ELSEVIER

Insertion sequences in prokaryotic genomes

Patricia Siguier, Jonathan Filée and Michael Chandler

Insertion sequences (ISs) are small DNA segments that are often capable of moving neighbouring genes. Over 1500 different ISs have been identified to date. They can have large and spectacular effects in shaping and reshuffling the bacterial genome. Recent studies have provided dramatic examples of such IS activity, including massive IS expansion during the emergence of some pathogenic bacterial species and the intimate involvement of ISs in assembling genes into complex plasmid structures. However, a global understanding of their impact on bacterial genomes requires detailed knowledge of their distribution across the eubacterial and archaeal kingdoms, understanding their partition between chromosomes and extra-chromosomal elements (e.g. plasmids and viruses) and the factors which influence this, and appreciation of the different transposition mechanisms in action, the target preferences and the host factors that influence transposition. In addition, defective (non-autonomous) elements, which can be complemented by related active elements in the same cell, are often overlooked in genome annotations but also contribute to the evolution of genome organisation.

Addresses
Laboratoire de Microbiologie et Génétique Moléculaires (UMR5100 CNRS) Campus Université Paul Sabatier 118, Route de Narbonne, F-31062 Toulouse Cedex, France

Corresponding author: Chandler, Michael (mike@ibcg.biotoul.fr)

Current Opinion in Microbiology 2006, **9**:526–531

This review comes from a themed issue on
Genomics
Edited by David W Ussery and Timothy D Read

Available online 28th August 2006

1369-5274/$ – see front matter

DOI 10.1016/j.mib.2006.08.005

Introduction

The impact of transposable genetic elements (TEs) on bacterial genomes was first appreciated several decades ago, with the discovery of their mutagenic properties and their capacity to transport accessory genes with them, at that time largely the sequestration and transmission of genes conferring resistance to antibiotics. Only a handful of TEs had been identified by 1989, and among these, only 50 bacterial insertion sequences (IS), arguably the simplest autonomous TEs containing a transposase gene flanked by inverted terminal DNA repeats (IRs), were identified. Although it was known that some bacteria harboured many hundreds of ISs [1], it was only with the advent of whole genome sequencing that the full extent of their distribution, diversity and (sometimes) high copy number was appreciated. Today a minimal but incomplete repertoire deposited in the ISfinder database (http://www-is.biotoul.fr) includes 1500 different ISs in over 295 eubacterial and archaeal species.

ISs are classified into about 20 families [2] on the basis of various shared features. An IS family is defined as a group of ISs with related transposases, strong conservation of the catalytic site, conservation of organisation and similar IRs. These divisions are rather subjective and the families will certainly change over time. Many families are relatively homogenous. Others such as IS*4* and IS*5* were initially less homogenous but must be redefined as more members are identified. The characteristics of these family groups are described in Table 1. Most specify transposase enzymes of the DDE group (named for the three amino acids, Asp, Asp and Glu involved in coordinating divalent metal ions involved in the chemical reactions required for transposition) although a growing number are composed of members whose transposases use different chemistries (e.g. tyrosine or serine nucleophiles).

Distribution within prokaryote genomes

ISs are found in most but not all eubacterial and archaeal genomes. The distribution of ISs in the ISfinder database is shown in Figure 1. This includes a large number of ISs from individual sequence files at the National Center for Biotechnology Information (NCBI), results of a complete analysis of all 27 archaeal genomes together with 144 plasmids, but only 20 of 325 eubacterial genomes (Table 2). This sample of ISs does not therefore represent a complete picture; in addition to our sampling bias of available genomes, bias is also introduced by the choice of genomes to be sequenced. For example, all IS families are present in the enterobacteria, and many in other members of the γ-proteobacteria. This is probably a much more complete picture than for certain other phyla because, historically, more attention has been focused on this group.

