region (see Additional file 1: Figure S3). The downstream motifs share conserved serine residues and are part of a larger conserved domain (see below). The absence of conserved core motifs in certain sequences suggests that these correspond to non-functional proteins. To retain proteins that are most likely to be functionally active, we selected those that included a complete core domain defined by PF01797 or $\mathrm{TnpA_{REP}}$-like profiles encoding both the HuH and Y motifs. We obtained a sample of 722 from the original 924 sequences: 172 class 1, 545 class 2 and 5 class 3. In spite of the variations from class to class, the conservation of key catalytic HuH and Y residues suggest that these proteins are all involved in cleavage and rejoining ssDNA [1].

Phylogenetic relationships between MCL classes

To gain insight into the relationship between these proteins, multiple alignments of the core domains of each class were first computed to take into account the variability in sequence length observed between MCL classes. These alignments were themselves aligned to produce a multiple alignment of the entire sequence set (Methods). This was trimmed to remove sites with rare indels and four unexpectedly divergent sequences. The final edited alignment contained 718 sequences with 121 sites. A phylogenetic tree (Methods) was computed and rooted on the branch that separates class 1 and 2 (Fig. 3). We observed a strong correspondence between MCL subclasses and the tree topology although the former used complete sequences while the latter was based only on the core domain. Most MCL subclasses formed a coherent subtree with a single root and strong bootstrap support. However, on the basis of this tree, subclass 2.1 can be clearly subdivided into two subclasses as it included two well supported subtrees (2.1.1 and 2.1.2 in Fig. 3). There are several exceptions: subclass 2.3 is embedded in the larger subclass 2.1, suggesting that the 2.1 outgroup is misclassified (grey arc d); and the well-supported subclass 2.2 is framed by divergent sequences originating from the center of the tree (black arcs a and b). In addition, in the same region, a small number of sequences among the deepest branches have an MCL classification in disagreement with their topological location on the tree (black arc c).

Taxonomic distribution of the subclasses

To determine the taxonomic distribution of the different subfamilies, the sequences of the tree were annotated according to the phylum to which their species belongs by following NCBI taxonomy (Fig. 3, outer ring). The subclasses display very different phylum diversity. Subclass 1.1 ($\mathrm{TnpA}_{\mathrm{IS}\mathit{200}}$) exhibits high taxonomic diversity. However the phyla are fragmented into several patches. This patchwork distribution is expected for mobile elements such as IS*200*/IS*605* which undergo horizontal transfer between distantly related phyla. Subclasses 2.1.2, 2.4 and 2.7 also exhibit a similar trend. In contrast, subclasses 2.1.1, 2.3, 2.5, 2.6 and 2.8 are dominated by a single phylum implying a phylum-specific $tnpA_{\mathrm{Y1}}$ gene expansion. Indeed, sequences from the subclasses 2.1.1, 2.3 and 2.5 are mostly found in Proteobacteria, sequences from the subclass 2.6 in Bacteroidetes and sequences from the subclass 2.8 in Cyanobacteria.

The tree was computed on a subset of our initial sample composed of potentially functional proteins with a complete catalytic core domain. To obtain a global view of subclass distribution across bacterial phyla, we compiled the results from the original sample, retaining only one strain per species. Among the 1354 species, 57.2% do not encode a $tnpA_{\mathrm{Y1}}$ gene, 25.6% encode at least one member of class 1 (with 21.5% at least one member of subclass 1.1), 25.4% encode at least one member of class 2 and 0.3% at least a member of class 3. Multiple different class 2 genes can co-occur in several species. Of these, we have distinguished up to four genes from different class 2 subclasses in 12 of the species in the library. Figure 4 shows the presence or absence of subclass members in each strain. Subclass 1.1 members are well distributed throughout, with a few strongly represented phyla (e.g. Cyanobacteria, Clostridia...). Members of the other subclasses show sparse distribution but with a tendency to co-occur in the same genera genome set (Cyanobacteria, Bacteroidetes, Epsilon-, Delta-, Beta- and Gamma-proteobacteria). Subclass 2.8 appears restricted to Cyanobacteria. The assembled Tenericute genomes, represented by 35 Mollicute species in our sample, do not exhibit $tnpA_{\mathrm{Y1}}$ homologs (although a number of IS*200*/IS*605* can be found in unassembled genomes). Class 3, enlarged to 24 proteins, is principally present in Planctomycetes (11), Acidobacteria (6) and Proteobacteria (5).

In bacteria, genome size variation is associated with the gain and loss of accessory genes [39, 40]. To test whether $tnpA_{\mathrm{Y1}}$ gene distribution follows a similar trend, we compared genomes coding for at least one of these genes with those containing none. We observed that the mean genome size is significantly larger for the former than for the latter (p value $< 10^{-16}$; see Additional file 1: Figure S5A). This effect is stronger when the $\mathrm{TnpA}_{\mathrm{IS}\mathit{200}}$

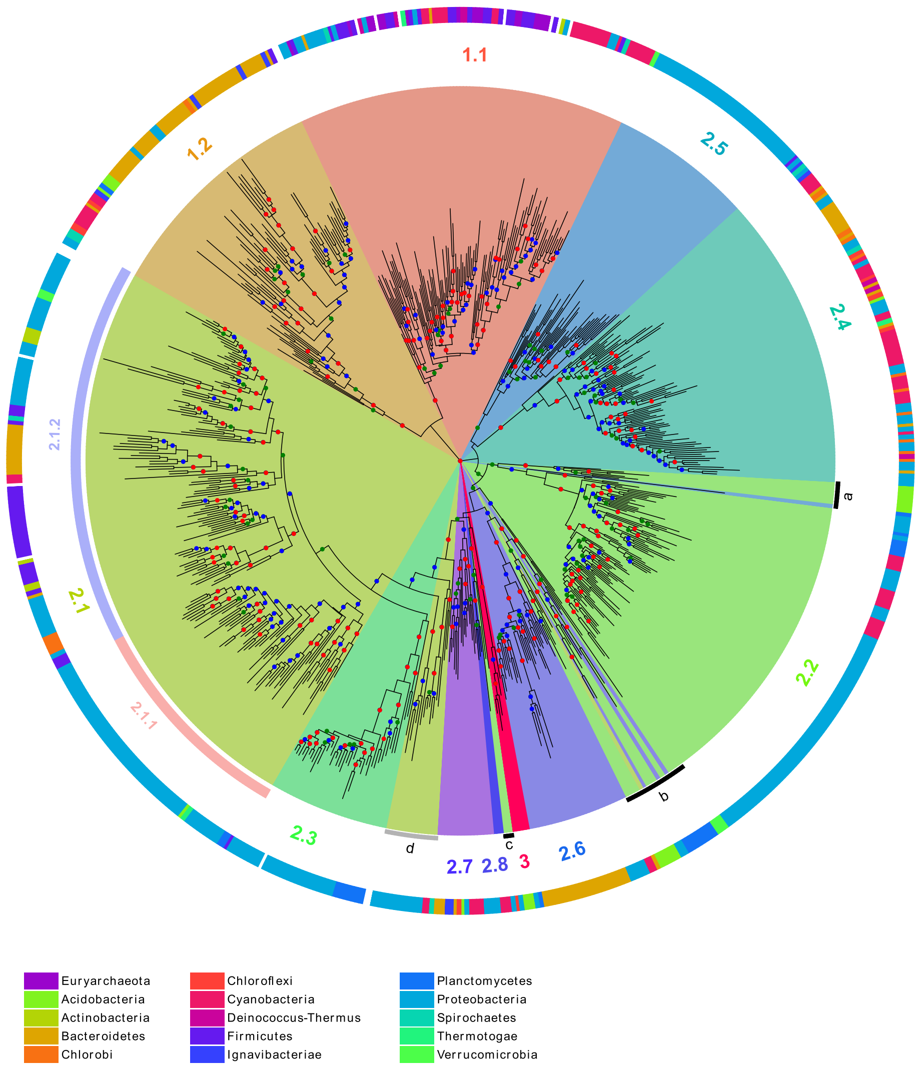

Fig. 3 Tree of the TnpA_{Y1} representative sequences. The tree is rooted to the branch that separates class 1 and 2. Color code for the bootstrap value (*b*) intervals: $0.7 < b \leq 0.8$ in green, $0.8 < b \leq 0.9$ in blue and $0.9 < b \leq 1$ in red. The MCL subclasses obtained with IF = 3.0 are colored and subclass numbers are added in the corresponding segments. Based on the tree topology, subclass 2.1 can be further divided in groups 2.1.1 and 2.1.2 and three subsets of sequences appear misclassified by MCL (highlighted by black and grey segments) according to their tree location. The taxonomic origin of the sequences is shown as a colored code in the outer ring (NCBI taxonomy phyla level). The phyla with less than three sequences on the tree are not colored. The Phyla are color-coded at the bottom of the figure. The tree display was obtained with online iTOL [83] (https://itol.embl.de/)

of subclass 1.1 are removed (see Additional file 1: Figure S5B and S5C). Strains with genome sizes greater than 5 Mb have a very high probability of encoding at least one *tnpA*$_{\text{Y1}}$ gene. Notable exceptions in our sample are the high G + C Gram-positive Actinobacteridae, that have among the largest genomes but carry only members of subclass 1.1, and Alpha-proteobacteria that include both species with small (such as the Rickettsiales) and large genomes (such as Rhizobiales) but exhibit none or only a few *tnpA*$_{\text{Y1}}$ homologs per genome regardless of the genome size. This general preponderance of *tnpA*$_{\text{Y1}}$ (excluding the IS-associated proteins) genes in larger genomes suggests that these genes might contribute to the adaptation of their host to new selective pressures and/or to diverse ecological conditions.

