In silico prediction of Genomic Islands in microbial genomes

Georgios S. Vernikos

Selwyn College

University of Cambridge

April 2008

A dissertation submitted for
the degree of Doctor of Philosophy
at the University of Cambridge

Declaration

The work presented in this dissertation was carried out at the Wellcome Trust Sanger Institute in Cambridge between April 2005 and April 2008. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. No part of this dissertation has been, or is being submitted for a degree or diploma or other qualification at this or any other university.

Georgios S. Vernikos
Cambridge
April 2008

Summary

Large inserts of horizontally acquired DNA that contain functionally related genes with limited phylogenetic distribution are often referred to as Genomic Islands (GIs). Integration of GIs can in a single step transform a microbial organism, changing radically the way it interacts with its niche, a process that is often referred to as "evolution in quantum leaps". A key aspect in the prediction of GIs is that at the time of their integration in the new host chromosome, their composition reflects mainly the composition of the donor, rather than the host.

The first part of this thesis concerns the design, development and implementation of a novel algorithm, that of the Interpolated Variable Order Motifs, for the composition-based prediction of GIs. This algorithm exploits compositional biases using variable order motif distributions and captures more reliably the local composition of a sequence compared with fixed order methods, overcoming the limitations of the latter. Furthermore, for the optimal localization of the boundaries of each predicted region, the Hidden Markov Model theory was implemented in a change point detection framework, predicting more accurately the true insertion point of candidate GIs.

In the second part of this thesis whole genome based comparative and phylogenetic techniques were used to study the acquisition of horizontally acquired genes in the *Salmonella* lineage in a time dependent manner. The compositional amelioration process was modelled and the relative time of acquisition of those genes was determined on different branches of the *S. enterica* phylogenetic tree.

The aim of the third part of this thesis is to explicitly quantify and model the contribution of genomic features to the GI structure, under a probabilistic framework. A hypothesis free, bottom-up search was implemented and identified approximately 700 genomic regions, including both GIs and randomly sampled regions from three different genera that form my training dataset. A Machine Learning approach was used to exploit the above dataset and study the structural variation of GIs.

The last part of this thesis focuses on the experimental validation of the *in silico* predictions made on a newly sequenced bacterial genome. Applying a PCR-based protocol, the presence and absence of the predicted candidate islands was probed in seventeen unsequenced closely and distantly related strains. The true borders of the predicted islands were confirmed by sequencing across the boundary site in strains lacking the island.

Acknowledgements

Firstly, my deepest personal thanks to my supervisor Julian Parkhill for being the busiest and at the same time the most available "teacher" that shaped, in a very profound and enjoyable way, my academic career; Julian has a remarkable ability to inspire people working next to him.

I would also like to thank members of my thesis committee, George Salmond for discussion on various biological aspects of my project, Richard Durbin and Alex Bateman for ideas and discussion on algorithmic techniques and Thomas Down for replying to my emails using pseudo-code! – It made our communication much easier and fun.

Many thanks to David Carter for making the source code of the biojava implementation of the Relevance Vector Machine available, Nick Thomson for long, late night philosophical (and not only) discussions on microbial genomes; Stephen Bentley for biological discussion and for contributing to a key step of my future career path, Matthew Holden for discussion on various aspects of microbial evolutionary dynamics and the other members of team 81 for advice on different steps of this project.

I also thank: Helena Seth-Smith and Paul Scott, for being extremely patient with my ignorance on lab-based techniques and Andrew Jackson for discussion on aspects of the phylogenetic analysis; Xavier Didelot for advice on whole genome sequence alignments and Matthew Avison for providing the *Stenotrophomonas maltophilia* strains; Manolis Dermitzakis for my very first, off-the-record, coffee-interview that took place in Greece; the Wellcome Trust Sanger Institute for my PhD studentship.

Special thanks to my parents, Stelios and Maria, for letting me take apart their washing machine, in order for me to discover what lies beneath; their very liberal approach to my intellectual curiosity shaped in a very profound way my academic and non-academic way of thinking; my brother, Christos, for convincing me that different is nice.

My love to Andriana (my wife to be) for standing by me no matter what, without whom at the end of this 4-year academic journey, I would only be a scientist.

Contents

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.

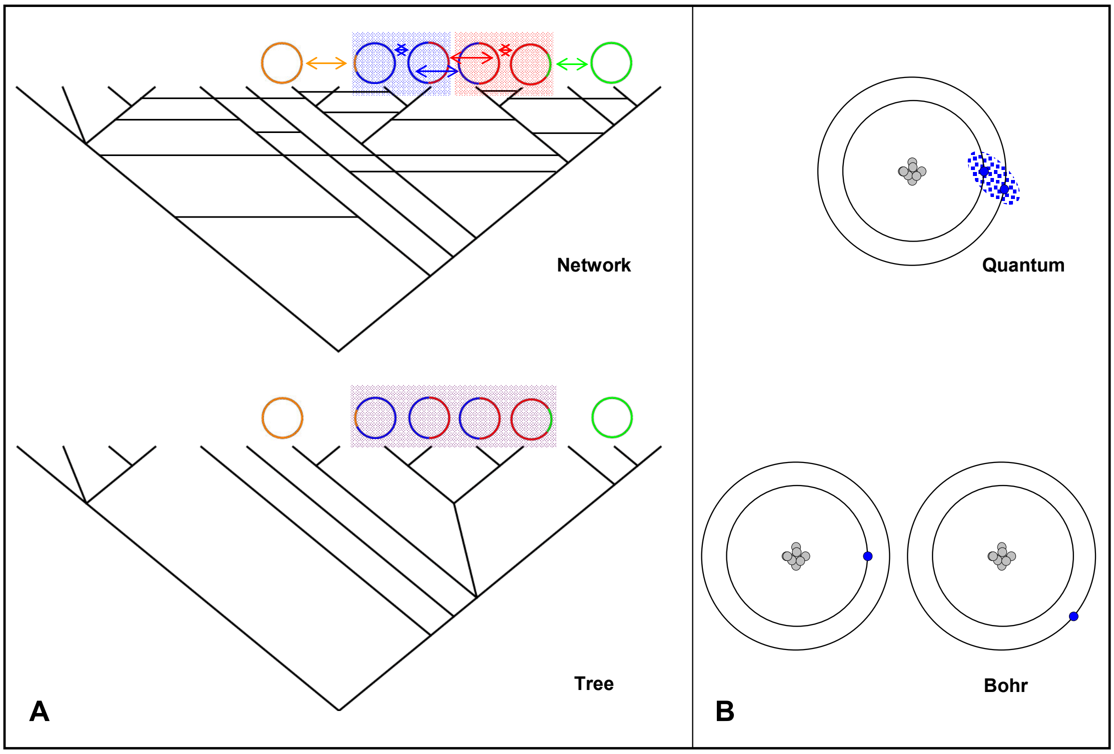


Figure 1.1: **A.** An example of genome mosaicism and the limitation of a bifurcating, tree-based classification system (bottom) for a reliable representation of the true phylogenetic histories of lineages exposed to high rates of genetic flux, compared to a phylogenetic network (top). **B.** Atom structure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

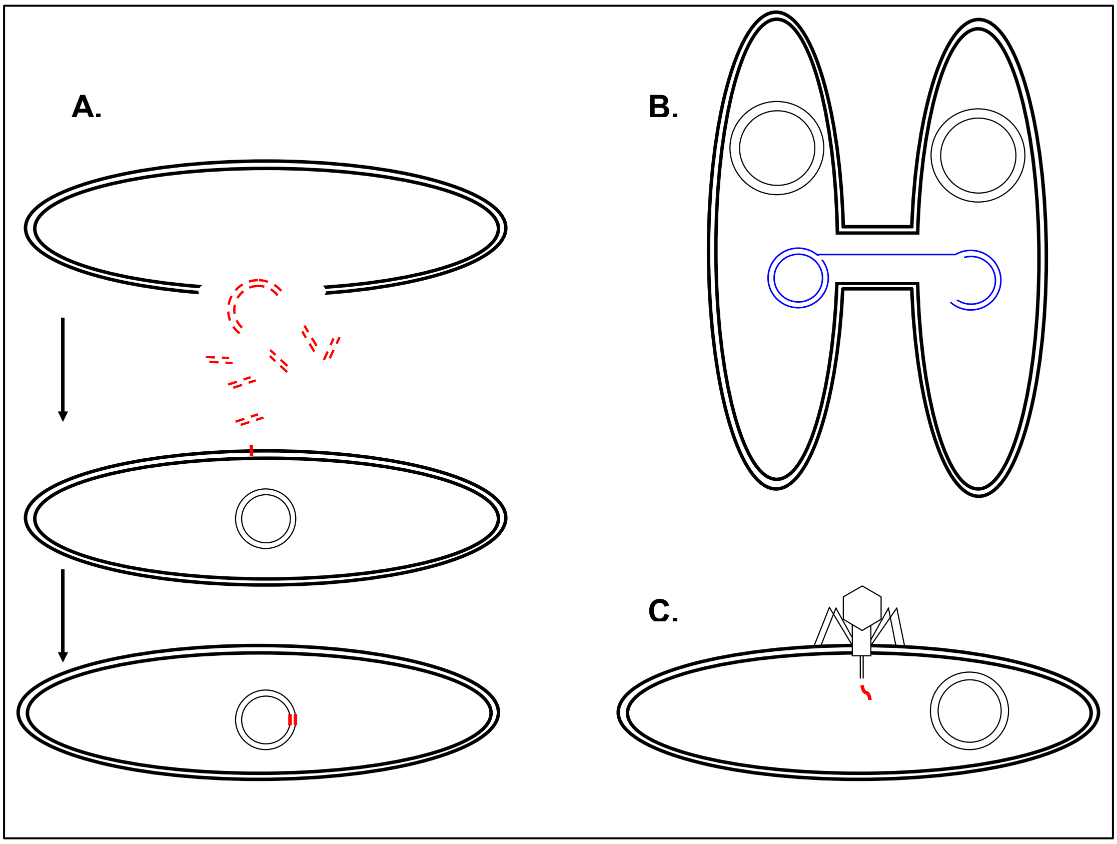


Figure 1.2: **A.** Uptake of naked DNA from the environment (transformation). **B.** Transfer of plasmid genetic material through the mating-pair pillus from a donor to a recipient bacterial cell (conjugation). **C.** Transfer of genetic material from a donor (not shown) to a recipient bacterial cell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

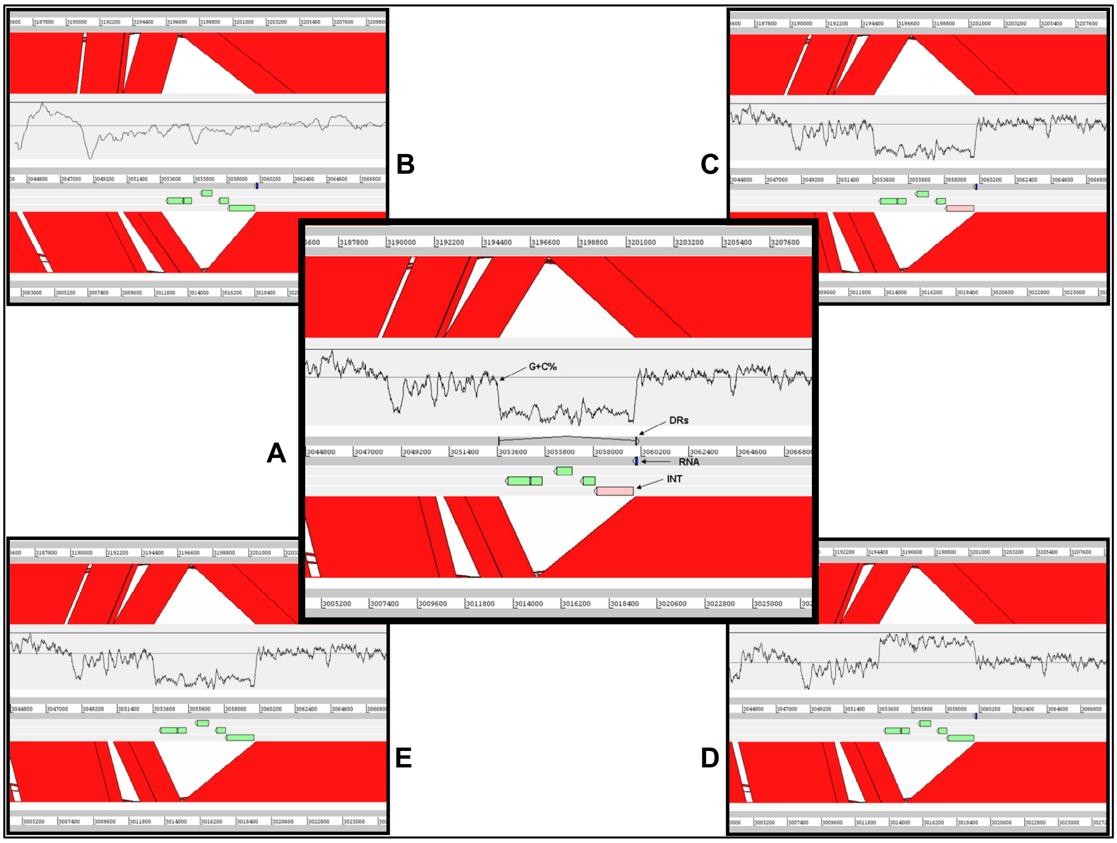


Figure 1.3: ACT (Carver *et al.*, 2005) screenshots: BLASTN comparison between three hypothetical bacterial strains (top, middle, bottom). Regions within the three strains with sequence similarity are joined by red coloured bands that represent the matching regions. **A.** Overview of the Genomic Island (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).

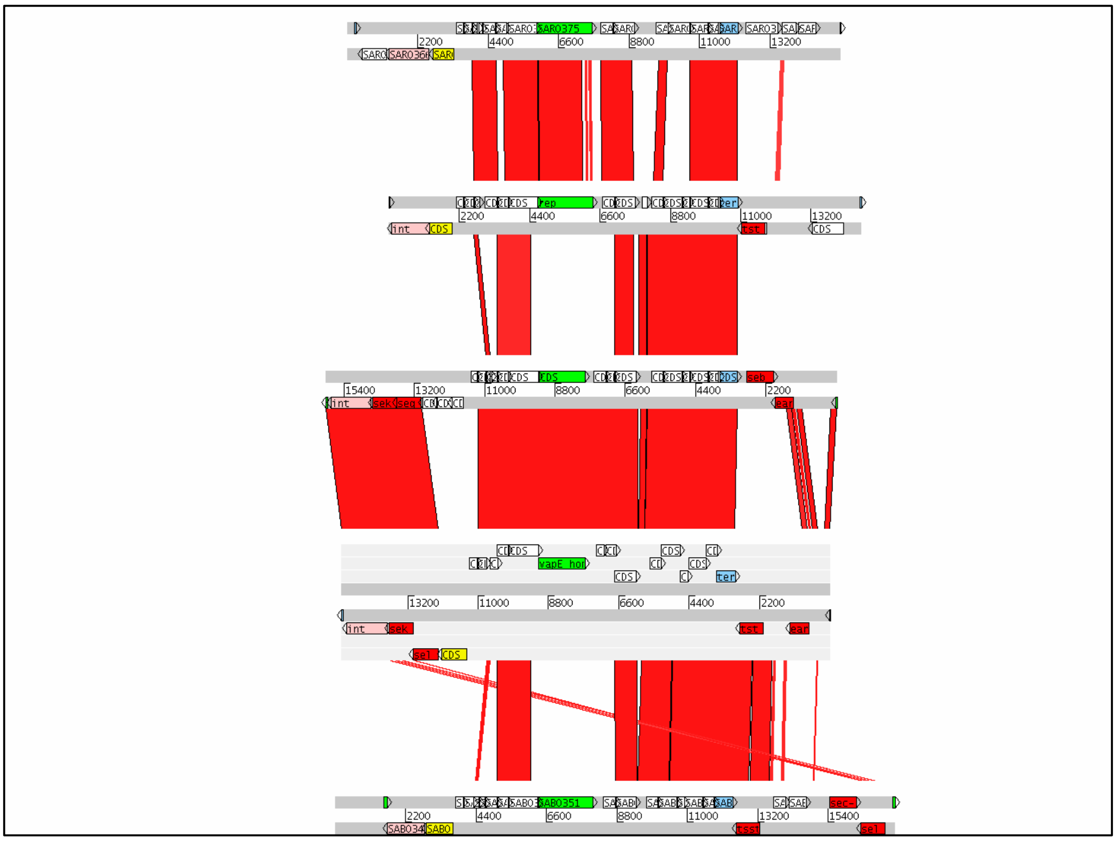


Figure 1.4: Comparison of SaPIs. From top to bottom: SaPI4 (*S. aureus* MRSA252), SaPI2 (*S. aureus* RN3984), SaPI3 (*S. aureus* COL), SaPI1 (*S. aureus* RN4282) and SaPIbov (*S. aureus* RF122). Regions within the five SaPI DNA sequences with sequence similarity are joined by red colored bands that represent 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).

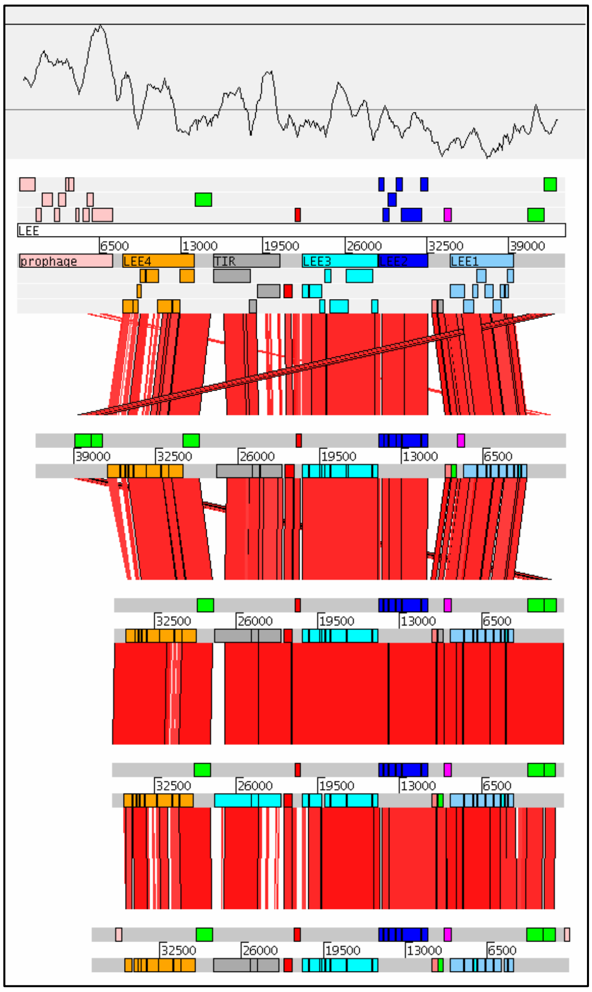


Figure 1.5: Comparison of the LEE island present in different genomes. From top to bottom: *E. coli* O157:H7 EDL933, *C. rodentium* DBS100, *E. coli* E2348/69 EPEC, *E. coli* 0181-6/86 EPEC and *E. coli* RDEC-1 EPEC. Operon colour scheme: Light blue; LEE1, dark blue; LEE2, cyan; LEE3, grey; TIR (LEE5), 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%).

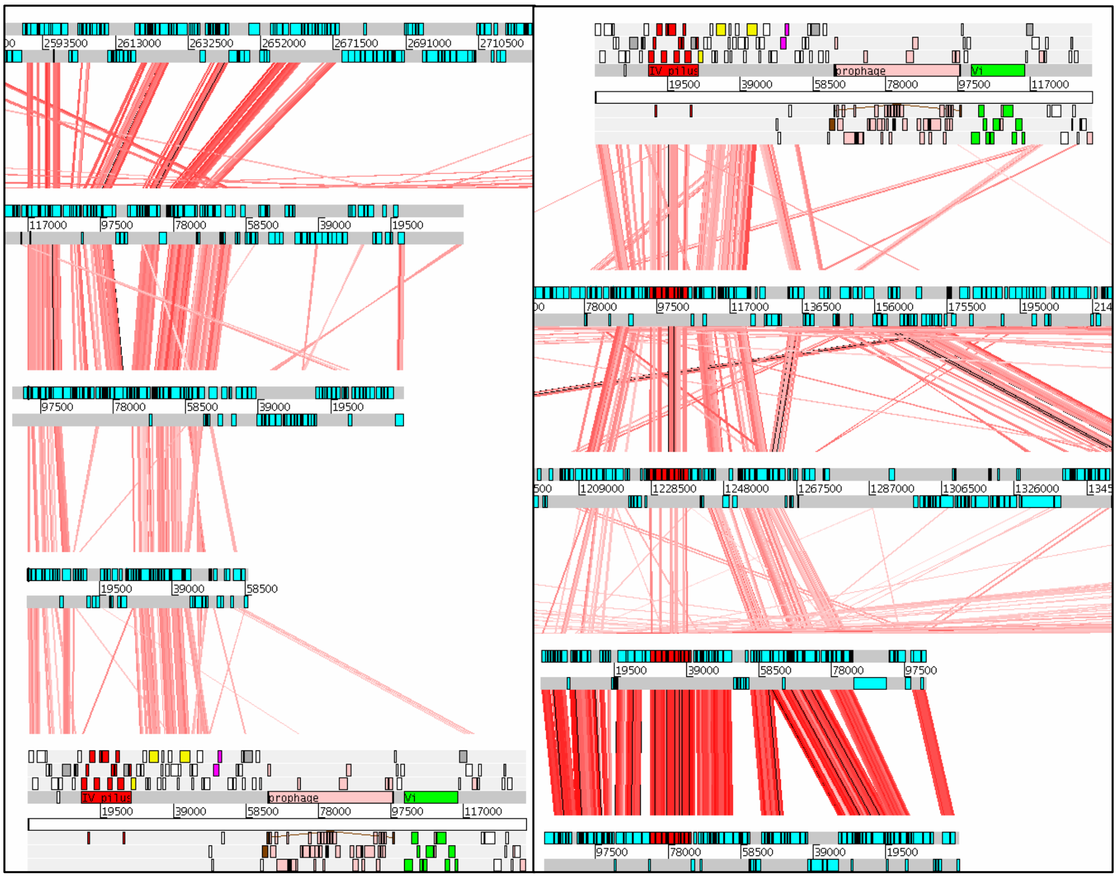


Figure 1.6: Comparison of SPI-7 with other GIs; from top to bottom and left to right: GI present in *Xanthomonas axonopodis* pv. Cistri 306, PAGI-3 island of *P. aeruginosa* SG17M, clc element in *Pseudomonas* sp. strain B13, ICEHin1056 in *H. influenza*, SPI-7 in *S. typhi* CT18 (bottom left and top 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).

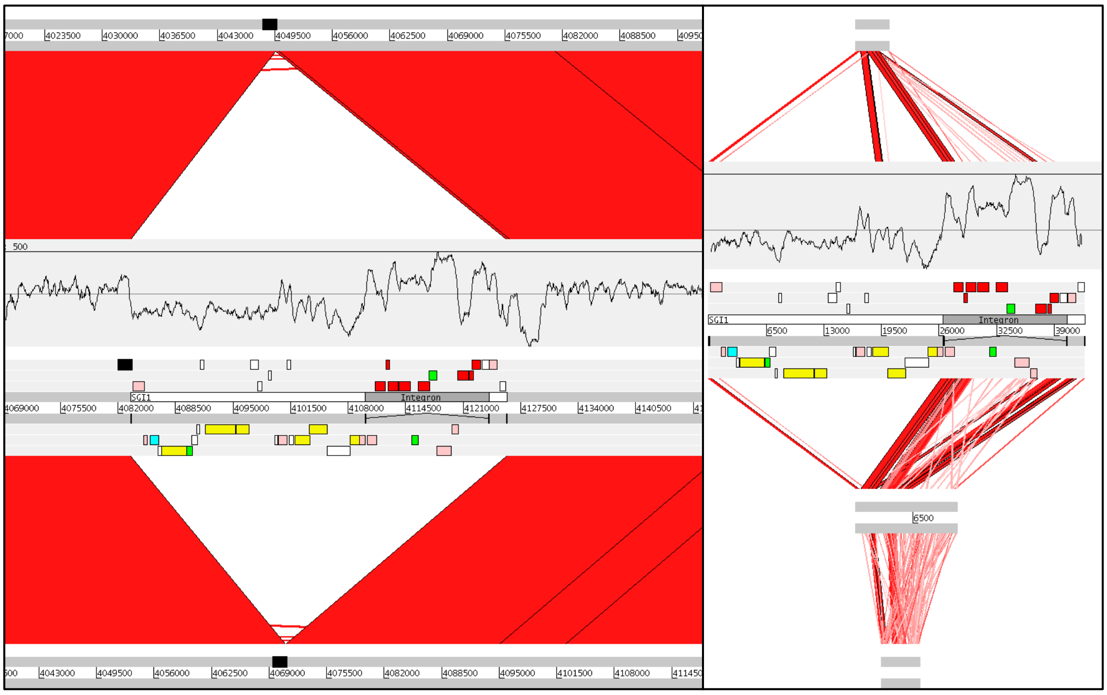


Figure 1.7: SGI-1 structure and phylogenetic distribution. **Left:** ACT comparison showing the presence of SGI-1 in Typhimurium DT104 (middle) and its absence in Typhimurium LT2 (top) and SL1344 (bottom). The G+C% content is embedded as a plot above SGI-1 with a window size of 0.5kb. Colour scheme: 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).

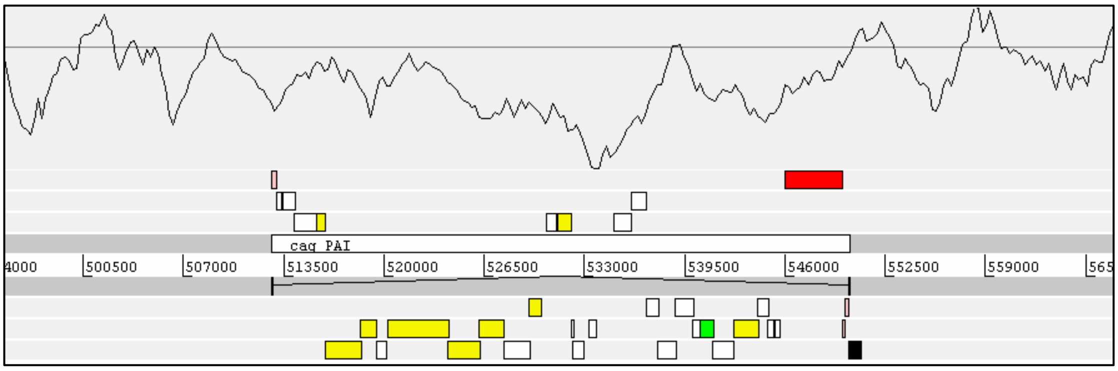


Figure 1.8: cag PAI structure. Colour scheme: Yellow; T4SS components, green; putative chaperone, red; cytotoxin-associated protein A – *cagA* gene, light pink; transposase, and black; *glr* gene – insertion point of the cag PAI. The 41bp DRs are shown as two joined black features flanking the cag PAI. 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).

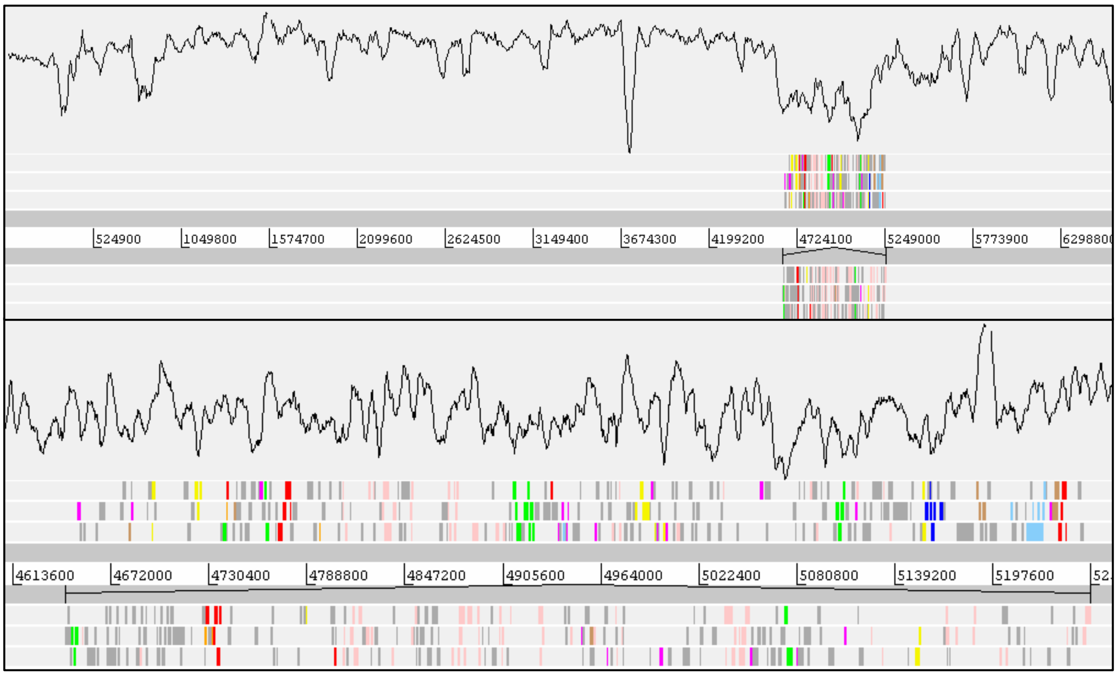


Figure 1.9: Structure of the symbiosis island (SI) in *M. loti*. Colour scheme: Light pink; integration/recombination, red; nitrogen fixation, yellow; nodulation, green; ABC transporters, light blue; conjugation, dark blue; T3SS translocation components, brown; cytochrome, orange; ferredoxin, and pink; 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).

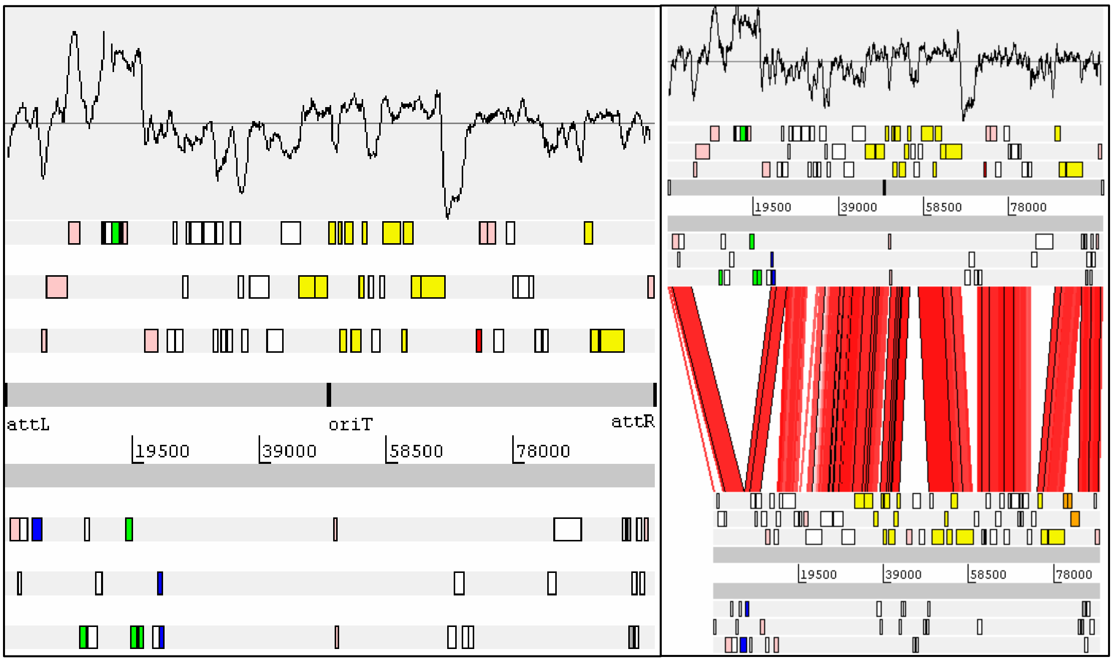


Figure 1.10: **Left.** SXT structure. Colour scheme: Pink; integrase, transposase and phage related, yellow; conjugation, green; antibiotic resistance, dark blue; UV repair, red; single-stranded DNA binding protein, grey; transcriptional activator, and orange; mercury resistance. **Right.** ACT comparison 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).

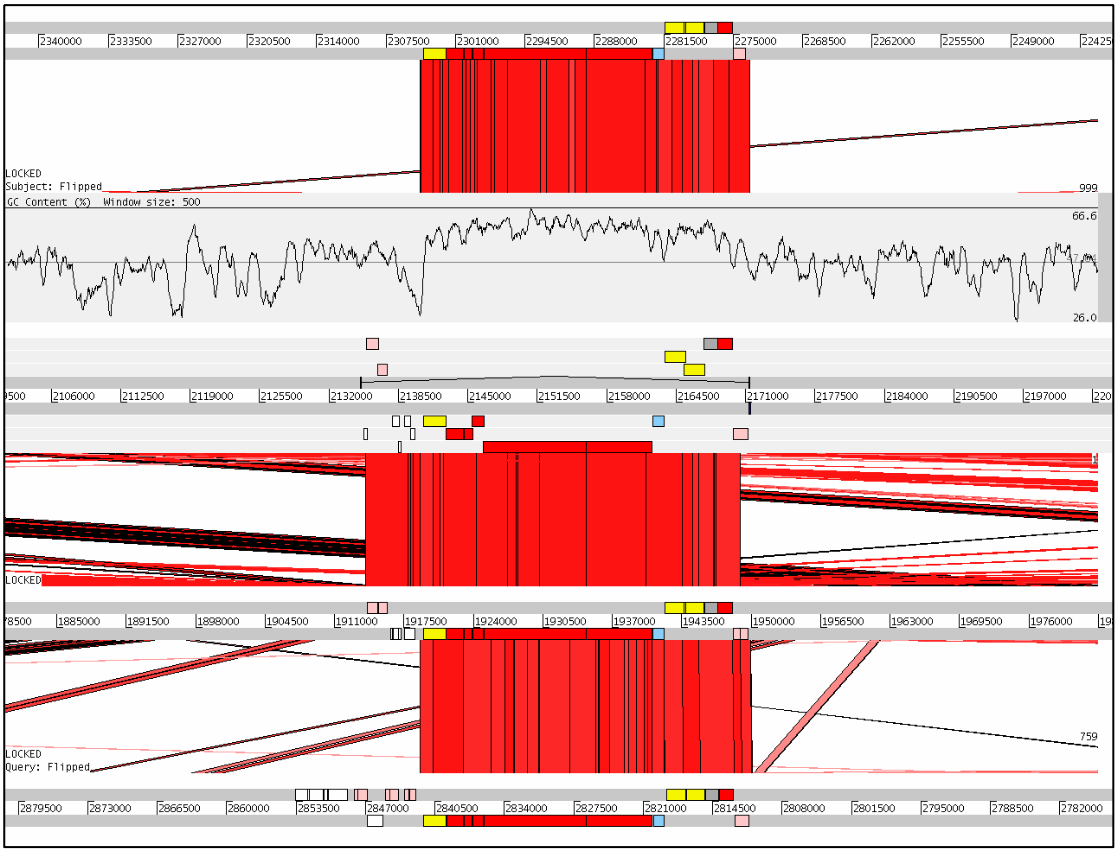


Figure 1.11 Comparison of the HPI present in different genomes. From top to bottom: Enteroaggregative *E. coli* 042, *Y. pestis* CO92, *Y. pseudotuberculosis* IP32953 and *Y. enterocolitica* 8081. Colour scheme: Light pink; integrase, transposase, red; yersiniabactin biosynthesis, yellow; yersiniabactin 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).

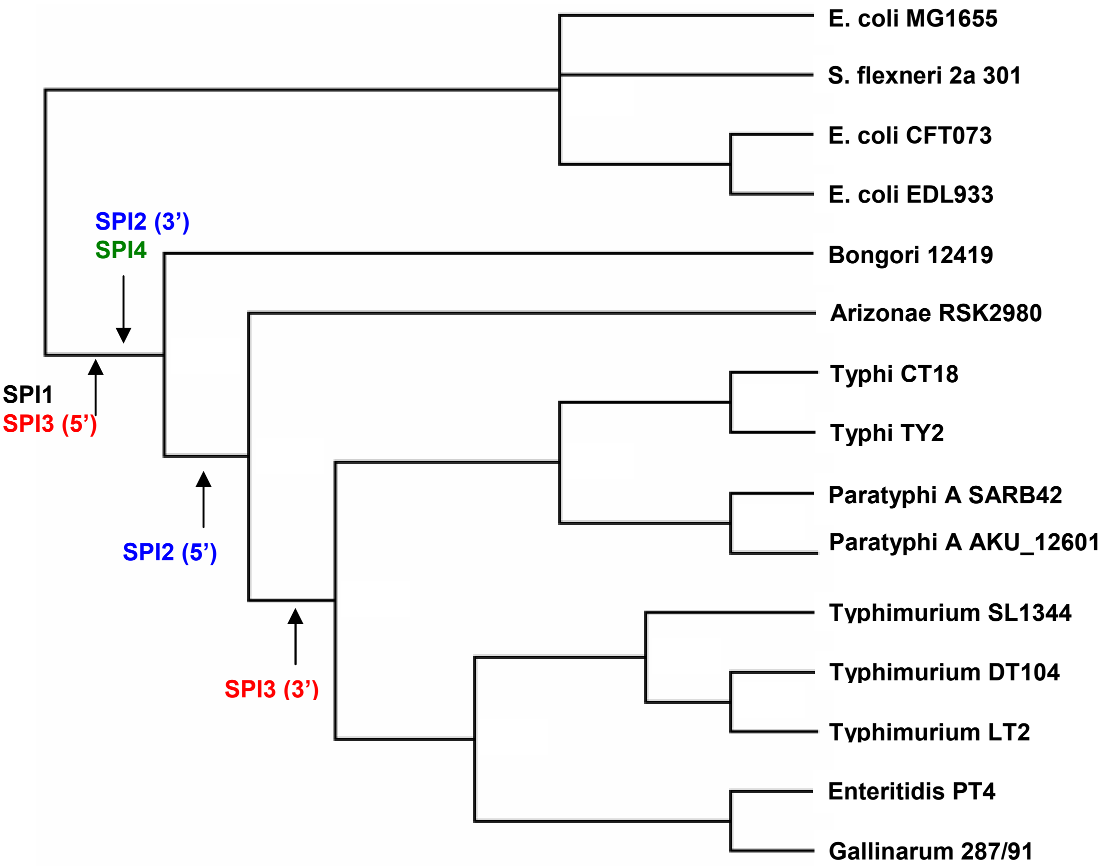


Figure 1.12: Phylogenetic distribution of four *Salmonella* Pathogenicity Islands: SPI-1, SPI-2, SPI-3 and SPI-4 in the *Salmonella* lineage. The cladogram shows the phylogenetic relationship between 11 *Salmonella* strains and four outgroups (*E. coli* MG1655, *S. flexneri* 2a 301, *E. coli* CFT073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

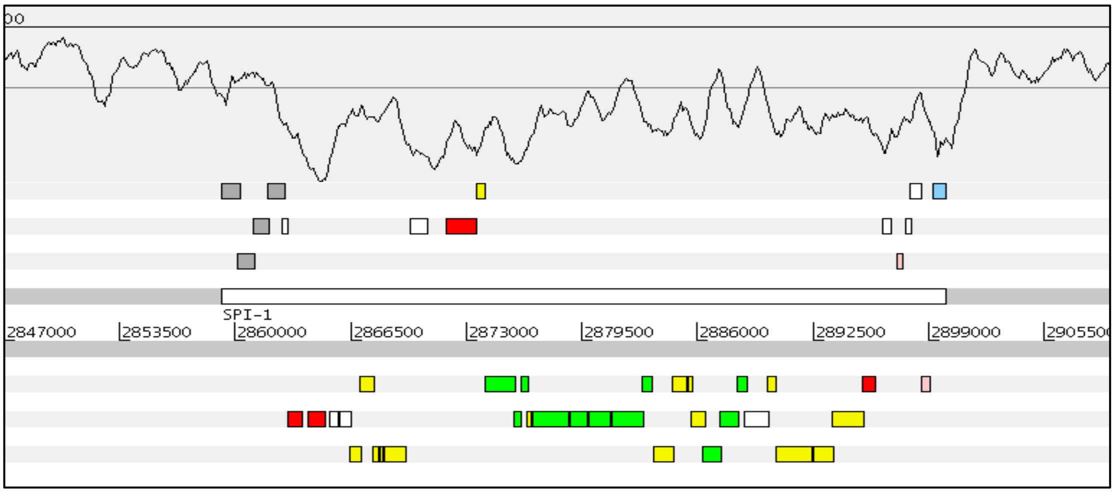


Figure 1.13: Gene content of SPI-1. Colour scheme: Grey; iron transport, red; regulation, yellow; Type III Secretion System, green; secreted proteins-effectors and chaperones, light pink; transposase, light blue; serine/threonine phosphatase, and white; hypothetical. At the top of the figure the G+C%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

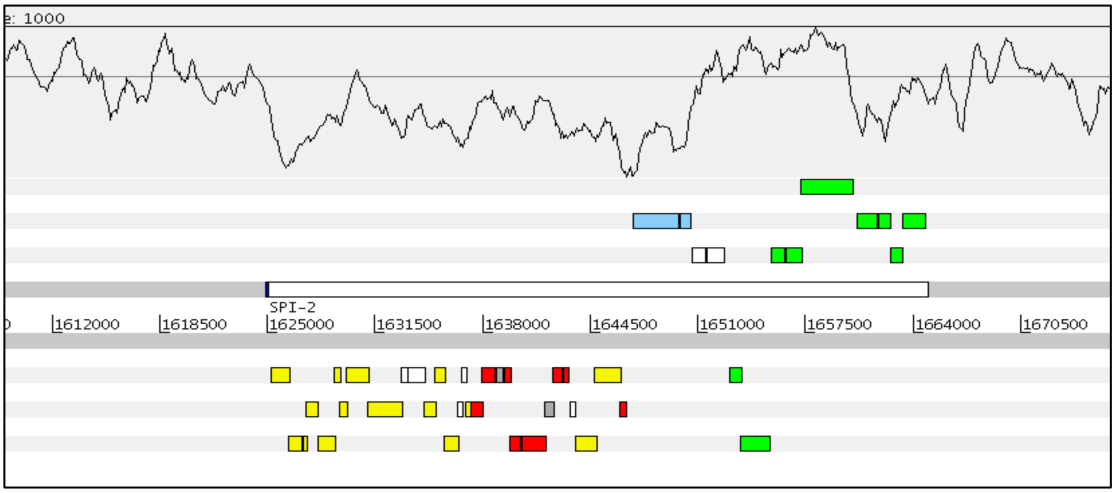


Figure 1.14: Gene content of SPI-2. Colour scheme: Grey; chaperones, red; secreted proteins-effectors, yellow; Type III Secretion System, green; tetrathionate reductase operon, dark blue; tRNA, light blue; two-component response regulator, and white; hypothetical. At the top of the figure the G+C% content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

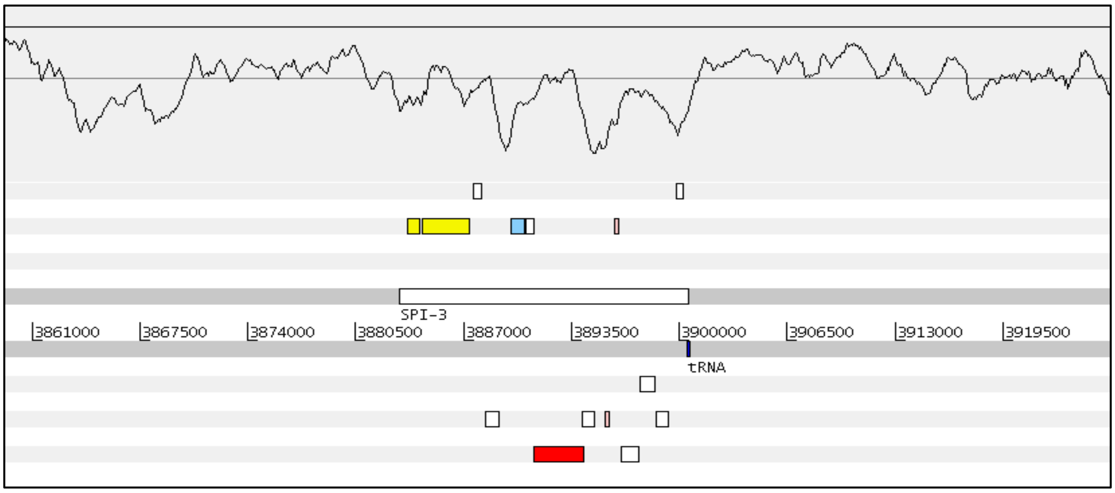


Figure 1.15: Gene content of SPI-3. Colour scheme: Red; autotransporter/T5SS, yellow; magnesium transport, dark blue; tRNA, light blue; transcriptional regulator, white; hypothetical, and light pink; transposase. At the top of the figure the G+C% content is shown with a window size of 1kb.

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).

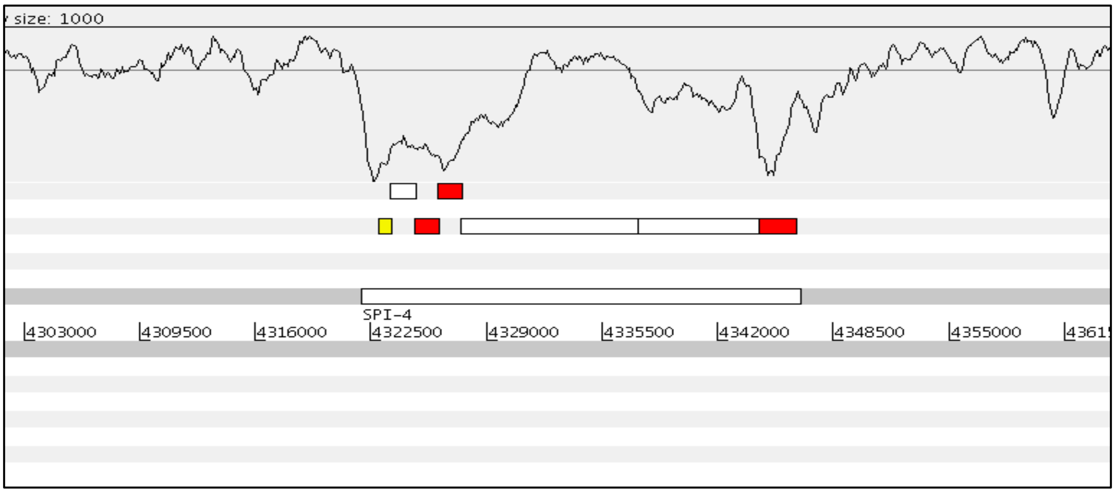


Figure 1.16: Gene content of SPI-4. Colour scheme: Red; Type I Secretion System, yellow; putative exported protein, and white; hypothetical. At the top of the figure the G+C% content is shown with a window size of 1kb.

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.

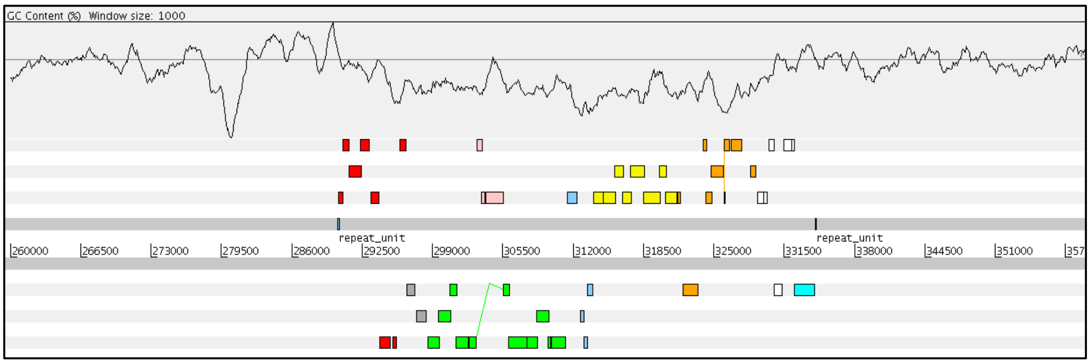


Figure 1.17: Gene content of cci. Colour scheme: Red; arsenic and antibiotic resistance, yellow; amino acid metabolism, light blue; transcriptional regulator, white; stress-response, light pink; transposase, green; fatty acid biosynthesis, grey; autoinducer synthesis, cyan; putative sulphate transporter, 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).

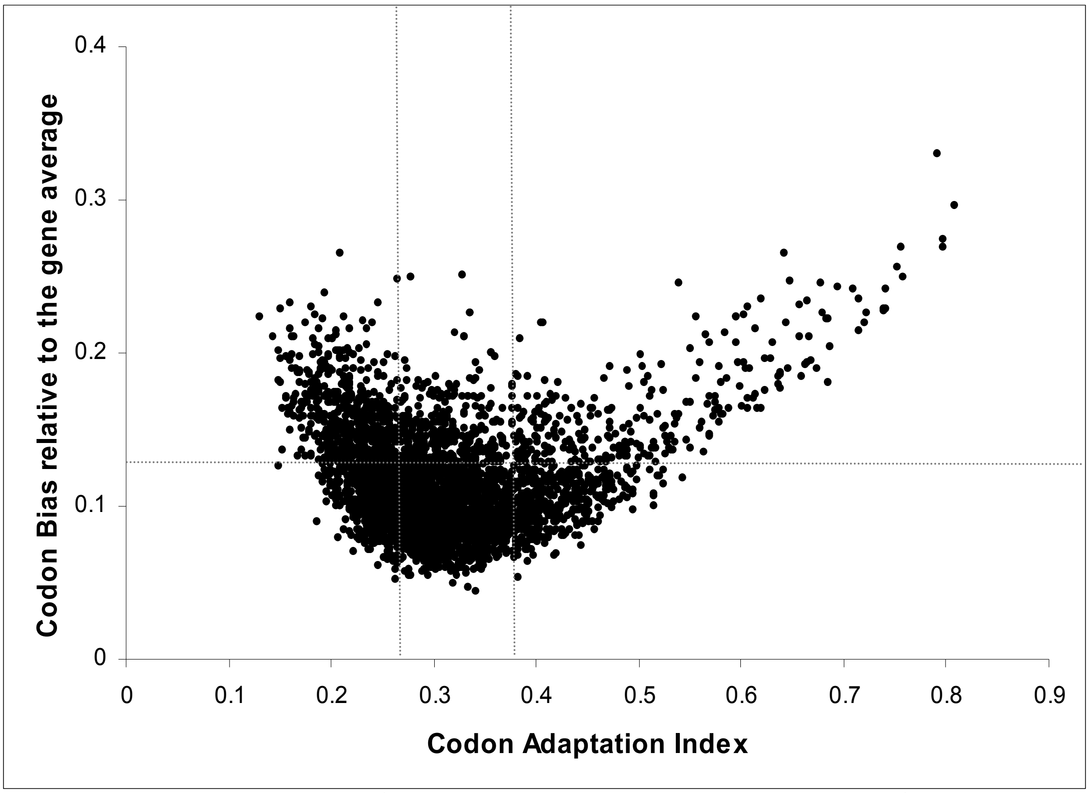


Figure 1.18: The codon bias (relative to the gene average codon usage) of the genes (>300bp) present in *S. typhi* CT18 is plotted against their codon adaptation index (CAI) value; CAI values have been calculated using the reference set of highly expressed genes, proposed by Sharp and Li (Sharp and Li,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

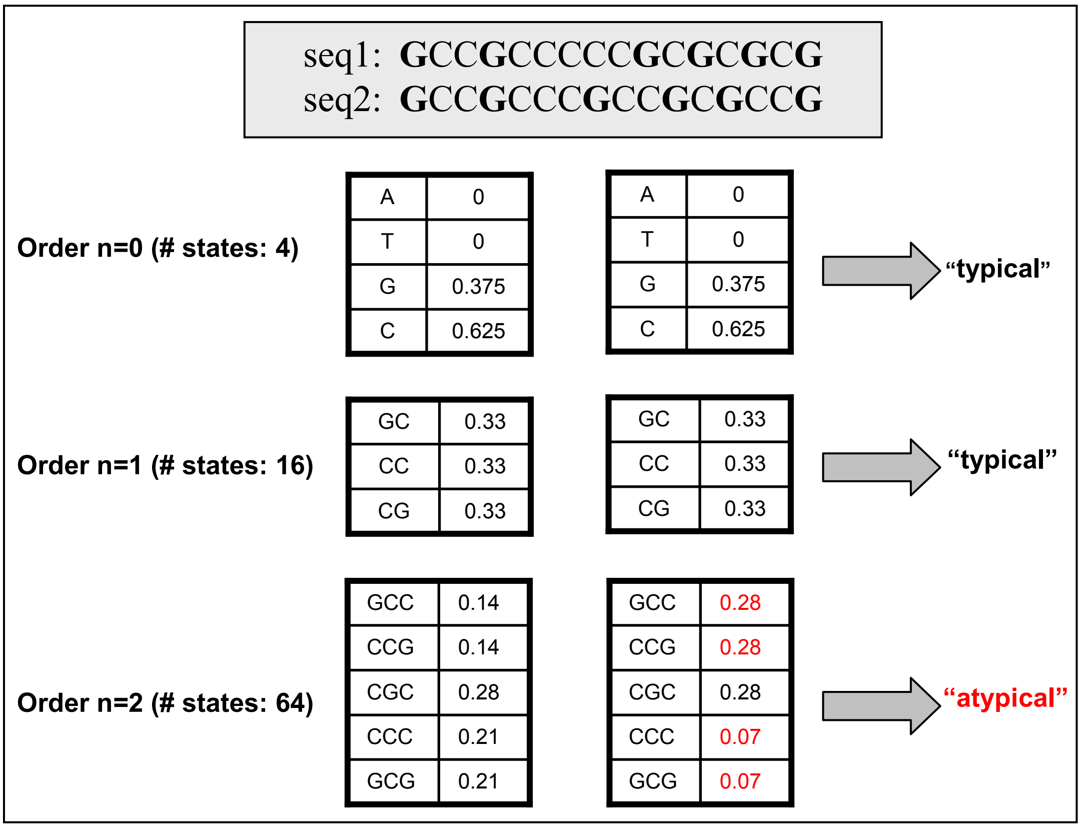
Chapter 2

Alien_Hunter algorithm

2.1 Introduction

There is a growing literature on the detection of Horizontal Gene Transfer (HGT) events by means of compositional-based, non-comparative methods (Garcia-Vallve *et al.*, 2003; Hsiao *et al.*, 2003; Karlin, 2001; Lio and Vannucci, 2000; Merkl, 2004; Sandberg *et al.*, 2001; Tsirigos and Rigoutsos, 2005). Such approaches rely only on sequence information and utilize different low (e.g. G+C% content) or high order (e.g. 8mers) indices to capture deviation from the genome backbone composition. The superiority of high order over lower order indices, in detecting local compositional bias, has been shown previously (Sandberg *et al.*, 2001; Tsirigos and Rigoutsos, 2005).

More specifically Tsirigos *et al.* (Tsirigos and Rigoutsos, 2005) simulated HGT events by inserting (*in silico*) genes from a gene pool into a query genome and analyzed the sensitivity of increasing order indices in predicting the simulated, manually inserted genes. Overall, using low order indices (e.g. single-nucleotides or di-nucleotides), 40-55% of the manually inserted genes were correctly predicted as HGT events, whereas higher order indices e.g. 8mers showed overall a much higher sensitivity, predicting correctly 50-65% of those genes, depending on the number of simulated HGT events. Another example showing the increased sensitivity of high order indices is shown in Figure 2.1; two compositionally very similar sequences (seq1 and seq2) can only be predicted as compositionally distinct, if indices of order two (i.e. tri-nucleotides) or higher are exploited, while zero or first order compositional analysis predicts those two sequences to be compositionally identical. However, given the increased dimensionality of the compositional alphabet, in a fixed-order based implementation of compositional distributions even high order indices may actually be poor estimators of the local sequence composition; this is likely to be the case when insufficient information is available, e.g. in short sequence samples or sliding windows, or when local, low-order compositional biases exist (Figure 2.2). Consequently methods exploiting multiple, different order indices can be more powerful in detecting compositional biases at various levels (Karlin, 2001).



seq1: **G**CC**G**CCCCC**G**C**G**C**G**C**G**
seq2: **G**CC**G**CCC**G**CC**G**C**G**CC**G**

A	0
T	0
G	0.375
C	0.625

A	0
T	0
G	0.375
C	0.625

GC	0.33
CC	0.33
CG	0.33

GC	0.33
CC	0.33
CG	0.33

GCC	0.14
CCG	0.14
CGC	0.28
CCC	0.21
GCG	0.21

GCC	0.28
CCG	0.28
CGC	0.28
CCC	0.07
GCG	0.07

Figure 2.1: An example of two different sequences (seq1 and seq2) and the discrimination efficiency of increasing order indices. Only second (or higher) order indices can discriminate the two sequences as compositionally distinct.

In this chapter I describe a novel algorithm for the prediction of putative horizontally transferred regions by means of variable order compositional distributions with the aim of overcoming the limitations of fixed-order compositional approaches and exploiting the advantages of both low order (small-alphabet) and high order (increased sensitivity) compositional indices. This approach does not require pre-existing annotation (e.g. gene prediction), and can therefore be applied directly to newly sequenced genomes and used as a supplementary tool in the annotation pipelines. Moreover I discuss the application of two different methods for determining a genome-specific score threshold, as well as the implementation of region specific two-state, second-order Hidden Markov Models (HMMs) to optimize the localization of the boundaries of the predicted regions. Finally I describe the pipeline followed to obtain a test dataset of manually curated putative horizontally transferred regions, the performance benchmarking against other, existing methods and the biological significance of the *in silico* predictions.

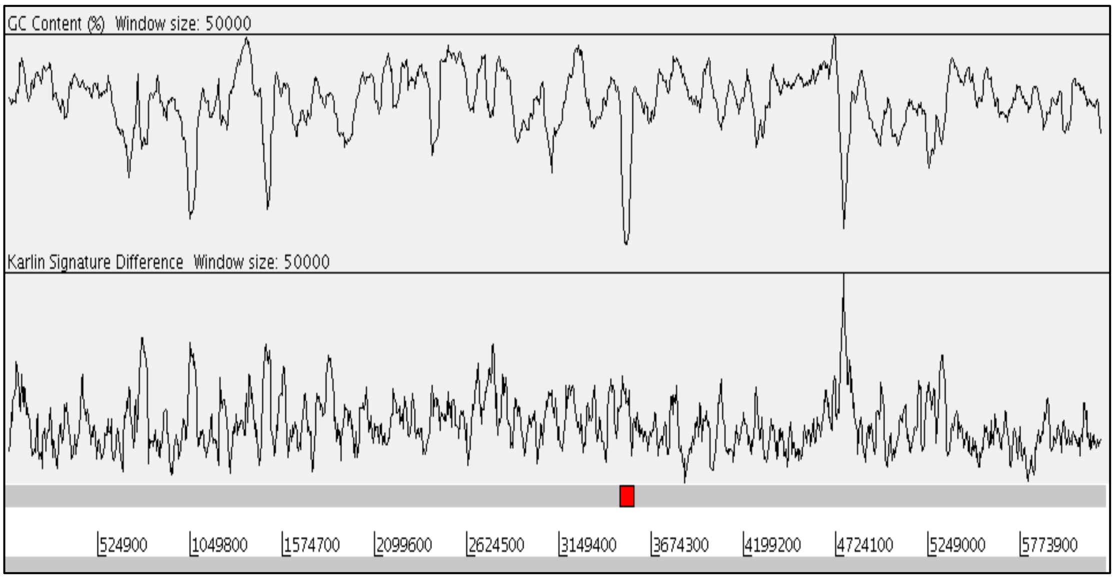


Figure 2.2: *Pseudomonas aeruginosa* PAO1 genome. The G+C% content and the di-nucleotide signature *δ** (Karlin, 2001) have been plotted genome-wide, with a window size of 50kb. A region carrying CDSs encoding products involved in the lipopolysaccharide (LPS) biosynthesis is shown as a red coloured feature. The LPS region deviates from the backbone composition mainly due to low order compositional bias (enriched in Adenine and Thymine); a compositional bias not captured by higher order indices e.g. di-nucleotides.

2.2 Methods

2.2.1 Interpolated Variable Order Motifs

Usage of low order compositional indices may not provide sufficient discrimination of regions with atypical high order (e.g. 6mers) composition. The total number of all different possible motifs (or indices) increases exponentially with the size *k* of the motifs. For *k*-mers of size *k* (e.g. k=6) there are 4^k (e.g. 4096) different possible k-mers (parameters). Consequently, because a much higher number of parameters are exploited, utilizing high order motifs is more likely to capture deviation from the genome background compositional distribution, as long as there is enough data to produce reliable probability estimates. However for high order motifs in short or biased sequences, a significant amount of data is likely to be missing. For example, using 8mers in a sliding window of 5kb, approximately 60,000 out of 65,536 (4^8) different possible 8mers will have an observed frequency of zero. Even for 8mers of non-zero frequency the information may not be enough to provide reliable estimates of the local sequence composition of a region, e.g. most 8mers will be present only once in a 5kb window.

An Interpolated Variable Order Motif (IVOM) approach (Vernikos and Parkhill, 2006) overcomes this problem, implementing variable order k-mers, "preferring" information derived from high order motifs, but when this information is insufficient, relying more on lower order motifs. Let B be the DNA alphabet, defined as: B = {a, t, g, c}. In an IVOM approach all k-mers with $1 \leq k \leq 8$ are exploited. Each k-mer can be seen as a linear combination of its component lower order motifs including itself. In a first step, for each k-mer m_k in the sequence S, its observed frequency $P_{m_k}(S)$ is calculated as follows:

$$P_{m_k}(S) = \frac{A_{m_k}(S)}{N - k + 1} \tag{2.1}$$

where $A_{m_k}(S)$ is the number of occurrences of m_k in the sequence S and N is the size of S. Generally a high order motif occurs less frequently (small number of occurrences) in a sequence compared to motifs of lower order, given that the total number of all different possible motifs is higher (larger alphabet) in the first case. In order to use in combination the different order k-mers, both the difference in the number of occurrences and in the total number of different possible k-mers have to be taken into account.

For each m_8 a weight is calculated for all ($1 \leq k \leq 8$) its interpolated k-mers, including itself, as follows:

$$W_{m_k^{m_8}}(S) = \frac{A_{m_k^{m_8}}(S) \cdot |B|^k}{\sum_{i=1}^{8} A_{m_i^{m_8}}(S) \cdot |B|^i} \tag{2.2}$$

where $m_k^{m_8}$ denotes the interpolated k-mer m_k starting at position 8-k+1 and ending at position 8 in m_8; $|B|^k$ denotes the total number of all different possible motifs of size k. In this framework a high and a low order motif have equal chances of producing bias given that both number of counts and dimensionality have been taken into account. For example, for a given 8mer if the number of occurrences of the corresponding interpolated 3mer ($|B|^3 = 64$) and 5mer ($|B|^5 = 1,024$) is 128 and 8 respectively, an IVOM approach treats the two k-mers as equally reliable estimates (64 x 128 = 1,024 x 8) of the local sequence composition of a region. Having computed the weights for each k-mer, in a second step the IVOM frequency for each 8mer m_8, as well as all its interpolated k-mers $m_k^{m_8}$, in the sequence S is calculated as follows:

$$\mathrm{IVOM}(S, m_k^{m_8}) = \begin{cases} W_{m_k^{m_8}}(S) \cdot P_{m_k^{m_8}}(S) + [1 - W_{m_k^{m_8}}(S)] \cdot \mathrm{IVOM}(S, m_{k-1}^{m_8}) & \text{if } k \geq 2 \\ W_{m_k^{m_8}}(S) \cdot P_{m_k^{m_8}}(S) & \text{if } k = 1 \end{cases} \tag{2.3}$$

The IVOM frequency of each interpolated k-mer is calculated step-wise, starting with the shortest interpolated k-mer (i.e. 1mer) and progressively moving towards longer k-mers all the way up to the 8mer itself. Using the above equation, it is possible for the observed frequencies of all the interpolated motifs to be combined linearly in such a way that if high order motifs are reliable (sufficient counts) estimates of the local sequence composition, then the corresponding weight will be high enough for the contribution of the lower motifs to be ignored and *vice versa*.

A similar equation is implemented by Salzberg (Salzberg *et al.*, 1998) in GLIMMER, a widely used gene prediction method. In GLIMMER however the above equation is used in a Markov model-based context i.e.

Interpolated Markov Models (IMMs). Moreover GLIMMER uses two different criteria in order to calculate the weight for each *k*-mer. The first is number of occurrences; if that number exceeds a pre-determined threshold value, then the weight is set to 1.0 (the default threshold value is 400). The second is a predictive value determined by a X^2 test comparing the observed base frequencies with the IMM probabilities derived from the immediately shorter context. In the IVOM algorithm however through equation 2.2 the weight for each *k*-mer is determined on-the-fly directly from the underlying local compositional landscape avoiding the incorporation of arbitrary threshold values (Table 2.1).

2.2.2 Relative entropy

In order to predict putatively horizontally transferred regions in microbial genomes, it is assumed that each genome exhibits a reasonably constant (although exceptions may apply – e.g. the rRNA operon) background sequence composition that is the result of the same mutational pressure applied throughout its sequence. Consequently regions of "atypical" composition within a genome are likely to have been horizontally acquired from a donor genome of different composition.

