Chapter 1

Introduction

1.1 Horizontal Gene Transfer

Perhaps very few themes in the study of microbial evolution have been as contentious as Horizontal Gene Transfer (HGT) (Kurland, 2000; Lawrence and Hendrickson, 2003). HGT is defined as the transfer of genetic material between a donor and a recipient, in which no asexual (or sexual) reproduction is involved; the donor need not be physically present. Early discussion on HGT came from Griffith, in a study focused on the ability of pneumococci to exchange genetic material through direct uptake of DNA from the environment (transformation) (Griffith, 1928); later on Anderson and Syvanen discussed the concept of gene transfer across species boundaries (Anderson, 1970; Syvanen, 1985).

HGT as a concept has fuelled very strong and ongoing debate about its impact, extent, gene and host repertoire affected and frequency throughout the evolution of species (Kurland, 2000; Lawrence and Hendrickson, 2003). The controversy stems mainly from the fact that HGT is a counterintuitive concept that threatens to reject (Doolittle and Papke, 2006; Gevers *et al.*, 2005; Lawrence, 2002) the universality of a very fundamental biological concept, that of the biological species (Mayr, 1942); furthermore it brings into question the Tree of Life (Darwin, 1859), i.e. the representation of the phylogenetic history and evolution of species through a strictly bifurcating tree-like structure.

In terms of its impact, views range (Lawrence and Hendrickson, 2003) from HGT being a valid but nonetheless rare mechanism of gene transfer with marginal impact on genome phylogeny (Kurland *et al.*, 2003), to HGT being a major driving force that enables accelerated microbial evolution, often referred to as "evolution in quantum leaps" (Groisman and Ochman, 1996); for example two single-step events of HGT enabled *Salmonella* to evade successfully the host defence mechanisms

and invade epithelial cells (Hacker *et al.*, 1997). Supporters of the first view put forward the idea that the evolutionary history of a species can still be reliably represented through a bifurcating tree-like structure that reflects mainly the organismal phylogeny (Daubin *et al.*, 2003; Kurland *et al.*, 2003; Lerat *et al.*, 2005; Woese, 2000) since HGT frequency is not high enough to obscure the true phylogenetic signal of a given species. Supporters of the second opinion, however, believe that HGT can obfuscate the organismal phylogenetic signal to such an extent (i.e. mosaic genomes that contain genes with different histories) that the reliable representation of the organismal phylogeny violates the strictly bifurcating structure of the Tree of Life; instead reticulate, network-like structures can more reliably represent the true phylogenetic relationships between species that extensively exchange genetic material (Doolittle, 1999; Gogarten and Townsend, 2005; Kunin *et al.*, 2005).

For example, two distantly related species that have extensively exchanged genetic material with each other, now having mosaic genomes with patches of DNA with different histories, will probably map (wrongly) on very close branches on the phylogenetic tree, since their phylogenetic histor(y)ies are forced to fit in a strictly binary (i.e. either they belong to the same species or not) classification system. On the other hand, acknowledging that mosaicism is a valid genomic state, we can allow genomes to belong to more than one species at the same time (Doolittle, 1999); under a phylogenetic network representation the same two genomes will map correctly on their respective species/genera branches but their extensive genetic exchange will also be taken into account, represented through multiple branches connecting the two lineages. It should be noted that similar results of genome mosaicism with patches of very similar DNA shared between very closely related taxa may also be attributed to genetic exchange via homologous recombination (Didelot *et al.*, 2007; Feil *et al.*, 2001).

An example that illustrates the extent of viable genomic mosaicism, and at the same time questions the true boundaries of the biological

species concept, comes from the model bacterial organism *Escherichia coli*; a three way comparison between the laboratory strain MG1655, the uropathogenic (UPEC) strain CFT073 and the enterohemorrhagic (EHEC) strain EDL933, shows that less than 40% of their common gene pool is shared between those three strains, although their high sequence similarity places them under the same species (Welch *et al.*, 2002).

At this point it may be useful to draw a parallel with quantum mechanics to discuss further the limitations of a binary classification system when describing complex biological processes. According to the classical Bohr model (Bohr, 1913) of the atom, electrons (in our case genomes) are allowed to belong only to one of the well-defined orbits (in our case species) around the nucleus. Later on, however, the quantum mechanics theory (Dirac, 1958) introduced a new, more realistic representation of the atom structure: the electrons surrounding the nucleus belong to a cloud (in our case phylogenetic network) of probable positions, rather than single well-defined orbits. The existence of the first atom model (in our case the tree of life) was due to our inability to study in a more detailed and realistic way the true structure of the atom (in our case the history of species); more sophisticated, non-binary methods bring a more realistic view in our understanding and modelling of the history of species evolution (Figure 1.1).

From the host point of view, the extent of HGT ranges from 0% in *Buchnera aphidicola* (Tamas *et al.*, 2002) to 24% in *Thermotoga maritima* (Nelson *et al.*, 1999); from the donor point of view, the extent of HGT might be up to 100%, i.e. whole genome transfer of a donor to a recipient cell (Hotopp *et al.*, 2007).

Examples of HGT events exist in all three domains of life, i.e. bacteria (Baumler, 1997; Lawrence and Ochman, 1997), archaea (Deppenmeier *et al.*, 2002; Gribaldo *et al.*, 1999) and eukaryota (Hotopp *et al.*, 2007), including humans, although the extent of HGT in the latter is not very well documented (Andersson *et al.*, 2001; Stanhope *et al.*, 2001).

In terms of gene repertoire, again HGT seems to affect a wide range of functional gene classes including genes encoding products involved in the translation machinery (e.g. aminoacyl-tRNA synthetases, ribosomal proteins) (Brochier *et al.*, 2000; Wolf *et al.*, 1999), ribosomal RNA (rRNA) genes (Nomura, 1999; Yap *et al.*, 1999), components of biosynthetic pathways (e.g. cytochrome c biogenesis system I and II) (Goldman and Kranz, 1998) and major metabolic components (e.g. glyceraldehyde-3-phosphate dehydrogenase) (Doolittle *et al.*, 1990); a good review on how HGT might have affected major metabolic pathways is given by Boucher (Boucher *et al.*, 2003). Although in theory all genes can be horizontally exchanged, some functional classes (e.g. operational genes) may be more frequently transferred than others (e.g. informational) (Jain *et al.*, 1999).

Estimates of the actual frequency of HGT events in microbial genomes exist and suggest that HGT can be indeed a very frequent mechanism of gene transfer. Lawrence and Ochman (Lawrence and Ochman, 1997) studying the effects of HGT in *E. coli* and *S. enterica* estimated the HGT rate to be 31 kb per million years (Myr); this rate is close to the frequency of DNA being introduced by point mutations. Applying this rate of HGT, the two sister lineages were predicted to have each gained and lost over 3Mb of alien DNA, since their divergence, approximately 100-140 million years (Myr) ago (Doolittle *et al.*, 1996; Ochman and Wilson, 1987).

Although horizontally acquired DNA enters a different, completely new genomic environment of a another host, the expression of horizontally acquired genes is not random or unrestrained; on the contrary the expression of alien DNA can be extremely sophisticated and fine-tuned. For example in *Salmonella* the quorum sensing mechanism that controls the cell population density directly affects the expression of genes that have been *en block* horizontally acquired under a single event (Choi *et al.*, 2007). Similarly SlyA, a virulence-related transcriptional regulator, participates in the regulation of another block of alien genes present in *S. enterica* (Linehan *et al.*, 2005). Recently a putative master regulator of the

expression of horizontally acquired DNA has been recognized in enterobacteriaceae (Navarre *et al.*, 2006): H-NS, a histone-like nucleoid structuring protein has been proposed to be responsible for selectively silencing horizontally acquired DNA of lower G+C% content relative the backbone composition of the host. It is worth noting that SlyA acts as an antagonist to H-NS, displacing the H-NS from promoter loci (Wyborn *et al.*, 2004), adding one extra level of complexity to the regulatory network controlling the expression of alien DNA in microbial genomes.

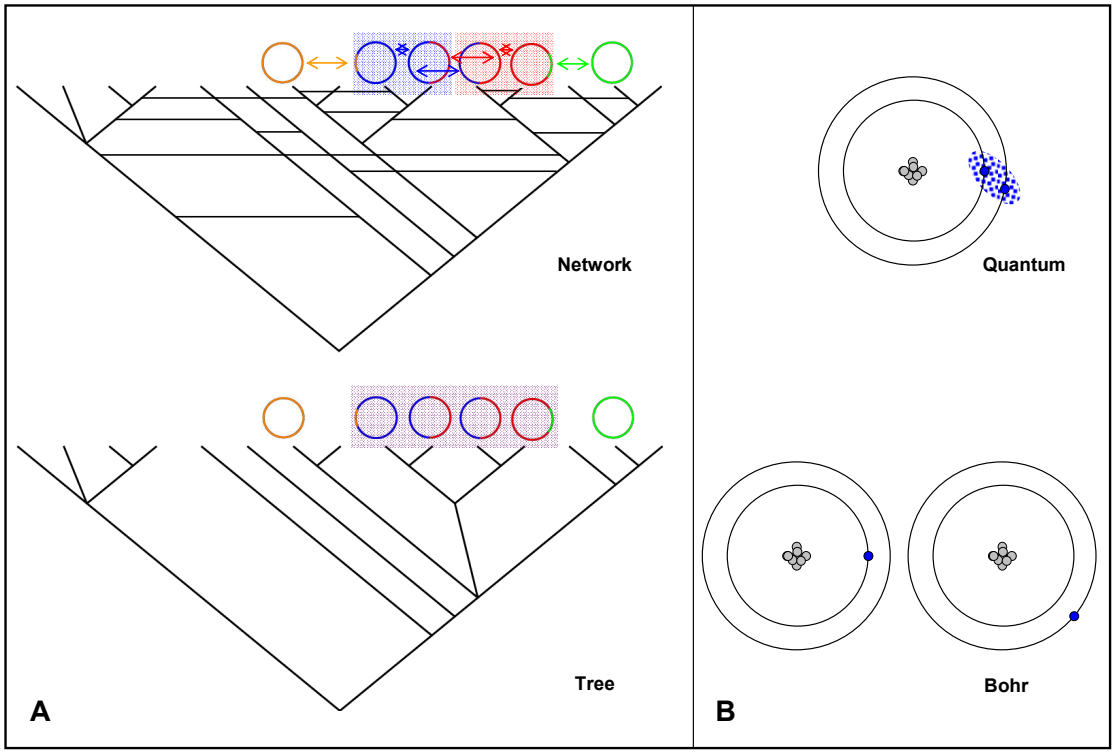

Figure 1.1: **A.** An example of genome mosaicism and the limitation of a bifurcating, tree-based classification system (bottom) for a reliable representation of the true phylogenetic histories of lineages exposed to high rates of genetic flux, compared to a phylogenetic network (top). **B.** Atom structure representation under the Bohr model (bottom) and the quantum theory (top).

There are three reported major mechanisms of HGT (Figure 1.2), namely transformation (Griffith, 1928), conjugation (Lederberg and Tatum, 1946) and transduction (Morse *et al.*, 1956). A major difference between conjugation and the other two types of gene transfer, in terms of the donor and the recipient, is that in transduction and transformation

there is no actual need for the donor to be physically present either in terms of time or in terms of space.

The recognition and uptake of naked DNA directly from the environment (transformation) is a widespread DNA transfer mechanism, present in many archael and bacterial species including Gram positive and Gram negative representatives (Lorenz and Wackernagel, 1994). In order for natural transformation to occur, a physiological state of competence must be reached; some bacteria species develop competence as a response to certain environmental changes whereas others, such as *Neisseria gonorrhoeae* and *Haemophilus influenzae* are constantly competent to accept naked DNA (Dubnau, 1999). Transformation in *Neisseria* and *H. influenzae* is selective and requires the presence of specific DNA Uptake Sequences (DUS) of approximately 10bp in length (Goodman and Scocca, 1988) that are scattered throughout the bacterial chromosome at frequencies up to 2,000 copies per chromosome (Parkhill *et al.*, 2000). Naked DNA binds non-covalently to binding sites on the cell surface (Lorenz and Wackernagel, 1994) prior to the translocation within the bacterial cell; double stranded DNA however needs to be converted to single stranded in order to be translocated successfully through the inner membrane (Chen and Dubnau, 2004).

DNA transfer between bacterial genomes can occur also through a different mechanism (i.e. transduction) that presupposes the presence of intermediates that fail to fit within the actual definition of a living organism, namely bacteriophages. Bacteriophages are viruses specialized to infect bacteria and a recent estimate suggests that approximately 10^{30} tailed bacteriophages exist on our planet, a number that far exceeds the population of any "living" organism (Brussow and Hendrix, 2002). There are two major types of transduction, generalized and specialized. In the first case, random fragments of the host bacterial chromosome can be packaged within the phage capsid during the replication and maturation process of the particles of a lytic bacteriophage. Some phage particles

carry exclusively bacterial DNA, and upon a second infection they can transfer genetic material from one bacterium to another.

Alternatively temperate bacteriophages integrate their genetic material into the bacterial chromosome, forming prophage elements. Upon induction a small part of the bacterial chromosome, close to the attachment site of the bacteriophage, is picked up and substitutes a small part of the actual prophage DNA; during the phage replication process the bacterial fragment replicates along with the phage DNA, such that every phage particle at the end will contain the same bacterial DNA fragment (specialized transduction). Upon a second infection, the DNA fragment of the previous host can now be transferred to a new bacterial recipient. The amount of transferable DNA through transduction depends on the actual dimension of the phage capsid and can be up to 100kb (Ochman *et al.*, 2000). Different bacteriophages infect certain bacterial species, and their specificity depends on the presence of distinct cell surface receptors on the bacterial cell.

The impact and extent of transduction as a mechanism of HGT can be concluded from a previous study (Canchaya *et al.*, 2003) focused on 56 sequenced Gram positive and Gram negative bacteria: 71% of those bacterial chromosomes contain at least one prophage sequence while prophages may account for up to 16% of the bacterial chromosomal DNA (Ohnishi *et al.*, 2001).

Conjugation is another mechanism of cell-to-cell DNA transfer that presupposes the physical co-occurrence of both the donor and the recipient cell. Conjugation is a widespread mechanism that allows the exchange of genetic material between distantly related lineages and even between different domains of life, e.g. bacteria-plant transfer (Buchanan-Wollaston et al., 1987). Conjugation frequently involves the transfer of a mobilizable or self-transmissible plasmid through a cell-to-cell bridge (mating pillus) from a donor to a recipient cell under a rolling-circle replication process (Khan, 1997).