Subclass domain organization

To better define the organization of protein domains, a multiple-alignment of the core domain was performed (Fig. 5) using the sequence order from the maximum likelihood tree (Fig. 3) and trimmed for sites with rare insertions. This figure shows that core domain length and the distance between the HuH and Y motifs (Fig. 5 top) is quite different between families but is conserved

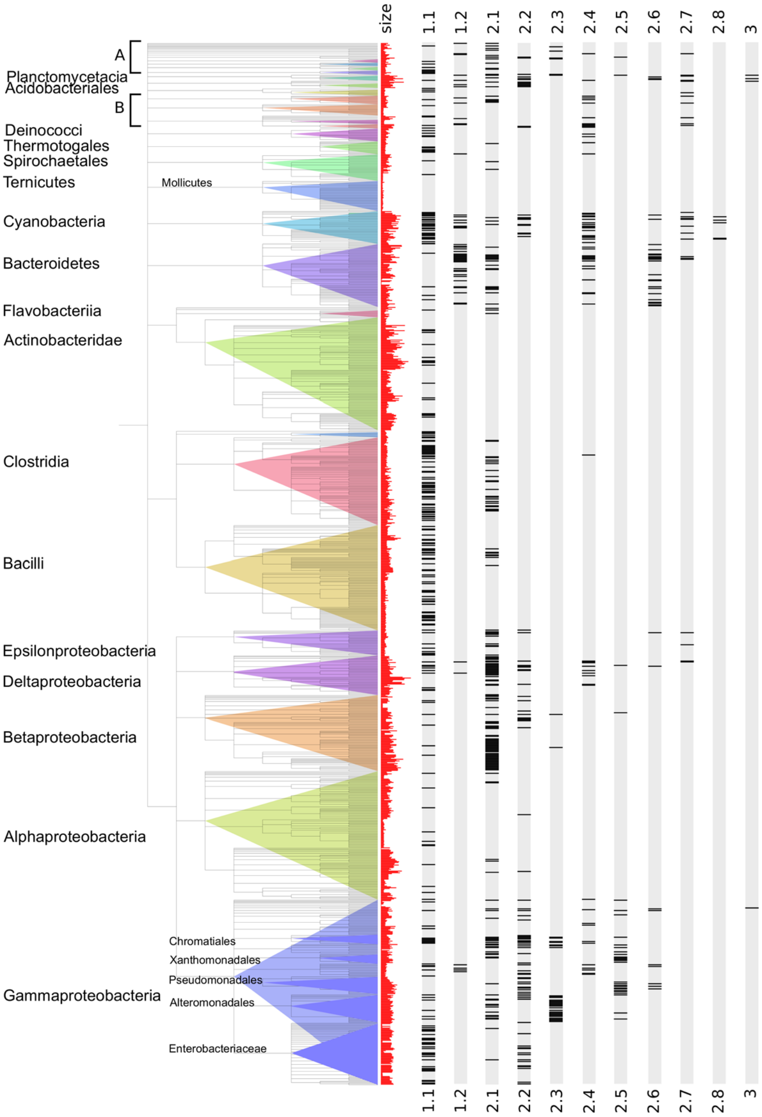

Fig. 4 Taxonomic distribution of the TnpA$_{Y1}$ subclasses. Each species of our sample of complete bacterial genomes is represented by one strain and displayed according to the NCBI taxonomic tree (phyla levels). Names of the largest phyla are indicated, first column, histogram of the genome sizes (red), next columns gene occurrence for each subclass. Black bands indicate the presence of at least one copy of the subclass in the genome. **a** includes Deferribacterales, Nitrospirales, Synergistia and Fusobacteriales. **b** includes Aquificales, Chlorobia, Chlamydiales and Chloroflexi

within each family. To facilitate further analyses the core domain was split into upstream and downstream sub-domains (annotated in blue; Fig. 5 top). The upstream sub-domain ends a few residues after the HuH motif and the downstream sub-domain starts several residues before the conserved W residue. The major length variations are concentrated between the strongly conserved structural elements, β3 and αC and between αC and β4. The insertion domains between β3 and αC are restricted to subclasses 2.3 and 2.4 but differ from each other. The sequence of the 2.3 insertion domain (55 AA) is strongly conserved while the 2.4 insertion domain is highly variable both in length and sequence. Inspection of the DNA sequences of members of this subclass with MEME revealed short repeats that could explain the observed heterogeneity in subclass 2.4 domain (see Additional file 1: Figure S6).

Additional small insertion domains are observed, between β1 and αA (10 AA) in subclass 2.3 and at the end of αB (8 AA) in subclass 2.4. These are located in similar parts of the 3D structure (see Additional file 1: Figure S7A and S7B respectively). These domains occur far from the projected DNA/protein interaction surface suggesting that they may not be involved in DNA binding or processing. The short insertions observed in subclasses 2.2 (10 AA) (see Additional file 1: Figure S7D)

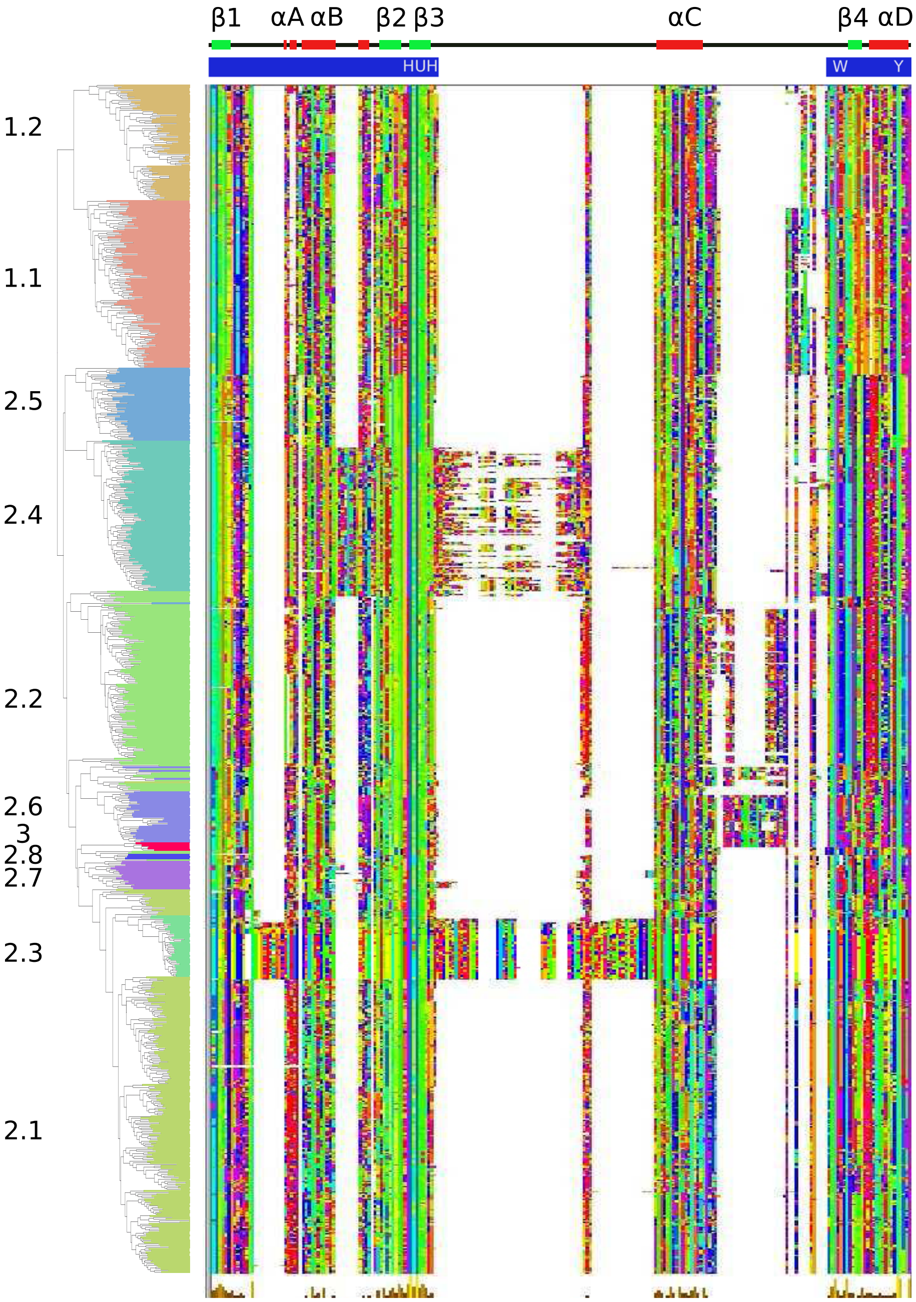

Fig. 5 Multiple alignment of the conserved core region. The alignment was reordered according to the phylogenetic tree. The secondary structure of *E. coli* TnpA$_{REP}$ (PDB 4ER8) is shown above the alignment. The location of the core upstream and downstream sub-domains is shown by blue bands. The alignment was trimmed to remove rare insertions and the display was obtained with Jalview and the Taylor color option

and 2.6 (21 AA) (see Additional file 1: Figure S7E) between αC and β4 are located close to a structural region involved in REP recognition (see Additional file 1: Figure S7C) suggesting they might contribute to DNA/protein interactions.

To characterize the overall class-specific domain organization and sequence conservation, we designed new HMMER profiles for the full length proteins. The protein domain map was constructed with reference to the phylogenetic tree (Fig. 6a). The different proteins are schematized as horizontal grey lines with color segments highlighting the different domains (Fig. 6b-d). This reveals that class 1 members are shorter and more homogeneous in size than are class 2 proteins. Note that sequences of subclass 2.1.2 have a higher heterogeneous length distribution.

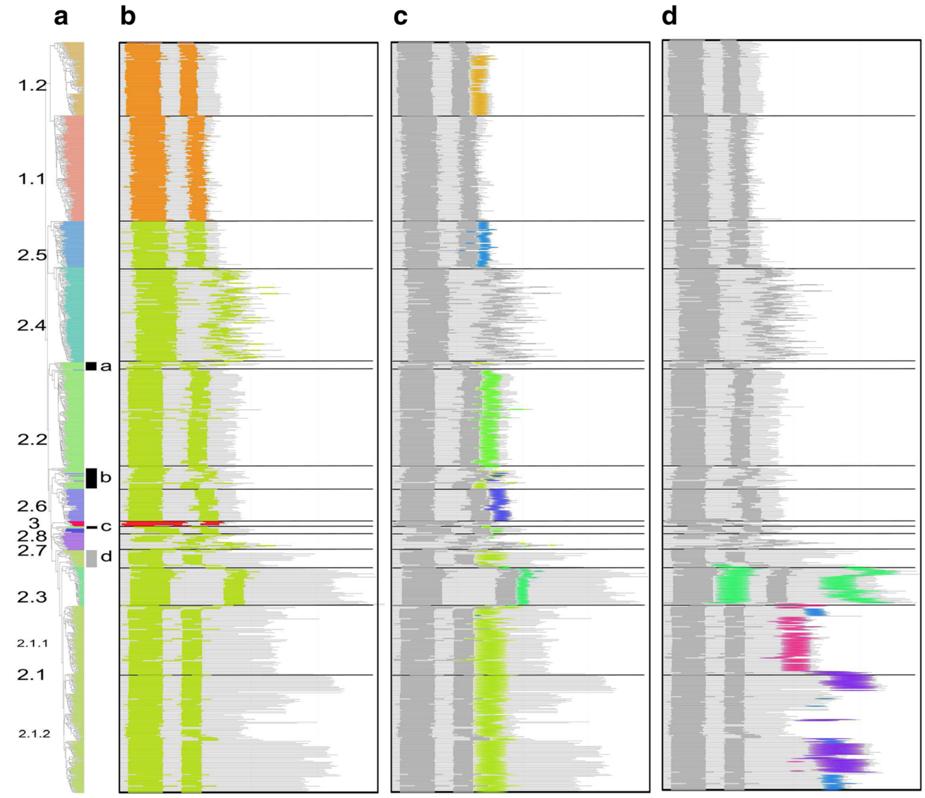

Fig. 6 Global domain organization. Proteins are shown as horizontal grey lines with colored patterns that refer to distinct features. Each panel highlights sequential protein domains. **a** Phylogenetic tree with subclass annotations. The black and grey segments highlight subsets of sequences misclassified by MCL (Fig. 3). **b** Positions of the upstream and downstream sub-domains of the core domain are shown in orange for TnpA$_{IS200/IS605}$, green for TnpA$_{REP}$ and red for class 3. **c** Short conserved domains found next to the core domain. **d** Additional conserved domains found in C-term regions and between two half of the core domain