In order to detect compositionally deviating regions, a sliding window approach over raw genomic sequence is applied. In this framework the analysis of atypical regions can be applied both on annotated and newly sequenced genomes without any level of annotation (e.g. pre-existing gene prediction).

Obviously in a sliding window based approach, different window sizes and moving steps can be exploited. In order to converge over the optimal sliding window size *L*, I experimented on different *L* values, implementing a Receiver Operating Characteristic (ROC) curve analysis and the results (Appendix A) showed that the greatest Area Under the Curve (AUC), which is a measure of the accuracy of the classifier, for *k*-mers of $k \leq 8$ is achieved when the sliding window size and step is set to 5kb and 2.5kb respectively.

Table 2.1: Example of two different 8mers, present in the sequence of *Salmonella* Pathogenicity Island (SPI)-7 inserted in the chromosome of Typhi CT18. The interpolated, variable order motifs of each 8mer are shown along with their observed frequency *A*, calculated based on the compositional analysis of SPI-7. The weight *W* of each interpolated *k*-mer has been also calculated. In the first 8mer (GCCAGCGC), the interpolated *k*-mer with the highest compositional information is the 8mer itself whereas for the second 8mer (AAAACATG) the most informative interpolated *k*-mer is the di-nucleotide 'TG'.

8mer	Interpolated *k*-mer	*A*	$\|B\|^k$	*A* x $\|B\|^k$	*W*
GCCAGCGC	GCCAGCGC	24	4^8	1572864	**42.10**
	CCAGCGC	49	4^7	802816	21.51
	CAGCGC	116	4^6	475136	12.73
	AGCGC	238	4^5	243712	6.53
	GCGC	738	4^4	188928	5.06
	CGC	2452	4^3	156928	4.20
	GC	9784	4^2	156544	4.19
	C	33854	4^1	135416	3.63
AAAACATG	AAAACATG	1	4^8	65536	7.38
	AAACATG	6	4^7	98304	11.08
	AACATG	26	4^6	106496	12.00
	ACATG	81	4^5	82944	9.35
	CATG	474	4^4	121344	13.67
	ATG	2110	4^3	135040	15.21
	TG	9243	4^2	147888	**16.66**
	G	32499	4^1	129996	14.65

It should be noted that increasing the order of the utilized *k*-mers causes the optimal window size to increase too (Wu *et al.*, 2005). The same authors concluded that for symmetric Kullback-Leibler discrepancy as a similarity measure and $2550 \leq L \leq 4950$ the optimal word size *k* is 8, confirming the rationale behind the selection of a 5kb sliding window used in the current analysis. The step of the sliding window is set to 2.5kb; however, increasing the step size too much will increase the uncertainty about the real boundaries of the predicted "atypical" regions. This technical issue and how it can be handled efficiently will be discussed in the next section. Both for the sliding window *w* and the genome *G* a compositional vector, defined as:

$$\overrightarrow{\mathbf{IVOM}(\boldsymbol{S},\boldsymbol{m_8})} = \{\mathrm{IVOM}(S,m_8) \mid m_8 \in B^8\} \tag{2.4}$$

is built. This vector extends over all ($|B|^8$) the different possible 8mers m_8 in the sequence S. In order to compare the two vectors (of w and G) a distance similarity measure has to be applied. In the current methodology, the relative entropy (Kullback-Leibler – KL distance), defined as:

$$d_G(w) = \sum_{m_8 \in B^8} \mathrm{IVOM}(w,m_8)\log_2 \frac{\mathrm{IVOM}(w,m_8)}{\mathrm{IVOM}(G,m_8)} \tag{2.5}$$

is implemented. The KL distance is a reasonable similarity measure in this case, since the task at hand is to compare two probability distributions; moreover KL is always non-negative and equals zero only if the two distributions are identical. Implementing equation 2.5, a sequence region of “atypical” composition will have high relative entropy while native-typical regions will have relative entropy close to zero (compositional distribution closer to the genome); it should be noted that the compositional vector of the genome IVOM(G,m_8), extends over all 8mers present in the genome sequence, including those of the current sliding window w.

2.2.3 Score threshold

Given that the current implementation of the Alien_Hunter algorithm is unsupervised, the very specific compositional landscape of different query, previously unseen genomic sequences will determine the exact value of the score threshold; above this threshold value, regions that deviate from the backbone composition of the query chromosome will be reported as putative horizontally acquired candidates. Consequently a pre-determined value of a score threshold, based on a supervised training on a test dataset is not applicable in the case of chromosomal compositional analysis, given that some chromosomes may consist of almost zero (Tamas *et al.*, 2002) up to 24% of alien DNA (Nelson *et al.*, 1999); moreover some bacterial chromosomes, that contain several HGT events, show a fairly constant backbone composition (e.g. *Salmonella*) while other genomes (e.g. *Staphylococcus*) display a highly mosaic composition.

Generally, for a typical microbial chromosome, the compositional distribution is a long-tail one of the form shown in Figure 2.3. The majority of the regions present in a bacterial genome will have a compositional distribution very close to the genome backbone composition (low IVOM score – blue coloured in Figure 2.3), a few will deviate (red colour) and very few will deviate strongly (green colour). A reasonable value for the score threshold is a value close to the point in the distribution where the transition from the "typical" (backbone – blue coloured) to the "atypical" (compositional deviating – red coloured) compositional score population occurs.

There are different approaches for capturing dynamically the optimal score threshold for any given, previously unseen microbial chromosome. In the current implementation of Alien_Hunter, I exploit two different methods. The first relies on a derivative-based approach, similar to the one exploited by Tsirigos and Rigoutsos (Tsirigos and Rigoutsos, 2005); in this approach, the transition, in the compositional score distribution f, from the "typical" to the "atypical "scores can be captured by calculating the derivative f' of the distribution.

Starting from the highest scoring regions moving (sliding window based) towards low-scoring ones, the point in the distribution, where the value of f' (calculated through the current sliding window) starts to remain steady (after several iterations) represents a good score threshold value that discriminates compositional deviating from non-deviating regions within a chromosome; the score threshold can be dynamically determined on-the-fly for each query genome. However, this approach can be quite sensitive to data noise depending on the actual shape of the compositional distribution. Moreover a derivative-based threshold often over-predicts (i.e. very low threshold); for example in the distribution shown in Figure 2.3 the point in the score distribution where the derivative starts to remain steady results in a score threshold of 7.6 well below the value of 13.2 where a much stronger transition from the "typical" to "atypical" scores occurs (see next paragraph).

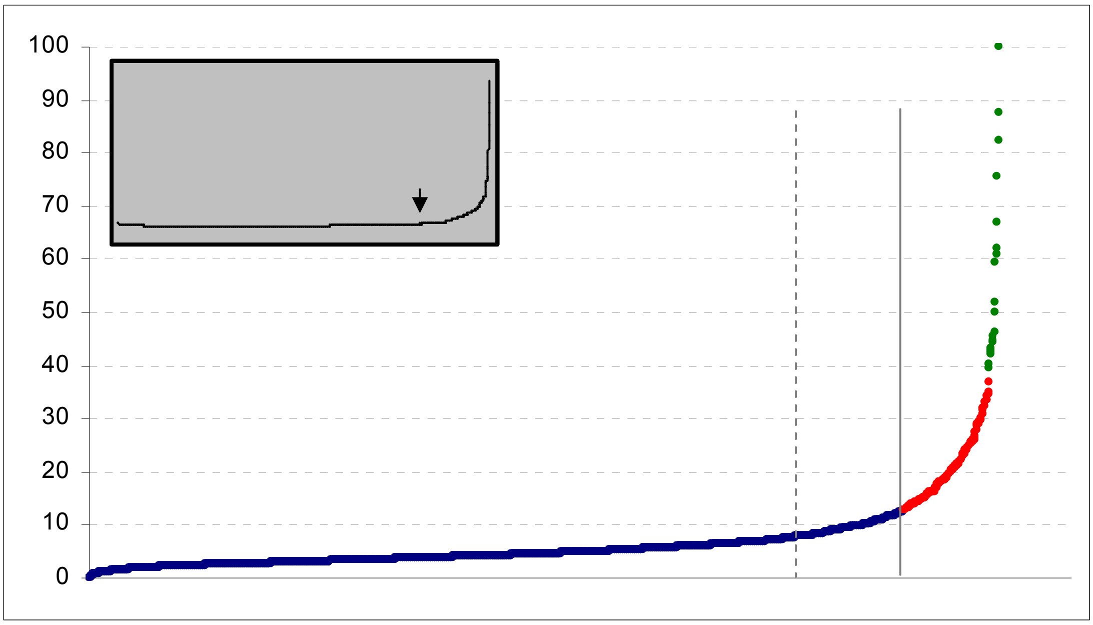


Figure 2.3: An example of the compositional score distribution of *E. coli* MG1655 chromosome. The IVOM score of all the sliding windows is plotted, sorted by increasing order. A three-colour scheme has been used to highlight the three distinct compositional populations, blue (backbone), red (intermediate compositional deviation), green (high compositional deviation). The dashed and solid, vertical grey line represents the score threshold determined by a derivative based (T=7.6) and a K-means clustering method (T=13.2), respectively. The derivative of the score distribution is plotted in the inset, with an arrow highlighting the value of the derivative used to determine the score threshold.

The second approach for determining dynamically the score threshold is based on the K-means clustering algorithm (MacQueen, 1967). K-means clustering is a non-hierarchical, supervised method, given that the initial number K of clusters is fixed and determined prior to the learning process. In the current implementation, in order to model properly the three distinct compositional populations (backbone, atypical and very atypical) a K-means clustering with three different clusters is exploited. The pseudocode describing the K-means clustering implementation is shown below.

```
Algorithm: K-means clustering.
C: number of re-initializations.
F: objective function.
i = 1.
1. Determine the number of clusters, K = 3.
2. Initialize the value of the 3 centroids.
3. Assign each point to the cluster with the nearest centroid value.
4. When all points have been assigned to one of the 3 clusters, update the new centroid
   values.
5. Re-iterate steps 3 and 4 until the 3 centroids do not change;
   convergence criteria: Last_$F_i$ –Current_$F_i$ < 0.1.
6. If i < C do
      if $F_i > F_{i\_max}$ then $F_{i\_max} = F_i$
      i++
      goto step 2 re-initializing the 3 centroids with different values.
7. Set the score threshold to the value where the transition from cluster 1 → 2 occurs,
   for the iteration with $F_{i\_max}$.
end
```

Although the K-means clustering algorithm always converges, the final clustering strongly depends on the values used to initialize the centroids of the three clusters. For those reasons different starting points are used to initialize the centroids and the iteration with the maximum (i.e. the one that separates the three clusters the most) objective function value is used to determine the score threshold. Through steps 1-5, the algorithm is trying to minimize the following objective function:

$$F = \sum_{j=1}^{K} \sum_{i=1}^{n} \left\| x_i - c_j \right\|^2 \qquad (2.6)$$

where, K is the number of clusters (i.e. 3), n the total number of sliding windows, x_i the IVOM score of each sliding window and c_j the centroid value. As the clustering algorithm proceeds, the value of the F function decreases and converges over a minimum value; at this stage the clustering terminates.

Table 2.2: Performance benchmarking of the Alien_Hunter algorithm implementing a derivative and a K-means clustering based algorithm to determine the score threshold (6.1 and 11.3 respectively) for the genome of Typhi CT18.

Performance metric	Derivative	K-means
Specificity	0.653	0.746
Sensitivity	0.649	0.511
Accuracy	0.764	0.775
Matthews correlation coefficient	0.473	0.473

Overall, the K-means clustering determines more accurately the optimal score threshold (Figure 2.3) and is less affected by the noise in the compositional data and the shape of the distribution. The accuracy of the Alien_Hunter algorithm, implementing the derivative-based and the K-means clustering algorithm, was benchmarked using a manually curated (see below) dataset of 1560 putative horizontally acquired genes in Typhi CT18 (Table 2.2); the data confirm the higher accuracy of the second method compared to the first one, although a derivative-based threshold results in an overall more sensitive method, capable of detecting older HGT events with composition very close to the genome backbone.

2.2.4 Change-point detection

As mentioned in section 2.2.2 the choice of the step for the sliding window approach is crucial, given that the window slides over raw genomic sequence (consequently the gene boundaries are unknown), decreasing the window step will increase the computation required, and increasing the window step will reduce the accuracy of the localization of the predicted "atypical" regions. For these reasons, upon the completion of the first round of the window-based prediction, a second-order, two-state HMM is implemented in a change-point detection framework. HMM is a statistical model widely used in speech and music recognition (Rabiner, 1989; Raphael, 1999) as well as in several bioinformatics tasks, e.g. gene prediction (Burge and Karlin, 1997). HMMs can be thought as finite state machines in which each state emits symbols governed by an emission probability distribution over a given alphabet of allowed symbols; at each stage, the model can either stay in the same state, or make the transition to a new one, a process governed by a distribution of transition probabilities; both the emission and transition probability distributions are state-specific.

HMMs can be described by two sequences (Durbin *et al.*, 1998). The hidden state sequence $\pi = (\pi_1,\ldots,\pi_L)$ also known as the "path" and the observed sequence $x = (x_1,\ldots,x_L)$ which corresponds to the observed symbols; in our case the bases of a DNA sequence. In an n-th order HMM each base x_i depends on the previous $(x_{i-n},\ldots, x_{i-1})$ bases as well as on the ith state π_i in the path. In the current study, two states are exploited: the "native" (N) state that corresponds to regions of typical (i.e. close to the genome backbone) composition and the "alien" (A) state that models compositionally deviating, "atypical" regions. Under this framework, a change-point corresponds to switching from one state to the other; in the current implementation the aim is to infer the boundaries of the predicted regions, where a state transition occurs. This change-point will represent the new optimized boundary of each prediction, offering higher predictive accuracy in terms of boundary localization (see results section). In order to detect the point where the transition from the native to the alien state occurs and *vice versa*, the following approach is pursued.

Each predicted "atypical" region is extended further upstream in order to incorporate sequence of typical composition. This hybrid sequence of one typical and one atypical subsequence, is used to train the HMM on-the-fly (the same approach is also applied on the downstream boundary – Figure 2.4). Implementing the Baum-Welch (BW) algorithm (Baum, 1972), the parameters (transition and emission probabilities) of the model are trained, in an iterative fashion until some convergence criteria are met. The BW algorithm is an Expectation Maximization (EM) technique that estimates transition and emission probabilities calculating the expected number of times each transition and emission is used, given the training sequence; this is done iteratively, until convergence, by considering probable paths within the sequence exploiting each time the current/updated parameters of the model. However different starting parameter values strongly affect the local maxima which the BW will converge over. One straightforward solution to this problem is to start multiple times from different initial model parameters, an approach that is implemented in this analysis.

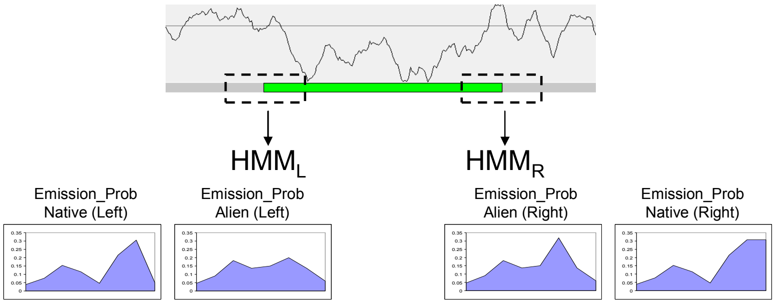

HMM$_L$
HMM$_R$
Emission_Prob Native (Left)
Emission_Prob Alien (Left)
Emission_Prob Alien (Right)
Emission_Prob Native (Right)
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0

Figure 2.4: Two side-specific HMMs trained for the left (HMM$_L$) and for the right (HMM$_R$) boundary of each predicted GI.

Given that we do not know beforehand for how long the system remains in the native state before it makes the transition to the alien state (and *vice versa* for the downstream boundary) the algorithm starts with multiple starting points (*prior* expectations) over the transition probability:

$$\alpha_{NA} = P(\pi_i = A \mid \pi_{i-1} = N) \tag{2.7}$$

where a_{NA} denotes the transition probability from the native (N) to the alien (A) state; for each starting point, the model is trained using the BW algorithm until convergence.

In a change-point detection framework with a single change-point, once the a_{NA} transition occurs, the model persists at the alien state until the end. For this reason only the a_{NA} transition probability is trained, while the transition probability from the alien to the native state is set to be zero (a_{AN} = 0, un-trainable). For the emission probabilities, given that the composition of the native and the alien DNA sequence is not known *a* *priori*, two trainable, uniform, second-order compositional distributions are exploited (Figure 2.5).

In a second step for each starting point, upon BW training, the Viterbi algorithm (Viterbi, 1967) is implemented with the updated-trained parameters. The Viterbi algorithm is a dynamic programming algorithm widely used in inferring the most probable state path π^* (in our case the most probable sequence of native/alien hidden states) given the observations (DNA bases) and the model parameters (emission and transition probabilities):

Algorithm: Viterbi.

1. Initialisation ($i = 0$): $v_0(0) = 1$, $v_N(0) = 0$ for $N > 0$.

2. Recursion ($i = 1 \ldots L$): $v_A(i) = e_A(x_i)\max_N(v_N(i-1)a_{NA})$; (2.8)

$ptr_i(A) = \text{argmax}_N(v_N(i-1)a_{NA})$.

3. Termination: $P(x,\pi^*) = \max_N(v_N(L)a_{N0})$;

$\pi_L^* = \text{argmax}_N(v_N(L)a_{N0})$.

4. Traceback ($i = L \ldots 1$): $\pi_{i-1}^* = ptr_i(\pi_i^*)$.

Source: (Durbin *et al.*, 1998).

Briefly a score $v_A(i)$ for each DNA base x_i in the state A (with the previous base x_{i-1} being in state N) is calculated (equation 2.8). The first part of this equation consists of the emission probability $e_A(x_i)$ of x_i in state A; the second part consists of the maximum value (over all values of N) of the product of the maximal score at the previous (i-1) base position and the transition probability from state N to A. The optimal path can be found by backtracking over an array of pointers ($ptr_i(A)$) that keep track, at each x_i in state A, of the maximum score of the previous state thus revealing the most probable sequence of hidden states that gave "birth" to the observed sequence.

In the Alien_Hunter algorithm, keeping track of the probability of the most probable path predicted by the Viterbi algorithm, the iteration (over different starting points of a_{NA}) with the highest probable path, among all the most probable state paths, will be the one which best describes the data (the true transition point). An example is given in (Table 2.3) and the algorithm is summarized in the following pseudocode:

Algorithm: Change-point detection.

```
C: number of iterations
Init: i = 1;
$a'_{NA}$: initial starting point for $a_{NA}$
1. extend the predictions upstream and downstream
2. set initial model:
   2.1. prior distribution for the emission probabilities:
        2.1.1. N state: trainable second order uniform ($e_N$) distribution
        2.1.2. A state: trainable second order uniform ($e_A$) distribution
   2.2. prior transition probabilities:
        2.2.1. $a_{NA} = a'_{NA}$ (multiple starting points - trainable)
        2.2.2. $a_{AN} = 0$ (untrainable)
3. BW training until convergence:
   3.1. stopping criteria: LastScore - CurrentScore < 0.001
   3.2. updated-trained emission, transition probabilities
4. Viterbi: most probable path $\pi^*$, with score $S_i$
        4.1.1. if $S_i > S_{imax}$ then $S_{imax} = S_i$
5. if i < C do
        5.1.1. i++;
        5.1.2. new starting point $a'_{NA}$
        5.1.3. goto step 2
6. report the path $\pi^*$ with $S_{imax}$
7. set predicted boundary = transition point in the path $\pi^*$ with $S_{imax}$
end
```

Table 2.3: An example of multiple starting points for the transition probability a_{NA} and the corresponding score of the most probable path π^* predicted by the Viterbi algorithm for a test hybrid sequence.

iteration	score S_i of path π^*	*prior* over a_{NA}	change-point (bp)
1	-9643.868804	500^{-1}	1720
2	-9643.868873	1000^{-1}	1720
3	-9627.033373	2000^{-1}	4870
4	-9627.033077	2500^{-1}	4870
5	-9627.033131	3000^{-1}	4870

In the example described in Table 2.3 the best model (highest scored π^*) for estimating the position in the sequence where the transition from the N to the A state occurs, is the one in which the prior expectation over the α_{NA} value is 2500^{-1}. In the first two starting points, the predicted change-point occurs at 1720bp (starting from the 5' end of the test hybrid sequence) whereas in the remaining cases the change-point is predicted at 4870bp. In the current version of the Alien_Hunter software (http://www.sanger.ac.uk/Software/analysis/alien_hunter/) the BW and the Viterbi algorithms are implemented using the relevant Biojava libraries (http://www.biojava.org).

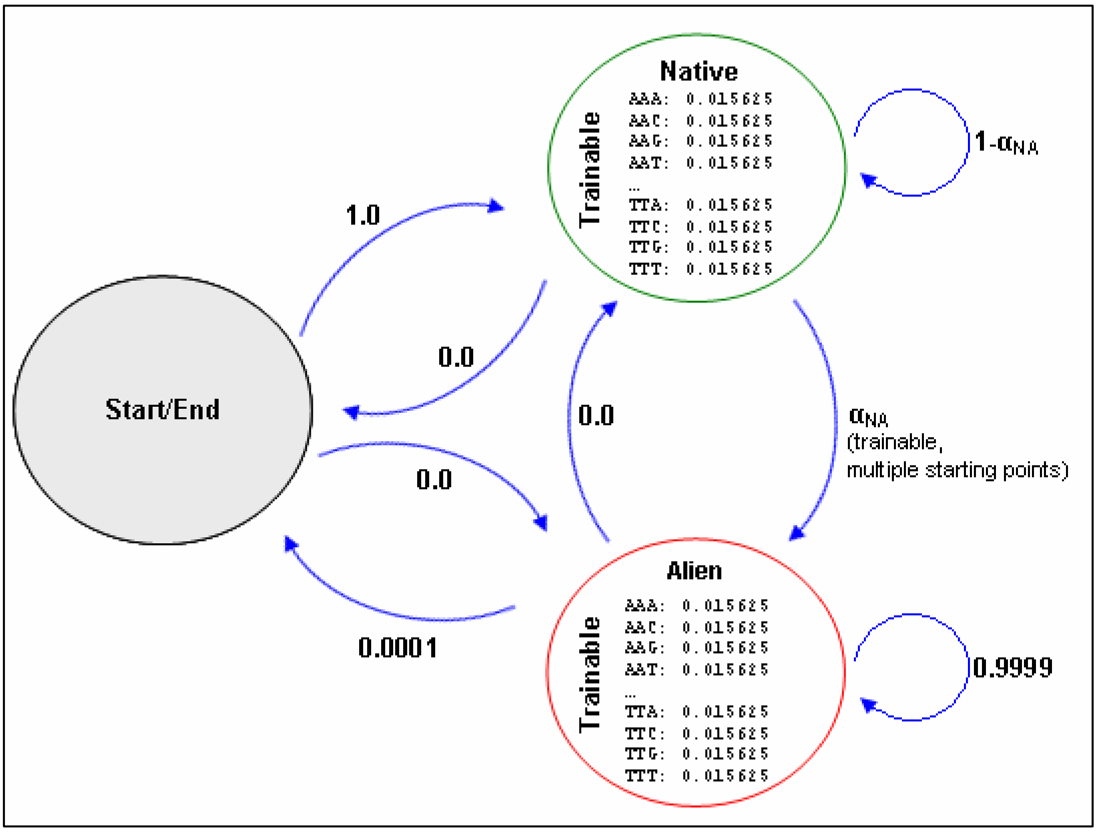


Figure 2.5: The architecture of the two-state (Native, Alien), second order HMM, used in a change-point detection framework.

2.2.5 Reciprocal FASTA

In order to evaluate the performance of Alien_Hunter, a test dataset of putative horizontally transferred genes was built. Previous approaches

(Azad and Lawrence, 2005; Tsirigos and Rigoutsos, 2005) involved simulation of HGT events by inserting genes from various donor genomes into the genome under study; such approaches simulate only very recent HGT events, thus they do not take into account the amelioration (Lawrence and Ochman, 1997) of horizontally acquired DNA, a time-dependent process. For this reason I chose to build a test dataset of putative HGT events, based on real data.

The genome of *S. typhi* CT18, a well-studied prokaryote in terms of HGT events, was used as the reference genome. *S. typhimurium* LT2 was selected as a sister lineage to Typhi while the genome of *E. coli* MG1655 was chosen as an outgroup of Typhi and Typhimurium. The main idea is that genes that are present in all the three genomes form a set of core genes, while the rest of the genes represent either species or strain specific genes thus are considered putative candidates for HGT. The choice of two sister lineages and one outgroup increases the chances of capturing older HGT events, which otherwise might be indistinguishable; for example SPI-1 and SPI-2 are species-specific, but not strain-specific islands. Moreover a comparative analysis between two sister taxa and one outgroup, enables a more reliable discrimination between gene loss and gene gain, two events that can equally explain a limited phylogenetic distribution of a gene, within a lineage. *E. coli* seems to form a good outgroup organism, given that the estimated divergence of *E. coli* and *S. enterica* from the common ancestor occurred approximately 100-140 Myr ago (Doolittle *et al.*, 1996; Ochman and Wilson, 1987). In order to extract all the putative horizontally transferred genes in Typhi, the following approach was pursued.

Each CDS (a) from the genome (A) was searched, with FASTA, against the CDSs of the other genome (B). If the top hit covered at least 80% of the length of both sequences with at least 30% identity, a reciprocal FASTA search of the top hit sequence (b) was launched against the CDSs of the first genome. If the reciprocal top hit is the same as the original query CDS then (a) and (b) are considered orthologous genes of (A) and

(B). Genes that are unique in, or are orthologs between Typhi and Typhimurium but do not have an ortholog in *E. coli* form the initial dataset of putative HGT events. In a second step, in order to validate the results, a BLASTN and TBLASTX comparison between the three genomes was carried out, to check for a syntenic relationship among the putative orthologs and the results were visualized using ACT (Carver *et al.*, 2005). It should be recognized that this procedure will also identify genes that have been uniquely deleted in *E. coli* as putative HGT events (see results section).

2.3 Results

2.3.1 Manually curated HGT dataset

Implementing the reciprocal FASTA approach described above, four different groups of genes present in Typhi were identified: The first group involves 725 genes that are unique in Typhi. The second and third group includes orthologous genes between Typhi and *E. coli* (52) and Typhi and Typhimurium (903). In the last group are 2920 core genes that are shared between all the three genomes (Figure 2.6).

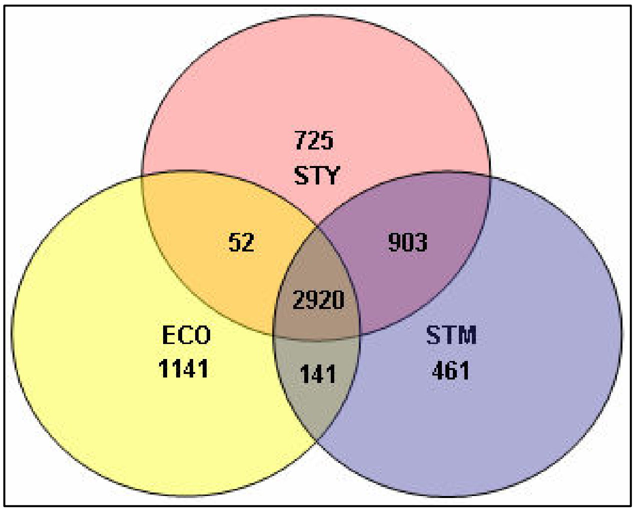


Figure 2.6: Venn diagram illustrating the unique and the orthologous genes present in the genome of *E. coli* (ECO), *S. typhi* (STY) and *S. typhimurium* (STM).

Excluding the 2920 predicted core genes and the 52 Typhi and *E. coli* unique orthologs, the remaining gene set (1628 genes) forms the initial dataset of putatively horizontally transferred genes in Typhi. In a second step, the above dataset was manually curated for gene position consistency using ACT, and the initial number was reduced to 1560 manually curated putative horizontally transferred genes which form the basis of the analysis described in the following sections.

It should be noted that this analysis yields a significantly high number of putative HGT events in the genome of Typhi CT18. The reliable estimation of true HGT events strongly depends on the evolutionary sample at hand; going well back in the evolutionary history of an organism offers more reliable detection of sequences that have been transferred horizontally from other sources. For example, some of the *Salmonella* lineage-specific genes might not necessarily represent HGT events (gene loss in *E. coli*). However this analysis provides a more reliable estimation of putative HGT events (taking into account the amelioration process), given that it is based on real data rather on simulated events. A more robust approach of discriminating putative gene gain from gene loss events will be described and discussed in chapter 3.

2.3.2 Three novel SPIs

Running the Alien_Hunter algorithm on the genome of Typhi CT18, all the previously annotated SPIs (SPI-1 to SPI-10) and bacteriophages were successfully predicted. Moreover this analysis revealed three novel putative SPIs, SPI-15, SPI-16 and SPI-17 (Table 2.4); SPI-11, 12 and SPI-13, 14 have been previously described is other *Salmonella* serovars (Chiu *et al.*, 2005; Shah *et al.*, 2005). SPI-15 represents an insertion of approximately 6.3kb, inserted in the 3' end of a tRNA^{Gly}; the insertion has duplicated a 22bp tRNA fragment, which forms the downstream boundary of SPI-15. Adjacent to the tRNA, there is an integrase gene of putative phage origin and further downstream four hypothetical protein-coding genes. Among the eight *Salmonella* genomes analyzed, SPI-15 is only present in Typhi CT18 (Figure 2.7); in Typhi TY2, there is a similar insertion of different gene content, at the same position, which is also flanked by two DRs, 22bp long.

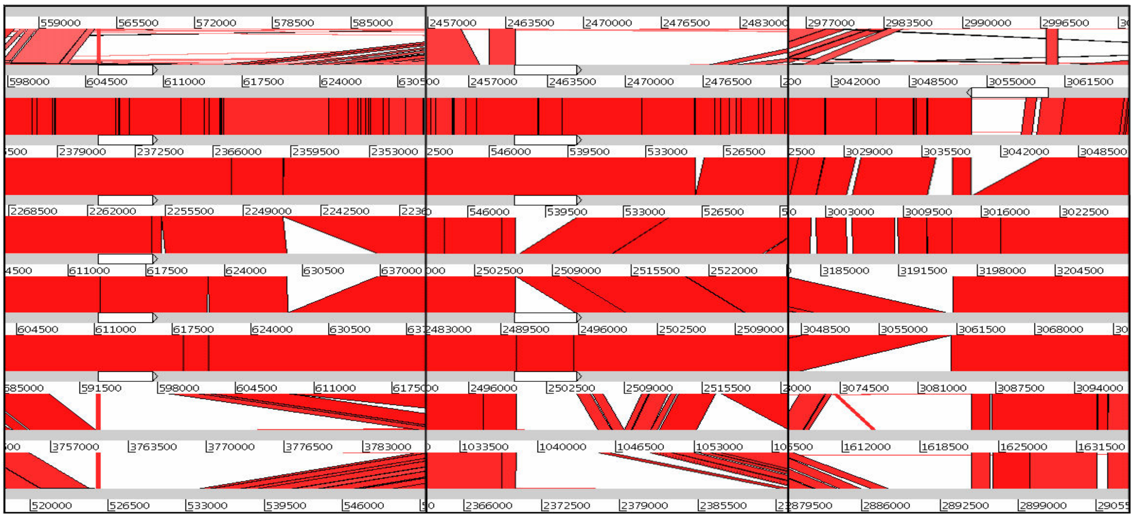


Figure 2.7: ACT screenshot: BLASTN comparison between *E. coli* and 8 *Salmonella* genomes (from top to bottom): *E. coli* MG1655, *S. typhi* CT18, *S. typhi* TY2, *S. paratyphi* A, *S. typhimurium* LT2, *S. gallinarum* 287/91, *S. enteritidis* PT4, *S. arizonae* RSK2980, *S. bongori* 12419. Regions within the nine genomes with sequence similarity are joined by red coloured bands that represent the matching regions. The three novel SPIs are illustrated as white coloured features (from left to right: SPI-16, SPI-17, SPI-15). The above screenshot is a mosaic picture of three individual screenshots at different locations along the genomes that have been concatenated for ease of visualization.

Although SPI-15 is a 6.3kb island, Alien_Hunter predicts a much larger (~18kb) region overlapping the SPI-15 locus. The predicted region starts at the exact 5' end of SPI-15, at the insertion point within the tRNA locus, but extends the 3' end 12kb further to the left (Figure 2.8), questioning the accuracy of the predicted boundaries. SPI-15 represents a very recent insertion, present only in the genome of Typhi CT18; however the comparison between Typhi CT18 and *E. coli* MG1655 suggests that the entire (~18kb) predicted region is absent from the genome of the latter. Possibly, it represents a mosaic region of more than one independent HGT events; a fairly old insertion (~12kb, G+C content = 50.1%) upstream of SPI-15 and a very recent insertion (SPI-15, G+C content = 48.9%). This observation suggests that Alien_Hunter shows increased sensitivity, predicting correctly even old HGT events or regions of mosaic compositional profile. A closer look at the 5' end of the entire 18kb region reveals that the predicted boundary starts immediately downstream of CDS STY3169 (encoding a histidine rich hypothetical protein); STY3169 is a pseudogene with an in-frame stop codon at position 110.

Table 2.4: Characteristics of the three novel predicted SPIs (SPI-15, SPI-16 and SPI-17) in the genome of Typhi CT18.

SPI	Location	Insertion site	Repeats	Integrase	Score	Size (bp)	Potential virulence determinants
SPI-15	3053654..3060017	tRNA^{Gly}	22nt (DR)	phage integrase	18.893	6364	unknown
SPI-16	605515..609992	tRNA^{Arg}	43nt (DR)	phage integrase	20.949	4478	serotype conversion by O-antigen glucosylation
SPI-17	2460793..2465914	tRNA^{Arg}	-	-	23.953	5122	serotype conversion by O-antigen glucosylation

The overall G+C content of STY3169 is 52.6% (genome average 52.09%), while the G+C contents from the 5' end up to the in-frame stop codon, and from the stop codon to the 3' end of this CDS are 49.8% and 54.3% respectively. Based on the comparison between Typhi CT18 and *E. coli* MG1655 the true 5' boundary of the 18kb locus is upstream of STY3169, suggesting perhaps that STY3169 is expected to be part of the 18kb locus. Perhaps, STY3169 as a non functional CDS, carrying an internal in-frame stop codon, has been subject to accelerated amelioration, a likely scenario, taking into account its mosaic composition (upstream and downstream of the stop codon) and the nonetheless very similar overall composition to the genome average (52.6% and 52.09% respectively). This further explains why the boundary predicted by the Alien_Hunter algorithm does not encompass the STY3169 CDS.

The second novel SPI, SPI-16 is a 4.5kb long island, inserted in a tRNA^{Arg} gene. Two DRs of 43bp form the boundaries of SPI-16 while a phage integrase (pseudogene) is located near the tRNA gene. Encoded within this island are two bactoprenol-linked glucose translocases (*gtrA* and *gtrB*) that along with the integrase pseudogene show high percentage identity (93%, 97% and 78% respectively) to homologous genes in the genome of bacteriophage P22 (Figure 2.9). *gtrA* and *gtrB* have been previously described to be involved in serotype conversion through O-antigen glycosylation mediated by bacteriophages (Guan *et al.*, 1999; Mavris *et al.*, 1997).

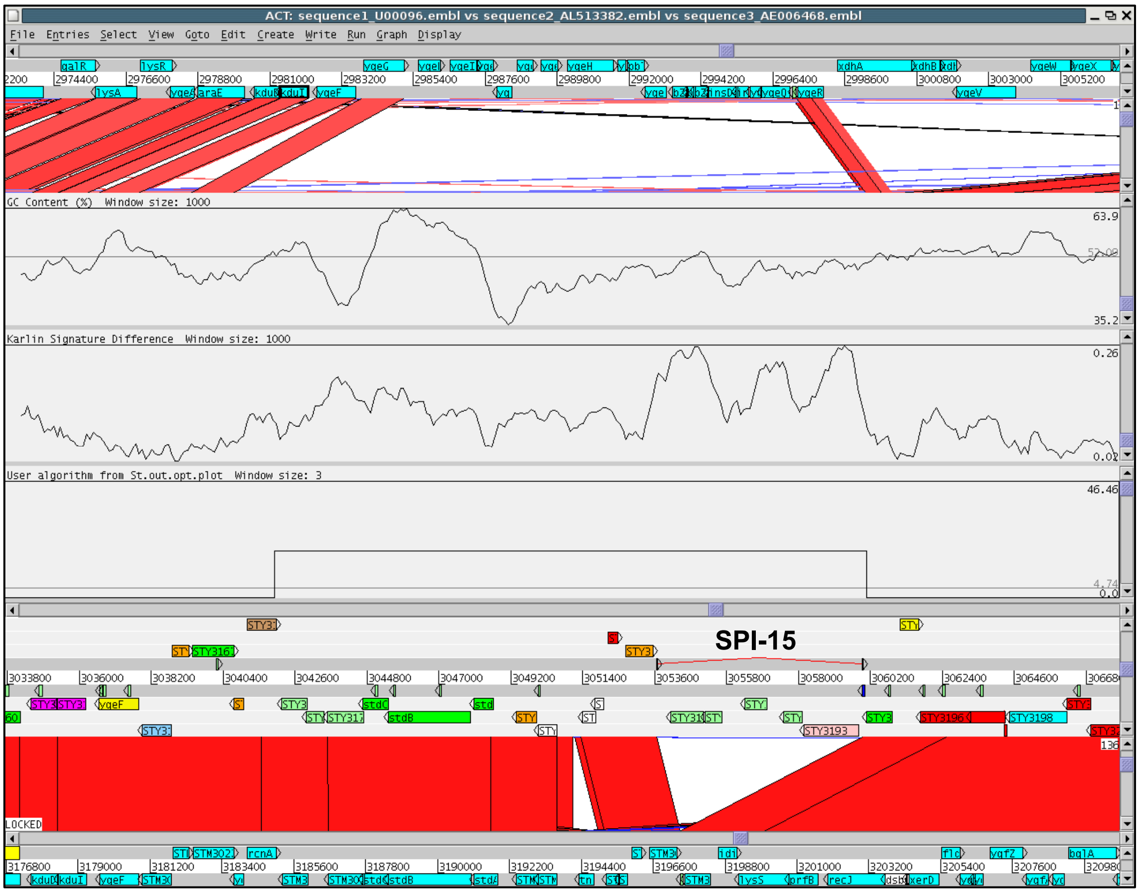
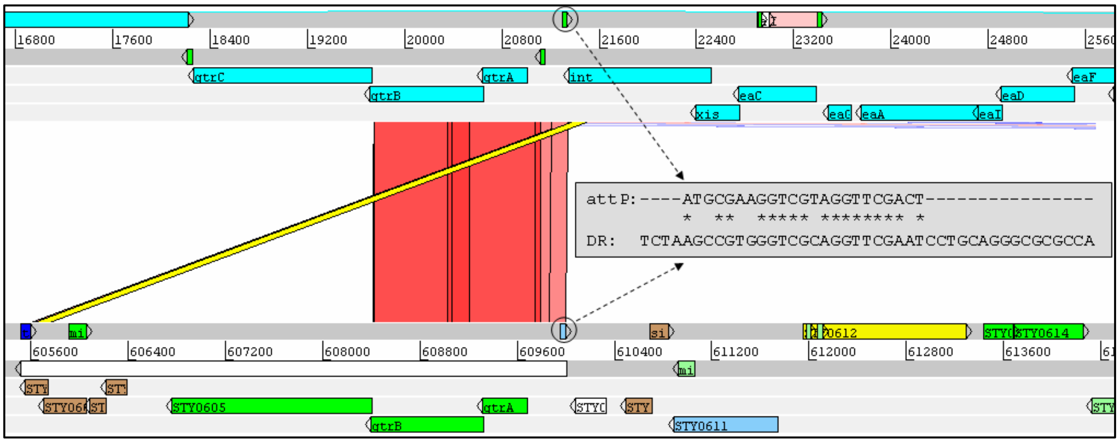

ACT: sequence1_U00096.embl vs sequence2_AL513382.embl vs sequence3_AE006468.embl
File Entries Select View Goto Edit Create Write Run Graph Display
GC Content (%) Window size: 1000
63.9
35.2
Karlin Signature Difference Window size: 1000
0.26
User algorithm from St.out.opt.plot Window size: 3
46.46
0.0
SPI-15
STY3193
STY3196
STY3198
LOCKED

Figure 2.8: ACT screenshot: BLASTN between *E. coli* MG1655, Typhi CT18 and Typhimurium LT2 (from top to bottom) at the genomic locus encompassing SPI-15 (flanked by DRs – red-coloured joined features). Plots (from top to bottom): G+C% content, di-nucleotide bias (δ^* difference) (window size = 1kb) and IVOM score. Regions within the three genomes with sequence similarity are joined by red coloured bands that represent the matching regions. The brown coloured CDS (STY3169), encoding for a histidine rich hypothetical protein, is a pseudogene with an in-frame stop codon at position 110.

This observation leaves open the possibility that SPI-16 and SPI-17 (see next paragraph) are GIs probably involved in driving the variation of the cell surface structure of Typhi and perhaps the way this bacterium interacts with its host (i.e. humans) or “parasitic” mobile elements, e.g. bacteriophages.

Also present in SPI-16 is STY0605 that encodes a putative membrane protein with nine predicted transmembrane segments (TMs). Although there is no sequence similarity to the *gtrC* gene in P22 bacteriophage (data not shown), both genes encode proteins with TMs in equivalent positions (the same applies for STY2629 of SPI-17 – see next paragraph). It seems possible that those proteins have similarity on the structural rather on the sequence level which might indicate similar function. Moreover the DR at the 5’ end of SPI-16 has significant sequence similarity (74% in 23nt) with the 23bp P22 bacteriophage attP attachment site (alignment in Figure 2.9).



```
attP:----ATGCGAAGGTCGTAGGTTCGACT----------------
         *  **  ***** ******** *
DR:  TCTAAGCCGTGGGTCGCAGGTTCGAATCCTGCAGGGCGCGCCA
```

Figure 2.9: ACT screenshot: BLASTN comparison between bacteriophage P22 and Typhi CT18 (from top to bottom). The highlighted yellow band represents the sequence similarity between the P22 phage integrase and the integrase pseudogene in SPI-16. Within the grey text box the sequence alignment between the DR of SPI-16 and the P22 bacteriophage attP is provided; identical bases are indicated with an asterisk.

These data support the phage origin of SPI-16 and indicate that this island seems to have been originated from a phage that shares similarities with P22 bacteriophage family. SPI-16 is absent from *E. coli*, *S. bongori* and *S. arizonae* while it is present in the rest of the *Salmonella* lineage (Figure 2.7). Interestingly in *S. bongori* at the same tRNA location, there is a different insertion (8155bp) with a phage integrase gene, suggesting that this tRNA locus might represent a hotspot for integration of different GIs in the *Salmonella* lineage.

The third novel island, SPI-17 is 5.1kb long, inserted in a tRNA^{Arg} gene. An integrase gene and DRs/IRs seem to be absent from this island, which is present in all the *Salmonella* genomes used in this study, apart from *S. bongori*, *S. arizonae*, and *S. typhimurium*; this observation may indicate a possible recent deletion event that took place in the genome of *S. typhimurium* (Figure 2.7). SPI-17 seems to belong to the same phage family as SPI-16 given that the two serotype converting genes (*gtrA* and *gtrB*) are also present in the former island and both show high similarity with homologous genes in P22 bacteriophage; moreover in SPI-17 there is a pseudogene (STY2631a) with sequence similarity to the P22 phage bifunctional tail protein coding gene (TSPE_BPP22), suggesting an island of phage origin with two well defined boundaries (*gtrA* and the phage tail protein coding gene).

2.3.3 Predicted boundary optimization

As mentioned earlier, given that the current method is sliding window-based, the step of the window significantly affects the accuracy of the localization of the predicted boundaries.

The implementation of a HMM model in a change-point detection framework seems to provide an effective way of dealing with this problem (Table 2.5). Indeed the average absolute error δx for the predicted boundaries with the implementation of the HMMs is much lower (3830bp) than that without the boundary optimization (4936bp). Interestingly the HMM-based approach gives an average δx quite close to the W8 method (Tsirigos and Rigoutsos, 2005) (3543bp); W8 is a gene-based method, thus it is expected to provide quite accurate predicted boundaries of HGT events.

Overall this indicates that the implementation of HMMs in a change-point detection framework significantly improves (22%) the localization of the predicted boundaries; an example is illustrated in

Figure 2.10. This region is absent from the genome of *E. coli* and *S. typhimurium* and the BLASTN comparison indicates a well defined putative horizontally transferred region, 5223bp long, consisting of four genes (STY3343, STY3344, STY3345, STY3347: putative membrane and putative hypothetical genes of no significant database hits). As illustrated in the score plot in Figure 2.10, the unoptimized boundaries (green coloured plot) were predicted in the middle of STY3343 and STY3349 genes.

Table 2.5: Absolute error of the Alien_Hunter algorithm for the predicted boundaries with (optimized) and without (unoptimized) the implementation of HMMs in a change point detection framework. In addition the absolute error of the W8 (gene-based) method is provided as a control set. The absolute error is defined as $\delta x = | x - x_0 |$, where x is the annotated boundary and x_0 is the predicted one.

Annotated HGT	**Boundaries(bp)**								**Absolute Error (bp)**					
	Annotated		**Optimized (HMM)**		**Unoptimized**		**W8**		**Optimized**		**Unoptimized**		**W8**	
	left	right	left	right	left	right	left	right	left	right	left	right	left	right
SPI-6	302172	361067	302445	358919	302500	362500	306935	360757	273	2148	328	1433	4763	310
Prophage10	1008747	1051266	999914	1053088	1000000	1055000	1001995	1055793	8833	1822	8747	3734	6752	4527
SPI-5	1085156	1092735	1081688	1091828	1082500	1095000	1085337	1094839	3468	907	2656	2265	181	2104
Bacteriophage	1538899	1572919	1539019	1572916	1537500	1577500	1538899	1574581	120	3	1399	4581	0	1662
SPI-2	1625084	1664823	1624923	1650692	1622500	1652500	1622537	1667392	161	14131	2584	12323	2547	2569
Bacteriophage	1887450	1933558	1872930	1933953	1870000	1937500	1870173	1939495	14520	395	17450	3942	17277	5937
SPI-9	2743495	2759190	2743818	2754300	2742500	2755000	none	2759190	323	4890	995	4190	none	0
Bacteriophage 27	2759733	2782364	2759506	2787702	2757500	2787500	2759733	2783554	227	5338	2233	5136	0	1190
SPI-1	2859262	2899034	2862660	2900872	2860000	2902500	2861845	2900586	3398	1838	738	3466	2583	1552
SPI-8	3132606	3139414	3133940	3151951	3130000	3152500	3134156	3149714	1334	12537	2606	13086	1550	10300
Bacteriophage	3515397	3549055	3514572	3558310	3512500	3562500	3512700	3552416	825	9255	2897	13445	2697	3361
SPI-3	3883111	3900458	3888383	3904602	3887500	3907500	3888370	3902214	5272	4144	4389	7042	5259	1756
SPI-4	4321943	4346614	4321935	4348906	4320000	4350000	4321410	4349963	8	2292	1943	3386	533	3349
SPI-7	4409511	4543072	4402961	4541642	4402500	4545000	4401582	4542913	6550	1430	7011	1928	7929	159
SPI-10	4683690	4716539	4685054	4723629	4682500	4727500	4683853	4728101	1364	7090	1190	10961	163	11562
ALL (left/right)									3112	4548	3811	6061	3731	3356
ALL (left+right)									**3830**		**4936**		**3543**	

Applying the HMM approach, the true transition points were successfully identified (red plot), predicting the exact downstream and upstream boundaries of this region, diminishing the uncertainty of the localization of the predicted regions caused by the sliding window approach. The reason why I chose not to apply a purely HMM-based approach in the first place was the fact that a significant number of GIs (e.g. SPI-2) show a very mosaic structure, a result of several individual acquisitions, perhaps of different origin. Given that a HMM implementation requires the properties of the regions modeled to remain constant throughout their whole length, such an approach is not readily applicable to the prediction of GIs in microbial genomes.

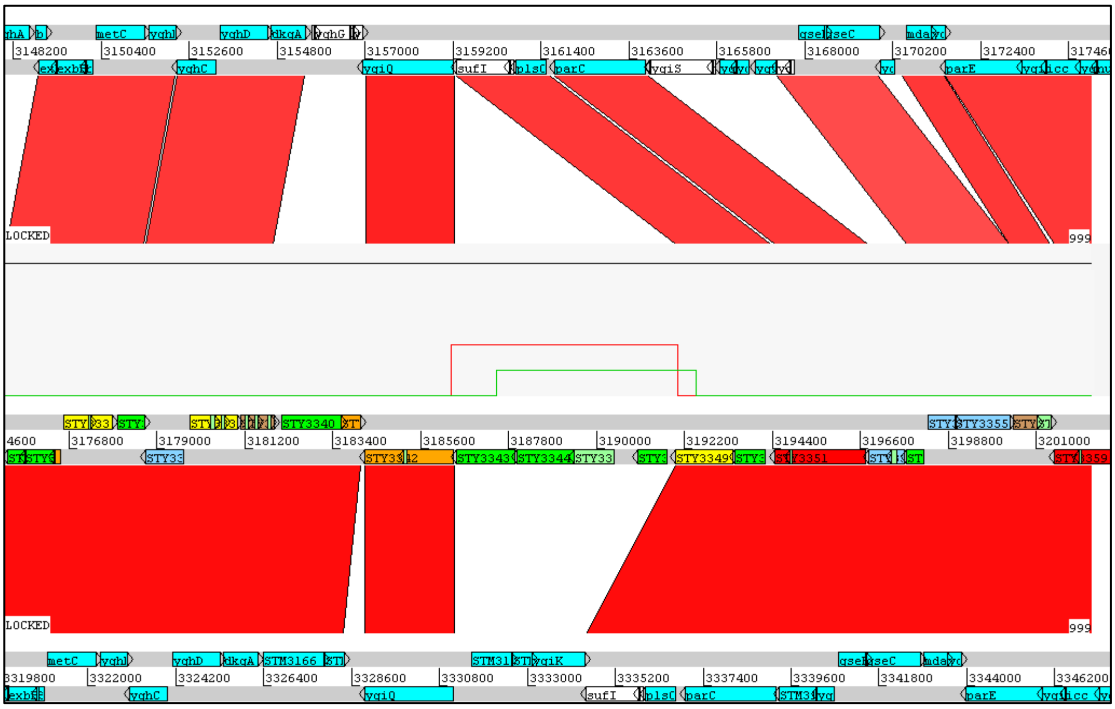


Figure 2.10: ACT screenshot: BLASTN comparison between (from top to bottom): *E. coli* MG1655, *S. typhi* CT18 and *S. typhimurium* LT2. An example of a predicted putative horizontally transferred region in the genome of *S. typhi* is indicated with two peaks in the IVOM score plot (above *S. typhi*). This region seems to be absent in the other two genomes compared. The red and the green coloured IVOM score plots represent the predictions of Alien_Hunter with optimized (HMM) and unoptimized boundaries respectively.

2.3.4 Performance benchmarking

In order to test the performance of the Alien_Hunter algorithm, a dataset of 1560 manually curated putative horizontally transferred genes in the genome of Typhi was used. Alien_Hunter was compared against four other published methods for the prediction of putative HGT events (Table 2.6): Islander (Mantri and Williams, 2004), IslandPath (Hsiao *et al.*, 2003), HGT-DB (Garcia-Vallve *et al.*, 2003), and the W8 method (Tsirigos and

Rigoutsos, 2005). Furthermore the above methods and the method for the prediction of PAIs introduced by Yoon *et al.* (Yoon *et al.*, 2005) were tested in terms of percentage coverage of the 10 previously described SPIs (SPI-1 to SPI-10) and the five annotated bacteriophages (Table 2.7).

Overall, Alien_Hunter shows the highest predictive accuracy (AC=0.764) compared with the other four methods (Table 2.6). Interestingly, the second most accurate method is W8, which utilizes higher order motifs (i.e. 8mers). These data suggest that the utilization of interpolated variable order motifs, improves both the sensitivity (SN) (Alien_Hunter: 0.649, W8: 0.62) and the specificity (SP) (Alien_Hunter: 0.653, W8: 0.643) compared with fixed-order methods; similarly this analysis confirms the superiority of higher order motif methods, discussed in the introduction.

Table 2.6: Performance comparison of the Alien_Hunter algorithm with other prediction methods. The comparison was based on the manually curated dataset of 1560 putative horizontally transferred genes, described in the text. TP: true positives, FP: false positives, TN: true negatives, FN: false negatives, SN: sensitivity, SP: specificity, AC: accuracy, CC: Matthews correlation coefficient. The performance of IslandPath was evaluated based on two compositional indices: G+C% content and di-nucleotide bias (δ* difference).

Method	TP	FP	TN	FN	Number of Predictions	SN	SP	AC	CC
Alien_Hunter	1013	539	2501	547	1552	0.649	0.653	0.764	0.473
W8	968	538	2502	592	1506	0.620	0.643	0.754	0.447
HGT-DB	435	116	2924	1125	551	0.279	0.789	0.730	0.351
Islander	275	89	2951	1285	364	0.176	0.755	0.701	0.258
IslandPath (GC)	611	467	2573	949	1078	0.392	0.567	0.692	0.266
IslandPath (δ*)	301	492	2548	1259	793	0.193	0.380	0.619	0.039

The sensitivity of Alien_Hunter is much higher compared to the other four methods which in turn reflects an increased ability to predict novel, putative horizontally transferred regions as well as already known examples. In terms of specificity Alien_Hunter is third from the top, following the Islander and the HGT-DB. Perhaps this can be attributed to the increased number of predictions provided by Alien_Hunter (1552) compared to the Islander (364) and HGT-DB (551) as well as to the fact that Alien_Hunter runs on raw genomic sequence without gene position information. Compared to the W8 method, although Alien_Hunter provides higher number of predictions, both its sensitivity and specificity are higher. In the second performance analysis, based on the percentage coverage of previously described HGT events, the Alien_Hunter predictions overlap with 91.2% of the CDSs present in SPIs and bacteriophages giving the highest number of complete GIs in Typhi, followed by the W8 method with 80.7% coverage.

Table 2.7: Performance comparison of the Alien_Hunter algorithm with other prediction methods based on a dataset of 10 previously described SPIs (SPI-1 to SPI-10) and five annotated bacteriophages (SopE and P4 bacteriophages were ignored because they overlap with SPI-7 and SPI-10 respectively). For each annotated island or phage the % CDS coverage by each method has been calculated. The genomic locations of annotated bacteriophages (from top to bottom) are: 1008747..1051266, 1538899..1572919, 1887450..1933558, 2759733..2782364 and 3515397..3549055.

Annotated HGT	**# CDS**	**Alien_Hunter**	**Islander**	**IslandPath**		**HGT-DB**	**W8**	**Yoon *et al.***
				GC	**δ^***			
SPI-6	60	81.7	0.0	51.7	41.7	40.0	70.0	0.0
Prophage10	63	81.0	100.0	23.8	39.7	25.4	96.8	0.0
SPI-5	8	100.0	100.0	75.0	100.0	100.0	100.0	100.0
Bacteriophage	53	100.0	0.0	39.6	5.7	34.0	86.8	0.0
SPI-2	44	77.3	0.0	61.4	18.2	68.2	77.3	100.0
Bacteriophage	71	88.7	0.0	33.8	8.5	35.2	94.4	0.0
SPI-9	4	25.0	0.0	25.0	50.0	0.0	25.0	0.0
Bacteriophage	19	89.5	0.0	36.8	0.0	5.3	73.7	26.3
SPI-1	44	95.5	0.0	54.5	25.0	77.3	88.6	40.9
SPI-8	16	100.0	0.0	68.8	0.0	68.8	68.8	0.0
Bacteriophage	46	89.1	0.0	37.0	6.5	23.9	60.9	0.0
SPI-3	14	85.7	0.0	28.6	0.0	14.3	42.9	100.0
SPI-4	7	100.0	0.0	85.7	0.0	100.0	100.0	100.0
SPI-7	149	100.0	31.5	31.5	32.2	28.2	81.2	10.1
SPI-10	29	100.0	44.8	44.8	62.1	6.9	72.4	0.0
ALL	627	91.2	20.9	40.5	25.0	36.8	80.7	17.7

These data suggest that Alien_Hunter is capable of detecting not only novel GIs but also of identifying the majority of the already known regions of "alien" origin. Overall Alien_Hunter predicts six complete structures (SPI-5, the bacteriophage at 1538899..1572919, SPI-8, SPI-4, SPI-7 and SPI-10), while in the case of SPI-2 it predicts 34 out of 44 genes; it has been shown previously (Hensel *et al.*, 1999) that SPI-2 is a mosaic island of at least two independent acquisitions (see chapter 3). The mosaic nature of this SPI is also apparent in the G+C content (44.08% and 52.85% for the two parts of the island). This observation might explain the fragmented prediction for this SPI by all the methods except for the method of Yoon *et al.* (Yoon *et al.*, 2005). The latter combines a method for capturing sequence deviation and similarity matches to already known PAIs to predict PAIs instead of GIs in general. Such methods can be powerful approaches in the detection of complete PAI structures of similar gene content with previously annotated ones, but are not directly applicable in the detection of novel PAIs or GIs.

Overall the W8 method only outperforms the Alien_Hunter algorithm twice: in the first case it predicts 96.8% (Alien_Hunter: 81%) of the complete structure of prophage10 and in the second case 94.4% (Alien_Hunter: 88.7%) of the bacteriophage located at position 1887450..1933558. The Islander provides the lowest number of predictions (364) perhaps due to the fact that it is restricted to predict only complete GI structures. In the case of known Typhi islands, Islander predicts three SPIs (SPI-5, SPI-7, SPI-10) and one bacteriophage (prophage 10); the rest of the already known SPIs were not predicted by this method although some of them (e.g. SPI-8) have identifiable tRNA and integrase genes.

2.4 Discussion

In this chapter, I introduced and described a novel computational method for the prediction of putative horizontally transferred regions. This method, IVOM, exploits compositional biases at various levels (e.g. codon, di-nucleotide and amino acid bias, structural constraints) by implementing variable order motif distributions. Under this framework, the local sequence composition can be captured more reliably, compared to fixed-order methods. The IVOM approach relies more on higher order motifs to make more accurate predictions, but when the underlying information is insufficient for high order motifs, it takes into account information obtained from lower order motifs. Moreover, an IVOM approach can be applied even on newly sequenced genomes, given that it does not require any level of pre-existing annotation or gene position information.

I also discussed the implementation of a HMM-based approach in a change-point detection framework for the optimization of the boundaries of the predicted regions and showed that the uncertainty of the localization of the predictions caused by a sliding window method can be sufficiently handled by such an approach enabling more accurate localization of putative HGT events. Applying the IVOM method on the genome of Typhi, all the previously annotated SPIs and bacteriophages were successfully predicted; moreover, the analysis of Typhi revealed the presence of three novel SPIs, SPI-15 to SPI-17, that have not been previously described. SPI-16 and SPI-17 represent islands of putative phage origin that may be implicated in serotype conversion by O-antigen glycosylation.

The performance benchmark of the Alien_Hunter algorithm against four published methods indicates that this method is more sensitive in detecting compositionally deviating, putative HGT regions. On the other hand Alien_Hunter shows fairly poor specificity compared with HGT-DB and Islander. This observation seems to indicate that the last two methods are more reliable in terms of SP compared to Alien_Hunter. One obvious reason behind the lower SP of Alien_Hunter is the increased number of predictions (1552). HGT-DB and Islander show the highest SP due to the low number of predictions (551 and 364 respectively); in other words they sacrifice SN for SP, predicting only a small fraction of the already annotated HGT regions (Table 2.7).

However if both SP and number of predictions are taken into account, Alien_Hunter provides the highest number of predictions and at the same time its SP is even higher than W8's, although the latter provides a lower number of predictions (1506). Overall this indicates that Alien_Hunter can be more sensitive and accurate compared to other methods that provide equally high number of predictions. It should be noted that this performance benchmark is based on a reciprocal FASTA approach that might penalize older HGT regions that were inserted prior to the divergence of *E. coli* and *Salmonella* lineages and were predicted by the Alien_Hunter algorithm. Such cases are considered False Positives based on this analysis, although they might represent true HGT events, and significantly affect the assigned SP of this algorithm.

Furthermore, the approach for the identification of orthologous genes, exploiting a reciprocal FASTA methodology, would in theory fail to correctly predict true orthologs in the following cases: A. One or both orthologs are pseudogenes, B. one of the orthologs has been deleted in one of the two genomes, C. a gene duplication event has created extra copies of the corresponding ortholog(s), D. one of the orthologs has not been annotated in one of the two genomes – although being present, E. one of the orthologs has been mis-annotated (truncated or extended) to such an extend that the condition of the minimum length of the region being similar in the two sequences is violated, and F. one or both orthologs are fast evolving, to such an extend that it is impossible, relying purely on sequence information, to predict them as true orthologs. With the exception of case F, all the other cases can be manually inspected and corrected, exploiting the genome annotation and gene position information; therefore, the results discussed in this chapter as well as in chapters 3 and 4, relative to the number of horizontally acquired genes, should be treated as an upper bound to the true number of HGT events, since fast evolving orthologs, could in theory be incorrectly classified as horizontal acquired genes.

The prediction of the three novel SPIs in Typhi CT18, raises the following question: *What is the minimum size of PAIs or GIs that still maintain their ability to mobilize (integrate-excise)?* Usually GIs are expected to be large (≥ 10kb), distinct chromosomal regions (Schmidt and Hensel, 2004). The three novel SPIs described in this analysis seem to represent exceptions to this rule, with a size of 4-6kb. For example SPI-17 is a minute PAI, and is absent from the genome of Typhimurium LT2, possibly indicating a recent deletion or recombination event. The size of these regions may be the reason why they have not been previously reported.

SPI-15 encodes four hypothetical protein-coding genes with unknown function. Moreover while SPI-15 is only present in Typhi CT18 and TY2, it can also be found in *Shigella flexneri* serovar 2a, strains 301 and 2457T. Given that SPI-15 or similar structures are present in *S. flexneri* and *S. typhi* but not in *E. coli* (MG1655, EDL933, O157:H7 and CFT073) or other *Salmonella*, it would be interesting to further investigate the functionality of SPI-15 with respect to the biology of *S. typhi* and *S. flexneri*, given that both organisms are human-restricted enteric pathogens.

The annotation of horizontally transferred regions (e.g. GIs, phages) is a key task in annotation pipelines, especially in the case of pathogens since it can reveal pathogenic aspects and characteristics of newly sequenced genomes. Prediction methods that reliably detect regions of “alien” origin, requiring a minimum level of annotation, can form a powerful tool for the understanding and analysis of the biology for the genome at hand, revealing key evolutionary steps in becoming a “successful” pathogen (see chapter 3).

Chapter 3

Genetic flux over time

3.1 Introduction

The divergence of *Salmonella* and *E. coli* lineages from their common ancestor has been estimated to have occurred approximately 100-140 million years (Myr) ago (Doolittle *et al.*, 1996; Ochman and Wilson, 1987). Using models of amelioration (i.e. the change of the sequence composition over time) to estimate the time of Horizontal Gene Transfer (HGT) events it has been previously inferred (Lawrence and Ochman, 1997) that the entire *E. coli* chromosome contains more than 600 kilobases (kb) of horizontally transferred, protein-coding DNA and that the two sister lineages (*E. coli* and *S. enterica*) have each gained and lost more than 3 megabases (Mb) of novel DNA since their divergence.

DNA sequences of recent HGT events can deviate strongly from the genome background composition while older insertions have often lost their donor-specific sequence signature (Lawrence and Ochman, 1997). Generally, each genome exhibits a reasonably constant background sequence composition; however some genes, traditionally considered part of the core-gene dataset, such as rRNA and ribosomal protein-coding genes often deviate compositionally from the genome background sequence composition mainly due to specific, well-conserved functional constraints rather than alien origin (although some of them can be horizontally exchanged (Nomura, 1999; Yap *et al.*, 1999)). In those cases the effect of the amelioration over time is expected to be limited since strong selection applies.

Base composition and specifically G+C content is known to be related to phylogeny (Forsdyke, 1996). Consequently closely related organisms tend to have similar G+C content; for example the average G+C content of *E. coli*, *Shigella* and *Salmonella* lineages is approximately 50%, 51% and 52% respectively while for the Gram positive *Staphylococcus* and

Streptococcus lineages the average G+C content is 33% and 38%, respectively.

Usually horizontally acquired genes are introduced into a single lineage, and therefore the acquired DNA sequence will be limited to the descendents of the recipient strain and absent from closely related ones. For example *Salmonella* Pathogenicity Island (SPI) 1, a 40kb island carrying a type-III secretion system (T3SS) that enabled the invasion of epithelial cells (Galan, 1996) is present in both *Salmonella* species, *S. bongori* and *S. enterica* while it is absent from the genome of *E. coli*. Consequently SPI-1 represents an ancient HGT event that took place close to the divergence of the two genera (*E. coli* and *Salmonella*) (Baumler, 1997).