Partition between chromosomes and plasmids

Naturally occurring bacterial plasmids are generally non-essential accessory genetic elements, ranging in size from several kilobases to hundreds of kilobases. An initial survey of a 'random' plasmid sample indicates that the percentage of IS DNA in plasmids smaller than ~20 kb is generally zero. There is an abrupt increase within

Table 1

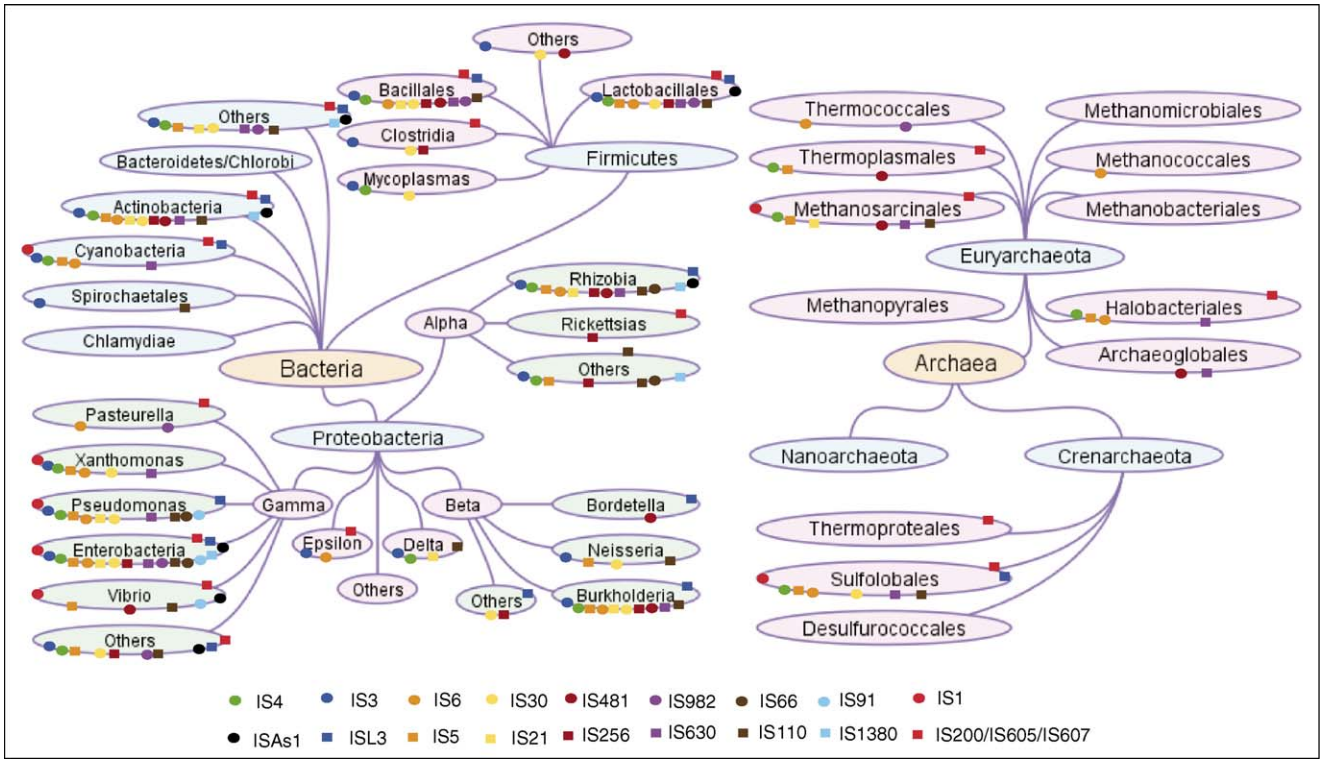
Some characteristics of the IS families.