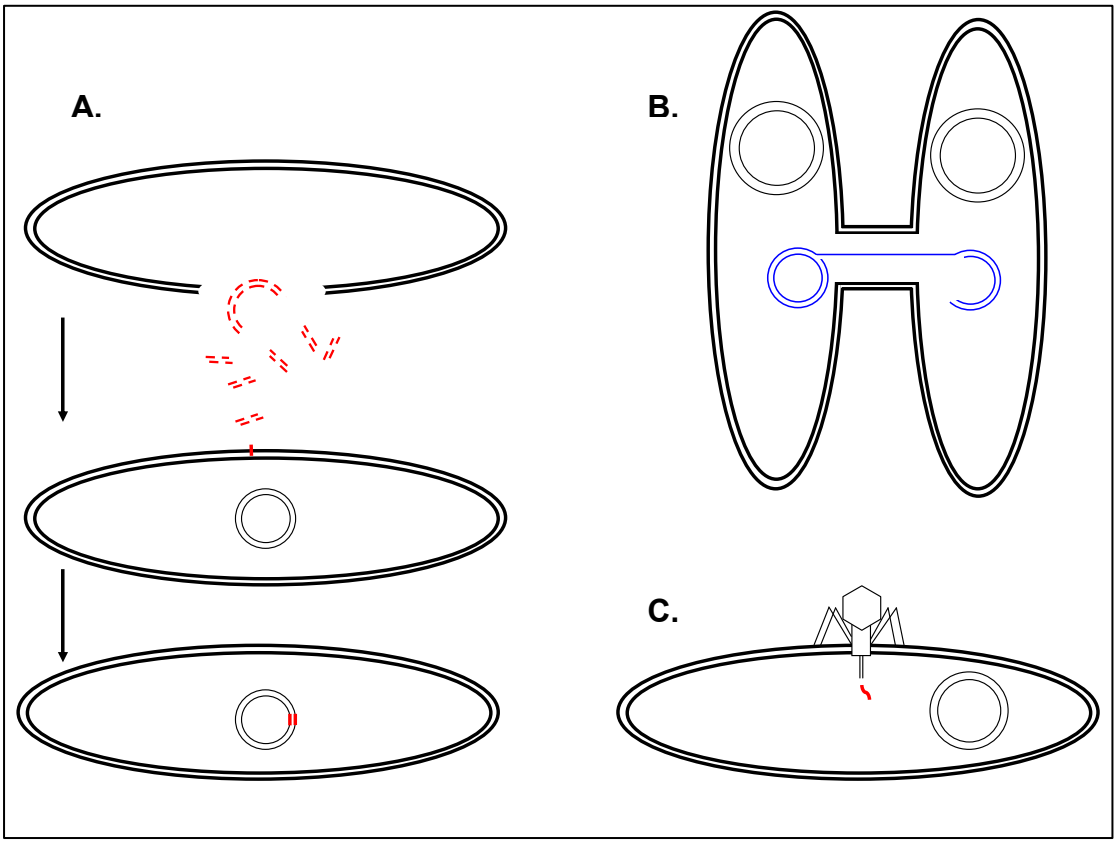

Figure 1.2: **A.** Uptake of naked DNA from the environment (transformation). **B.** Transfer of plasmid genetic material through the mating-pair pillus from a donor to a recipient bacterial cell (conjugation). **C.** Transfer of genetic material from a donor (not shown) to a recipient bacterial cell through a bacteriophage intermediate (transduction).

Some plasmids of Gram negative bacteria build the mating pillus utilizing a type IV secretion system (T4SS) and the specificity of the actual conjugation is determined by several factors including the interaction of the pillus with the outer membrane and the cell surface structure of the recipient cell (Anthony et al., 1994). If prior to the conjugation event, the plasmid had been inserted within the actual chromosome of the donor, e.g. via a recombination event between sequences of the plasmid and the chromosome, it is possible for DNA fragments of the donor chromosome to be captured by the plasmid and get transferred to the recipient cell; a subsequent recombination between the donor DNA fragment and the recipient chromosome represents the final step in the HGT event via a conjugation mechanism.

1.2 Genomic Islands

Horizontally acquired DNA sequences that contain functionally related genes with limited phylogenetic distribution, i.e. present in some bacterial genomes while being absent from closely related ones, are often referred to as genomic islands (GIs). The location of those mobile elements often correlates with distinct structural features such as tRNA genes, direct repeats (DRs) and mobility genes (e.g. integrase, transposase), which has lead to a definition of the GI structure that includes these features (Figure 1.3), (Hacker *et al.*, 1997; Hacker and Kaper, 2000; Schmidt and Hensel, 2004).

The range of the GI size is very wide, leaving almost no space for common consensus; for example GIs can be less than 4.5kb in length e.g. the *Salmonella* Pathogenicity Island (SPI)-16 in *S. typhi* CT18 (Vernikos and Parkhill, 2006) reaching up to 0.6Mb e.g. the symbiosis island in *Mesorhizobium loti* accounting for almost 10% of the total size of the chromosome (Sullivan and Ronson, 1998).

Some of the GI associated features are shared by other genomic elements such as integrated plasmids, bacteriophages, extracellular polysaccharide biosynthesis loci (Hacker and Kaper, 2000; Zhang *et al.*, 1997) and other gene clusters under specific constraints; these may or may not be recently horizontally acquired. However, GIs usually differ from bacteriophages and plasmids in the lack of autonomous replication origin (Schmidt and Hensel, 2004).

Pathogenicity islands (PAIs) constitute a specific type of GIs that provide virulence properties to bacterial strains. The concept of PAI was established in the late 1980s by Jörg Hacker and colleagues studying the virulence properties of uropathogenic strains of *E. coli* (UPEC) 536 and J96 (Hacker *et al.*, 1990; Knapp *et al.*, 1986). Clusters of virulence genes were previously described under the term “virulence gene blocks” (Hacker, 1990; High *et al.*, 1988; Low *et al.*, 1984). The observation that a group of genes could be deleted as a unit led to the definition of the term

"pathogenicity DNA islands" and later on to "pathogenicity islands" (Blum *et al.*, 1994; Hacker *et al.*, 1990).

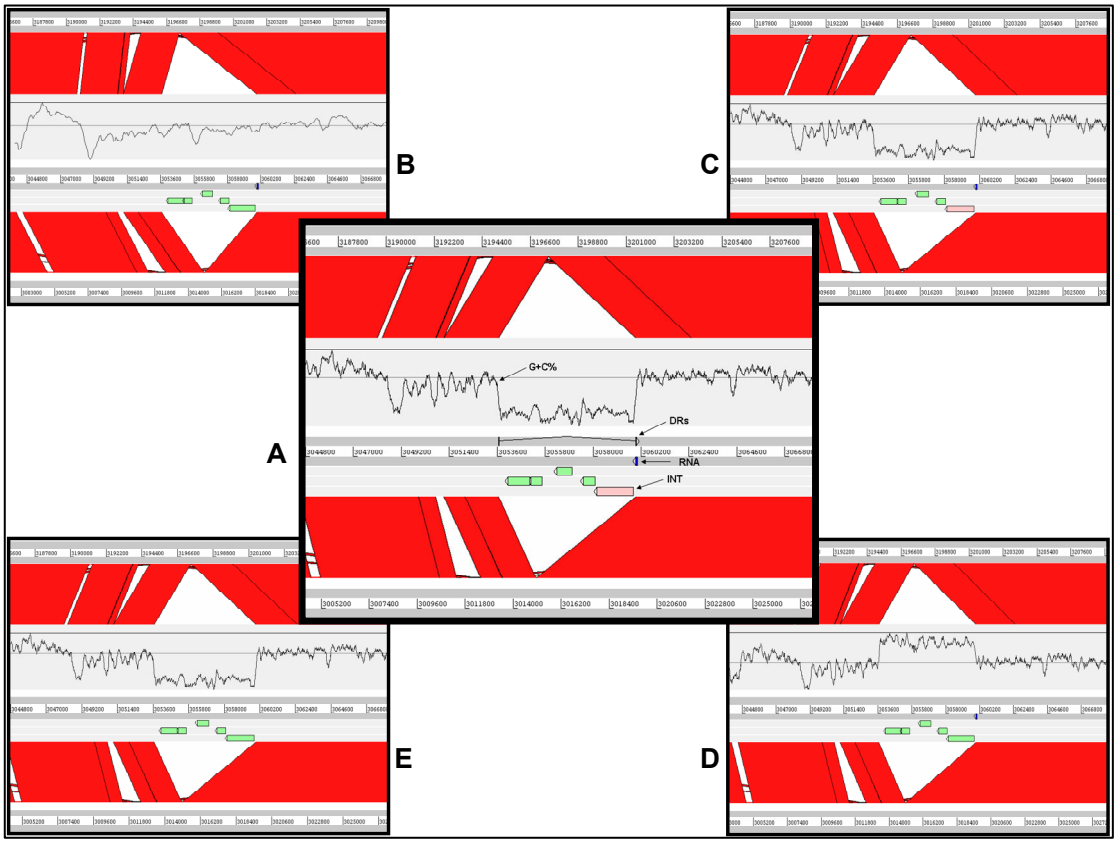

Figure 1.3: ACT (Carver *et al.*, 2005) screenshots: BLASTN comparison between three hypothetical bacterial strains (top, middle, bottom). Regions within the three strains with sequence similarity are joined by red coloured bands that represent the matching regions. **A.** Overview of the Genomic Island (GI) structure: pink and green coloured features represent integrase and functionally related genes respectively. The RNA gene in the proximity of the GI is detailed as a blue coloured feature, while the direct repeats flanking this island are shown as two joined features. The G+C% composition is shown in the graph plot above the island, using a 0.5kb window size. Structural variation of the GI structure: **B.** A hypothetical GI structure with no significant compositional deviation from the backbone composition, inserted adjacent to an RNA locus (e.g. *Salmonella* Pathogenicity Island (SPI)-6). **C.** A hypothetical GI structure with significant (low G+C%) compositional deviation from the backbone composition, inserted adjacent to an RNA locus, carrying an integrase gene at the 5' end (e.g. SPI-5). **D.** A hypothetical GI structure with significant (high G+C%) compositional deviation from the backbone composition, inserted adjacent to an RNA locus (e.g. SPI-9). **E.** A hypothetical GI structure with significant (low G+C%) compositional deviation from the backbone composition (e.g. SPI-4).

The phenotypic properties of GIs depend not only on their actual genetic alphabet i.e. their functional modules but also on the ecological context i.e. the host niche. In other words the same GI may confer completely different phenotype depending on the host organism. For example the iron uptake system present in *Yersinia* spp. is also present in non pathogenic bacteria found in the soil; in the first case it enables the survival of the bacterium within the host (resulting into pathogenic phenotypes) while in the second case it is used occasionally only under conditions of limited iron (Hacker and Carniel, 2001; Schmidt and Hensel, 2004).

Examples of other types of GIs include the symbiosis island in *M. loti* (Sullivan and Ronson, 1998), the metabolic islands in *Burkholderia cepacia* and *Pseudomonas aeruginosa* (Arora *et al.*, 2001; Baldwin *et al.*, 2004) and the antibiotic resistance island in *Salmonella* (Doublet *et al.*, 2005).

GIs are also present in Gram-positive bacteria but they can differ structurally from those present in Gram-negative bacteria; overall they do not exhibit specific junction sites (e.g. DRs), they are rarely inserted adjacent to RNA loci and they are often stably integrated in the host genome due to the lack of mobility genes (Hacker et al., 1997).

Insertion of GIs into the bacterial chromosome is often a site-specific event. About 75% of GIs currently known have been inserted at the 3' end of a tRNA locus including the acceptor-TψC stem-loop and often the CCA end (Hacker *et al.*, 2002; Hou, 1999; Williams, 2002). For example the $tRNA^{selC}$ locus has been extensively used as an integration hot spot for many different GIs in enteric bacteria (Ritter *et al.*, 1995). The length of the DRs ranges from 9 bp (e.g. PAI-1 in UPEC CFT073) to 135 bp (e.g. the locus of enterocyte effacement (LEE) PAI in EHEC). An average length for the DRs is approximately 20 bp (Kaper and Hacker, 1999; Schmidt and Hensel, 2004).

Three theories have been proposed to explain the predominant use of tRNAs as insertion sites:

1. Certain tRNAs might read more efficiently (atypical) codons of the associated GI, e.g. the rare $tRNA^{LeuX}$ (Ritter *et al.*, 1997).

2. Multiple copies of tRNA genes provide alternative insertion sites for GIs.

3. Integrases recognize specific motifs in the conserved tRNA structure/sequence to facilitate the integration and excision of GIs (Reiter *et al.*, 1989).

Other genes though may also act as insertion sites for GIs e.g. the cag PAI has been inserted within the *glr* (glutamate racemase) gene of *Helicobacter pylori* (Censini et al., 1996); another example is SPI-9 (Parkhill *et al.*, 2001) that has been inserted close to a tmRNA (also known as 10Sa RNA, (Williams, 2003)) locus of *S. typhi* CT18.

There are many different ways of describing and illustrating the structural variation of GIs; for example GIs can be classified based on their mechanism of mobilization e.g. transduction or conjugation. A similar classification can be based on the actual family of mobility genes, e.g. tyrosine or serine recombinases and DDE transposases. Another classification scheme could be based on the overall phenotypic properties of GIs, e.g. virulence, metabolism and antibiotic resistance. Under the same framework of classification but in a higher resolution, GIs can be classified based on their functional modules, e.g. Type III Secretion Systems (T3SS) and T4SS or even based on the origin of those modules, e.g. phage, plasmid and transposon-derived; the combination of those modules creates a continuum of mobile element mosaicism, increasing even further the structural complexity of GIs (Osborn and Boltner, 2002; Toussaint and Merlin, 2002). GIs can also be classified based on their composition (e.g. high or low G+C%), the host repertoire (e.g. Gram positive, Gram negative), the level of structural mosaicism (e.g. more than one independent insertion event) or even the insertion point preference (e.g. tRNA, tmRNA or coding sequences – CDSs). It is worth noting that often some GI modules are interrelated with others; for example most of

the tyrosine recombinases catalyse site-specific insertion at the 3' end of a tRNA locus (Burrus *et al.*, 2002).

Furthermore a more abstract and broadly applicable classification scheme can be based simply on the actual GI sequence-structural components i.e. presence or absence of mobility genes, repeats and phage related domains, compositional deviation and proximity to tRNA loci; under this classification scheme, any type of GI can be described regardless of the mechanism of integration, the host repertoire or other specific properties. For example a GI can be described with a binary profile: e.g. [1,0,1,0,0,1,0] for repeats, integrase, tRNA, phage-domains, leading or lagging strand bias, high or low gene density and low or high G+C% content respectively, in a similar way that a given isolate of a species is described by a given allelic profile using the Multi Locus Sequence Typing (MLST) method (Maiden *et al.*, 1998). Under this framework GIs with the same structural [1,0,1,0,0,1,0] profile would be grouped under the same GI family.

In the following section, I aim to provide a short review on representative examples of distinct GI structures, summarizing key features that reveal the structural variability and conservation that describes this superfamily of mobile elements. I will focus on a broad selection of thirteen GI structures providing a representative sampling of the structural variation, rather than attempting to be comprehensive, providing an extensive list of all the known examples of GIs. There are many existing reviews discussing most of the known GIs, notably by Hacker (Hacker and Kaper, 2000; Kaper and Hacker, 1999), and Schmidt (Schmidt and Hensel, 2004). Furthermore, an online database namely PAI-DB (http://www.gem.re.kr/paidb/) provides a very comprehensive list of known PAI structures (Yoon *et al.*, 2007). Some of the issues summarised here are discussed in more detail in other chapters of this thesis.