On the other hand SPI-2, which is important for systemic infection, is a mosaic of two independent acquisitions: The tetrathionate reductase (*ttr*) gene cluster (Hensel *et al.*, 1999), a 15kb region present in *S. bongori* and *S. enterica*, and a 25kb region, encoding an additional T3SS (Hensel *et al.*, 1997), present only in *S. enterica*. Consequently, using a reference tree topology, HGT events can be distributed into phylogenetic branches of increasing depth; moreover their relative time of insertion, i.e. the most ancient branch in the tree topology that shares a putative horizontally acquired (PHA) gene present only in descendant lineages, can be inferred. Based on this principle, Daubin and Ochman (Daubin and Ochman, 2004), identified sequences unique to monophyletic groups at increasing phylogenetic depths, and studied the characteristics of sequences with no detectable database match (ORFans) using *E. coli* MG1655 as a reference genome.

A key step in inferring the relative time of insertion of PHA genes is the construction of phylogenetic trees that will capture reliably the evolutionary history of the organisms being studied. rRNA genes have been extensively used as molecular chronometers for inferring phylogeny and building tree topologies (Woese, 1987). However, it has been shown that even these traditionally core components of the cell can be horizontally transferred (Nomura, 1999; Yap *et al.*, 1999). Consequently more reliable phylogenies can be built based on approaches exploiting larger sequence samples, e.g. whole-genome sequence (Doolittle, 1999; Doolittle and Papke, 2006). Moreover homologous recombination might well complicate the inference of the true evolutionary history of the genomes under study (Doolittle and Papke, 2006; Feil *et al.*, 2001; Smith *et al.*, 1993). Many closely related bacteria exchange a significant amount of DNA via homologous recombination through highly similar patches throughout their genome sequence (Didelot *et al.*, 2007). Therefore different regions within those genomes might well have different evolutionary histories that cannot be reliably captured by phylogenies relying on a single tree topology (Doolittle and Papke, 2006).

In this chapter, I describe a comparative analysis (Vernikos *et al.*, 2007) between eleven *Salmonella*, three *E. coli* and one *Shigella* strain in order to infer the relative time of insertion of putative HGT events in three strains of the *S. enterica* lineage, by implementing a whole-genome sequence alignment to construct the phylogenetic tree topology of the organisms under study. The relative time of insertion is inferred taking into account the most parsimonious sequence of events i.e. allowing for deletions or independent acquisitions in some of the descendant or ancestral branches. Moreover I discuss and analyse data suggesting that prophages in the *Salmonella* lineage are shared only between very recently diverged lineages but that their sequence composition is very similar to their host's. Finally I describe the implementation of G+C content, the Codon Adaptation Index (CAI) (Sharp and Li, 1987) and high order compositional vectors (Vernikos and Parkhill, 2006), in order to monitor the amelioration process over time.

3.2 Methods

3.2.1 Whole Genome Alignment

The extent of intra-species diversity of bacterial populations was shown recently in a study focused on whole-genome sequence comparisons of eight *Streptococcus agalactiae* isolates (Tettelin *et al.*, 2005); the results show that the genome of a bacterial species (i.e. pan-genome) (Medini *et al.*, 2005) consisting of core and dispensable (i.e. partially shared) genes, may be many times larger than the genome of a single isolate; consequently single-genome sequences may represent poor samples of the overall complexity and structure of a bacterial species.

In the current analysis the phylogenetic relationship of 15 genomes will be inferred by pursuing a whole-genome based comparative genomics approach; the rationale behind a whole-genome based methodology as opposed to marker-based or core-gene based approaches relies on two important aspects of comparative genomics; gene content information and phylogenetic resolution.

In the first case, the dispensable gene pool of a species often encodes components that drive host-adaptation, antigenic variability and determine pathogenicity and virulence related properties of different isolates; for example, HGT can in a single step transform a normally benign organism into a pathogen, a process often referred to as “evolution in quantum leaps” (Groisman and Ochman, 1996); in *Salmonella* two single-step HGT events enabled the invasion of host cells, evading the host defence system, while its close relative *E. coli* evolved as an opportunistic and commensal pathogen.

In terms of phylogenetic resolution, traditional classification systems geared towards analyzing a handful of genetically distinct, often non-overlapping species representatives are capturing only a tiny fraction (Table 3.1) of the species variation (Medini *et al.*, in press); as such they struggle to cope with the increasingly complex structure, the overlapping (fuzzy) boundaries and the dynamic nature of bacterial populations. Moving from single-gene (e.g. 16s rRNA (Woese, 1987)) phylogenies trying to capture the phylogenetic history of an entire bacterial species exploiting only a tiny sequence sample (~0.07%) of a genome, to approaches using a much larger sequence sample (~0.2%) (e.g. multilocus sequence typing – MLST (Maiden *et al.*, 1998)) and recently to whole-genome (Tettelin *et al.*,

2005) comparative genomics (100% coverage), is definitely a big step closer to understanding and more reliably reconstructing the phylogenetic history of bacterial populations.

Table 3.1: Properties of four methods for the comparative analysis of microbial genomes. Estimates have been calculated based on: [a]*Neisseria meningitidis*: genome size ~2.2 Mb (Bentley *et al.*, 2007), 16S rRNA length ~1.5kb (Sacchi *et al.*, 2002), length of MLST loci ~4kb (Maiden *et al.*, 1998). [b]*Salmonella typhi*: genome size ~4.8 Mb (Deng *et al.*, 2003), SNPs on gene fragments covering ~89 Kb (Roumagnac *et al.*, 2006). Source: (Medini *et al.*, in press).

Method	Genome coverage (%)	Core genes	Dispensable genes
16s rRNA	0.07[a]	Yes	No
MLST	0.2[a]	Yes	No
SNPs	2[b]	Yes	Yes
Whole-genome	100	Yes	Yes

Taking into account the increased genome fluidity and sequence mosaicism of bacterial chromosomes due to genome rearrangements, gene gain, gene loss and recombination events, conventional multiple sequence alignment methods, e.g. ClustalW (Thompson *et al.*, 1994) and MUSCLE (Edgar, 2004) that assume sequence co-linearity are not directly applicable on building whole-genome sequence alignments.

MAUVE (Darling *et al.*, 2004), is a genome comparison tool that merges chromosomal sequence rearrangement analysis with conventional multiple sequence alignment methods, providing an efficient method of building whole-genome multiple sequence alignments taking into account extensive chromosomal reordering. In the case of non collinear chromosomal sequences, MAUVE firstly identifies locally collinear regions (termed locally collinear blocks – LCBs) within the chromosomes that represent regions of sequence similarity shared between two or more genomes; in a second step the identified LCBs are progressively aligned using conventional multiple sequence alignment methods (i.e. ClustalW or MUSCLE). The overall algorithm is summarized in the following pseudocode:

Algorithm: MAUVE.

1. Identify multiple maximal unique matches (multi-MUMs), i.e. local alignments of exactly matching (single-copy) sequences that are shared between 2 or more chromosomes.
2. Calculate a phylogenetic guide tree based on the multi-MUMs sequences.
3. Partition a subset (anchors) of the multi-MUMs into LCBs.
4. Do recursive anchoring to identify new anchors within and outside the LCBs.
5. Align each LCB based on the guide tree.

Source: (Darling *et al.*, 2004).

Note: Formally, an LCB is a sequence of multi-MUMs that satisfies a total ordering property, such that the left end of the *i*th multi-MUM occurs before the left end of the *i*+1 multi-MUM, for all multi-MUMs in the LCB and for all the genomes compared.

In the current analysis, 15 genomic sequences (Table 3.2) were aligned by implementing the MAUVE algorithm with the default parameters. An example of the whole-genome sequence alignment of the 15 chromosomes, visualized through the alignment viewer of MAUVE is shown in Figure 3.1. The whole genome sequence alignment of 122 LCBs shared between the 15 genomes was used to build a whole-genome based phylogenetic tree (discussed in the next section).

Table 3.2: The list of 15 strains used in this comparative analysis.

Organism	Reference	Accession Number
Escherichia coli K-12 MG1655	(Blattner *et al.*, 1997)	U00096
E.coli O157:H7 EDL933	(Perna *et al.*, 2001)	AE005174
E. coli CFT073	(Welch *et al.*, 2002)	AE014075
Shigella flexneri serotype 2a 301	(Jin *et al.*, 2002)	AE005674
Salmonella bongori 12419	http://www.sanger.ac.uk/Projects/Salmonella/	N/A
S. arizonae RSK2980	http://genome.wustl.edu/genome_index.cgi	N/A
S. enterica serovar Typhi CT18	(Parkhill *et al.*, 2001)	AL513382
S. enterica serovar Typhi TY2	(Deng *et al.*, 2003)	AE014613
S. enterica serovar paratyphi A SARB42	(McClelland *et al.*, 2004)	CP000026
S. enterica serovar paratyphi A AKU_12601	http://genome.wustl.edu/genome_index.cgi	N/A
S. enterica serovar Typhimurium SL1344	http://www.sanger.ac.uk/Projects/Salmonella/	N/A
S. enterica serovar Typhimurium LT2	(McClelland *et al.*, 2001)	AE006468
S. enterica serovar Typhimurium DT104	http://www.sanger.ac.uk/Projects/Salmonella/	N/A
S. enterica serovar Enteritidis PT4	http://www.sanger.ac.uk/Projects/Salmonella/	N/A
S. enterica serovar Gallinarum 287/91	http://www.sanger.ac.uk/Projects/Salmonella/	N/A

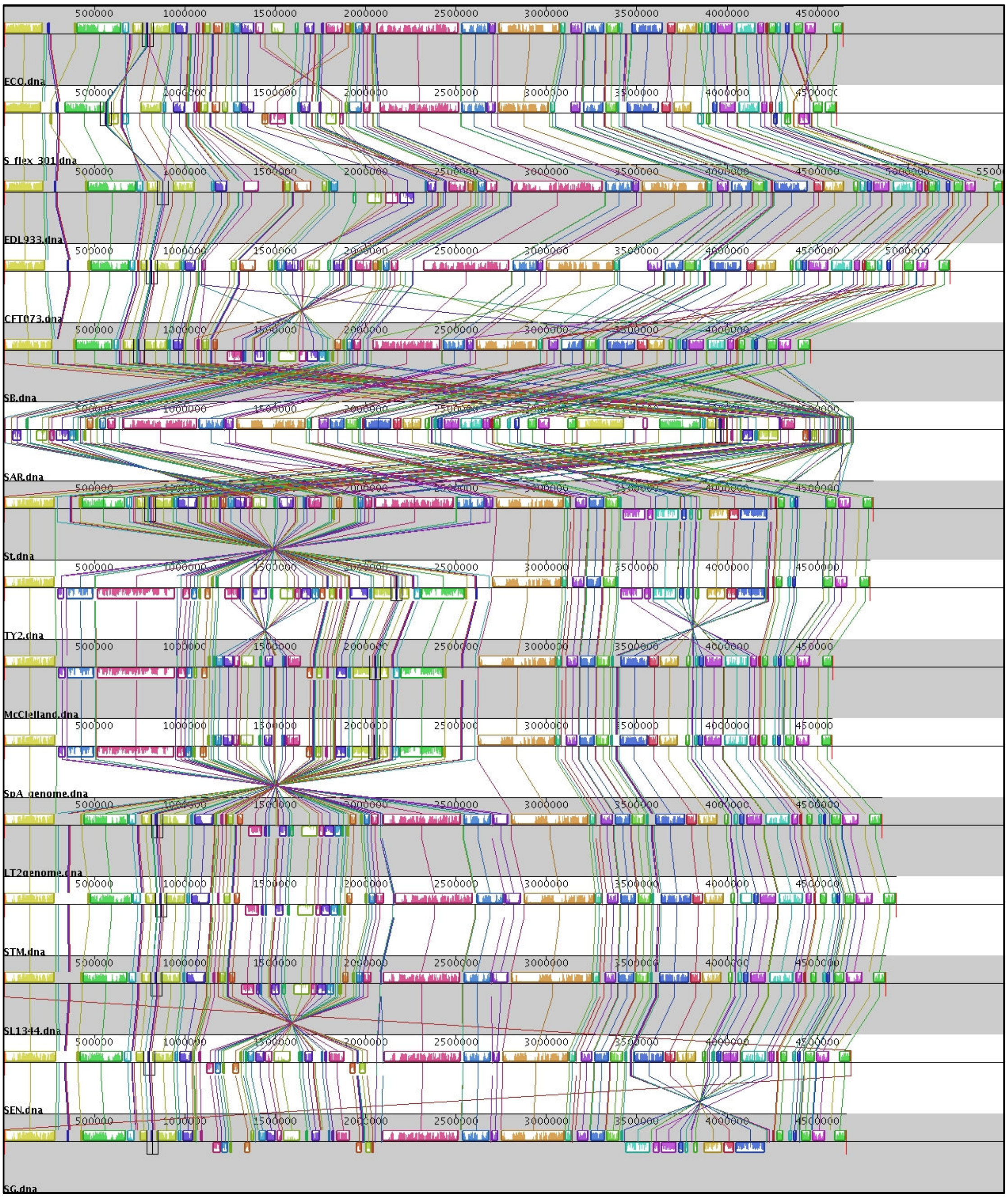


Figure 3.1: MAUVE alignment viewer screenshot: 15 genomes have been aligned (from top to bottom): *E. coli* MG1655, *E. coli* EDL933, *E. coli* CFT073, *S. flexneri* 2a 301, *S. bongori* 12419, *S. arizonae* RSK2980, *S. typhi* CT18, *S. typhi* TY2, *S. paratyphi* A SARB42, *S. paratyphi* A AKU_12601, *S. typhimurium* SL1344, *S. typhimurium* LT2, *S. typhimurium* DT104, *S. enteritidis* PT4, *S. gallinarum* 287/91. Coloured boxes (LCBs) above (forward) and below (reverse orientation) the line represent regions within each chromosome aligned with other regions with sequence similarity, present in the other chromosomes. Inside each aligned box, a similarity profile plot shows the average level of conservation of that sequence. Vertical lines connect the corresponding (same colour) LCBs present in the 15 chromosomes.

3.2.2 Phylogenetic tree building methods

In the current phylogenetic analysis, the aim is to estimate the phylogenetic tree topology that best describes the evolutionary history of the 15 enteric bacteria, taking into account their increased level of genetic fluidity. There are three major "schools" of tree-building methodologies (Table 3.3) widely used to infer the most likely phylogenetic tree: distance-based methods, e.g. unweighted pair-group method with arithmetic mean (UPGMA) (Michener and Sokal, 1957) and Neighbor-Joining (NJ) (Saitou and Nei, 1987); maximum parsimony (MP) methods; and statistical methods, e.g. maximum likelihood (ML) (Felsenstein, 1981) and Bayesian inference (Holder and Lewis, 2003; Huelsenbeck *et al.*, 2001).

Table 3.3: Properties of four widely used tree-building methods.

Method	**Pros**	**Cons**
Neighbor-Joining	Very fast, $O(n^3)$ for *n* taxa.	Does not necessarily produce the minimum-evolution (optimal) tree.
Maximum-Parsimony	Provides information on the ancestral sequences.	Ambiguous results if homoplasy is common ("long branch attraction"). Underestimates branch lengths.
Maximum-Likelihood	Site-specific likelihoods. Accurate branch lengths.	Computationally intensive.
Bayesian-inference	Faster than ML. Accurate branch lengths.	Relies on the prior distribution over the parameters of the model.

There are numerous previous studies arguing for or against the accuracy and reliability of those methods, exploiting different test datasets and model parameters (Huelsenbeck, 1995; Saitou and Imanishi, 1989; Tateno *et al.*, 1994). Although their evaluation leads to different conclusions, they all converge over the superiority of the NJ and ML methods over the MP method. For those reasons, both of those methods were exploited in the current methodology implementing the DNAML and NEIGHBOR modules of the PHYLIP software (Felsenstein, 1989), discussed in more detail in the following sections.

Generally speaking, the number of all different possible tree topologies grows rapidly with the number of taxa. It can be shown (Felsenstein, 1978) that the number of alternative topologies for an unrooted tree as a function of the number of taxa (*T*), is:

$$A(T) = \prod_{i=3}^{T} (2i-5) \quad ,$$

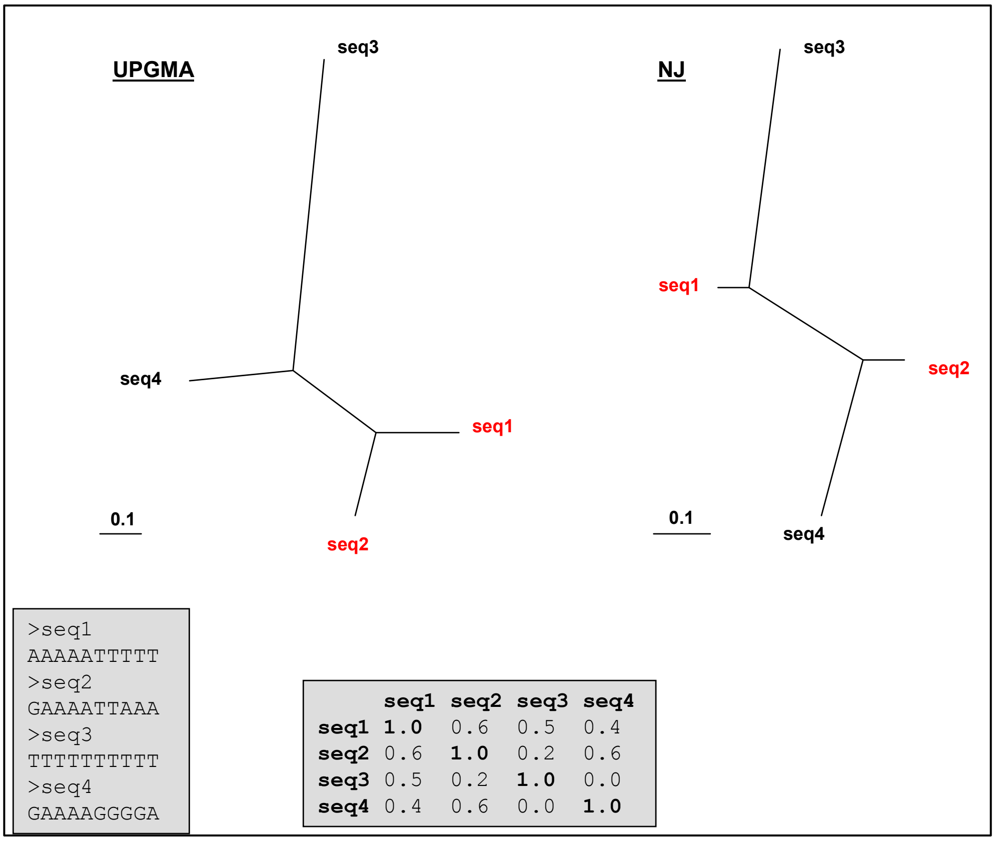
while for a rooted tree, that number is:

$$A(T) = (2T-3)\prod_{i=3}^{T} (2i-5)$$

That means that for 10 and 20 taxa, there are approximately 2 x 10^6 and 2.2 x 10^{20} alternative unrooted tree topologies, respectively.

3.2.2.1 UPGMA

The simplest (and less efficient) tree-building method is UPGMA; this method, exploits a sequential clustering algorithm that starts by identifying the two most similar (given a distance matrix) operational taxonomic units (OTUs) and then builds step-wise the phylogenetic tree topology, evaluating the similarities between the remaining OTUs; the two most similar OTUs of the previous step, are treated as a single OTU in subsequent clustering steps. The main disadvantage of the UPGMA method is that it is based on the assumption that the rate of evolution is constant over time in all the evolutionary lineages (molecular clock hypothesis); in other words, the UPGMA clustering finds the correct tree topology only if the distances between the different taxa are ultrametric, i.e. d(A,B) ≤ *max* [d(A,C), d(B,C)], for all A, B and C; where $d(x,y)$ is the distance metric between OTUs x and y (Figure 3.2).



```
>seq1
AAAAATTTTT
>seq2
GAAAATTAAA
>seq3
TTTTTTTTTT
>seq4
GAAAAGGGGA
```

	seq1	seq2	seq3	seq4
seq1	**1.0**	0.6	0.5	0.4
seq2	0.6	**1.0**	0.2	0.6
seq3	0.5	0.2	**1.0**	0.0
seq4	0.4	0.6	0.0	**1.0**

Figure 3.2: An example of four hypothetical sequences and the inferred true topology, exploiting the UPGMA (left) and the NJ (right) method. The four sequences and the similarity matrix are shown at the bottom of the figure.

3.2.2.2 Maximum Parsimony

MP exploits the concept of parsimony that favours generally simpler over more complicated hypotheses. As such, MP is based on the assumption that the best tree topology is the one that requires the minimum number changes to explain the observed differences between the taxa, and searches for the topology with the minimal cost. If $S_k(a)$ denotes the minimal cost for assignment of character a to node k, such that:

$$S_k(a) = \min_b(S_i(b) + S(a,b)) + \min_b(S_j(b) + S(a,b))$$

the topology with the minimal cost can be found by minimizing the above function for all characters *a* and all nodes *k* of the tree; *i* and *j* denote the daughter nodes of node *k*, and $S(a,b)$ denotes the cost of substituting *a* with *b*.

The MP algorithm consists of two steps: 1) the computation of the cost for a given tree and 2) a search through all trees, to find the overall minimum of this cost; for a small number of taxa e.g. (< 10), an exhaustive search of all the possible tree topologies can be carried out; for a higher number of taxa, however, heuristic methods have to be exploited. Broadly speaking there are two major MP algorithms; weighted parsimony and traditional parsimony (Fitch, 1971). In the first algorithm, each character substitution is assigned a cost while the second algorithm counts simply the number of character substitutions.

3.2.2.3 Bayesian inference

A Bayesian approach produces the tree (or a set of equally optimal trees) that is most likely to be explained by the data (i.e. sequences); in other words it estimates the posterior probability P(H/D) of the hypothesis given the data. This is different from ML that finds the tree that is most likely to have produced the data, evaluating the probability of seeing the data given the hypothesis, i.e. P(D/H). The posterior probability, in a Bayesian implementation, is calculated exploiting Bayes' theorem:

$$P(\vartheta / D) = \frac{P(\vartheta) \cdot P(D / \vartheta)}{P(D)}$$

where *P(θ/D)* is the posterior probability of the tree, *P(θ)* is the prior probability of the tree, *P(D/θ)* is the likelihood of the data given the tree and *P(D)* is the probability of the data (can be calculated as a marginal probability and serves as a normalizing constant, i.e. the sum of the posterior probabilities is 1). The posterior probabilities can be approximated by a Markov Chain Monte Carlo (MCMC) approach (Hastings, 1970; Metropolis *et al.*, 1953) that performs a random walk through the parameter space, randomly modifying the parameters (e.g. the tree topology, a branch length or a substitution model parameter) accepting or rejecting proposed moves based on their posterior probability. If the new posterior computed is larger than the current one, the proposed move is taken, otherwise depending on the level of decrease the move is rejected or accepted; therefore, the Markov chain visits the different regions in the parameter space proportionally to their posterior probability.

3.2.2.4 Neighbor – Joining

NJ (Saitou and Nei, 1987) exploits the concept of minimum evolution (Rzhetsky and Nei, 1993), i.e. at each step the topology with the minimum total branch length is preferred. The NJ algorithm is a star-decomposition algorithm, i.e. the initial tree is a star-like topology that does not however guarantee that the optimal tree topology will be found (greedy algorithm) given that it is prone to converge over a local rather than a global maxima.

NJ is a distance-based, tree building algorithm like UPGMA that nonetheless overcomes the limitation of assuming a constant evolutionary rate for all lineages. This property is very important, and can efficiently avoid converging over the wrong tree topology in case of different evolutionary rates (i.e. the ultrametric condition does not apply); instead of selecting simply the taxa with the minimum distance $d(x,y)$ (that might well not be true neighbouring taxa, see Figure 3.2), NJ builds a new distance matrix (that corrects for different rates) by subtracting from $d(x,y)$ distance the average distances of the two taxa x and y to all the other taxa. The pseudo-code describing the NJ algorithm is given below:

Algorithm: Neighbor-Joining.

Define:

$D_{ij} = d_{ij} - (r_i + r_j)$, where

$$r_i = \frac{1}{|L|-2}\sum_{k \in L} d_{ik}$$

$|L|$ denotes the size of the set L of leaves, and d_{ij} is the distance between taxa i and j.

Initialization:

Define T to be the set of leaf nodes, one per sequence, and set $L = T$.

Iteration:

Pick a pair i, j in L for which D_{ij} is minimal.

Define a new node k, and set $d_{km} = ½\ (d_{im} + d_{jm} - d_{ij})$ for all $m \in L$.

Add k to T, with edges of lengths $d_{ik} = ½\ (d_{ij} + r_i - r_j)$, $d_{jk} = d_{ij} - d_{ik}$, joining k to i and j respectively. Remove i, j from L and add k.

Termination:

When L consists of two leaves, i and j, add the remaining edge between them, with length d_{ij}.

Source: (Durbin *et al.*, 1998).

3.2.2.5 Maximum Likelihood

As mentioned earlier, the aim in a maximum likelihood approach is to maximize the likelihood of a tree P(data|tree), i.e. the probability of the data given a tree topology and a model of evolution (see next section). For a set x of n sequences x_i, for $i = 1 \ldots n$, given a model of evolution, the aim is two-fold: (1) to search through all the possible tree topologies T with the n sequences assigned at the corresponding leaves of the tree and (2) to search over all possible branch lengths t, with the objective of finding the maximum likelihood tree, i.e. the tree with topology T and branch lengths t that maximizes $P(x \mid T, t)$.

In the case of two sequences x_1 and x_2, there is only one possible rooted tree topology T, therefore the likelihood of the tree will vary relative to the branch lengths t_1 and t_2. In this example, let $x_{1,\ m}$ and $x_{2,\ m}$ denote the residues at the mth site of the two sequences. Assigning a residue a to the root of the tree, we can calculate the probability (q_a) of having a at the root of T and of having substitutions of a by $x_{1,\ m}$ and $x_{2,\ m}$, as follows:

$$P(x_{1,m},x_{2,m},a\,|\,T,t_1,t_2)=q_aP(x_{1,m}\,|\,a,t_1)P(x_{2,m}\,|\,a,t_2)$$

In a second step, in order to calculate the probability of generating $x_{1,\ m}$ and $x_{2,\ m}$ residues at the two leaves of T, we have to sum over all different possible values of a, since we do not have any prior knowledge of what the residue at the root of the tree is:

$$P(x_{1,m},x_{2,m}\,|\,T,t_1,t_2)=\sum_a q_aP(x_{1,m}\,|\,a,t_1)P(x_{2,m}\,|\,a,t_2)$$

The final step is to calculate the full likelihood over the entire length (M) of the two sequences x_1 and x_2:

$$P(x_1,x_2\,|\,T,t_1,t_2)=\prod_{m=1}^{M}P(x_{1,m},x_{2,m}\,|\,T,t_1,t_2)$$

In order to calculate the probability P($z\,|\,y,t$) of a sequence z arising from an ancestral sequence y over the branch length t, we need a model of evolution that describes how residues are substituted by others. Details of such evolutionary models will be discussed in the next section. It can be shown that given a *transition-probability* matrix $P(t) = e^{Qt}$ that determines the probability that a given residue a will become b after time t (Q denotes the *substitution-rate* matrix that determines the rate of change between pairs of nucleotides in an infinitely small time interval dt), we can compute the maximum likelihood estimate (MLE) of a given branch length t, i.e. the value of t that maximizes the likelihood of the tree.

For example, in the case of two hypothetical nucleotide sequences x_1 and x_2, each 95 nucleotides long with 9 different nucleotides, exploiting the simplest evolutionary model of Jukes and Cantor (Jukes and Cantor, 1969) (see next section for details) the MLE of the branch length between x_1 and x_2 can be estimated (Figure 3.3) applying an expectation maximization (EM) algorithm. Generally in the case of n sequences x_1, ..., x_n with m residues, the probability of generating those residues at the n leaves of T with branch lengths t can be calculated by taking the product of the probabilities of substitutions on all branches of the tree:

$$P(x_{1,m}...x_{n,m} \mid T,t) =$$
$$\sum_{a_{n+1},a_{n+2},...a_{2n-1}} q_{a_{2n-1}} \prod_{i=n+1}^{2n-2} P(a_i \mid a_{a(i)},t_i) \prod_{i=1}^{n} P(x_{i,m} \mid a_{a(i)},t_i)$$

where $a(i)$ denotes the parent node of node i. Note that the sum is over all possible assignments of a_k to non-leaf nodes k, i.e. nodes n+1 ... 2n–1. The above probability can be calculated pursuing a post-order traversal (i.e. leaves → root direction) of the tree, exploiting the *pruning algorithm* introduced by Felsenstein (Felsenstein, 1981). If the residue at node k is a then the probability of all the leaves below k is $P(L_k \mid a)$. Having computed the probabilities $P(L_i \mid b)$ and $P(L_j \mid c)$ of all b and c, at the daughter nodes i and j of k, the probability $P(L_k \mid a)$ can be calculated as follows:

```
**Algorithm:** Maximum-Likelihood (Felsenstein).
**Initialise:**
Set: $k = 2n - 1$.
**Recursion:** Compute $P(L_k \mid a)$ for all $a$ as follows:
If $k$ is leaf node:
        Set $P(L_k \mid a) = 1$ if $a = x_{k,m}$, $(L_k \mid a) = 0$ if $a \neq x_{k,m}$.
If $k$ is an internal node:
        Compute $P(L_i \mid a)$ and $P(L_j \mid a)$ for all $a$ at the daughter nodes $i$ and $j$, and set:
$P(L_k \mid a) = \left[ \sum_b P(b \mid a,t_i) P(L_i \mid b) \right] \times \left[ \sum_c P(c \mid a,t_j) P(L_j \mid c) \right]$   (3.1)
**Termination:**
Likelihood at site $m$:
$P(x_m \mid T,t) = \sum_a q_a P(L_{2n-1} \mid a)$
```

Source: (Durbin *et al.*, 1998).

Assuming that all *M* sites are independent, the full likelihood is:

$$P(x \mid T,t) = \prod_{m=1}^{M} P(x_m \mid T,t)$$

Note that the pruning algorithm of Felsenstein's calculates successively the probabilities of the data on each subtree of the tree topology *T*. Therefore it is crucial to sum over all the ancestral states of a node only after having done so for all of its child nodes. In equation 3.1, the two terms represent the probability that residue *a* will become *b* (or *c*) over the branch length t_i (or t_j) times the probability of observing the tips of node *i* (or *j*) given the state *b* (or c), summed over all possible states *b* (or c).

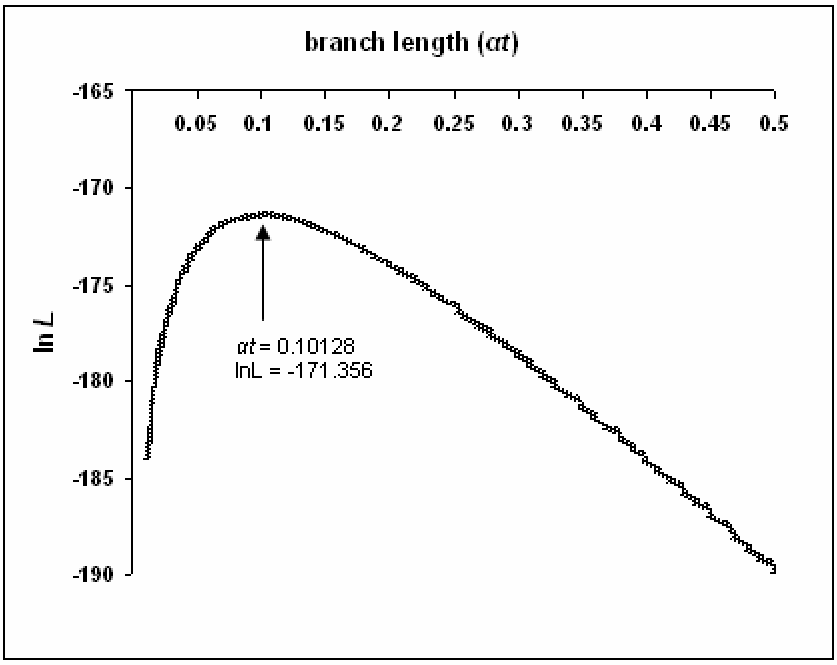


Figure 3.3: The *log* likelihood P(x_1,x_2 | *T*,*t*) for two sequences x_1, x_2 with 9 different nucleotides (nt) and a total length of 95nt, exploiting the Jukes and Cantor model. The MLE (0.10128) of the branch length is shown.

3.2.3 Nucleotide substitution models

Generally, DNA sequences derived from a common ancestor will, over time, gradually diverge due to substitution of their nucleotides. The distance between two sequences reflects the expected number of nucleotide substitutions per site, and assuming a constant over time evolutionary rate, the distance is a linear function of the time of divergence. The simplest estimate of the distance between two sequences is the proportion (p) of sites at which the two sequences differ. For example for two sequences, each 100nt long with 20 different sites, $p = 20\% = 0.2$. However because over time, the two sequences will accumulate more and more substitutions and some sites will have changed multiple times, the observed differences do not necessarily represent the true number of substitutions that have occurred since the divergence of the two sequences.

Therefore, for sequences diverged long time ago, p underestimates the number of substitutions, since it does not take into account multiple substitutions (Figure 3.4). For that reason, more sophisticated and realistic evolutionary models have to be exploited in order to estimate more reliably the true evolutionary time elapsed since the divergence of two sequences, taking into account the various aspects of the dynamics dictating the substitutions of nucleotide residues.

3.2.3.1 Jukes-Cantor model

The simplest evolutionary model (Figure 3.5), introduced by Jukes and Cantor (Jukes and Cantor, 1969), assumes that every nucleotide changes into any other nucleotide with exactly the same rate a. For two nucleotide residues i and j (where i, j = T, C, A or G), let q_{ij} denote the instantaneous rate of substitution of i by j. Those substitution rates for all 16 different combinations of nucleotide pairs can be represented in the form of a *substitution-rate* matrix Q:

$$Q = \{q_{ij}\} = \begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix}$$

Note that for any nucleotide *i* the total rate of substitution is 3α, and the order of nucleotides in the matrix is: T, C, A, G.

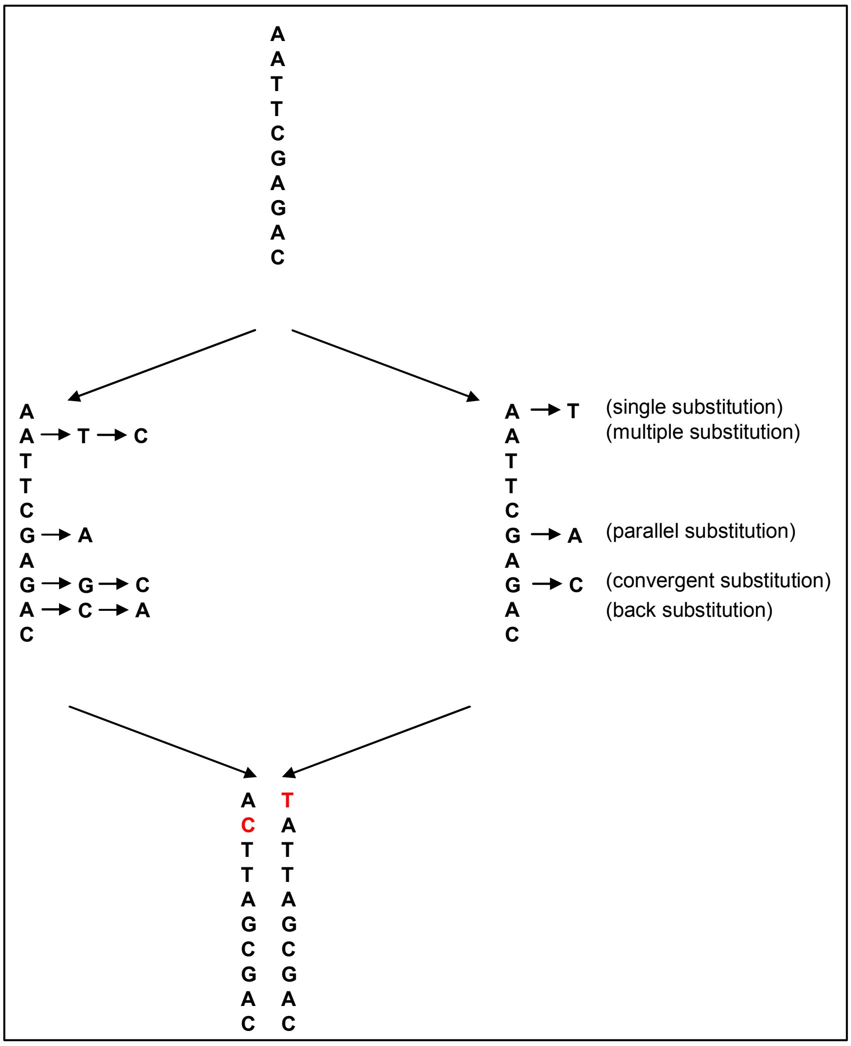


Figure 3.4: An example of multiple substitutions at the same site for a set of two hypothetical sequences diverged from a common ancestral sequence (top). Only two observed substitutions (p = 0.2) are inferred, while the true number of substitutions is 10, i.e. 1.0 substitutions per site.

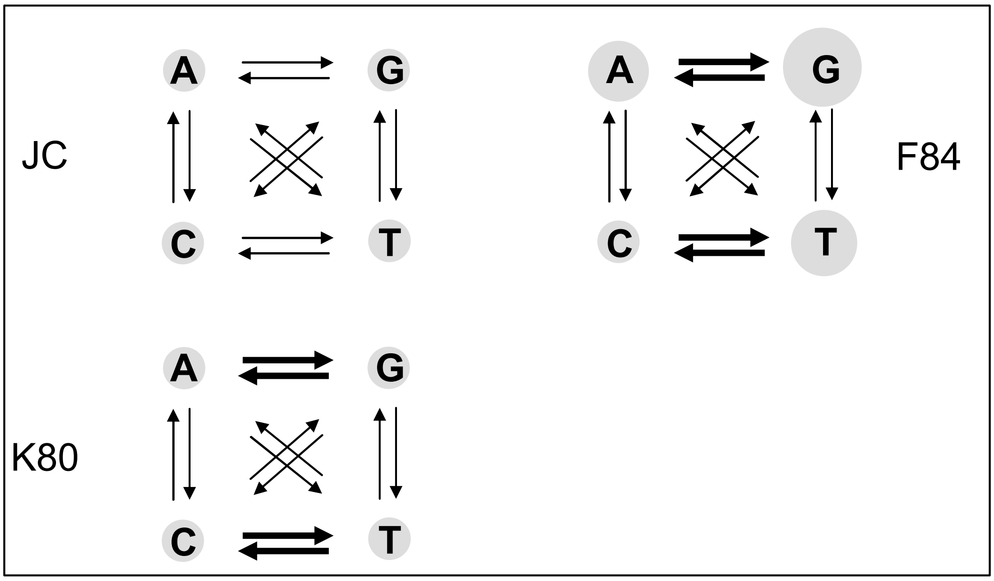


Figure 3.5: Three models of nucleotide substitution; JC (Jukes and Cantor, 1969), K80 (Kimura, 1980) and F84 (Kishino and Hasegawa, 1989).Arrows of different thickness represent different substitution rates and circles of different size the different nucleotide equilibrium frequencies.

$q_{ij}\ dt$ represents the probability of $i \rightarrow j$ change over an infinitely small time interval *dt*. However in the case of biological sequences, we are more interested in longer time t ($t > 0$) periods, over which residue substitutions occur. In other words we want to estimate the transition probability p_{ij} (t) of i being substituted by j after time t. The 16 different transition probabilities $p_{ij}(t)$ can be represented in the form of a *transition-probability* matrix:

$$P(t) = e^{Qt} = \begin{bmatrix} p_r(t) & p_s(t) & p_s(t) & p_s(t) \\ p_s(t) & p_r(t) & p_s(t) & p_s(t) \\ p_s(t) & p_s(t) & p_r(t) & p_s(t) \\ p_s(t) & p_s(t) & p_s(t) & p_r(t) \end{bmatrix},$$

where:

$$p_r(t) = \frac{1}{4}\left(1 + 3e^{-4at}\right)$$

$$p_s(t) = \frac{1}{4}\left(1 - e^{-4at}\right) \quad .$$

Using the *transition-probability* matrix $P(t)$ we can calculate over the time period t, the probability of nucleotide i having being substituted by j (Figure 3.6). Note that for $t\rightarrow\infty$, $p_r(t) = p_s(t) = ¼$, suggesting that the nucleotide equilibrium frequencies according to the JC model are $q_T = q_C = q_A = q_G = ¼$. In other words, after time $t\rightarrow\infty$, at every site of the sequence so many substitutions have occurred that the target nucleotide is random (i.e. with equal probability of observing any of the four nucleotides).

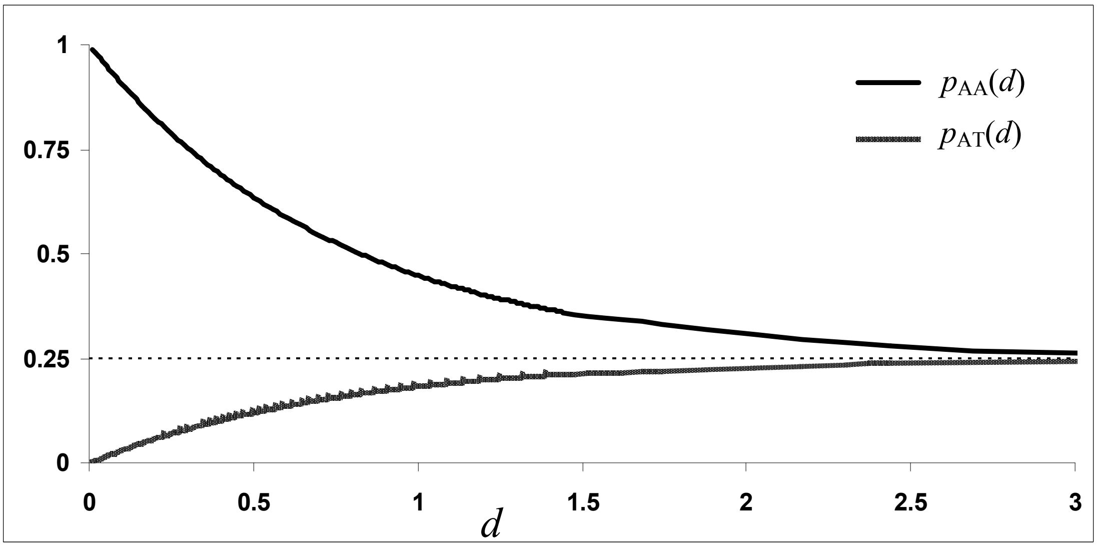


Figure 3.6: Jukes and Cantor model: transition probabilities $p_r(t)$ and $p_s(t)$ plotted against distance d (=3at); d is expressed as the expected number of substitutions per site. Assuming that for any nucleotide, the total substitution rate is 3a (see *substitution-rate* matrix Q), if two hypothetical sequences are separated by time t (i.e. diverged from their common ancestor t/2 ago) the distance d between them is 3at.

Under the JC model, for any nucleotide the total substitution rate is 3*a*, while the probability *p* of a nucleotide being different from the nucleotide of the ancestral sequence is:

$$p = 3p_s(t) = \frac{3}{4}\left(1 - e^{-4at}\right) = \frac{3}{4}\left(1 - e^{-\frac{4}{3}d}\right)$$

Consequently, if we know the proportion $\hat{p}$ of different sites between two sequences, we can estimate their distance:

$$\hat{d} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}\hat{p}\right)$$

The above equation represents the MLE (Figure 3.3) of the distance between the two sequences. Note that if two sequences are different in over 75% of their sites, the above estimate is not applicable, since their estimated distance becomes infinite.

3.2.3.2 Kimura – 2 parameter model

The JC model fails to capture a very important parameter driving the dynamics behind nucleotide substitutions; purine to purine (A ↔ G) or pyrimidine to pyrimidine (T ↔ C) substitutions (i.e. transitions) occur more frequently than substitutions between purines and pyrimidines (A,G ↔ G,C), i.e. transversions. A slightly more complex model of nucleotide substitutions that accounts for different transition and transversion rates, was introduced by Kimura (Kimura, 1980). However this model is still far from realistic, since it assumes (as the JC model does) that the nucleotide equilibrium frequencies are equal. The *substitution-rate* matrix for the Kimura 2-parameter model (K80) is:

$$Q = \begin{bmatrix} -(\alpha+2\beta) & \alpha & \beta & \beta \\ \alpha & -(\alpha+2\beta) & \beta & \beta \\ \beta & \beta & -(\alpha+2\beta) & \alpha \\ \beta & \beta & \alpha & -(\alpha+2\beta) \end{bmatrix},$$

where α denotes the transition and β the transversion substitution rates, respectively. Note that the distance d between two sequences is now $(\alpha + 2\beta)t$, and the total substitution rate for each nucleotide is $\alpha + 2\beta$. In a similar principle to the one used for the JC model, it can be shown that the estimate of the distance between two sequences is:

$$\hat{d} = -\frac{1}{2}\ln\left(1 - 2S - V\right) - \frac{1}{4}\ln\left(1 - 2V\right) \quad ,$$

where S and V are the fractions of transitions and transversions in the alignment of two sequences, respectively. Exploiting the K80 model with transition/transversion rate (k = 0.75), for the same example of the two sequences (each 95nt long with 9 different nucleotides) used in Figure 3.3, the MLE of their distance is 0.10136, (JC distance = 0.10128); note that the K80 model with k = 0.5 reduces to the JC model, giving the same distance estimate.

3.2.3.3 F84 model

A more sophisticated model (F84) of substitution with five free parameters, allowing different transition and transversion substitution rates ($\alpha \neq \beta$), as well as different nucleotide equilibrium frequencies ($q_T \neq q_C \neq q_A \neq q_G$) was proposed by Felsenstein; this model is the one exploited by the DNAML module of the PHYLIP package (Felsenstein, 1989) and the transition probabilities for this model were firstly described by Kishino and Hasegawa (Kishino and Hasegawa, 1989). The F84 model reduces to the K80 model for $q_T = q_C = q_A = q_G$, and the JC model for $2\alpha = \beta$ and $q_T = q_C = q_A = q_G$.

3.2.3.4 Substitution rate variation

So far all the evolutionary models discussed rely on a very simplifying assumption; each site in the sequence is evolving with the same rate, i.e. a single substitution matrix describes all the different nucleotide sites. However in biological sequences, this assumption rarely holds; for example, in the case of protein coding genes for each codon there are three different nucleotide positions, i.e. position 1, 2 and 3, and because of the genetic code degeneracy each position is under different mutational pressure. In the case of RNA coding genes, secondary loop and stem structures evolve with different substitutions rates. Therefore, assuming a single evolutionary rate across all the nucleotide sites underestimates the true distance between two sequences.

The rate variation among sites can be approximated by a statistical distribution, in which case the rate *r* for any site is a random variable drawn from that distribution. It has been shown that the rate variation among sites approximates the gamma distribution (Yang, 1994; Yang, 1996):

$$g(r,\alpha,\beta)=\frac{e^{-\beta r}r^{\alpha-1}\beta^{a}}{\Gamma(\alpha)}$$

for $0 < r, \alpha, \beta < \infty$, where α and β are the shape and the scale parameters, respectively. The mean of the distribution is $E(r)=\alpha / \beta$ and the variance $\text{var}(r) = \alpha / \beta^2$. The rate variation among sites is inversely correlated with the α parameter (Figure 3.7):

- If $\alpha \leq 1$, then most sites have very low substitution rates, and very few have very high rates,
- if $\alpha \rightarrow \infty$, then all sites have the same rate,
- if $\alpha > 1$, then most sites have intermediate rates and few sites have either very high or very low rates.

I will give an example showing that ignoring the rate variation among sites, leads to underestimation of the true distance between two sequences. Considering again the hypothetical sequences (length: 95nt, mismatches: 9nt) discussed in the Maximum Likelihood section above, the JC distance with the α parameter set to 0.5 (i.e. most sites have very low substitution rate), is 0.11627, much higher than the JC distance (= 0.10128) ignoring the rate variation among sites.

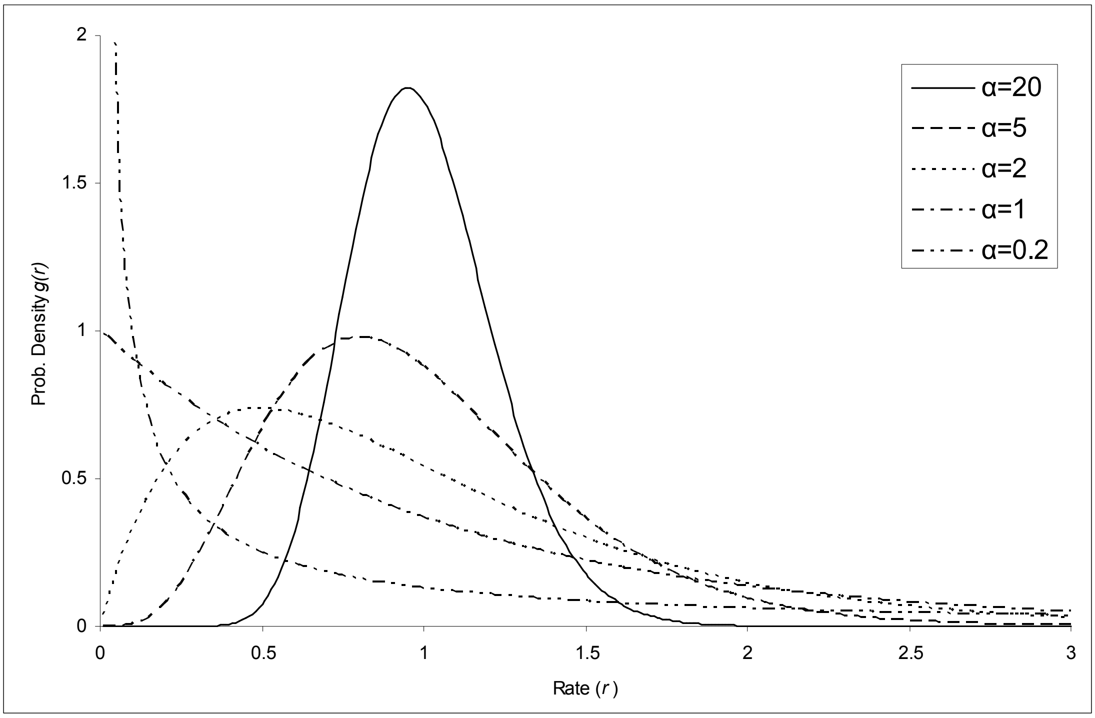


Figure 3.7: *Gamma* distribution $g\,(r,\ a,\ \beta)$; probability densities for different values of the a parameter. In this example, $a = \beta$. The mean of the distribution is $E(r)=a\ /\ \beta = 1$ and the variance $\mathrm{var}(r) = a\ /\ \beta^2 = 1/a$.

One way of estimating the different substitution rates of different sites in a multiple-alignment of sequences, is to treat the unknown r_i rate of each site i as the hidden state and the residues of each column in the alignment as the observed state in a Hidden Markov Model (HMM). With a HMM implementation, we can estimate the most probable state (i.e. rate) path that best describes the data. Defining the number of expected number k of different rates r_i and a prior probability distribution that determines the probabilities of occurrence of each rate, we can infer for each site i the most probable rate r_i. An EM technique, e.g. the Baum-Welch algorithm (Baum, 1972) can be used to estimate the parameters (i.e. emission and transition probabilities) of the HMM and a dynamic programming approach, e.g. the Viterbi algorithm (Viterbi, 1967) can be used to estimate the most probable rate path (Figure 3.8). For details about the Viterbi and the Baum-Welch algorithm refer to chapter 2. A HMM-based implementation for inferring different rates of evolution at different sites, was introduced by Felsenstein and Churchill (Felsenstein and Churchill, 1996) and implemented in the DNAML module of the PHYLIP package (Felsenstein, 1989).

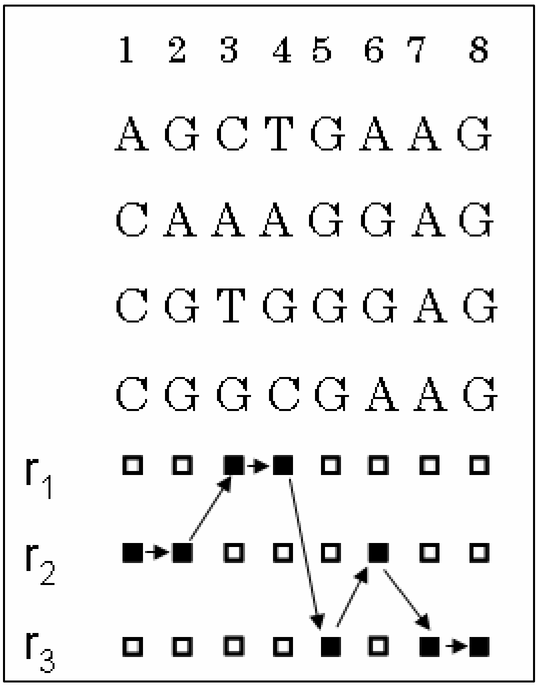


Figure 3.8: An example of four hypothetical sequences, each 8nt long. Each nucleotide site evolves under a different substitution rate (r_1 > r_2 > r_3). Assuming that there are k (=3) different substitution rates, implementing a Hidden Markov Model (HMM) approach, we can infer the most likely rate r_i for each site.

3.2.3.5 Parameter estimation

Although the Maximum Likelihood method can produce a very reliable tree topology with all the parameters (e.g. node/branch order and branch length) optimized, in the case of a large number of sequences it can be very computationally intensive. The overall aim is two-fold; search through all the possible tree topologies and then for each topology compute the maximum likelihood estimate of its branch lengths. Although the ML method is not applicable in the case of a large number of sequences, searching for the ML tree for a set of four (nucleotide or protein) sequences is a very straight forward computation (15 different rooted tree topologies).

This concept is exploited by the quartet puzzling algorithm (Strimmer and von Haeseler, 1996) and implemented by the TREE-PUZZLE software (Schmidt *et al.*, 2002). The quartet puzzling algorithm consists of three steps: 1. All possible quartet ML trees are reconstructed (ML step), 2. The quartet trees are repeatedly combined to an overall intermediate tree (puzzling step) adding sequences step-wise (with multiple input orders), 3. In the consensus step, a majority rule consensus of all intermediate trees is constructed. Because the quartet puzzling algorithm is efficiently fast, the parameters e.g. the α shape-parameter of the gamma distribution for among site rate variation, the transition/transversion rate and the nucleotide frequencies can be accurately estimated from the data, prior to the tree building (e.g. NJ or ML) method.

Using the whole-genome sequence alignment of the 15 (11 *Salmonella* and four outgroup strains) reference genomes, built by the MAUVE method, and running the TREE-PUZZLE algorithm the parameters of the evolutionary model were estimated from the data (Table 3.4). The multiple sequence alignment and the estimated model parameters were fed into the NEIGHBOR and the DNAML modules of PHYLIP (Felsenstein, 1989) to build the Neighbor-Joining and the Maximum Likelihood tree topology of the dataset, respectively.

Table 3.4: Evolutionary model parameters, estimated from the data, by the TREE-PUZZLE method, exploiting a whole-genome based multiple sequence alignment of 11 *Salmonella* and four outgroup strains.

Model of substitution	HKY85 (Hasegawa *et al.*, 1985)	
Expected transition/transversion ratio	2.22	
Expected pyrimidine transition/purine transition ratio	1.01	
Rate matrix R	A-C rate	1.00000
	A-G rate	4.38068
	A-T rate	1.00000
	C-G rate	1.00000
	C-T rate	4.38068
	G-T rate	1.00000
Nucleotide frequencies	pi(A)	23.9%
	pi(C)	26.2%
	pi(G)	26.0%
	pi(T)	23.9%
Gamma distribution – alpha parameter	a = 0.26, S.E. 0.00	
	Number of Gamma rate categories: 4	
	Category	Relative rate
	1	0.0008
	2	0.0696
	3	0.5975
	4	3.3321
	Categories 1-4 approximate a continuous Gamma-distribution with expectation 1 and variance 3.87.	
Quartet Puzzling	Number of puzzling steps	1000
	Analysed quartets	1365
	Fully resolved quartets	1365
	Partly resolved quartets	0
	Unresolved quartets	0

3.2.4 Relative time of HGT events

In order to differentiate more reliably gene loss from gene gain (HGT) in the *Salmonella* lineage, a genomic dataset of three *E. coli* (MG1655, EDL933, CFT073) and one *S. flexneri* strain was used; those four genomes form the outgroup lineage in the reference tree topology. For example a gene that is present in the *Salmonella* lineage and absent from *E. coli*

MG1655 might well be either a true HGT in the former or deletion in the latter. However, if for example, the same gene is also present in *E. coli* EDL933 and *E. coli* CFT073 then we can infer more reliably that this event probably represents a deletion (in *E. coli* MG1655) rather a true HGT in the *Salmonella* lineage. Conversely, a sequence that is confined to one lineage is more likely to have been horizontally acquired than to have been deleted independently from multiple lineages (Lawrence and Ochman, 1998).

In a parsimony model the least complex (i.e. with the lowest cost) interpretation of an observation is always favoured; in our case this is the minimum number of events or changes within a phylogenetic tree that can explain the current state of phylogenetic relationships between the taxa compared.

In the current analysis, the tree topology was used as the reference phylogenetic history of the 15 genomes compared, and genes present in three representative *S. enterica* strains (i.e. Typhi CT18, Paratyphi A SARB42 and Typhimurium LT2) were distributed on increasing depth branches of the phylogenetic tree. For example, in the case of CT18, a gene X in CT18 that has orthologs only in TY2 and the four outgroup genomes, is more likely to represent an independent HGT event in the parent node of CT18 and TY2, rather than the result of multiple deletions in the other nine genomes (Figure 3.9 A). Similarly, a gene X in CT18 that has no ortholog in the four outgroups and the *S. bongori* genome but has orthologs in the other nine genomes is more likely to have been acquired on the branch predating the divergence of *S. enterica* from *S. arizonae*. (Figure 3.9 B).

The algorithm for inferring the most likely relative time of acquisition of a PHA gene in the *Salmonella* lineage, taking into account the most parsimonious sequence of events, is summarized in the following pseudocode.

```
Algorithm: Maximum Parsimony for inferring the relative time of HGT events.
Define: $k$ is the number of the node. $a$ is the state of $k$ (“0” or “1” for gene absence or presence, respectively).

A. Ancestral state reconstruction:
Iteration (post-order tree traversal, i.e. leaves → root direction):
If $k$ is a leaf node:
        Set $S_k = a$.
If $k$ is an internal node:
        Compute $S_i$ and $S_j$ for all $a$ at the daughter nodes $i$ and $j$, of $k$.
        $S_k$:
        For $a = 0$, compute:
                $A = | a - S_i | + | a - S_j |$    (1)
        For $a = 1$, compute:
                $B = | a - S_i | + | a - S_j |$    (2)
        if (A<B) then set $S_k = 0$
        elsif (A>B) then set $S_k = 1$
        else set $S_k = [0,1]$
Note: In case of equally parsimonious ancestral states, i.e. $S_x = [0,1]$ then compute (1) and (2) for both states of $S_x$.
Termination:
If $k = 2n - 1$, where $n$ is the number of taxa.

B. Relative time of acquisition inference:
For all $k$ in the node path leading from the root of the tree to the node of the reference genome:
If $S_k = 1$ then set $t^* = k$, (break loop).
else $k--$;
where $t^*$ denotes the relative time of HGT in the Salmonella lineage (relative to the reference genome).
```

This algorithm consists of two parts; in the first part (A), the ancestral states of gene presence/absence are reconstructed in a post-order tree traversal, starting from the leaves moving towards the root of the tree. If the presence (or absence) of a gene X on ancestral branches can be unambiguously inferred, a state character 1 (or 0) is assigned on the node following the corresponding branch; alternatively both state characters (1, 0) are assigned. In the second part (B) of this algorithm, for each reference genome the relative time of acquisition of the gene in question is inferred following the node path leading from the root of the tree to the node of the reference genome; the relative time of acquisition is assigned to be the first node (more specifically the parental branch of this node) of the reference path, for which a character state 1 has been assigned.

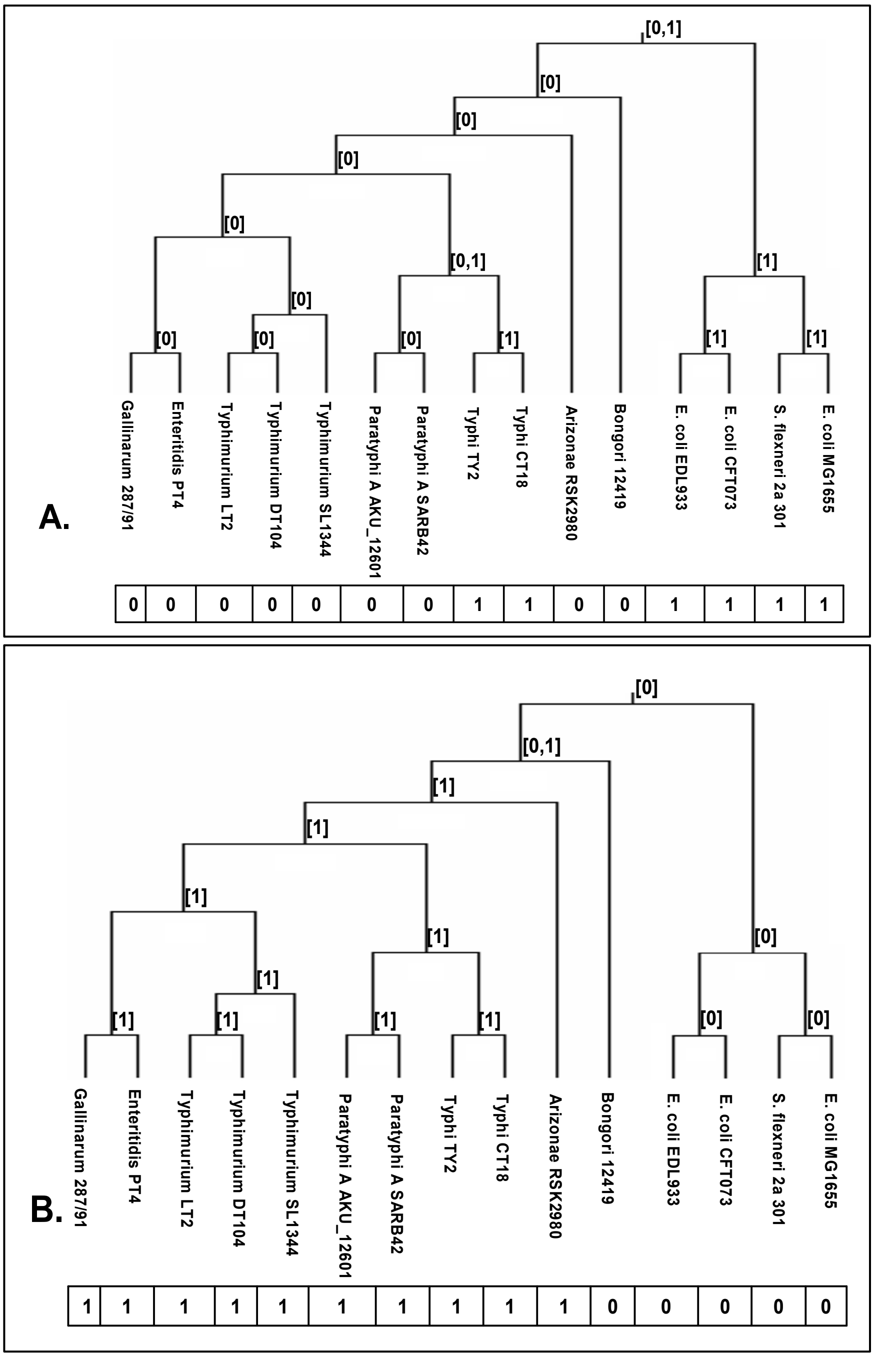


Figure 3.9: The phylogenetic distribution of a hypothetical gene X, present only in the four outgroups and the two Typhi genomes (A). The phylogenetic distribution of a hypothetical gene X, absent from the four outgroups and the Bongori genome (B). In the columns below the tree topology the presence or absence of the gene X is shown as "1" and "0" respectively. On each node, the inferred ancestral state that gives the minimum cost is shown in a binary fashion, i.e. [1] or [0]. In case of equally parsimonious ancestral states, both possible states are assigned to each node. In the first case, the inferred relative time of insertion is the branch predating the Typhi node, while in the second case it is the branch predating the divergence of *S. enterica* from *S. arizonae*.

3.2.5 Compositional analysis

In order to monitor the level of amelioration with respect to the inferred relative time of insertion for each gene in each of the three query genomes, the overall as well as the codon-position specific G+C content was calculated. The G+C content of the second codon position is generally very constrained to similar values across species (Lawrence and Ochman, 1997), given that most possible nucleotide substitutions would result in a change in the encoded aminoacid residue (non-synonymous substitutions); therefore calculating the codon-position specific G+C content increases the compositional resolution.

Furthermore in order to increase the sensitivity of capturing compositionally deviating genes (genes that do not deviate in terms of G+C content but show higher order compositional bias), I implemented the Interpolated Variable Order Motifs (IVOMs) method (Vernikos and Parkhill, 2006). In order to differentiate horizontally acquired from highly expressed genes that can also deviate compositionally, I also performed a CAI analysis (for details refer to section 1.3 of the introduction), measuring the adaptation of each gene to the codon usage of a reference set of highly expressed genes, proposed by Sharp and Li (Sharp and Li, 1986).