Family[a]	Groups[b]	Size range[c]	DR[d]	Ends[e]	IRs[f]	ORFs[g]	Comments[h]
IS*1*	-	770	9(8–11)	GG(T)	Y	2	DD(35)E; –1 frameshift
IS*3*	IS*2*	1300–1350	5	TG(A/T)	Y	2	DD-E; –1 frameshift
	IS*3*	1200–1300	3(4)				
	IS*51*	1300–1400	3(4)				
	IS*150*	1400–1550	3–5				
	IS*407*	1200–1250	4				
IS*4*	-	1300–1950	9–12	C(A)	Y	1	DD-E
IS*5*	IS*5*	1100–1350	4	GG	Y	1	DD-E
	IS*427*	800–1000	2(4)	GA/G		(2)	Possible frameshift
	IS*903*	1000–1100	9	GGC		1	
	IS*1031*	850–950	3	GAG		1	
	IS*H1*	900–1150	8	-		1	Archaea-specific
	IS*L2*	800–1100	2–3	-		1	
IS*6*	-	750–900	8	GG	Y	1	DD(34)E
IS*21*	-	1950–2500	4(5,8)	TG	Y	2	DD-E
IS*30*	-	1000–1250	2–3	-	Y	1	DD(33)E
IS*66*	-	2500–2700	8	GTA	Y	3	-
IS*91*	-	1500–1850	0	-	N	1	ss-DNA Rep
IS*110*	-	1200–1550	0	-	N	1	DD-E (?)
IS*200/IS605*	IS*200*	600–800	0	-	N	1	Relaxase (Y)
	IS*605*	1500–2300	0	-	N	2	Relaxase (Y)
	IS*607*	1350–2600	0	-	N	2	Recombinase - S
IS*256*	-	1300–1500	8–9	GG/A	Y	1	DD-E
IS*481*	-	950–1100	5–6	TGT(A/G)	Y	1	DD-E
IS*630*	-	1100–1200	2	-	Y	1	DD-E
IS*982*	-	1000	(7)	AC(C/G)	Y	1	DD-E
IS*1380*	-	1650	4	CC/G	Y	1	DD-E
IS*As1*	-	1200–1350	8	C	Y	1	DD-E?
IS*L3*	-	1300–1550	8	GG	Y	1	DD-E
Tn*3*	-	>3000	5	GGGG	-	-	DD-E

[a]major IS families; [b]subgroups where defined; [c]typical length in base pairs; [d]length of direct repeats (DR) generated on insertion in base pairs; [e]most frequent sequence found at the ends; [f]presence of IRs; [g]number of ORFs involved in transposition; [h]transposase types.
The numbers in parentheses represent [d]the length range of the DR, and [h]the number of amino acid residues separating the D from the E. [e]The letters in parentheses represent highly but not completely conserved amino acids.
We have not included the emerging divisions in the IS4 family nor the emerging families from both eubacteria and archaea. These will be presented and described in detail elsewhere.

plasmids above this size threshold, averaging between 5–15%, reaching 20% in some cases and, in one extreme case (pW100; see *IS modules* below) exceeding 40%. The most trivial explanation is that a minimal plasmid gene set representing a 'minimal' DNA length is necessary for plasmid viability. Plasmids below a certain size range would, by definition, not be able to include insertions. A related explanation is that the smaller plasmids would tend to carry fewer nonessential genes and therefore be more recalcitrant to insertions. A more interesting explanation is that plasmids capable of self-transfer between various species and genera — and thus are larger because they carry transfer functions — would be capable of acquiring transposons, ISs and accessory genes during passage between different host genomes. Indeed, there is some evidence that certain transposable elements (TEs) specifically target transmissible plasmids (see below). In the case of bacterial chromosomes, however, the density of ISs is generally below 3%, except in a few rare cases (see below, *IS expansion and genome reduction*). Plasmid properties that distinguish them from bacterial chromosomes can include high copy numbers, differences in vegetative replication (e.g. replication by rolling-circle mechanisms), self-transmissibility, which also generally occurs using rolling-circle replication, or the capacity to be mobilized by self-transmissible plasmids. Some of these features might specifically suit TEs. Thus in one of its two transposition pathways, transposon Tn7 preferentially inserts into transmissible plasmids in an orientation-specific manner. Its transposition apparatus targets the transposon to the DNA ends provided, for example, by Okazaki fragments generated on the lagging-strand of DNA synthesis during plasmid transfer, in which, unlike in chromosomal replication, leading- and lagging-strand DNA synthesis are physically separated. Indeed, one Tn7 transposition protein binds to 3′ recessed DNA ends. Tn7 is also attracted to artificially induced double-strand chromosomal breaks and to distorted DNA [3]. Related results have been obtained for IS*903* (IS*5* family). For this IS, insertion into a transmissible plasmid showed a