The rationale behind the classification scheme discussed in the following section will be based on the distinct building blocks and

functional modules carried by a broad selection of GIs which in return affect both their phenotypic properties and the mechanism of mobilization (Table 1.1).

Table 1.1: A list of 13 representative Genomic Islands and their functional modules, used in this analysis.

Functional module	Genomic Island	Reference
Toxin	SaPIs	(Fitzgerald *et al.*, 2001; Holden *et al.*, 2004; Lindsay *et al.*, 1998; Novick *et al.*, 2001; Novick and Subedi, 2007)
T1SS	SPI-4	(Gerlach *et al.*, 2007; Morgan *et al.*, 2007)
T2SS and Iron Uptake	HPI	(Carniel, 1999; Carniel, 2001)
T3SS	LEE, SPI-1, SPI-2	(Jerse *et al.*, 1990; McDaniel *et al.*, 1995)
T4SS	cag PAI	(Censini *et al.*, 1996)
Type IV pilus	SPI-7	(Parkhill *et al.*, 2001; Pickard *et al.*, 2003)
T5SS	SPI-3	(Blanc-Potard *et al.*, 1999)
Symbiosis (nodulation and nitrogen fixation)	Symbiosis Island ICEMlSymR7A	(Sullivan and Ronson, 1998)
Metabolism	cci	(Baldwin *et al.*, 2004)
Antibiotic resistance	SGI-1, Vibrio SXT	(Beaber *et al.*, 2002; Doublet *et al.*, 2005; Hochhut and Waldor, 1999)

1.2.1 SaPIs

Staphylococcus aureus is a common commensal organism present on the respiratory tract and skin of 30-70% of the human population. However *S. aureus* can also act as a pathogen, shows high resistance to antibiotics and is commonly associated with nosocomial infections, including, but not limited to, bacteraemia, endocarditis and syndromes caused by a wide range of toxins, such as exotoxins and superantigens.

Superantigens including the toxic shock syndrome toxin-1 (TSST-1) (Lindsay *et al.*, 1998) are frequently carried by the staphylococcal pathogenicity islands (SaPIs), a structurally very well conserved family of phage-related GIs (Figure 1.4) present in all but one (MSSA476) of the

sequenced *S. aureus* strains and in many other staphylococci (Novick and Subedi, 2007). Six different sites on the staphylococcal chromosome are occupied by the already identified SaPIs in a similar orientation to prophage elements, i.e. with the majority of the genes oriented in the same direction as chromosomal replication (Novick and Subedi, 2007).

SaPIs are induced to excise and replicate by specific types of temperate staphylococcal bacteriophages and the replicated SaPI DNA is encapsulated into phage-encoded capsids and spread with high frequency within the staphylococci lineage by means of generalized transduction (Maiques *et al.*, 2007); SaPIs are stably integrated in the chromosome in the absence of helper phages due to the lack of SaPI-encoded excisionase (Novick, 2003). SOS induction, triggered by antibiotics, may result in a wide spread of SaPIs, raising an issue of how effective antibiotic treatment can really be in the case of bacteria with highly mobile, virulent elements like SaPIs (Ubeda *et al.*, 2005).

In terms of gene content SaPIs are extremely well conserved; a conserved (encapsidation) module consisting of five genes is present at the right end of SaPIs (Figure 1.4); within this module the *ter* gene encodes the terminase small subunit commonly found in phages of Gram-positive bacteria. Another conserved gene present in SaPIs is the *rep* gene encoding a helicase/primase-like protein that is important for the replication of the SaPI DNA and is located on the left side of the encapsulation module. Further on the left there are two genes with helix-turn-helix motifs, encoding putative regulatory proteins (Novick and Subedi, 2007) while at the left most end of the SaPIs, adjacent to the attachment (att), site is the integrase gene. Often in some SaPIs, two superantigens are located next to the integrase gene. The two superantigens, TSST-1 (*tst* gene) and enterotoxin B (*seb* gene), are located at the same position but on opposite orientation in SaPI1 (Lindsay *et al.*, 1998) and SaPI3 (Novick *et al.*, 2001) (Figure 1.4).

SaPIs have a size of 15-17kb and are flanked by DRs, consisting of a highly conserved core of 15-22bp flanked by variable sequences. The

average G+C content of SaPIs is 31%, lower than the chromosome G+C content of 33%. Similarly in terms of gene density, SaPIs have a much higher gene density of 1.45 genes/kb, compared to the genome average of 0.9 genes/kb, suggesting chromosomes of higher gene density than that characterizing the *Staphylococcus* lineage as the potential source of those GIs, one obvious possibility being bacteriophage genomes (Vernikos and Parkhill, 2007).

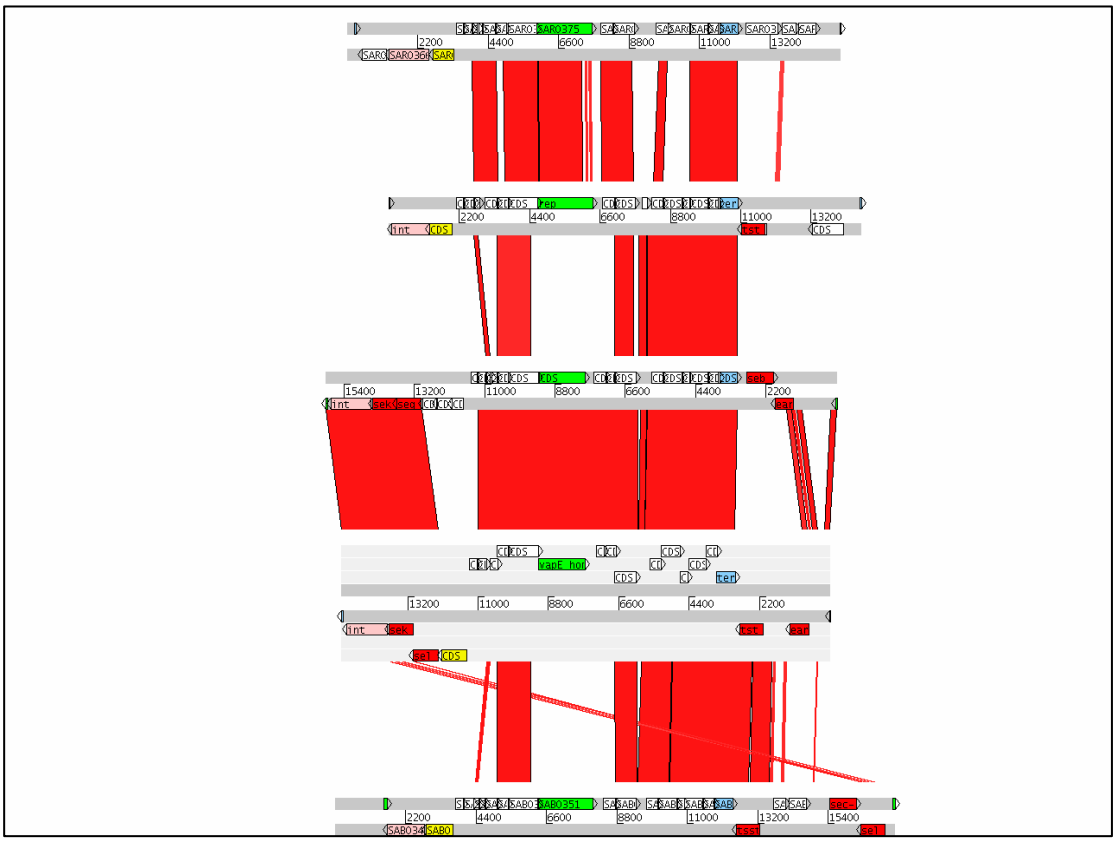

Figure 1.4: Comparison of SaPIs. From top to bottom: SaPI4 (*S. aureus* MRSA252), SaPI2 (*S. aureus* RN3984), SaPI3 (*S. aureus* COL), SaPI1 (*S. aureus* RN4282) and SaPIbov (*S. aureus* RF122). Regions within the five SaPI DNA sequences with sequence similarity are joined by red colored bands that represent the matching regions under a BLASTN comparison (ACT screenshot). CDS colouring scheme: Pink; integrase, green; helicase/primase, red; superantigens, yellow; helix-turn-helix motif, and light blue; terminase.

1.2.2 LEE

The pathological intestinal phenotype “attaching and effacing” (A/E) that is described as the effacement of the intestinal epithelial microvilli and the

intimate attachment of the bacterium to the epithelial cells (Jerse *et al.*, 1990; Kaper and Hacker, 1999; Wales *et al.*, 2005), is attributed almost exclusively to a mobile element, termed locus of enterocyte effacement (LEE) (McDaniel *et al.*, 1995). Enteropathogenic *E. coli* (EPEC), a major cause of infantile diarrhea in the developing world, EHEC responsible for food and water-borne poisoning causing hemorrhagic colitis and *Citrobacter rodentium*, responsible for murine colonic hyperplasia, are hosts of the LEE island and produce A/E lesions (Deng *et al.*, 2001; McDaniel *et al.*, 1995; Perna *et al.*, 1998). In commensal *E. coli* K-12 transfer of the LEE confers the full A/E phenotype (McDaniel and Kaper, 1997).

LEE is a very well conserved family of GIs, with an average G+C content of 38%, significantly lower than the genome average G+C content of *E. coli* and *C. rodentium* (50% and 54% respectively). The size of the LEE island ranges from 35kb (in EPEC E2348/69) to 43kb (in EHEC O157:H7) although in the last case the almost 8kb difference is attributed to a putative P4 prophage element integrated between the $tRNA^{selC}$ locus and the LEE island, suggesting an independent acquisition event after the acquisition of the LEE island (Perna *et al.*, 1998); however the same prophage is also present in O55:H7 EPEC (Kaper and Hacker, 1999). In most cases the LEE island is integrated adjacent to the $tRNA^{selC}$ gene, although in some EPEC strains and in *C. rodentium* it is integrated in different loci (Gal-Mor and Finlay, 2006), leaving open the possibility of multiple independent acquisitions of this GI throughout the evolution of A/E bacteria. The LEE island completely lacks DRs, and phage-or plasmid-related domains and its mechanism of mobilisation remains unknown; LEE has an overall gene density slightly higher than the genome average (1.15 and 0.97 genes/kb respectively).

In terms of gene content, LEE contains 41 genes, organized in five polycistronic operons, namely LEE1, LEE2, LEE3, LEE4 and LEE5 (Figure 1.5). LEE1 contains the *ler* (LEE-encoded regulator) gene, whose product is homologous to the H-NS transcriptional regulators (Kaper and

Hacker, 1999); the *ler* gene activates the transcription of LEE2-LEE4. Also present in the LEE1 are genes encoding components of the T3SS. LEE2 contains also components of the T3SS and additionally the *cesD* gene encoding a chaperone important for the secretion of EspD and EspB (Elliott *et al.*, 1998). Components of the T3SS are also present in the third operon (LEE3).

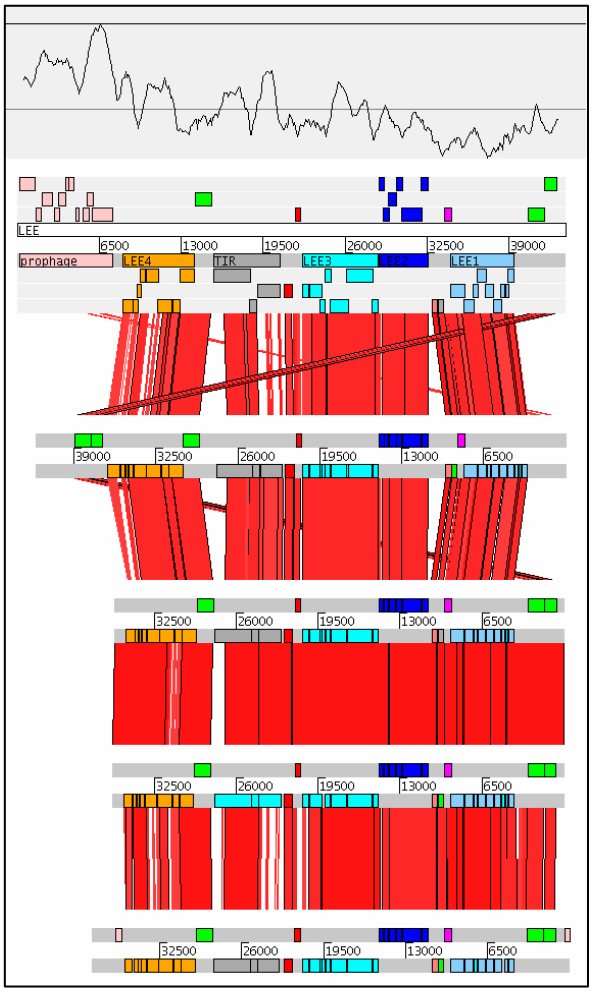

Figure 1.5: Comparison of the LEE island present in different genomes. From top to bottom: *E. coli* O157:H7 EDL933, *C. rodentium* DBS100, *E. coli* E2348/69 EPEC, *E. coli* 0181-6/86 EPEC and *E. coli* RDEC-1 EPEC. Operon colour scheme: Light blue; LEE1, dark blue; LEE2, cyan; LEE3, grey; TIR (LEE5), orange; LEE4, and light pink; P4 prophage. Graph: G+C% content with a window size of 1kb.

LEE4 contains genes (*espA, espD, espB and espF*) encoding products secreted by the T3SS. The fifth operon (TIR) carries three genes namely

cesT, *eae* and *tir*; *cesT*, like *cesD*, encodes a chaperone. *eae* gene was the first characterised gene of the LEE island (Jerse *et al.*, 1990) and encodes an outer membrane protein (intimin), an important intestinal adherence factor (Donnenberg *et al.*, 1993) that binds on the translocated intimin receptor (Tir). The intimin receptor is encoded by the *tir* gene and is translocated into the host cell via the T3SS (Kenny *et al.*, 1997).

Although the gene content of the LEE island is very well-conserved, in terms of sequence composition and similarity, LEE shows a mosaic structure. The most highly conserved genes encode the components of the T3SS (*esc* genes), whereas the genes encoding secreted proteins (*esp*) are more divergent (90% and 80% average nucleotide sequence similarity respectively). The *tir* gene is one of the most divergent loci in the LEE island with an average sequence similarity of 68%. The mosaic nature of LEE is also evident from the G+C content; LEE1, LEE2 and LEE3 have an average G+C content of 33.3%, 38.2% and 39.5% as opposed to a much higher G+C content of the TIR and the LEE4 operon of 43.6% and 42.6% respectively (Figure 1.5).

1.2.3 SPI-7

Salmonella enterica serovar Typhi (*S. typhi*), a human restricted, host adapted pathogen is the aetiological agent of typhoid fever (Parry *et al.*, 2002) and most *S. typhi* isolates carry the viaB locus that encodes the Vi capsular polysaccharide (Hashimoto *et al.*, 1993; Hornick *et al.*, 1970; Robbins and Robbins, 1984). In some *Salmonella* serovars Typhi, Paratyphi C and Dublin isolates the viaB locus is located on a PAI, termed SPI-7 (Parkhill *et al.*, 2001).