3.2.6 Orthologous genes

In order to identify orthologous genes, each genome in the reference dataset was compared against all the other genomes, by means of best reciprocal FASTA (Pearson, 1990) approach; overall an all-against-all comparison of 67,553 genes was performed (details of the best reciprocal FASTA algorithm are given in section 2.2.5 of chapter 2). The results were manually curated taking into account the syntenic relationship among the putative orthologs by visualizing the comparison using ACT (Carver *et al.*, 2005).

3.3 Results

3.3.1 Time distribution of PHA genes

In order to construct the tree topology that best describes the phylogenetic history of the strains studied in this analysis, I implemented the Neighbor Joining (Saitou and Nei, 1987) and the Maximum Likelihood (Felsenstein and Churchill, 1996) method, exploiting three different models of nucleotide substitution, namely JC, K80 and F84. Both (NJ and ML) methods resulted in identical tree topology illustrated in Figure 3.10. These data suggest that using whole-genome sequence information the true phylogeny of the organisms at hand can be captured reliably (see discussion for more details).

For each of the three query genomes the total number of PHA genes, as well as their relative time of insertion was inferred (Appendix B, C and D). The results are summarized in Table 3.5 and Figure 3.10. Using each of the three query genomes, on the branches prior to nodes 1, 2 and 3, I inferred similar numbers of PHA genes for the corresponding relative time of insertion (for the sake of simplicity, from this point on I will refer to the branch prior to node X as branch X). The different number of PHA genes is principally due to small differences in the number of genes in each genome (insertions, deletions, gene-remnants) as well as differences in the genome annotation.

From this point on I assign on branches 1, 2 and 3 the intersection of the respective number of genes determined on each branch using each one of the three query genomes. Overall, this reciprocal FASTA analysis suggests that approximately 2,500 orthologous genes form a core gene set shared by all the 11 *Salmonella* strains; this number reduces to approximately 2,000 orthologous genes shared by the *E. coli*, *S. flexneri* and *Salmonella* strains, used in this study (Figure 3.11). Interestingly this figure is very close to the 2,049 native genes in the γ-Proteobacteria, proposed by Daubin and Ochman (Daubin and Ochman, 2004).

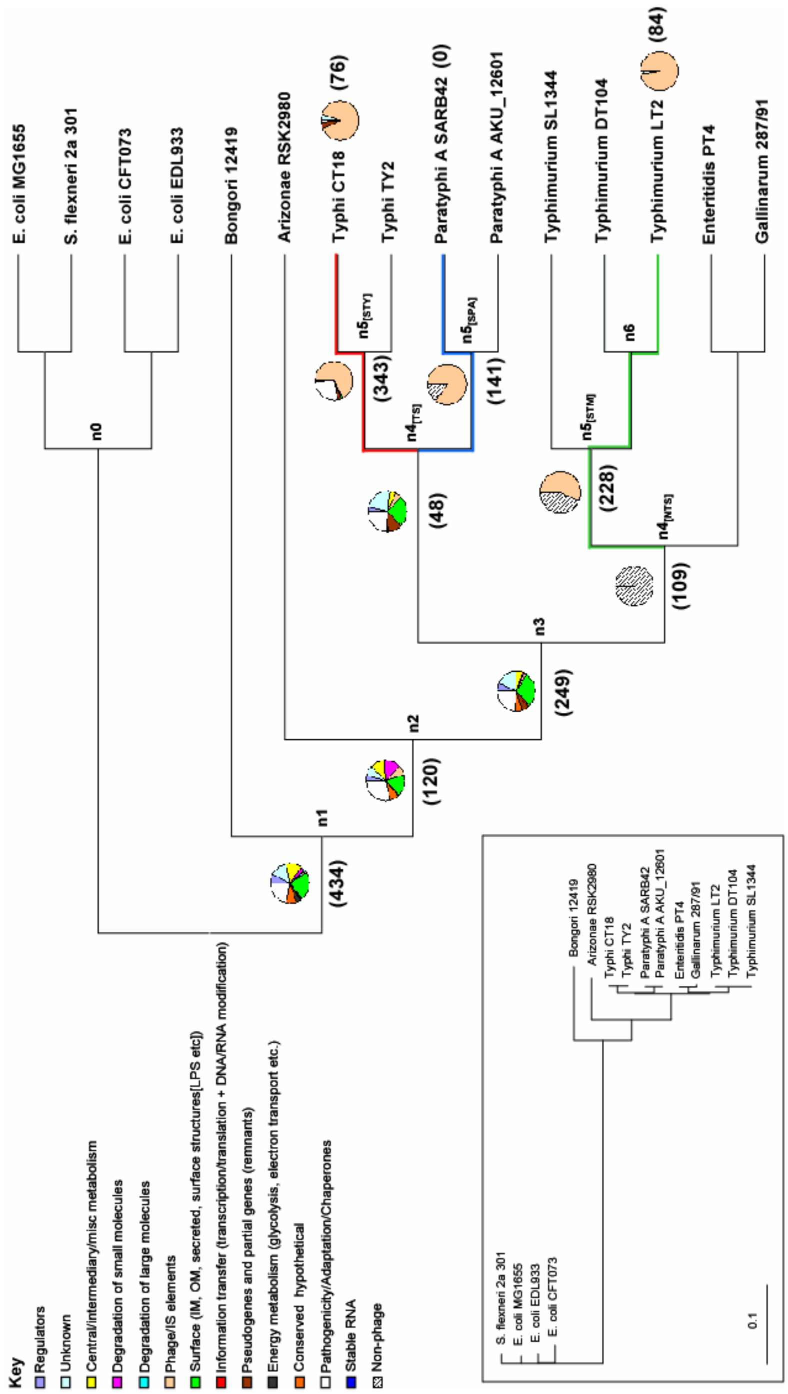


Figure 3.10: Numerical and functional distribution of PHA genes. The cladogram (main) shows the phylogenetic relationship between the 15 genomes used in this study, ignoring branch length. The topology of the tree is based on whole-genome sequence alignment. For the true phylogenetic distance with the respective branch lengths drawn to scale refer to the phylogram detailed in the inset of this figure; the phylogram is built using the ML method exploiting the F84 model. Numbers within parenthesis (main) reflect the number of PHA genes. Pie charts on each branch represent the functional classification of genes based on the colour-class detailed in the key. The non-phage functional class (black and white diagonal hatching) was introduced to classify CDSs without colour-coded functional classification in their annotation; those CDSs assigned into the "non-phage" pseudo-class represent CDSs that belong to any of the thirteen functional classes apart from the phage class. Numbers of genes on branches 1, 2 and 3 reflect the intersection of the respective number of genes determined on each branch using one of the three query genomes; the same applies for genes assigned to branch 4[TS].

Table 3.5: A list of PHA genes, and their inferred relative time of insertion.

S. typhi CT18		*S. paratyphi* A SARB42		*S. typhimurium* LT2	
Relative time of insertion	**PHA genes**	**Relative time of insertion**	**PHA genes**	**Relative time of insertion**	**PHA genes**
Branch 1	493	Branch 1	434	Branch 1	473
Branch 2	124	Branch 2	120	Branch 2	128
Branch 3	316	Branch 3	268	Branch 3	249
Branch 4 [TS]	62	Branch 4 [TS]	48	Branch 4 [NTS]	109
Branch 5 [STY]	343	Branch 5 [SPA]	141	Branch 5 [STM]	228
Branch CT18	76	Branch SARB42	0	Branch LT2	84
Total	1,414	Total	1,011	Total	1,271

This analysis revealed a surprisingly high number of 434 PHA genes inserted at the base of the *Salmonella* lineage (branch 1). Based on two independent previous studies (Doolittle *et al.*, 1996; Ochman and Wilson, 1987) the divergence of the *E. coli* and *Salmonella* lineage occurred approximately 100-140 Myr ago. Consequently putative HGT events on branch 1 represent ancient insertions, close to the divergence of these two lineages and include 76 coding sequences (CDSs) of “ancient” SPIs such as SPI-5, SPI-4, a part of SPI-2 (ttr-region), SPI-9, SPI-1 and a part of SPI-3 (magnesium transport ATPase – mgt region).

The *cob* operon of *S. enterica*, which encodes vitamin B12 biosynthesis, has been previously shown to be horizontally acquired in the *Salmonella* lineage following its divergence from the *E. coli* lineage (Lawrence and Roth, 1995; Lawrence and Roth, 1996). In a later study, Lawrence and Ochman (Lawrence and Ochman, 1997) showed, using a model of reverse amelioration, that the *cob* operon was probably introduced into the *Salmonella* lineage 71 Myr ago. The current analysis assigned the *cob* operon to branch 2 which predates the divergence of *S. arizonae* from the *S. enterica* lineage. Based on the data available, we can infer that the divergence of *S. arizonae* from the *S. enterica* lineage occurred approximately 100-71 Myr ago, and further suggest that the 120 inferred PHA genes assigned to branch 2 have an absolute time of insertion of the same order of magnitude.

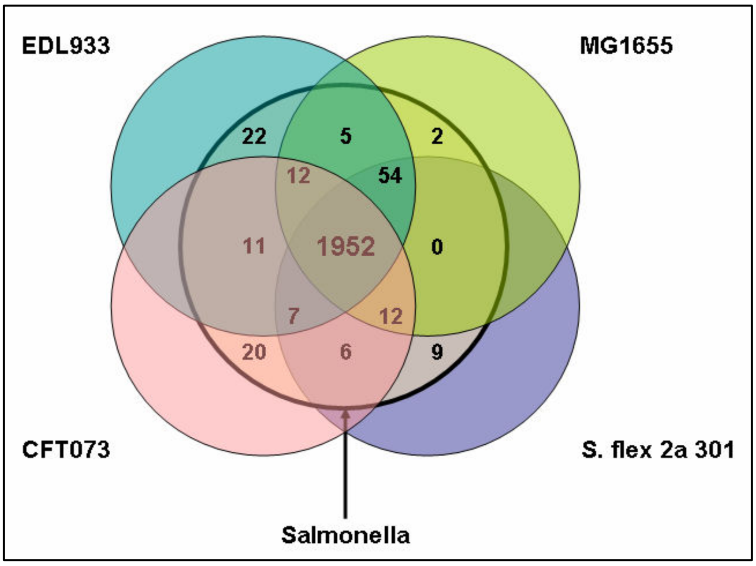


Figure 3.11: Venn diagram illustrating the orthologous genes shared between all the 11 *Salmonella* strains (bold circle in the middle) and the genomes of *E. coli* MG1655, *E. coli* EDL933, *E. coli* CFT073 and *S. flexneri* 2a 301. The number highlighted in bold, represents the total number of orthologous genes (core genes) shared between the 15 genomes used in this study.

On branch 3 (*S. enterica* lineage), there are 249 inferred PHA genes. On this branch are found SPIs that are restricted to the *S. enterica* lineage, such as part of SPI-3 (3' end), part of SPI-10 (fimbrial-sef operon), SPI-6, SPI-16 and SPI-17. Finally on more recent branches, i.e. branch $5_{[STY]}$ (STY: *S. typhi*), branch $5_{[SPA]}$ (SPA: *S. paratyphi* A), branch $5_{[STM]}$ (STM: *S. typhimurium*) and strain-specific genes, (relative to each of the three query genomes), I have inferred a significant number of putative HGT events which are mainly dominated by CDSs that belong to annotated prophage structures (discussed in more detail below).

3.3.2 Functional analysis of PHA genes

Implementing a classification of 14 functional classes, listed in Figure 3.10, each of the PHA genes, with a given relative time of insertion, was assigned into one of the 14 colour-coded functional classes. The results are summarized, via pie charts assigned to each branch, in Figure 3.10. Overall, from this functional classification, it is clear that PHA genes on branches 1-3, branch $4_{[TS]}$ (Typhoidal *Salmonella*) and branch $4_{[NTS]}$ (Non-Typhoidal *Salmonella*) show a wide distribution over almost all the 13 functional classes (e.g. cell-surface, regulation, central metabolism, pathogenicity), while gene-remnants/pseudogenes are mainly restricted to recently diverged lineages, i.e. the *S. enterica* species. Moreover CDSs that belong to annotated structures of prophages (light pink-coloured functional class) are predominant in very recent lineages (i.e. on branches $5_{[STY], [SPA], [STM]}$, or strain-specific CDSs).

On branch $4_{[TS]}$, which predates the Typhi-Paratyphi A divergence, overall 24% of PHA genes have unknown function, 26% encode cell surface-related components, 11% are remnants/pseudogenes and 24% are related to pathogenicity or adaptation. Also on this branch are the CDSs of a 8.5kb, previously uncharacterized, Genomic Island (GI) at position 2187521-2195992bp, of very low G+C (36.29%) content that encodes 16 CDSs (STY2349-STY2364 in CT18) of unknown function, without significant similarity with previously annotated CDSs. Furthermore, this novel GI does not have any of the "classical" GI-related features e.g. direct/inverted repeats, integrase gene or insertion adjacent to RNA locus. Details about the composition of this putative GI and other genes assigned to branch $4_{[TS]}$ will be discussed in the following section.

The functional analysis of the PHA genes assigned to recent branches (branches $5_{[STY], [SPA], [STM]}$ and strain-specific) is in line with a previous study focused on *E. coli* MG1655 showing that IS elements and prophage remnants represent mostly very recent insertion events in MG1655 (Lawrence and Ochman, 1998); the same study suggests that very few acquired DNA sequences are maintained for more than 10 Myr in the genome of *E. coli* MG1655. In the current study, there is no complete-intact prophage structure, inserted at the base of *Salmonella* lineage that is present in all the 11 *Salmonella* strains or even prophages inserted in the *S. enterica* lineage that are shared between the Typhi, Paratyphi A and the Typhimurium strains. Using Typhi CT18 as a query genome, on branch $5_{[STY]}$, 67% (231) of PHA CDSs belong to prophage structures, while 93% (71) of CT18-restricted PHA CDSs are of phage origin. Similarly, in the case of Typhimurium LT2, 57% and 98% of PHA genes that are on branch $5_{[STM]}$ and LT2-restricted, respectively, belong to annotated prophage structure. In the lineage of Paratyphi A, 85% of PHA CDSs acquired on branch $5_{[SPA]}$ are of phage origin; interestingly there are no SARB42-specific CDSs relative to Paratyphi A AKU_12601.

In a previous study, Thomson *et al.* (Thomson *et al.*, 2004) provided data showing that many prophage structures present in Typhi CT18 are predicted to be Typhi-specific, further suggesting that these bacteriophages have a level of specialization for their host and play a key role in generating genetic diversity in the *S. enterica* lineage. Moreover the same authors suggested that Typhi has indeed a unique pool of prophage elements that distinguish it from other serovars, in contrast with the *Salmonella* specific SPIs which show a wider distribution within the *Salmonella* lineage (Ochman and Groisman, 1996).

Generally in microbial genomes, some PHA genes are retained over long evolutionary distances and therefore contribute to species diversification (Lawrence, 1999; Lawrence, 2001), while PHA genes that might be detrimental, or not advantageous for the host are rapidly removed (Lawrence *et al.*, 2001; Lawrence and Ochman, 1998). Horizontally acquired DNA is more likely to be deleted than are native, core genes; for example, prophage structures often harbor direct repeats forming their endpoints i.e. phage attachment sites attL (left) and attR (right) that can, via homologous recombination, efficiently remove those “parasitic” elements. Furthermore, some prophage genes can be detrimental (e.g. the *N* gene of bacteriophage λ), neutral (e.g. integrases) or advantageous (e.g. immunity repressors) (Lawrence *et al.*, 2001). Based on this model, parasitic-detrimental DNA sequence (e.g. prophage elements) is removed by sequential deletion over time. This bias of deletion over insertion (Andersson and Andersson, 1999) can equilibrate HGT events, and this is further supported by the comparable genome size of closely related genomes (Bergthorsson and Ochman, 1998). Overall the current study suggests that indeed prophage structures are not retained for a long time in the *Salmonella* lineage, while complete, intact prophage structures represent very recent insertions in the Typhi, Paratyphi A and the Typhimurium lineage, which based on their impact (detrimental, neutral or advantageous) on the host, will eventually be retained or removed from those genomes.

3.3.3 Compositional analysis

The aim of the compositional analysis in this study was to determine if there is any clear trend for genes assigned to relatively old branches in the reference tree topology to show sequence composition closer (compared to more recent insertions) to the average composition of the host genome, thus supporting the effect of amelioration as a time-dependent process. It should be noted that because this analysis is focused on the effects of the amelioration in the *Salmonella* lineage which diverged fairly recently from *E. coli* and the rest of the enteric bacteria, we expect to identify, if any, mild effects of the amelioration on the sequence composition of the gene datasets under study. For example, Daubin and Ochman (Daubin and Ochman, 2004) applying a similar approach on a much broader phylogenetic sample (the γ-Proteobacteria), showed a strong correlation between the G+C content and different phylogenetic depths in their reference tree topology.

As a starting point for the compositional analysis of PHA genes, I applied the Alien_Hunter algorithm, which implements the Interpolated Variable Order Motifs (IVOMs) method (Vernikos and Parkhill, 2006), to the three query genomes, and performed a benchmarking analysis of its sensitivity versus the inferred relative time of insertion of PHA genes; the results are shown in Figure 3.12. Overall it can be concluded that the sensitivity of this HGT prediction method correlates strongly with the relative time of insertion. Indeed, in all the three query genomes regression analysis showed a correlation ($0.45 \leq R^2 \leq 0.74$) between the sensitivity and the relative time of insertion. For example, PHA genes inserted at the base of *Salmonella* lineage, e.g. on branch 1, can be identified with a False Negative (FN) rate of 0.55 while more recent insertions with a much lower FN rate of 0-0.2. It is worth noting that the high sensitivity of Alien_Hunter on very recent branches is in contrast with the drop in the IVOMs score distribution (Figure 3.13); the majority of the PHA genes assigned to these branches belong to prophage structures, consequently their clustering and not their composition should mainly explain the high sensitivity of this algorithm on these branches. It is important to note that the analysis of the sensitivity of this algorithm relies on the assumption that all the PHA genes identified in the current analysis are true horizontally acquired genes and the conclusions drawn about its performance are specific for this set of PHA genes.

Calculating the G+C content, both overall and codon position specific, as well as higher order compositional biases, implementing the IVOMs method, the amelioration process versus the relative time of insertion of PHA genes was monitored (Figure 3.13, Figure 3.14). Using Typhi CT18 and Paratyphi A SARB42 as query genomes this analysis revealed that there is a clear correlation (R^2 = 0.98 for branches 1-3, R^2 = 0.65 for branches 1-$4_{[TS]}$) between the G+C content or the IVOMs score of PHA genes and the relative time of their insertion on the earlier branches; however, this strong correlation seems to "break down" in the case of very recent putative HGT events, i.e. insertions that took place after the divergence of Typhi and Paratyphi A lineages (Figure 3.13, Figure 3.14).

For example genes assigned to branches 1 and 2 show an average G+C content of 51.4% and 50.6% respectively, close to the average gene G+C content of 53.2% and 53.3% (CT18 and SARB42 respectively). The same observation becomes much clearer when calculating higher order compositional biases (Figure 3.13). Based on the IVOMs score, genes on branches 1 and 2 have an average score of 0.06 and 0.063 respectively while more recently acquired genes, i.e. on branches 3 and 4[TS] have a score of 0.072 and 0.093 respectively; the average, genome-wide IVOMs score in Typhi CT18 is 0.059.

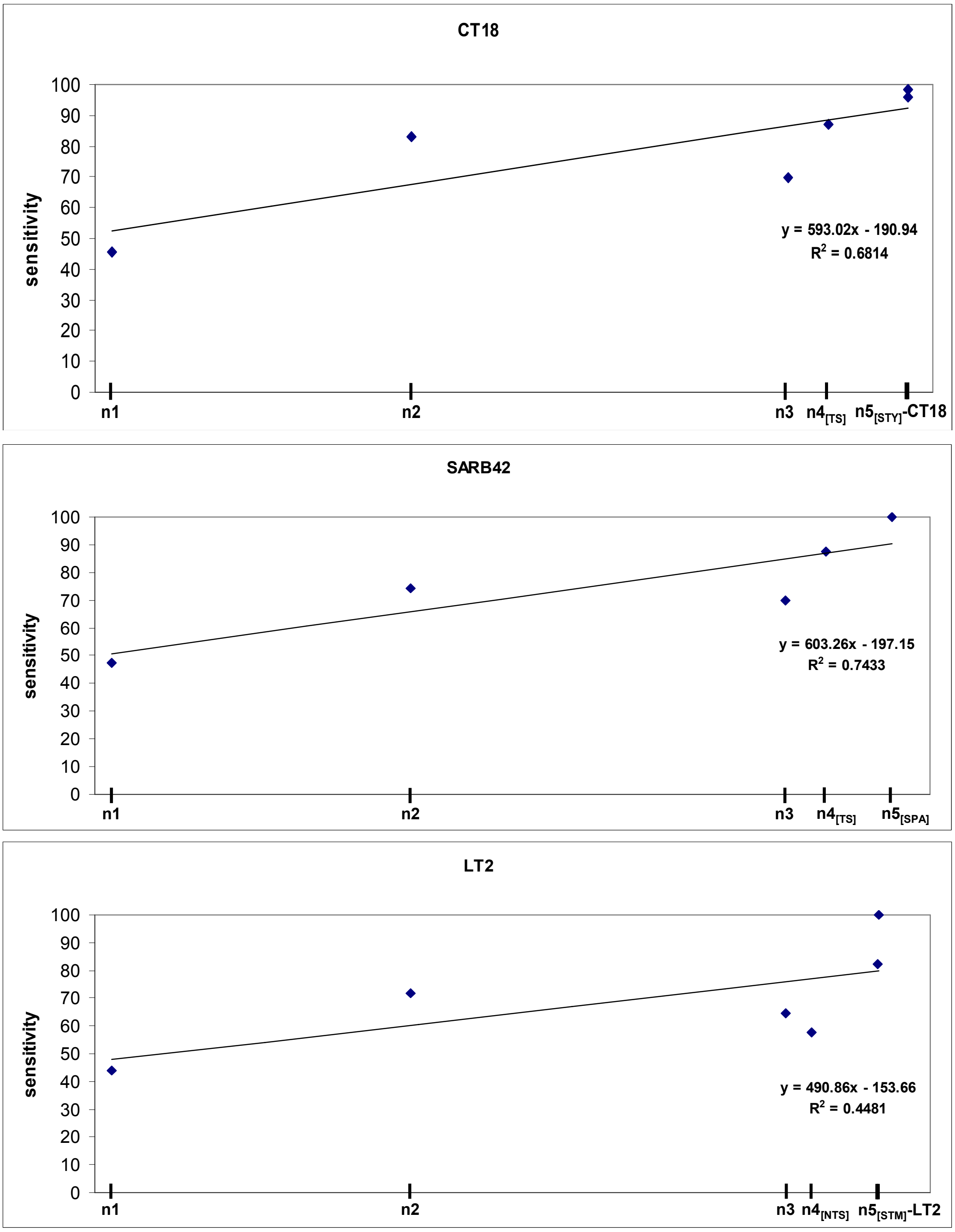


Figure 3.12: Sensitivity of the Alien_Hunter algorithm, which implements the IVOMs method, versus the inferred relative time of insertion of PHA genes for the three query genomes: *S. typhi* CT18 (top), *S. paratyphi* A SARB42 (middle), *S. typhimurium* LT2 (bottom). The nodes on the X axis are scaled according to the respective branch lengths of the tree topology shown in the inset of Figure 3.10. Regression analysis is provided embedded within the three graphs; *p*-values: 0.04, 0.05 and 0.14 respectively.

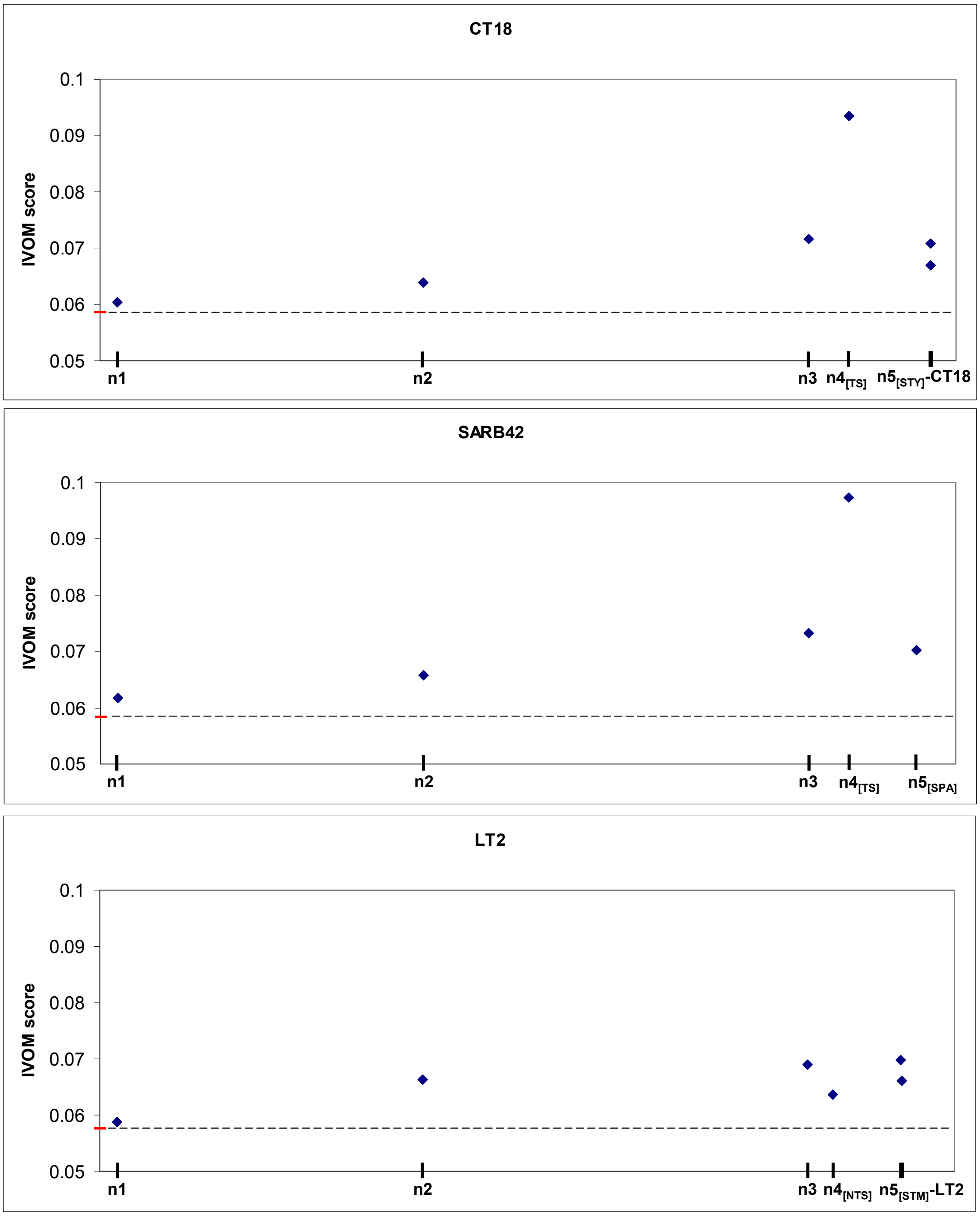


Figure 3.13: Average score, taking into account higher order compositional biases, of putative horizontally acquired genes, versus the inferred relative time of insertion, in the three query genomes: *S. typhi* CT18 (top), *S. paratyphi* A SARB42 (middle), *S. typhimurium* LT2 (bottom). The score is calculated implementing the IVOMs method. The average score for the three query genomes is highlighted in red (the embedded dashed line is provided for ease of comparison). The nodes on the X axis are scaled according to the respective branch lengths of the tree topology shown in the inset of Figure 3.10.

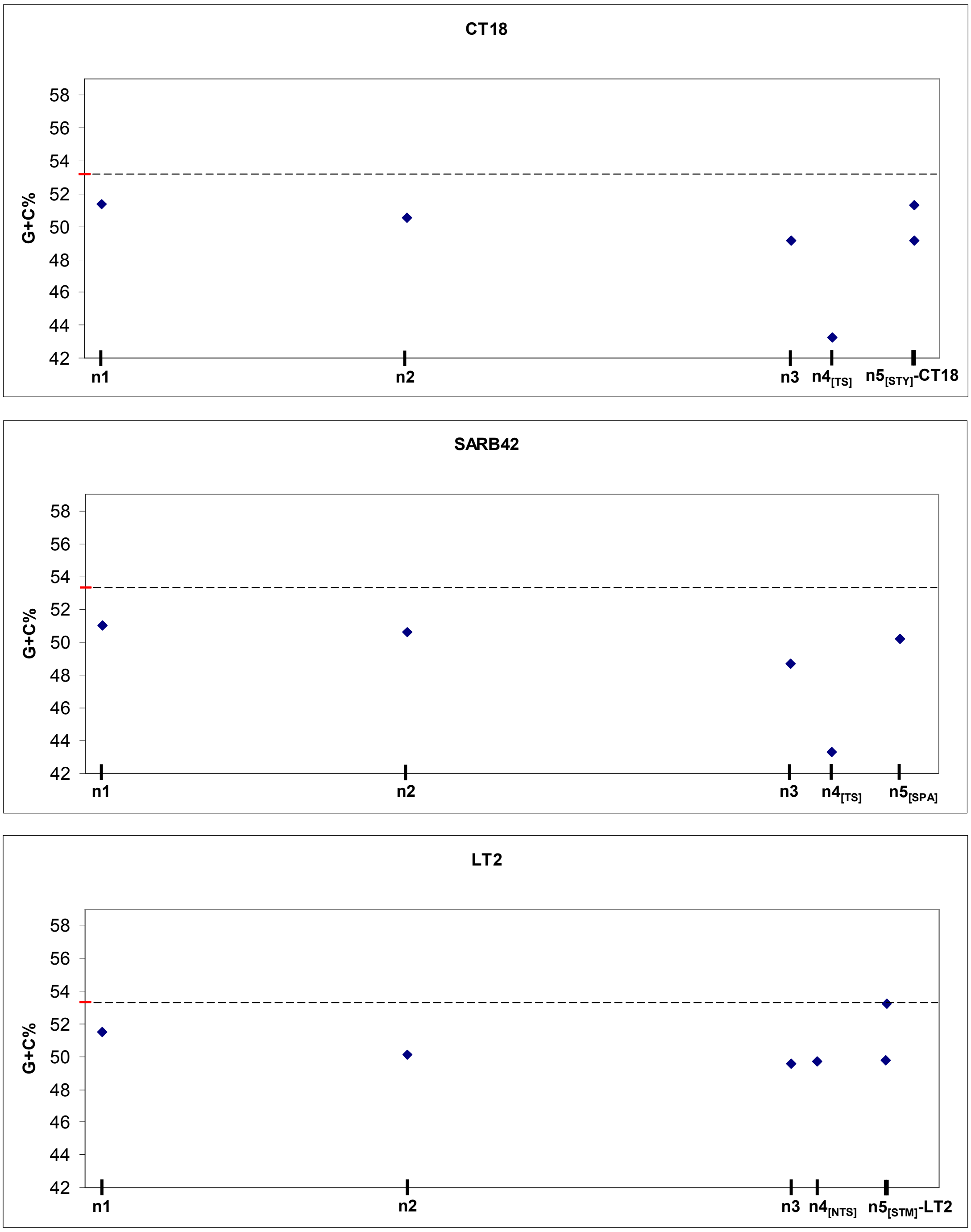


Figure 3.14 Average G+C content of putative horizontally acquired genes, versus the inferred relative time of insertion, in the three query genomes: *S. typhi* CT18 (top), *S. paratyphi* A SARB42 (middle), *S. typhimurium* LT2 (bottom).The average G+C content for the three query genomes is highlighted in red (the embedded dashed line is provided for ease of comparison). Error bars could not be visualized (the standard deviation is in the range of 0.05-0.08). The nodes on the X axis are scaled according to the respective branch lengths of the tree topology shown in the inset of Figure 3.10.

A similar observation can be made for Typhimurium LT2. More specifically, there is a very strong correlation ($R^2 = 0.89$) between G+C content or IVOMs score and the relative time of insertion which breaks-down on branches descendent of node 3 (Figure 3.13, Figure 3.14). More specifically, the average G+C content of genes assigned to branches 1, 2 and 3, is 51.5%, 50% and 49.6% respectively while for genes on the branch $4_{[NTS]}$, the average G+C content is 49.7%. Similarly, using the IVOMs method, the corresponding scores for the four branches are: 0.059, 0.066, 0.069 and 0.064 respectively.

PHA genes assigned to branch $4_{[TS]}$ on the Typhi-Paratyphi A lineage show a very strong compositional deviation, indicated both by their very low G+C content of 43.3% (gene average: 53.2%) and the IVOMs score of 0.093 (genome average: 0.059). Furthermore, the codon-position specific G+C content of genes assigned to branch $4_{[TS]}$, deviates strongly (GC_1 = 49%, GC_2 = 37%, GC_3 = 43%) (Figure 3.15) from the expected values (GC_1 = 59%, GC_2 = 41%, GC_3 = 56%, respectively) based on the three linear equations (13, 14, 15) provided by Lawrence and Ochman (Lawrence and Ochman, 1997). Those three linear equations are based on the observation that the G+C% content at the three codon positions shows a linear positive correlation with the genomic (i.e. genome average) G+C% content, although with different rates of correlation (Muto and Osawa, 1987), Figure 3.15:

$$GC_1 = 0.615 \times GC_{Genome} + 26.9$$

$$GC_2 = 0.270 \times GC_{Genome} + 26.7$$

$$GC_3 = 1.692 \times GC_{Genome} + 32.3$$

Source: (Lawrence and Ochman, 1997).

Therefore, the above linear equations can be used to infer the level of departure from those expected values of codon-position specific G+C% content and evaluate the level of amelioration of PHA genes; the codon-position specific G+C% content of PHA genes that have been recently horizontally acquired, follow the expected GC_1, GC_2 and GC_3 values based on the GC_{Genome} of their donor; on other hand PHA genes that have been acquired a long time ago, follow the expected GC_1, GC_2 and GC_3 values based on the GC_{Genome} of the their "new" host; PHA genes still undergoing the amelioration process are expected to fall somewhere in between.

The G+C content of the second codon position is generally very constrained to similar values across species (Lawrence and Ochman, 1997). Interestingly, genes assigned to branch $4_{[TS]}$ in Typhi-Paratyphi A lineage show a significant deviation also in this compositionally well-conserved codon position possibly suggesting a distantly related donor genome (Figure 3.15).

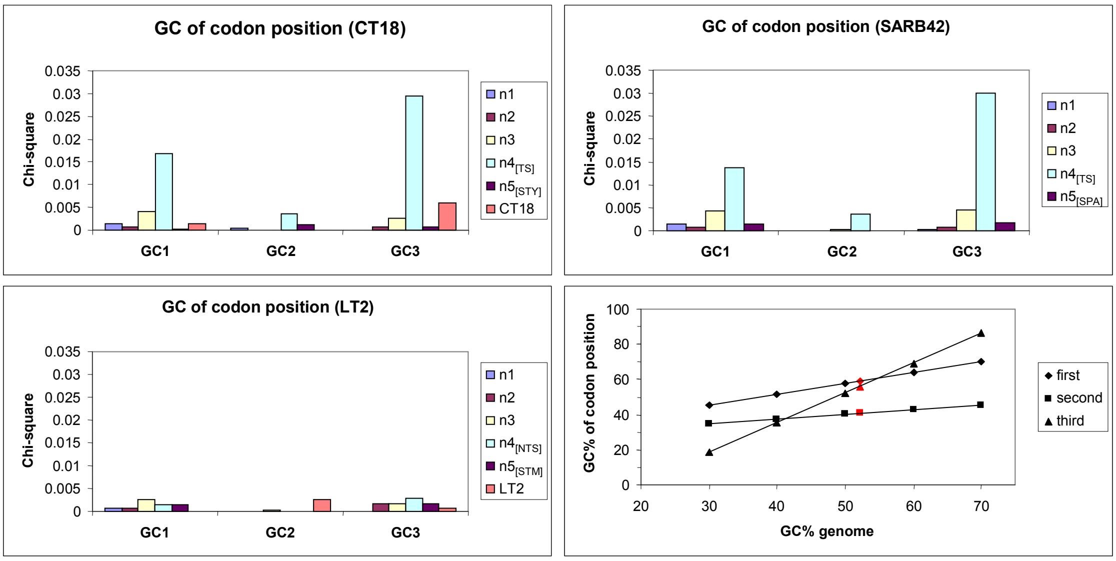


Figure 3.15 Chi-square values of G+C content over the three codon positions, for genes assigned to lineages of increasing depth in the reference tree topology. Chi-square values are calculated using the expected G+C codon-position values derived from the three linear equations (13, 14, 15) provided by Lawrence and Ochman (Lawrence and Ochman, 1997). At the right-bottom side of the figure, the correlation between genomic G+C content and G+C content at the three codon positions based on the data provided by Muto and Osawa (Muto and Osawa, 1987), is provided. Genes that are still under the amelioration process are expected to deviate from those expected values. The expected G+C content for each codon position in the *Salmonella* lineage is highlighted in red.

Codon usage analysis revealed that genes on branch $4_{[TS]}$ show a bias towards A+T rich codons (Figure 3.16). For example, the 'AAA' codon is overrepresented in CDSs of this branch, compared to its average frequency in the genome; the AAA codon (encoding lysine) has been previously shown to be overrepresented in highly expressed genes (Sharp and Li, 1987). To test further whether genes on this branch deviate compositionally due to their highly expressed pattern, rather than their alien origin, I performed a CAI analysis (summarized in Table 3.6). It can be clearly seen that genes on branch $4_{[TS]}$ deviate compositionally from the genome background composition, more likely due to their alien origin, rather than their high rate of expression, representing the "left ear" in the "rabbit-like" codon bias vs CAI plot described in (Karlin *et al.*, 1998).

Indeed, genes on branch $4_{[TS]}$ show an average CAI value of 0.221, significantly lower (*p*-value = 4.95 10^{-13}) than the average gene CAI value (= 0.31) and much lower than the CAI values of highly expressed genes, e.g. ribosomal protein coding (CT18: 0.554, SARB42: 0.560, LT2: 0.561) and aminoacyl-tRNA synthetase genes (CT18: 0.437, SARB42: 0.453, LT2: 0.434). Furthermore, the CAI analysis revealed that genes inferred in this study of being PHA do not show CAI values of highly expressed genes, and overall their CAI values are significantly lower (*p*-value = 3.75 10^{-74}) than the average gene CAI values.

Overall, using any of the three query genomes (CT18, SARB42, LT2) this analysis indicates that very recent acquisitions e.g. on branches $5_{[STY], [SPA], [STM]}$ seem to have been equally "ameliorated" with acquisitions on older branches e.g. branches 1, 2; moreover, in the case of LT2 genome, strain specific acquisitions (see LT2 branch) show sequence composition very close to the genome composition. Very recent acquisitions are expected to deviate strongly from the host backbone composition, unless the donor is very close compositionally to the host. Amelioration, a time-dependent process, can not have significantly affected their sequence composition, which should still reflect mostly the donor rather than the host specific compositional signature. However, recent acquisitions identified in this study either show very close composition to the host backbone composition, for example PHA genes on LT2 branch have an average G+C content of 53.26% very close to the gene average G+C of

53.33%, or deviate compositionally equally to PHA genes acquired on older branches; for example the G+C content of PHA genes in CT18 on branches 1 and $5_{[STY]}$ is 51.4 and 51.3 respectively. Similarly the G+C content of PHA genes in SARB42 on branches 2 and $5_{[SPA]}$ is 50.6% and 50.2% respectively.

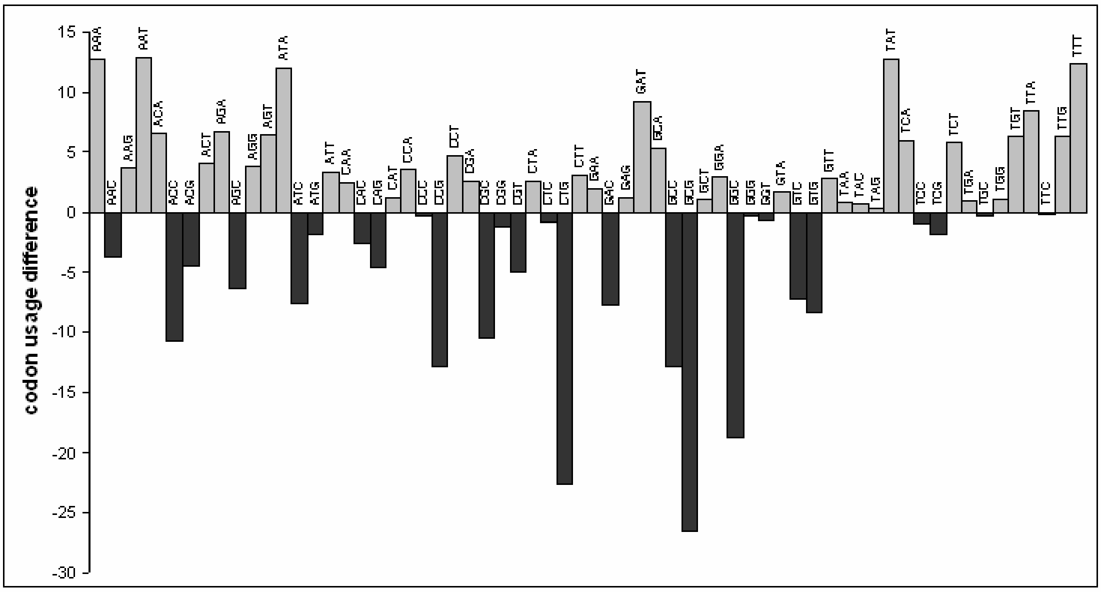


Figure 3.16: Codon usage difference of CDSs assigned on branch $4_{[TS]}$ relative to the average codon usage in Typhi CT18. Positive values in the Y axis indicate overrepresentation (light grey bars) of certain codons in CDSs of this branch relative to the average codon usage and vice versa.

Interestingly, branches descendant of nodes $4_{[TS]}$ and $4_{[NTS]}$ are dominated by genes of phage origin (57-98% of genes at the given relative time of insertion), Figure 3.10. For example on branch $5_{[STY]}$, 67% of Typhi CT18 genes assigned to this branch belong to one of the six prophage structures present both in Typhi CT18 and TY2. On branch $5_{[STY]}$, SPI-7 and the phage-related gene G+C content is 50.87% and 51.98% respectively. In a previous study, it has been shown that the last common ancestor of Typhi existed 15,000-150,000 years ago, during the human hunter-gatherer period (Kidgell *et al.*, 2002); consequently PHA genes assigned to branch $5_{[STY]}$, have a time of insertion of the same order of magnitude. Similarly, in Typhimurium LT2, there are two prophage (Fels-1, Fels-2) structures that represent very recent acquisitions (LT2-specific), and are absent from the other two Typhimurium strains. CDSs of these prophage elements have an average G+C content of 53.57% and 52.94% respectively, while their CAI value is 0.307, very close to the LT2 genome average CAI of 0.313.

Table 3.6: Average CAI values for genes of different inferred relative time of insertion for the three query genomes. Average CAI values for all genes in the genome, ribosomal protein coding and aminoacyl-tRNA synthetase genes, are also provided as a reference. Genes ≤ 300bp were excluded. The reference gene set of highly expressed genes was the one proposed by Sharp and Li (Sharp and Li, 1986) using the genome of *E. coli*.

S. typhi CT18		*S. paratyphi* A SARB42		*S. typhimurium* LT2	
Genes	**CAI**	**Genes**	**CAI**	**Genes**	**CAI**
PHA on branch 1	0.264	PHA on branch 1	0.264	PHA on branch 1	0.264
PHA on branch 2	0.258	PHA on branch 2	0.258	PHA on branch 2	0.258
PHA on branch 3	0.256	PHA on branch 3	0.256	PHA on branch 3	0.256
PHA on branch 4 [TS]	0.221	PHA on branch 4 [TS]	0.221	PHA on branch 4 [NTS]	0.275
PHA on branch 5 [STY]	0.283	PHA on branch 5 [SPA]	0.297	PHA on branch 5 [STM]	0.269
PHA on branch CT18	0.282	PHA on branch SARB42	NA	PHA on branch LT2	0.307
All genes	0.310	All genes	0.315	All genes	0.313
Ribosomal	0.554	Ribosomal	0.560	Ribosomal	0.561
tRNA synthetase	0.437	tRNA synthetase	0.453	tRNA synthetase	0.434

3.4 Discussion

The aim of this analysis was to study the distribution of PHA genes in a time-dependent manner i.e. to infer the relative time of insertion based on the reference tree topology, throughout the *Salmonella* lineage, applying an extensive comparative analysis between 11 *Salmonella*, three *E. coli* and one *Shigella* strain. The selection of four genome sequences that form an outgroup of the *Salmonella* lineage was made in order to differentiate gene loss from gene gain more reliably, two mechanisms that could explain the presence of a gene in one lineage and its absence from a sister, closely related lineage.

However, because the *E. coli* and *Salmonella* lineage represent very closely related, sister lineages the 434 PHA genes inferred to have been acquired at the base of the of the *Salmonella* lineage might equally represent deletion events in the *E. coli* lineage subsequent to the common ancestor with *Salmonella*. To investigate further this alternative scenario, I used a set of three more distantly related enteric outgroup genomes: *Erwinia carotovora* SCRI1043 (accession number: BX950851), *Yersinia enterocolitica* 8081 (accession number: AM286415) and *Y. pseudotuberculosis* IP32953 (accession number: BX936398). Less than 5% of the 434 PHA genes inferred to have been acquired on branch 1 have orthologous genes present in this distant outgroup. These data suggest that the majority (>95%) of 434 PHA genes most likely represent true HGT events that occurred quite early in the evolution of the *Salmonella* lineage, rather than deletion events in the *E. coli* lineage.

In the current study I exploited a much larger sequence sample, i.e. whole genome sequence, rather than selected gene/protein sequences to serve as "molecular chronometers", thus the phylogenetic signature seems to be strong enough for the NJ and ML method to result in identical tree topologies, inferring the same phylogenetic history of the query genomes at hand. However, care should be taken when interpreting whole-genome sequence based phylogenies, since extensive HGT events, homologous recombination or other homoplastic events might well obscure the true phylogenetic history of the genomes under study (Doolittle and Papke, 2006) whose phylogeny may therefore be more efficiently described using phylogenetic nets rather than single tree topologies (Doolittle, 1999; Hilario and Gogarten, 1993; Martin, 1999).

For example, Didelot *et al.* (Didelot *et al.*, 2007) showed that Paratyphi A and Typhi genomes are, over 75% of their sequence, distantly related *S. enterica* members (both in terms of nucleotide divergence and gene content), while the remaining 25% of their sequence is much more similar (average nucleotide divergence 0.18%, instead of 1.2%); the authors suggested that the two genomes have recently exchanged, via homologous recombination, a significant amount of DNA, now representing seemingly similar lineages (convergence evolution). In the current analysis, the two lineages, have been mapped on very close branches of the *Salmonella* phylogenetic tree (Figure 3.10), representing perhaps a limitation of the applied, strictly bifurcating tree topology; using instead a reticulate phylogenetic network, would have probably (correctly) mapped the two seemingly similar lineages on more distant branches, taking also into account (by means of multi-furcating branches connecting the two lineages) the extensive amount of exchanged homologous DNA.

It is worth noting that whole-genome based phylogenetic approaches are capturing the "overall" phylogenetic signal based on whole chromosome sequences. In the case of very closely related organisms, e.g. strains of the same serovar, minor differences in terms of gene content (e.g. prophages, GIs) cannot be reliably represented in the "overall" phylogenetic signal. In other words, whole genome-based phylogenies focusing on a wide range of strains may suffer from low resolution in the case of very closely related genomes. Moreover mobile elements may show similarity on the sequence level (e.g. prophages) but differ on the structural level (i.e. different phage types). Relying on sequence information only, these seemingly similar mobile elements will bias the relatedness of closely related strains (e.g. the three Typhimurium strains used in this study).

The reason why I pursued a comparative, rather than a compositional based approach (i.e. defining PHA genes based simply on their compositional deviation, but ignoring their distribution throughout the lineage of interest), was the fact that compositional based approaches frequently underestimate the true number of HGT events (Lawrence and Ochman, 1997), either due to the amelioration process, in the case of ancient insertions, or due to compositionally similar donor genomes, in the case of new insertions. The current comparative analysis suggests that approximately 30, 25 and 28% of protein-coding sequences in Typhi CT18, Paratyphi A SARB42 and Typhimurium LT2 respectively, represent putative HGT events. The distribution of those PHA genes on different branches of the reference tree topology reveals that approximately 35-40%

of them were acquired at the base of the *Salmonella* lineage (branch 1), very close to its divergence from *E. coli*, reflecting perhaps the acquisition of genes that enabled the exploration of new niches e.g. the acquisition of SPI-1 which enabled *Salmonella* to invade epithelial cells (Galan, 1996). Moreover, 20% of those genes were acquired at the base of the *S. enterica* lineage (branch 3); overall 60-70% were inserted after the divergence of the *Salmonella* from the *E. coli* lineage and prior to the divergence of the *S. enterica* subspecies. This suggests that approximately 60-70% of the putative HGT events are probably shared between most of the subspecies of the *S. enterica* lineage.

Based on the functional classification of genes assigned to branches 1, 2 and 3 that predate the *S. enterica* lineage, it becomes evident that generally, genes within almost all functional classes, e.g. regulation, energy metabolism, cell surface, virulence-related, have been horizontally acquired. Moreover the functional distribution of genes assigned to branches 1 and 3 correlates strongly with the functional distribution of all the genes in the genome (R: 0.71 and 0.63 – p-value: 0.01 and 0.03 respectively) whereas this correlation is weaker (and not significant) for genes on branches 2 and 4 (R: 0.54 and 0.45 – p-value: 0.07 and 0.1 respectively) and disappears (R<0.01, p-value: 0.98) completely in the case of very recent branches (branch 5 or genome-specific).

On branches 1-4 there is a fairly constant percentage of genes encoding cell-surface structures (18-28%), genes related to pathogenicity and adaptation (22-29%) and regulatory elements (4-8%). Furthermore, the percentage of genes with unknown function ranges from 8-18%, while fragmented gene-remnants (pseudogenes) account for 6% and 11% on branches 3 and 4[TS] respectively with almost no pseudogenes (< 0.1%) on branches 1 and 2. The increased number of genes acquired at the base of *S. enterica* lineage that have been inactivated suggests that some of these early-acquired functions are no longer necessary, and are being lost in these serovars. The increased number of pseudogenes (11%) in the Typhi-Paratyphi A lineage that are absent from the Typhimurium lineage supports a genome degradation process via pseudogene formation suggested to be due to the recent change in niche of these serovars (Parkhill *et al.*, 2001).

The compositional analysis of the inferred PHA genes indicates that there is indeed a strong correlation between the time of insertion and amelioration towards the host-specific genomic signature. In other words, anciently horizontally acquired genes have ameliorated more towards the host composition, compared to more recent acquisitions. However, even HGT events inferred to have inserted at the base of the *Salmonella* lineage still preserve some of their donor genome sequence signature, as indicated by their overall and codon-position specific G+C content, suggesting that these genes are still undergoing the amelioration process.

On the other hand, in the case of very recent acquisitions that represent mostly insertion of prophage elements, it seems that their sequence composition is already much closer to the host background composition, presumably not due to the amelioration process, since they have been acquired fairly recently, but rather due to an adaptation to the specific sequence signature of the their host. Perhaps the compositional adaptation of those prophages is pivotal for the masking of their alien sequence identity in order to successfully integrate into the bacterial chromosome, without being detected by the histone-like nucleoid structuring (H-NS) protein that selectively silences horizontally acquired DNA of lower G+C content than the host genome (Navarre *et al.*, 2006).

If we take into account both the absence of complete-intact prophage structures from old branches (1-3, $4_{[TS]}$ and $4_{[NTS]}$), and the significant compositional similarity of those prophage-related genes to the host sequence composition, when the effects of the amelioration process are expected to be mild, it would be tempting to speculate that prophage elements in the *Salmonella* lineage have undergone an adaptation to specific serotypes. However this hypothesis does not explain why anciently inserted prophages e.g. those inserted at the base of *Salmonella* lineage, prior to the divergence of *S. bongori* and *S. arizonae* from the *S. enterica*, have not been retained in descendent lineages e.g. the Typhi, Paratyphi A and Typhimurium strains.

Perhaps anciently inserted bacteriophages at the base of the *Salmonella* lineage carried genes that were either neutral or detrimental, providing no profound advantage to the host, and over time the host has lost those parasitic elements via a deletion process which has left behind molecular fossils of those elements. This observation is further supported by the absence of pseudogenes on very old branches, i.e. branches 1 and 2; perhaps the ongoing time-dependent process of deleting redundant or detrimental DNA sequence has already removed a much higher proportion of pseudogenes on very old branches, compared to recent ones further suggesting that genome degradation is still a continuous process in the *Salmonella* lineage (Lawrence *et al.*, 2001).

3.5 Conclusions

Overall the current analysis has shown that the impact of amelioration, a time-dependent process, is still detectable even in fairly recent HGT events, e.g. that occurred 100-140Myr ago; moreover it sheds more light on the relative time of insertion of HGT events in the *Salmonella* lineage, and presents data that show that prophage structures are not retained for long periods in the *Salmonella* lineage.

Whether this last observation is related to an ongoing genome degradation process that over time removes redundant or detrimental DNA sequences, equilibrating the horizontal influx of genes, maintaining a fairly constant genome sequence size, still remains to be clarified. Perhaps the study of the very recently acquired prophage elements which seem to account for the majority of the strain or serovar specific genes (McClelland *et al.*, 2004; Thomson *et al.*, 2004), and their impact (detrimental, neutral, advantageous) on the evolution, life-style and host adaptation of the *Salmonella* strains, might shed more light on the underlying principles of the observed genome degradation process.

The prophage elements present in the *Salmonella* lineage show a very close sequence composition to the host-specific background composition, strongly suggesting that those parasitic elements have specialized and adapted to their hosts, playing a key role in driving bacterial evolution (Thomson *et al.*, 2004), or even speciation itself supporting the notion of “evolution in quantum leaps”, introduced by Groisman and Ochman (Groisman and Ochman, 1996). Overall, the distribution of PHA genes in the *Salmonella* lineage coincides strongly with the divergence of the major *Salmonella* species, underlining the major impact of horizontal transfer in the evolution of the salmonellae.

Chapter 4

Resolving the structure of Genomic Islands

4.1 Introduction

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, which has lead to a definition of the GI structure that includes these features (Table 4.1), (Hacker *et al.*, 1997; Hacker and Kaper, 2000; Schmidt and Hensel, 2004).

GIs present in Gram-positive bacteria may 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).

Several web-based suites exploit the GI structural definition (Table 4.1) with the aim of implementing and automating the *in silico* prediction of genomic regions that share some or all of the GI-related signatures; those regions are subsequently annotated as novel GIs. For example Islander (Mantri and Williams, 2004) and IslandPath (Hsiao *et al.*, 2003), two web-based suites, combine and overlap several GI-related features trying to predict genomic regions as close as possible to the GI structural definition.

Although a large number of mobile elements fall well within the GI definition, there are several concerns about the structural consensus of GIs: Firstly, the current definition of the GI structure was put forward 11 years ago (Hacker *et al.*, 1997) when only 12 complete bacterial genomes were available; in May 2007 there were 558 complete published genomes and 1144 ongoing, enabling a more realistic sampling of the GI structural space for any potential structural variation to be captured. Secondly, there are a large number of GIs that deviate strongly from the GI definition (Table 4.2). Thirdly, *in silico* prediction methods that assume a full or partial structure similar to the GI structural definition, or search for GIs with some level of similarity to already known GI structures, bias the sampling of the GI structural space towards "well-structured" GIs.

Table 4.1: Common features of Genomic Islands.

Large inserts of horizontally acquired DNA (10 to 200kb)
Sequence composition different from the core backbone composition
Insertion usually adjacent to RNA genes
Often flanked by direct repeats or insertion sequence (IS) elements
Limited phylogenetic distribution i.e. present in some genomes but absent from closely related ones
Often mosaic structures of several individual acquisitions
Genetic instability
Presence of mobility genes (e.g. integrase, transposase)

A fundamental property of GIs, independent of any *a priori* structural definition, is their origin: GIs are horizontally acquired mobile elements of limited phylogenetic distribution. Based on this concept, a search of the GI structural space is feasible in a hypothesis-free framework without the need to make any *a priori* assumptions about the GI structure which rely on previously seen examples of GIs.

The aim of this analysis is to study the structural variation of GIs and revisit the GI definition, taking into account only the fundamental property of GIs i.e. their horizontal origin. Instead of exploiting a top-down approach searching for GIs that follow the GI structural definition, I reverse this framework by pursuing a hypothesis-free, bottom-up search (Vernikos and Parkhill, 2008); in a first step GIs are defined as genomic regions with limited phylogenetic distribution consistent with recent acquisition (as identified by maximum parsimony), and in a second step those regions are structurally annotated. In a third step, the structural features sampled from this hypothesis-free search are exploited in a machine learning approach with the aim of explicitly quantifying and modelling their contribution to the GI structural definition.

A similar approach of a hypothesis-free identification of GIs, defined as genomic regions with limited phylogenetic distribution, was applied in eight *Streptococcus agalactiae* strains (Tettelin *et al.*, 2005). Gene loss and gene gain are two distinct mechanisms that can both lead to limited phylogenetic distribution of a DNA sequence. However, Tettelin *et al.* did not apply any restriction (e.g. maximum parsimony) in order to differentiate gene gain from gene loss and defined as putative GIs any region (>5kb) that was absent from at least one of the eight reference genomes.

In the current study I focus on three different bacterial genera i.e. *Salmonella*, *Staphylococcus* and *Streptococcus* for four major reasons: there are enough (>10) sequenced genomes for each genus, this collection of strains covers both Gram-negative and Gram-positive groups and has both commensal and pathogenic representatives, and HGT plays a key role in the evolution of those three lineages (Broker and Spellerberg, 2004; Lawrence and Ochman, 1997; Novick and Subedi, 2007; Rosini *et al.*, 2006; Tettelin *et al.*, 2005; Towers *et al.*, 2004; Vernikos *et al.*, 2007; Waterhouse and Russell, 2006).

Table 4.2: A selection of annotated Genomic Islands that show structural variation. Features of GIs that deviate from the GI structural definition (Table 4.1) are highlighted in grey. For the G+C% deviation (GC – GC_{mean}), GIs that deviate less than 1% from the average G+C% content are highlighted as compositionally non-deviating regions. The representation of the repeats, integrase and RNA features is binary: "1" if present, "0" if absent.

Coordinates	Host	GI	Size	G+C% deviation	Repeats	Integrase	RNA	Gram
839352..853808	*S. aureus* MW2	vSa3	14457	-4.49	1	1	1	+
1891660..1923796	*S. aureus* MW2	vSaß	32137	-4.24	0	0	1	+
1932974..1959426	*S. aureus* Mu50	vSaß	26453	-4.16	0	1	1	+
2133112..2148791	*S. aureus* Mu50	vSa4	15680	-2.56	1	1	0	+
2251120..2266138	*S. epidermidis* RP62A	vSe1	15019	-1.43	1	0	0	+
1519667..1558081	*S. epidermidis* ATCC15305	vSe2	38415	-6.4	1	1	1	+
1012154..1023023	*S. haemolyticus* JCSC1435	vSh1	10870	-2.87	1	1	0	+
2117669..2133994	*S. haemolyticus* JCSC1435	vSh2	16326	-4.06	1	1	1	+
2578642..2593348	*S. haemolyticus* JCSC1435	vSh3	14707	-1.74	0	1	0	+
385739..432833	*S. agalactiae* NEM316	PAI3	47095	1.64	1	0	0	+
711791..759003	*S. agalactiae* NEM316	PAI7	47213	1.62	1	0	0	+
1013026..1060093	*S. agalactiae* NEM316	PAI8	47068	1.66	0	0	0	+
1163554..1197443	*S. agalactiae* NEM316	PAI10	33890	2.04	0	0	1	+
1255736..1261279	*S. agalactiae* NEM316	PAI11	5544	-6.37	1	1	1	+
302172..361067	*S. typhi* CT18	SPI-6	58896	-0.57	0	0	1	–
605515..609992	*S. typhi* CT18	SPI-16	4478	-9.98	1	1	1	–
1085156..1092735	*S. typhi* CT18	SPI-5	7580	-8.52	0	1	1	–
1625084..1664823	*S. typhi* CT18	SPI-2	39740	-4.91	0	0	1	–
2460780..2465939	*S. typhi* CT18	SPI-17	5122	-13.39	0	0	1	–
2742876..2759156	*S. typhi* CT18	SPI-9	16281	4.62	0	0	1	–
2859262..2899034	*S. typhi* CT18	SPI-1	39773	-6.22	0	0	0	–
3053654..3060017	*S. typhi* CT18	SPI-15	6364	-3.01	1	1	1	–
3132606..3139414	*S. typhi* CT18	SPI-8	6809	-14.03	1	1	1	–
3883111..3900458	*S. typhi* CT18	SPI-3	17348	-5	0	0	1	–
4321943..4346614	*S. typhi* CT18	SPI-4	24672	-7.74	0	0	0	–
4409511..4543072	*S. typhi* CT18	SPI-7	133562	-2.42	1	1	1	–
4683690..4716539	*S. typhi* CT18	SPI-10	32850	-5.51	0	1	1	–

4.2 Methods

The methodology followed throughout this analysis is summarized as flowchart in (Figure 4.1), and described in the following sections.

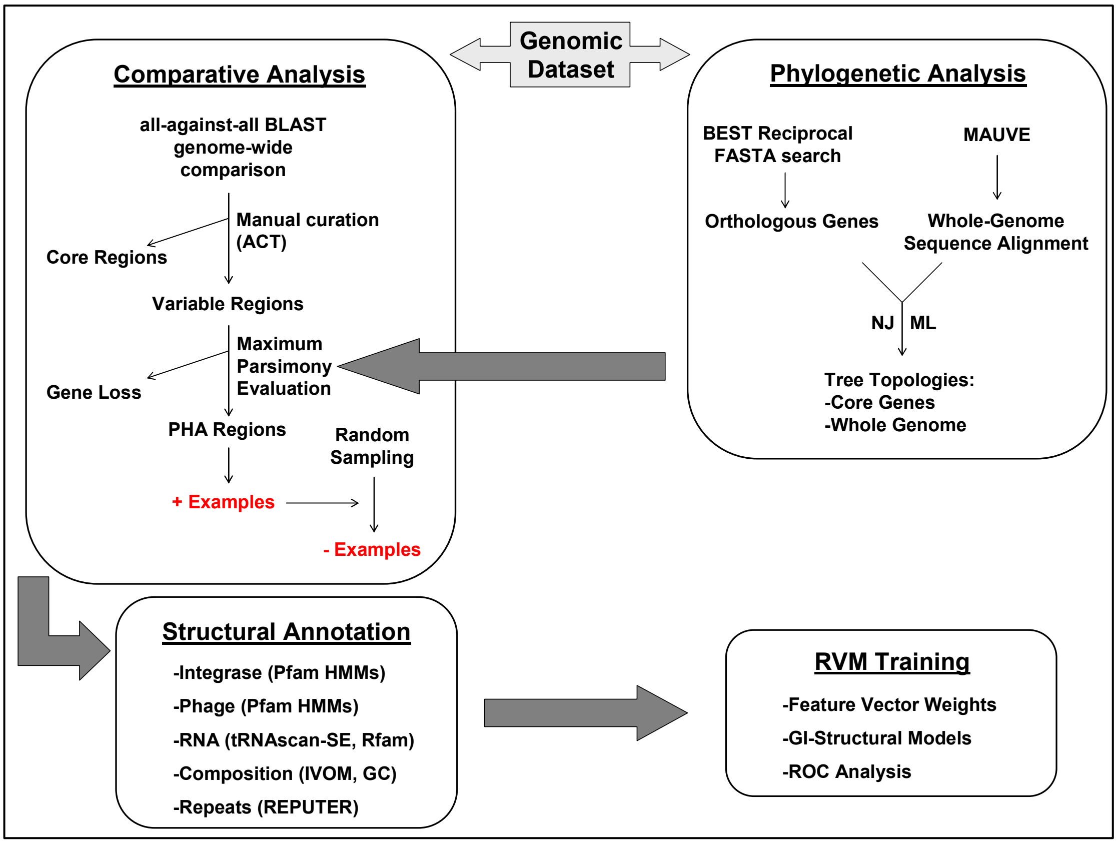


Figure 4.1: Flowchart summarizing the major steps in the methodology followed throughout this analysis: A phylogenetic analysis using both whole-genome sequence (if applicable) and the amino acid sequence of the core gene products was carried out enabling the construction of the reference tree topology for each genus. In a second step, a comparative analysis (genome-wise) was performed between the chromosomes of each genus and the corresponding outgroups, leading to the identification of regions with limited phylogenetic distribution. In a third step, a maximum parsimony model (based on the reference tree topology) was applied in order to differentiate gene gain from gene loss events and exclude regions with limited phylogenetic distribution due to a gene loss event. The remaining regions formed the positive control dataset (i.e. putative horizontally acquired – PHA regions) of this analysis. The negative control dataset (i.e. non GIs), was built implementing a random sampling approach, sampling regions only within the inter-GI parts of the chromosome; both positive and negative examples were annotated structurally. In a final step, the structural features of each region were used as input vectors to a machine learning method (Relevance Vector Machine – RVM) leading to the construction of structural GI models.

4.2.1 Genomic Dataset

A list of all the 49 strains used in this comparative analysis is provided in Table 4.3. Throughout this analysis, I focused on the analysis of 37 reference bacterial strains from three different genera, namely *Salmonella*, *Staphylococcus* and *Streptococcus*. In order to differentiate a limited phylogenetic distribution pattern due to a gene gain or a gene loss event (under a maximum parsimony evaluation), 12 more distantly related bacterial strains that formed outgroups for the three reference genera were also included in this analysis.