Figure 1



Distribution of IS families in the Eubacteria and Archaea. The different families are colour-coded. This is only a partial distribution extracted from ISs present in the ISfinder database (http://www-IS.biotoul.fr).

pronounced orientation preference. This was lost if the transfer origin of the plasmid was inverted [4].

The authors are not aware of similar studies with other TEs but think that these types of approaches are essential to understand the impact of TEs in shaping genomes.

Another area which has received little attention is genomic regional specificity. Some bacterial genomes exhibit distinct IS clusters (data from ISfinder). These might result from horizontal acquisition of blocks of DNA, reflect TE exclusion from other regions, result from TE extinction or stem from TE insertion specificity. Tn*7* has again provided an interesting model: studies have revealed its preferential insertion into the terminus region of the chromosome. This is a region that exhibits high recombination activity, possibly indicating DNA fragility [5]. Moreover, the orientation depends on the direction of chromosome replication [3]. IS*903* also shows distinct regional preferences in insertion into the *E. coli* chromosome. These are significantly less marked in the absence of the histone-like nucleoid structuring protein H–NS [6]. These types of effect require further examination not only for IS*903* but also for other IS paradigms.

Table 2

Number of genomes and plasmid sequences in the public databases and those analysed in ISfinder.

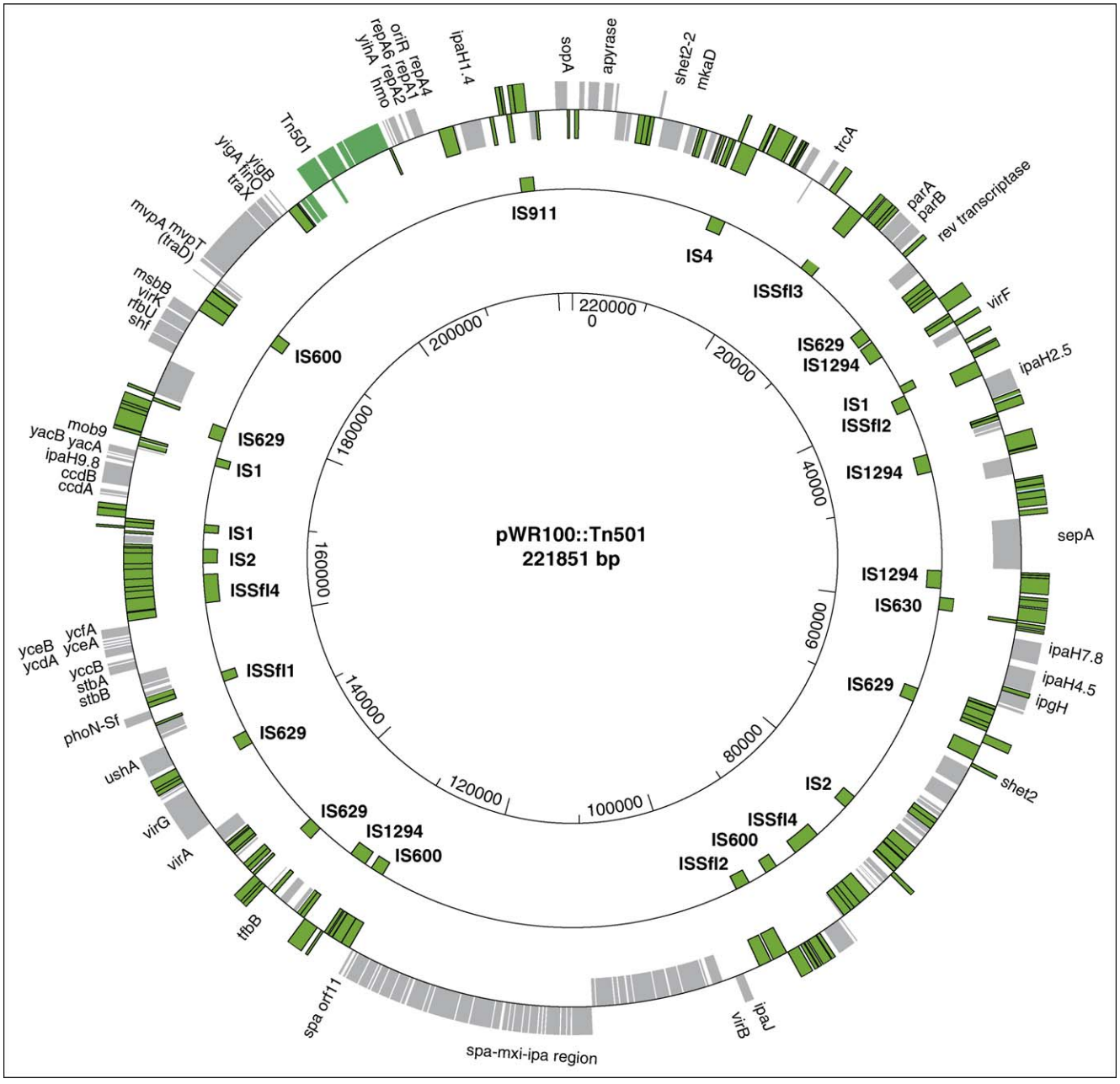
Genomes	Sequenced	Analysed for ISs	
Bacteria		**Finished**	**In progress**
Chromosome	344	21	24
Plasmid	759	101	18
Draft Assembly	237	-	-
Archaea			
Chromosome	28	28	-
Plasmid	41	43	-
Draft Assembly	2	-	2