In general the distance between conserved core upstream and downstream sub-domains is well preserved within each family although in subclass 2.4 it is quite variable as previously mentioned (Fig. 6b). Short length variations between subclasses are observed at the protein N-terminus. These are longer and more variable in the small subclass 2.7. Figure 6c highlights the domain following the core domain. Most classes, except subclass 1.1 and subclasses 2.4 and 2.7, exhibit such a core domain extension with different subclass-specific conserved domains as judged by their HMMER profiles (see Additional file 1: Figure S8). Moreover, domains of subclasses 2.1, 2.2, 2.3 and 2.6 exhibit similarities in their extension regions (see Additional file 1: Figure S8) including a conserved serine. These short motifs were also detected by MEME (see Additional file 1: Figure S3). The sequences highlighted by the arc a (Figs. 3 and 6c) that appear misclassified in subclass 2.2 do not actually share the specific subclass 2.2 extension. The sequences featured by arc b (Figs. 3 and 6c) show a non-homogeneous extension with some sequences sharing the subclass 2.6 C-terminus. Finally, the 2.1 proteins placed as an outgroup of subclasses 2.1 and 2.3 (arc d, Figs. 3 and 6c) possess the subclass 2.1 extension motif.

Additional conserved domains were identified in the longest sequences (Fig. 6d). Subclass 2.3 carries a specific C-term domain (see Additional file 1: Figure S9A) in addition to the conserved domain (Fig. 5) between upstream and downstream core segments (see Additional file 1: Figure S9A). Division of subclass 2.1 into subgroups 2.1.1 and 2.1.2 suggested by the tree topology (Fig. 3) is reinforced by the C-term organization. Subgroup 2.1.1 carries a conserved C-term extension, while subgroup 2.1.2 is more heterogeneous both in length and by the presence or absence of conserved sequences (purple and blue, Fig. 6d). The 2.1.2 subgroup additional C-term domains include a region with similarity to Pfam Bac_DnaA_C domain (purple), a motif found in DnaA proteins which binds 9-bp repeats upstream of the oriC replication origin and activates bacterial DNA replication initiation. Other C-term ends have similarities with either the Pfam HTH_23 or HTH_28 and are likely to be involved in DNA binding. The Bac_DnaA_C (purple) and HTH (blue) related domains (see Additional file 1: Figure S9B) share conserved residues and may correspond to a long and short version of an ancestral helix-turn-helix domain.

REP-like sequences in the vicinity of *tnpA*$_{\mathrm{REP}}$

Full length *tnpA*$_{\mathrm{REP}}$ genes are often flanked by REP sequences present in multiple copies dispersed in intergenic regions of the host genome [5, 22, 36, 37]. Each *tnpA*$_{\mathrm{REP}}$ is associated with a specific REP sequence family. A genome may carry more than one type of *tnpA*$_{\mathrm{REP}}$ gene but these are all associated with their own specific REP sequence family [22, 34, 35, 37]. A given REP sequence family can include different subfamilies, as observed in *E. coli* MG1655 where three different related REP sequences occur (Y, Z1 and Z2) [31], and its single REPtron is composed of Y and Z2 REPs arranged in three REP sequence pairs in inverted orientation (three REPINs). It has been proposed that TnpA$_{\mathrm{REP}}$ is involved in the mobility and/or amplification of REP sequences [5, 22, 32, 33] and the specific nuclease activity of *E. coli* TnpA$_{\mathrm{REP}}$ [5, 23] strongly supports this hypothesis.

To determine whether the other *tnpA*$_{\mathrm{REP}}$-like genes are also associated with REPs, we searched for putative REP sequences in each genome with RNAMotif [41] (Methods). Canonical REP sequences, with a 5′ GT[AG]G tetranucleotide sequence at the foot of a secondary structure, were found in the close neighborhood (+/− 1000 nt) of the large majority of the subclasses 2.2, 2.4, 2.5, 2.7 and 2.8 genes.

The results are shown according to the reference phylogenic tree (Fig. 7a). This program recovered the vast majority of REP sequences which had been identified manually (Fig. 7b, red rectangles). However, in a few instances the program did not detect manually identified putative REP sequences (Fig. 7b, violet rectangles). The detailed organization of REP sequences found by automatic annotation and belonging to the more frequent REP sequence family are shown on Fig. 7c.

For classes 2.6 and 3 a manual inspection was necessary to identify putative motifs since these diverge significantly: they do not carry the consensus GT[AG]G sequence and have a weaker secondary structure (see Additional file 2: Table S1). In nearly half the sample, as in *E. coli*, each gene was flanked by more than one REP sequence subfamily (Fig. 7d). Neither automatic nor manual searches identified REP-like sequences associated with the 2.1 and 2.3 TnpA$_{\mathrm{Y1}}$ subclasses.

Class specific REP sequence features

REP sequences can have variable lengths with either long or short hairpins (Methods). The REP sequence stem length distribution is TnpA$_{\mathrm{Y1}}$ subclass specific. Subclass 2.2, which contains the *E. coli* TnpA$_{\mathrm{REP}}$, includes a majority of REP sequence families with longer stems showing miss-pairing (bulges on the secondary structures, Fig. 7e). The *tnpA*$_{\mathrm{REP}}$ genes assumed to be misclassified in subclass 2.2 (Figs. 5 and 6, black arcs a, b and c) are enriched in a GTGG motif confirming their marginality (Fig. 7a black bars, a, b and c).

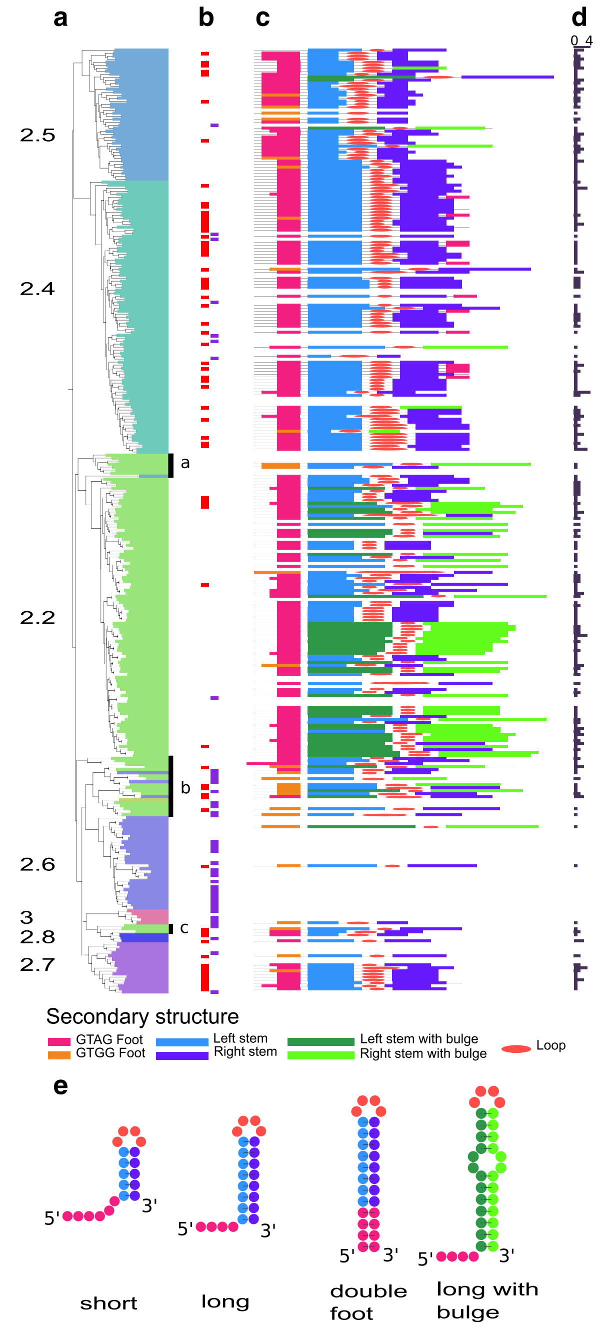

Fig. 7 REP secondary structures and distribution. **a** Phylogenetic tree with subclass annotations. **b** REP sequences annotated by manual inspection: red rectangles, REP sequences also found by automatic annotation and violet rectangles, REP sequences missed by automatic annotation. **c** Schematic representation of REP sequences found by automatic annotation, only the more frequent REP sequence family is reported for each REPtron. **d** Histogram of the number of REP sequences families per REPtron. **e** Distinguishable REP sequence secondary structure types

Subclass 2.4 contains two REP types: canonical long REP sequences, where the bottom of the hairpin stem is composed of 3 consecutive G-C base pairs, and particular long REP sequence families with a 5′ GT[AG]G and its reverse complement, C[TC]AC 3′, at the end of the sequence ("double foot", Fig. 7e). Note, that a double foot REP sequence could be generated simply by central deletion between an inverted REP sequence dimer. A unique feature of subclass 2.4 is that, in addition to the REP sequences flanking the *tnpA*$_{\mathrm{REP}}$ genes, REP sequences can be found inserted within the *tnpA*$_{\mathrm{REP}}$ genes themselves (see Additional file 1: Figure S10). Remarkably, these insertions do not disrupt the reading frame (Figs. 5 and 6). At the protein level, the insertions are generally located between the HuH and Y motifs. REP sequence insertions in a variety of unrelated genes have also been reported. These insertions may not affect the activity of the corresponding proteins if they are located in flexible linkers or loops [36]. On the predictive structural model of subclass 2.4 (see Additional file 1: Figure S7B), the variable region appears as a peripheral extension away from the ssDNA binding and catalytic sites which may have little impact on protein activity.