The G+C content of SPI-7 is 49.7%, slightly lower that the genome average G+C content of Typhi CT18 (52%). The overall gene density of SPI-7 is 0.99 genes/kb while the genome average is 0.91 genes/kb. SPI-7 is inserted at the 3' end of the $tRNA^{Phe}$ gene which has been displaced at the 3' end of SPI-7; however the insertion has fully restored the displaced DNA fragment of the tRNA at the point of insertion.

In terms of sequence composition and gene content, SPI-7 is a highly mosaic GI with distinct functional modules that were acquired in several independent HGT events (Pickard *et al.*, 2003). At the 3' end of SPI-7 (Figure 1.6), close to the displaced tRNA fragment, are a set of genes encoding proteins for conjugation and DNA replication such as single-stranded DNA binding protein (ssb), DNA helicase (dnaB), chromosome partitioning protein (parB) and topoisomerase (topB); the average G+C content of this region is 49.9%. Further on the right of this module there is a group of 14 genes (*pil*) that encode a type IVB pilus system. The type IVB pilus system is likely to play a role in the intestinal cell attachment of *S. typhi* (Zhang *et al.*, 2000) and may initially have served as a mating pair formation system of a conjugative plasmid; this gene cluster is similar to the one present in plasmid R64 (Zhang *et al.*, 1997).

Further on the right of the type IVB pilus system is a set of genes involved in DNA transfer, frequently plasmid-encoded, like *traE*, *traG* and *traC*. These three functional modules present at the left end of SPI-7 (Figure 1.6) have been previously suggested to play a key role in the mobilization of the entire SPI-7 locus via conjugation and form probably an independently acquired part of SPI-7; the similarity of this locus to plasmid R64 (Zhang *et al.*, 1997) suggests a possible plasmid-related origin (Pickard *et al.*, 2003). This observation is in line with other studies showing strong similarity of this SPI-7 locus to a wide range of structurally well conserved GIs (Figure 1.6), discussed in the following paragraph (Hensel, 2004; Mohd-Zain *et al.*, 2004).

A bacteriophage encoding a SPI-1 effector protein, SopE which is important for the invasion of *Salmonella* in the epithelial cells (Friebel *et al.*, 2001; Mirold *et al.*, 1999; Wood *et al.*, 1996) represents a 33.5kb insertion next to the conjugation locus of SPI-7 with a G+C content very close to the genome average (51.57% and 52.09% respectively) and it is flanked by a set of 9bp DRs.

This prophage element represents probably an independent HGT event in Typhi, given that it is absent from SPI-7 present in Dublin and

Paratyphi C, adding an extra level of mosaicism to SPI-7 (Pickard *et al.*, 2003). Next to the SopE prophage lies the viaB locus consisting of 10 genes, *vexA-E* and *tviA-E* encoding Vi polysaccharide export and biosynthesis proteins respectively; the average G+C content of this ~15kb region is 45.2%; significantly lower than the genome average and the overall G+C content of the entire SPI-7 (49.7%).

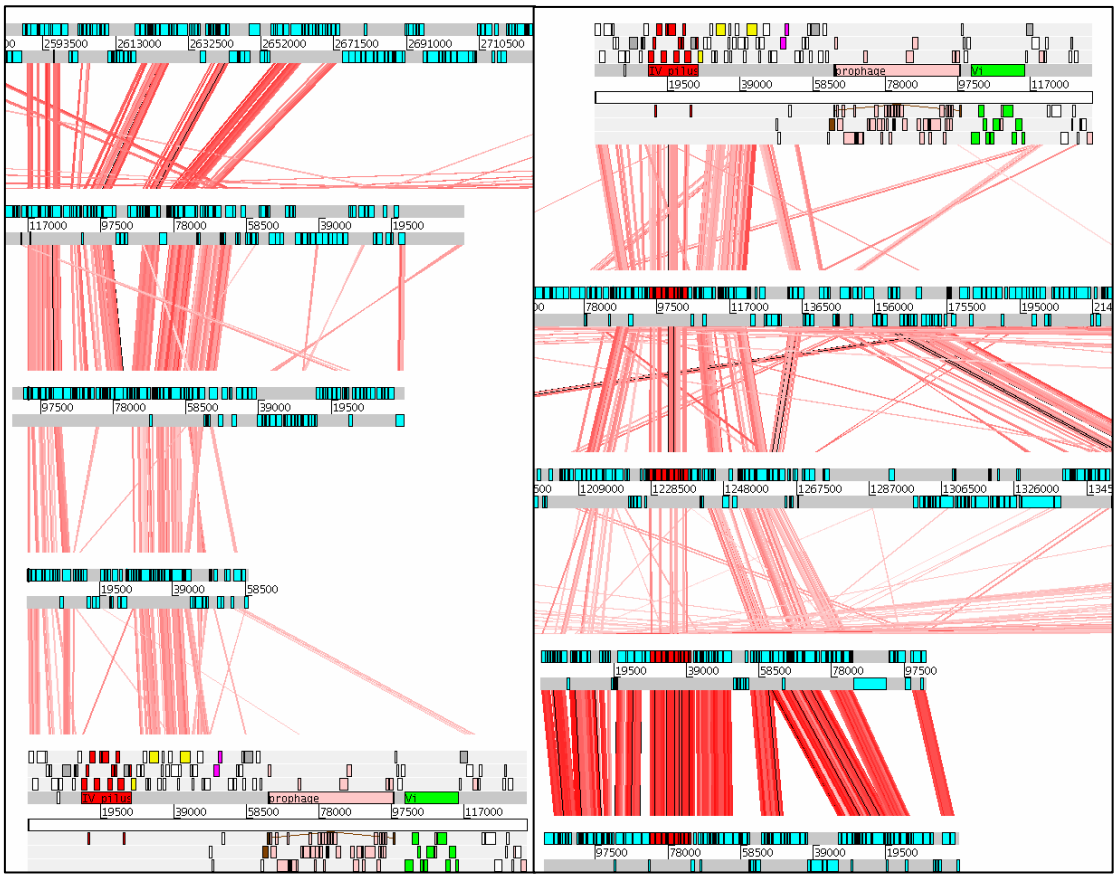

Figure 1.6: Comparison of SPI-7 with other GIs; from top to bottom and left to right: GI present in *Xanthomonas axonopodis* pv. Cistri 306, PAGI-3 island of *P. aeruginosa* SG17M, clc element in *Pseudomonas* sp. strain B13, ICEHin1056 in *H. influenza*, SPI-7 in *S. typhi* CT18 (bottom left and top right), YAPI island in *Y. enterocolitica* 8081, GI in *Photorhabdus luminescens* TT01, pKLC102 plasmid of *P. aeruginosa* C, PAP1 island in *P. aeruginosa* PA14 . Colour scheme: Grey; DNA replication, red; type IVB pilus, yellow; DNA transfer, light pink; prophage, pink; regulatory proteins ibrA and ibrB, brown; UV protection, and green; Vi antigen biosynthesis and export.

The first 62kb (conjugation) region of SPI-7 constitutes a very well conserved genetic locus similar to GIs and other mobile elements present in a wide range of hosts including β and γ-Proteobacteria and plant pathogen representatives (Figure 1.6), (Mohd-Zain *et al.*, 2004; Pickard *et*

al., 2003). Members of this diverse GI family include SPI-7 (134kb) (Parkhill *et al.*, 2001; Pickard *et al.*, 2003), ICEHin1056 (59kb) in *H. influenza* (Mohd-Zain *et al.*, 2004), the 65kb YAPI island in *Y. enterocolitica* 8081 (Thomson *et al.*, 2006), a 140kb GI in *Photorhabdus luminescens* TT01 (accession number NC_005126), the 105kb clc element in *Pseudomonas* sp. strain B13 (Ravatn *et al.*, 1998), the 106kb PAP1 island in *P. aeruginosa* PA14 strain (He *et al.*, 2004), the pKLC102 plasmid of *P. aeruginosa* C (Klockgether *et al.*, 2004), the 114kb PAGI-3 island of *P. aeruginosa* SG17M strain and a 86kb island present in *Xanthomonas axonopodis* pv. Cistri, strain 306 (accession number NC003919). The members of this family of mobile elements integrate into a wide range of distinct tRNA loci, have an average G+C content ranging from 40 to 70% and share a set of approximately 33 core genes (Mohd-Zain *et al.*, 2004). So far only the clc element and ICEHin1056 have been shown to be able to conjugate at frequencies of 10^{-6}-10^{-7} (Dimopoulou *et al.*, 1992; Ravatn *et al.*, 1998). The fact that these seemingly similar GI structures are conserved in terms of gene content but are extremely diverged in terms of sequence composition and integration site, leaves open the possibility of convergent evolution rather than recent common ancestry.

1.2.4 SGI-1

The multidrug resistance (MDR) phenotype of *S. enterica* serovar Typhimurium strain DT104 is attributed to a 43kb *Salmonella* genomic island 1 (SGI-1) integrated at the 3' end of the *thdF* gene encoding a thiophene and furan oxidation protein (Doublet *et al.*, 2005; Mulvey *et al.*, 2006). The MDR phenotype was acquired by DT104 in the early 1980s upon the integration of SGI-1; DT104 is resistant to chloramphenicol, ampicilin, streptomycin, tetracycline and sulfonamides (Threlfall, 2000).

The antibiotic resistance genes are located at the 3' end of SGI-1 on a 13kb class 1 integron. Generally, class 1 integrons consist of a conserved 5' end (integrase gene) and 3' end (quaternary ammonium compound and sulphonamide resistance genes) while the central recombination site

contains a variable gene cassette(s) (Recchia and Hall, 1995). The G+C content of this integron is significantly higher than the remaining part of SGI-1 and the overall genome average G+C content of DT104 (58.7%, 44.2% and 52% respectively). The SGI-1 integron is flanked by 26bp DRs (Figure 1.7), while an internal set of 5bp inverted repeats are flanking the antibiotic resistance gene cassette. The structure of this class 1 integron is highly variable, due to the integrase-mediated exchange of gene cassettes (Boyd *et al.*, 2002; Carattoli *et al.*, 2002; Doublet *et al.*, 2003). There are at least two different types of variation, including loss of a single resistance gene and exchange of the entire gene cassette (Levings *et al.*, 2005).

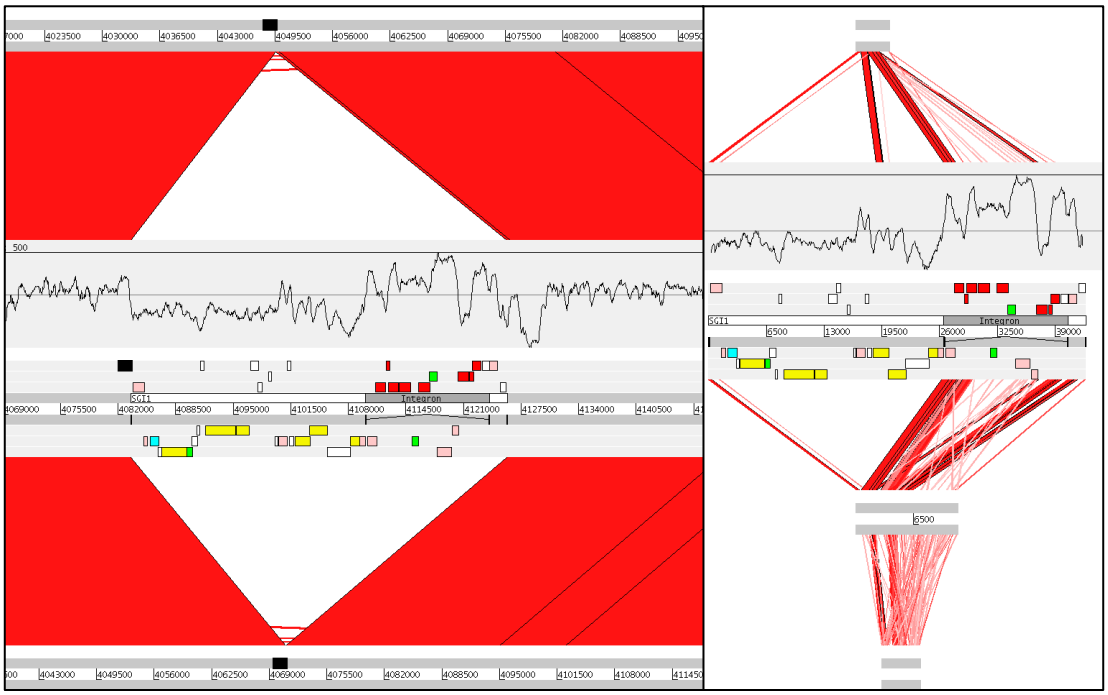

Figure 1.7: SGI-1 structure and phylogenetic distribution. **Left:** ACT comparison showing the presence of SGI-1 in Typhimurium DT104 (middle) and its absence in Typhimurium LT2 (top) and SL1344 (bottom). The G+C% content is embedded as a plot above SGI-1 with a window size of 0.5kb. Colour scheme: Black; chromosomal gene *thdF* (SGI-1 insertion point) present in all 3 Typhimurium strains, yellow; conjugation, red; drug-resistance, green; regulation, light pink; recombination/integration, and cyan; DNA replication. **Right:** Comparison of SGI-1 (middle) against SGI1-J (top) present in *S. enterica* Emek (Levings *et al.*, 2005), SGI1-K (below SGI-1) present in *S. enterica* Kentucky (Levings *et al.*, 2007) and SGI1-L (bottom) present in *S. enterica* Newport (Cloeckaert *et al.*, 2006).

A third set of repeats (DRs) of 19bp are flanking the entire SGI-1 element and the attachment site on the left border (attL) of SGI-1

corresponds to the 3' end of the *thdF* gene which is fully restored at the site of insertion (Figure 1.7). The overall G+C content of SGI-1 is 49.2% and the average gene density is very close to the DT104 genome average (0.99 and 0.98 respectively). In terms of gene content, there are at least six functional classes of genes, including recombination, replication, conjugation, regulation, drug resistance and other genes of unknown function (Figure 1.7). The conjugation related genes (including a mating pair protein coding gene) suggest that SGI-1 might have, at least originally, been acquired through a cell-to-cell contact mechanism (conjugation). Recently is has been shown that SGI-1 can be conjugally transferred *in trans* by a helper plasmid (R55), with an extrachromosomal circular intermediate at frequencies of 10^{-5}-10^{-6} transconjugants/donor; for these reasons SGI-1 has been classified as a mobilizable, non-self-transmissible element (Doublet *et al.*, 2005).

Apart from Typhimurium DT104, SGI-1 has been also described in *Proteus mirabilis* (Ahmed *et al.*, 2007) and other *S. enterica* serovars including Albany, Agona, Paratyphi B, Meleagridis and Newport (Boyd *et al.*, 2001; Cloeckaert *et al.*, 2000; Doublet *et al.*, 2003; Ebner *et al.*, 2004; Meunier *et al.*, 2002; Mulvey *et al.*, 2004).