IS impact on genomes: IS gene assembly versus IS expansion

IS modules

Identification and localization of full-length and fragmented ISs on plasmids clearly implies that they are involved in creating modular assemblies of genes (the simplest being concatenation within compound transposons). A good example is the 221 kb virulence megaplasmid of *Shigella flexneri*, pW100 (Figure 2) [7,8]. IS material represents 46% of the plasmid including 26 full-length ISs and an extensive array of IS fragments indicative of scars from ancestral rearrangements. Similar assemblies are also detected in other megaplasmids including those from

Figure 2



Map of the *Shigella flexneri* virulence plasmid pW100. This is derived from the data of Buchreiser *et al.* [7] and Ventkatesan *et al*. [8]. Note IS elements are highlighted in green and boxed by black lines. Other genes are shown in grey. The outer circle shows both full length ISs and fragments. Boxes inside and outside the circle distinguish IS and gene orientation. The copy of Tn501 is not a natural resident. It was introduced as a marker by Ventkatesan *et al.* [8]. The second circle show the full length IS copies only. The inner circle shows the plasmid coordinates in base pairs.

the Rhizobiaceae (see [9]). Indeed, these are probably responsible for multiple plasmid-chromosome exchanges [10].

IS expansion and genome reduction

Genome comparisons of strains from a single species or genera have provided important information concerning genome evolution and the emergence of pathogenicity. The role of ISs in this process was illustrated in a key publication comparing members of the Bordetellae: *Bordetella bronchiseptica*, *Bordetella parapertussis* and *Bordetella pertussis* [11•]. Phylogeny results suggested that *B. bronchiseptica* was the ancestral *Bordetella* species. *B. bronchiseptica* possesses the largest genome (5.34 Mb) of the three, and carries no known ISs, but does have several prophages. In comparison, *B. parapertussis* has lost all