Subclass 2.5 TnpA$_{\mathrm{Y1}}$ are associated with short REP sequences. In some of these, the 5′ conserved tetranucleotide is located 1–2 nt from the secondary structure rather than directly at its foot but are "compensated" by shorter stems. Thus, the sum of the foot and stem length is distributed in a narrow range (10 to 13 nt). In *Stenotrophomonas maltophilia*, a subtree of subclass 2.5, REP sequences are more similar to the long canonical REP sequences of subclass 2.4. TnpA$_{\mathrm{Y1}}$ genes of subclas 2.6 are flanked by atypical secondary structures identifiable by manual inspection (Fig. 7). Copies of these secondary structures can be found within the host genome. Genes of subclasses 2.7 and 2.8 are flanked by long canonical REP sequences with 5′ GT[AG]G tetranucleotides.

Recognition of REP guide sequence

The residues involved in TnpA$_{\mathrm{REP}}$ interaction with the GTAG guide sequence have been identified in the case of *E. coli* (see Additional file 1: Figure S11A) [23]. Although sequence alignment revealed that these residues are not all strictly conserved (see Additional file 1: Figure S11B), W 94, Q95 and R104 residues are well conserved in subclass 2.2, 2.4, 2.5 and 2.7. E100 is fully replaced by D in subclass 2.5 and 2.7 and H101 is sometimes replaced by R in the same subclasses. Despite the very high level of W94 conservation its role is at presently unclear. Therefore, subclasses with *tnpA*$_{REP}$ associated with typical REPs have sequence conservation in this region, suggesting that, as in the case of *E. coli* TnpA$_{\mathrm{REP}}$, it is involved in recognition of the GTAG sequences within members of these subclasses. The sequences of subclass 2.1 and 2.3 members have different motifs and sequences of subclass 2.6 are not conserved in this region.

Subclass specific features summary

Class 1

Class 1, composed of subclasses 1.1 and 1.2, is present in 25.6% of our sample of species, close to the 26% estimated previously [38].

Subclass 1.1 Subclass 1.1 is the largest group, present in 22.4% of all species (Table 1) and includes typical IS*200* tranposases (ISfinder [11, 42, 43]). It shows the highest taxonomic diversity of all TnpA$_{\mathrm{Y1}}$ proteins (Figs. 3 and 4) and it is the only subclass observed in Archaea. A characteristic signature is the Y motif (Y- - -Q) (see Additional file 1: Figure S3) and the hydrophobic residue (u) of the HuH motif is either V or I. These are the shortest TnpA$_{\mathrm{Y1}}$ family members, composed essentially of the core without additional domains. These *tnpA* genes are often found in microsynteny with a *tnpB* gene whose function is not yet entirely clear [16] and flanked by the typical small subterminal DNA hairpin structures (ISfinder). The high genomic copy number (see Additional file 1: Figure S12), the presence of partial sequences, the high sequence conservation observed between copies within the same genome (see Additional file 1: Figure S13) and their wide distribution in Archaeal and Bacterial phyla are consistent with the IS nature of these sequences [11, 44, 45].

Subclass 1.2 Subclass 1.2 proteins are closely related to those of subclass 1.1 (Table 1; Figs. 2 and 3) but with distinctive features: a characteristic Y motif signature, Y- - -Q- -HH (see Additional file 1: Figure S3); a shorter core domain (see Additional file 1: Figure S4), a conserved specific C-term domain (Fig. 7c); a low genomic copy number (see Additional file 1: Figure S12); generally without typical terminal IS*200* secondary structure features; and a taxonomic distribution mostly restricted to species of Cyanobacteria and Bacteroidetes phyla (covering less than 4% of all species). Therefore, despite their close relationship, this subclass does not share the IS properties of subclass 1.1. It was not described by Bertels et al. [38] and to our knowledge, this is its first characterization.

Class 2

Class 2, present in 25.5% of our species sample, includes the known TnpA$_{\mathrm{REP}}$ proteins and is composed of 8 subclasses 2.1 to 2.8. Class 2 genes have low intra-genomic copy number (except subclass 2.3, see Additional file 1: Figure S12) and low sequence conservation between copies within the same genome (see Additional file 1: Figure S13). The common Y-motif signature of all class 2 members is Y- - -NP. The hydrophobic residue (u) of

Table 1 Summary of the TnpA$_{Y1}$ family subclasses properties

Subclass[a]	Subclass size	Subclass coverage	Median full length	Key motifs[b]			Length	Core domain				Protein tail domain			Intra-genomique copies		REP	Taxonomic distribution
								Sequence insertion[c]				Proximal		Distal				
				HuH	W	Y		β1:αA	αB	β3:αC	αC:β4	hmm[d]	motif	hmm[d]	Mean	Identity[e]		
1.1	2620	22.42	152	D**H**[VI]**H**	**W**[TS]	**Y**[IV]E**NQ**	117								4.8	99.4	small subterminal DNA hairpin structures	Archaea Bacteria
1.2	101	3.83	148	D**H**[IV]**H**	**W**Q	**Y**IK**NQ**KE**HH**	112					30			1.7	36.2	no	Cyanobacteria Bacteroidetes
2.1 (1)	493	14.9	238	N**H**V**H** N**H**[YF]**H**	**W**[EQ] **F**Q	**Y**IEL**NP** **Y**[IV]HL**NP**	111					53	SS	49 37 69	1.9	49.2	no	Bacteria prevalence in ε, β, δ proteobacteria
2.2 (2)	270	6.71	165	D**H**L**H**	**W**E	**Y**IHY**NP**	111				10	33	SS		1.5	42.3	long long with bulge	Proteobacteria
2.3	224	2.88	325	N**H**Y**H**	**W**E	**Y**VDL**NP**	172	10		55		21	TS	61	3.9	74.8	no	Chromatiales Alteromonadales
2.4 (4)	124	4.94	194	N**H**[IFV]**H**	**W**Q	**Y**IxN**NP**	146		8	yes[f]					1.5	40.1	long double foot inserted in *tnpAREP*	Cyanobacteria Bacteroidetes and various phyla
2.5 (3)	87	2.21	151	D**H[**LF]**H**	**W**Q	**Y**I[VI]A**NP**	111					20			1.6	55.3	short few long	*Xanthomonad* *Pseudomonad*
2.6 (5)	63	2.51	179	N**H**[LIV]**H**	**W**Q	**Y**IH[QN]**NP**	135				21	29	SS		1.8	40.7	atypical secondary structures	Bacteroidetes
2.7	36	1.84	213	[ND]**H**V**H**	**W**Q	**Y**I[LR][NQ]**NP**	102								1.2	37.4	long	Cyanobacteria and few other phyla
2.8	8	0.37	199	N**H**Y**H**	**W[**YH]	**Y**IHY**NP**	112								1.0	na	long	Cyanobacteria
3	14	0.29	178	N**H**V**H**	**W[**TA]	**Y**V[VI][AENY]E**Q**	155								1.2	60.7	atypical secondary structures	Planctomycetes Acidobacteria Proteobacteria

Bold entries correspond to conserved residues

[a] in brackets, Bertels's groups

[b] MEME consensus sequences, conserved AA are in bold face

[c] insertion length (AA)

[d] HMM length (AA)

[e] mean pairwise indentity (%) and

[f] variable length and sequence insertions

the HuH motif is either L, V, I, F, Y or M (ranked by decreasing frequencies) (see Additional file 1: Figure S3).

Subclass 2.1 Subclass 2.1 is the second largest subclass (14.9% of all strains, Table 1). Their members are among the longest proteins of our sample. They carry an additional C-term conserved domain (Fig. 6 and see Additional file 1: Figure S8B). However this subclass can be subdivided into two groups, 2.1.1 and 2.1.2, according both to the tree topology (Fig. 3) and the C-term domain nature that appears conserved in the 2.1.1 members while it is more heterogeneous in 2.1.2 sequences (Fig. 6d). This C-term extension shows similarities with helix-turn-helix motifs (Bac_DnaA_C, HTH_23 or HTH_28 Pfam profiles, see Additional file 1: Figure S9B). They share few conserved residues and may correspond to a long and short version of an ancestral helix-turn-helix domain. Neither automatic nor manual searches identified REP-like sequences in the vicinity of the $tnpA_{\mathrm{REP}}$ gene. Previously, a collection of Y1 proteins homologous to $\mathrm{TnpA}_{\mathrm{REP}}$, was called TIRYT (Terminal Inverted Repeat associated tYrosine Transposase) because they were associated with terminal inverted repeats rather than typical secondary structures [36]. Of the 26 examples described [36], 20 belong to subclass 2.1. All include one of the three HTH domains. However, the consensus sequence GGGG[AT][CG]A[CG] observed in the majority of the flanking inverted repeats does not resemble the consensus TT[AT]TNCACA of the high affinity DnaA boxes [46] suggesting that these C-term domains may not be directly involved in DNA target recognition or have evolved to recognize a new target sequence. The other 6 TIRYT examples are distributed between subclasses 2.2 (3 from arc b and one from arc a) and 2.5 (2 proteins). The TIRs flanking subclass 2.5 $tnpA_{\mathrm{REP}}$ in *P. putida* and *P. fluorescens* originated from fusion of degenerate REP elements with unrelated sequences. Subclass 2.1 therefore probably corresponds to the TIRYT described by De Nocera et al. [36]. Further analyses are necessary to confirm this.

Subclass 2.2 Subclass 2.2 includes the first $tnpA_{\mathrm{REP}}$ identified in *E. coli* [5, 22, 36] and represents the third largest of the groups (6.7% of all strains). They are predominantly found in Proteobacteria (Figs. 3 and 4; Table 1). Members exhibit short insertions between αC and β4, a region neighboring that involved in REP sequence hairpin recognition in the crystal structure [23] (see Additional file 1: Figure S7D). It is noteworthy that this subclass includes REP sequences with a hairpin-bulge-hairpin secondary structure (Fig. 7) suggesting that these short conserved domains might contribute to an original REP/protein interaction.