1.2.5 cag PAI

H. pylori is a gram negative spiral shaped bacterium that colonizes half of the human population and causes peptic ulcer and gastric neoplasia (Dunn *et al.*, 1997). Almost 10% of infected individuals develop peptic ulcer and only 1% gastric cancer (Gal-Mor and Finlay, 2006). Those severe pathogenic phenotypes of *H. pylori* are mainly attributed to the presence of a mobile genetic element namely cag PAI (Censini *et al.*, 1996). Overall *H. pylori* shows extreme genetic variation, attributed to direct uptake of DNA from the environment (transformation) (Hofreuter *et al.*, 2001), high rates of mutation (Bjorkholm *et al.*, 2001) and frequent recombination (Suerbaum and Achtman, 1999).

The cag region is a 37-40kb PAI with a G+C content significantly lower than the average *H. pylori* G+C content (35.7% and 39% respectively). Its average gene density (0.88 genes/kb) is also lower than the genome average gene density of 0.96 genes/kb. The cag PAI has been inserted at the 3' end of the *glr* gene that encodes a glutamase racemase protein; the insertion has not disrupted the *glr* gene but rather its 3' end has been fully restored upon the integration of the cag PAI island. The displaced and the restored fragment of the *glr* gene form a set of 41bp DRs that are flanking the boundaries of the cag PAI (Figure 1.8). Based on an *in silico* model, the cag PAI has been estimated to have been acquired by *H. pylori* approximately 50Myr ago (Kaper and Hacker, 1999).

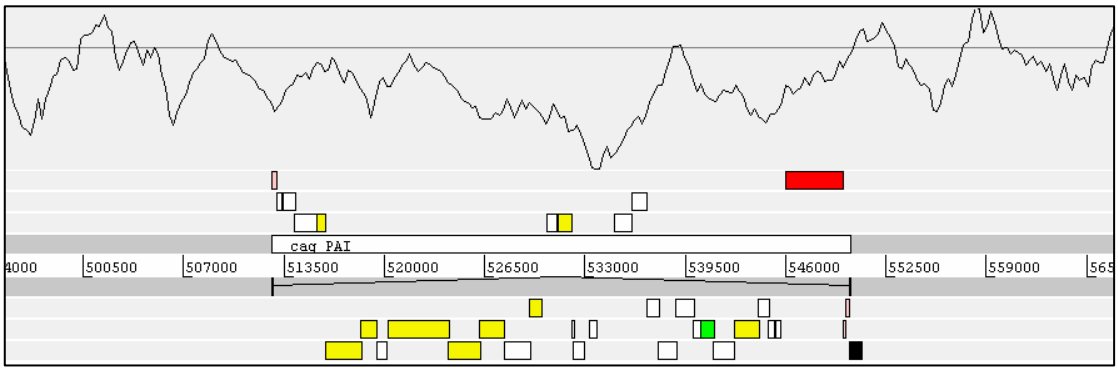

Figure 1.8: cag PAI structure. Colour scheme: Yellow; T4SS components, green; putative chaperone, red; cytotoxin-associated protein A – *cagA* gene, light pink; transposase, and black; *glr* gene – insertion point of the cag PAI. The 41bp DRs are shown as two joined black features flanking the cag PAI. Graph: G+C% content with a window size of 2.5kb.

In terms of gene content cag PAI encodes 27-33 genes and in some strains two IS elements are present at the 5' and 3' end of this PAI next to the attL and attR. Nine of the cag PAI genes show sequence similarity to the T4SS components of the *Agrobacterium tumefaciens* virB operon (Gal-Mor and Finlay, 2006). One of those genes (*cagE*) shows sequence similarity to the *virB4* gene of *A. tumefaciens* (Ward *et al.*, 1988) and the *trbE*, *traB* and *ptlC* genes of plasmids RP4 and pKM101 and *Bordetella petussis* respectively (Lessl and Lanka, 1994; Weiss *et al.*, 1993; Winans *et al.*, 1996). CagE, like, PtlC, VirB4, TrbE and TraB contains a Walker box

(Walker *et al.*, 1982), a type A nucleotide-binding site, that is required for ATP hydrolysis (Censini *et al.*, 1996); CagE has been shown to stimulate bacterial-mediated epithelial IL-8 secretion (Maeda *et al.*, 2001; Sharma *et al.*, 1998; Tummuru *et al.*, 1995).

The *cagF* gene product has been previously shown to interact with CagA and has been hypothesized to encode a chaperone-like protein (Couturier *et al.*, 2006). The product of the *cagA* gene (cytotoxin-associated protein A) is an immunodominant antigen and is the only known effector protein delivered by the *H. pylori* T4SS (Segal *et al.*, 1999). CagA interacts with host proteins and affects the cell junctions, the cytoskeleton and signal transduction pathways (Bourzac and Guillemin, 2005) leading to actin polymerization and anomalous epithelial cell proliferation. Overall the cag PAI up-regulates the expression of proinflammatory chemokines (e.g. IL-8) which in return contribute to the stomach tissue damage and inflammation (Crabtree *et al.*, 1995).

1.2.6 Symbiosis island

M. loti species can be differentiated from other *Mesorhizobium* and *Rhizobium* species based on the fact that the entire genetic information required for symbiotic lifestyle is chromosomally rather plasmid encoded (Chua *et al.*, 1985; Sullivan *et al.*, 1995). The species name nomenclature "loti" comes from its host species *Lotus* on which nodules containing the bacteria, are formed. It has been shown that the symbiosis information in *M. loti* is encoded on a mobile genomic element, termed symbiosis island (SI) (Sullivan and Ronson, 1998). SI is perhaps the largest known example of GI, constituting almost 10% (611kb) of the entire *M. loti* chromosome (7Mb). The average G+C content of SI is lower than the average genome G+C content (59.7% and 62.7% respectively) surprisingly lower if the exceptionally large size of SI is also taken into account. The average gene density is slightly lower than the genome average (0.94 and 0.96 respectively) suggesting perhaps a phylogenetically closely related species-donor. SI has been integrated at the 3' end of the $tRNA^{Phe}$ gene

reconstructing the displaced fragment at the point of insertion; the two tRNA fragments (17bp) form the attL and attR, flanking the entire 611kb SI region (Figure 1.9).

In terms of gene content, SI hosts 576 genes that encode for a wide range of functional products including, but not limited to, integration/recombination (111 CDSs), nitroxen fixation (19 CDSs), nodulation (18 CDSs), ABC transporters (21 CDSs), conjugation (11 CDSs), T3SS components (6 CDSs), cytochromes (10 CDSs), ferredoxin (4 CDSs) and transcriptional regulators (20 CDSs). It is worth noting that of the 111 integration/recombination CDSs 89 encode transposases (Figure 1.9).

Figure 1.9: Structure of the symbiosis island (SI) in *M. loti*. Colour scheme: Light pink; integration/recombination, red; nitrogen fixation, yellow; nodulation, green; ABC transporters, light blue; conjugation, dark blue; T3SS translocation components, brown; cytochrome, orange; ferredoxin, and pink; transcriptional regulation. The screenshot at the top illustrates the same SI structure in a lower resolution allowing an overview of the SI region within the *M. loti* chromosome. The G+C% content (graph at the top of each panel) was drawn with a window size of 50kb and 5kb (top and bottom respectively).

The entire 611kb SI region can excise, forming a circular intermediate and get transferred via conjugation from symbiotic to non-symbiotic *Mesorhizobium* species (Hentschel and Hacker, 2001; Ramsay *et*

al., 2006; Sullivan and Ronson, 1998). A P4 phage integrase located at the 5' end of SI is necessary for the integration and excision of SI, with higher frequencies of excision during the exponential and stationary phase (Ramsay *et al.*, 2006). The fact that two genes of SI have sequence similarity to the quorum sensing *traR* and *traI* genes of *A. tumefaciens* leaves open the possibility of a fine-tuned mechanism that controls the excision of the SI elements relative to the bacterial population density (Fuqua and Winans, 1994; Ramsay *et al.*, 2006).

SI belongs to a diverse family of mobile elements, termed integrative and conjugative elements (ICEs) defined by their ability to excise site-specifically, to be mobilized and transferred via conjugation from one host to another, forming a circular extra-chromosomal intermediate, and to integrate in the recipient genome (Burrus *et al.*, 2006; Burrus *et al.*, 2002; Burrus and Waldor, 2004). The term "ICE" describes a very diverse family of GIs that share a mix of both phage and plasmid related functions; ICEs, like plasmids, transfer via conjugation and like prophages integrate and replicate along with the recipient chromosome. Known examples of ICEs include the symbiosis island of *M. loti* (Sullivan and Ronson, 1998), the *Vibrio cholerae* SXT element (see below) (Beaber *et al.*, 2002; Hochhut and Waldor, 1999), the ICEHin1056 element in *H. influenza* (Mohd-Zain *et al.*, 2004) and the clc element of *Pseudomonas* spp. strain B13 (Ravatn *et al.*, 1998).

1.2.7 SXT

A typical representative of the ICE GI family is the SXT element initially described in *V. cholerae* O139 (Burrus *et al.*, 2006). Since the first (1992) characterization of the SXT element, 25 other members of this family, including the R391 ICE present in *Providencia rettgeri* (Coetzee *et al.*, 1972), have been described (Burrus *et al.*, 2006). SXT is a 99.5kb ICE with a G+C content very close to the genome average G+C content of *V. cholerae* (47.05% and 47.69% respectively). The average gene density of SXT is lower than the genome average (0.87 and 0.93 genes/kb

respectively). The insertion point of SXT is the 5' end of the *prfC* gene that encodes a peptide chain release factor 3 (RF3) (Hochhut and Waldor, 1999); upon integration, the *prfC* gene is disrupted and the SXT element replaces this fragment with a different, novel 5' coding sequence such that the *prfC* gives a functional RF3 product after the integration of SXT (Burrus *et al.*, 2006). The SXT element is flanked by a set of 17bp DRs. The same *prfC* locus serves as an insertion point for the closely related ICE element R391 (Hochhut *et al.*, 2001); however, SXT can integrate in different loci if the *prfC* gene is absent (Burrus and Waldor, 2003).

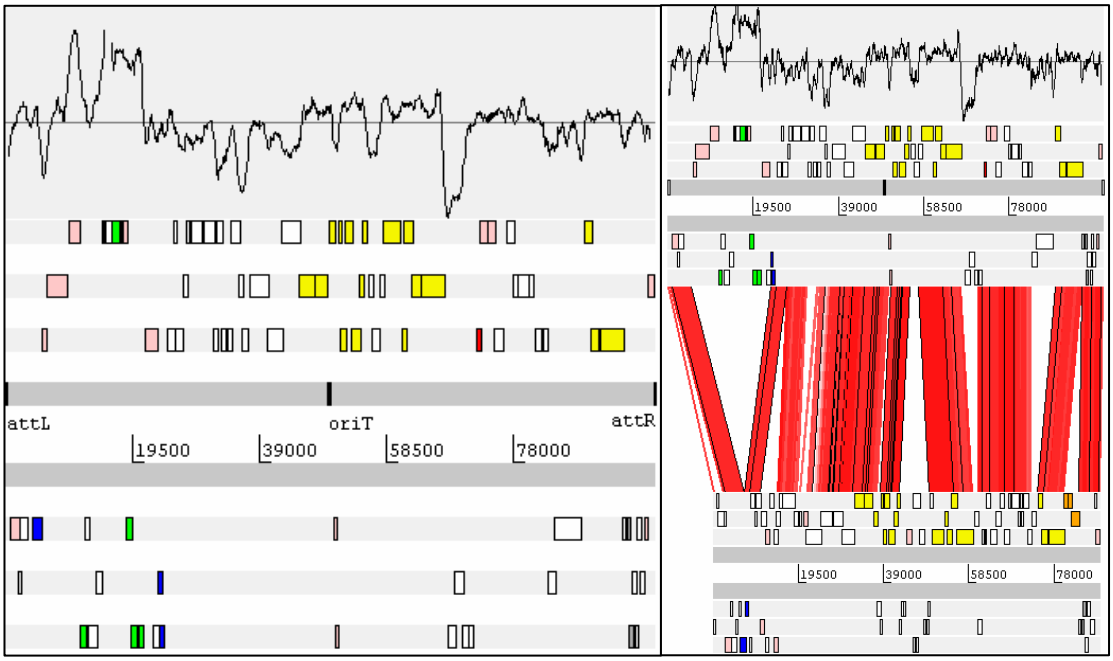

Figure 1.10: **Left.** SXT structure. Colour scheme: Pink; integrase, transposase and phage related, yellow; conjugation, green; antibiotic resistance, dark blue; UV repair, red; single-stranded DNA binding protein, grey; transcriptional activator, and orange; mercury resistance. **Right.** ACT comparison of the SXT element in *V. cholerae* (top) and the R391 in *P. rettgeri* (bottom).

SXT carries a tyrosine recombinase integrase gene that belongs to the λ family of integrases that is key for the excision and integration of the SXT element (Burrus *et al.*, 2006). The transfer of SXT involves excision, the formation of a circular extra-chromosomal intermediate, conjugal transfer via the T4SS to the recipient cell and integration in the host

chromosome. SXT can also mobilize *in trans* other non-conjugative plasmids as well as chromosomal genomic sequence under an Hfr-like mechanism (Burrus *et al.*, 2006). SXT is very similar to the SGI-1 island of Typhimurium DT104, both carrying antibiotic resistance genes, both integrating site-specifically at a given locus via an integrase encoded gene product and they both have mosaic structure (Mulvey *et al.*, 2006); however SGI-1 is a mobilizable but not self-transmissible ICE in contrast to the SXT element.

In terms of gene content almost 50% of the SXT genes are not involved in the mobilization process. A set of genes encoding for antibiotic resistance (streptomycin, chloramphenicol, trimetroprim and suflamethoxazole) is located at the 5' end of the SXT element (Figure 1.10). Further downstream there is a set of two genes involved in UV repair processes; those two genes are also present in R391. Towards the middle and the 3' end of SXT there are three *tra* gene operons. The first is involved in DNA processing and the other two in pilus assembly and mating pair stabilization; all three *tra* operons are very well conserved in terms of gene content and sequence similarity between the SXT and the R391 element (Figure 1.10). The components of the SXT element responsible for conjugation show also sequence similarity to the R27 plasmid of Typhi (Sherburne *et al.*, 2000). Genes involved in the regulation of the SXT are located at the 3' end of this element; two genes (*setC* and *setD*) show sequence similarity to the flagellar regulators *flhC* and *flhD* (Kutsukake *et al.*, 1990) and are key factors for the transcriptional regulation of the SXT (Beaber *et al.*, 2002), while a third gene (*setR*) shows sequence similarity to the CI repressor of λ phages (Beaber *et al.*, 2002).