The 12 outgroup genomes were used only in the maximum parsimony evaluation of the predicted regions and do not form part of the actual dataset for which the data were produced. Briefly, 11 *Salmonella* strains with four outgroups (*E. coli*, *Shigella*), 13 *Staphylococcus* strains with four outgroups (*Bacillus*, *Listeria*) and 13 *Streptococcus* strains with four outgroups (*Lactobacillus*, *Lactococcus*, *Enterococcus*) were analyzed.

4.2.2 Best reciprocal FASTA

For each of the three genera, all genomes were (pair-wise) compared against the others including the four outgroups. In order to infer the orthologous genes in each pair of genomes compared, I applied a best reciprocal FASTA (Pearson, 1990) method (details of the best reciprocal FASTA algorithm are given in section 2.2.5 of chapter 2). Overall 1952, 741 and 429 orthologous genes were identified in the *Salmonella*, *Staphylococcus* and *Streptococcus* datasets (including the corresponding four outgroups) respectively (Figure 4.2).

Table 4.3: The list of 49 strains used in this comparative analysis.

Organism	**Reference**	**Accession Number**
Escherichia coli K-12 MG1655	(Blattner *et al.*, 1997)	U00096
E.coli O157:H7 EDL933	(Perna *et al.*, 2001)	AE005174
E. coli CFT073	(Welch *et al.*, 2002)	AE014075
Shigella flexneri serotype 2a 301	(Jin *et al.*, 2002)	AE005674
Salmonella bongori 12419	http://www.sanger.ac.uk/Projects/Salmonella/	N/A
S. arizonae RSK2980	http://genome.wustl.edu/genome_index.cgi	N/A
S. enterica serovar Typhi CT18	(Parkhill *et al.*, 2001)	AL513382
S. enterica serovar Typhi TY2	(Deng *et al.*, 2003)	AE014613
S. enterica serovar paratyphi A SARB42	(McClelland *et al.*, 2004)	CP000026
S. enterica serovar paratyphi A AKU_12601	http://genome.wustl.edu/genome_index.cgi	N/A
S. enterica serovar Typhimurium SL1344	http://www.sanger.ac.uk/Projects/Salmonella/	N/A
S. enterica serovar Typhimurium LT2	(McClelland *et al.*, 2001)	AE006468
S. enterica serovar Typhimurium DT104	http://www.sanger.ac.uk/Projects/Salmonella/	N/A
S. enterica serovar Enteritidis PT4	http://www.sanger.ac.uk/Projects/Salmonella/	N/A
S. enterica serovar Gallinarum 287/91	http://www.sanger.ac.uk/Projects/Salmonella/	N/A
Bacillus subtilis 168	(Kunst *et al.*, 1997)	AL009126
Bacillus anthracis Ames	http://cmr.tigr.org/tigr-scripts/CMR/GenomePage.cgi?org=gba	AE017334
Listeria innocua Clip11262	(Glaser *et al.*, 2001)	AL592022
Listeria monocytogenes EGD-e	(Glaser *et al.*, 2001)	AL591824
Staphylococcus saprophyticus ATCC 15305	(Takeuchi *et al.*, 2005)	AP008934
Staphylococcus haemolyticus JCSC1435	(Takeuchi *et al.*, 2005)	AP006716
Staphylococcus epidermidis ATCC 12228	(Zhang *et al.*, 2003)	AE015929
Staphylococcus epidermidis RP62A	(McGillivary *et al.*, 2005)	CP000029
Staphylococcus aureus MRSA252	(Holden *et al.*, 2004)	BX571856
Staphylococcus aureus RF122	(Herron-Olson *et al.*, 2007)	AJ938182
Staphylococcus aureus Mu50	(Takeuchi *et al.*, 2005)	BA000017
Staphylococcus aureus N315	(Takeuchi *et al.*, 2005)	BA000018
Staphylococcus aureus MSSA476	(Holden *et al.*, 2004)	BX571857
Staphylococcus aureus MW2	(Takeuchi *et al.*, 2005)	BA000033
Staphylococcus aureus USA300	(Diep *et al.*, 2006)	CP000255
Staphylococcus aureus COL	(McGillivary *et al.*, 2005)	CP000046
Staphylococcus aureus NCTC 8325	http://www.genome.ou.edu/staph.html	CP000253
Lactobacillus johnsonii NCC 533	(Pridmore *et al.*, 2004)	AE017198
Lactobacillus plantarum WCFS1	(Kleerebezem *et al.*, 2003)	AL935263
Enterococcus faecalis V583	(Paulsen *et al.*, 2003)	AE016830
Lactococcus lactis IL1403	(Bolotin *et al.*, 2001)	AE005176
Streptococcus pneumoniae R6	(Hoskins *et al.*, 2001)	AE007317
Streptococcus pneumoniae TIGR4	(Tettelin *et al.*, 2001)	AE005672
Streptococcus suis P1/7	http://www.sanger.ac.uk/Projects/S_suis/	N/A
Streptococcus thermophilus CNRZ1066	(Bolotin *et al.*, 2004)	CP000024
Streptococcus thermophilus LMG 18311	(Bolotin *et al.*, 2004)	CP000023
Streptococcus agalactiae NEM316	(Glaser *et al.*, 2002)	AL732656
Streptococcus agalactiae A909	(Tettelin *et al.*, 2005)	CP000114
Streptococcus uberis 0140J	http://www.sanger.ac.uk/Projects/S_uberis/	N/A
Streptococcus equi 4047	http://www.sanger.ac.uk/Projects/S_equi/	N/A
Streptococcus pyogenes MGAS10750	(Beres *et al.*, 2006)	CP000262
Streptococcus pyogenes MGAS2096	(Beres *et al.*, 2006)	CP000261
Streptococcus pyogenes MGAS9429	(Beres *et al.*, 2006)	CP000259
Streptococcus pyogenes Manfredo	(Ramsden *et al.*, 2007)	AM295007

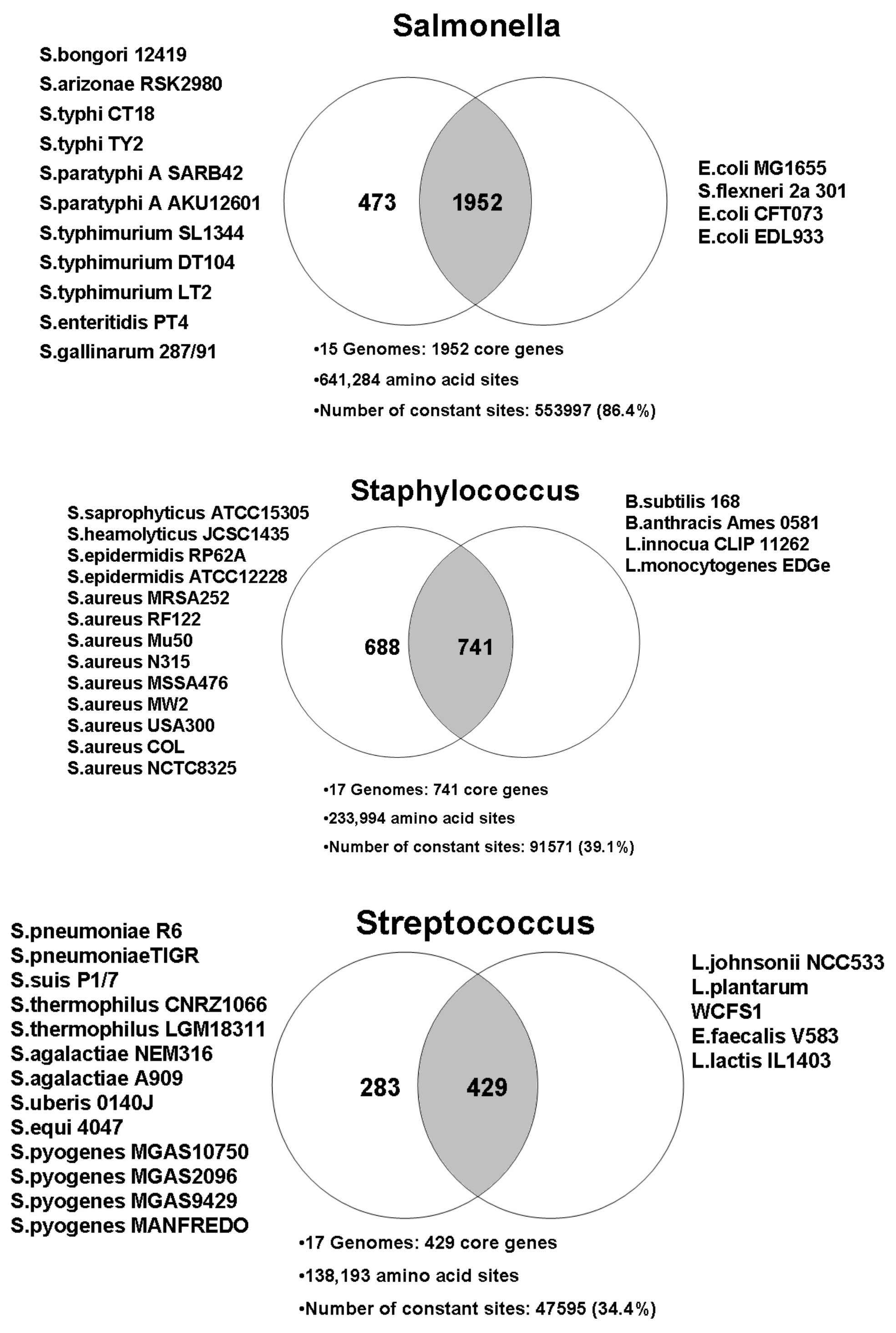


Figure 4.2: Venn diagram illustrating the orthologous genes shared between each of the three reference genera and the corresponding outgroup strains: 473 *Salmonella*-specific and 1952 core genes (genes shared between the *Salmonella* and the four outgroup strains) (top), 688 *Staphylococcus*-specific and 741 core genes (middle), 283 *Streptococcus*-specific and 429 core genes (bottom).

4.2.3 Multiple Sequence Alignments

Whole genome sequence alignments were made using the MAUVE algorithm (Darling *et al.*, 2004); for details about this algorithm see section 3.2.1 of chapter 3. The complete chromosome sequence of the 11 *Salmonella* strains and the four outgroups were aligned. For the *Staphylococcus* dataset, only the 13 *Staphylococcus* chromosomes were aligned, excluding the four outgroup sequences due to the overall low sequence similarity to the *Staphylococcus* genomes.

For the *Streptococcus* dataset the overall low sequence similarity between the different strains did not allow the construction of whole genome sequence alignments. Moreover, for each genus, amino acid sequence alignments of the core gene (i.e. orthologous genes shared by all the strains of a given genus and the corresponding outgroups) products were also built using the CLUSTALW (Thompson *et al.*, 1994) software; the alignments were manually inspected and curated.

4.2.4 Phylogenetic analysis

For the construction of the reference tree topology, modules of the PHYLIP package version 3.65 (Felsenstein, 1989) were implemented. More specifically, for the whole genome sequence alignments (*Salmonella* and *Staphylococcus* datasets), the DNADIST module with the method for correcting the rate heterogeneity among sites was used. I also used the NEIGHBOR module, which implements the Neighbor-Joining (NJ) method (Saitou and Nei, 1987) and the DNAML module which implements the Maximum Likelihood (ML) method for DNA sequences (Felsenstein and Churchill, 1996); the models of nucleotide substitution were those described in chapter 3, i.e. F84 (Kishino and Hasegawa, 1989), K80 (Kimura, 1980) and JC (Jukes and Cantor, 1969); for details about the NJ and the ML methods, see section 3.2.2 of chapter 3.

For the construction of NJ and the ML tree topologies utilizing the amino acid sequence alignment of the core gene products for each genus (and the corresponding outgroups) the PROTDIST, NEIGHBOR and

PROML modules of the PHYLIP package were used, exploiting two models of evolution (see next paragraph), i.e. the JTT model (Jones *et al.*, 1992) and the approximation method proposed by Kimura (Kimura, 1983). Different tree topologies for a given lineage were evaluated further through the PROML module of PHYLIP and the TREE-PUZZLE (Schmidt *et al.*, 2002) software, exploiting the model with the highest number of parameters; for each genus the tree topology with the highest likelihood was selected as the reference. All the parameters were determined from the data using the TREE-PUZZLE software.

Models of amino acid substitution, as opposed to nucleotide substitutions, are mainly based on empirically derived parameters. Such empirical models describe the amino acid substitutions by analyzing multiple sequences from existing protein sequence databases; i.e. sequence alignments between very similar proteins are used to obtain estimates of the relative substitution rates between different amino acid pairs. In other words these empirical models, as opposed to mechanistic models, do not model explicitly the dynamics driving the amino acid substitution, e.g. mutational biases, translation of codons into amino acids and constraints at the amino acid level.

The first attempt to construct an empirically derived amino acid substitution model was that described by Dayhoff *et al.* (Dayhoff *et al.*, 1978). Phylogenetic trees were constructed exploiting the sequences of 71 protein families available at that time; their ancestral protein sequences were reconstructed implementing a parsimony method and the most likely residues at each position in the ancestral sequences were inferred.

In order to reduce the impact of multiple substitutions, the authors focused only on very similar protein sequences, i.e. each pair of sequences differed in less than 15% of their residues. The frequencies of all pairing of residues between sequences and their (reconstructed) ancestral sequences were counted and extrapolated to longer times to derive substitution probabilities. Dayhoff *et al.* approximated the *transition-probability* matrix by defining the substitution matrix to be 1 PAM (i.e. point accepted mutations) matrix if the expected number of substitutions per site was 0.01. Therefore, the unit "1 PAM" can be seen as the amount of evolution which changes, on average, 1% of amino acids in a protein sequence; for different sequence distances, e.g. t = 1 or 2.5 substitutions per site, different PAM matrices (i.e. PAM_{100} or PAM_{250}, respectively) can be derived.

Later on, Jones *et al.* (Jones *et al.*, 1992) exploiting the same principle as did Dayhoff *et al.* (Dayhoff *et al.*, 1978) updated the Dayhoff matrix analyzing a much larger collection of proteins sequences and this updated matrix is known as the JTT matrix.

An approximation of the PAM distance was proposed by Kimura (Kimura, 1983), and this is simply a distance formula that measures the proportion (p) of different amino acid residues between two sequences, as follows:

$$D = -\ln(1 - p - 0.2p^2)$$

The Kimura distance has the advantage of being very fast, but does not take into account the different types of amino acid residue and substitution; furthermore, the distance between two sequences becomes infinite if more than 85.41% of their amino acid residues are different.

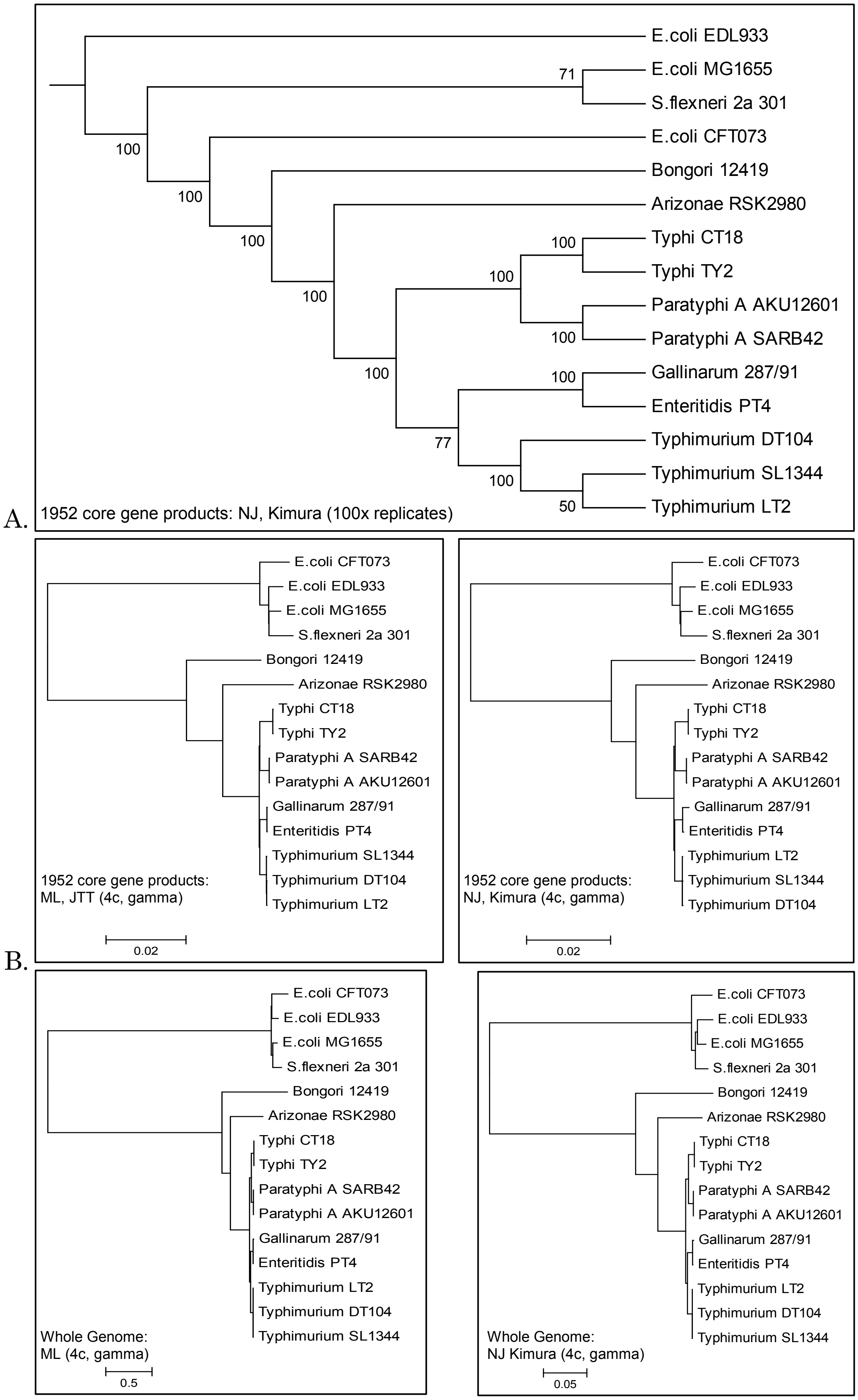
4.2.5 Comparative analysis

The genomic sequences of each genus and the corresponding outgroups were compared using a genome-wide, all-against-all BLAST (Altschul *et al.*, 1997) comparison; the results were visualized through ACT (Carver *et al.*, 2005) and manually inspected. Genomic regions (≥ 2 coding sequences – CDSs) of limited phylogenetic distribution that are present in some of the strains while being absent from the rest are processed further (at this stage core genomic regions, shared by all strains, are excluded).

In a second step regions of limited phylogenetic distribution are analyzed applying a maximum parsimony model (for details see section 3.2.4 of chapter 3), in order to differentiate gene gain (HGT) from gene loss; the maximum parsimony model is based on the reference tree topology of each genus (Figure 4.3, Figure 4.4 and Figure 4.5). Genomic regions identified under this framework as being putative horizontally acquired, formed the positive control set of this analysis; overall 331 putative GIs were sampled from the 37 reference chromosomes (Table 4.4, Figure 4.6).

Table 4.4: A list of the positive (putative GIs) and the negative (non-GIs) control regions, sampled from the 37 reference chromosomes used in this analysis.

Datasets	**Positive examples**	**Negative examples**	**Total**
Salmonella	211	210	421
Streptococcus	54	53	107
Staphylococcus	66	74	140
Gram –	211	210	421
Gram +	120	127	247
Gram +/–	331	337	668


A.
1952 core gene products: NJ, Kimura (100x replicates)
E.coli EDL933
E.coli MG1655
S.flexneri 2a 301
E.coli CFT073
Bongori 12419
Arizonae RSK2980
Typhi CT18
Typhi TY2
Paratyphi A AKU12601
Paratyphi A SARB42
Gallinarum 287/91
Enteritidis PT4
Typhimurium DT104
Typhimurium SL1344
Typhimurium LT2
71
100
100
100
100
100
100
100
100
100
77
100
50
B.
1952 core gene products:
ML, JTT (4c, gamma)
0.02
E.coli CFT073
E.coli EDL933
E.coli MG1655
S.flexneri 2a 301
Bongori 12419
Arizonae RSK2980
Typhi CT18
Typhi TY2
Paratyphi A SARB42
Paratyphi A AKU12601
Gallinarum 287/91
Enteritidis PT4
Typhimurium SL1344
Typhimurium DT104
Typhimurium LT2
1952 core gene products:
NJ, Kimura (4c, gamma)
0.02
E.coli CFT073
E.coli EDL933
E.coli MG1655
S.flexneri 2a 301
Bongori 12419
Arizonae RSK2980
Typhi CT18
Typhi TY2
Paratyphi A SARB42
Paratyphi A AKU12601
Gallinarum 287/91
Enteritidis PT4
Typhimurium LT2
Typhimurium SL1344
Typhimurium DT104
Whole Genome:
ML (4c, gamma)
0.5
E.coli CFT073
E.coli EDL933
E.coli MG1655
S.flexneri 2a 301
Bongori 12419
Arizonae RSK2980
Typhi CT18
Typhi TY2
Paratyphi A SARB42
Paratyphi A AKU12601
Gallinarum 287/91
Enteritidis PT4
Typhimurium LT2
Typhimurium DT104
Typhimurium SL1344
Whole Genome:
NJ Kimura (4c, gamma)
0.05
E.coli CFT073
E.coli EDL933
E.coli MG1655
S.flexneri 2a 301
Bongori 12419
Arizonae RSK2980
Typhi CT18
Typhi TY2
Paratyphi A SARB42
Paratyphi A AKU12601
Gallinarum 287/91
Enteritidis PT4
Typhimurium LT2
Typhimurium DT104
Typhimurium SL1344

C.

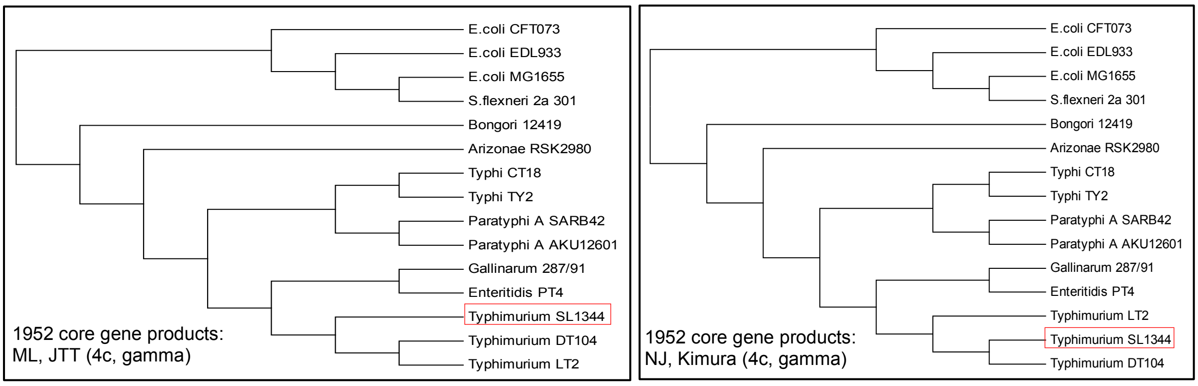
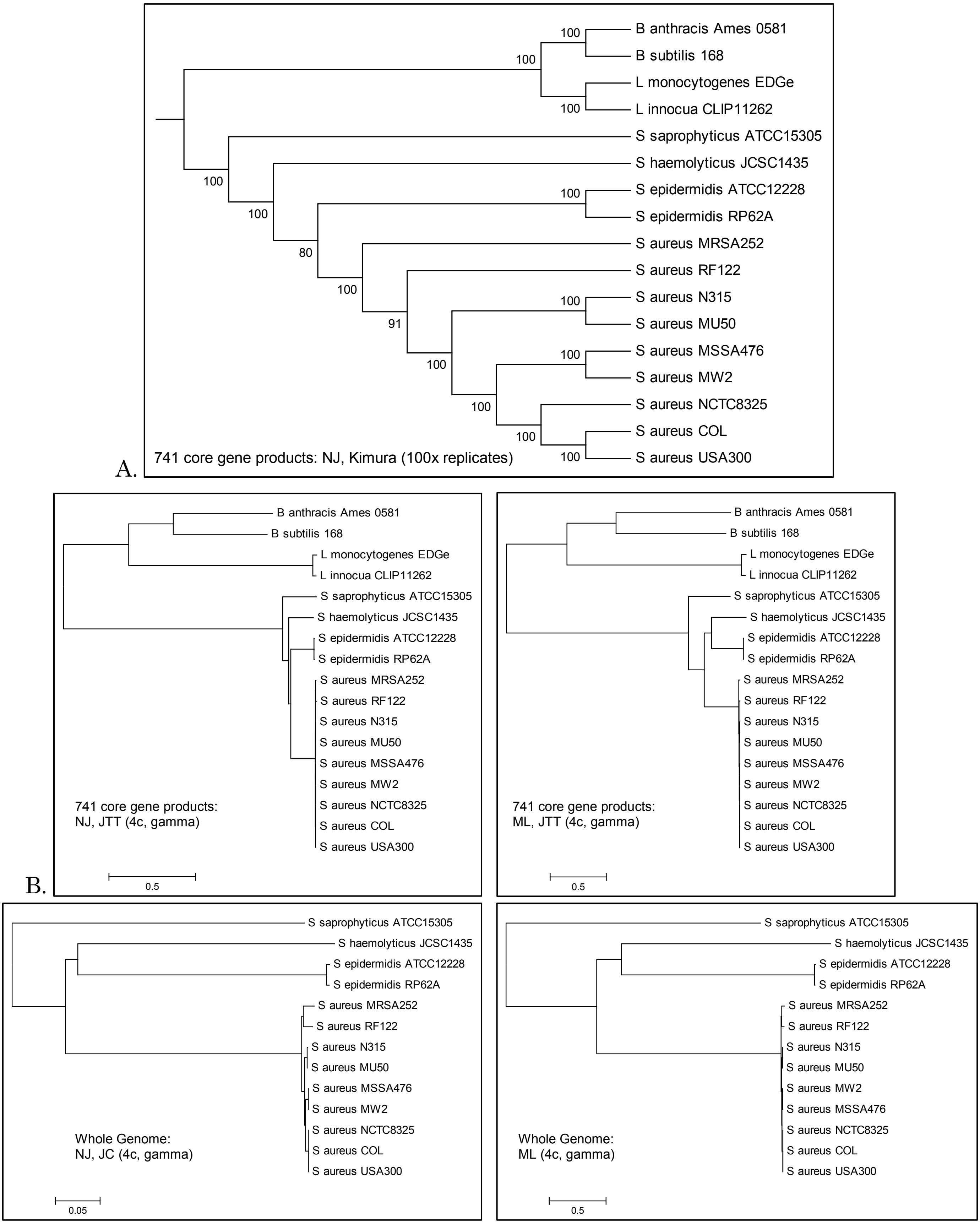
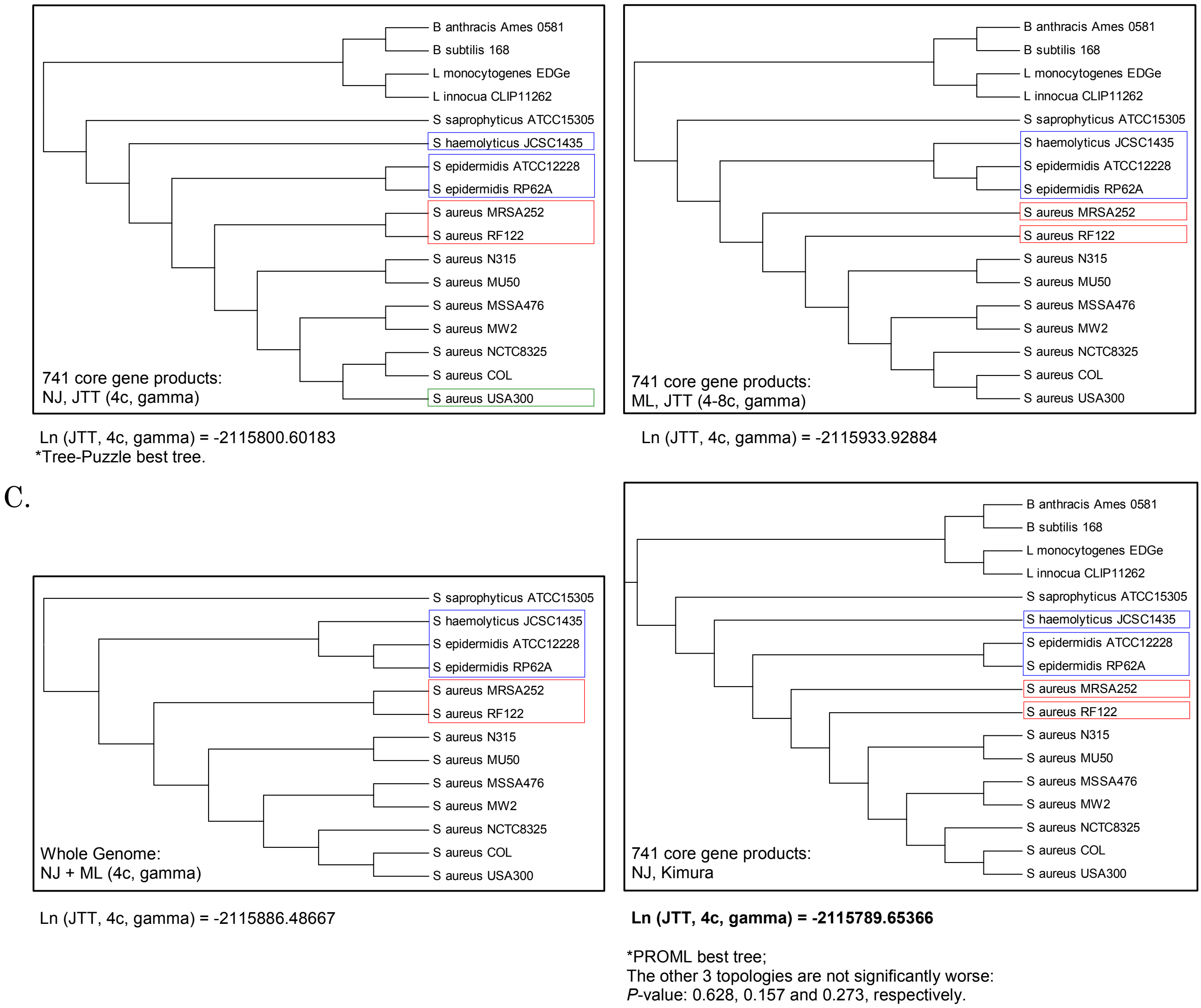


Figure 4.3: **A.** The phylogenetic relationship between the 11 *Salmonella* and the four outgroup genomes (ignoring branch length), is shown as cladogram. Bootstrap values (proportions out of 100) are given for each node. The tree topology is based on the amino acid sequence of 1952 core gene products shared by the 15 genomes. **B.** Phylogenetic tree topologies using the ML (left) and the NJ (right) method, based on the alignment of the 1952 core gene products (top) and the whole chromosome sequences (bottom) of 11 *Salmonella* and four outgroup genomes. **C.** Differences between the tree topologies (core gene products) given by the ML and the NJ methods are highlighted; the only difference in terms of node topology lies within the Typhimurium lineage. In the ML topology, DT104 and LT2 are grouped together, while in the NJ topology DT104 is grouped together with SL1344. The bootstrap value of 50 supports the observed ambiguity.


B anthracis Ames 0581
B subtilis 168
L monocytogenes EDGe
L innocua CLIP11262
S saprophyticus ATCC15305
S haemolyticus JCSC1435
S epidermidis ATCC12228
S epidermidis RP62A
S aureus MRSA252
S aureus RF122
S aureus N315
S aureus MU50
S aureus MSSA476
S aureus MW2
S aureus NCTC8325
S aureus COL
S aureus USA300
100
80
91
741 core gene products: NJ, Kimura (100x replicates)
A.
741 core gene products:
NJ, JTT (4c, gamma)
741 core gene products:
ML, JTT (4c, gamma)
0.5
B.
Whole Genome:
NJ, JC (4c, gamma)
Whole Genome:
ML (4c, gamma)
0.05



Ln (JTT, 4c, gamma) = -2115800.60183
*Tree-Puzzle best tree.

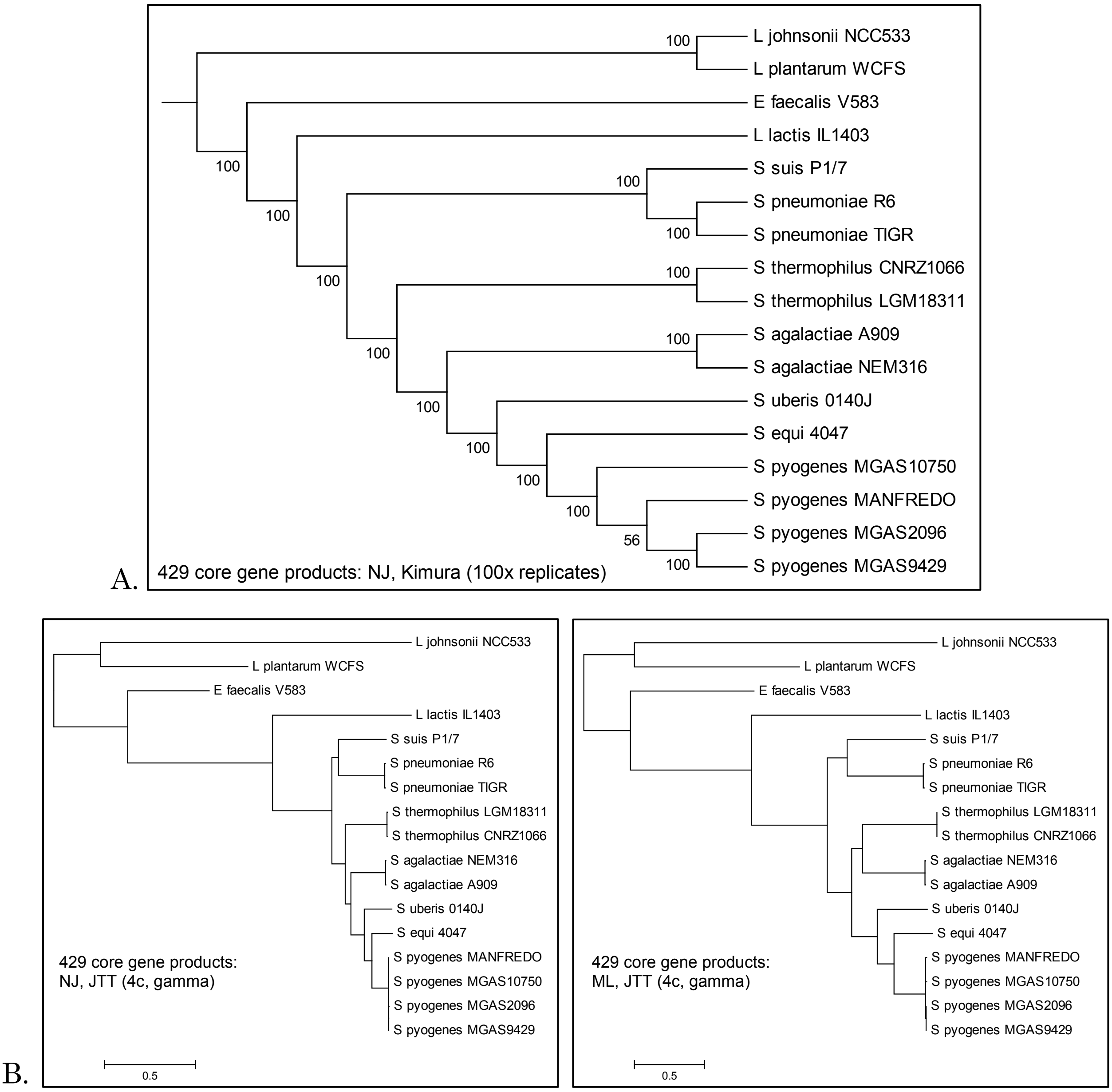
Ln (JTT, 4c, gamma) = -2115933.92884

Ln (JTT, 4c, gamma) = -2115886.48667

Ln (JTT, 4c, gamma) = -2115789.65366

*PROML best tree;
The other 3 topologies are not significantly worse:
P-value: 0.628, 0.157 and 0.273, respectively.

Figure 4.4: **A.** The phylogenetic relationship between the 13 *Staphylococcus* and the four outgroup genomes (ignoring branch length), is shown as cladogram. Bootstrap values are given for each node. The tree topology is based on the amino acid sequence of 741 core gene products shared by the 17 genomes. **B.** Phylogenetic tree topologies using the NJ (left) and the ML (right) method, based on the alignment of the 741 core gene products (top) and the whole chromosome sequences (bottom) of 13 *Staphylococcus* and four outgroup genomes.
C. Differences between the tree topologies given by the ML and the NJ methods are highlighted. The likelihood (Ln) for each tree topology, under a JTT model with four categories of sites (4c) and a Gamma distribution for modelling the rate variation among sites (gamma), is provided for each topology. Based on TREE-PUZZLE, the best tree topology is the one given by the NJ method (JTT, 4c, gamma). Based on the tree topology evaluation of *PROML*, the NJ (*Kimura* model) method gives the best tree topology (highest Ln); however the other three topologies are not significantly worse (*p*-value: 0.628, 0.157 and 0.273 respectively), suggesting that the observed differences are close to the systematic error of those methods.


A.
429 core gene products: NJ, Kimura (100x replicates)
L johnsonii NCC533
L plantarum WCFS
E faecalis V583
L lactis IL1403
S suis P1/7
S pneumoniae R6
S pneumoniae TIGR
S thermophilus CNRZ1066
S thermophilus LGM18311
S agalactiae A909
S agalactiae NEM316
S uberis 0140J
S equi 4047
S pyogenes MGAS10750
S pyogenes MANFREDO
S pyogenes MGAS2096
S pyogenes MGAS9429
100
100
100
100
100
100
100
100
100
100
100
100
100
56
100
B.
429 core gene products:
NJ, JTT (4c, gamma)
L johnsonii NCC533
L plantarum WCFS
E faecalis V583
L lactis IL1403
S suis P1/7
S pneumoniae R6
S pneumoniae TIGR
S thermophilus LGM18311
S thermophilus CNRZ1066
S agalactiae NEM316
S agalactiae A909
S uberis 0140J
S equi 4047
S pyogenes MANFREDO
S pyogenes MGAS10750
S pyogenes MGAS2096
S pyogenes MGAS9429
0.5
429 core gene products:
ML, JTT (4c, gamma)
L johnsonii NCC533
L plantarum WCFS
E faecalis V583
L lactis IL1403
S suis P1/7
S pneumoniae R6
S pneumoniae TIGR
S thermophilus LGM18311
S thermophilus CNRZ1066
S agalactiae NEM316
S agalactiae A909
S uberis 0140J
S equi 4047
S pyogenes MANFREDO
S pyogenes MGAS10750
S pyogenes MGAS2096
S pyogenes MGAS9429
0.5

C.

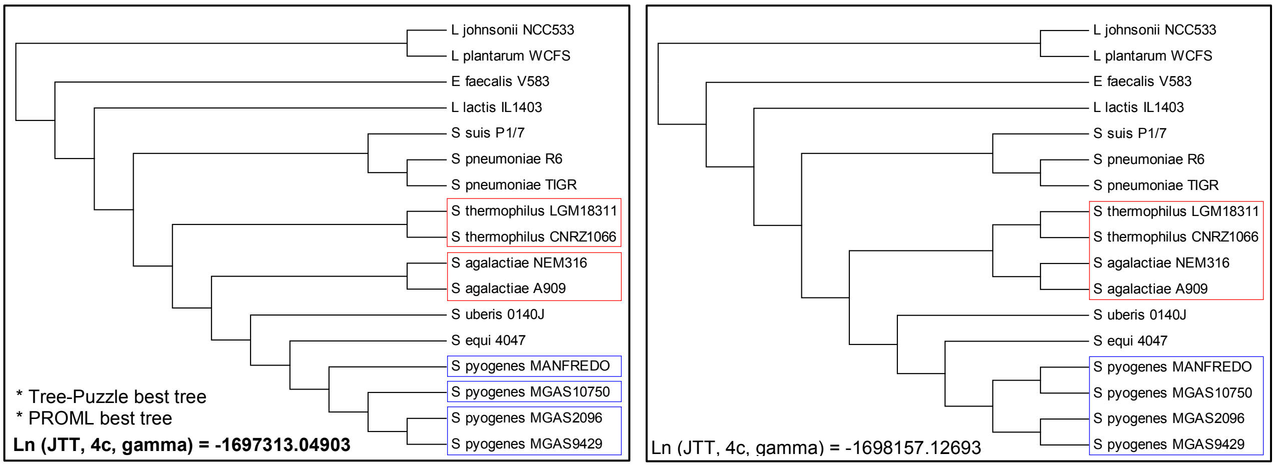
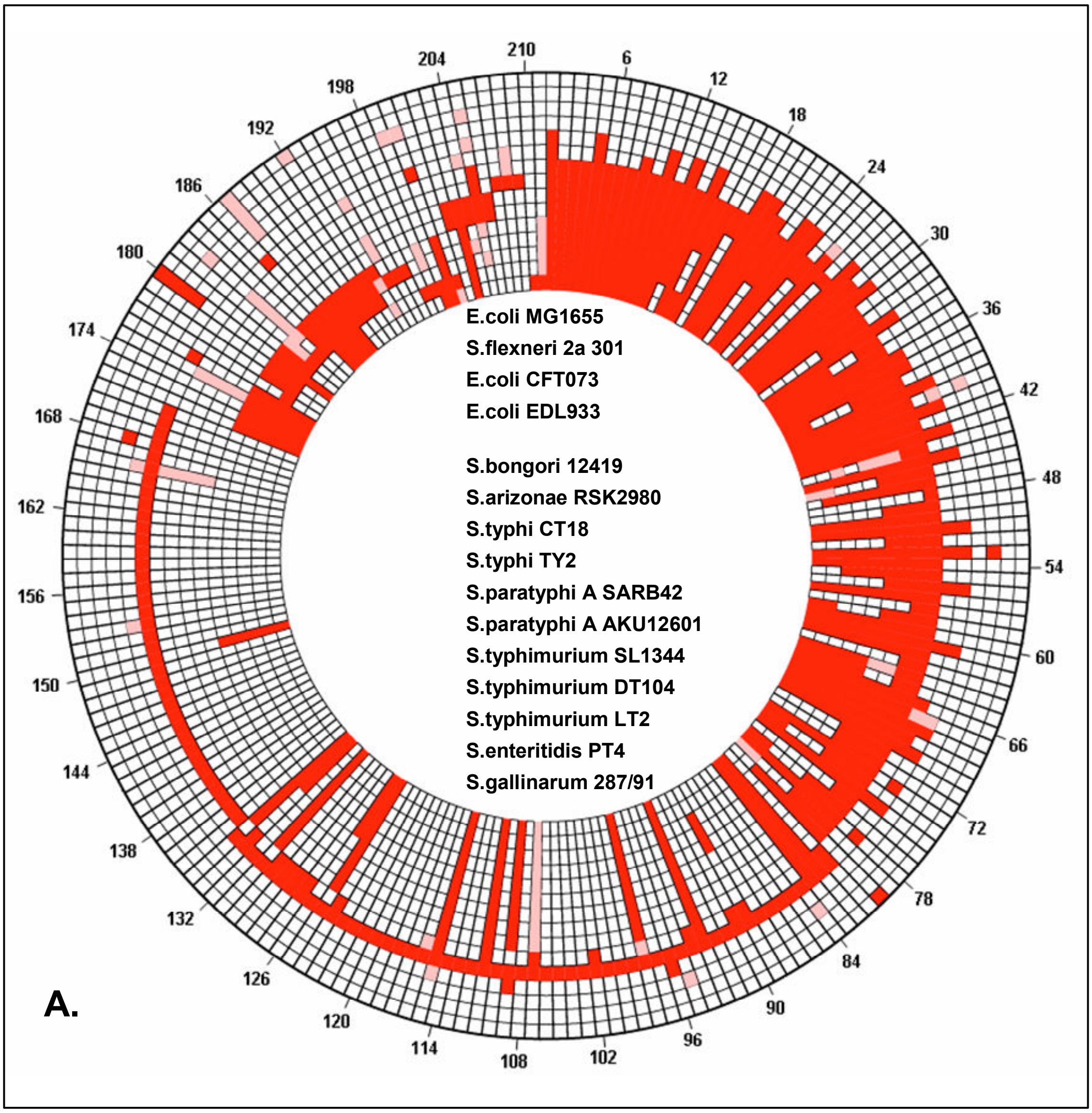
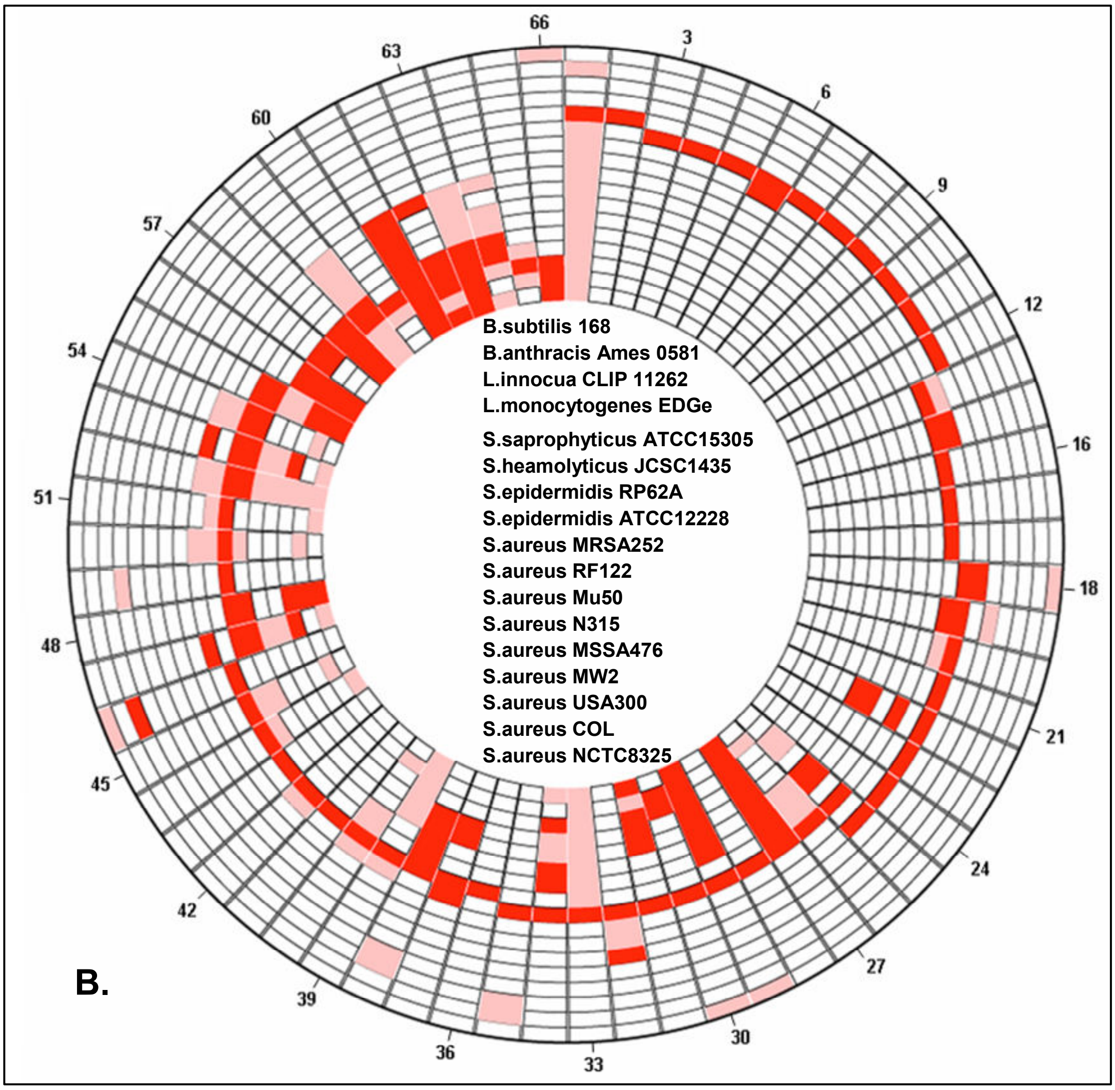


Figure 4.5: **A.** The phylogenetic relationship between the 13 *Streptococcus* and the four outgroup genomes (ignoring branch length), is shown as cladogram. Bootstrap values are given for each node. The tree topology is based on the amino acid sequence of 429 core gene products shared by the 17 genomes. **B.** Phylogenetic tree topologies using the NJ (left) and the ML (right) method, for the 429 core gene products of the 13 *Streptococcus* and the four outgroup genomes. **C.** Differences in the tree topology given by the ML and the NJ methods are highlighted: based on the tree topology evaluation (TREE-PUZZLE and PROML) the NJ method (left) gives the topology with the highest likelihood (best tree).


210
6
12
18
24
30
36
42
48
54
60
66
72
78
84
90
96
102
108
114
120
126
132
138
144
150
156
162
168
174
180
186
192
198
204
E.coli MG1655
S.flexneri 2a 301
E.coli CFT073
E.coli EDL933
S.bongori 12419
S.arizonae RSK2980
S.typhi CT18
S.typhi TY2
S.paratyphi A SARB42
S.paratyphi A AKU12601
S.typhimurium SL1344
S.typhimurium DT104
S.typhimurium LT2
S.enteritidis PT4
S.gallinarum 287/91
A.


66
3
6
9
12
16
18
21
24
27
30
33
36
39
42
45
48
51
54
57
60
63
B.subtilis 168
B.anthracis Ames 0581
L.innocua CLIP 11262
L.monocytogenes EDGe
S.saprophyticus ATCC15305
S.heamolyticus JCSC1435
S.epidermidis RP62A
S.epidermidis ATCC12228
S.aureus MRSA252
S.aureus RF122
S.aureus Mu50
S.aureus N315
S.aureus MSSA476
S.aureus MW2
S.aureus USA300
S.aureus COL
S.aureus NCTC8325
B.

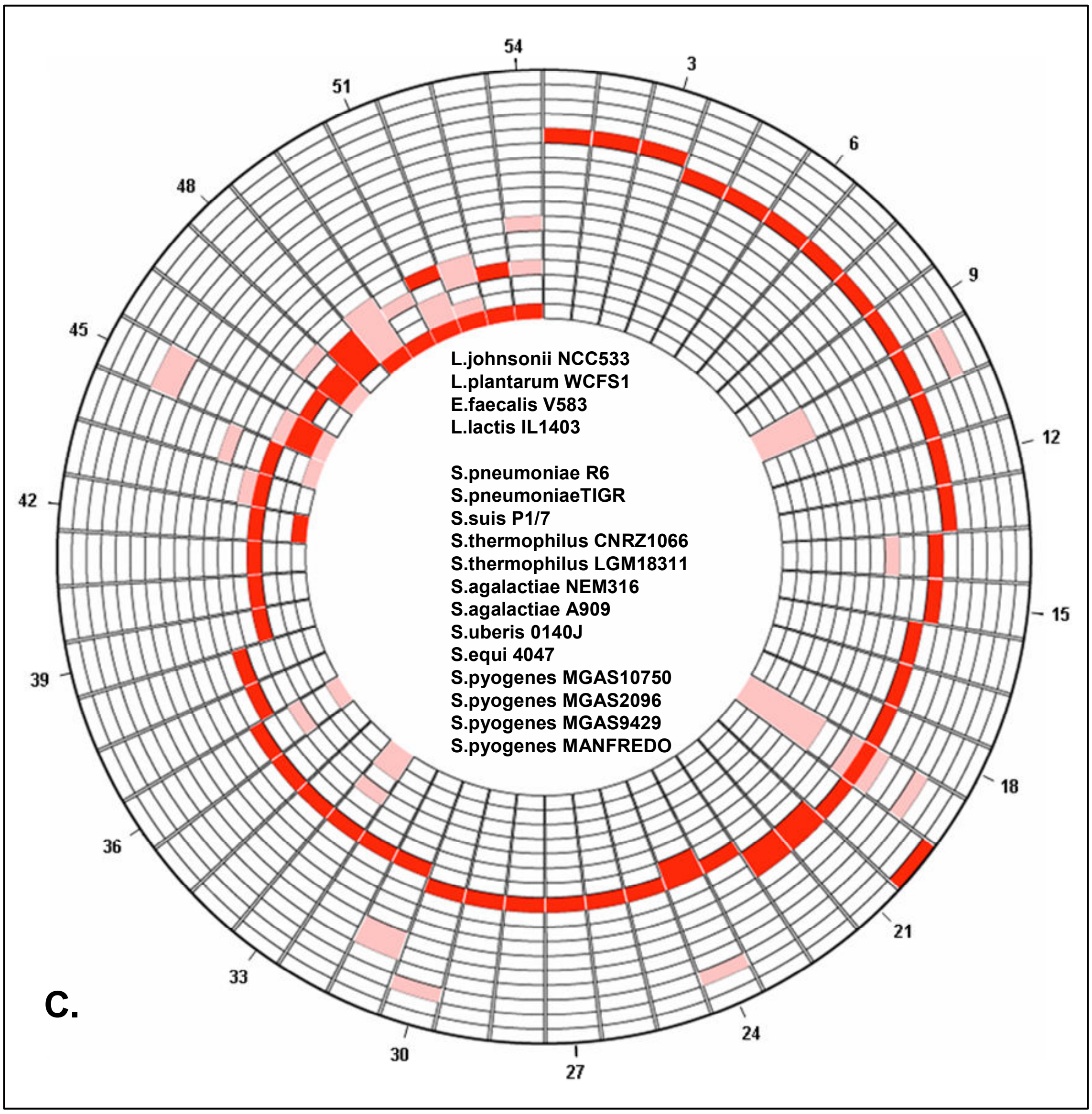


Figure 4.6: Circular map of the *Salmonella* (A), *Staphylococcus* (B) and *Streptococcus* (C) "mobilome", illustrating the phylogenetic distribution of the putative GIs identified in the three reference lineages (red: presence, pink: partial presence, white: absence). The list of strains (outwards-inwards orientation relative to the map) is embedded at the centre of the circular map. The regions are arbitrarily numbered based on the strain first found.

4.2.6 Random sampling

For the construction of the negative control dataset, i.e. genomic regions that are not GIs, a random sampling approach was followed. For each genome with identified putative GIs, an equal number of non-GI regions were randomly sampled, sampling the size distribution of the corresponding genus-specific GIs (Figure 4.7). Overall, this analysis yielded 337 non-GIs, giving a total number of 668 training sets (Table 4.4 and Appendix E, F and G). Random sampling was "forced" to occur only within inter-GI regions of each chromosome. The results of the random sampling approach were manually curated, removing randomly sampled regions that had been already sampled from other chromosomes of different strains of the same genus; the manual curation filtered out any redundancy in the training set that could possibly affect the training and evaluation process. For theses reasons, the numbers of positive and negative examples for each genus are slightly different.

4.2.7 Structural annotation

4.2.7.1 Integrase(-like) protein domains

Each query genome (six frame translation) was searched against 15 integrase(-like) Pfam (Sonnhammer *et al.*, 1998) Hidden Markov Models (HMMs), using the HMMER software (http://hmmer.janelia.org/). Throughout this analysis, 15 protein domains (Appendix H) that are frequently found in proteins involved in the mobilization of DNA are referred to as integrase-like domains, or simply "integrase".

4.2.7.2 Phage-related protein domains

In order to predict CDSs of putative phage origin, the *hmmpfam* search option of the HMMER package was used and each query genome (six frame translation) was searched against a manually constructed database of 191 phage-related Pfam HMMs (Appendix H).

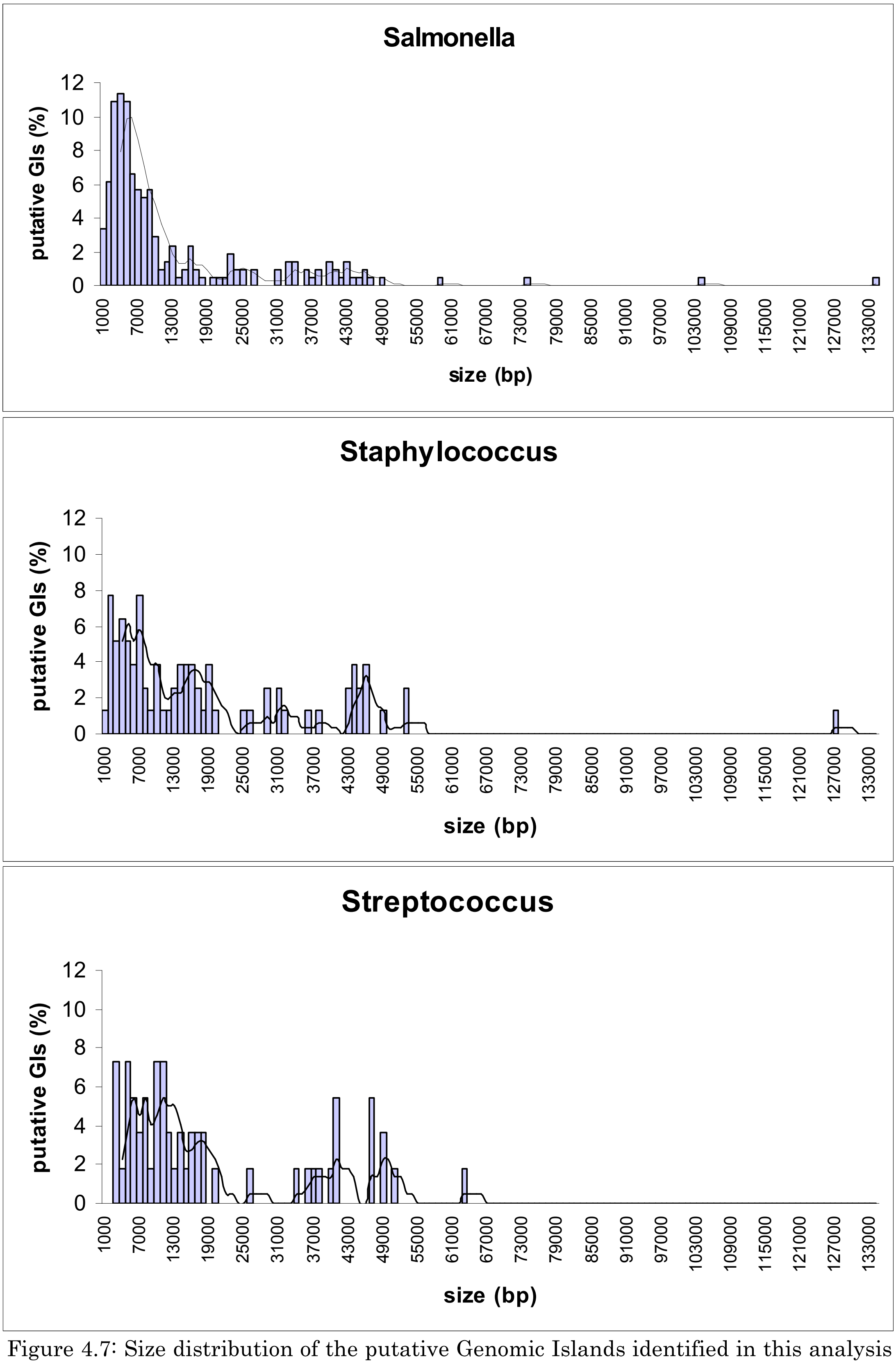


Figure 4.7: Size distribution of the putative Genomic Islands identified in this analysis for the three reference genera.

4.2.7.3 Non-coding RNA

Each query genome was searched against the non-coding RNA families of the Rfam database (Griffiths-Jones *et al.*, 2003). This methodology was followed in order that putative associations of GIs with other non-coding RNA families (apart from the tRNA and tmRNA genes) could be captured.

4.2.7.4 Compositional analysis

For all the 668 regions identified in this analysis, their Interpolated Variable Order Motif (IVOM) score (Vernikos and Parkhill, 2006) was calculated, using the Alien_Hunter algorithm. The IVOM frequency is a weighted sum of compositional biases derived from different size ($1 \leq k \leq 8$) k-mers that captures both low and high order compositional deviation from the backbone composition. The IVOM score is expressed as the relative entropy between the query and the genome-backbone (variable order) compositional distribution, i.e. the higher the IVOM score is, the stronger the compositional deviation.

4.2.7.5 Repeat analysis

Repeat analysis at the boundaries of each of the 668 regions was performed, using the REPuter software (Kurtz and Schleiermacher, 1999). The REPuter parameters used are as follows: Type of repeats (= Forward, Complemented), minimum size of repeats (= 18bp), number (hamming distance) of mismatches for degenerate repeats (= 3).

4.2.7.6 Other

All 668 regions were further annotated in terms of size (bp), gene density (number of genes per kb) and their insertion point; in the latter case two distinct (binary) states were evaluated: insertion point within a CDS locus (disrupting the corresponding CDS) or insertion within an intergenic part of the chromosome.

4.2.8 Machine Learning

In order to build structural models of GIs, eight features were taken into account: The IVOM score (relative entropy), insertion point (1 if within a

CDS locus, 0 otherwise), size of each region (bp), gene density (genes/kb), repeats (binary: 1 if present, 0 otherwise), phage-related protein domains (binary), integrase(-like) protein domains (binary) and non-coding RNA (binary). Furthermore, the RNA feature was further divided into tRNA and misc_RNA subcategories; the same applies for the repeats feature that was further divided into DRs and inverted repeats (IRs) subcategories.

The aim of the machine learning in this analysis is dual: GI structural models will be trained in order to quantify (i.e. assign weights to) the relative contribution of each feature to the GI structure and in a second step the derived models will be used to classify previously unseen examples (GIs and non-GIs) enabling evaluation of the generalization properties of each model and capturing of any potential variation in the GI structure. For this purpose, 668 training sets were used to train 11 GI models using a Biojava (http://www.biojava.org) implementation of the Relevance Vector Machine (RVM) (Tipping, 2001).

The RVM is a method for sparse, Bayesian-based learning with applications in classification and regression analysis; sparse learning algorithms are methods that integrate the selection of features with learning of the optimal model parameters. The RVM is a model of identical functional form to the well-known Support Vector Machine (SVM) (Schölkopf *et al.*, 1999) that nonetheless overcomes a few of the limitations of the latter (Tipping, 2001). The RVM models exploit overall fewer basis functions relative to an SVM model, offering the advantage of increased sparsity, building simpler models with better generalization properties on unseen data. Moreover the RVM exploits a probabilistic Bayesian learning framework, i.e. the model gives estimates of the posterior probability of membership in one of the two classes (in classification analysis) rather than trying to make an "absolute" binary decision (as in the case of the SVM). The RVM method has been previously applied in detecting binding sites in human protein-coding sequences (Down *et al.*, 2006), in the identification of transcriptional start sites in mammalian DNA (Down and

Hubbard, 2002) and in a vertebrate gene finding method (Carter and Durbin, 2006).

Given a set of *N* examples (training set) along with their corresponding class (i.e. GI, non-GI) we are trying to build a model of how the input vectors $\{\mathrm{x}_i\}_{i=1}^{N}$ affect the corresponding classification $\{c_i\}_{i=1}^{N}$, with the aim of making predictions of the class for unseen input data, based on the model parameters (weights) $\{w_j\}_{j=1}^{K}$ calculated during the training; *K* denotes the number of basis functions (in our case structural features e.g. repeats, RNA, IVOM, etc) used to describe the data. Throughout this analysis, I will refer to the RVM model parameters *w* as "weights" because they quantify the relative contribution of each feature to the model, i.e. the higher the feature weight the higher its contribution to the model; note that for the model parameters *w* there is no actual upper or lower bound. In order to build structural GI models, the Generalized Linear Models (GLMs) (McCullagh and Nelder, 1989), a form of model suitable for classification and regression analysis, are exploited. A GI structural model (S_i) is the weighted sum of *K* basis functions of the form:

$$S_i = U + \sum_{j=1}^{K} w_j \cdot x_{ij} \tag{4.1}$$

For two-class classification (in our case class 1 corresponds to GI and class 0 to non-GI) the aim is to predict the posterior probability that a given input *x* is a true GI, given the model. In the case of a binary classification task, a commonly used link function for the GLMs is the logistic function:

$$\sigma(S_i) = \frac{1}{1+e^{-S_i}} \tag{4.2}$$

The logistic function (Figure 4.8) normalizes ($0 \leq \sigma(S_i) \leq 1$) the output of model S_i and can be considered as an estimate of the probability that a given structure is a true GI, given the model. In function 4.1, *U* is a constant that controls the output of this function, in such a way that the final score (assuming the logistic function) can take any value between 0 and 1.

The feature weight w is indicative of the actual feature contribution to the given model, (i.e. the higher the weight the higher the feature contribution), however it does not take into account the dispersion of the actual values of a given feature in the training set. A more reliable estimate of the actual feature importance can be calculated through the following function:

$$R_j = w_j \cdot SD_j \tag{4.3}$$

where R_j is the "importance" of feature j with weight w_j and standard deviation SD_j (the standard deviation of the actual values of a given basis function in the training set). Under this framework, a basis function with significant SD_j will be more important (higher R) than a basis function with comparable weight but with lower SD_j.