Subclass 2.3 The members of subclass 2.3 are found in less than 3% of all strains but prevail in Chromatiales and Alteromonadales where more than five copies per genome can be identified (see Additional file 1: Figure S12; Table 1). However, those copies do not exhibit the high intra genomic sequence conservation observed for members of subclass 1.1 (see Additional file 1: Figure S13). The subtree corresponding to this class is deeply rooted in subclass 2.1 (long branches) (Fig. 3). However, the leaves corresponding to the Proteobacteria are linked by short branches which suggest a recent and active spreading of elements of this distinct $\mathrm{TnpA}_{\mathrm{Y1}}$ subclass in Chromatiales and Alteromonadales strains. As for subclass 2.1, the gene is not flanked by identifiable REP-like DNA secondary structures. Members have the longest sequence length (median equal to 325 AA; see Additional file 1: Figure S1) resulting from multiple domain additions. Two domain insertions occurred between conserved secondary structure elements of the conserved core region (β1 and αA; β3 and αC). These domains are highly conserved and predicted to be exposed to the surface of the protein (see Additional file 1: Figure S7A). Moreover, the long C-terminal tail is composed of two domains. The proximal domain includes a conserved serine residue observed in other subclasses while the long distal domain is subclass specific. The accessibility of all these additional domains suggests that they may be involved in as yet unidentified molecular interactions and/or processes.

Subclass 2.4 Subclass 2.4 is abundant in Cyanobacteria and Bacteroidetes but also scattered across a number of phyla (Figs. 3 and 4; Table 1). Subclass 2.4 contains two REP types: canonical long and double foot REP sequences. A remarkable and unique feature of subclass 2.4 members is the predominance of REP sequence insertions within the $tnpA_{\mathrm{REP}}$ genes (see Additional file 1: Figure S10).

Subclass 2.5 The cognate REPs of subclass 2.5 belong to the larger REP sequence families with typical REP sequence organization. Elements of this subclass are mostly limited to *Xanthomonad* and *Pseudomonad* species (Figs. 3 and 4; Table 1) and some have been described in previous studies [32–34, 47–50]. Comparative genome analysis suggests a recent invasion of $tnpA_{\mathrm{REP}}$ genes in these genomes and that REP sequence diversification and proliferation are still ongoing leading to high numbers of REP copies [37]. All these observations are in agreement with the suggestion that $tnpA_{\mathrm{REP}}$ genes coevolve with their cognate REPs [22].

Subclass 2.6 Subclass 2.6 members are mostly present in Bacteroidetes (Figs. 3 and 4; Table 1). Their genes are flanked by atypical secondary structures identifiable by manual inspection (Fig. 7), copies of which can be found elsewhere in the genome. Like subclass 2.2, they also contain short insertions between αC and β4 (see Additional file 1: Figure S7E). This domain may contribute to the recognition of the unusual REP sequences.

Subclasses 2.7 and 2.8 Members of small subclasses 2.7 and 2.8 are variable in length (Fig. 6; Table 1) and genes are flanked by long canonical REP sequences with 5′ GT[AG]G tetranucleotides. Subclass 2.8 is restricted to Cyanobacteria in our sample but closely related sequences are found from other unrelated phyla in NCBI non-redundant (nr) database. They are characterized by a conserved Y residue in the HuH motif (HYH) and a proline rich N-term domain (see Additional file 1: Figure S3). Some carry a REP sequence insertion.

Class 3

Class 3 proteins are relatively rare (< 0.3% of all species) and found principally in the Planctomycetes, Acidobacterial and Proteobacterial phyla (Table 1). They include a Y-motif signature, Y- - - -Q related to that of class 1 (see Additional file 1: Figure S3), and a distinctive core domain profile (Fig. 6b). Their genes are flanked by several atypical secondary structures with a 5′ tetranucleotide foot CnGA identifiable by manual inspection (Fig. 7, see Additional file 2: Table S1). This feature is more reminiscent of a REPtron than an IS. Moreover, the associated DNA secondary structures can be present in high copy number in a given host genome (e.g. ~ 600 in *Solibacter usitatus* Ellin6076), again reminiscent of REP sequences. Interestingly, there are three closely related class 3 genes in this bacterium. Thus class 3 examples resemble both IS and REPtrons and are possibly evolutionary intermediates between class 1 and 2 (Fig. 2c). Further experimental analyses are required to determine their origin and behaviour.

Discussion

Previous studies have underlined the similarities between IS*200*/IS*605* family transposons (their HuH-Y1 transposases and their terminal DNA secondary structures) and the bacterial REP systems (the TnpA_{REP} proteins and their associated REP sequences) [5, 22, 23]. Here we have examined the diversity of Y1 HuH proteins related to those of the IS*200*/IS*605* and REPtron families in the complete genomes of archaea and bacteria. We refer to these collectively as TnpA_{Y1}. Based on overall sequence similarities, the sample of full-length proteins can be divided into three classes and 11 subclasses (Fig. 2). This classification is in agreement with the phylogenetic tree obtained from the multiple alignments of the conserved core alone (Fig. 3). There is strong conservation of the regions corresponding to secondary structure elements of *E. coli* TnpA_{REP} [23] and the presence of subclass-specific insertions located between these elements (Fig. 5). The identification of subclass-specific short conserved motifs and of conserved domains flanking the core domain confirms this classification (Fig. 6).

Subclass partition: Strength and weakness of different approaches

Bertels et al. (2017) defined six TnpA_{Y1} groups, with at least 30 or more members, for both $\text{TnpA}_{\text{IS}\mathit{200}}$ and TnpA_{REP} (RAYTs) families. The structure of these two families appeared very different. The $\text{TnpA}_{\text{IS}\mathit{200}}$ family was more homogeneous and the groups (clusters) were very close to each other while those of the TnpA_{REP} (RAYTs) family were more distinct. Our analysis structures the TnpA_{REP} (RAYTs) family into eight groups. To establish the correspondence between both classifications we used the sequences provided by Bertels and collaborators (supplementary data) for their five larger RAYT groups. The IS group sequences and the RAYT group 6 sequences were not available. 291 of the 292 RAYT group 1 sequences are assigned to our subclass 2.1; 238 of the 239 RAYT group 2 sequences belong to our subclass 2.2; 113 of the 120 RAYT group 3 sequences are allocated to our subclass 2.5; the 100 RAYT group 4 sequences are assigned to our subclass 2.4 and finally the 40 RAYT group 5 sequences are found in our subclass 2.6. Our classification is therefore only partially covered by the groups defined in Bertels et al. (2017). One of our larger subclasses (2.3) and the smallest (2.7, 2.8) were not identified. Neither did they detect class 3, whose characteristics suggest that it may represent an evolutionary intermediary between classes 1 and 2. In addition, the well supported distinction between $\text{TnpA}_{\text{IS}\mathit{200}}$ subclass 1.1 and the related subclass 1.2 was not found and the evolutionary relationships between $\text{TnpA}_{\text{IS}\mathit{200}}$ and TnpA_{REP} (RAYTs) families was not clearly established.

These differences could be explained by the different methodological workflows implemented. First for the identification of the TnpA_{REP} (RAYT) and $\text{TnpA}_{\text{IS}\mathit{200}}$ protein candidates, Bertels et al. (2017) used two independent tblastn searches based on two queries, one for each protein family ($\text{TnpA}_{\text{IS}\mathit{200}}$ and TnpA_{REP}) while we undertook a more sensitive profile based search (hmmsearch) using the transposase IS*200* like (Y1_Tnp) profile from Pfam. At this step, an e-value threshold of 0.01 was used to maximize detection of remote homologs. Sequences of subclasses 2.3, 2.7, 2.8 and 3 could be too distant from the Bertels query sequence (*Pseudomonas fluorescens* SBW25, PFLU_RS20900 belonging to subclass 2.5) to be identified by their tblastn approach (e-value < 0.02). In addition, we then reduced the sequence redundancy (due to recent gene duplications and the unbalanced representation of bacterial strains in the dataset) by clustering closely related sequences (cut off of 70% of identity), each cluster was then further reduced to one representative sequence, the medoid.

Exploration of sub-family organization was performed in both analyses using MCL graph partitioning but the similarity relationships between pairwise sequences were calculated differently. We generated a reciprocal all-against-all BlastP comparison (e-value <1e^{-5}) on our data set including both TnpA$_{IS\mathit{200}}$ and TnpA$_{REP}$ (RAYTs) proteins and the BlastP results were reformatted to an undirected and weighted graph (log of e-value). Different MCL inflate factor values were tested to select the optimal partition. Our approach is inspired by TRIBE-MCL [51] (https://micans.org/mcl/), a method for detecting protein families in large databases. Bertels et al. (2017) computed relationships between proteins of the TnpA$_{IS\mathit{200}}$ or TnpA$_{REP}$ (RAYTs) family independently by pairwise sequence comparisons obtained with a global alignment algorithm (Needleman–Wunsch). The results were filtered to retain only the pairs with more than 26% identity and then converted into a weighted graph (pairwise identity). The use of such a global alignment does not appear appropriate since, as we describe above, the proteins show variable domain organization and length heterogeneity (Table 1). To explore the relationships between the groups identified independently in both families they then randomly selected 30 sequences in each group and partitioned the graph using MCL as explained above. An inflate factor of 2.0 was invariantly used for graph partitioning with MCL.

Finally, we explored the evolutionary relationships between the proteins of the family by constructing a tree based on the conserved core domains shared by the entire sequence set. Only potentially functional sequences (with the conserved HuH and Y motifs) were retained because pseudogenes with a higher evolution rate could disturb tree topology. Bertels et al. (2017) computed a phylogenetic tree by randomly selecting three sequences from each of the TnpA$_{IS\mathit{200}}$ and RAYT groups.

Vertical versus horizontal inheritance

Although there is no evidence that class 2 genes are mobile, their presence in distantly related species suggests that the ancestor appeared early in evolution. However, members of this class are observed in only 25.5% of species of our sample with a patchy taxonomic distribution, in agreement with Bertels et al. (2017). The hypothesis of vertical inheritance would imply multiple independent gene loss events during the course of evolution.