1.2.8 HPI

There are 11 species of *Yersinia*, a Gram-negative, rod-shaped bacterium. *Y. pestis* is the causative agent of plague, a systemic invasive disease (Perry and Fetherston, 1997), known as "The Black Death". Throughout human history three human pandemics (6-8th centuries, 14-19th centuries,

19th century-today) attributed to *Y. pestis*, have been reported (Perry and Fetherston, 1997). *Y. pestis* is a blood-borne pathogen that evolved from the gastrointestinal pathogen *Y. pseudotuberculosis* approximately 1,500-20,000 years ago (Achtman *et al.*, 1999).

The availability of iron in the environment of microorganisms is essential for their survival. In mammals, iron is usually bound to proteins such as ferritin, haemoglobin, lactoferrin and transferrin. For the direct utilization of iron from the environment, bacteria often synthesise for low-molecular weight binding modules with high affinity to iron termed siderophores (Mietzner and Morse, 1994) that carry iron atoms into the bacterial cytosol via periplasmic, outer and inner membrane transport proteins (Carniel, 2001).

In *Yersinia*, the siderophore system termed yersiniabactin is encoded on a genetic locus, the High Pathogenicity Island (HPI) (Carniel *et al.*, 1996), a PAI that enables *Yersinia* to kill mice in very low dosages. The size of HPI varies from 36kb in *Y. pestis* and *Y. pseudotuberculosis* to 43kb in *Y. enterocolitica* with an average G+C content of 56%, significantly higher than the average genome-wide G+C content (46-50%). The gene density of HPI is lower than the genome average (0.52 and 0.84 respectively), and the boundaries of HPI are defined by a set of imperfect DRs of 24bp on each side of the island corresponding to the DNA sequence at the 3' end of a $tRNA^{Asn}$ locus; although three copies of the $tRNA^{Asn}$ gene exist in the *Yersinia* chromosome, only in *Y. pseudotuberculosis* HPI can insert into any of those three loci, whereas for the other two species (*Y. pestis* and *Y. enterocolitica*) only one, specific $tRNA^{Asn}$ locus serves as the integration site (Buchrieser *et al.*, 1998).

In terms of gene content, the core of HPI consists of 12 CDSs (Figure 1.11) that show overall higher G+C content than the remaining CDSs present in HPI (58% and 46% respectively), suggesting that HPI is probably a mosaic mobile element. Five CDSs encode products for the biosynthesis of the yersiniabactin system, including two high molecular-weight proteins (HMWP1 and HMWP2), a putative salicylate synthetase

(YbtS), a putative thioesterase (YbtT), YbtE that adenylates salicylate and YbtU, a protein of unknown function (Carniel, 2001). Components for the transport of the yersiniabactin-iron complex are encoded by three other CDSs, namely fyuA (encoding an outer membrane protein), ybtP and ybtQ (encoding inner membrane ABC transporters). Two other CDSs, ybtA and ybtX encode for a transcriptional regulator (AraC type) and a signal transducer respectively. At the 5' end of the HPI a bacteriophage P4-like integrase is located immediately downstream of the $tRNA^{Asn}$ locus; this observation leaves open the possibility that HPI might have been originally acquired from another bacterium via a bacteriophage element (transduction) (Carniel, 1999), although conjugation is another proposed mobilization mechanism of the HPI (Antonenka *et al.*, 2005). A putative candidate donor of the HPI is the genome of *Klebsiella* (in which HPI is also found integrated) with an average G+C content of 56-57%, very close to the HPI G+C% content (Carniel, 1999).

In terms of sequence relatedness, there are two distinct evolutionary forms of HPI, HPI present in *Y. pestis* and *Y. pseudotuberculosis* and HPI present in *Y. enterocolitica* (Rakin and Heesemann, 1995). The 3' end of the latter ends 12.8kb downstream of fyuA gene and includes four IS elements and seven more CDSs of unknown function (Carniel, 2001); the overall G+C content of the additional 12.8 DNA sequence is significantly lower (38.6%) than the remaining of the HPI (58%).

In terms of phylogenetic distribution, apart from the Yersinia genus, HPI has also been found in *E. coli* (including commensal strains), *Citrobacter diversus* and *Klebsiella* isolates (Bach *et al.*, 2000; Schubert *et al.*, 1998), suggesting that HPI although firstly described in the genome of highly pathogenic *Yersinia*, is also present in non-pathogenic organisms. This observation further supports the concept that the phenotypic properties of GIs are strongly depended on the specific niche of the bacterial host (Hacker and Carniel, 2001; Schmidt and Hensel, 2004); indeed no direct pathogenic components are present on the HPI, rather

the HPI cargo makes the survival of the bacterium possible in this niche, without itself producing directly damaging effects on the host cells (Kaper and Hacker, 1999).

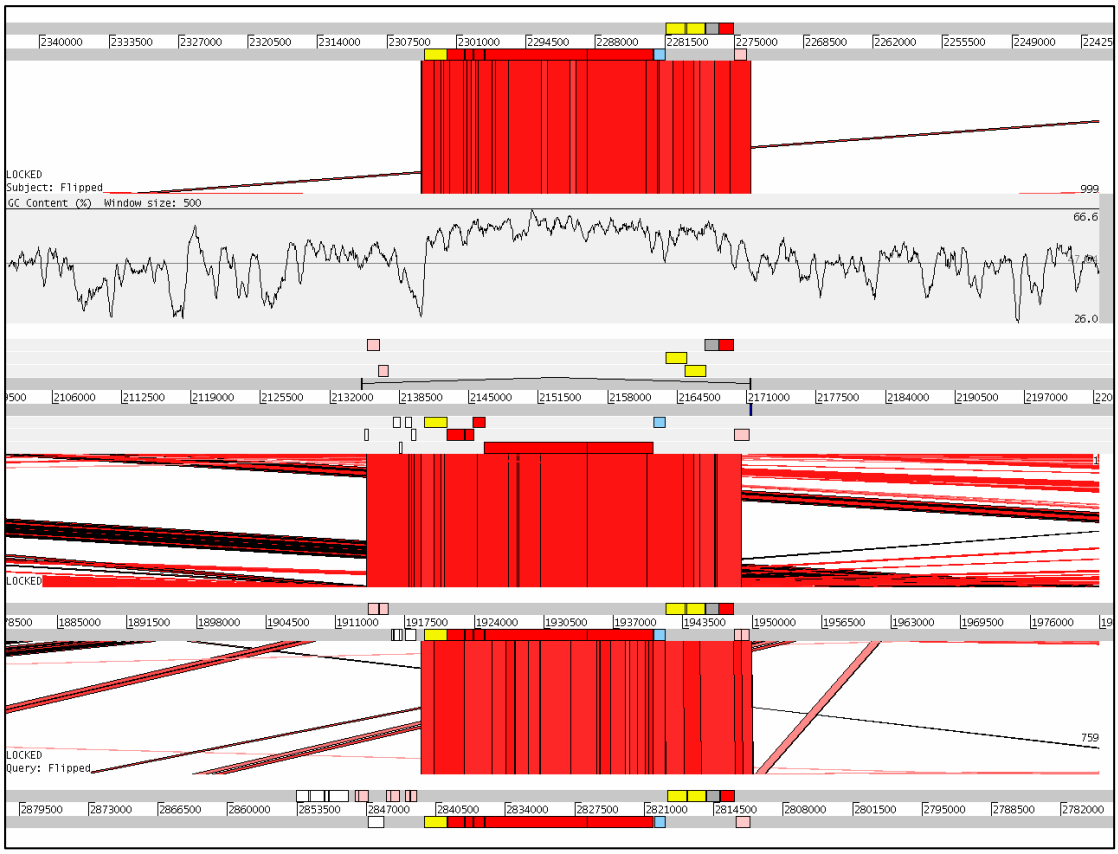

Figure 1.11 Comparison of the HPI present in different genomes. From top to bottom: Enteroaggregative *E. coli* 042, *Y. pestis* CO92, *Y. pseudotuberculosis* IP32953 and *Y. enterocolitica* 8081. Colour scheme: Light pink; integrase, transposase, red; yersiniabactin biosynthesis, yellow; yersiniabactin transport, grey; signal transducer, light blue; transcriptional regulator, dark blue; tRNA, and white; unknown function. The imperfect DRs flanking the HPI are shown as two joined black-coloured features.

Although HPI is a non self-transferable GI with unknown mechanism of dissemination, it is a highly unstable mobile element (Carniel, 2001). In *Y. pseudotuberculosis* the HPI excises, via the P4-like integrase precisely and spontaneously at a frequency of 10^{-4} (Buchrieser *et al.*, 1998), while in *Y. pestis*, the excision occurs but it is not precise and involves an extended genomic region of 102kb that includes not only the HPI but also the pigmentation (pgm) locus (Fetherston *et al.*, 1992); the 102kb deletion is probably the result of homologous recombination between the two IS100 elements present at the boundaries of this region,

rather than the result of P4-integrase mediated excision. In *Y. enterocolitica* the P4-like integrase is a pseudogene which further explains the stable integration of HPI in some isolates of that species.

HPI is not the only known example of PAI carrying iron uptake systems; other examples include the aerobactin system in *Shigella flexneri*, two putative siderophore systems in UPEC and the iron transport system of SPI-1 in *Salmonella*, discussed in more detail in the following section.

1.2.9 SPI-1, 2, 3, 4

The majority of SPIs are exceptional PAIs, not only due to their phylogenetic distribution (Figure 1.12), i.e. they are species but not strain specific islands (Hacker *et al.*, 1997), but also due to their atypical PAI structure (Table 1.2) that is not described by the classical GI definition (Hacker *et al.*, 1997; Hacker and Kaper, 2000; Schmidt and Hensel, 2004). The fact that many SPIs lack identifiable repeat elements flanking their boundaries and mobility genes, e.g. integrases, suggests that those mobile elements are probably stably integrated in the *Salmonella* chromosome, perhaps representing very ancient insertions that over time lost their ability to mobilize (Wong *et al.*, 1998).

Table 1.2: Structural features and properties of four *Salmonella* Pathogenicity Islands: SPI-1, SPI-2, SPI-3 and SPI-4. For easy of comparison, the genome average G+C content and gene density of *Salmonella* is 52.09% and 0.913 genes/kb respectively.

SPI	SIZE	G+C%	DRs	Integrase	RNA	Gene Density	Mosaic	Virulence properties	Functional module
SPI1	39773	45.9	-	-	-	1.058	No	Invasion of epithelial cells	T3SS
SPI2	39740	47.18	-	-	$tRNA^{Val}$	1.006	Yes	Intracellular proliferation	T3SS
SPI3	17348	47.09	-	-	$tRNA^{SelC}$	0.81	Yes	Intramacrophage survival	T5SS
SPI4	24672	44.35	-	-	-	0.28	No	Intramacrophage survival	T1SS

SPI-1 products are essential for the internalization of *Salmonella* into epithelial cells, through a cascade of events that include the

rearrangement of the actin cytoskeleton, membrane ruffling and signal transduction interference (Patel and Galan, 2005). Those host responses are initiated by effector proteins secreted via a T3SS (known as Inv-Spa) present on SPI-1 (Figure 1.13); some effector proteins (e.g. sopB and sopE) are encoded on other genomic regions (SPI-5 and an integrated bacteriophage respectively) within the *Salmonella* chromosome (Hardt *et al.*, 1998; Wood *et al.*, 1998).

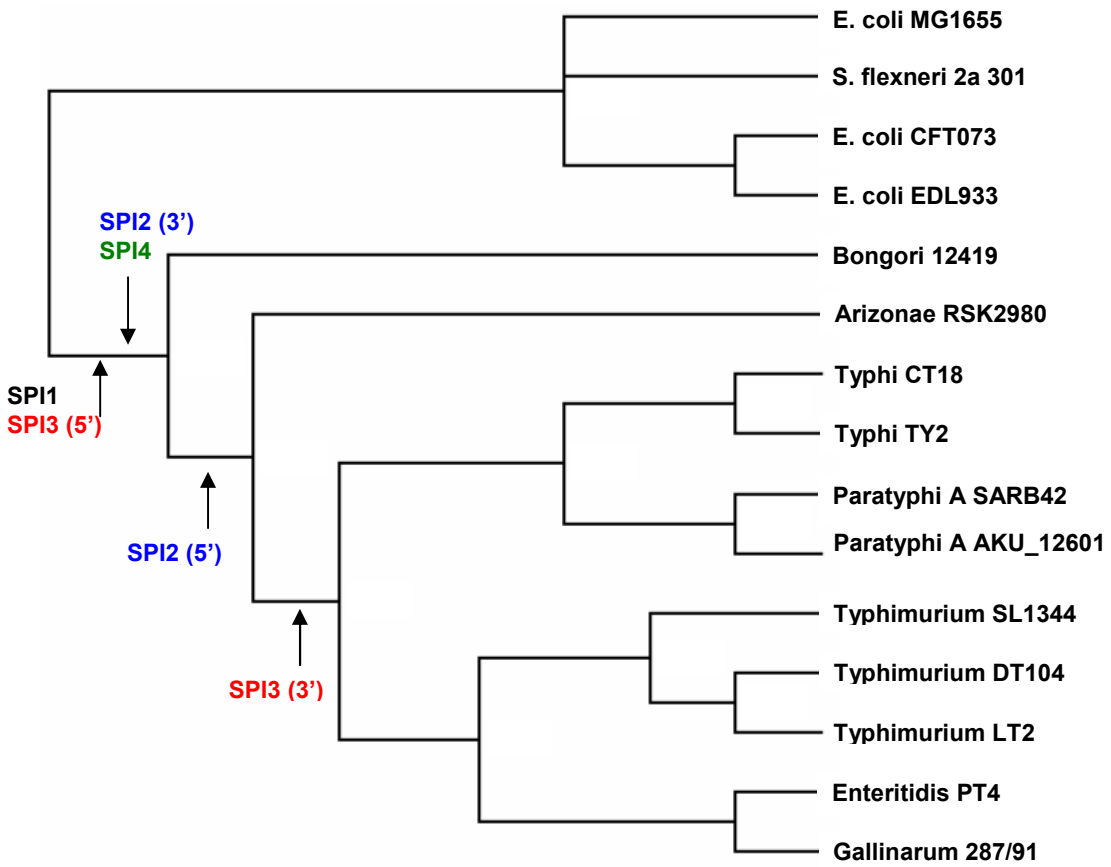

Figure 1.12: Phylogenetic distribution of four *Salmonella* Pathogenicity Islands: SPI-1, SPI-2, SPI-3 and SPI-4 in the *Salmonella* lineage. The cladogram shows the phylogenetic relationship between 11 *Salmonella* strains and four outgroups (*E. coli* MG1655, *S. flexneri* 2a 301, *E. coli* CFT073 and *E. coli* EDL933) ignoring branch length.

The expression of SPI-1 products is under the regulation of the PhoP-PhoQ two component system (Groisman, 2001) via the transcriptional down-regulation of SPI-1 master regulator HilA (Bajaj *et al.*, 1996). SPI-1 represents a very old HGT event in the evolution of

salmonellaea, acquired at the bottom of the *Salmonella* lineage (Ochman and Groisman, 1996), very close to its divergence time from its sister lineage *E. coli* approximately 100-140 Myr ago (Figure 1.12) (Doolittle *et al.*, 1996; Ochman and Wilson, 1987). Although two pseudogenes with transposase domains exist at the 5' end of SPI-1, the mobilization mechanism of SPI-1 remains unknown; however partial deletions of the SPI-1 locus have been observed in environmental *Salmonella* serovars (Ginocchio *et al.*, 1997).