prophages but has acquired many ISs (22 copies of IS*1001* and 90 of IS*1002*, both ISL3 family members; Table 1) and its genome is only 4.77 Mb in size. *B. pertusis*, the agent of whooping cough, has an even smaller genome (4.1 Mb), with more than 260 ISs (six copies of IS*1002* and 17 of IS*1663*, both members of the IS*110* family; Table 1) and over 238 copies of IS*481*. A comparison of synteny between these genomes suggested that the ISs have provoked a considerable degree of genome rearrangement in addition to the reduction in genome size, and this is accompanied by a high level of gene inactivation [11•]. Many *B. bronchiseptica* genes, absent in either *B. parapertussis* or *B. pertussis*, are involved in gene regulation or in surface structure synthesis (probably involving surface antigens). Parkhill *et al.* ([11•] and personal communication) proposed that such IS expansions result from an evolutionary bottleneck resulting from population isolation.

A similar, but less extreme, scenario can be painted for the Yersinae [12,13], where the relationship between *Yersinia pseudotuberculosis* and the different *Yersinia pestis* pathovars (Antiqua, Medievalis and Orientalis). Medievalis (KIM10+) is similar to that of the Bordetellae. Also, Orientalis (CO92) strains show amplification of IS*100* (a member of the IS*21* family), IS*285* (a member of the IS*256* family), and to a lesser extent IS*1661* (a member of the IS*3* family) and IS*1541*, and IS285 (a member of the distinct, non-DDE, IS*200* family). Other potential examples of this can be found and in *E. coli* and the Shigellae and in the Archaea. For example, *Sulfolobus solfataricus* carries at least 130 full IS copies and at least 200 partial IS copies, whereas the related *Sulfolobus acidocaldarius* carries only a few partial IS copies [14] (J Filée, P Siguier and M Chandler, submitted).

Whether such IS expansion occurs in a step-by-step process, or by a global activation of transposition remains to be determined.

IS extinction

Analysis of a diverse set of 18 bacterial genomes revealed that the intra-genomic sequence diversity of a given IS is very low, suggesting that most ISs in an individual genome are evolutionarily young and might have been recently acquired [15•]. This observation might be explained if there was a period of IS expansion followed by a series of IS 'extinctions' in bacterial lineages. This implies that ISs could bring transitory selective advantages to their host, such as lateral gene transfers and genomic rearrangements, but might be detrimental to their host in the long term [15•].

MITES

Miniature inverted repeat transposable elements (MITES), first defined in plant genomes [16], are generally less than 300 by transposases of full-length parental genomic copies. Many eukaryotic MITES are related to the Tc-mariner family elements (distantly related to bacterial IS*630* family). IS*630*-related MITES were the first described bacterial examples [17–19]. MITES showing similarities to other IS families have now been observed in the Eubacteria and Archaea [20] (J Filée, P Siguier and M Chandler, unpublished results). These include IS*1*, IS*4*, IS*5* and IS*6* among ISs with DDE transposases and IS*200*/IS*605* among elements with other types of transposase. Full-length copies the parental IS might not be available or might be so divergent as to escape detection in a standard BLAST (basic local alignment search tool) analysis. This is the case for certain MITES from the Archaea [20] and will probably also be true for the Eubacteria. The detection and analysis of MITES is therefore arduous. We note also that our analyses have revealed a significant number of solo IS-related IRs in various genomes (ISfinder). At present the overall impact of these elements in shaping bacterial genomes is difficult to assess.

Certain MITES are able to influence gene expression. The two types of IS*630*-related MITES identified in *Neisseria* species carry outward-directed promoters [18]. In one type the promoters are constitutive. Members of the other type carry a functional integration host factor (IHF) site and the promoters are regulated negatively by IHF (N Buisine and R Chalmers, personal communication). Interestingly, IS*630*-related MITES (IS*Stso1* and IS*Spn2*) from *Streptococcus pneumoniae* and *Streptococcus mitis* have been recently identified in intergenic regions. These elements, called 'BOX', affect gene expression when placed upstream or downstream of the gene [21], possibly by changing mRNA stability. Also, other classes of highly repeated elements with similar structures to MITES have also been identified. These include REP (repetitive extragenic palindrome) sequences in the Enterobacteriacae, known to influence mRNA stability and transcription termination, as well as carrying binding sites for IHF and gyrase [22]. Whereas no equivalent full length IS has yet been observed, it is possible that REP sequences are also MITEs.