A more parsimonious model is to assume that dissemination of class 2 genes occurred by multiple horizontal gene transfers (HGTs). Support for this hypothesis comes from the observation of *tnpA*$_{REP}$ gene transfer from a Pseudomonad to marine gammaproteobacteria [22] and evidence that HGT is likely to have occurred between fluorescent pseudomonad strains [37]. In the absence of formal evidence for autonomous mobility of these elements, *tnpA*$_{REP}$ genes might use the same routes as accessory genes for transfer (e.g. transformation [52], conjugation [53], transduction [54], and gene transfer agents [55, 56]). The taxonomic distribution of class 2 genes (Fig. 4) is consistent with the observation that HGT events are more frequent between closely related species [57]. Indeed, except for subclass 2.1, and to a lesser extent for subclass 2.4, *tnpA*$_{REP}$ genes are mostly confined to a specific taxonomic group. This proximity would also favour integration of class 2 proteins into host cell metabolism by, for example, providing domains for interaction with other host proteins or influencing gene expression. The observation here, that intra-genome gene copies belonging to the same subclass are distantly related (see Additional file 1: Figure S13), is also compatible with multiple HGTs. Gene loss should also be considered as an evolutionary force since such events are suspected in genomes encoding REP sequences in the absence of *tnpA*$_{REP}$ gene [36] and is well documented in *E. coli* strains, where the REPtron locus had been replaced by an operon encoding toxin/antitoxin genes [23]. Gene loss may have occurred after the acquisition of *tnpA*$_{REP}$ by HGT in the ancestor. The contribution of HGT and gene loss can be estimated with reconciliation methods that compare gene trees and species trees to recover the history of gene families [58], but this is outside the scope of this work.

Functional outcomes

All sequences share a conserved structural core domain including HuH and Y motifs involved in the distinctive site-specific ssDNA cleavage and ligation mechanism of this protein superfamily [10, 15, 23]. However, the local context of the catalytic tyrosine is class-specific: Y- - -Q, Y- - -NP and Y- - - -Q for classes 1, 2 and 3, respectively. The position of the conserved glutamine residue (Q131 in the *Helicobacter pylori* IS*608* transposase, Fig. 1), on the same face of αD helix as the catalytic tyrosine residue, is essential for transposition in vivo and this residue is part of the divalent metal ion binding site with the histidines of the HuH motif [10]. The same role was assigned to the asparagine residue (N119) in TnpA$_{REP}$ from *E. coli* MG1655 [23]. The sequence conservation observed suggests that proteins of different classes have the same single-stranded DNA editing activities.

The Y1 transposases including TnpA$_{IS\mathit{200}}$ and TnpA$_{REP}$ of *E. coli* require ssDNA. Although this can be provided by a variety of different processes [16, 17, 21, 23, 59], the lagging strand of the replication fork is an important source in vivo [60]. The propensity for excision from and insertion into the lagging strand and the intimate coupling of single strand transposition to replication has been well documented for two members of the IS*200*/IS*605* family [11–13, 16, 19].

Transposase targeting can be mediated by direct interaction with proteins of the replication fork complex. For example, TnpA of IS*608* ($\text{TnpA}_{\text{IS}608}$) and other $\text{TnpA}_{\text{IS}200}$ family members are thought to interact with DnaN [21, 61], the β sliding clamp replisome processing factor. Moreover, $\text{TnpA}_{\text{IS}608}$ shows affinity for DNA structures resembling replication forks [21]. It has been proposed that the region of $\text{TnpA}_{\text{IS}608}$ interaction with DnaN is located in the C-term region, next to the catalytic tyrosine, in the αE helix [61]. A survey of enzymes that interact with the β sliding clamp within the replication fork reveals that they share a short and poorly conserved binding motif (consensus QL[SD]LF) [62]. These motifs are often located in a highly flexible C-terminal tail of the protein [62–64]. However, this sequence is not conserved within $\text{TnpA}_{\text{IS}200}$ members. In spite of this, it seems possible that the short serine motif located next to the catalytic tyrosine in subclasses 2.1, 2.2, 2.3, and 2.6 could play a role of targeting the protein to the replication fork via protein-protein interactions. Other subclasses may have evolved alternative subclass-specific motifs or domains for interaction with other host targets. For example, the DnaA-like DNA binding domain in subclass 2.1 proteins suggests an alternative, more direct, way of targeting the replication fork. Interaction of transposase with universal and highly conserved proteins such as sliding clamps would facilitate transfer between phylogenetically distant organisms [65]. Indeed, the transposition pathways of a number of IS and transposons require host enzymatic functions [65–68], including DNA polymerases and other factors implicated in DNA replication, suggesting a functional link between transposition and replication [17, 21].

The discontinuous taxonomic distribution of class 2 proteins supports the notion of a transitory positive impact of REPtrons on host cell fitness. Although their function has yet to be formally elucidated, their possible involvement in single-stranded DNA editing and in the proliferation of small dispersed repeated sequences could be a beneficial factor in adaptive transition phases. After this phase, gene silencing or gene loss could restore genome stability. Their phylogenetic proximity to $\text{TnpA}_{\text{IS}200}$ suggests that class 2 proteins may represent mobile element domestication [5, 22, 32, 38].

Conclusions

Previous studies have underlined the similarities between IS*200*/IS*605* family transposons, their HuH-Y1 transposases and their terminal DNA secondary structures, and the bacterial REP systems, the TnpA_{REP} (RAYT) proteins and their associated REP sequences [5, 22, 23]. Here, we performed a genome-wide analysis of TnpA proteins related to those of the IS*200*/IS*605* and REPtron families in complete genomes of archaea and bacteria. We refer to these collectively as TnpA_{Y1}. All sequences share a conserved structural core domain including HuH and Y motifs involved in the distinctive site-specific ssDNA cleavage and ligation mechanism of this protein superfamily [10, 15, 23]. Based on sequence similarity, these proteins can be arranged in classes and subclasses. Subclass 1.1 includes sequences similar to IS*200*/IS*605* transposases while proteins of the other subclasses are not. Here, we identify and characterize a new subclass (subclass 1.2) closely related to IS*200*/IS*605* transposases but which does not share IS properties, as well as a previously unknown group of HuH-Y proteins (class 3). Each subclass is characterized by specific additional sequence domains possibly involved in protein/DNA or protein/protein interactions and in targeting these proteins to single-stranded DNA. The taxonomic distribution reveals that more than 25% of the analyzed species encode at least one $tnpA_{\text{REP}}$-like gene, but with accumulation in some phyla. The average size of genomes containing at least one $tnpA_{\text{REP}}$-like gene is significantly larger than that of genomes without any of these genes. Their patchy taxonomic distribution is in agreement with dissemination by multiple horizontal gene transfers followed by gene loss. The genes, of a closely related subset of subclasses, are flanked by typical REP sequences in a structure called REPtron. The proteins encoded by these genes are assumed to be responsible for REP sequence proliferation in the genome. Proteins of all subclasses share a common catalytic domain involved in single strand DNA editing. The disparity observed in taxonomic distribution and the diversity of domain arrangements of the proteins belonging to different subclasses suggest that they evolved to different cell physiology and raise the possibility of their domestication in cell function related to single strand DNA editing.

Methods

Building TnpA_{Y1} dataset

Completely sequenced genomes of 172 archaea and 2178 bacteria covering 1351 distinct species were downloaded from EBI (http://www.ebi.ac.uk/genomes/). The complete genomes of these 2355 strains, their proteomes and EMBL features were managed with an in house mySQL database. We also downloaded the collection of hidden Markov models (HMM) from Pfam (http://pfam.xfam.org, release 29.0).

To retrieve transposase IS*200*-like proteins, we adopted a two step approach. First, the Pfam entry Y1_Tnp (PF01797), covering the transposase IS*200*-like protein family, was used as query with hmmsearch from the HMMER3 package (http://hmmer.org/; [69]) against the 7,223,104 protein sequences from our sample. Only proteins showing an alignment with an e-values < 0.01 (recommended threshold) were retained. A first set of

4081 proteins was obtained. Second, we reduced the sequence redundancy inherent to our sample due to i) the multiple replicates from strains of the same species (e.g. *E. coli* includes 56 strains in our sample), ii) the unbalanced distribution of bacterial phyla in public databases, leading to the overrepresentation of some closely related species and iii) finally the presence of a high number of identical copies in some genomes due to the repetitive nature of the element (e.g. 99 copies of IS*200* in *Yersinia pestis* Angola). To minimize this bias which can affect the performances of multiple alignment and phylogenetic methods, closely related sequences were represented by one sequence, the medoid. To identify this, we performed all-against-all BlastP comparisons of our initial set of proteins. An identity cut off of 70% was chosen for retaining similarities as it offers the best compromise between the sample size reduction and the conservation of the overall $\mathrm{TnpA}_{\mathrm{Y1}}$ family structure. Protein relationships were then converted into a graph in which the vertices represent protein sequences, and the edges represent their relationships. The graph was further processed by a graph-partitioning approach based on the Markov Clustering algorithm (MCL, [70]). This identified 687 clusters composed of a unique sequence and 244 clusters composed of closely related sequences. For each of the 244 clusters, we computed the medoid, i.e., the sequence whose average dissimilarity to all the other proteins in the cluster is minimal, for which we add the constraint that its length should be close the median length of all sequences of the cluster. This resulted in a set of 931 proteins composed of 687 unique sequences and 244 medoids. Finally, all-against-all BlastP comparisons (e-value $<1e^{-5}$) of the restricted 931 protein sequence set were performed. The graph obtained was composed of a large connected component of 924 vertices and of singletons or two vertices. Pfam protein annotation revealed that, contrary to the 924 proteins, those belonging to the small clusters had best hits with conserved domains unrelated to the $\mathrm{TnpA}_{\mathrm{Y1}}$ Pfam domain. They were considered as false positives and removed from our sample to yield a final sample of 924 proteins for further analyses.

HMM profiles

To characterize the different protein families and domain organization, we used the HMMER package (http://hmmer.org). This implements methods using probabilistic models, profile hidden Markov models (profile HMMs), that assign a position-specific scoring system for substitutions, insertions, and deletions [71–73]. This is used for construction of protein domain models and is the basis of the protein domain model database, Pfam [74, 75]. As seeds, we used the representative sequences described above. Alignments were obtained with mafft version 7 [76] with default parameters "except –localpair –maxiterate 1000" parameters to ensure a high accuracy. The multiple alignments were edited with Jalview [77] and the core domain was retained following deletion of left and right flanking regions. Partial or highly divergent sequences were removed. The HMMER profiles were built with hmmbuild from the input multiple alignments and assembled in an HMM database with hmmpress. Protein sequences were scanned against this database with hmmscan. The domain annotation file was parsed with a Perl program to select the best non-overlapping domains in each sequence. The residues which aligned with the conserved motifs in the HMMER profiles were extracted and the results of the domain annotation saved in the database.