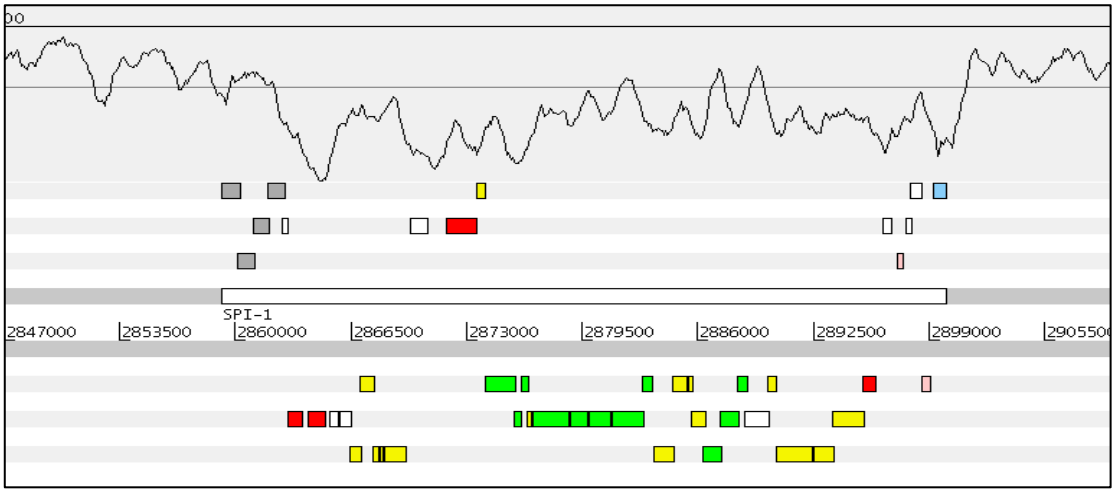

Figure 1.13: Gene content of SPI-1. Colour scheme: Grey; iron transport, red; regulation, yellow; Type III Secretion System, green; secreted proteins-effectors and chaperones, light pink; transposase, light blue; serine/threonine phosphatase, and white; hypothetical. At the top of the figure the G+C% content is shown with a window size of 1kb.

SPI-2 carries 32 CDSs that encode a T3SS (known as Spi-Ssa) apparatus, effector proteins, chaperones and a two-component regulatory system (Figure 1.14) (Hensel *et al.*, 1997; Hensel *et al.*, 1998; Worley *et al.*, 2000); the latter, known as SsrAB, controls the expression of SPI-2 in response to at least two environmental stimuli (pH and inorganic phosphate) (Lober *et al.*, 2006). A chromosomal region in the *Y. pestis* genome shows sequence and gene order similarity to the T3SS of SPI-2 (Parkhill *et al.*, 2001). Moreover the T3SS of SPI-2 is also very similar to the T3SS encoded on the LEE island (Kaper and Hacker, 1999). Although SPI-1 is essential for cell invasion, the components of the SPI-2 locus are

important for the intracellular proliferation of *Salmonella* within the *Salmonella* containing vacuole (SCV) by means of the T3SS effectors that affect the vesicular trafficking within the host cell by preventing the action of phagocyte oxidase and nitric oxide synthetase (Chakravortty *et al.*, 2002; Kuhle *et al.*, 2006; Uchiya *et al.*, 1999; Vazquez-Torres *et al.*, 2000).

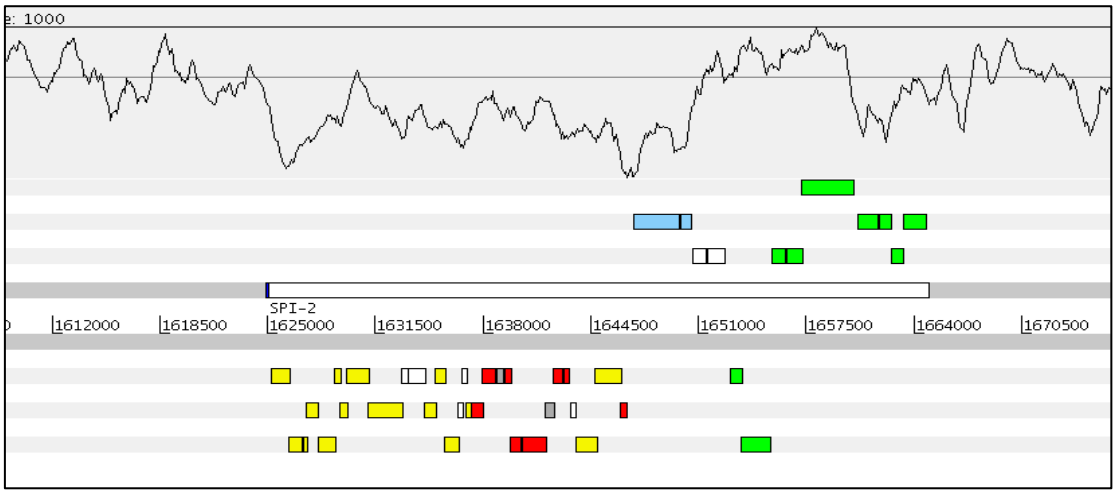

Figure 1.14: Gene content of SPI-2. Colour scheme: Grey; chaperones, red; secreted proteins-effectors, yellow; Type III Secretion System, green; tetrathionate reductase operon, dark blue; tRNA, light blue; two-component response regulator, and white; hypothetical. At the top of the figure the G+C% content is shown with a window size of 1kb.

SPI-2 is a mosaic PAI of at least two independent acquisitions, which occurred at distinct times throughout the evolution of salmonellae (Hensel *et al.*, 1999; Vernikos *et al.*, 2007). The first part (~25kb) of SPI-2 that includes the components of the T3SS and the two-component regulatory system is present only in *S. enterica* species while the second part (~14kb) of SPI-2 (towards the 3' end) encoding for a tetrathionate reductase gene cluster (*ttr*) (Hensel *et al.*, 1999) represents a much older insertion present both in *S. bongori* and *S. enterica* species (Figure 1.12). The mosaic nature of SPI-2 is also evident from its % G+C content; the recent insertion has a G+C content of 44%, while the older one (ttr operon)

has a G+C content much closer to the genome average G+C (52.8% and 52.1% respectively) (Figure 1.14).

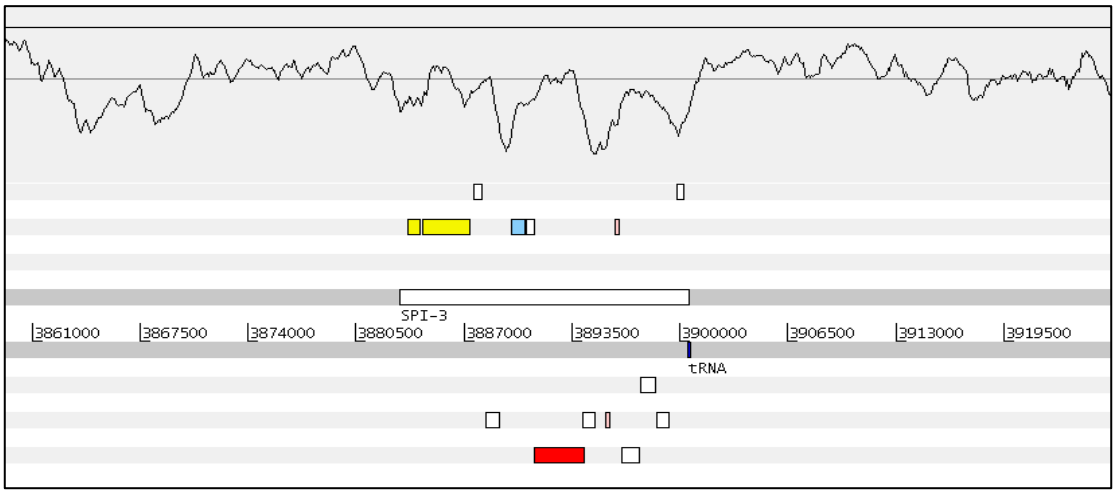

Figure 1.15: Gene content of SPI-3. Colour scheme: Red; autotransporter/T5SS, yellow; magnesium transport, dark blue; tRNA, light blue; transcriptional regulator, white; hypothetical, and light pink; transposase. At the top of the figure the G+C% content is shown with a window size of 1kb.

SPI-3 is another example of a mosaic PAI that is the result of more than one independent acquisition (Blanc-Potard *et al.*, 1999; Vernikos *et al.*, 2007). The first part of SPI-3 carries the *mgtCB* operon, a high-affinity magnesium transport system that is essential for the intramacrophage survival of *Salmonella* in the low Mg^{2+} environment of the phagosome (Snavely *et al.*, 1991) and a set of two CDSs of unknown function. This first part (~6kb) of SPI-3 represents an old insertion present at the bottom of *Salmonella* lineage (*S. bongori* and *S. enterica* species) (Figure 1.12) with an average G+C content of 50.3%.

The second, low G+C (45.6%) part (~11kb) of SPI-3 carries 10 CDSs encoding a putative transcriptional regulator (marT), a T5SS (misL; T5SS are also known as autotransporters (Henderson *et al.*, 1998)), two transposases, a putative exported protein (fidL) and five CDSs of unknown function (Figure 1.15); both *marT* and *misL* are pseudogenes in Typhi CT18. The second part of SPI-3 is a more recent HGT event present in *S. enterica* species but absent from *S. bongori* and *S. arizonae* (Figure 1.12).

The mechanism of mobilization of SPI-3 remains unknown as although it is inserted in the proximity of a tRNA locus, it lacks an identifiable integrase and repeats flanking its boundaries.

SPI-4 is an exceptional SPI in that it shows sequence composition-based but not phylogenetic mosaicism (Figure 1.16). The first (~9kb) 5' part of SPI-4 has an average G+C content of 38% while the middle (~7kb) and the 3' part (~9kb) of SPI-3 have a G+C content of 54% and 44% respectively. In terms of gene content, SPI-4 encodes a putative T1SS, a fairly simple secretion apparatus (consisting of an ABC transporter, a periplasmic protein and an outer membrane protein) that is important for intramacrophage survival (Wong *et al.*, 1998).

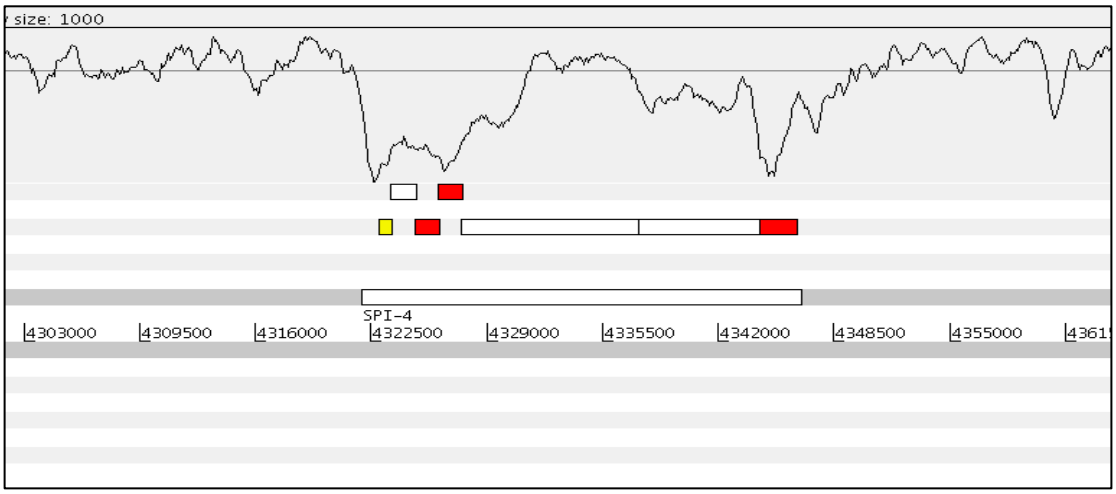

Figure 1.16: Gene content of SPI-4. Colour scheme: Red; Type I Secretion System, yellow; putative exported protein, and white; hypothetical. At the top of the figure the G+C% content is shown with a window size of 1kb.

A large repetitive protein (SiiE) that is the putative substrate of the T1SS present on SPI-4 is a very large nonfimbrial adhesin that binds to the surface of the epithelial cells (Gerlach *et al.*, 2007). The expression of SPI-4 that mediates adhesion and SPI-1 that mediates invasion in the epithelial cells seem to be co-regulated (Gerlach *et al.*, 2007) by the invasion response regulator (sirA) (Ahmer *et al.*, 1999), illustrating the fine-tuning, "cross-talk" of mobile elements. In terms of structure SPI-4 lacks most of the classical GI-related structures including mobility genes,

repeats and tRNA and represents a very early HGT event in the evolution of salmonellaea present at the bottom of the *Salmonella* lineage (Figure 1.12).

1.2.10 Metabolic Island

The *B. cenocepacia* island (cci), is an unusual GI that shows distinct functional mosaicism, encoding both metabolic and pathogenicity-related components (Figure 1.17) (Baldwin *et al.*, 2004). cci is a 44kb, low G+C content (62%, genome average 67.3%) island that encodes 50 CDSs and shows overall significantly higher gene density than the rest of the *B. cenocepacia* chromosome (1.13 and 0.872 respectively); cci is flanked by a set of 17bp DRs on each side.

The *B. cepacia* epidemic strain marker (BCESM), a 1.4kb DNA sequence that encodes a putative transcriptional regulator (esmR) (Mahenthiralingam *et al.*, 1997) is part of the cci element. BCESM constitutes an epidemiological marker characterizing virulent *B. cenocepacia* isolates that infect individuals with cystic fibrosis.

In terms of gene content there are at least seven functional classes of CDSs present on cci. At the 5' end of cci, there is a cluster of eight CDSs encoding products involved in arsenic and antibiotic resistance (Figure 1.17). Further to the right of this locus there is a region carrying an N-acyl homoserine lactone (AHL) synthase gene and its transcriptional regulator (autoinducer synthesis loci) and a cluster of CDSs with sequence similarity to fatty acid biosynthesis components and a set of three transposases. Four putative transcriptional regulators, including the *esmR* gene are located downstream of this region. The remaining half of cci carries eight CDSs encoding products involved in amino acid metabolism and transport, eight conserved hypothetical CDSs of unknown function and a cluster of stress response CDSs.

Clearly, the cci element represents indeed a functional mosaic; the autoinducer synthesis locus is associated with quorum sensing and has been shown to be involved in the pathogenicity related phenotypic

properties of *B. cenocepacia* (Lewenza *et al.*, 1999; Sokol *et al.*, 2003), while at the same time components involved on a wide range of metabolic properties including amino acid and fatty acid biosynthesis are also present. This observation along with the overall analysis of mobile elements discussed in this section raises an issue of how to develop a reliable classification system of GIs; not only do compositional, phylogenetic and structural-based classification systems collapse due to the extensive mosaicism of GIs, but also functional-based systems fail to provide a universally applicable classification model.