Conclusions — perspectives

We have attempted to provide a short integrated overview of the diversity, distribution and activity of ISs and their derivatives within prokaryotic genomes. This is only a partial picture because we only have the full annotations of a small fraction of the available eubacterial genomes. In many cases these annotations are only approximate and it is rare that the entire IS (with both its ends) is included. One major problem in assessing the impact of ISs lies in the accurate identification of DNA components. Generally, genome annotations successfully identify full-length transposase genes, although these are sometimes mislabelled as integrases. The availability of the ends of the corresponding element is less common and a full

annotation of IS-related DNA fragments including those without open reading frames (ORFs; e.g. pW100; Figure 2) is quite rare. This is perhaps not surprising because automated annotation protocols are laborious and at present the results require a meticulous assessment by the human eye. Moreover, it is clear that additional types of TE remain to be identified. Recent additions include the IS*200*/IS*605*/IS*607* group [23] and IS*CR* elements (a newly defined group of elements related to IS91, thought to transpose using a rolling circle replication mechanism and able to transport downstream flanking genes) [24]. In this light we believe that a collaborative effort to compile all the identified ISs into the ISFinder database is crucial. We have also tried to highlight areas of investigation which we think are important to address experimentally: these include the relative attractiveness of plasmids and chromosomes for IS insertion, as well as the properties of the growing numbers of MITEs, MITE-like elements and solo-IRs.

Acknowledgements

We would like to thank members of the Chandler laboratory, B Tong-Hoang, P Rouseau, G Duval-Valentin, C Guynet, N Pouget and E Gueguen, for fruitful discussion. Intramural funding was provided by the Centre National de la Recherche Scientifique (CNRS; France) and extramural funding by European contract: LSHM-CT-2005-019023. JF was supported by the CNRS and by the European contract.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Ohtsubo H, Nyman K, Doroszkiewicz W, Ohtsubo E: **Multiple copies of iso-insertion sequences of IS1 in *Shigella dysenteriae* chromosome**. *Nature* 1981, **292**:640-643.

2. Chandler M, Mahillon J: **Insertion sequences revisited**. In *Mobile DNA*, vol II. Edited by Craig NL, Craigie R, Gellert M, Lambowitz A. ASM press; 2002:305-366.

3. Peters JE, Craig NL: **Tn7: smarter than we thought**. *Nat Rev Mol Cell Biol* 2001, **2**:806-814.

4. Hu WY, Derbyshire KM: **Target choice and orientation preference of the insertion sequence IS903**. *J Bacteriol* 1998, **180**:3039-3048.

5. Louarn J, Kuempel PL, Cornet F: **The terminus region pf the *Escherichia coli* chromosome, or, All's Well That Ends Well**. In *The Bacterial Chromosome.* Edited by Higgins NP. ASM Press; 2005.

6. Swingle B, O'Carroll M, Haniford D, Derbyshire KM: **The effect of host-encoded nucleoid proteins on transposition: H-NS influences targeting of both IS903 and Tn10**. *Mol Microbiol* 2004, **52**:1055-1067.

7. Buchrieser C, Glaser P, Rusniok C, Nedjari H, D'Hauteville H, Kunst F, Sansonetti P, Parsot C: **The virulence plasmid pWR100 and the repertoire of proteins secreted by the type III secretion apparatus of *Shigella flexneri***. *Mol Microbiol* 2000, **38**:760-771.

8. Venkatesan MM, Goldberg MB, Rose DJ, Grotbeck EJ, Burland V, Blattner FR: **Complete DNA sequence and analysis of the large virulence plasmid of *Shigella flexneri***. *Infect Immun* 2001, **69**:3271-3285.