Details about the training and technical aspects of the RVM are discussed in detail in (Down and Hubbard, 2003; Tipping, 2001). Briefly, the probability that a given dataset is correctly classified given the model is given by the following function:

$$P(c \mid x, w) = \prod_{i=1}^{N} \sigma(S_i)^{ci}(1-\sigma(S_i))^{1-ci} \tag{4.4}$$

where S_i is the output of the linear model for the i-th data in the training set; note that for binary classification $c \in \{0,1\}$.

Exploiting Bayes' theorem (for details see section 3.2.2.3 of chapter 3), we can use the likelihood function 4.4 to infer possible weight values given the training dataset:

$$P(w \mid x, c) \propto P(w)P(c \mid x, w) \tag{4.5}$$

where $P(w)$ is the prior probability distribution over the weight values. Generally, the prior distribution can be a very broad, non-informative distribution, however in the case of the RVM, we are more interested in very sparse models (in order to avoid substantial over-fitting to the training set), as such we aim to favour simple over complex models. For this reason a new vector of parameters *a* is introduced that controls the width of the prior (i.e. *Gaussian* distribution *G*) over each weight:

$$P(w) = \prod_i \mathcal{G}(w_i \mid 0, a_i^{-1}) \tag{4.6}$$

Moreover, for the purposes of the Bayesian inference, a very broad (non-informative) *Gamma* distribution is used as the hyperprior over the *a* parameter. Note that the *a* parameter can be seen as the inverse variance of the *Gaussian* distribution (equation 4.6).

During the training process, the RVM is estimating appropriate values of the model weights in an iterative fashion, with the aim of maximizing the likelihood function (4.4). If a given basis function is informative when classifying the training dataset, then by setting its weight to a non-zero value, will increase the number of correctly classified data, which in turn will increase the likelihood function (4.4), and therefore the probability of the model given the training set. On the other hand, if a basis function is not informative (or has redundant information) for the classification task, there is no actual weight value that would increase the likelihood.

However, by setting the *a* parameter to a large value, the prior distribution becomes peaked around zero; as such the posterior probability of the model is maximized by setting the corresponding weight value to zero. When the value of the *a* parameter of a basis function is sufficiently high the corresponding basis function becomes irrelevant, and is removed from the model. Under this increased sparsity framework, RVM models avoid efficiently overfitting to the training dataset, selecting only a small number of "relevance" vectors, with good generalization properties on unseen datasets.

The issue of finding optimal model parameters, exploiting equation 4.5, can be solved by different approaches e.g. Maximum Likelihood estimation (Tipping, 2001) or the Metropolis-Hastings (Hastings, 1970; Metropolis *et al.*, 1953) algorithm; for details see section 3.2.2.3 of chapter 3.

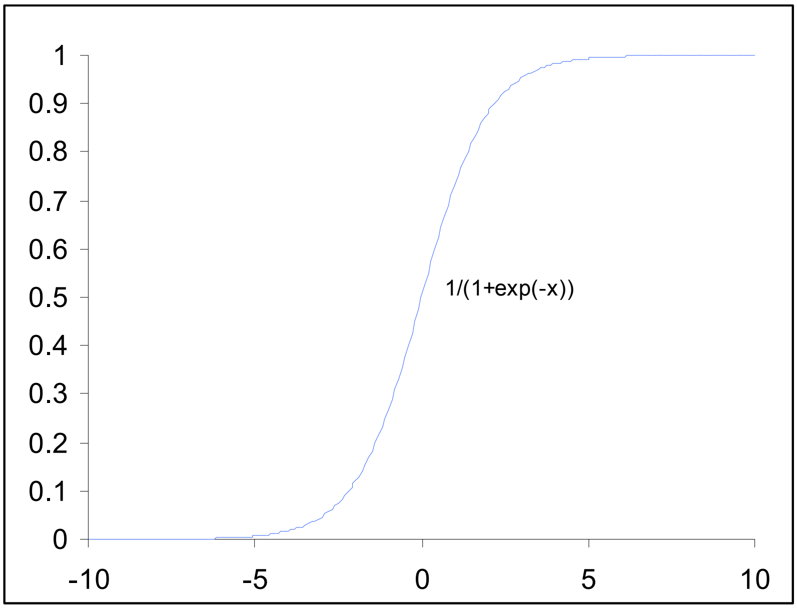


Figure 4.8: The logistic function $f(x) = 1/(1+e^{-x})$.

4.2.9 ROC curve

In order to evaluate the performance of the RVM classifier under different GI models, I implemented a receiver operating characteristic (ROC) curve analysis (Appendix I). The ROC curve illustrates graphically the performance of a classifier, under different cut-off values showing the trade-off between sensitivity and specificity. More specifically, in a ROC curve the True Positive rate (Sensitivity) is plotted against the False Positive rate (1-Specificity) for increasing values of the score cut-off of a binary classifier. The area under the (ROC) curve (AUC) is a measure of accuracy: The closer the curve follows the left-hand and the top border of the ROC space, the more accurate the classification model. A perfect classifier (AUC=1) would predict correctly all the True Positives

(Sensitivity = 1) giving no False Positives (Specificity = 1). A classifier that makes a random guess would result in an AUC of 0.5.

4.2.10 Cross-Validation

Cross-validation is a method for estimating generalization error based on resampling. It provides an indication of how well the classifier performs in making new predictions for previously unseen data. Some of the data is removed prior to the training; after the training, the data that was removed is used to test the performance of the learned model on unseen data. That involves the division of the data into *m* subsets of (approximately) equal size; then training the method *m* times, each time leaving out one of the subsets from the training and using that (omitted) subset for testing; in this analysis I pursued a five-fold cross validation approach dividing each dataset into five subsets.

4.3 Results

Implementing a whole-genome based comparative analysis between 37 reference strains of three different genera and 12 outgroup genomes, a training set of 668 regions was built (Table 4.4). This training set, that includes both putative GIs (differentiated from gene loss events by a maximum parsimony approach) and randomly sampled regions (non-GIs), was used to study the structural variation of GIs and quantify the contribution of each feature to a GI structural model. As a starting point, GI structural models for each genus were built implementing the RVM method (Tipping, 2001). In addition, in order to capture potential genus-specific signatures as well as to evaluate the ability of the RVM models to make generalizations on unseen data from different lineages, cross-genus GI models were built using different mixtures of training and test datasets. Overall 11 structural GI models were built and analyzed (Table 4.5); the structural details of each model are discussed in detail in the following sections.

Table 4.5: A list of 11 structural GI models, built based on different training sets: 1) 421 *Salmonella* regions, 2) 107 *Streptococcus* regions, 3) 140 *Staphylococcus* regions (including 2 regions overlapping rRNA operons), 4) 138 *Staphylococcus* regions (no rRNA operons), 5) 245 *Staphylococcus-Streptococcus* regions, 6) 559 *Salmonella-Staphylococcus* regions, 7) 528 *Salmonella-Streptococcus* regions, 8) 666 *Salmonella-Staphylococcus-Streptococcus regions*. Training sets 9-11 include three subsets of approximately 140 different *Salmonella*-specific regions combined with the *Staphylococcus* and *Streptococcus*-specific regions. Each model, expressed through function S_i, is the weighted sum of eight basis functions (structural features): The Interpolated Variable Order Motif (IVOM) score that measures both low and high order compositional deviation from the backbone composition and is expressed as the relative entropy between the query and the genome-backbone (variable order) compositional distribution, the insertion point (INSP) of each genomic region; two states were (binary) evaluated: insertion point within a CDS locus (disrupting the corresponding CDS) or insertion within an intergenic part of the chromosome, the size (SIZE) of each genomic region, the gene density (DENS = number of genes per kb) of each region, presence or absence (binary) of direct/inverted repeats (REPEATS) flanking the boundaries of each genomic region, presence or absence (binary) of integrase and/or integrase-like (INT) protein domains, presence or absence (binary) of phage-related protein domains (PHAGE), presence or absence (binary) of non-coding RNA (RNA) in the proximity of each region.

1) Si = -0.764 + 6.203 (x)**IVOM** + 0.000(x)**INSP** + -4.956(x)**SIZE** + 0.000(x)**DENS** + 0.635(x)REPEATS + 0.995(x)INT + 2.086(x)**PHAGE** + 1.968(x)**RNA**
2) Si = -2.978 + 4.151 (x)**IVOM** + 3.219(x)**INSP** + 0.000(x)**SIZE** + 0.000(x)**DENS** + 2.185(x)REPEATS + 3.351(x)INT + 0.000(x)**PHAGE** + 0.000(x)**RNA**
3) Si = -0.005 + 0.000 (x)**IVOM** + 0.000(x)**INSP** + -4.324(x)**SIZE** + 0.000(x)**DENS** + 0.360(x)REPEATS + 1.303(x)INT + 3.995(x)**PHAGE** + 0.000(x)**RNA**
4) Si = -4.583 +12.752 (x)**IVOM** + 0.000(x)**INSP** + -2.843(x)**SIZE** + 2.486(x)**DENS** + 0.000(x)REPEATS + 1.552(x)INT + 2.157(x)**PHAGE** + 0.000(x)**RNA**
5) Si = -1.544 + 3.756 (x)**IVOM** + 2.842(x)**INSP** + -2.583(x)**SIZE** + 0.000(x)**DENS** + 1.297(x)REPEATS + 1.892(x)INT + 2.554(x)**PHAGE** + 0.000(x)**RNA**
6) Si = -0.923 + 6.528 (x)**IVOM** + 0.000(x)**INSP** + -4.462(x)**SIZE** + 0.000(x)**DENS** + 0.771(x)REPEATS + 1.404(x)INT + 2.441(x)**PHAGE** + 1.159(x)**RNA**
7) Si = -0.763 + 4.330 (x)**IVOM** + 2.516(x)**INSP** + -4.941(x)**SIZE** + 0.000(x)**DENS** + 1.030(x)REPEATS + 1.630(x)INT + 2.027(x)**PHAGE** + 1.842(x)**RNA**
8) Si = -0.879 + 4.659 (x)**IVOM** + 2.795(x)**INSP** + -4.434(x)**SIZE** + 0.000(x)**DENS** + 0.897(x)REPEATS + 1.553(x)INT + 2.433(x)**PHAGE** + 1.319(x)**RNA**
9) Si = -1.293 + 5.285 (x)**IVOM** + 3.072(x)**INSP** + -3.914(x)**SIZE** + 0.000(x)**DENS** + 1.007(x)REPEATS + 1.668(x)INT + 2.847(x)**PHAGE** + 0.000(x)**RNA**
10) Si = -1.057 + 4.234 (x)**IVOM** + 3.003(x)**INSP** + -3.396(x)**SIZE** + 0.000(x)**DENS** + 0.927(x)REPEATS + 1.722(x)INT + 1.664(x)**PHAGE** + 1.539(x)**RNA**
11) Si = -1.627 + 3.552 (x)**IVOM** + 0.000(x)**INSP** + -4.138(x)**SIZE** + 0.727(x)**DENS** + 1.449(x)REPEATS + 1.728(x)INT + 3.685(x)**PHAGE** + 0.000(x)**RNA**

4.3.1 GI structural models

Each GI model (Table 4.5) is the weighted sum of K basis functions, where K denotes the number of features used to describe a GI structure. In this analysis, eight structural features were used (IVOM, INTEGRASE, PHAGE, SIZE, RNA, DENSITY, REPEATS and INSP). Each feature is evaluated during the training process of the RVM, and its overall contribution to the structural model is expressed by the corresponding feature weight.

For example a feature frequently related to GI structures (but absent from randomly sampled regions), receives typically higher weight (i.e. contributes more to the model) compared to a feature found equally frequently both in GIs and non-GIs; in the latter case the feature weight will be lower or even zero (i.e. feature ignored).

In the following section the contribution of each structural feature to the corresponding GI model is evaluated through a function (*R*) that quantifies the relative feature importance, rather than the actual feature weight (*w*). Briefly the importance *R* of each feature is expressed as the product of the corresponding weight *w* and the corresponding standard deviation (*SD*) of the feature values in the training set.

I prefer to assess the feature contribution to the model, through the *R* rather than the *w* value, because *R* takes into account the variability of the dataset, normalizing the values with the corresponding *SD*. Consider for example two different structural features; the values of the first feature in the training set have higher dispersion relative to the values of the second feature. If both features have comparable *w* values, then the first feature will be more important than the second one meaning that, because of its variability, it is more informative than the second feature. Based on that, it is not unusual for some features to have a very high value of *w* but a low value of *R*.

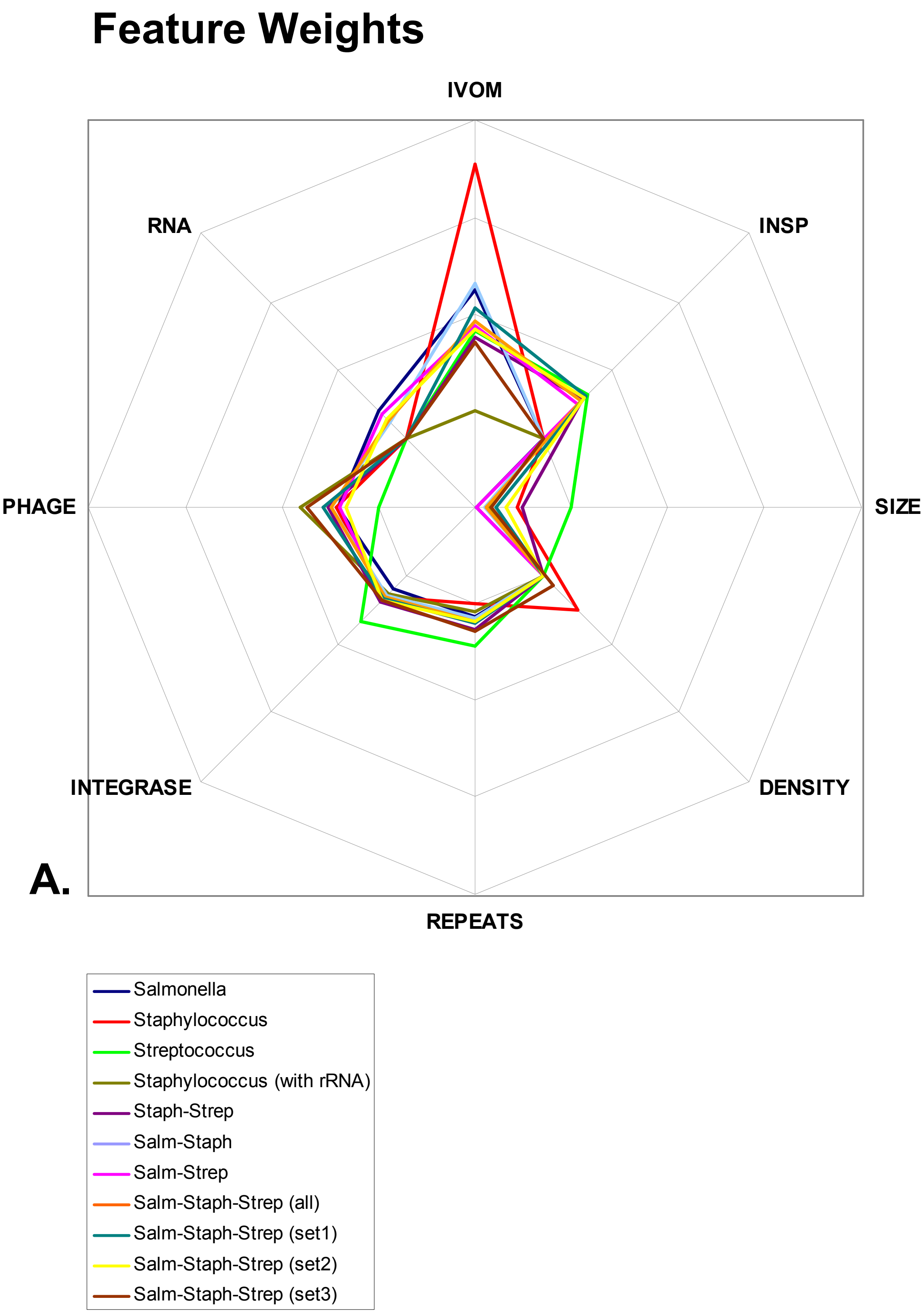
4.3.1.1 Genus-specific

4.3.1.1.1 Salmonella

Using 211 positive (putative GIs) and 210 negative (randomly sampled) examples (Table 4.4, Appendix E) a model that describes the structure of GIs present in the *Salmonella* lineage was built (Figure 4.9, Table 4.5). Overall under this model, the most “important” (informative) features are: IVOM (R_{IVOM} = 0.65), SIZE (R_{SIZE} = 0.38), PHAGE (R_{PHAGE} = 0.27), RNA (R_{RNA} = 0.26), INTEGRASE (R_{INT} = 0.13) and REPEATS ($R_{REPEATS}$ = 0.085); in this model, the DENSITY and INSP features were ignored. Note that the SIZE feature received a negative weight (W_{SIZE} = -4.956); the same applies for all the other GI models apart from the one built based on the *Streptococcus* dataset (see below) in which the SIZE feature is completely ignored (W_{SIZE} = 0). A more detailed discussion about the negative weight of the SIZE feature is provided in section 4.4.

In order to investigate further the structural variation of GIs, in terms of preference for insertion within a specific locus and for different type of repeats flanking their boundaries, the RNA feature was further subdivided into tRNA and misc_RNA (any kind of non-coding RNA apart from tRNA) features; the same applies for the REPEATS feature that was further divided into DRs and IRs. The relative "importance" of those six structural features was evaluated pair-wise: (RNA, INSP), (tRNA, misc_RNA) and (DRs, IRs) (Figure 4.10).

The results show that for GIs present in *Salmonella* chromosomes, insertion within an RNA (R_{RNA} = 0.72) rather than a CDS locus (R_{INSP} = 0.0) is the most informative feature when classifying unknown regions as GIs. In the case of RNA locus, insertion of GIs within a tRNA (R_{tRNA} = 0.60) is slightly more informative than insertion within a misc_RNA locus ($R_{miscRNA}$ = 0.51). In terms of type of repeats flanking the boundaries of GIs, DRs (R_{DRs} = 0.63) rather than IRs (R_{IRs} = 0.0) is the most informative feature.


Feature Weights
IVOM
INSP
SIZE
DENSITY
REPEATS
INTEGRASE
PHAGE
RNA
A.
Salmonella
Staphylococcus
Streptococcus
Staphylococcus (with rRNA)
Staph-Strep
Salm-Staph
Salm-Strep
Salm-Staph-Strep (all)
Salm-Staph-Strep (set1)
Salm-Staph-Strep (set2)
Salm-Staph-Strep (set3)

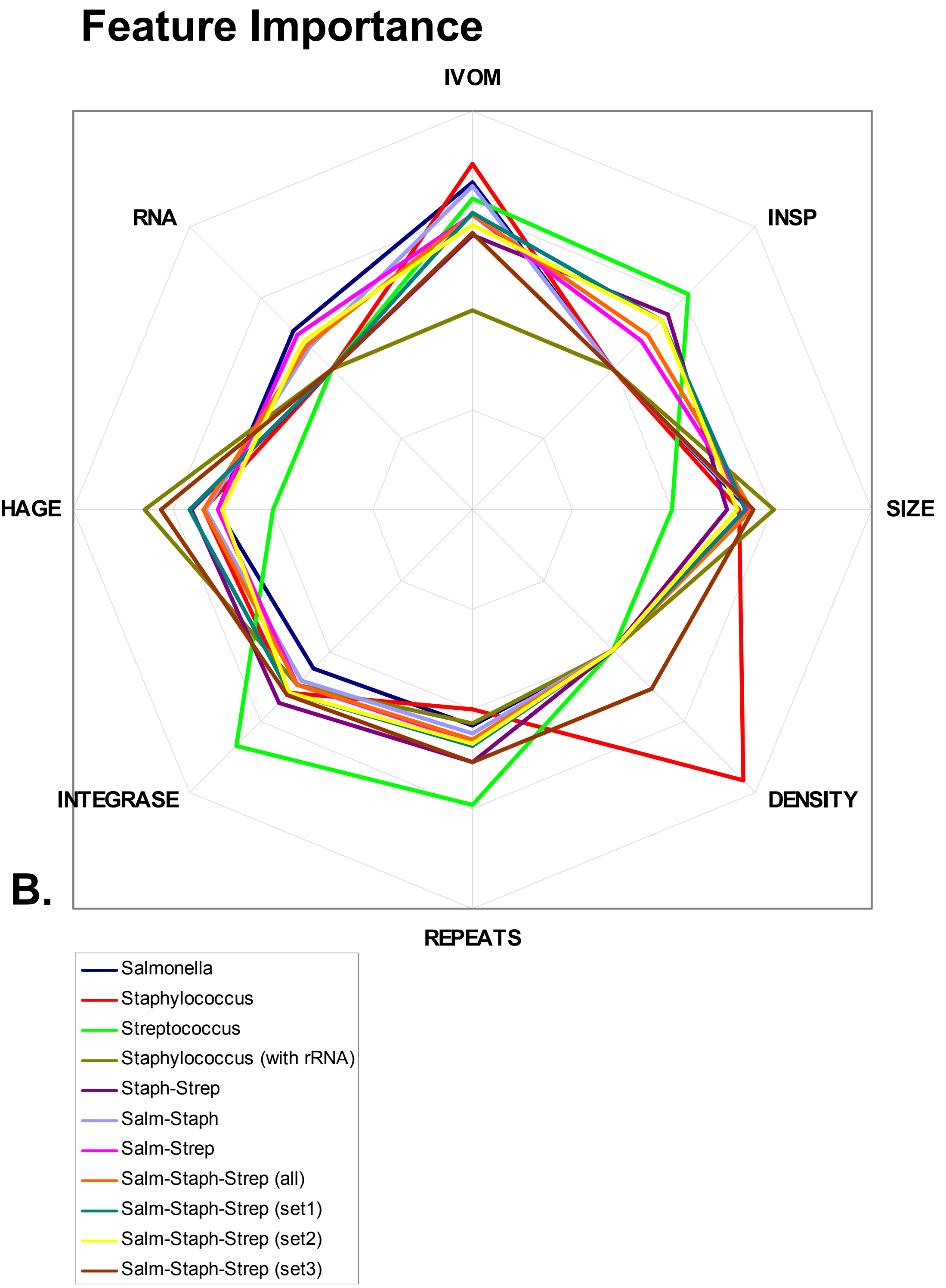


Figure 4.9: Radar diagram illustrating the feature weight (A) and "importance" (B) of the eight structural features under different GI models, based on 11 training datasets. Features: IVOM (feature composition), INSP (insertion point), SIZE (the size of each region), DENSITY (gene density), REPEATS (repeats flanking each region), INTEGRASE (integrase-like protein domains), PHAGE (phage-related protein domains), RNA (non-coding RNAs). Each apex in the octagon-like diagram corresponds to one of the eight structural features, while the height of the plot at the corresponding apex is indicative of the actual feature weight (A) or importance (B).

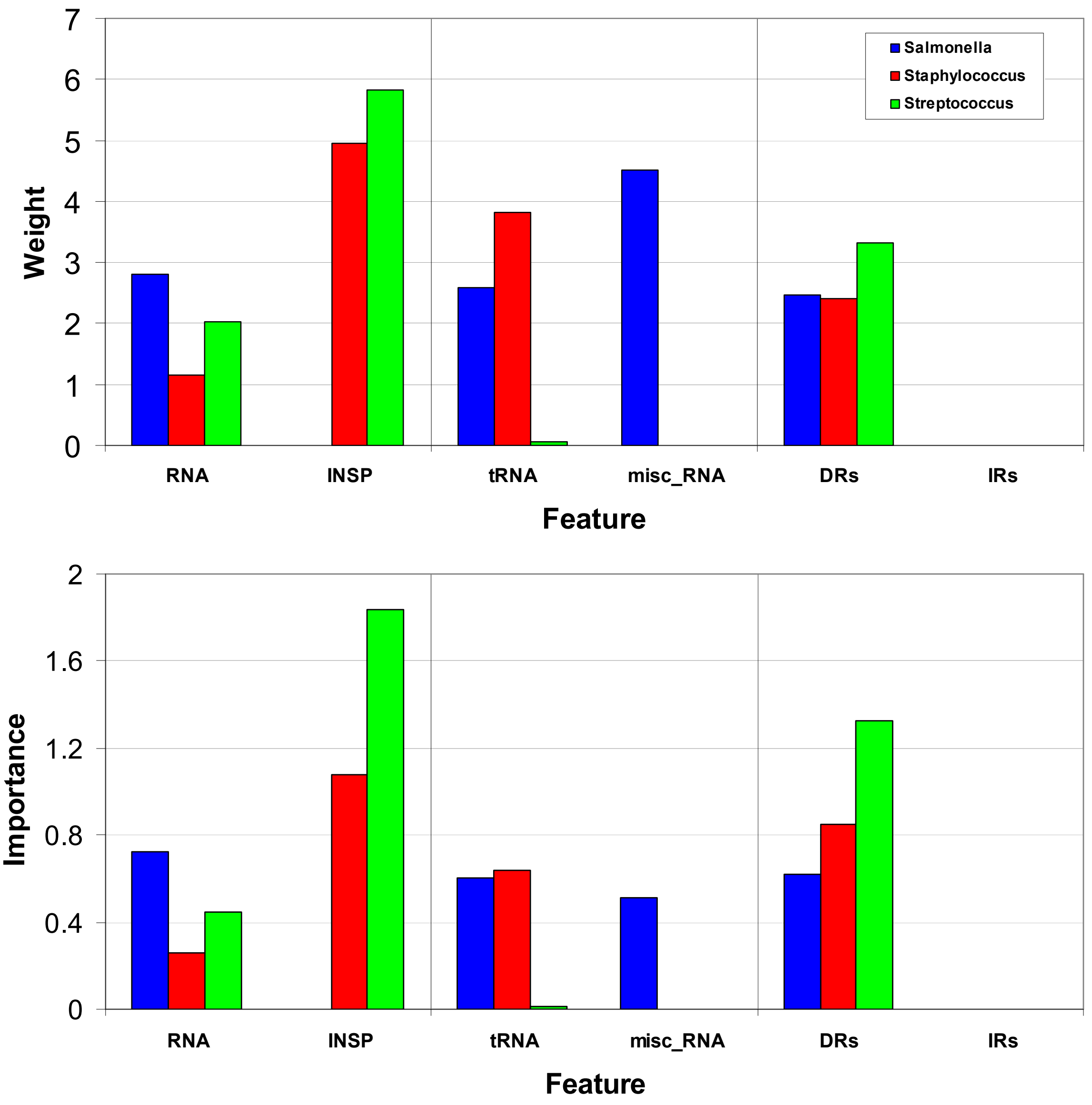


Figure 4.10: Bar chart illustrating the feature weight (top) and "importance" (bottom) of six structural features (evaluated pair-wise), under three different dual-featured GI models, trained on: *Salmonella*, *Staphylococcus* and *Streptococcus*-specific regions respectively. Features: [RNA, INSP], [tRNA, misc_RNA], [DRs, IRs].

4.3.1.1.2 Staphylococcus

The model that describes the structure of GIs present in *Staphylococcus* genomes was built based on 66 putative GIs and 74 randomly sampled regions (Table 4.4, Appendix F). Overall under this model, the most predictive informative structural features are: PHAGE (R_{PHAGE} = 0.65), SIZE (R_{SIZE} = 0.51), INTEGRASE (R_{INT} = 0.25) and REPEATS ($R_{REPEATS}$ = 0.07); the remaining features were ignored. Two randomly sampled regions had the two highest IVOM scores in this dataset of 140 examples.

These two regions (*Staph.Epid_RP62.non.12* and *Staph.MRSA252.non.21* in Appendix F) overlap with two rRNA operons. rRNA operons often deviate compositionally from the genome backbone composition mainly due to specific, well-preserved functional constraints rather than their horizontal origin (Vernikos and Parkhill, 2006; Vernikos *et al.*, 2007). Excluding those two regions and repeating the training, the GI model assigned weights to previously ignored features and modified each weight overall: DENSITY ($R_{DENS} = 0.92$), IVOM ($R_{IVOM} = 0.74$), PHAGE ($R_{PHAGE} = 0.35$), SIZE ($R_{SIZE} = 0.34$), INTEGRASE ($R_{INT} = 0.30$); the rest of the features were ignored (Figure 4.9, Table 4.5).

When GI models are trained (pair-wise) only on selected structural features, insertion within a CDS locus ($R_{INSP} = 1.1$) is more informative than insertion within an RNA locus ($R_{RNA} = 0.26$). Between the different type of non-coding RNAs, insertion within a tRNA ($R_{tRNA} = 0.64$) rather than a misc_RNA ($R_{miscRNA} = 0.0$) is the most informative feature. In terms of type of repeats, again DRs is the most informative feature ($R_{DRs} = 0.85$, $R_{IRs} = 0.0$) (Figure 4.10). It is worth noting that under these three partial GI models, some previously ignored (under the full GI model above) structural features, i.e. RNA, INSP and REPEATS, are now informative predictors, further suggesting those features were redundant predictors under the full model in which all eight features were evaluated.

4.3.1.1.3 Streptococcus

The training set for the *Streptococcus* genus consists of 54 and 53 positive and negative control examples respectively (Table 4.4, Appendix G). Under this model, the most informative GI structural features are: INTEGRASE ($R_{INT} = 0.67$), IVOM ($R_{IVOM} = 0.56$), INSP ($R_{INSP} = 0.53$) and REPEATS ($R_{REPEATS} = 0.48$). The remaining four features were ignored (Figure 4.9, Table 4.5), giving the highest sparsity GI model that exploits only four (of the eight) basis functions.

In terms of pair-wise evaluation of selected structural features (Figure 4.10), GIs present in *Streptococcus* genomes follow the same pattern of insertion point preference with the *Staphylococcus* GIs, i.e.

insertion within a CDS locus (R_{INSP} = 1.84) is more informative than insertion within an RNA locus (R_{RNA} = 0.45); the same applies for the type of non-coding RNAs (R_{tRNA} = 0.013, $R_{miscRNA}$ = 0.0) and the type of repeats (R_{DRs} = 1.33, R_{IRs} = 0.0).

4.3.1.2 Cross-genus

4.3.1.2.1 Staphylococcus-Streptococcus

Combining 138 *Staphylococcus* and 107 *Streptococcus* genomic regions, a dataset of 245 (Gram positive) examples was built in order to study the structural variation of GIs across genus/species boundaries. In this cross-genus GI model the most informative features are: PHAGE (R_{PHAGE} = 0.41), INSP (R_{INSP} = 0.39), IVOM (R_{IVOM} = 0.374), INTEGRASE (R_{INT} = 0.37), SIZE (R_{SIZE} = 0.272) and REPEATS ($R_{REPEATS}$ = 0.270); the remaining structural features were ignored (Figure 4.9, Figure 4.11 and Table 4.5).

4.3.1.2.2 Salmonella-Staphylococcus

A cross-genus dataset of 421 *Salmonella* and 138 *Staphylococcus* specific regions was built and used to train a GI structural model; under this model the most informative features, are: IVOM (R_{IVOM} = 0.62), SIZE (R_{SIZE} = 0.40), PHAGE (R_{PHAGE} = 0.34), INTEGRASE (R_{INT} = 0.21), RNA (R_{RNA} = 0.15) and REPEATS ($R_{REPEATS}$ = 0.12). The remaining features were ignored (Figure 4.9, Figure 4.11 and Table 4.5).

4.3.1.2.3 Salmonella-Streptococcus

Combining the *Salmonella* and *Streptococcus*-specific regions, a dataset of 528 examples was built. Under this cross-genus GI model, the most informative structural features are: IVOM (R_{IVOM} = 0.48), SIZE (R_{SIZE} = 0.39), PHAGE (R_{PHAGE} = 0.28), INTEGRASE (R_{INT} = 0.25), RNA (R_{RNA} = 0.24), INSP (R_{INSP} = 0.20) and REPEATS ($R_{REPEATS}$ = 0.16) (Figure 4.9, Figure 4.11 and Table 4.5).

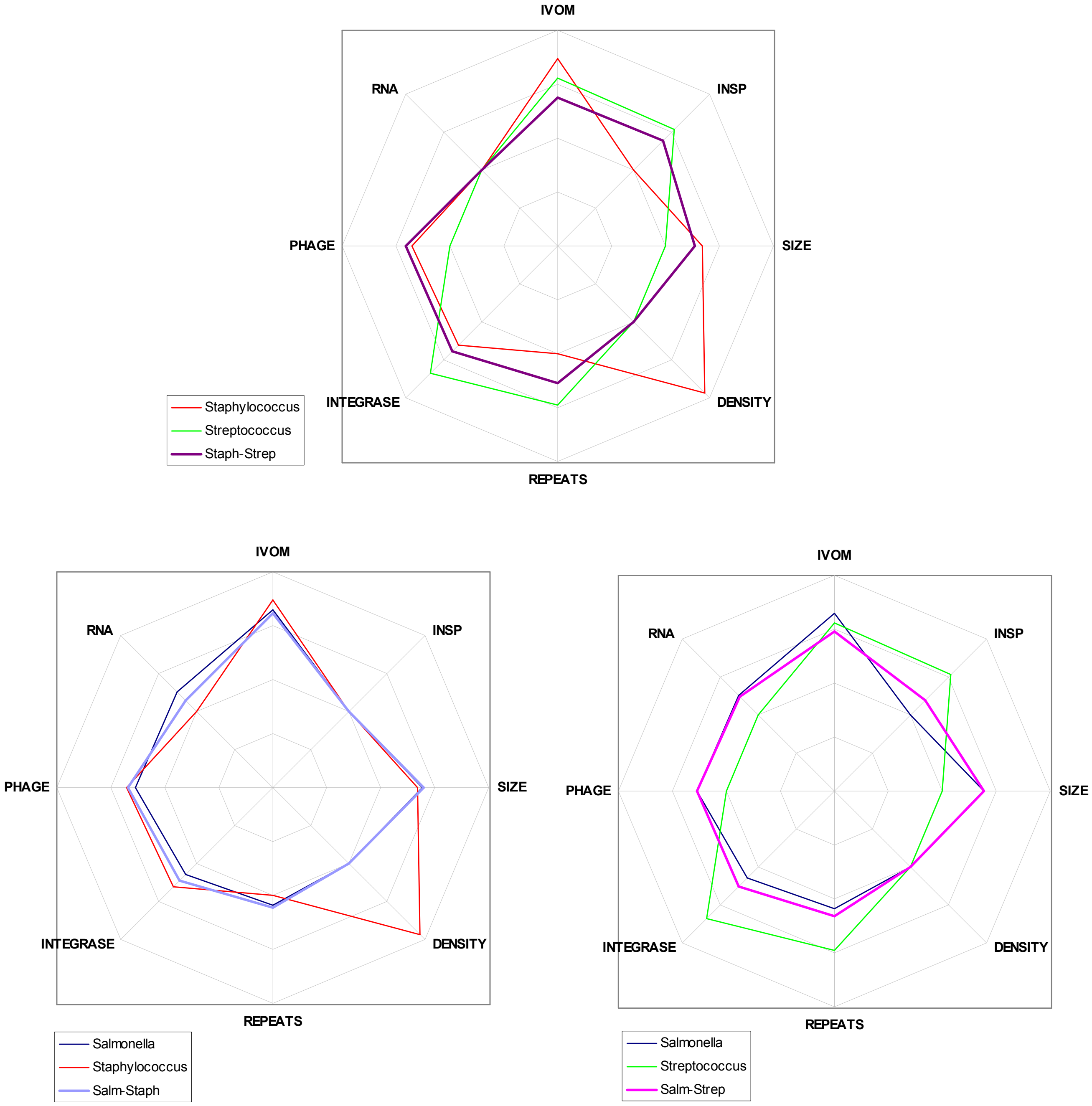


Figure 4.11: Radar diagram illustrating the "importance" of eight structural features under different genus-specific and cross-genus (2 genera) GI models: *Staphylococcus-Streptococcus* (top), *Salmonella-Staphylococcus* (bottom-left) and *Salmonella-Streptococcus* (bottom-right).

4.3.1.2.4 All three genera

In order to study the structural variation of GIs across the three genera, taking into account the difference in the dimensionality of the three genus-specific datasets (421 *Salmonella*, 138 *Staphylococcus* and 107 *Streptococcus*-specific regions), two different approaches were followed: In the first approach a training set ($N = 666$) was built combining the full *Salmonella* and the other two genus-specific datasets; in the second approach the *Salmonella* dataset was split into three subsets ($N \approx 140$ each) each of which was combined with the full *Staphylococcus* and *Streptococcus* datasets giving three training sets (namely set1, set2 and set3) of approximately 385 examples each; in each set the three different genera contribute approximately the same number of examples.

Training the RVM on the full ($N = 666$) cross-genus dataset (all), the most informative GI structural features are: IVOM ($R_{IVOM} = 0.48$), SIZE ($R_{SIZE} = 0.39$), PHAGE ($R_{PHAGE} = 0.35$), INTEGRASE ($R_{INT} = 0.25$), INSP ($R_{INSP} = 0.24$), RNA ($R_{RNA} = 0.17$) and REPEATS ($R_{REPEATS} = 0.15$) (Figure 4.9, Figure 4.12 and Table 4.5).

Using each of the three smaller datasets (set 1-3) to train the RVM, the most informative features under the three GI models are (for each model, respectively): IVOM [$R_{IVOM} = 0.49, 0.43, 0.39$], PHAGE [$R_{PHAGE} = 0.42, 0.25, 0.56$], SIZE [$R_{SIZE} = 0.37, 0.32, 0.41$], INTEGRASE [$R_{INT} = 0.29, 0.30, 0.31$], INSP [$R_{INSP} = 0.34, 0.34, 0.0$], REPEATS [$R_{REPEATS} = 0.19, 0.17, 0.27$], DENSITY [$R_{DENS} = 0.0, 0.0, 0.26$] and RNA [$R_{RNA} = 0.0, 0.19, 0.0$]. Based on the four RVM trainings (all, set1, set2 and set3), the four models that capture the structural variation of GIs across the three genera have converged over fairly similar GI structures, with the exception of genus-specific features, i.e. the RNA feature for *Salmonella*, the INSP feature for *Streptococcus* and the DENSITY feature for *Staphylococcus* (see discussion and Figure 4.12).

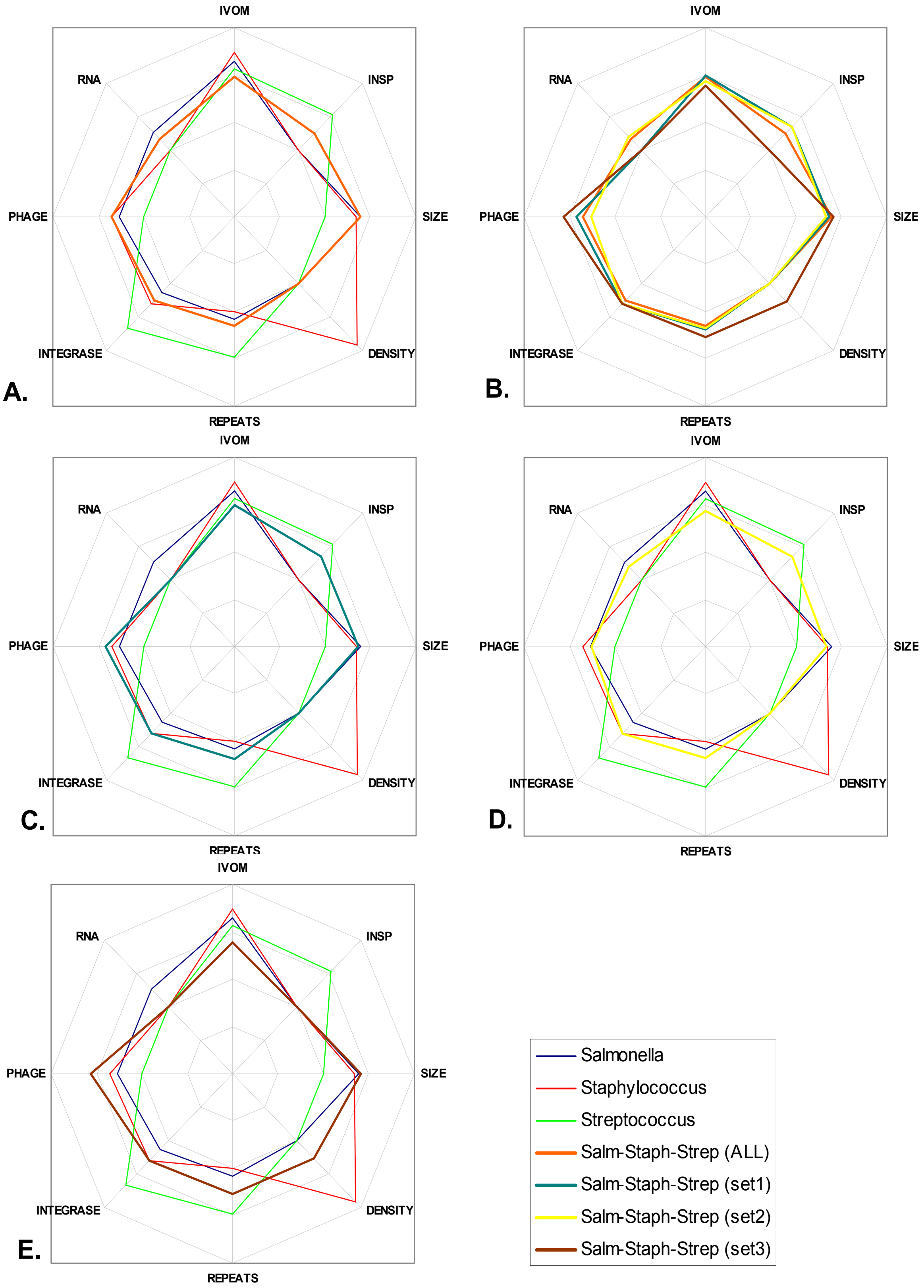


Figure 4.12: Radar diagram illustrating the "importance" of eight structural features under different genus-specific and cross-genus (3 genera) GI models: The *Salmonella* complete dataset (A), set1 (C), set2 (D) and set3 (E) are combined with the complete *Staphylococcus* and *Streptococcus* training datasets. The above four cross-genus GI models are shown together in the same diagram (B) for ease of comparison.

4.3.2 Prediction accuracy

In order to evaluate the prediction accuracy of the RVM classifier each dataset was split into five smaller subsets of approximately the same size and the RVM was trained on the 4/5 of the dataset and tested on the remaining 1/5; this process was repeated five times (for each dataset), classifying each time non overlapping test sets (five-fold cross validation). Moreover, in order to evaluate further the generalization properties of each GI structural model I performed six "genus-blind" cross validations, training a model only on examples of one genus and testing it on examples of the other two. This blind test was performed in order to investigate how different genus-specific models would perform in classifying regions from unknown taxa. In order to estimate the relative accuracy and generalization properties of each model, I performed a ROC curve analysis, evaluating the AUC.

Overall, throughout the 10 five-fold cross validations the different GI models made good generalizations on unseen data, classifying with high accuracy (AUC: 0.82-0.94) unknown examples (GIs and non-GIs) (Figure 4.13 and Appendix I). Between the three different genus-specific GI models, the *Streptococcus* (Strep) model is the most accurate, followed by the *Salmonella* (Salm) and the *Staphylococcus* (Staph) models (AUC: 0.94, 0.83 and 0.82 respectively).

Between the three different GI models, trained on a mixture of examples from two different genera, the Staph-Strep (Gram-positive) model is the most accurate, followed by the Salm-Staph and the Salm-Strep models (AUC: 0.88, 0.85 and 0.84 respectively). Overall the Salm-Staph model performs better than the corresponding two genus-specific Salm and Staph models (Figure 4.13); similarly the Salm-Strep and Staph-Strep models are overall more accurate than the Salm and Staph models respectively.

GI models trained on a mixture of examples from all the three genera show fairly similar performance (AUC: 0.84-0.88). More specifically the three GI models trained on datasets in which the three genera are equally represented (i.e. set1, set2 and set3), perform equally well (AUC: 0.87, 0.86, 0.88) and slightly better than the model trained on all ($N = 666$) examples (AUC: 0.84), underlining the increased sparsity property of the RVM method.

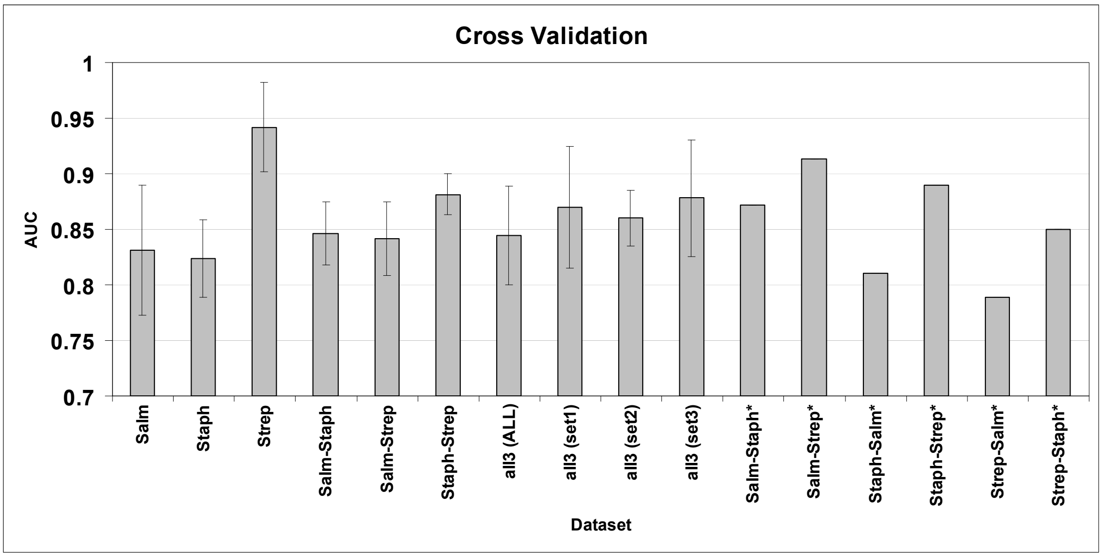


Figure 4.13: A Bar chart illustrating the average performance of the RVM classifier, under different training and test datasets. Each dataset is split into five subsets of approximately equal size; four of the five subsets are used to train an RVM model while the omitted subset is used to test the performance of this model. This process is repeated five times on non overlapping test sets (five-fold cross-validation). The performance of the RVM models was evaluated through the receiver operating characteristic (ROC) curve. The average value and ±1 S.D. of the AUC over the five subsets of the five–fold cross-validation is calculated for the first ten datasets. The AUC values for the last six datasets (with the asterisk) summarize the performance of the RVM, when trained on the whole dataset of the first genus and tested on the whole dataset of the second genus, e.g. for the Salm-Strep* dataset, the 421 *Salmonella*-specific regions were used to train a GI model that was tested on the 107 *Streptococcus*-specific regions.

The evaluation of the three genus-specific GI models, under a "genus-blind" cross-validation framework indicates that the RVM classifier can very accurately predict unseen examples from close or distantly related genera that are not included in the training set (Figure 4.13). More specifically, using the Salm model to classify *Staphylococcus* and *Streptococcus*-specific regions can be overall more (AUC: 0.87 vs 0.82) or similarly (AUC: 0.91 vs 0.94) accurate compared to the corresponding genus-specific models, respectively. The Staph model shows high accuracy (AUC: 0.81 and 0.89) in classifying *Salmonella* and *Streptococcus*-specific regions respectively; overall this model is slightly less accurate than the corresponding genus-specific models (AUC: 0.83 and 0.94 respectively). Similar conclusions can be drawn for the performance (AUC: 0.79 and 0.85) of the Strep model when classifying *Salmonella* and *Staphylococcus*-specific regions respectively. Again this model is more accurate in classifying *Staphylococcus*-specific regions than the Staph model (AUC: 0.85 and 0.82 respectively), but is less accurate in classifying *Salmonella*-specific regions than the Salm model (AUC: 0.79 and 0.83 respectively).

4.4 Discussion

The aim of this analysis was to study the structural variation of GIs, quantifying and modelling the "importance" of genetic features that can be informative when classifying GIs and non-GI regions, enabling a quantitative rather than a descriptive definition of the actual GI structure to be proposed. The basic principle behind this analysis is a hypothesis-free framework, in which no *a priori* assumptions are made about the GI structure.

Implementing a machine learning oriented approach, genomic regions (both GIs and randomly sampled regions) from 37 chromosomes of three different genera were exploited in order to build genus-specific as well as cross-genus GI structural models. Overall the three genus-specific GI models show both core and variable structural features with distinct genus-specific signatures. For example, the IVOM and INT features are informative in all three GI models; on the other hand the RNA, INSP and DENSITY features are *Salmonella*, *Streptococcus* and *Staphylococcus*-specific features respectively (Figure 4.9, Table 4.5).

Moreover, in the Strep model apart from the INSP feature, the INT and REPEATS features contribute more to the overall structural model compared to the other two genus-specific models, while the SIZE and

PHAGE features seem to be informative only in the Salm and Staph structural GI models.

Care should be taken when interpreting the “importance” of each of the eight structural features. In this analysis the GI models are built by evaluating how informative each feature is, taking into account cross-feature relationships and information redundancy. Mapping the eight features in a high dimensional space enables cross-feature relationships to be captured: if some features contain information present already in other features (redundant information) then for the sake of model-sparsity those features (basis functions) will be ignored by setting their weight to zero value. That however does not necessarily mean that those features may not be informative when seen on their own, i.e. in single-featured GI models (Figure 4.14).

Therefore it is more intuitive to interpret the “importance” of each feature as its relative (in combination with the rest of the features) rather than its absolute “importance” under a GI model. For example in the Strep model, the PHAGE feature is ignored when building a model evaluating all the eight features. However when the PHAGE feature is evaluated in a single-featured model, it turns out to be the second most informative feature (Figure 4.14); this observation is in line with previous studies showing the impact of bacteriophage elements in the evolution of Streptococci (Banks *et al.*, 2003; Broudy *et al.*, 2001; Fischetti, 2007). Perhaps some of the information in the PHAGE feature is already present in some other features (e.g. phage integrase protein domains of the INTEGRASE feature) making the PHAGE feature a redundant predictor under a multi-featured GI model.

The same observation applies for the SIZE feature. In a multi-featured model, SIZE is a very informative feature for the Salm and Staph models; however in a single-featured model (i.e. evaluated on its own) the SIZE feature is ignored in all three genera models (Figure 4.14). This further suggests that in multi-featured models some structural features correlate with the SIZE feature. Moreover throughout this analysis, the

SIZE feature received a negative weight in all GI models apart from the Strep model. Generally, during the training process some features may correlate positively or even negatively (e.g. the SIZE feature) with class membership. This does not necessarily suggest that true GIs are always of small size, but rather that the SIZE feature is negatively correlated with some other features.

This observation becomes much clearer in the case of the Strep model in which both the SIZE and the PHAGE features received a weight of zero. However in the other 10 models, the same two features received a negative and a positive weight respectively (Table 4.5). Perhaps the SIZE feature is inversely correlated with the PHAGE feature, suggesting that GIs of phage origin are on average larger than GIs of different origin. Indeed for the *Salmonella* and the *Staphylococcus* dataset the average size of GIs of phage origin is significantly larger than the size of GIs of different origin (*p*-value = 1.17 x 10^{-7} and 1 x 10^{-5} respectively). In order for the reverse correlation of the SIZE and some features to be captured in the model, the SIZE feature has to have a negative weight.

The fact that in the Strep GI model, three structural features (i.e. INTEGRASE, REPEATS and INSP) are unusually highly informative (relative to the other two genus-specific models) while at the same time those three features are frequently involved in the mobilization of genomic DNA (i.e. integration/excision), leaves open the possibility of a GI model that is capturing a distinct *Streptococcus*-specific mechanism of genetic element integration preferably within CDS loci.

It is worth noting that the Strep GI model shows the highest sparsity exploiting only half of the basis functions (4 out of the 8 structural features), compared to the Staph (5 out of 8) and the Salm (6 out of 8) GI models, proposing a much simpler structural model, in order to describe GIs in the *Streptococcus* lineage (Table 4.5); this observation is in line with the outstanding classification accuracy of the Strep GI model (AUC: 0.94 – Figure 4.13).

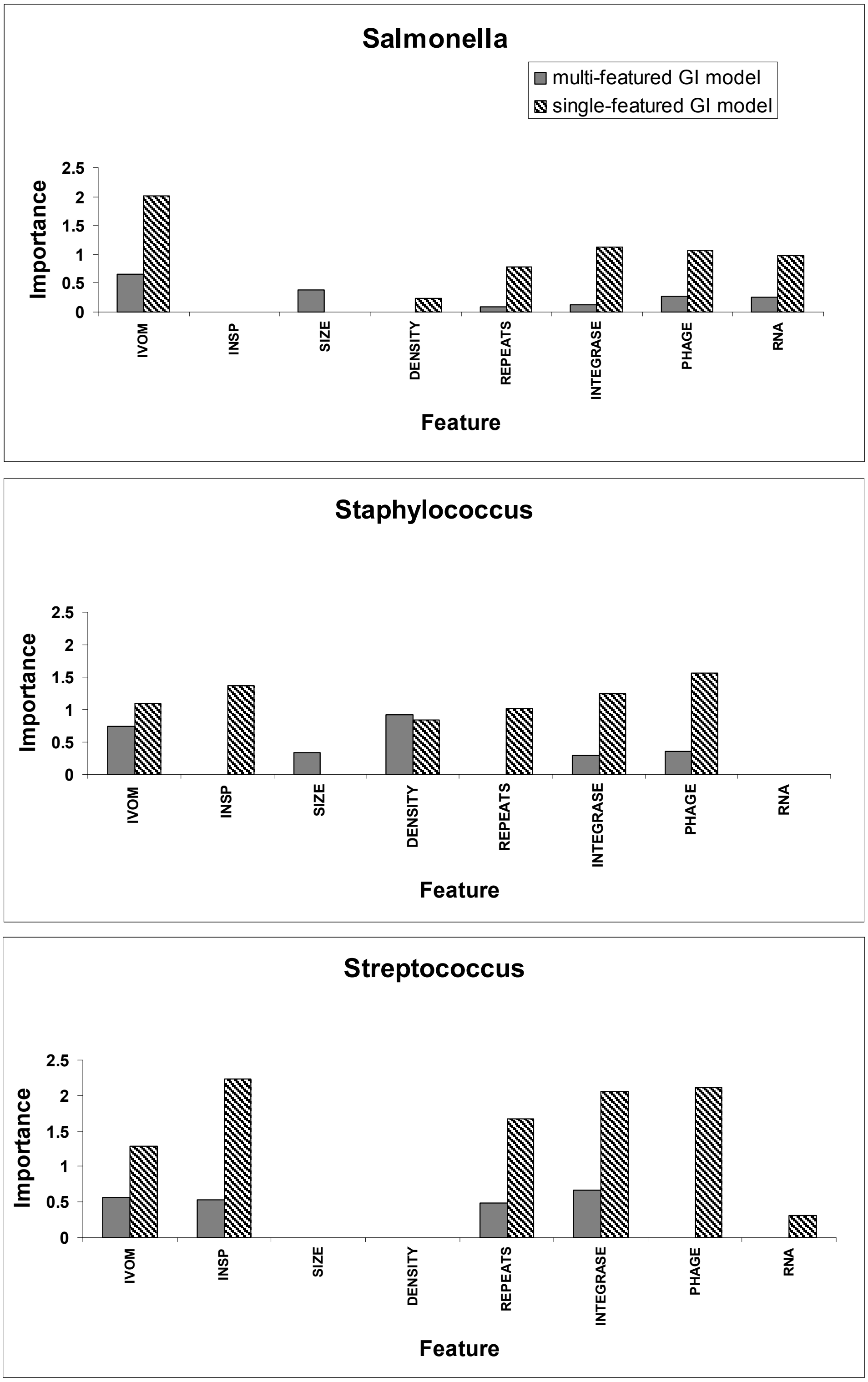


Figure 4.14: Bar chart illustrating the "importance" of eight structural features under a *Salmonella*, *Staphylococcus* and *Streptococcus* GI model. Grey-coloured bars show the "importance" of every feature, in a (multi-featured) GI model in which all eight features are taken into account (relative importance). Gradient black-coloured bars show the "importance" of each feature, in a (single-featured) GI model with only one structural feature evaluated each time (absolute importance).

The distinct structural feature with the highest contribution to the Staph GI model, while being ignored in the other two genus-specific models, is the DENSITY feature (Figure 4.9, Figure 4.15). Overall the average gene density of GIs present in *Staphylococcus* genomes, is significantly (*p*-value = 1.4 x 10^{-6}) higher than that of randomly sampled regions; in *Salmonella* and *Streptococcus* lineages this feature is less informative when predicting GIs (*p*-value = 1.7 x 10^{-3} and 1.3 x 10^{-2} respectively).

Again, it is possible that this genus-specific GI model is capturing the underlying origin of GIs present in *Staphylococcus* genomes, 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. For example, the staphylococcal pathogenicity islands (SaPIs) represent members of a structurally very well conserved family of phage-related GIs (Novick and Subedi, 2007); the structure of SaPIs is discussed in section 1.2.1 of chapter 1.

Increasing further the resolution within certain GI structural features (i.e. insertion within a CDS or RNA locus, tRNA or misc_RNA and DRs or IRs), training the RVM pair-wise only on those selected features, the genus-specific signatures of each model become more evident (Figure 4.10). For the prediction of GIs in the *Salmonella* lineage, integration within a non-coding RNA locus is much more informative than within a CDS locus.

The opposite observation can be made for the *Staphylococcus* and *Streptococcus* models. In the case of non-coding RNA, insertion within a tRNA or a misc_RNA locus are almost equally informative for the prediction of *Salmonella* GIs, while in *Staphylococcus* and *Streptococcus* lineages, insertion within a tRNA locus is much and slightly more informative than insertion within a misc_RNA respectively. In all three genera the predominant type of repeats associated with GIs are DRs.

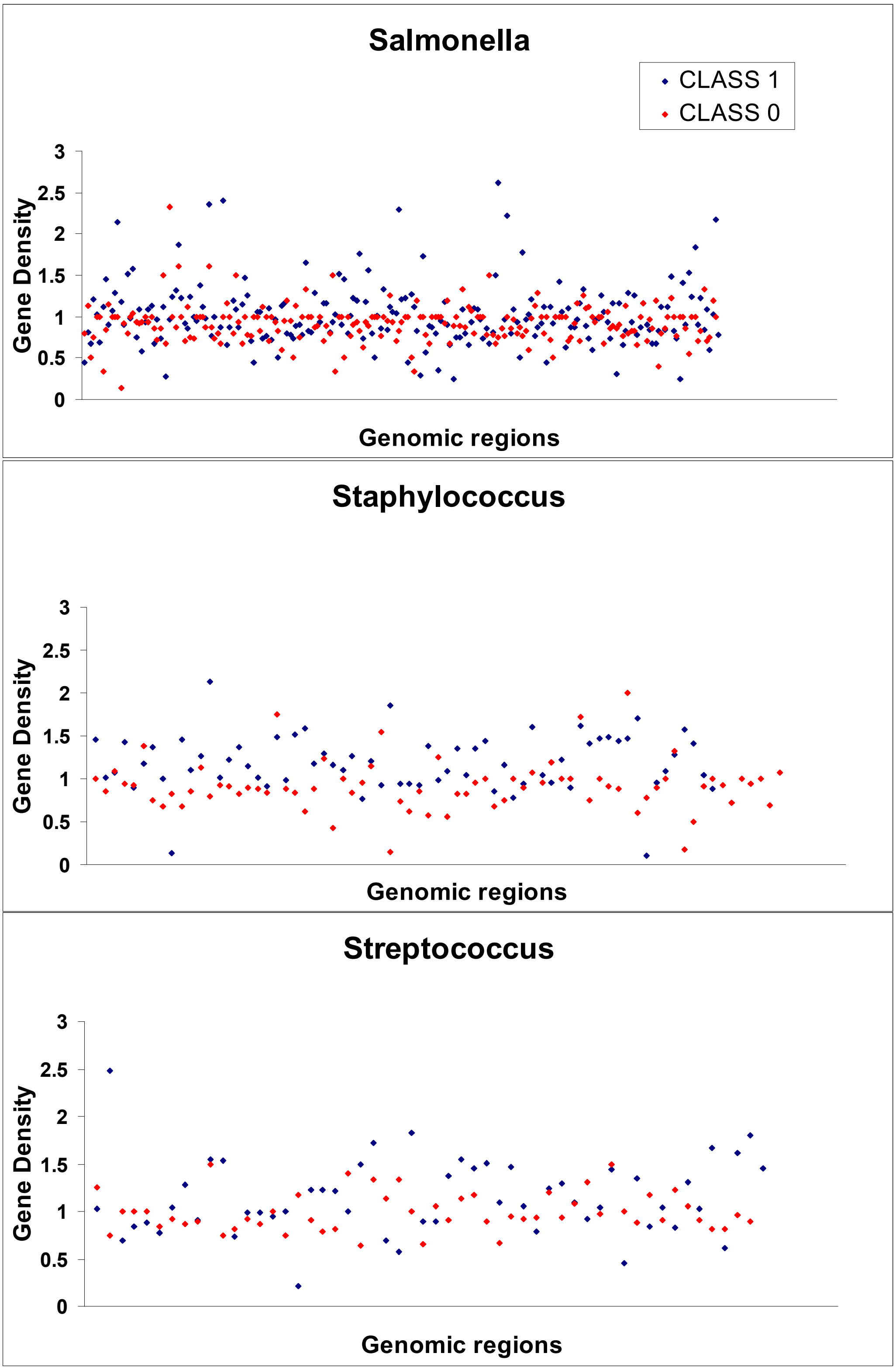


Figure 4.15: Gene Density of class "1" (GIs) and class "0" (randomly sampled) regions in *Salmonella* (top, *p*-value = 1.7 x 10^{-3}), *Staphylococcus* (middle, *p*-value = 1.4 x 10^{-6}) and *Streptococcus* (bottom, *p*-value = 1.3 x 10^{-2}) genera. The *p*-value has been calculated using a two-tailed *t*-test.

Although the three genus-specific GI structural models show distinct signatures, suggesting well-defined GI families with core and variable regions, when the RVM training takes place on a mixture of cross-genus examples, the various GI models converge over fairly similar GI structures (Figure 4.9). This observation supports further the idea that GIs overall represent a superfamily of mobile elements with significant structural variation, rather than a well defined family when looking across genus boundaries.

When the predictive accuracy and generalization properties of the cross-genus models are evaluated, many of those models perform overall equally well or better compared to the corresponding genus-specific models (Figure 4.13). This observation perhaps suggests that in some cases the RVM method has overfitted slightly on a subset of a genus-specific training dataset, misclassifying the remaining subset; when more training examples from other genera are included in the training dataset, models with much lower degree of overfitting are trained.

Between the cross-genus GI models, trained on a mixture of two different genera examples, the Staph-Strep model shows the highest accuracy compared to the Salm-Staph and Salm-Strep. Perhaps this cross-genus GI model is capturing structural properties of GIs found in Gram positive bacteria that are less or not informative for the prediction of GIs in Gram negative bacteria (Hacker *et al.*, 1997).

Even when the cross validation is based on a GI model that is trained on a genus-specific dataset and tested on examples of a different genus, the prediction accuracy remains remarkably high, further supporting the concept of the GI superfamily. For example, the accuracy of the model trained on *Salmonella* examples and tested on *Streptococcus* examples, is very similar to that of the *Streptococcus*-specific model. Moreover, the genus-specific GI model with the highest sparsity i.e. the Strep model discriminates remarkably well GIs from randomly sampled regions when tested on examples from the other two genera (Figure 4.16).