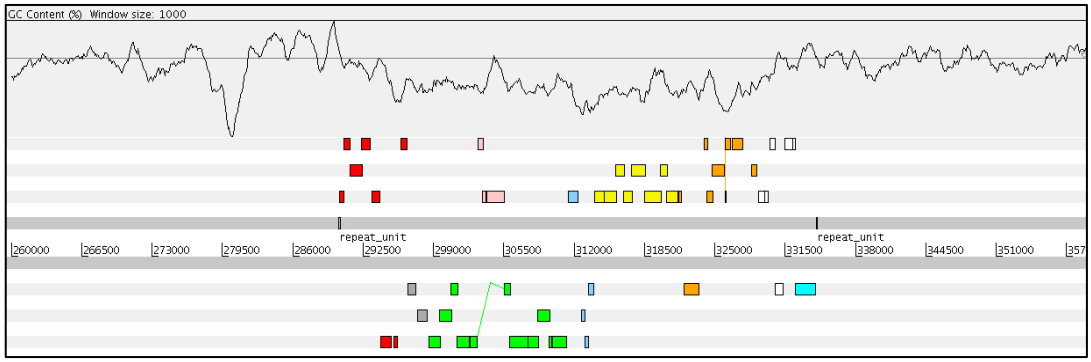

Figure 1.17: Gene content of cci. Colour scheme: Red; arsenic and antibiotic resistance, yellow; amino acid metabolism, light blue; transcriptional regulator, white; stress-response, light pink; transposase, green; fatty acid biosynthesis, grey; autoinducer synthesis, cyan; putative sulphate transporter, and orange; conserved hypothetical. The 17bp DRs flanking the cci element are shown as black-coloured features. At the top of the figure the G+C% content is shown with a window size of 1kb.

1.3 *In silico* prediction of GIs

This section reviews some of the current methodologies for the computational prediction of GIs discussing the limitations and advantages of each method.

At the time of insertion, horizontally acquired DNA reflects mainly the sequence composition of its donor. Over time this horizontally acquired DNA converges towards the sequence composition of its new host and eventually becomes compositionally indistinguishable from the backbone of the host genome, a time dependent process known as amelioration

(Lawrence and Ochman, 1997). Consequently recent HGT events are, in theory, easier to predict, by means of compositional analysis, compared to older insertions. However several exceptions apply; if the sequence composition of the donor genome is very close to the acceptor then even in the case of very recent HGT events, their prediction will be non-trivial. Conversely, core components of the host genome that are not horizontally acquired but deviate compositionally due to specific well-preserved functional constraints (e.g. the rRNA genes) can be falsely predicted as HGT events (Vernikos and Parkhill, 2006; Vernikos *et al.*, 2007).

Based on this principle, i.e. the majority of horizontally acquired DNA sequences are likely to deviate from the host backbone composition, several indices have been exploited to capture compositional biases (Table 1.3). These indices can lead to the identification of GIs; however, for the prediction of PAIs further analysis is required to investigate the contribution of these elements to virulence.

Often combination of more than one index can be used for a more efficient identification of "alien" regions. For example Lawrence and Ochman (Lawrence and Ochman, 1997) and Karlin *et al.* (Karlin *et al.*, 1998) utilized the codon bias and the codon adaptation index (CAI) (Sharp and Li, 1987) to identify atypical regions. For a native gene with atypical G+C content resulting from selection over preferred codons, both the chi-square (codon bias) and CAI values will be high. On the other hand, for a highly biased "alien" gene, the chi-square value will be high but the CAI value will be low, given that it is biased but not in a host-specific manner resulting in the well-known "rabbit-like" codon bias-CAI plot (Figure 1.18).

In a similar multi-index approach, Karlin (Karlin, 2001) applied the % G+C content, dinucleotide frequency difference, codon and amino acid bias to detect alien gene clusters. Most of these indices cause overlapping peaks predicting the same atypical regions; however, there are cases in which one or more indices might perform poorly in the detection of compositionally deviating regions, depending on the level of compositional bias (see Figure 1c therein).

Yoon *et al.* (Yoon *et al.*, 2005) combined sequence similarities and composition abnormalities to predict PAIs rather than GIs in general.

Table 1.3: Commonly used indices for the identification of regions with atypical composition.

Indices	Description
Codon Adaptation Index (CAI) (Sharp and Li, 1987)	A measure of the relative adaptiveness of the codon usage of a gene towards the codon usage of highly expressed genes.
Frequency of Optimal codons (Fop) (Ikemura, 1981)	The ratio of optimal codons to synonymous codons.
Codon Bias Index (CBI) (Bennetzen and Hall, 1982)	A measure of directional codon bias; it measures the extent to which a gene uses a subset of optimal codons.
Effective number of codons (NC) (Wright, 1990)	This index quantifies how far the codon usage of a gene departs from equal usage of synonymous codons. A gene utilizing only one codon per aa has the strongest bias and the minimum index value, 20; a gene using all codons equally has a value of 61.
GC content	Measures the frequency of guanine or cytosine.
GC_1 and GC_3 content	These indices measure the frequency of guanine or cytosine in the 1st and 3rd codon position respectively.
δ* difference (Karlin, 1998)	Is the average absolute dinucleotide relative abundance difference. Dinucleotide relative abundance values are calculated as the dinucleotide frequency normalized over the product of the frequencies of the two mononucleotides of the given dinucleotide.
Codon usage contrasts (Karlin, 2001; Karlin and Mrazek, 2000)	Compares codon biases of the gene set of each window to the average gene codon usages.
Amino acid contrasts (Karlin, 2001)	Compares amino acid biases of proteins in each window relative to the average proteome amino acid frequencies.
High order motifs (Sandberg *et al.*, 2001; Tsirigos and Rigoutsos, 2005)	Compares the frequency of words W of size n of a sliding window against the corresponding ones of the genome e.g. for words of size n=8 , the total number of different words is 4^8 (= 65,536).
Translational efficiency (P2) (Gouy and Gautier, 1982)	This index describes the proportion of codons conforming to the intermediate strength of codon-anticodon interaction energy rule of Grosjean and Fiers. For a gene with uniform codon usage P2 = 0.5.
Intrinsic codon bias index (ICDI) (Freire-Picos *et al.*, 1994)	An index to estimate codon bias of genes from species in which optimal codons are unknown. Its correlation with other index values, like CBI or NC, is high.
Scaled Chi-square (Shields and Sharp, 1987)	This index measures the degree of bias due to a non uniform use of synonymous codons of a gene, using uniform synonymous codon usage as the expectation. These values are scaled by division by the number of codons in the gene.

They utilized BLAST (Altschul *et al.*, 1997) and BLAT (Kent, 2002) to identify homologues of known PAIs in a given genome, and G+C% content and codon usage bias to identify compositional deviating regions from the genome backbone. Overlapping (atypical composition and PAI homologs)

regions were reported as candidate PAIs. Although this approach goes one step ahead, predicting PAIs instead of GIs it is restricted to predict PAIs that have similar gene content to previously identified ones, thus it is unsuitable for the prediction of novel PAIs. A very comprehensive web resource of PAIs, utilizing this methodology is available at http://www.gem.re.kr/paidb/ (Pathogenicity Island Database – PAI DB) (Yoon *et al.*, 2007).

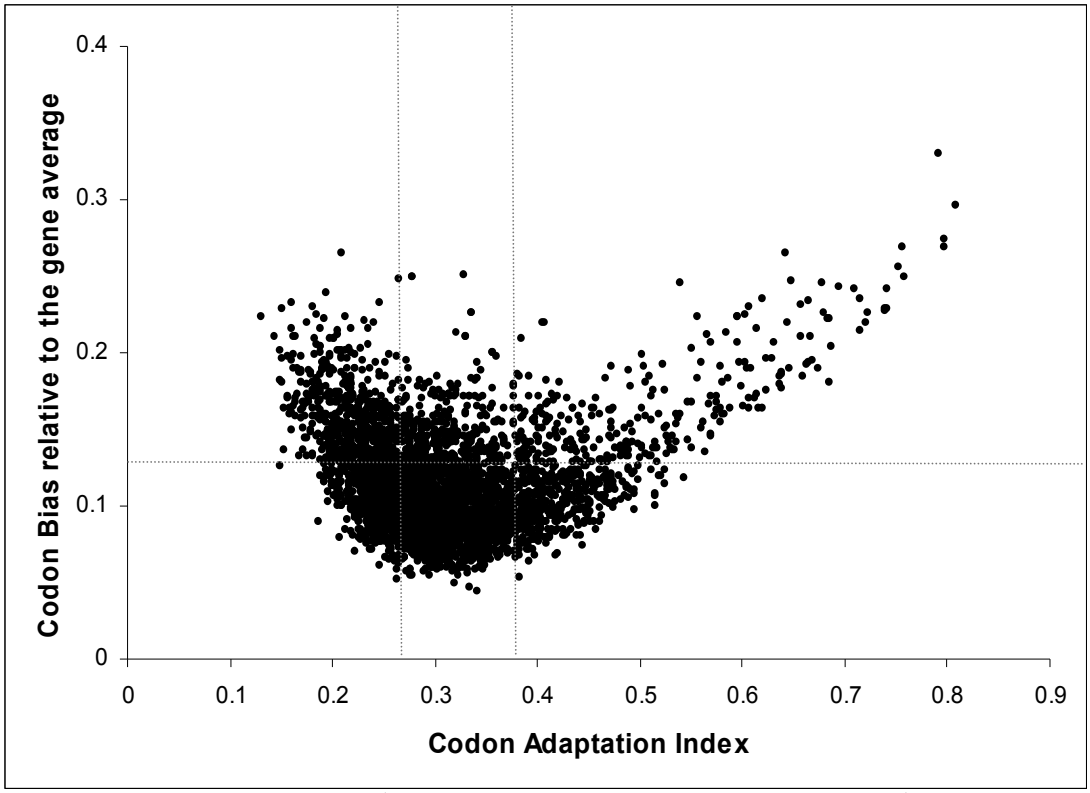

Figure 1.18: The codon bias (relative to the gene average codon usage) of the genes (>300bp) present in *S. typhi* CT18 is plotted against their codon adaptation index (CAI) value; CAI values have been calculated using the reference set of highly expressed genes, proposed by Sharp and Li (Sharp and Li, 1986), using the genome of *E. coli.*

Garcia-Vallve *et al.* (Garcia-Vallve *et al.*, 2003) developed a statistical method for the prediction of horizontally transferred genes, using the G+C% content, codon usage, amino acid usage, and gene position analysis; a web resource (HGT-DB) implementing this methodology is accessible through http://www.tinet.org/~debb/HGT/. It is a useful

database of HGT events however there is no user-defined option for uploading the sequence of interest thus it is restricted to the genome collection, updated by the authors. Moreover this approach relies on gene finding methods as it utilizes sliding window over genes and not over raw genomic sequence, consequently it depends on existing annotation. Methods for the prediction of HGT events are normally expected to precede the annotation procedure, aiding/supporting the annotation pipelines, rather than extending pre-existing annotation.

Mantri *et al.* (Mantri and Williams, 2004) developed an algorithm, Islander, exploiting the principle that islands tend to be preferentially integrated within RNA loci. Islander produces a list of tRNA and tmRNA genes and uses each as a query for a BLAST search. Although it is an innovating approach, in terms of independence from compositional indices, it is restricted only to very well-structured islands; for example candidate islands that do not contain an integrase gene or are longer than 200kb or the integration site is not a tRNA locus are rejected. Islander is accessible at http://kementari.bioinformatics.vt.edu/cgi-bin/islander.cgi.

IslandPath (Hsiao *et al.*, 2003) is another web-based suite (http://www.pathogenomics.sfu.ca/islandpath/) for the prediction of GIs utilizing the G+C% content, dinucleotide bias, RNA and mobility (e.g. integrase, transposase) gene information but it depends mainly on other resources and requires manual intervention. For example, RNA location information is obtained from NCBI (http://www.ncbi.nlm.nih.gov/), mobility genes are identified by keyword scanning against NCBI and supplemented with COG (Tatusov *et al.*, 2001) classification information and the overall methodology is reliant on gene-finding methods.

Tsirigos and Rigoutsos (Tsirigos and Rigoutsos, 2005; Tsirigos and Rigoutsos, 2005) and Sandberg *et al.* (Sandberg *et al.*, 2001) utilized higher-order nucleotide sequences (motifs) to overcome the weak discrimination power of di- and trinucleotide models. Both studies provide data in favour of the higher order motifs, with the optimal template size to be 8-9 nucleotides (nt). Tsirigos *et al.* used sliding window over genes

(reliant on gene prediction) while Sandberg *et al.* used windows over raw genomic sequence. Moreover Tsirigos *et al.*, in order to evaluate the performance of their method, simulated HGT events inserting genes from a gene pool into several genomes. This kind of approach however does not take into account the amelioration process that takes place over time on horizontally transferred genes; thus it is rather focused on recently integrated genes. The results of this analysis are accessible at http://cbcsrv.watson.ibm.com/HGT/.

A score-based identification of GIs (SIGI) is another methodology for the *in silico* prediction of GIs and their putative origin, utilizing a codon frequency-based approach (Merkl, 2004). A single-gene sliding window is implemented to search a query genome and the codon usage of each gene is compared to a non-redundant set of 400 codon usage tables representing different microbial species; clusters of consecutive genes that deviate from the genome average codon usage are determined as putative GIs and the species with the closest codon usage is inferred to be the most likely donor of those putative GIs. SIGI represents a novel approach for the prediction of GIs that reports also the putative source of horizontally acquired genes; however it is dependent on gene finding methods, is restricted to a collection of microbial genomes updated by the authors (no user-defined sequence uploading option) and utilizes only the codon usage bias to detect atypical regions. The results of this analysis are accessible at http://www.g2l.bio.uni-goettingen.de/software/sigi/sigi.htm.

MobilomeFINDER (http://mml.sjtu.edu.cn/MobilomeFINDER) is an interactive web suite for the prediction of GIs or mobile elements (mobilome) in general (Ou *et al.*, 2007). MobilomeFINDER's novelty relies on the fact that the actual prediction of GIs is based on a multi-factorial methodology exploiting comparative, composition and structural-based approaches integrating not only *in silico* but also experimental data making it applicable both on fully sequenced but also on unsequenced query strains. The limitation of such comparative-based methodologies

however relies on the fact that for newly sequenced genomes without identified close relatives the prediction of GIs is not possible.

In the first introductory part of this thesis, I have given a broad overview of various aspects related to HGT, the biological interest behind the study of such DNA exchanges and their impact in driving bacterial evolution; I have also discussed the GI structural definition, its limitations, and I have described several representative examples of GIs showing the extreme compositional, functional and structural variation of those mobile elements. These discussions have the aim of addressing the problem at hand and the focus of this thesis: How do we reliably predict and model GI structures in microbial genomes using *in silico* methods given their extreme mosaic nature? In the following chapters of this thesis I will focus mainly on three different, novel approaches for the prediction of GIs; a compositional, a comparative and a structural-based methodology, discussing their advantages and limitations.