Phylogenetic tree

Multiple alignments were processed with trimAl [78] to eliminate aligned positions with a high frequency of gaps. ProtTest [79] was used to select the optimal parameter combination. The resulting parameters included the LG model of sequence evolution with γ-correction (four categories of evolutionary rates), shape parameter and proportion of invariant sites estimated from the data. Phylogenetic trees were computed with PhyML [80, 81]. Branch supports were estimated with parametric bootstrap analysis. Trees were drawn and annotated with the interactive Tree Of Life web server (iTOL, http://itol.embl.de/) [82, 83].

Motif identification

The search for conserved motifs used MEME program (Multiple Em for Motif Elicitation, http://meme-suite.org/) [84]. Individual MEME motifs do not contain gaps. The unaligned sequences were submitted to MEME with a motif width between 3 and 10 amino acids. We searched for a maximum of 20 motifs with an occurrence of zero or one motif per sequence since we did not expect that each motif would be present in all sequences. Repeats were predicted in $tnpA_{\mathrm{REP}}$ genes with 'any number of repetitions' option of MEME. The results of MEME were used by MAST (Motif Alignment & Search Tool) program to annotate the motifs in sequences of our database.

REP sequence prediction

REPs are not conserved in sequence between species (species-specific) and within a genome copies change slightly in sequence and length (imperfect/partial repeats). However, they share a common structural feature: a tetranucleotide guide sequence located 5′ to the foot of a short hairpin. We used the RNAMotif program [41] to search for REP sequence candidates in $\mathrm{TnpA}_{\mathrm{REP}}$-encoding genomes. This program requires a file that describes the secondary structure interactions and the set of rules used to filter matching sequences which was constructed from the analyses of REPs reported in [36]. It includes 66 previously described REP sequence families found mostly in Proteobacteria. All include a GTAG or GTGG tetranucleotide located between 0 to 2 nt 5′ from a secondary structure and include a basal stem of fully paired 7 to 20 nt and a loop of 2 to 5 nt (short hairpins) or an additional fully paired stem of 1 to 10 nt connected to the first by a bulge of 1 to 3 nt (long hairpins) (see Fig. 7e). In addition, the first four basal stem nucleotides include at least three G or C bases (short hairpins) or the GC content of the stem is > = 50% (long hairpins). Only four REP sequences out of 66 do not follow these rules. The RNAMotif program was applied. The REP sequence candidates predicted in DNA sequences of complete genomes by RNAMotif were clustered into families with BlastN and MCL [70]. For each family a multiple alignment was obtained with mafft [76] and the secondary structure computed with RNAalifold [85]. Only families with at least one occurrence located in the vicinity (1000 nt upstream and downstream) of a $tnpA_{\mathrm{REP}}$ gene were retained. The predicted REPs were assembled in dimers (BIMEs) or clusters, if the distance between consecutive REPs on the genome is less or equal to 150 nt. This in silico analysis was supervised by human expertise.

Additional files

Additional file 1: Figure S1. Distribution of protein length in each subclass; **Figure S2.** Coverage of HMM profiles and proteins for each subclass; **Figure S3.** *ab initio* search for conserved motifs with MEME in proteins of each subclass; **Figure S4.** Distribution of the distances (AA) between the HuH and Y motifs in proteins of each subclass; **Figure S5.** Genome size and the occurrence of $\mathrm{TnpA}_{\mathrm{Y1}}$; **Figure S6.** Short repeated sequences in subclass 2.4 $\mathrm{TnpA}_{\mathrm{REP}}$; **Figure S7.** Homology modelling of proteins with extra domains in the conserved core domain; **Figure S8.** C-terminal subclass specific domains; **Figure S9.** Additional subclass specific domains; **Figure S10.** Subclass 2.4, REP insertions in coding sequences; **Figure S11.** Analysis of conservation in subclasses of the key residues involved in 5′ GTAG guide sequence binding; **Figure S12.** Number of intra genomic copies of each subclass; **Figure S13.** Percentage of sequence alignment identities between pairs of $\mathrm{TnpA}_{\mathrm{Y1}}$ sequences. (DOCX 2899 kb)

Additional file 2: Table S1. REP manually identified in REPtrons. (XLSX 66 kb)

Abbreviations

AA: Amino acids; BIME: Bacterial interspersed mosaic element; C-term: End of a protein or polypeptide; HGT: Horizontal gene transfer; nt: Nucleotides; N-term: Start of a protein or polypeptide; RAYT: REP-Associated tYrosine Tansposases; REP: Repeated Extragenic Palindromes; REPIN: REP doublets forming hairpIN; TIRYT: Terminal Inverted Repeat associated tYrosine Transposase

Acknowledgements

The authors thank Bao Ton-Hoang for discussions.

Funding

This work was supported by intramural funding from the French Centre National de la Recherche Scientifique (to Y.Q., P.S., M.C. and G.F.) and by grants from the French Agence National de Recherche (ANR-12-BSV8–0009-01; partners Y.Q., P.S., M.C. and G.F.)

Availability of data and materials

The sequence data as well as the phylogeny data, including alignments, used and/or analyzed during the current study are available in the TreeBASE repository, http://purl.org/phylo/treebase/phylows/study/TB2:S22805.

Authors' contributions

YQ, PS, and MC made substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data. YQ, PS, MC and GF have been involved in drafting the manuscript or revising it critically for important intellectual content. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 13 March 2018 Accepted: 29 May 2018
Published online: 19 June 2018

References

1. Chandler M, de la Cruz F, Dyda F, Hickman AB, Moncalian G, Ton-Hoang B. Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. Nat Rev Microbiol. 2013;11:525–38.
2. Kornberg A, Baker TA. DNA replication. 2nd ed. New York: W.H. Freeman; 1992.
3. Hickman AB, Ronning DR, Kotin RM, Dyda F. Structural unity among viral origin binding proteins: crystal structure of the nuclease domain of adeno-associated virus Rep. Mol Cell. 2002;10:327–37.
4. Hasselmayer O, Nitsche C, Braun V, von Eichel-Streiber C. The IStron CdISt1 of Clostridium difficile: molecular symbiosis of a group I intron and an insertion element. Anaerobe. 2004;10:85–92.
5. Ton-Hoang B, Siguier P, Quentin Y, Onillon S, Marty B, Fichant G, et al. Structuring the bacterial genome: Y1-transposases associated with REP-BIME sequences. Nucleic Acids Res. 2012;40:3596–609. Available from: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkr1198. Cited 19 June 2014
6. Ilyina TV, Koonin EV. Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. Nucleic Acids Res. 1992;20:3279–85.
7. Koonin EV, Ilyina TV. Computer-assisted dissection of rolling circle DNA replication. Biosystems. 1993;30:241–68.
8. Mahillon J, Chandler M. Insertion sequences. Microbiol Mol Biol Rev MMBR. 1998;62:725–74.
9. Ronning DR, Guynet C, Ton-Hoang B, Perez ZN, Ghirlando R, Chandler M, et al. Active site sharing and subterminal hairpin recognition in a new class of DNA transposases. Mol Cell. 2005;20:143–54.
10. Barabas O, Ronning DR, Guynet C, Hickman AB, Ton-Hoang B, Chandler M, et al. Mechanism of IS200/IS605 family DNA transposases: activation and transposon-directed target site selection. Cell. 2008;132:208–20.
11. He S, Corneloup A, Guynet C, Lavatine L, Caumont-Sarcos A, Siguier P, et al. The IS200/IS605 family and "peel and paste" single-strand transposition mechanism. Microbiol Spectr. 2015;3(4). PMID:26350330. https://doi.org/10.1128/microbiolspec.MDNA3-0039-2014.
12. Ton-Hoang B, Guynet C, Ronning DR, Cointin-Marty B, Dyda F, Chandler M. Transposition of ISHp608, member of an unusual family of bacterial insertion sequences. EMBO J. 2005;24:3325–38.
13. Guynet C, Hickman AB, Barabas O, Dyda F, Chandler M, Ton-Hoang B. In vitro reconstitution of a single-stranded transposition mechanism of IS608. Mol Cell. 2008;29:302–12.
14. Guynet C, Achard A, Hoang BT, Barabas O, Hickman AB, Dyda F, et al. Resetting the site: redirecting integration of an insertion sequence in a predictable way. Mol Cell. 2009;34:612–9.
15. Hickman AB, James JA, Barabas O, Pasternak C, Ton-Hoang B, Chandler M, et al. DNA recognition and the precleavage state during single-stranded DNA transposition in D. radiodurans. EMBO J. 2010;29:3840–52.
16. Pasternak C, Ton-Hoang B, Coste G, Bailone A, Chandler M, Sommer S. Irradiation-induced Deinococcus radiodurans genome fragmentation