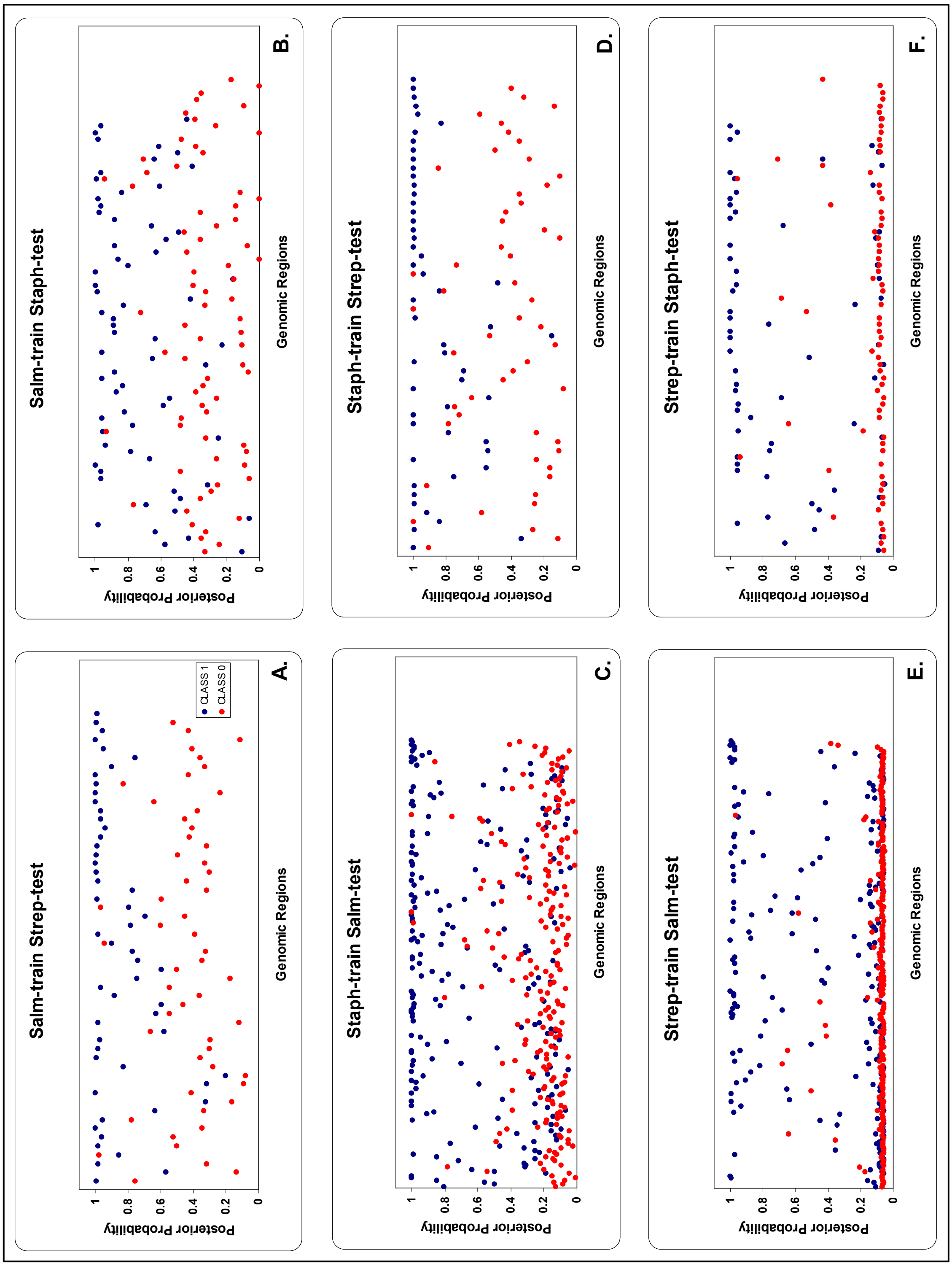


Figure 4.16: Scatter plot showing the posterior probability of a given region of being a true GI, given the model. Each genus specific dataset (e.g. Salm) is used to train an RVM model (e.g. Salm-train) that is then tested on the dataset of one of the other two genera (e.g. Strep-test). Each point in the scatter plot represents the posterior probability of either a GI (class 1, blue coloured) or a randomly sampled region (class 0, red coloured) of being a true GI given the model. For example in scatter plot A, a model trained on the *Salmonella* dataset was tested on the *Streptococcus* dataset: GIs (blue coloured points) in the test-set were correctly classified with a high probability very close to 1 while randomly sampled regions (red coloured points) in the test-set received on average a much lower probability.

Overall, the evaluation of the eight structural features across the 11 training datasets shows that the IVOM, PHAGE, SIZE and INTEGRASE features are on average the most informative ones, followed by the INSP, REPEATS, DENSITY and RNA features (Figure 4.17). It seems that the four most informative structural features are important predictors when classifying GIs from any of the three genera, suggesting that there are core features of a superfamily of mobile elements, whereas the other four, less informative features are capturing genus-specific properties of GIs (being informative only when predicting GIs from a single genus), suggesting these may be variable features of distinct genus-specific GI families.

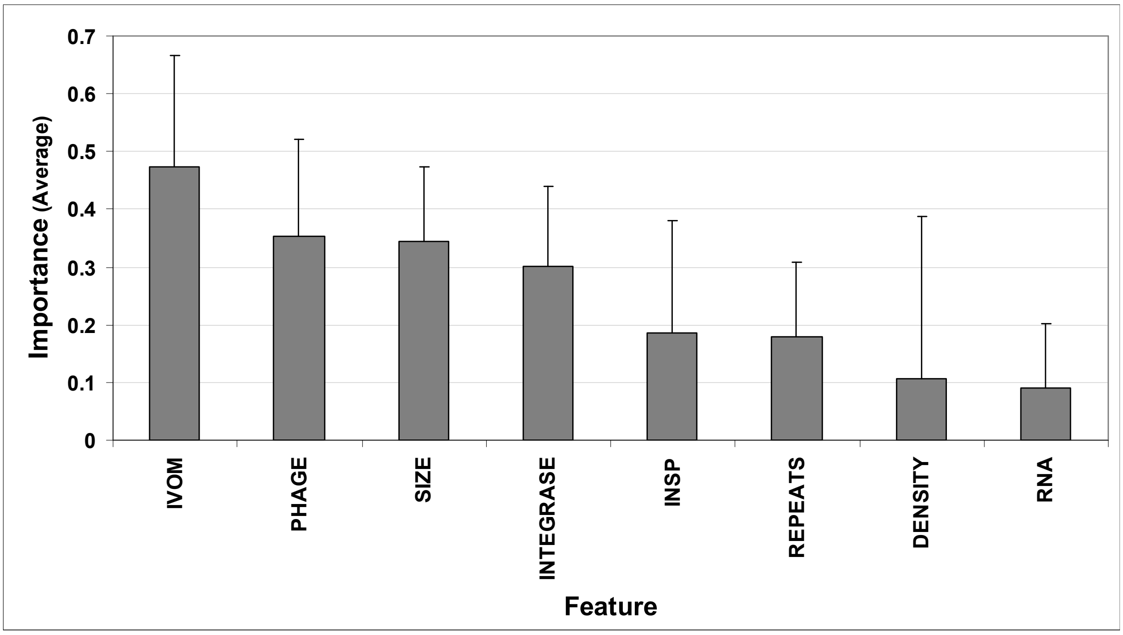


Figure 4.17: Bar chart illustrating the average "importance", across 11 structural GI models, of the eight structural features evaluated in this analysis. The eight features have been sorted (in decreasing order) based on their average "importance". Error bars show 1 SD.

The analysis carried out in this chapter forms the first attempt to quantify the actual GI structure in a probabilistic framework taking into account the contribution of all the informative structural features. Instead of vaguely describing putative GIs we can explicitly quantify our level of confidence that they fit an empirically-derived structure. This probabilistic scoring framework enables a systematic description of GI elements, which can be ranked based on their underlying structural information and subsequently classified into distinct structural families.

Although this methodology provides some new insights about the structural variation of GIs, there are some limitations that have to be taken into account: 1) the RVM method shows increased sparsity, providing simple models that can very accurately capture the underlying structural variation in some cases (e.g. the Strep model). On the other hand, the RVM method overfitted twice, to some extent, to the *Staphylococcus* dataset: firstly, the two Staph models (with and without the two rRNA operons in the control dataset) show significantly different weights, and secondly the Staph model models the *Staphylococcus* dataset more poorly than any of the other two genus-specific models (Salm and Strep), perhaps overfitting to the DENSITY feature. To test whether this is indeed the case for the Staph model, the DENSITY feature was removed from the training and test datasets and the cross validation was repeated using the three models (Salm, Staph, Strep), re-evaluating their performance on the *Staphylococcus* dataset.

The data supports the suggestion that the poorer performance of the Staph model on the *Staphylococcus* dataset, relative to the other two genus-specific models, is due to overfitting of the model to 20% of the dataset that had examples with significantly higher gene density than the rest of the dataset. The new Staph model outperforms the other two models when tested on the *Staphylococcus* dataset; more specifically, the AUC before and after the removal of the DENSITY feature for the three models, is as follows: (Staph = 0.824, 0.875), (Salm = 0.872, 0.865), (Strep = 0.850, 0.850). 2) The RVM method, as implemented in the current study, gave an error margin of 10-20%.

Possible sources of this error margin include: Significant structural intersection of the GIs and the randomly sampled regions; some randomly sampled regions were sampled close to classical GI-related structural features (e.g. tRNA) simply by chance while a few GIs lack most (or all) of the classical GI-related features (since no *a priori* structural assumptions were made). Moreover, the phylogenetic sample used in the current study strongly affects the validity of the training datasets; overall 11-13 strains and four outgroups were analyzed for each reference genus.

Regions of limited phylogenetic distribution (under a maximum parsimony evaluation) were defined as GIs, while inter-GI chromosomal regions were randomly sampled. Under this framework there are two possibilities to be taken into account: Firstly, some predicted GIs might not actually represent true GIs, if the phylogenetic resolution is further increased, i.e. including more reference strains and more distantly related outgroups. Secondly, some randomly sampled regions might have been sampled over “ancient” GIs that were acquired prior to the divergence of the reference and the outgroup lineages. Consequently, care should be taken when interpreting the results of this analysis; the parameters of the RVM models and the validity of the actual training datasets directly affect the conclusions drawn about the structural variation of GIs. These conclusions are specific only for the three datasets analyzed, the structural annotation methodology and the machine learning method implemented in this study.

The species sample used in this analysis is inevitably small in the context of a wide, representative sampling of the GI structural space. However, it forms a proof of concept showing that the components of a GI structure can explicitly be quantified through a probabilistic framework. Under this concept more species and many more structural components (e.g. the distance of GIs from the origin of replication oriC, their relative time of acquisition, number of pseudogenes per island and coding strand bias) can be taken into account and evaluated, enabling the construction of more sophisticated and more detailed structural models.

Overall in this analysis, I showed that GIs tend to fall within structural families with well defined signatures when looking within certain lineage boundaries, but when the taxa resolution decreases, i.e. looking at GIs across different species, universally distributed structural

GI components emerge. Perhaps overall, GIs should be seen as a superfamily of mobile elements with unifying and variable structural features rather than a single, well-defined family.

Chapter 5

Experimental validation of the predictions

5.1 Introduction

So far I have discussed three different methodologies for the prediction of Genomic Islands (GIs), i.e. a compositional-based (chapter 2), a comparative-based (chapter 3) and a structural-based (chapter 4) approach. For each method I have used an *in silico* derived, manually curated test-dataset in order to validate the results and benchmark the prediction accuracy. However, what I have not yet discussed is a "real-life", combined application of these methods on un-annotated datasets, derived from very early stages in the annotation pipelines; this reveals the true strengths/weaknesses of this multifactorial, integrative approach in aiding and/or guiding (rather than extending pre-existing) annotation methodologies, especially when the genome sequences of closely related strains are not available to identify horizontally acquired regions.

This challenge forms the focus of this chapter; using a newly sequenced, un-annotated bacterial genome, the aim is to make *in silico* predictions of horizontally acquired regions, exploiting an integrative compositional and structural-based approach, and use experimental, rather than *in silico*, protocols to confirm the putative origin (vertical or horizontal) of the predicted genomic regions. Applying a Polymerase Chain Reaction (PCR) protocol, the presence and absence of the predicted islands will be probed in 17 un-sequenced closely and distantly related strains and the true borders of these islands will be confirmed by sequencing across the boundary site in strains lacking the island.

At the time that this project was conceived, the genome sequence of *Stenotrophomonas maltophilia*, strain K279a became available. *S. maltophilia*, previously taxonomically classified as *Xanthomonas maltophilia* or *Pseudomonas maltophilia*, is a gram-negative, aerobic, nonfermentative bacillus (Denton and Kerr, 1998). *S. maltophilia*, is an important nosocomial pathogen, especially in immunocompromised patients, it has an unclear route of acquisition, little is known about its virulence properties (Denton and Kerr, 1998) and it shows resistance to broad-range antimicrobial agents, including β-lactam (Saino et al., 1982) and aminoglycoside antibiotics (Muder *et al.*, 1996). Clinical manifestations related to *S. maltophilia*, include, but are not limited to, endocarditis (Mehta *et al.*, 2000), bacteremia (Muder *et al.*, 1996), meningitis (Libanore *et al.*, 2004) and pneumonia (Fujita *et al.*, 1996).

Therefore the genome sequence of *S. maltophilia* K279a forms an excellent test-dataset for the purposes of this analysis; *S. maltophilia* is an important life-threatening pathogen, with unknown virulence properties, and there is only one complete genome sequence of this species available, rendering benchmarking based on *in silico* comparative genomics inapplicable.

5.2 Methods

Given my very limited previous experience in lab-based techniques and protocols, the experimental methodology followed in this analysis was designed to be effective and at the same time simple and easy to implement, without requiring special training and extensive supervision. The aim was to validate the *in silico* predictions by exploiting the PCR protocol, using primers designed to flank the borders of the candidate islands predicted in the sequenced genome; this methodology made it feasible to sample the presence/absence of those GIs in closely and distantly related un-sequenced *S. maltophilia* clinical isolates, draw conclusions about their phylogenetic distribution and estimate the accuracy of the predicted boundaries.

The *in silico* and experimental methods pursued in this analysis are described in the following sections. It should be noted that the conclusions drawn will be purely based on the results confirming both the presence of the candidate islands in some strains and their absence from at least one of the remaining strains; in case the data cannot confirm these two requirements, I will not make any inferences about the possible phylogenetic distribution and the origin of those predicted regions (see discussion section).

5.2.1 *In silico* prediction of GIs

The genome sequence (size: 4.85Mb, G+C%: 66.32) of *S. maltophilia* strain K279a (http://www.sanger.ac.uk/Projects/S_maltophilia/) was used as input to the Alien_Hunter (Vernikos and Parkhill, 2006) software (see chapter 2) and candidate GIs were predicted exploiting only compositional-based information. In a second step, the predicted candidate GIs were structurally annotated as discussed in chapter 4 and their structural annotation was used as input to the relevance vector machine (RVM) classifier (Tipping, 2001); RVM assigned a score to each prediction, quantifying our posterior belief that those structures are likely to be true GIs.

For the classification purposes, the three genus-specific structural GI models of *Salmonella*, *Staphylococcus* and *Streptococcus* described in chapter 4, as well as a model trained on all three datasets (Table 5.1, Table 5.2) were exploited. A sample of eight predictions with both highly and less probable GI structures with a score range of 0.2371–0.9997 formed the test-dataset of this analysis.

5.2.2 Comparative analysis

For the *in silico* sequence comparisons between the predicted boundaries of the putative GIs in the reference strain and the sequenced DNA fragments across the predicted insertion point in the un-sequenced *S. maltophilia* strains, a BLASTN (Altschul *et al.*, 1997) comparison was implemented and the results were visualized using ACT (Carver *et al.*, 2005).

Table 5.1: Structural annotation of eight genomic regions predicted as candidate GIs in the genome of *S. maltophilia*, strain K279a. Eight structural features were evaluated: The Interpolated Variable Order Motif (IVOM) score that measures both low and high order compositional deviation from the backbone composition and is expressed as the relative entropy between the query and the genome-backbone (variable order) compositional distribution, the insertion point (INSP) of each genomic region; two states were (binary) evaluated: insertion point within a CDS locus (disrupting the corresponding CDS) or insertion within an intergenic part of the chromosome, the size (SIZE) of each genomic region (bp), the gene density (DENS = number of genes per kb) of each region, presence or absence (binary) of direct/inverted repeats (REPEATS) flanking the boundaries of each genomic region, presence or absence (binary) of integrase and/or integrase-like (INT) protein domains, presence or absence (binary) of phage-related protein domains (PHAGE) and presence or absence (binary) of non-coding RNA (RNA) genes in the proximity of each region.

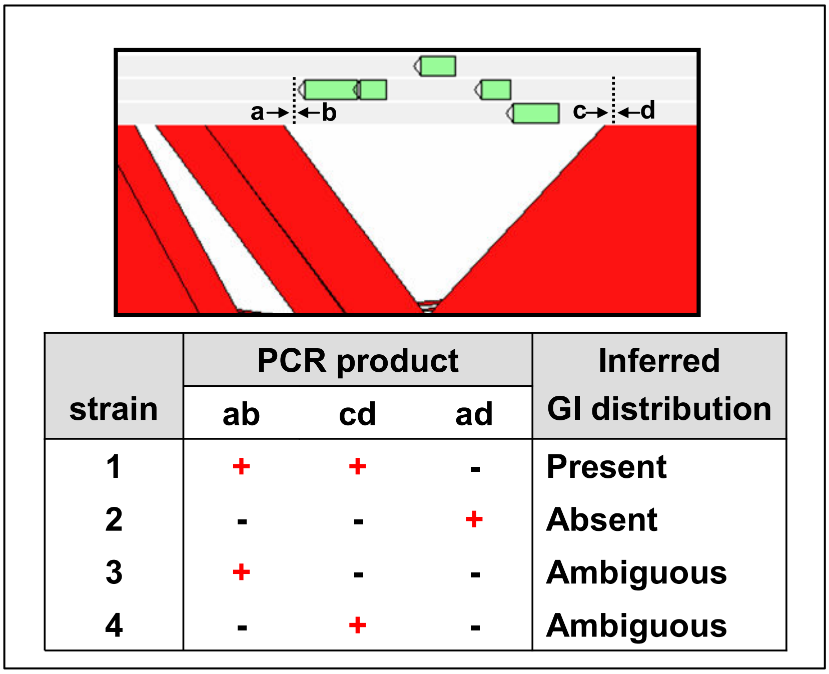
Location	Region	IVOM	INSP	SIZE	DENS	REPEATS	INT	PHAGE	RNA
60416..70829	R1	0.38128	1	10,413	1.3444	1	1	1	0
3089398..3127169	R16	0.74458	0	37,771	1.0060	1	1	1	1
299814..335480	R4	0.32642	0	35,666	1.2897	1	1	1	1
1323939..1367750	R12	0.55018	0	43,811	1.2325	1	1	1	0
1720046..1724493	R14	0.72176	0	4,447	1.7986	1	0	0	1
1945379..2002745	R15	0.28154	0	57,366	1.1854	1	1	1	1
3913072..3931089	R20	0.16626	0	18,017	0.6666	1	0	0	0
631285..661659	R7	0.27377	0	30,375	0.8559	0	0	0	0

Table 5.2: Posterior probability of being a true GI, for eight predicted genomic regions, exploiting four GI models, i.e. *Salmonella*-specific (Salm), *Staphylococcus*-specific (Staph), *Streptococcus*-specific (Strep) and the all-three (all3) genera model.

Region	Salm model	Staph model	Strep model	all3 model
R1	0.9918	0.9991	0.9994	0.9997
R16	0.9995	1.0000	0.9965	0.9992
R4	0.9944	0.9959	0.9804	0.9948
R12	0.9851	0.9997	0.9922	0.9903
R14	0.9978	0.9999	0.9005	0.9890
R15	0.9786	0.9826	0.9765	0.9835
R20	0.5023	0.2109	0.4742	0.4983
R7	0.3070	0.5223	0.1368	0.2371

5.2.3 Principle of the experimental approach

The principle of the experimental approach followed throughout this study is based on the analysis of the presence or absence of the amplified products for each set of primers, designed to flank the two boundaries of the predicted GIs (primers “a” and “b” for the left boundary; primers “c” and “d” for the right boundary), as well as for the “a” and “d” primers (Figure 5.1).



strain	PCR product			Inferred GI distribution
	ab	cd	ad	
1	+	+	-	Present
2	-	-	+	Absent
3	+	-	-	Ambiguous
4	-	+	-	Ambiguous

Figure 5.1: Screenshot summarizing the principle of the experimental approach followed in this study.

This experimental approach exploits the following three assumptions: A. If both the "a+b" and "c+d" products for a given GI-strain set are successfully amplified, then the predicted GI is inferred to be present in the corresponding strain; B. If only the "a+d" product is amplified, then the predicted GI is inferred to be absent from the corresponding strain; in this case the true boundaries can be determined by generating sequence from this product across the boundary site in strains lacking the island; C. Finally, amplified products for any other different combination of primers (e.g. only "a+b" or only "c+d" products) are inferred to be ambiguous results.

5.2.4 DNA purification

17 un-sequenced *S. maltophilia* clinical strains (Figure 5.2) were kindly provided by Dr Matthew Avison at the Department of Cellular and Molecular Medicine, University of Bristol. The 17 strains were grown overnight on Luria-Bertani broth (LB) media, at 37°C. Purification of genomic DNA was carried out using the *Wizard Genomic DNA*

Purification Kit of *Promega* according to the protocol for isolating genomic DNA from Gram negative bacteria (pages 16-17, *Promega* manual).

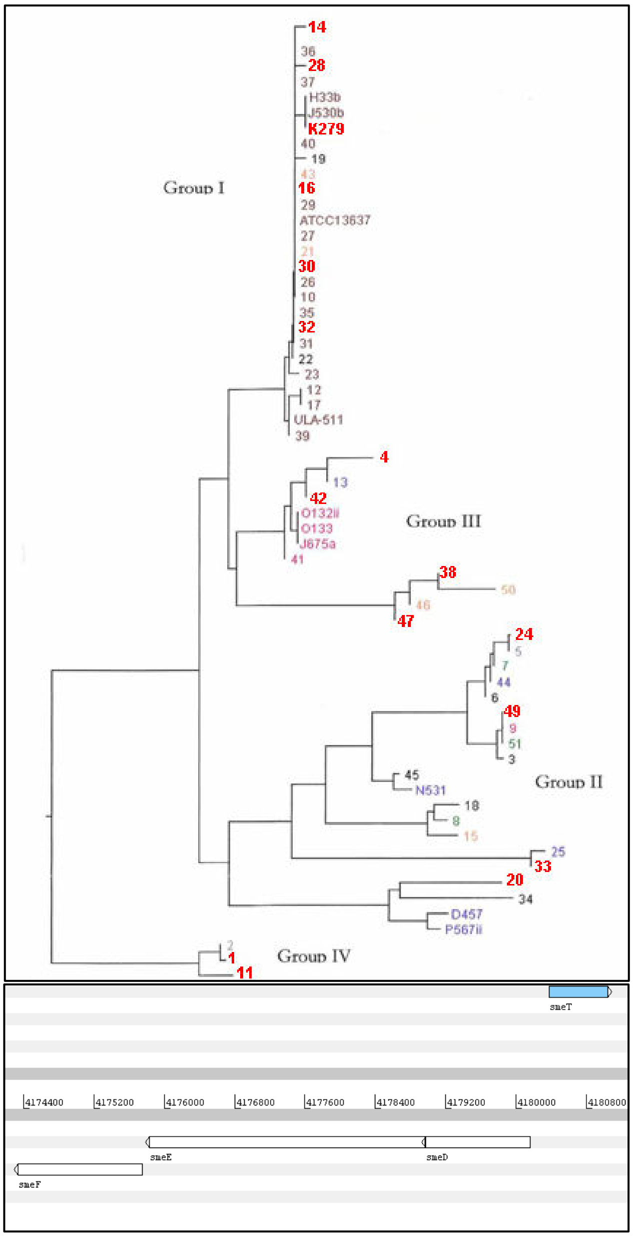


Figure 5.2: Phylogenetic tree of *S. maltophilia* isolates based on the *smeT–smeD* intergenic sequence (top); figure modified from (Gould *et al.*, 2006). The name of the strains used in this analysis, is highlighted in bold, red-coloured font. Three CDSs (smeD, smeE and smeF – accession number AJ252200), encode components of a multidrug efflux pump (Alonso and Martinez, 2000) present in *S. maltophilia*. The expression of the *smeDEF* locus (bottom), is regulated by a putative transcriptional repressor (smeT, belonging to the TetR and AcrR transcriptional regulator family), located upstream of the *smeDEF* locus (Sanchez *et al.*, 2002). The *smeT–smeD* intergenic region consists of a highly conserved and a hypervariable untranslated region (Gould *et al.*, 2004) and contains the putative promoters of *smeT* and *smeDEF*. The grouping of the *S. maltophilia* strains in the four (I, II, III and IV) phylogenetic groups has been based on the analysis of the 16s rRNA locus.

The concentration of the genomic DNA extracted from the 17 strains and the genomic DNA of the reference K279a strain was measured using the NanoDrop ND-1000 spectrophotometer; the results are shown in Table 5.3.

Table 5.3: Genomic DNA concentration of the 18 *S. maltophilia* strains used in this study.

Strain	Concentration (ng/µl)
K279a	200 (diluted to a final concentration of 20 ng/µl)
K279(1)	9.1
K279(2)	13.6
30	6.1
1	12.6
4	8.4
47	4.6
28	7.9
20	11.1
11	6.8
33	4.8
16	6.1
14	4.4
32	8.1
42	24.2
38	16.5
24	7.5
49	9.4

5.2.5 Primer design

For each of the eight candidate GIs, two sets of primers, one flanking the upstream and one flanking the downstream boundary were designed implementing the Primer3 software (Rozen and Skaletsky, 2000), available at http://frodo.wi.mit.edu/, using the default parameters. The 16 designed primers (Table 5.4) were ordered from SIGMA GENOSYS (http://www.sigmaaldrich.com/Brands/Sigma_Genosys.html).

Table 5.4: Primer sequences used in this analysis.

Genomic Region	Left boundary primer set	Right boundary primer set
R1	5′-gcagtgactcctgcagatcc-3′	5′-tcccccattacagcaggtag-3′
	3′-aggcttggtcttgcgaatag-5′	3′-ggagatccgaacatgcaatc-5′
R4	5′-ggcctgagcgactactacatc-3′	5′-gcaactccagctcatgctc-3′
	3′-ctgaaacatcggggaatcac-5′	3′-gcaagggctttcaagagttg-5′
R7	5′-agaagaccgagctgttcacc-3′	5′-cggtttcgaatatccagtgc-3′
	3′-gtttgacgtagctggcattg-5′	3′-ggatctgtttgcgatcctg-5′
R12	5′-cttcaagagctcgaccaacc-3′	5′-gactccatctcctggactgc-3′
	3′-tcgttcttgggctattatgg-5′	3′-accgtggccaatatcaagtc-5′
R14	5′-aatggtcgcgataccagttc-3′	5′-tacttgcttccctgccagac-3′
	3′-ctcgttcctcggcttcatag-5′	3′-atgacttcgggaatgcagac-5′
R15	5′-gagcgtagttgtcgtcgttg-3′	5′-acaggccttcgcagacatag-3′
	3′-gtttagccagagccgcatag-5′	3′-gcacgccaatactgagactg-5′
R16	5′-tgatccatccattctgcaag-3′	5′-atgcttgacgaaaggtttgc-3′
	3′-cctcccagattcgtgaaacc-5′	3′-tgtgcacgatgatctcaacc-5′
R20	5′-ggtggatgagaagccgatg-3′	5′-atctggccggagaagtacac-3′
	3′-cgtgtgctcaacgagaagg-5′	3′-acgagatcatgggctaccac-5′

5.2.6 Polymerase Chain Reaction – PCR

The purpose of PCR is the amplification of specific DNA fragments to a very large number of copies. The PCR protocol consists of three major steps (i.e. denaturation, annealing and extension), each of which is repeated 30-40 times.

The DNA fragments of interest were PCR amplified using the following reaction mixture (total volume 10μl): 0.2μl of genomic DNA, 0.1μl (100μM initial concentration) forward and reverse primers, 1μl (2 μM) dNTPs (dATP, dCTP, dGTP, and dTTP), 1μl PCR buffer (10x), containing 15mM of $MgCl_2$, 7.4μl of double-distilled water and 0.2μl (5units/μl) *Taq* polymerase (Amplitaq). PCR amplification was carried out using a PTC-225 peltier thermal cycler (MJ Research), implementing the program detailed in Table 5.5.

Table 5.5: The PCR protocol used in this analysis. At step 3 the optimal annealing temperature for each primer set was initially determined (and subsequently applied) using a gradient PCR protocol with a range of annealing temperature of 53-68 °C.

Step	Temperature (°C)	Time
1	95	10min
2	95	30sec
3	53-68	30sec
4	72	3min
5	goto step 2 (x39)	
6	72	10min
7	10	0min (for ever)

5.2.7 Gel electrophoresis

DNA fragments were separated on an agarose gel exploiting the electrophoresis protocol. The principle of this protocol is the separation of nucleic acids or proteins based on their charge and mass. Using an electric field, the macromolecules can be separated on a gel, with a rate of migration that depends on many factors, including the applied voltage, the hydrophobicity, size and shape of the molecules, the agorose gel concentration and the ionic strength of the buffer solution.

The agarose gel (1% w/v) was prepared by dissolving 0.5g of agarose (Sigma) in 50ml of Tris-acetate-EDTA (TAE) buffer (1x). The samples were loaded using 2µl of ficoll loading dye (0.25% bromophenol blue, 0.25% xylene cyanol FF, 15% Ficoll 400 in water) and run on the gel for 45 minutes applying a voltage of 60V. The samples were stained for 10-20 minutes by adding 5µl of ethidium bromide (10mg/ml) to the running buffer; the DNA bands on the gel were viewed under ultraviolet light. The size of the DNA fragments was determined by comparison with 1kb (Invitrogen) DNA ladder (1µg/µl).

5.2.8 Sequencing

In order to confirm the true borders of the predicted islands the boundary site in strains lacking the islands was sequenced (sequences are listed in Appendix J). All sequencing was performed by the core sequencing teams at the Sanger Institute, according to the protocols of the sequencing facility; briefly the templates were sequenced using AB BigDye terminator chemistry, and run on AB3730 machines. The resulting traces were base-called with in-house software (ASP), which also recognised and trimmed cloning vector and poor quality sequences.

5.3 Results

5.3.1 Genomic Island candidates

5.3.1.1 Genomic Island 1

The first candidate GI (R1), is a 10.5kb genomic region of low G+C content (63.82% – genome average 66.32%) and high gene density (1.34 – genome average 0.904) inserted within a coding sequence (CDS) (Smlt0055) encoding a putative alcohol dehydrogenase; this CDS is now disrupted by the integrated GI with the two CDS fragments flanking the 18bp direct repeats (DRs) of the island (Figure 5.3). R1 consists of 14 CDSs, the majority of which encode products with unknown function (Appendix K) and has the highest RVM score (0.9997, under the all3 model – Table 5.2), representing a highly probable GI structure.

R1 seems to represent a very recent acquisition in the *S. maltophilia* K279 lineage (Figure 5.4) since it is present only in the three *S. maltophilia* K279 strains (namely K279a, K279(1) and K279(2)). The absence of this GI was confirmed for strains 30, 28, 20, 14, 32, 24 and interestingly for strains K279(1) and K279(2) too (Figure 5.4); these results seem to contradict the presence of this GI in the latter two strains. However, sequencing across the boundaries of R1 in strain 28 (which lacks the island) and K279(1) shows that the two sequences are 97.5% identical

(711/729 identical residues – Appendix L) suggesting that the insertion point of R1 is present in all eight strains.

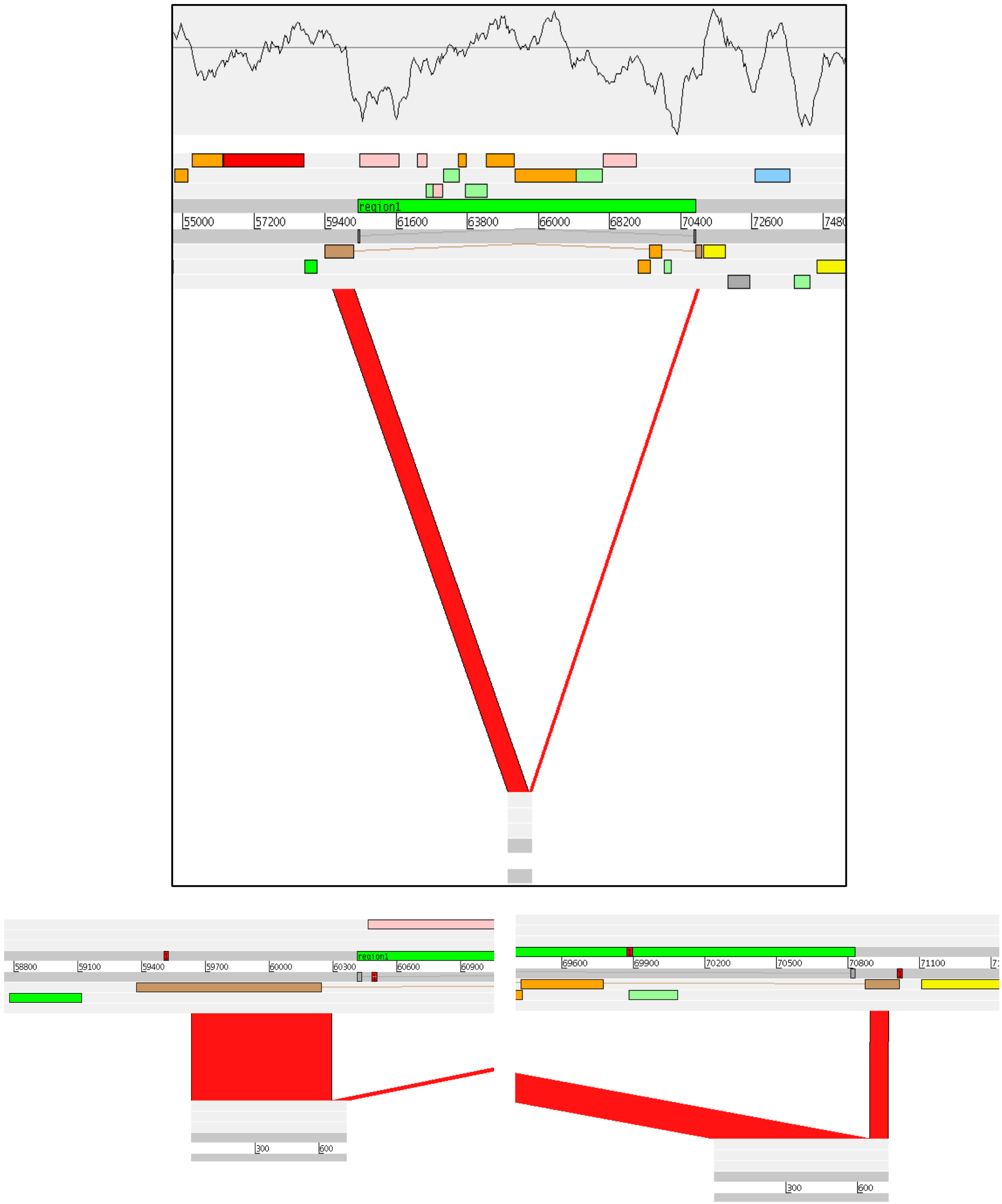
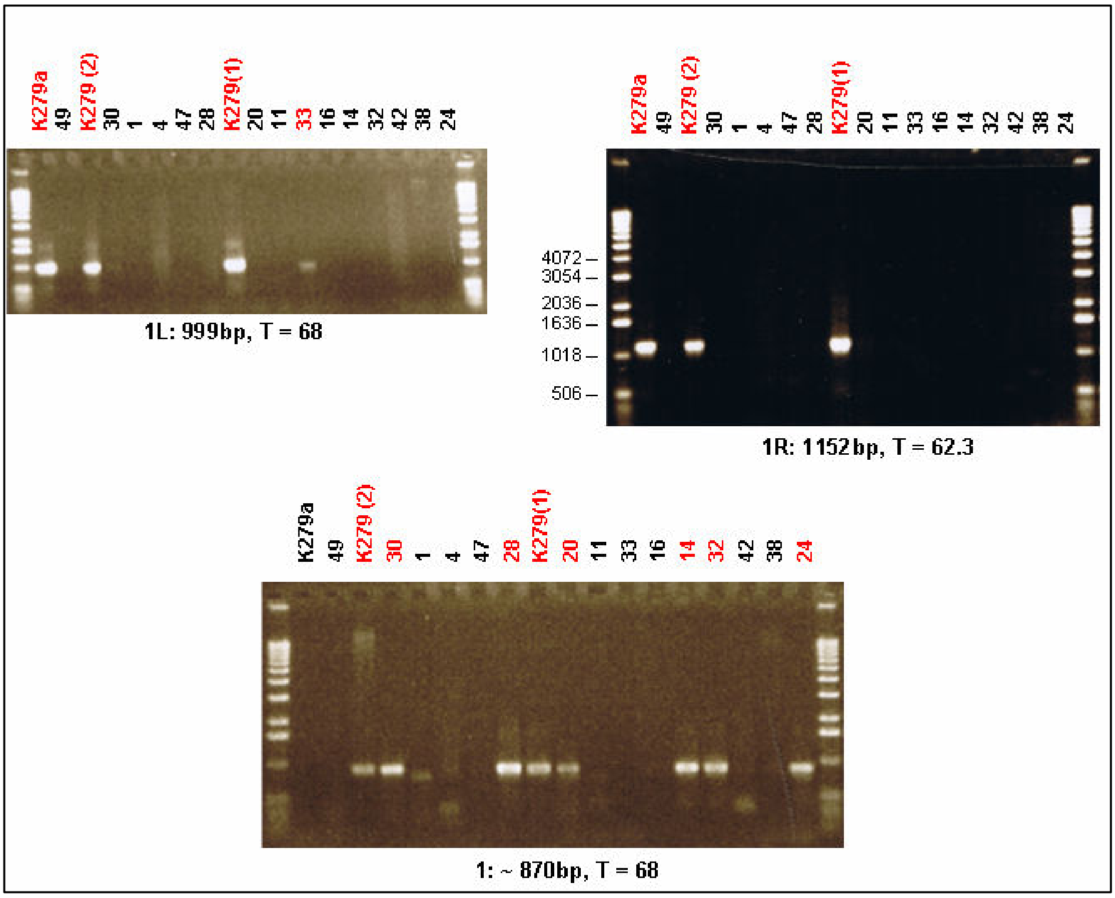


Figure 5.3: ACT screenshot: Predicted genomic island R1. Top: BLASTN comparison between *S. maltophilia* strain K279a and the sequenced fragment across the boundary site of R1 in strain 28. Regions within the two sequences with similarity are joined by red coloured bands that represent the matching regions. The G+C% content with a window size of 1kb is shown at the top of this screenshot. R1 is shown as green-coloured feature flanked by a set of 18bp DRs (grey coloured joined features). The DRs of R1 are flanked by the two fragments of Smlt0055 (brown-coloured joined features). Bottom: Higher resolution ACT screenshot showing the sequence similarity of the left and the right boundaries of R1 and the sequenced fragment of strain 28. The two sets of primers used to amplify the left and the right boundaries of R1 are shown as red-coloured features flanking the left and right attachment sites of this island.

Given that both the left and the right boundaries of R1 are present in all three K279 strains and, at the same time, the PCR results across the predicted insertion point of R1 suggest that R1 is also absent from K279(1) and K279(2) (Figure 5.4), it is likely that the insertion point (i.e. Smlt0055) of R1 has been duplicated in the latter two strains; the first copy has been disrupted by R1 while the second is intact.



1L: 999bp, T = 68

1R: 1152bp, T = 62.3

1: ~ 870bp, T = 68

Figure 5.4: PCR amplification of the left (1L) and the right (1R) boundaries (top) of genomic island R1 and of the region across the boundary site of R1 in strains lacking the island (bottom). The name of each strain is provided at the top of each lane and strains with amplified product, of the expected size, are highlighted in red. For each amplified product the expected sequence size (bp) and the optimal annealing temperature (T) is provided below each gel screenshot.

The global alignment between the insertion point of R1 in the reference strain K279a and the corresponding sequenced fragments in K279(1) and strain 28 (S28) shows that the three sequences are highly similar (K279a-K279(1): 99% identical – 723/730 identical residues; K279a-S28: 97.1% identical – 709/730 identical residues). An alternative hypothesis that might well explain the above ambiguity, is that a fraction of the K279(1) and the K279(2) populations, used to extract the genomic DNA for those two strain types, might have R1 inserted within Smlt0055 while the remainder of the population has the corresponding CDS intact (e.g. via a putative deletion event of R1).

Frequent deletion of GIs during population growth has been seen in other organisms (Buchrieser *et al.*, 1998; Bueno *et al.*, 2004; Nair *et al.*, 2004) and appears to occur via homologous recombination between the flanking DRs.

5.3.1.2 Genomic Island 16

R16, is a 37.7kb island of low G+C content (62.21% – genome average 66.32%) and similar gene density (1.006) to the genome average (0.904), inserted at the 3' end of a tRNA^{Ser} locus. R16 is flanked by a set of 21bp DRs with the terminal 13bp corresponding to the disrupted 3' end of the tRNA gene (Figure 5.5). R16 consists of 41 CDSs, the majority of which encode products of unknown function while three CDSs (Smlt3051, Smlt3053 and Smlt3069) encode two putative conjugal transfer proteins (traA and traD) and a putative plasmid partitioning protein, respectively (Appendix K). Based on the RVM score (0.9992, Table 5.2) R16 also represents a highly probable GI structure.

Similar to R1, R16 probably represents a recent acquisition in *S. maltophilia* K279 strains (Figure 5.6). The PCR results confirm that the same form of R16 is present in all three K279 strains, leaving open however the possibility that a variation of R16, with a different left boundary, might also be present in at least seven other strains (that gave amplified product, of the expected size, for the right boundary of R16) (Figure 5.6). Sequencing across the insertion site of R16, confirmed the complete absence of this island in at least one strain (strain 24), (Figure 5.6, Appendix J).

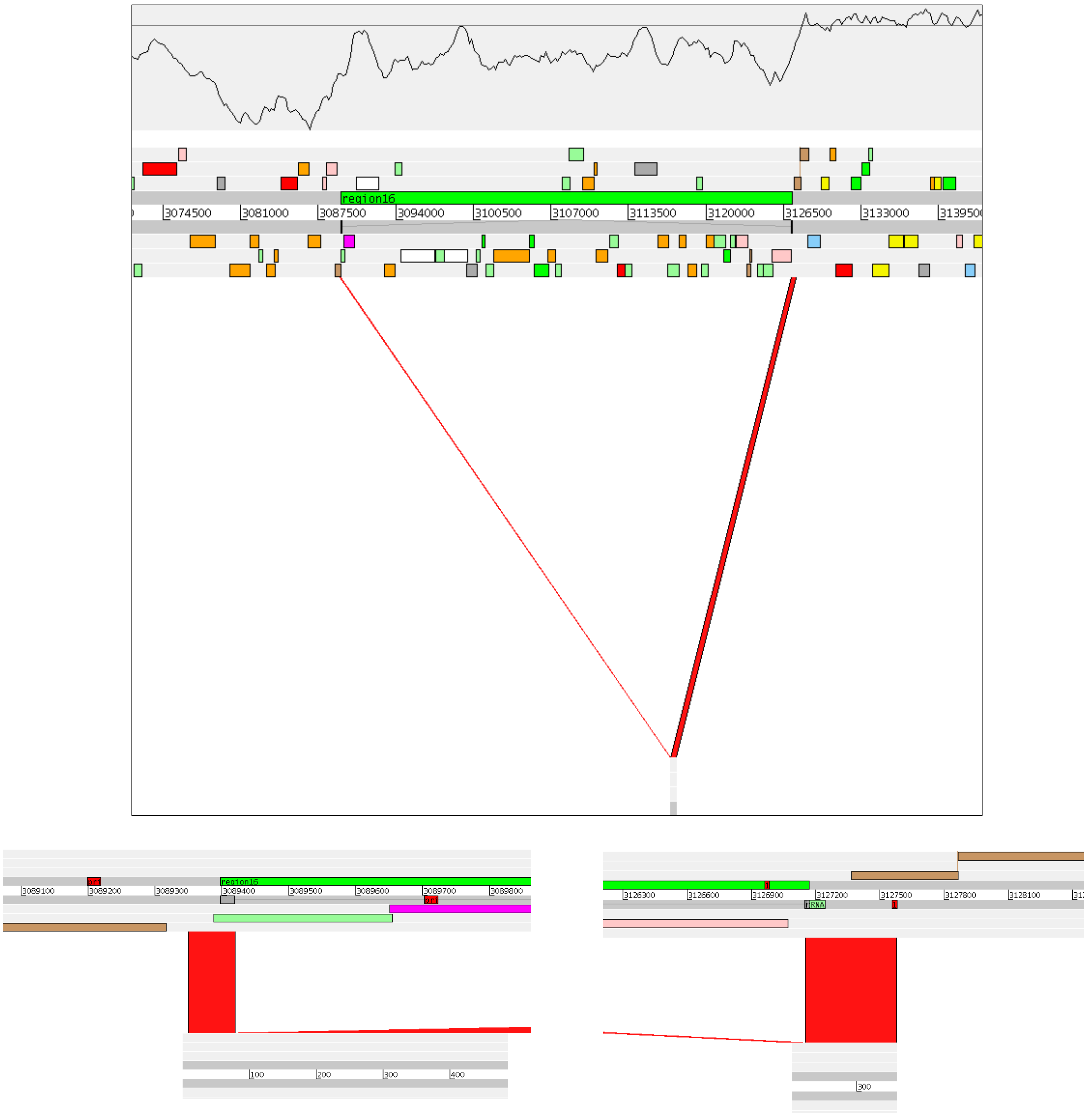


Figure 5.5: ACT screenshot: Predicted genomic island R16. Top: BLASTN comparison between *S. maltophilia* strain K279a and the sequenced fragment across the boundary site of R16 in strain 24. Bottom: Higher resolution ACT screenshot; R16 is shown as green-coloured feature flanked by a set of 21bp DRs (grey coloured joined features). The disrupted tRNA^{Ser} gene is shown as a light-green coloured feature overlapping with the DR at the right boundary of R16 (bottom-right screenshot). The two sets of primers used to amplify the left and the right boundaries of R16 are shown as red-coloured features flanking the left and right attachment sites of this island.

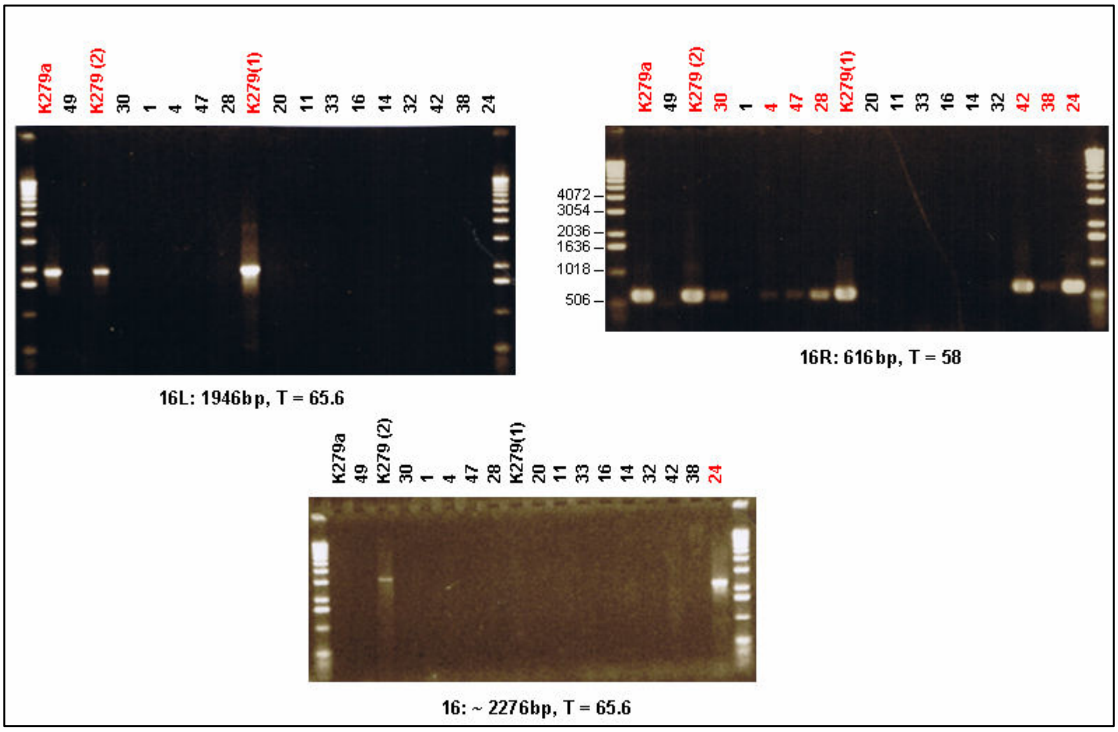


Figure 5.6: PCR amplification of the left (16L) and the right (16R) boundaries (top) of genomic island R16 and of the region across the boundary site of R16 in strains lacking the island (bottom).

5.3.1.3 Genomic Island 4

R4 is a 35.7kb island of low G+C content (63.42%) and high gene density (1.29) inserted at the 3' end of a tRNA^{Thr} locus (Figure 5.7). R4 is a putative prophage flanked by a set of 31bp DRs that correspond to the 3' end of the tRNA gene. R4 has a very high RVM score (0.9948) and consists of 45 CDSs, over half of which have sequence similarity to annotated phage-related CDSs (Appendix K).

R4 is present in the three *S. maltophilia* K279 strains (Figure 5.8) and a variation of this island with a different right boundary cannot be excluded from being present in strains 28, 32 and 24. PCR across the insertion point of R4 did not confirm the absence of this island in any of the 17 strains (see benchmarking section below); however the fact that for nine *S. maltophilia* strains the left boundary of R4 gave an amplified product of the expected size, and that the left primer set corresponds to the 3' end of Smlt0285 encoding a phage-related integrase (which is often conserved amongst related phages), leaves open the possibility that different or similar prophages might also be present in these strains.

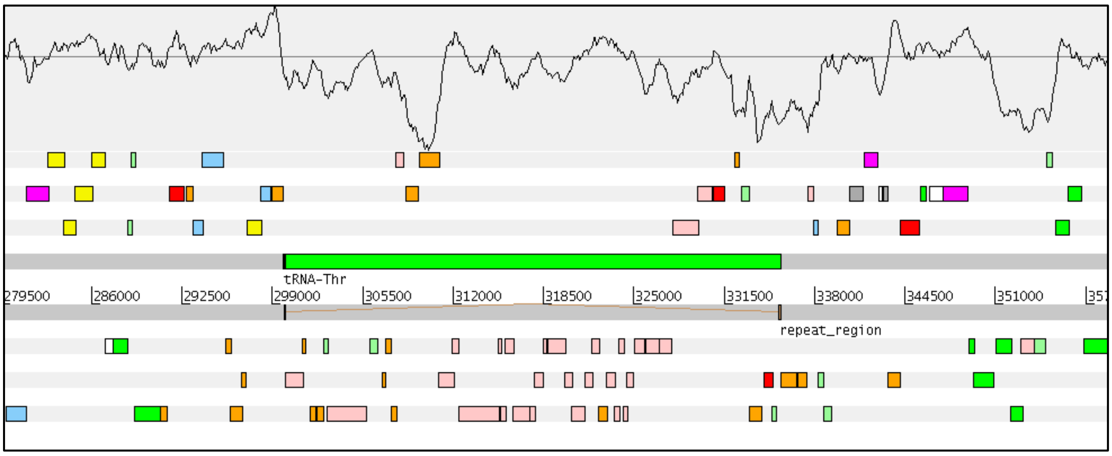


Figure 5.7: Artemis (Rutherford *et al.*, 2000) screenshot: Predicted genomic island R4, present in *S. maltophilia* strain K279a. The disrupted tRNA^{Thr} gene is shown immediately upstream of R4. The 31bp DRs are shown as brown-coloured joined features flanking the island. The G+C% content with a window size of 1kb is shown at the top of this screenshot.

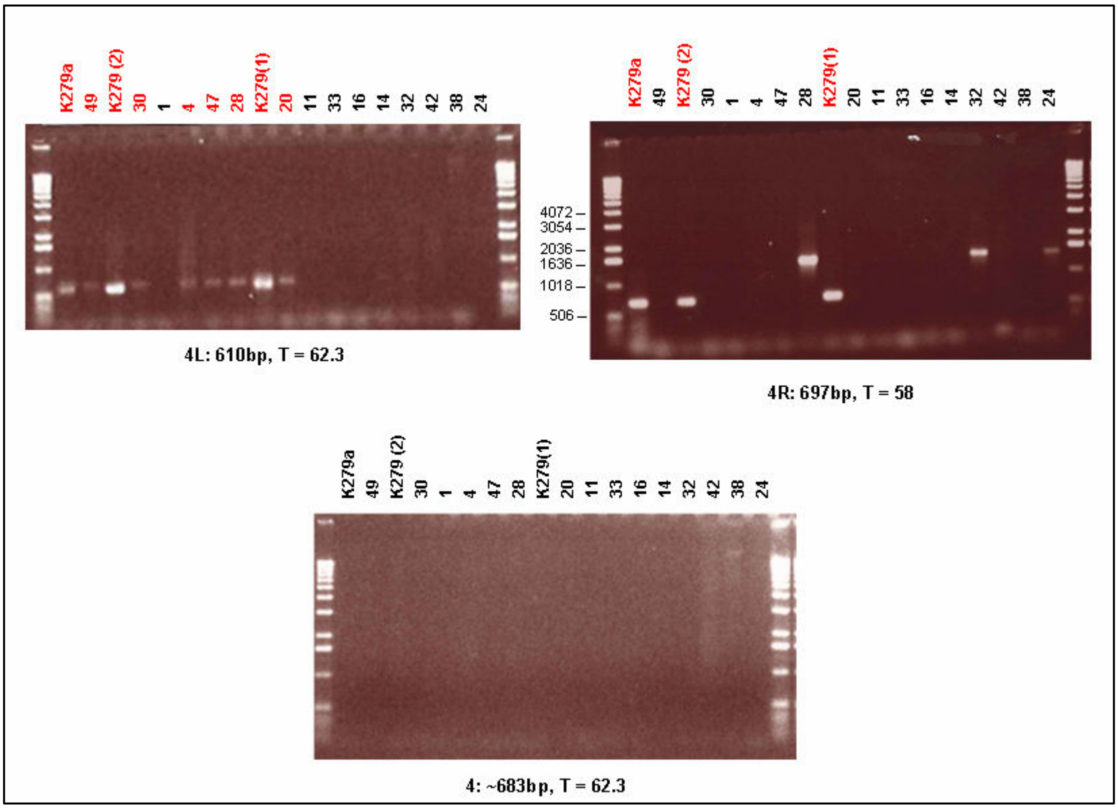


Figure 5.8: PCR amplification of the left (4L) and the right (4R) boundaries (top) of genomic island R4 and of the region across the boundary site of R4 in strains lacking the island (bottom).

5.3.1.4 Genomic Island 12

Similar to R16, R12 carries at least 10 CDSs encoding putative conjugal transfer proteins (Appendix K).

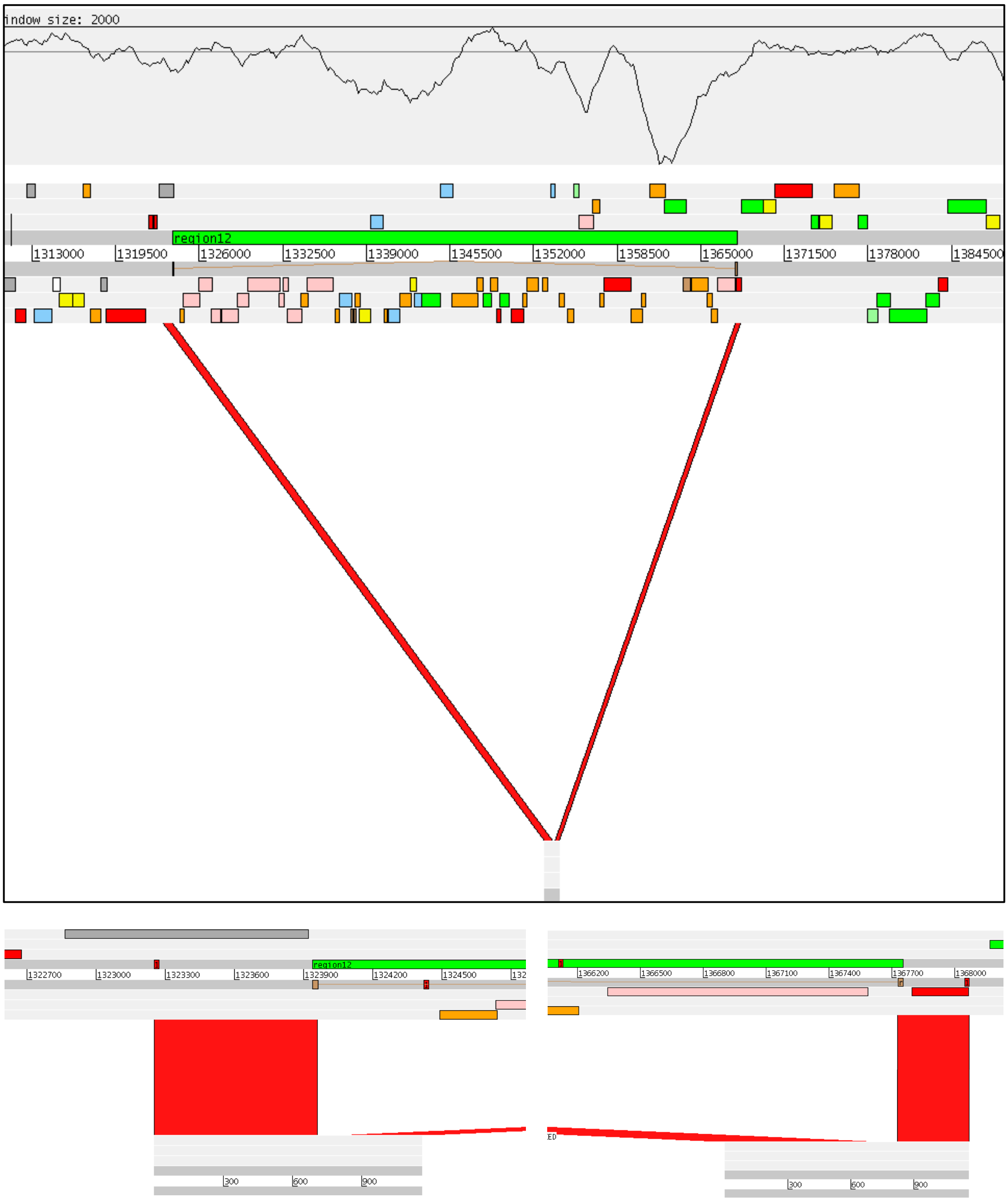


Figure 5.9: ACT screenshot: Predicted genomic island R12. Top: BLASTN comparison between *S. maltophilia* strain K279a and the sequenced fragment across the boundary site of R12 in strain 49. Bottom: Higher resolution ACT screenshot; R12 is shown as green-coloured feature flanked by a set of 22bp DRs (brown coloured joined features). The two sets of primers used to amplify the left and the right boundaries of R12 are shown as red-coloured features flanking the left and right attachment sites of this island.

R12 (Figure 5.9) is a 43.8kb island of low G+C content (62.69%) and high gene density (1.23) flanked by a set of 22bp DRs. R12 consists of 53 CDSs and it is inserted within a locus of three ribosomal protein coding genes (*smlt*1278 and *smlt*1279 encoding two putative 50S ribosomal proteins L21 and L27, located upstream of the left R12 boundary; *smlt*1337, encoding a putative 30S ribosomal protein S20, located downstream of the right R12 boundary). The posterior probability of this genomic region of being a true GI, under the all3 model, is 0.9903.

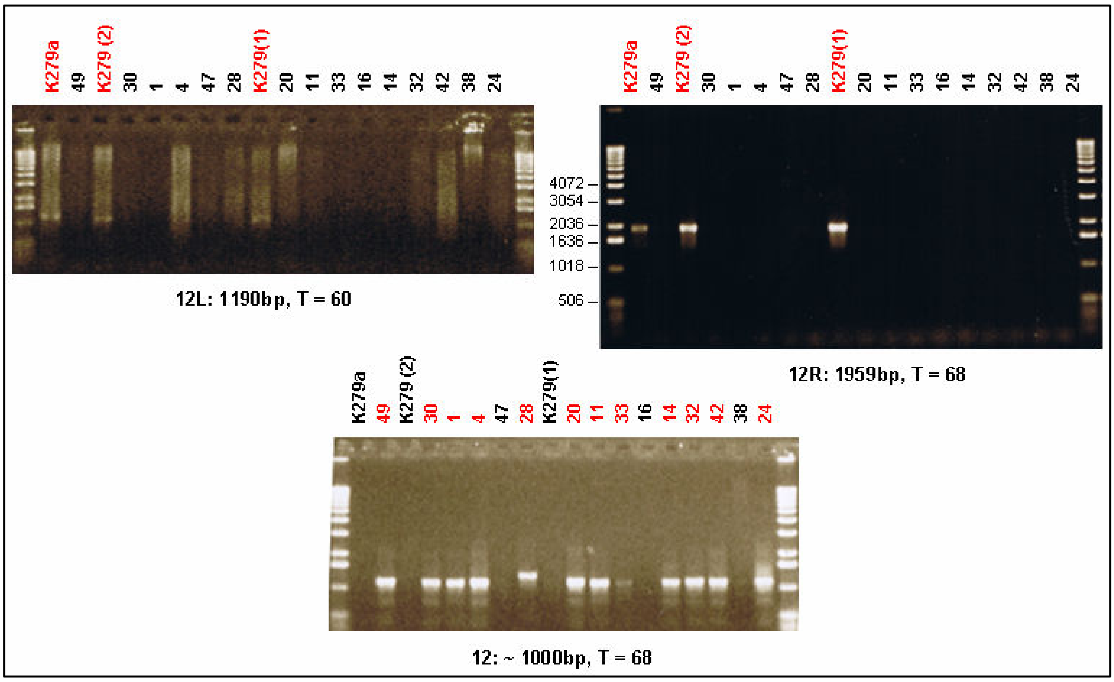


Figure 5.10: PCR amplification of the left (12L) and the right (12R) boundaries (top) of genomic island R12 and of the region across the boundary site of R12 in strains lacking the island (bottom).

R12 represents a very recent insertion, present in all three K279 strains (Figure 5.10), while its absence is confirmed, by PCR and sequencing across its insertion point, in 12 of the 17 *S. maltophilia* strains; these data suggest that most likely R12 is a K279-specific island and its insertion point is unoccupied in the majority of the un-sequenced isolates.

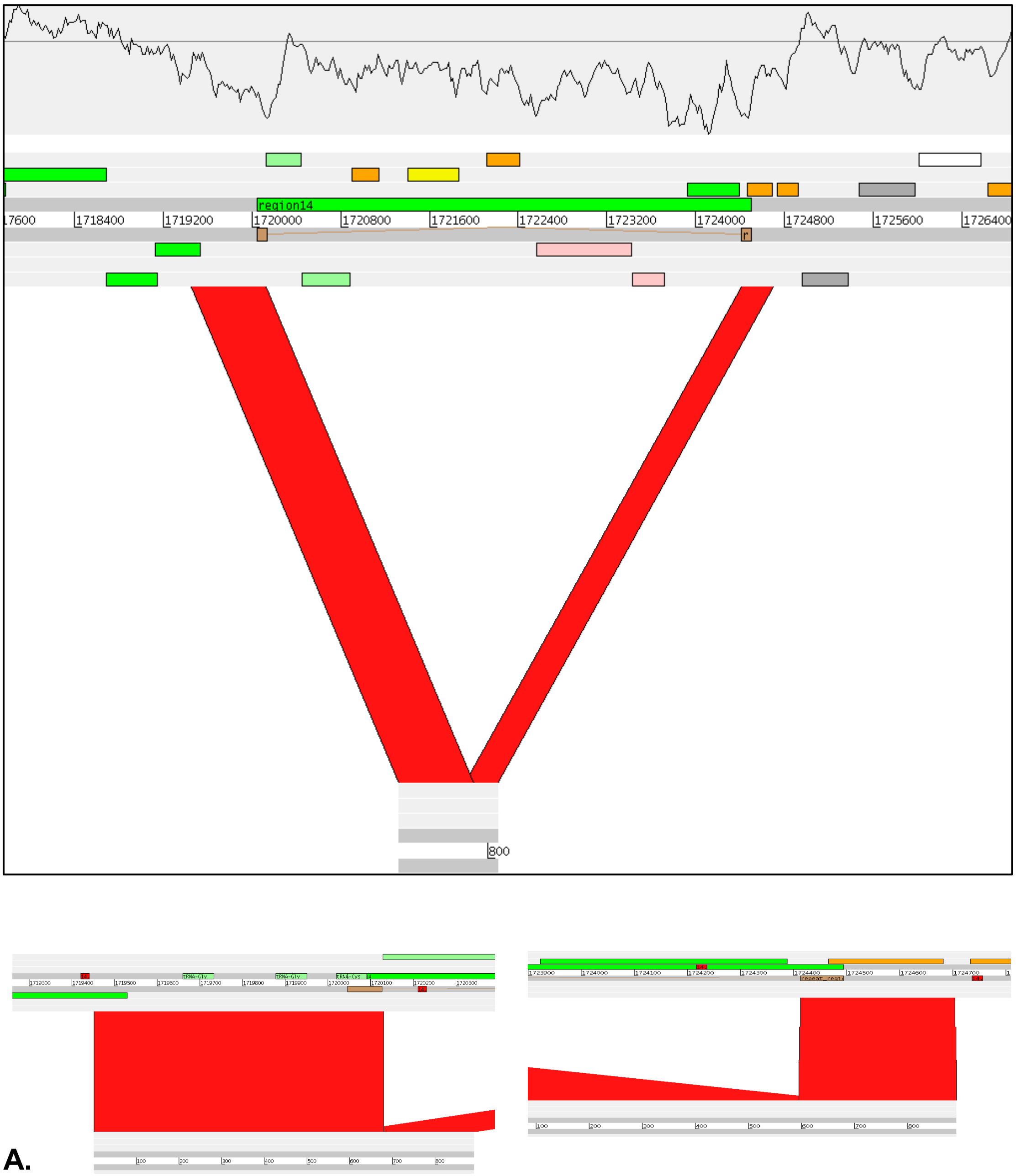
5.3.1.5 Genomic Island 14

R14 is a small island of 4.4kb and very high gene density (1.8), a value that is double the average gene density (0.904) characterising the genome of *S. maltophilia*, strain K279a. R14 has a very low G+C content (58.7%) and consists of eight CDSs, three of which (Smlt1662, Smlt1663 and Smlt1660) encode two insertion sequence (IS) Xac3-like transposases and a putative modification methylase, respectively (Appendix K). R14 is flanked by a set of very large (81bp) DRs, that overlap with 65% of the entire tRNA^{Cys} gene, located upstream of the left boundary of R14 (Figure 5.11); the insertion point of this island corresponds to the 3' end of this tRNA locus. The right boundary of R14 overlaps for 28bp with the 5' end of Smlt1665 (conserved hypothetical protein). The RVM score of R14 is 0.989.