9. Freiberg C, Fellay R, Bairoch A, Broughton WJ, Rosenthal A, Perret X: **Molecular basis of symbiosis between *Rhizobium* and legumes**. *Nature* 1997, **387**:394-401.

10. Mavingui P, Flores M, Guo X, Davila G, Perret X, Broughton WJ, Palacios R: **Dynamics of genome architecture in *Rhizobium* sp. strain NGR234**. *J Bacteriol* 2002, **184**:171-176.

11. • Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, Harris DE, Holden MT, Churcher CM, Bentley SD, Mungall KL *et al.*: **Comparative analysis of the genome sequences of *Bordetella pertussis, Bordetella parapertussis* and *Bordetella bronchiseptica***. *Nat Genet* 2003, **35**:32-40.
This is a spectacular demonstration of IS expansion.

12. Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MT, Prentice MB, Sebaihia M, James KD, Churcher C, Mungall KL *et al.*: **Genome sequence of *Yersinia pestis*, the causative agent of plague**. *Nature* 2001, **413**:523-527.

13. Chain PS, Carniel E, Larimer FW, Lamerdin J, Stoutland PO, Regala WM, Georgescu AM, Vergez LM, Land ML, Motin VL *et al.*: **Insights into the evolution of Yersinia pestis through whole-genome comparison with *Yersinia pseudotuberculosis***. *Proc Natl Acad Sci USA* 2004, **101**:13826-13831.

14. Brugger K, Torarinsson E, Redder P, Chen L, Garrett RA: **Shuffling of *Sulfolobus* genomes by autonomous and non-autonomous mobile elements**. *Biochem Soc Trans* 2004, **32**:179-183.

15. • Wagner A: **Periodic extinctions of transposable elements in bacterial lineages: evidence from intragenomic variation in multiple genomes**. *Mol Biol Evol* 2006, **23**:723-733.
This is a population dynamics view of ISs in bacterial genomes.

16. Feschotte C, Zhang X, Wessler S: **Miniature inverted repeat transposable elements and their relationship to established DNA transposons**. In *Mobile DNA*, vol II. Edited by Craig NL, Craigie R, Gellert M, Lambowitz A. ASM press; 2002:1147-1158.

17. Correia A, Pisabarro A, Castro JM, Martin JF: **Cloning and characterization of an IS-like element present in the genome of *Brevibacterium lactofermentum* ATCC 13869**. *Gene* 1996, **170**:91-94.

18. Buisine N, Tang CM, Chalmers R: **Transposon-like Correia elements: structure, distribution and genetic exchange between pathogenic *Neisseria* sp**. *FEBS Lett* 2002, **522**:52-58.

19. Oggioni MR, Claverys JP: **Repeated extragenic sequences in prokaryotic genomes: a proposal for the origin and dynamics of the RUP element in *Streptococcus pneumoniae***. *Microbiology* 1999, **145**:2647-2653.

20. Brugger K, Redder P, She Q, Confalonieri F, Zivanovic Y, Garrett RA: **Mobile elements in archaeal genomes**. *FEMS Microbiol Lett* 2002, **206**:131-141.

21. Saluja SK, Weiser JN: **The genetic basis of colony opacity in *Streptococcus pneumoniae*: evidence for the effect of box elements on the frequency of phenotypic variation**. *Mol Microbiol* 1995, **16**:215-227.

22. Espeli O, Moulin L, Boccard F: **Transcription attenuation associated with bacterial repetitive extragenic BIME elements**. *J Mol Biol* 2001, **314**:375-386.

23. Kersulyte D, Mukhopadhyay AK, Shirai M, Nakazawa T, Berg DE: **Functional organization and insertion specificity of IS607, a chimeric element of *Helicobacter pylori***. *J Bacteriol* 2000, **182**:5300-5308.

24. Toleman MA, Bennett PM, Walsh TR: **ISCR elements: novel gene-capturing systems of the 21st century?** *Microbiol Mol Biol Rev* 2006, **70**:296-316.