triggers transposition of a single resident insertion sequence. PLoS Genet. 2010;6:e1000799.
17. Ton-Hoang B, Pasternak C, Siguier P, Guynet C, Hickman AB, Dyda F, et al. Single-stranded DNA transposition is coupled to host replication. Cell. 2010;142:398–408.
18. He S, Hickman AB, Dyda F, Johnson NP, Chandler M, Ton-Hoang B. Reconstitution of a functional IS608 single-strand transpososome: role of non-canonical base pairing. Nucleic Acids Res. 2011;39:8503–12.
19. He S, Guynet C, Siguier P, Hickman AB, Dyda F, Chandler M, et al. IS200/IS605 family single-strand transposition: mechanism of IS608 strand transfer. Nucleic Acids Res. 2013;41:3302–13.
20. Pasternak C, Dulermo R, Ton-Hoang B, Debuchy R, Siguier P, Coste G, et al. ISDra2 transposition in Deinococcus radiodurans is downregulated by TnpB. Mol Microbiol. 2013;88:443–55.
21. Lavatine L, He S, Caumont-Sarcos A, Guynet C, Marty B, Chandler M, et al. Single strand transposition at the host replication fork. Nucleic Acids Res. 2016;44:7866–83.
22. Nunvar J, Huckova T, Licha I. Identification and characterization of repetitive extragenic palindromes (REP)-associated tyrosine transposases: implications for REP evolution and dynamics in bacterial genomes. BMC Genomics. 2010;11:44.
23. Messing SAJ, Ton-Hoang B, Hickman AB, McCubbin AJ, Peaslee GF, Ghirlando R, et al. The processing of repetitive extragenic palindromes: the structure of a repetitive extragenic palindrome bound to its associated nuclease. Nucleic Acids Res. 2012;40:9964–79. Available from: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gks741. Cited 19 June 2014
24. Gilson E, Perrin D, Hofnung M. DNA polymerase I and a protein complex bind specifically to E. coli palindromic unit highly repetitive DNA: implications for bacterial chromosome organization. Nucleic Acids Res. 1990;18:3941–52.
25. Boccard F, Prentki P. Specific interaction of IHF with RIBs, a class of bacterial repetitive DNA elements located at the 3′ end of transcription units. EMBO J. 1993;12:5019–27.
26. Espéli O, Boccard F. In vivo cleavage of Escherichia coli BIME-2 repeats by DNA gyrase: genetic characterization of the target and identification of the cut site. Mol Microbiol. 1997;26:767–77.
27. Khemici V, Carpousis AJ. The RNA degradosome and poly(A) polymerase of Escherichia coli are required in vivo for the degradation of small mRNA decay intermediates containing REP-stabilizers. Mol Microbiol. 2004;51:777–90.
28. Gilson E, Rousset JP, Clément JM, Hofnung M. A subfamily of E. coli palindromic units implicated in transcription termination? Ann Inst Pasteur Microbiol. 1986;137B:259–70.
29. Higgins CF, McLaren RS, Newbury SF. Repetitive extragenic palindromic sequences, mRNA stability and gene expression: evolution by gene conversion? A review. Gene. 1988;72:3–14.
30. Liang W, Rudd KE, Deutscher MP. A role for REP sequences in regulating translation. Mol Cell. 2015;58:431.
31. Bachellier S, Saurin W, Perrin D, Hofnung M, Gilson E. Structural and functional diversity among bacterial interspersed mosaic elements (BIMEs). Mol Microbiol. 1994;12:61–70.
32. Bertels F, Rainey PB. Within-genome evolution of REPINs: a new family of miniature mobile DNA in bacteria. Guttman DS, editor. PLoS Genet. 2011;7:e1002132. Available from: http://dx.plos.org/10.1371/journal.pgen.1002132. Cited 19 June 2014
33. Bertels F, Rainey PB. Curiosities of REPINs and RAYTs. Mob Genet Elem. 2011;1:262–8.
34. Rocco F, De Gregorio E, Di Nocera PP. A giant family of short palindromic sequences in Stenotrophomonas maltophilia. FEMS Microbiol Lett. 2010;308:185–92.
35. Loper JE, Hassan KA, Mavrodi DV, Davis EW, Lim CK, Shaffer BT, et al. Comparative genomics of plant-associated Pseudomonas spp.: insights into diversity and inheritance of traits involved in multitrophic interactions. Guttman DS, editor. PLoS Genet. 2012;8:e1002784. Available from: http://dx.plos.org/10.1371/journal.pgen.1002784. Cited 31 Jan 2018
36. Di Nocera P, De Gregorio E, Rocco F. GTAG- and CGTC-tagged palindromic DNA repeats in prokaryotes. BMC Genomics. 2013;14:522. Available from: http://www.biomedcentral.com/1471-2164/14/522. Cited 19 June 2014
37. Nunvar J, Licha I, Schneider B. Evolution of REP diversity: a comparative study. BMC Genomics. 2013;14:385.
38. Bertels F, Gallie J, Rainey PB. Identification and characterization of domesticated bacterial transposases. Genome Biol Evol. 2017;9:2110–21.
39. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, et al. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet. 2009;5:e1000344.
40. Bobay L-M, Ochman H. The evolution of bacterial genome architecture. Front Genet. 2017;8:72. Available from: http://journal.frontiersin.org/article/10.3389/fgene.2017.00072/full. Cited 13 Feb 2018
41. Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R. RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res. 2001;29:4724–35.
42. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 2006;34:D32–6.
43. Siguier P, Varani A, Perochon J, Chandler M. Exploring bacterial insertion sequences with ISfinder: objectives, uses, and future developments. Methods Mol Biol. 2012;859:91–103.
44. Siguier P, Gourbeyre E, Chandler M. Bacterial insertion sequences: their genomic impact and diversity. FEMS Microbiol Rev. 2014;38:865.
45. Siguier P, Gourbeyre E, Chandler M. Everyman's guide to bacterial insertion sequences. Microbiol Spectr. 2015;3:MDNA3-0030-2014. Available from: http://www.asmscience.org/content/journal/microbiolspec/10.1128/microbiolspec.MDNA3-0030-2014. Cited 16 Feb 2018
46. Schaper S, Messer W. Interaction of the initiator protein DnaA of Escherichia coli with its DNA target. J Biol Chem. 1995;270:17622–6.
47. Aranda-Olmedo I, Tobes R, Manzanera M, Ramos JL, Marqués S. Species-specific repetitive extragenic palindromic (REP) sequences in Pseudomonas putida. Nucleic Acids Res. 2002;30:1826–33.
48. Feil H, Feil WS, Chain P, Larimer F, DiBartolo G, Copeland A, et al. Comparison of the complete genome sequences of Pseudomonas syringae pv. syringae B728a and pv. tomato DC3000. Proc Natl Acad Sci. 2005;102:11064–9. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0504930102. Cited 5 Nov 2014
49. Tobes R, Pareja E. Repetitive extragenic palindromic sequences in the Pseudomonas syringae pv. tomato DC3000 genome: extragenic signals for genome reannotation. Res Microbiol. 2005;156:424–33.
50. Silby MW, Cerdeño-Tárraga AM, Vernikos GS, Giddens SR, Jackson RW, Preston GM, et al. Genomic and genetic analyses of diversity and plant interactions of Pseudomonas fluorescens. Genome Biol. 2009;10:R51.
51. Enright AJ. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002;30:1575–84. Available from: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/30.7.1575. Cited 15 Sept 2015
52. Griffith F. The significance of pneumococcal types. J Hyg (Lond). 1928;27:113–59. Available from: http://www.journals.cambridge.org/abstract_S0022172400031879. Cited 28 Sept 2017
53. Lederberg J, Tatum EL. Gene recombination in Escherichia coli. Nature. 1946;158:558.
54. Zinder ND, Lederberg J. Genetic exchange in salmonella. J Bacteriol. 1952;64:679–99.
55. Marrs B. Genetic recombination in Rhodopseudomonas capsulata. Proc Natl Acad Sci U S A. 1974;71:971–3.
56. Lang AS, Zhaxybayeva O, Beatty JT. Gene transfer agents: phage-like elements of genetic exchange. Nat Rev Microbiol. 2012;10:472–82.
57. Lawrence JG, Hendrickson H. Lateral gene transfer: when will adolescence end? Mol Microbiol. 2003;50:739–49.
58. Page RD, Charleston MA. Trees within trees: phylogeny and historical associations. Trends Ecol Evol. 1998;13:356–9.
59. Fricker AD, Peters JE. Vulnerabilities on the lagging-strand template: opportunities for mobile elements. Annu Rev Genet. 2014;48:167–86.
60. Duderstadt KE, Reyes-Lamothe R, van Oijen AM, Sherratt DJ. Replication-fork dynamics. Cold Spring Harb Perspect Biol. 2014;6:a010157.
61. Gómez MJ, Díaz-Maldonado H, González-Tortuero E, López de Saro FJ. Chromosomal replication dynamics and interaction with the β sliding clamp determine orientation of bacterial transposable elements. Genome Biol Evol. 2014;6:727–40.
62. Dalrymple BP, Kongsuwan K, Wijffels G, Dixon NE, Jennings PA. A universal protein-protein interaction motif in the eubacterial DNA replication and repair systems. Proc Natl Acad Sci U S A. 2001;98:11627–32.
63. Bunting KA, Roe SM, Pearl LH. Structural basis for recruitment of translesion DNA polymerase Pol IV/DinB to the beta-clamp. EMBO J. 2003;22:5883–92.
64. López de Saro FJ, Georgescu RE, Goodman MF, O'Donnell M. Competitive processivity-clamp usage by DNA polymerases during DNA replication and repair. EMBO J. 2003;22:6408–18.

65. Parks AR, Li Z, Shi Q, Owens RM, Jin MM, Peters JE. Transposition into replicating DNA occurs through interaction with the processivity factor. Cell. 2009;138:685–95.
66. Curcio MJ, Derbyshire KM. The outs and ins of transposition: from mu to kangaroo. Nat Rev Mol Cell Biol. 2003;4:865–77.
67. Turlan C, Loot C, Chandler M. IS911 partial transposition products and their processing by the Escherichia coli RecG helicase. Mol Microbiol. 2004;53:1021–33.
68. Jang S, Sandler SJ, Harshey RM. Mu insertions are repaired by the double-strand break repair pathway of Escherichia coli. PLoS Genet. 2012;8:e1002642.
69. Eddy SR. Accelerated profile HMM searches. Pearson WR, editor. PLoS Comput Biol. 2011;7:e1002195. Available from: http://dx.plos.org/10.1371/journal.pcbi.1002195. Cited 26 June 2017
70. Van Dongen S. Graph Clustering Via a Discrete Uncoupling Process. SIAM J Matrix Anal Appl. 2008;30:121–41. Available from: http://epubs.siam.org/doi/abs/10.1137/040608635. Cited 1 Sept 2014
71. Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. Hidden Markov models in computational biology. J Mol Biol. 1994;235:1501–31. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0022283684711041. Cited 26 June 2017
72. Durbin R, editor. Biological sequence analysis: probabalistic models of proteins and nucleic acids. Cambridge: Cambridge University Press; 1998.
73. Eddy SR. Profile hidden Markov models. Bioinforma Oxf Engl. 1998;14:755–63.
74. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins. 1997;28:405–20.
75. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44:D279–85.
76. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013;30:772–80. Available from: http://mbe.oxfordjournals.org/cgi/doi/10.1093/molbev/mst010. Cited 15 Sept 2015
77. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25:1189–91. Available from: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btp033. Cited 15 Sept 2015
78. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinforma Oxf Engl. 2009;25:1972–3.
79. Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. Bioinforma Oxf Engl. 2005;21:2104–5.
80. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 2003;52:696–704.
81. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59:307–21.
82. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinforma Oxf Engl. 2007;23:127–8.
83. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44:W242–5. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw290. Cited 26 June 2017
84. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37:W202–8.
85. Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF. RNAalifold: improved consensus structure prediction for RNA alignments. BMC Bioinformatics. 2008;9:474. Available from: https://doi.org/10.1186/1471-2105-9-474
86. Guex N, Peitsch MC, Schwede T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective. Electrophoresis. 2009;30(Suppl 1):S162–73.