R14 is present in the three K279 strains, while the PCR results confirmed its absence in at least eight *S. maltophilia* strains (Figure 5.12). However only two of those strains (30 and 14) gave the expected product size (~900bp) corresponding to the sequence across the insertion point of R14, while the remaining six strains (4, 47, 28, 20, 42 and 24) gave a product of slightly larger size (~1,200-1,300bp); these data leave open the possibility of a putative internal sequence variation of the corresponding R14 insertion site in the latter six strains, given that the sequencing of the two different products confirmed the same left and right boundaries of this island (Figure 5.11).

It is worth mentioning, that the sequencing of the corresponding region in strains 14 and 28 would, in theory, reveal (see R15 in the following section) the gene content of the ~400bp size difference between the amplicons; however the entire sequence of the amplified region was successfully determined only for strain 14, whereas in the case of strain 28, the sequence is missing a fragment from the left end of the corresponding amplicon. Based on the gene content information, showing three tRNA genes located immediately upstream of R14 (Figure 5.11b), it is likely that the ~400bp size difference might be due to the presence of extra copies of tRNA genes in strain 28.

Interestingly this size variation is consistent with the phylogenetic tree of the *S. maltophilia* lineage (Figure 5.2); indeed strains 14 and 30 (product size ~900bp) are members of the same taxonomic group (i.e. group I) with the three K279 strains. On the other hand, the remaining six strains (with the exception of strain 28) are more distantly related isolates and belong to the taxonomic groups II and III (Figure 5.2).



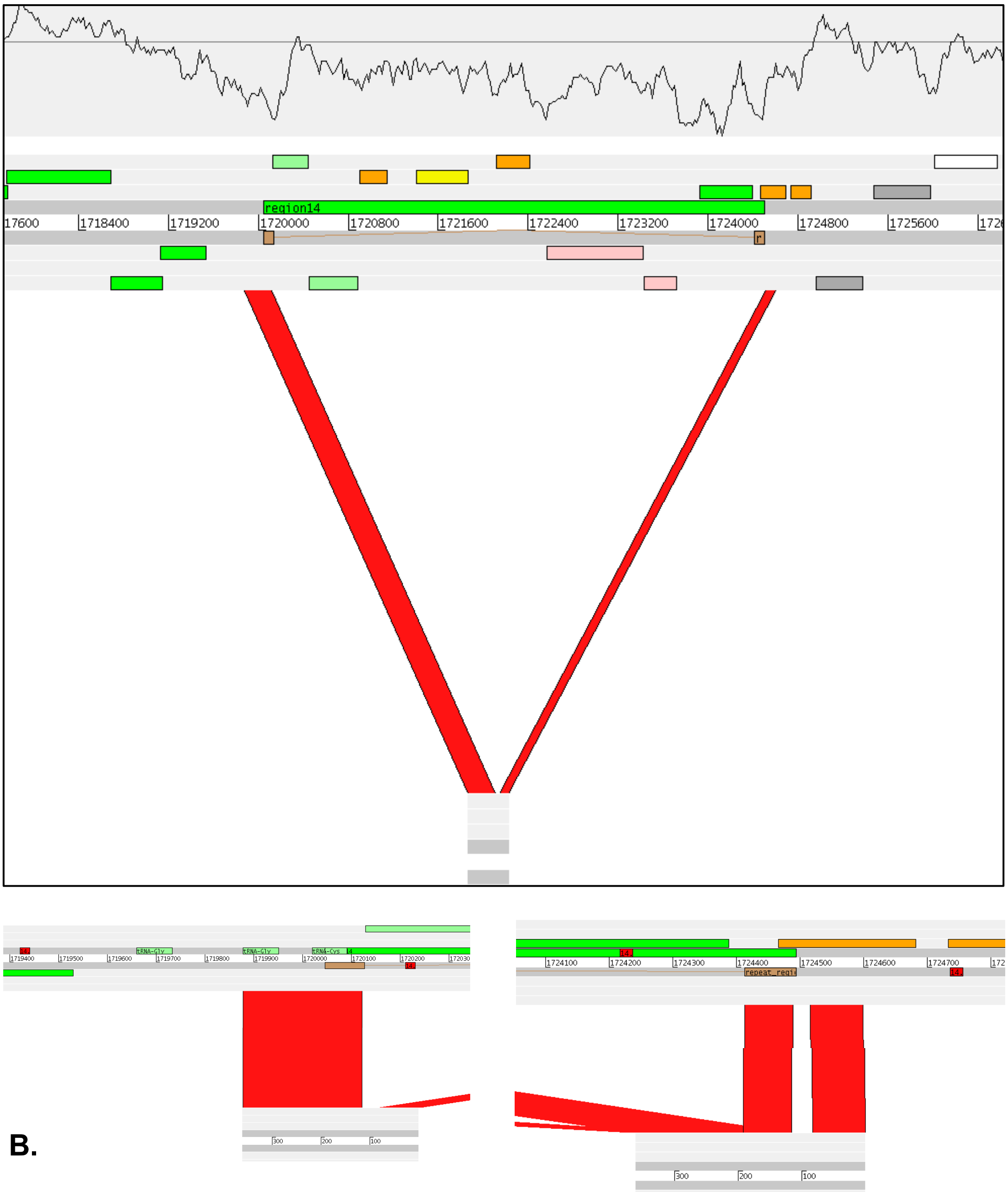
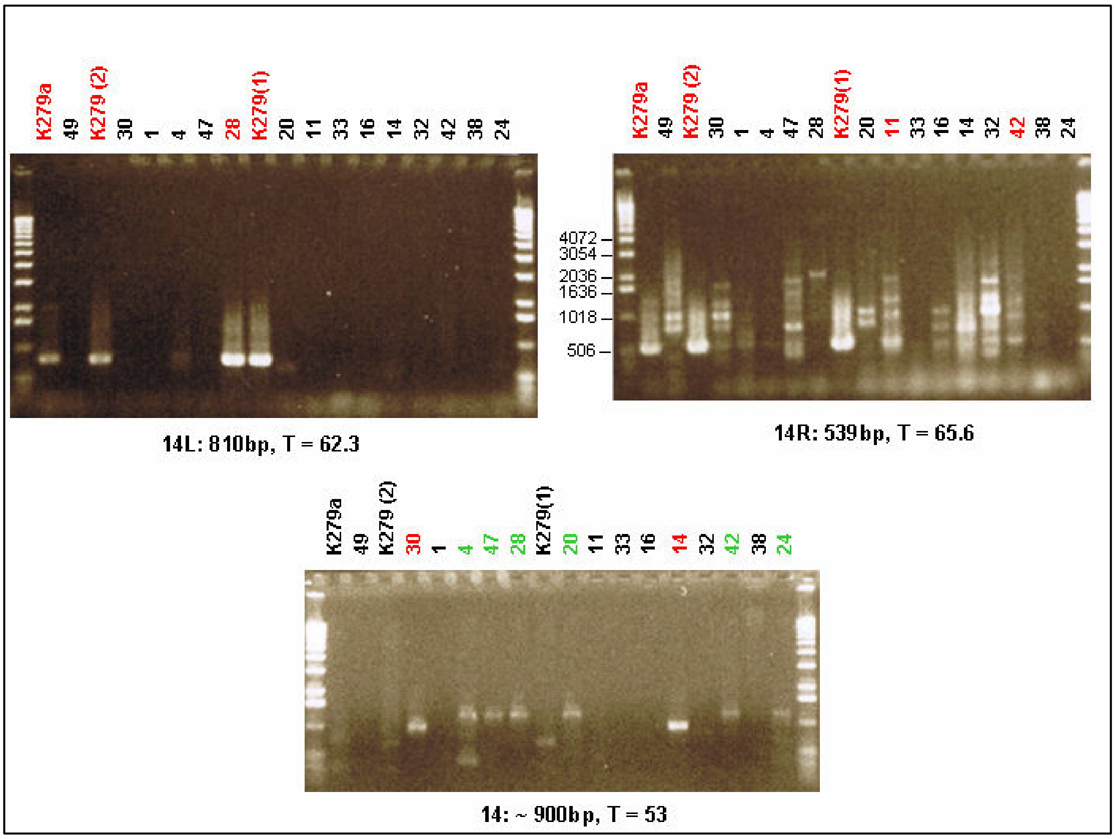


Figure 5.11: ACT screenshot: Predicted genomic island R14. Top: BLASTN comparison between *S. maltophilia* strain K279a and the sequenced fragment across the boundary site of R14 in strain 14 **(A)** and strain 28 **(B)**. Bottom: Higher resolution ACT screenshot; R14 is shown as green-coloured feature flanked by a set of 81bp DRs (brown coloured joined features). The two sets of primers used to amplify the left and the right boundaries of R14 are shown as red-coloured features flanking the left and right attachment sites of this island; the tRNA^{Cys} gene, upstream of R14 is shown as a light-green coloured feature overlapping with the left boundary of this island (bottom-left).



14L: 810bp, T = 62.3

14R: 539bp, T = 65.6

14: ~ 900bp, T = 53

Figure 5.12: PCR amplification of the left (14L) and the right (14R) boundaries (top) of genomic island R14 and of the region across the boundary site of R14 in strains lacking the island (bottom). Colour scheme (bottom): Strains whose product size is ~900bp (expected size) are red coloured while strains with a larger product size ~ 1,200-1,300bp are green coloured. Note: for a higher annealing temperature (T = 65.6 °C), only strains 30 and 14 gave an amplified product for the sequence fragment across the boundary site of R14.

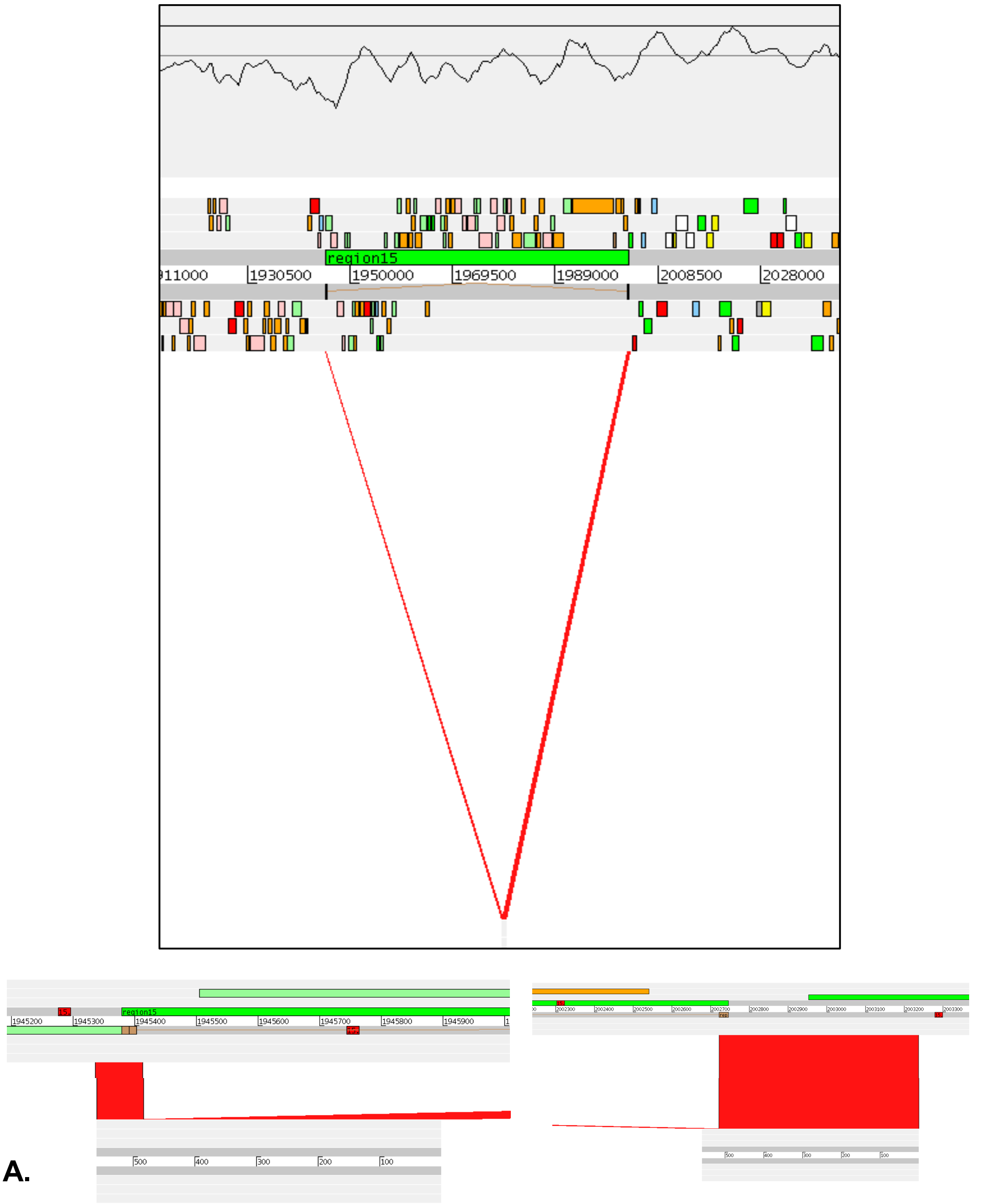
5.3.1.6 Genomic Island 15

R15 is a 57.4kb island of low G+C content (64.8%) and high gene density (1.18) inserted at the 5' end of a tmRNA (also known as 10Sa RNA) gene (Figure 5.13). R15 carries 68 CDSs, 11 of which have sequence similarity to annotated phage-related CDSs while the majority of the remaining CDSs encode for proteins of unknown function (Appendix K). R15 has a high RVM score (0.984) and is flanked by a set of 24bp DRs that overlap with the first 12 bases of the tmRNA locus; this tmRNA gene seems to represent an insertion site hot-spot, since its 3' end forms the insertion site of a second (52.9kb) genomic element of putative phage origin (data not shown), flanked by a set of (8bp) DRs, that is located immediately upstream of R15 in a head-to-head orientation; for these reasons, the left primer set for R15 was designed within the tmRNA locus to avoid possible problems with the differential presence of the other island. Interestingly R15 carries a tRNA^{Met} gene located (internally) 20.4kb downstream of the left boundary of this island; it is worth noting that overall there are five copies (including the R15 copy) of tRNA^{Met} genes present in the genome of *S. maltophilia*, strain K279a.

Unlike the previously discussed GIs, the presence of R15 was confirmed in four *S. maltophilia* strains, namely K279a, K279(1), K279(2) and strain 32 (Figure 5.14); the latter belonging to the same taxonomic group (group I) as the three K279 strains (Figure 5.2). The absence of R15 was confirmed in at least seven strains and, similarly to R14, there are two different product sizes that are phylogenetically consistent with the *S. maltophilia* phylogenetic tree; for strains 30, 28, 16, 14 and 24 the PCR amplified products had the expected size (~656bp) while strains 4 and 42 gave a product size of ~1,500bp (Figure 5.14). With the exception of strain 24 (taxonomic group II) all four strains that gave the expected product size belong to the taxonomic group I, while strains 4 and 42 are members of the taxonomic group III (Figure 5.2).

It is worth mentioning that the ~800bp size difference between the PCR products is almost exclusively attributed to the presence of two predicted CDS fragments present in strain 42 (and presumably in strain 4); those two CDS fragments, named herein CDS1 and CDS2 are very similar (Figure 5.13c) to SmalDRAFT_1529 (encoding a putative uncharacterized protein) and SmalDRAFT_1530 (encoding a GCN5-related N-acetyltransferase) present in *S. maltophilia* R551-3 ctg153 (Accession Number: AAVZ01000019). CDS1 and CDS2 along with CDS3 (encoding a putative transmembrane protein, similar to the 5' end of SmalDRAFT_1530 and Smlt1982 – present in K279a) are also sequentially located in the same orientation in *S. maltophilia* R551-3;

however based on the BLAST comparison, CDS1 is probably a remnant of SmalDRAFT_1529 (Figure 5.13c). These data confirm the absence of R15 from the corresponding predicted insertion site present in the available sequence of *S. maltophilia* R551-3 and further suggest that the gene content of this locus is conserved and unoccupied (by a GI) in at least three *S. maltophilia* strains.



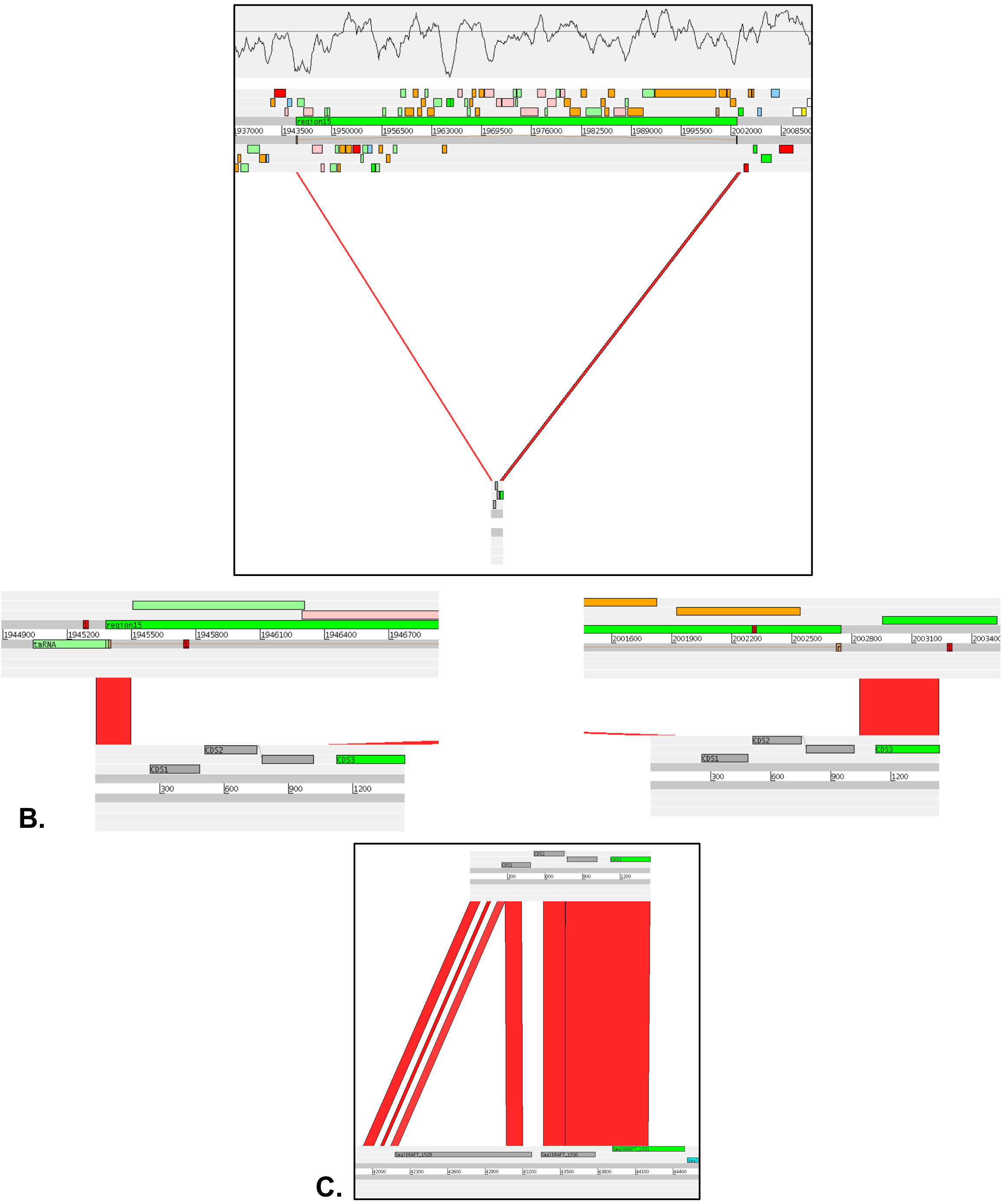


Figure 5.13: ACT screenshot: Predicted genomic island R15. Top: BLASTN comparison between *S. maltophilia* strain K279a and the sequenced fragment across the boundary site of R15 in strain 24 **(A)** and strain 42 **(B)**. Bottom: Higher resolution ACT screenshot; R15 is shown as green-coloured feature flanked by a set of 24bp DRs (brown coloured joined features). The two sets of primers used to amplify the left and the right boundaries of R15 are shown as red-coloured features flanking the left and right attachment sites of this island; the tmRNA gene, upstream of R15 is shown as a light-green coloured feature overlapping with the left boundary of this island. **C.** ACT comparison between the sequence across the boundary site of R15 in stain 42 and the corresponding sequence in *S. maltophilia* R551-3 ctg153 (Accession Number: AAVZ01000019).

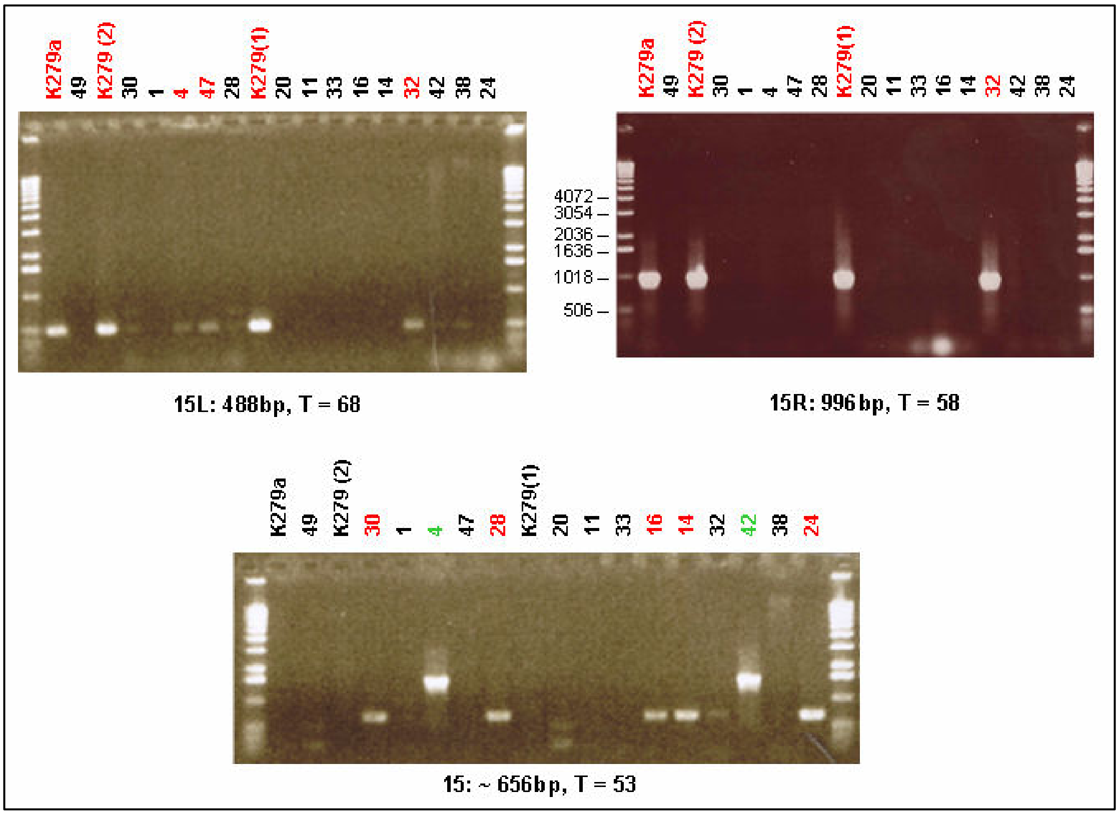


Figure 5.14: PCR amplification of the left (15L) and the right (15R) boundaries (top) of genomic island R15 and of the region across the boundary site of R15 in strains lacking the island (bottom). Colour scheme (bottom): Strains whose product size is ~656bp (expected size) are red coloured while strains with a larger product size ~ 1,500bp are green coloured.

5.3.1.7 Genomic Island 20

R20 is a medium size putative island of 18kb, low gene density (0.67) and very similar G+C content (65.8%) to the genome average (66.32%). R20 consists of 12 CDSs (Appendix K), encoding, among others, an autotransporter haemagglutinin-related protein (Smlt3829), two putative giant cable pilus-related proteins (Smlt3830 and Smlt3833), a putative outer membrane usher protein (Smlt3832) and a putative 50S ribosomal protein L31 (Smlt3836). R20 is flanked by a set of 24bp DRs (Figure 5.15) with the left DR being located immediately downstream of the termination codon of Smlt3827 (conserved hypothetical protein). The posterior probability of R20 of being a true GI is quite low (0.498).

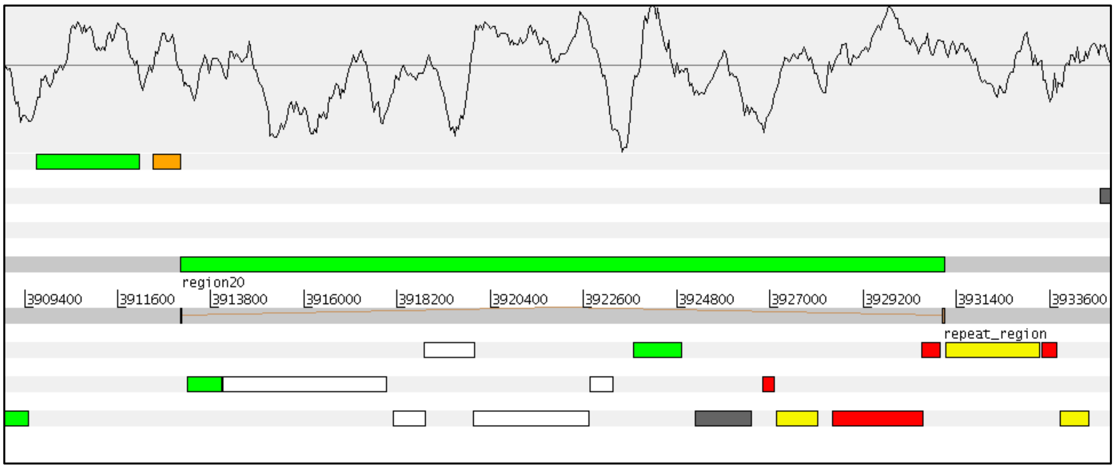


Figure 5.15: Artemis screenshot: Predicted genomic island R20, present in *S. maltophilia* strain K279a. The 24bp DRs are shown as brown-coloured joined features flanking the island. The G+C% content with a window size of 1kb is shown at the top of this screenshot.

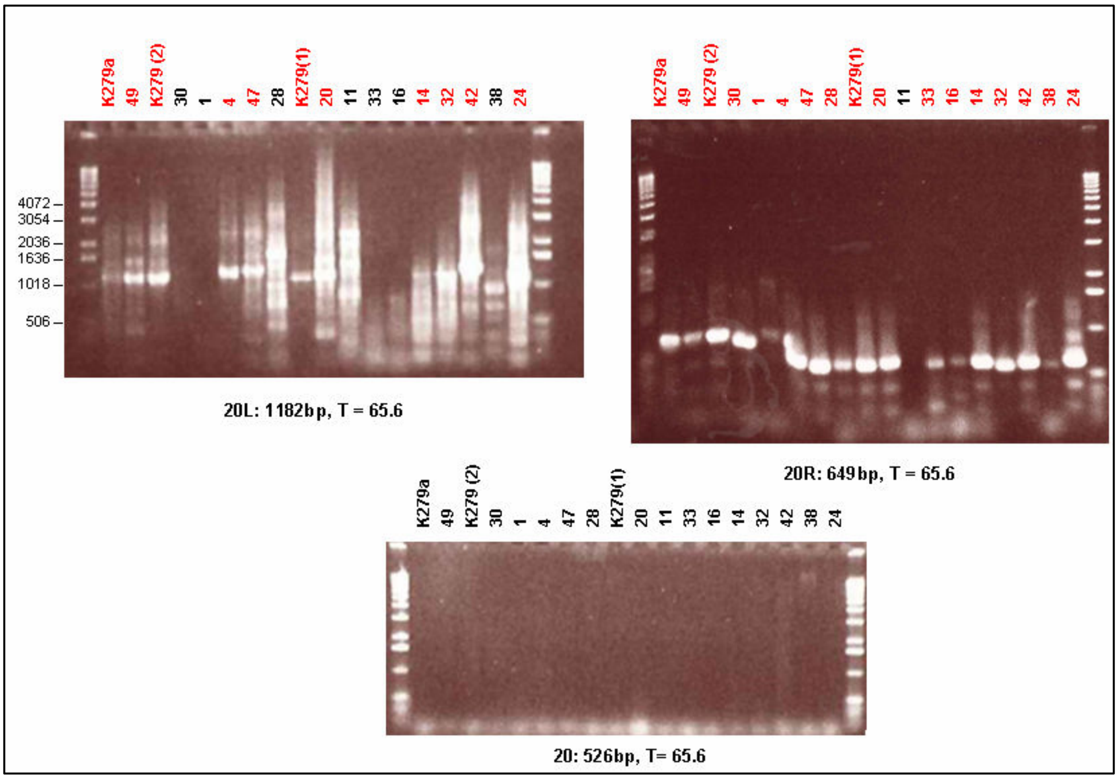


Figure 5.16 : PCR amplification of the left (20L) and the right (20R) boundaries (top) of genomic island R20 and of the region across the boundary site of R20 in strains lacking the island (bottom).

The fact that R20 region encompasses a major component (ribosomal protein L31) of the translation machinery (ribosome), in combination with its very low RVM score, makes it unlikely that this predicted region represents a true GI that has been horizontally acquired. Indeed, the PCR results suggest that this genomic region is present in the majority of the *S. maltophilia* strains used in this study (Figure 5.16), while the PCR across the insertion point of R20 did not indicate its absence in any of the 17 strains.

5.3.1.8 Genomic Island 7

R7 is a 30.4kb predicted island (Figure 5.17) of low G+C content (62.4%) and low gene density (0.86). R7 consists of 26 CDSs that encode proteins mainly involved in the lipopolysaccharide (LPS) biosynthesis (Appendix K) and represents a very unlikely GI structure with a very low posterior probability of 0.237.

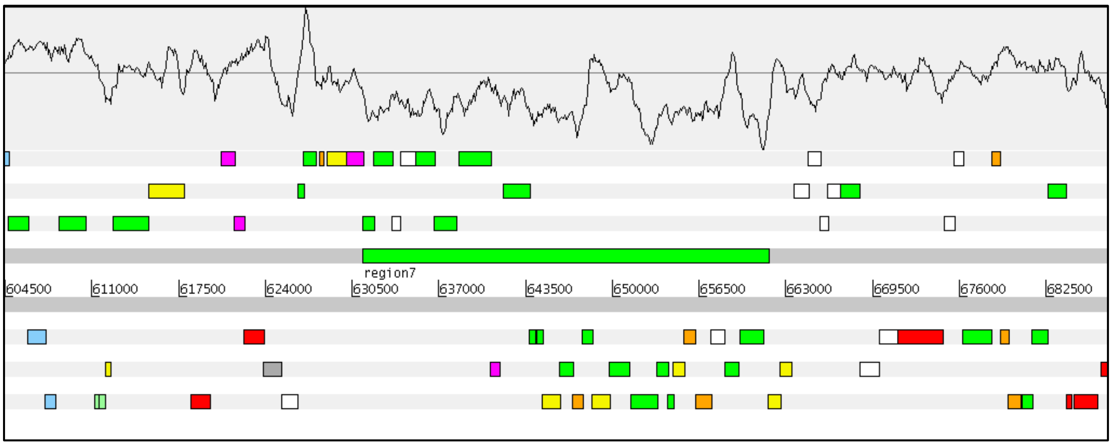


Figure 5.17: Artemis screenshot: Predicted genomic island R7, present in *S. maltophilia* strain K279a. The G+C% content with a window size of 1kb is shown at the top of this screenshot.

The presence of R7 was confirmed for the three K279 strains (Figure 5.18), although a small size variation of the left boundary of this island cannot be excluded in the case of strains 30 and 32; the absence of R7 was not identified in any of the 17 strains.

Failure to identify the left and right boundaries of this island in other strains, suggests that there is variation in the genes present in this locus; extensive variation in these types of loci is well known in other organisms (Bentley *et al.*, 2006).

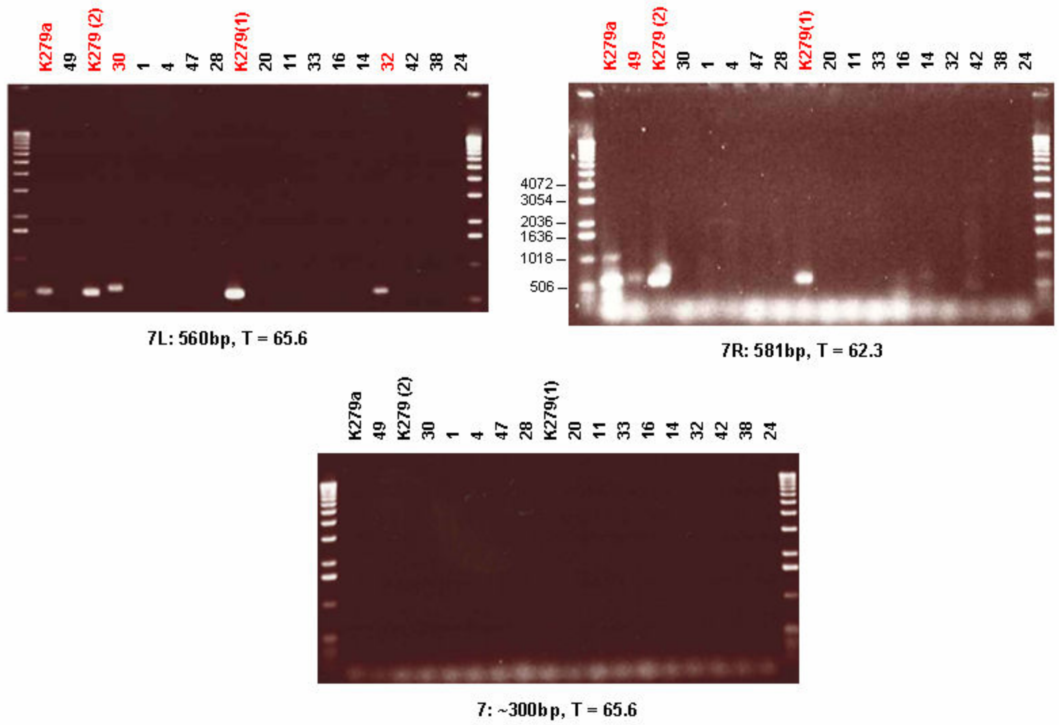


Figure 5.18: PCR amplification of the left (7L) and the right (7R) boundaries (top) of genomic island R7 and of the region across the boundary site of R7 in strains lacking the island (bottom).

5.3.2 Performance benchmarking

5.3.2.1 Prediction accuracy

In order to estimate the accuracy of this GI prediction pipeline I will make the following, three-fold assumption; a predicted region will be considered a true positive (TP) prediction if the following three conditions are met: A. For a predicted candidate GI a PCR product of the expected size, for the left and the right predicted boundary, is observed in at least one of the 17 un-sequenced strains; B. For the same candidate GI a PCR product, of the expected size, for the sequence across the predicted insertion point of this

GI, is observed in at least one of the 17 un-sequenced strains; C. The same predicted GI structure has a posterior probability (of being a true GI) higher than an arbitrarily determined threshold of 0.5.

If only conditions A and C are met, the predicted regions will be considered false positives (FP). If only condition A is met the predicted regions will be considered true negatives (TN). Finally if conditions A and B are met but condition C is not, the predicted regions will be considered false negatives (FN).

Exploiting the above rationale, we can get a naïve estimation of the predictive accuracy of the current pipeline, relying purely on an experimentally validated dataset of eight candidate GIs; five (R1, R12, R14, R15 and R16) of the eight candidate GIs represent TP, one region (R4) is a FP prediction and two (R7 and R20) are TN predictions, yielding a specificity of 0.83 (= TP/(TP+FP) = 5/6), a sensitivity of 1.0 (= TP/(TP+FN) = 5/5) and an overall accuracy of 0.875 (= (TP+TN)/(TP+TN+FP+FN) = 7/8).

5.3.2.2 Boundary accuracy

Assuming that the correct (observed) boundaries of a predicted candidate GI are the ones determined by sequencing across its insertion point in strains lacking the island, we can estimate the prediction accuracy of this methodology in terms of boundary optimization. In the current evaluation, I have used the absolute error defined as $\delta x = | x - x_0 |$, where x is the observed boundary determined by sequencing across the predicted insertion point in strains lacking the island (if applicable) and x_0 is the predicted one; the results (Table 5.6) show that the current methodology that integrates compositional-based (Alien_Hunter) and structural-based (RVM) prediction approaches gives a very small average, absolute error of 21bp; this number is significantly lower than the absolute error (3830bp) of Alien_Hunter (see section 2.3.3 of chapter 2) that relies purely on compositional information.

Table 5.6: Absolute error of the GI prediction pipeline (Alien_Hunter + RVM) for the predicted boundaries of the eight candidate islands.

Region	Boundaries				Absolute error (bp)	
	Left		Right		Left	Right
	Predicted	Observed	Predicted	Observed		
R1	60416	60293	70829	70894	123	65
R16	3089418	3089419	3127149	3127153	1	4
R4	299814	–	335480	–	–	–
R12	1323960	1323958	1367729	1367727	2	2
R14	1720126	1720130	1724413	1724413	4	0
R15	1945402	1945412	2002722	2002722	10	0
R20	3913072	–	3931089	–	–	–
R7	631285	–	661659	–	–	–
ALL (left/right)					28	14.2
ALL (left+right)					21.1	

5.4 Discussion

The aim of this analysis was three-fold. First, a blind-test exploiting an experimentally derived test-dataset of a single sequenced and 17 un-sequenced reference strains was carried out in order to sample the presence or absence of the predicted candidate islands in closely and distantly related *S. maltophilia* isolates; this approach would make it feasible to draw conclusions about the phylogenetic distribution of those putative GIs that in return would confirm or reject their horizontal origin.

Second, the integrative GI prediction pipeline described in this chapter was applied on the newly sequenced, un-annotated genome of *S. maltophilia*, strain K279a and used as a complementary methodology to the annotation pipelines developed in the pathogen sequencing unit (PSU) at the Sanger Institute. Predictions of putative GI structures were used to infer the likely origin of the initially un-annotated CDSs, overlapping with these predictions, as well as to more accurately determine the true boundaries of partially annotated putative horizontally acquired regions. Furthermore, while this anylisis was still in progress, the gene-content information derived from the ongoing annotation of K279a genome put the predicted insertion point of GIs into context; for example, in the case of R1 a set of 18bp DRs were predicted to flank the boundaries of this GI, and based on the gene prediction and subsequent manual curation, it was inferred that the insertion point of R1 was within the coding sequence of Smlt0055; this further suggests that *in silico* predictions and experimental protocols can mutually benefit from each other.

Third, the generalization properties of this prediction pipeline, which integrates compositional-based and structural-based techniques, in making accurate predictions for previously unseen examples of a newly sampled genomic dataset were evaluated, relying purely on an experimental rather than an *in silico* based benchmarking approach. This analysis evaluated two specific properties of the current GI prediction approach; how reliably this methodology predicts GIs in newly sequenced genomes and, for the predictions that are true positives, how accurately their boundaries can be determined.

For a sample of eight candidate GIs with a posterior probability range of 0.2371–0.9997, the data confirm that over half (5/8) of the predictions are likely to be true GI structures that have been probably acquired very recently in the lineage of the three *S. maltophilia* K279 strains. Moreover, the experimental validation of two, very low scoring (0.4983 and 0.2371) predicted GI structures (R20 and R7) suggests that those regions are probably not real GIs, in line with their very low posterior probability; these data confirm the increased specificity of the proposed method in reliably predicting true GIs.

Although the experimental methodology described in this chapter, along with the performance benchmarking, gives results showing a very good overall prediction accuracy for the described approach, even in the case of a previously unseen genomic dataset, there are several obvious limitations affecting the conclusions drawn from this analysis, that have to be taken into account.

Overall the experimental PCR protocol as implemented in the current analysis suffers from low resolution. Firstly, probing the presence or absence of the putative GIs, under the given methodology, is feasible only if the sequence of their predicted boundaries is highly conserved among the reference K279a strain and the 17 un-sequenced *S. maltophilia* strains.

Theoretically speaking, in the case of more distantly related strains this methodology would not necessarily give amplified products for the sequence that corresponds to the predicted GI boundaries since the low level of sequence similarity would prohibit the binding of the corresponding primer set to its genomic DNA template; however, because of the second assumption (section 5.2.3) of the experimental methodology exploited in this analysis, the requirement for an "a+d" amplicon acts as a control, since in the case of distantly related strains, the "a" and "d" primers will also fail to bind to the DNA template and give an amplified product.

An alternative PCR approach that could overcome this limitation, would involve the design of degenerate primers that would allow sequence ambiguity between the primers and the template. However, the results of this analysis suggest that this is probably not the case for the given genomic dataset of the 18 *S. maltophilia* strains since PCR amplified products, of the expected size, are successfully produced even in the case of distantly related isolates. For example the results in Figure 5.10 show that for phylogenetically distantly related strains (Figure 5.2), e.g. strain 11 (group IV) and 20 (group II) a PCR product of the expected size for the sequence across the insertion point of R12 was successfully obtained.

Secondly, this methodology does not provide any information about the actual gene content, size and internal structural variation of GIs inferred to be present in any of the 17 un-sequenced strains. For example a predicted GI of putative phage origin might have similar bacteriophage integrase and tail protein coding CDSs at the two boundaries with an inferred "identical" GI structure present in some of the un-sequenced genomes; clearly prophages of different type or family can have high sequence similarity at those flanking CDSs but do not necessarily represent the same prophage. In other words sequence similarity at the predicted boundaries between genomic regions present in different strains neither guarantees that those regions are of the same origin, or gene content, nor does it exclude internal size variation, e.g. in the case of GI remnants, or deletions.

An alternative, more sophisticated approach that would overcome those limitations is the Southern blotting protocol (Southern, 1975) that exploits a probe hybridization principle; however such a methodology is out of the scope of this analysis, for reasons discussed at the beginning of this chapter; it is worth mentioning that the protocol used in this analysis was only devised to check the predicted boundaries of GIs and not to completely explore the content of the GIs.

Thirdly, in the case of probing the absence of a given candidate GI in some of the un-sequenced strains by seeking to amplify the sequence across the predicted insertion point of this GI, again this methodology will fail to give an amplified product if a different GI has been inserted at the corresponding insertion point in the target strains. In that case, we will not be able to infer that the reference GI is absent from the target strains, although this is clearly the case. An alternative methodology would involve a long-range PCR protocol that could amplify longer genomic regions; however the results of this approach would still be conditional on the size of the intervening sequence between the left and the right ends of the corresponding insertion point.

Clearly the current experimental methodology exploits very simple concepts and principles and as such it provides a very rough evaluation of the true strengths and weaknesses of the discussed *in silico* pipeline. Nonetheless, this analysis forms a proof of concept that the *in silico* prediction of GIs can be integrated successfully in experimental methodologies and gives data suggesting that some of the *in silico* predictions have probably limited phylogenetic distribution and represent putative recent horizontal gene transfer events in the *S. maltophilia* lineage.

Moreover, the data presented in the current analysis show that prediction pipelines that merge compositional-based (low-level) with structural-based (high-level) approaches can yield more reliable predictions of putative GIs compared to methodologies exploiting either of those approaches. Overall it can be concluded that *in silico* prediction methods, relying on and exploiting a minimum level of pre-existing annotation, can be very powerful tools in aiding or guiding, in a high-throughput fashion, the annotation pipelines of microbial genomes (see next chapter).

Chapter 6

Discussion

6.1 Conclusions

In this thesis, I have introduced and discussed a multi-factorial methodology for modelling, predicting and analyzing mobile genetic elements, present in microbial genomes, termed genomic islands (GIs). The reason that this project was chosen, is the observation that GIs drive accelerated rates of evolution (Groisman and Ochman, 1996) in microbial populations that in return shape host-pathogen interactions, adaptations to specific niches and the overall population structure, in a way fundamentally different from the biological processes and dynamics shaping eukaryotic genomes.

In order to predict and study GIs I firstly introduced a novel compositional-based algorithm (chapter 2), exploiting the principle that at the time of insertion, horizontally acquired genomic DNA carries the sequence signature of its donor and often deviates compositionally from the sequence signature of its new host. Although this assumption might not hold in many cases (e.g. in the case of compositionally similar donor and host genomes, host genes under functional constraints and horizontally acquired genes that have converged to the host composition due to the time-dependent process of amelioration), it can be tolerated for the sake of developing unsupervised algorithms that can be directly applied on raw genomic datasets, with a minimal (if any) level of annotation (Table 6.1).

The novelty of this methodology relies on the fact that it exploits a new compositional algorithm, i.e. the Interpolated Variable Order Motifs (IVOMs) that overcomes the limitations of pre-existing, fixed (low or high) order compositional based methodologies, by introducing an interpolated variable order approach in analyzing local compositional biases. Under this principle, no *a priori* assumption is made about the order of the compositional distribution that best captures departures from the genome backbone compositional distribution.

Obviously relying more on a higher level of annotation (e.g. gene prediction and functional/structural annotation) more accurate algorithms, that capture more reliably the true origin of putative horizontally acquired genomic regions, can be devised. Exploiting this principle, I introduced in chapter 4 a machine learning approach that quantifies our posterior belief that a genomic structure is likely to be a true GI.

This methodology did not make any *a priori* assumptions about the structure of GIs, but instead implemented a bottom-up search, sampling both putative GIs and non-GI genomic regions from Gram positive and Gram negative bacteria, rather than relying on a previous GI structural definition (Hacker *et al.*, 1997). The data showed that GIs represent a superfamily of mobile elements with core and variable structural features, characterized by increased structural variation, approaching probably a structural continuum, under which families and subfamilies are distinguishable but also conditional on the assumptions made and the arbitrarily chosen criteria used.

The novelty of this methodology relies on the fact that traditional machine learning approaches were exploited under a "forward-reverse" concept; a training dataset was used to train structural GI models, and those models were exploited not only to make predictions ("forward" implementation) on unseen examples, but most importantly to use their estimated parameters (weights) in "reverse" to draw conclusions about the structural variation of GIs.

Although the benchmarking analysis showed that structural-based predictions of GIs can be more reliable than methodologies exploiting purely compositional based information, they form supervised solutions that require a higher level of annotation (Table 6.1).

A feature of GIs, independent of any *a priori* compositional or structural assumption, is their horizontal origin, i.e. GIs are horizontally acquired mobile elements of limited phylogenetic distribution. Exploiting this principle in chapter 3 I discussed a comparative-based approach for the prediction of GIs, the modelling of their compositional amelioration over time, and the mapping of their inferred relative time of acquisition on the phylogenetic history of the reference genomic dataset.

Comparative based methodologies, applied in the prediction of GIs, can be more accurate and reliable than structural and compositional based approaches, purely due to the fact that they make no *a priori* assumptions about how GIs should "look"; instead they utilize information about a more fundamental property, i.e. their origin. However comparative-based methods, require a very wide (sequenced) species sample, a prerequisite that might well prohibit the application of such approaches in the case of species with very few sequenced representatives (Table 6.1).

It becomes obvious that in the case of predicting genetic elements characterized by increased levels of mobility, exploiting information (e.g. composition) derived from a single genome sequence provides only a very narrow and static "snapshot" of their mobile life and history. On the other hand, capturing a dynamic rather than a static picture of a bacterial population, allowing inter- and intra-species genetic-flux (i.e. gene loss, gene gain, duplication, recombination and chromosomal-rearrangements), key evolutionary steps and host adaptations to be explicitly modelled, provides a more reliable description of those highly mobile genetic elements. Under this "genetic-flux" framework, a more comprehensive picture of bacterial populations can be built taking into account both static (e.g. sequence information) and dynamic (i.e. genetic-flux) parameters (see future work section below).

In chapter 5, I carried out a blind-test, applying, in an integrative fashion, the compositional and the structural-based techniques described in the previous chapters on a newly sequenced, un-annotated genome with the specific aim of performing a "real-life" implementation of this prediction pipeline utilizing only the minimum level of information, i.e.

raw genomic sequence. Exploiting an experimentally validated test dataset, I discussed results showing that such methodologies can be directly applicable on genomic datasets even at the very early stages of the annotation pipelines, acting as complementary tools to the currently existing annotation methods.

Table 6.1: Properties of three different *in silico* methods developed and discussed in this thesis, for the analysis and study of Genomic Islands.

Method	Annotation level	Information	Chapter	Pros	Cons
Alien Hunter	Low	Composition	2	▪ Automated ▪ Fast ▪ Unsupervised ▪ Applicable on newly sampled and sequenced, un-annotated genomic datasets	▪ Composition might "lie" (compositionally similar donors-hosts, genes under functional constrains, amelioration)
RVM	Medium	Structure	4	▪ Very fast ▪ Reliable ▪ Good generalization properties	▪ Supervised ▪ Requires known examples to form the training dataset ▪ Requires structural annotation
Phylogenetic tree	High	Gene content, phylogenetic distribution	3	▪ Very reliable predictions if the correct model of evolution is applied ▪ Gives estimates about the relative time of acquisition ▪ Allows mapping of key evolutionary events on the phylogenetic history of the genomes of interest	▪ Time consuming (phylogenies) ▪ Manual curation ▪ Requires pre-existing sequenced closely and distantly related genomes

6.2 Future work

Although methodologies exploiting the dynamic properties of bacterial populations have just started to emerge (Daubin and Ochman, 2004; Didelot *et al.*, 2007; Fuxelius *et al.*, 2008; Vernikos *et al.*, 2007) providing a step-wise decomposition of the evolutionary history of species over time, and revealing key evolutionary events that drive host-adaptation and pathogenicity, they are still far from being complete, efficiently automated, standardized and high-throughput.

This challenge could well form the focus of a future project; to perform a bioinformatic whole-genome based comparative study of bacterial genomes in order to quantify explicitly inter- and intra-species differences and interactions. The results could be used to implement a high-throughput *in silico* platform for fast and reliable step-wise decomposition of the evolutionary history of bacterial populations, focused on identifying virulence genes and potential vaccine candidates (Vernikos, 2008). In the following sections I provide a brief outline of how this methodology could be implemented.

6.2.1 High-throughput modelling of genetic flux

6.2.1.1 Selection of bacterial genomes

For the purposes of studying inter- and intra-species genetic-flux, a set of query as well as outgroup genomes is needed. Two options can be exploited:

A. Manual: The user can select manually a set of species and outgroup representative genomes, based on prior knowledge.

B. Composition-based: Variable-order compositional distributions can be used, implementing the Interpolated Variable Order Motifs (IVOMs) theory (Vernikos and Parkhill, 2006). IVOMs is a very powerful and sensitive method that can reliably estimate the relatedness, by means of compositional analysis, of different closely or distantly related bacterial chromosomes, overcoming the limitations of fixed-order compositional indices (e.g. % G+C content). Its increased resolution can discriminate even very similar genomes e.g. of the same serovar, while the fact that it is alignment-free makes it efficiently fast and automated. This method can automatically select appropriate closely and distantly related (i.e. outgroup) genomes for a reliable study of genetic-flux.

6.2.1.2 Whole-genome, all-against-all comparative analysis

A. Orthologous genes: In order to identify orthologous genes, each genome in the dataset can be compared against all the other genomes, by means of a best reciprocal FASTA (Pearson, 1990) approach. Although this methodology has been optimized and fine-tuned to predict reliably orthologous genes (Bentley *et al.*, 2007; Thomson *et al.*, 2006; Vernikos *et al.*, 2007), the best matches between genes of the different genomes may well be paralogs rather than true orthologs. However this limitation is a desired property in the current methodology; gene duplication is part of the genetic-flux concept and such prediction ambiguities can be analyzed in a second step taking into account their syntenic relationship to differentiate true orthologs from paralogs.

B. Phyletic profile: From the above all-against-all comparison the different patterns of presence or absence (i.e. phyletic profile) of all the genes in the pan-genome (i.e. the genome of a bacterial species consisting of core and dispensable genes, (Medini *et al.*, 2005)), can be grouped and coded in a binary fashion, i.e. [1,0] to denote [presence, absence] respectively. The phyletic profile can be analyzed for the purpose of a three-fold strategy:

1. The patterns of gene presence or absence can be grouped into core (shared among all genomes) and dispensable (partially shared and strain-specific) gene sets; modelling the number of strain-specific genes in the pan-genome as a function of adding step-wise new genomes, could enable us to draw conclusions about the pan-genome properties (i.e. open or closed pan-genome) and its rate of growth (Tettelin *et al.*, 2005).

2. The phyletic profile can be used to build the phylogenetic tree of the dataset relying on an alignment-free, distance-based approach (Fitz-Gibbon and House, 1999; Snel *et al.*, 1999). The phyletic profile can be converted into a distance matrix, in which the distance will reflect the fraction of genes that two genomes have in common. This alignment-free methodology is key for the development of a high-throughput approach since it is very fast compared to sequence-based techniques, exploits the

entire pan-genome and takes into account the various aspects of genetic-flux.

3. The phylogenetic tree of the dataset can be exploited as the reference tree topology, in order to infer putative gene gain and gene loss events, analyzing the phyletic profile by means of a maximum parsimony model (Mirkin *et al.*, 2003; Vernikos *et al.*, 2007). This methodology will enable us to estimate the relative time (Daubin and Ochman, 2004; Vernikos *et al.*, 2007) and rate of gene-transfer events on branches of increasing depth within the tree, revealing potential key host-adaptation strategies, e.g. genome-degradation (Gomez-Valero *et al.*, 2007; Parkhill *et al.*, 2001).

C. Recombination events: The first step for the detection of putative recombination events can be based on the following assumption: if the topology of individual gene trees is statistically different from the reference tree topology of the entire dataset, those genes can be considered candidates for inter or intra-species recombination (Dykhuizen and Green, 1991; Feil *et al.*, 2001).

In a second step, a sliding window can be exploited to analyze local discrepancies in the sequence similarity of consecutive genes with their corresponding orthologs in the other genomes. Significantly different (higher or lower) sequence similarity not expected by chance after evaluating the gene neighbourhood of the query and the target genomes can be combined with violations of the reference tree topology (previous step) in order to determine the possible direction of recombination (i.e. inter- or intra-species).

D. Chromosomal rearrangements: Analyzing the co-linearity of the orthologous gene sets between two genomes will enable us to detect “breaks” in the syntenic relationship between the two chromosomes and infer possible large-scale rearrangements (e.g. inversions) (Eisen *et al.*, 2000; Liu and Sanderson, 1995; Tillier and Collins, 2000); their location relative to the terminus and the origin of replication could reveal the level of selective pressure for maintaining the genome order.

6.2.1.3 Quantification of genetic-flux

Generalized Linear Models (GLMs) (McCullagh and Nelder, 1989) can be used to build species and cross-species specific models of genetic-flux, quantifying the genome fluidity of bacterial populations. In the current framework, each GLM will be the weighted sum of K basis functions, where K denotes the different parameters of genetic-flux (e.g. gene gain, gene loss, duplication, chromosomal rearrangements, and recombination) used to describe a bacterial population exploiting a generalized genetic-flux alphabet; a similar approach to that described in chapter 4.

In the current genetic-flux framework, GLMs can be trained using species and cross-species genomes, quantifying explicitly under a probabilistic framework the contribution of each of the genetic-flux parameters in shaping the dynamic structure of specific bacterial populations. Consequently each GLM will provide in a single linear equation a step-wise decomposition of the evolutionary history of those bacterial populations. The gene-flux GLMs can be used in a machine learning method in order to evaluate how reliably genomic datasets can be classified into different bacterial species, based on their genetic-flux profile. Misclassifications, due to overlapping genetic-flux properties of seemingly distinct bacterial species can be further analyzed to re-evaluate the relatedness of the latter.

6.2.1.4 Biological significance

The results of this study could be directly applicable to: 1. The identification and classification of different or similar adaptation mechanisms to the same or different hosts, respectively.

2. The study and characterization of the genetic boundaries between free-living and host-adapted bacteria, as well as between pathogenic and commensal bacteria.

3. Defining the minimum number of species isolates to be sequenced in order to have a reliable sample of the diversity of a given bacterial population (open or closed pan-genome).

4. Guiding the identification of new vaccine candidates, using the concept of “reverse vaccinology" (Rappuoli, 2000; Rappuoli and Covacci, 2003); whereby comparative genomics has enabled the successful development of novel vaccines against major pathogens (Behr *et al.*, 1999; Maione *et al.*, 2005; Pizza *et al.*, 2000).

5. Quantifying explicitly the genetic-fluidity of bacterial species using a single, linear equation. Utilising a generalized gene-flux alphabet, new whole-genome based classification systems can be devised.

6. Mapping the relative time of gene transfer events from the evolutionary history of bacteria to the evolutionary history of their host, enabling us to begin to understand how the interactions between key gene-transfer events in the evolution of pathogenic bacteria (Parkhill *et al.*, 2001) and behavioural or demographic changes in their host population (Thomson *et al.*, 2008), lead to the emergence of novel pathogens.

6.3 Final remarks

To end, I would like to make a comment on the application of quantitative or qualitative models in modern biology. Initially, when the very first steps towards understanding the rules and principles that govern biological systems were made, simplistic assumptions had to be introduced, to keep the complexity of the hypotheses low enough for biologists to be able to draw valuable and, most importantly, interpretable conclusions. During the last ten years, or so, the transition from single-isolate genomics to comparative genomics of entire biological populations, has introduced new (previously unknown) parameters that in some cases threaten to question or even to reject our initial assumptions about fundamental biological concepts and definitions. For example, in the current context of increased microbial genome fluidity, the fundamental definition of the biological species (Mayr, 1942), does not provide a realistic and representative description of the dynamic relationships that shape microbial evolution. Moving from intuition-driven or even macroscopic observation-driven hypotheses to data-driven hypotheses represents a more realistic approach in the study of biological systems, even when this requires revisiting and perhaps rejecting our initial, intuitively correct but biologically erroneous assumptions and definitions; a recent example, derived from microbial populations that extensively exchange genetic material, involves the rejection of the strictly bifurcating tree of life (Darwin, 1859) by a more realistic model-structure, that of the reticulate phylogenetic network (Huson and Bryant, 2006).

Bibliography

Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A. and Carniel, E. (1999) Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis, *Proc Natl Acad Sci U S A*, **96**, 14043-14048.

Ahmed, A.M., Hussein, A.I. and Shimamoto, T. (2007) Proteus mirabilis clinical isolate harbouring a new variant of Salmonella genomic island 1 containing the multiple antibiotic resistance region, *J Antimicrob Chemother*, **59**, 184-190.

Ahmer, B.M., van Reeuwijk, J., Watson, P.R., Wallis, T.S. and Heffron, F. (1999) Salmonella SirA is a global regulator of genes mediating enteropathogenesis, *Mol Microbiol*, **31**, 971-982.

Alonso, A. and Martinez, J.L. (2000) Cloning and characterization of SmeDEF, a novel multidrug efflux pump from Stenotrophomonas maltophilia, *Antimicrob Agents Chemother*, **44**, 3079-3086.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*, **25**, 3389-3402.

Anderson, N.G. (1970) Evolutionary significance of virus infection, *Nature*, **227**, 1346-1347.

Andersson, J.O. and Andersson, S.G. (1999) Insights into the evolutionary process of genome degradation, *Curr Opin Genet Dev*, **9**, 664-671.

Andersson, J.O., Doolittle, W.F. and Nesbo, C.L. (2001) Genomics. Are there bugs in our genome?, *Science*, **292**, 1848-1850.

Anthony, K.G., Sherburne, C., Sherburne, R. and Frost, L.S. (1994) The role of the pilus in recipient cell recognition during bacterial conjugation mediated by F-like plasmids, *Mol Microbiol*, **13**, 939-953.

Antonenka, U., Nolting, C., Heesemann, J. and Rakin, A. (2005) Horizontal transfer of Yersinia high-pathogenicity island by the conjugative RP4 attB target-presenting shuttle plasmid, *Mol Microbiol*, **57**, 727-734.

Arora, S.K., Bangera, M., Lory, S. and Ramphal, R. (2001) A genomic island in Pseudomonas aeruginosa carries the determinants of flagellin glycosylation, *Proc Natl Acad Sci U S A*, **98**, 9342-9347.

Azad, R.K. and Lawrence, J.G. (2005) Use of Artificial Genomes in Assessing Methods for Atypical Gene Detection, *PLoS Comput Biol*, **1**, e56.

Bach, S., de Almeida, A. and Carniel, E. (2000) The Yersinia high-pathogenicity island is present in different members of the family Enterobacteriaceae, *FEMS Microbiol Lett*, **183**, 289-294.

Bajaj, V., Lucas, R.L., Hwang, C. and Lee, C.A. (1996) Co-ordinate regulation of Salmonella typhimurium invasion genes by environmental and regulatory factors is mediated by control of hilA expression, *Mol Microbiol*, **22**, 703-714.

Baldwin, A., Sokol, P.A., Parkhill, J. and Mahenthiralingam, E. (2004) The Burkholderia cepacia epidemic strain marker is part of a novel genomic island encoding both virulence and metabolism-associated genes in Burkholderia cenocepacia, *Infect Immun*, **72**, 1537-1547.

Banks, D.J., Lei, B. and Musser, J.M. (2003) Prophage induction and expression of prophage-encoded virulence factors in group A Streptococcus serotype M3 strain MGAS315, *Infect Immun*, **71**, 7079-7086.

Baum, L.E. (1972) An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process, *Inequalities*, **627**, 1-8.

Baumler, A.J. (1997) The record of horizontal gene transfer in Salmonella, *Trends Microbiol*, **5**, 318-322.

Beaber, J.W., Hochhut, B. and Waldor, M.K. (2002) Genomic and functional analyses of SXT, an integrating antibiotic resistance gene transfer element derived from Vibrio cholerae, *J Bacteriol*, **184**, 4259-4269.

Behr, M.A., Wilson, M.A., Gill, W.P., Salamon, H., Schoolnik, G.K., Rane, S. and Small, P.M. (1999) Comparative genomics of BCG vaccines by whole-genome DNA microarray, *Science*, **284**, 1520-1523.

Bennetzen, J.L. and Hall, B.D. (1982) Codon selection in yeast, *J Biol Chem*, **257**, 3026-3031.

Bentley, S.D., Aanensen, D.M., Mavroidi, A., Saunders, D., Rabbinowitsch, E., Collins, M., Donohoe, K., Harris, D., Murphy, L., Quail, M.A., Samuel, G., Skovsted, I.C., Kaltoft, M.S., Barrell, B., Reeves, P.R., Parkhill, J. and Spratt, B.G. (2006) Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes, *PLoS Genet*, **2**, e31.

Bentley, S.D., Vernikos, G.S., Snyder, L.A., Churcher, C., Arrowsmith, C., Chillingworth, T., Cronin, A., Davis, P.H., Holroyd, N.E., Jagels, K., Maddison, M., Moule, S., Rabbinowitsch, E., Sharp, S., Unwin, L., Whitehead, S., Quail, M.A., Achtman, M., Barrell, B., Saunders, N.J. and Parkhill, J. (2007) Meningococcal Genetic Variation Mechanisms Viewed through Comparative Analysis of Serogroup C Strain FAM18, *PLoS Genet*, **3**, e23.

Beres, S.B., Richter, E.W., Nagiec, M.J., Sumby, P., Porcella, S.F., DeLeo, F.R. and Musser, J.M. (2006) Molecular genetic anatomy of inter- and intraserotype variation in the human bacterial pathogen group A Streptococcus, *Proc Natl Acad Sci U S A*, **103**, 7059-7064.

Bergthorsson, U. and Ochman, H. (1998) Distribution of chromosome length variation in natural isolates of Escherichia coli, *Mol Biol Evol*, **15**, 6-16.

Bjorkholm, B., Sjolund, M., Falk, P.G., Berg, O.G., Engstrand, L. and Andersson, D.I. (2001) Mutation frequency and biological cost of antibiotic resistance in Helicobacter pylori, *Proc Natl Acad Sci U S A*, **98**, 14607-14612.

Blanc-Potard, A.B., Solomon, F., Kayser, J. and Groisman, E.A. (1999) The SPI-3 pathogenicity island of Salmonella enterica, *J Bacteriol*, **181**, 998-1004.

Blattner, F.R., Plunkett, G., 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. and Shao, Y. (1997) The complete genome sequence of Escherichia coli K-12, *Science*, **277**, 1453-1474.

Blum, G., Ott, M., Lischewski, A., Ritter, A., Imrich, H., Tschape, H. and Hacker, J. (1994) Excision of large DNA regions termed pathogenicity islands from tRNA-specific loci in the chromosome of an Escherichia coli wild-type pathogen, *Infect Immun*, **62**, 606-614.

Bohr, N. (1913) On the constitution of atoms and molecules, *Philosophical magazine*, **26**, 1-25.

Bolotin, A., Quinquis, B., Renault, P., Sorokin, A., Ehrlich, S.D., Kulakauskas, S., Lapidus, A., Goltsman, E., Mazur, M., Pusch, G.D., Fonstein, M., Overbeek, R., Kyprides, N., Purnelle, B., Prozzi, D., Ngui, K., Masuy, D., Hancy, F., Burteau, S., Boutry, M., Delcour, J., Goffeau, A. and Hols, P. (2004) Complete sequence and comparative genome analysis of the dairy bacterium Streptococcus thermophilus, *Nat Biotechnol*, **22**, 1554-1558.

Bolotin, A., Wincker, P., Mauger, S., Jaillon, O., Malarme, K., Weissenbach, J., Ehrlich, S.D. and Sorokin, A. (2001) The complete genome sequence of the lactic acid bacterium Lactococcus lactis ssp. lactis IL1403, *Genome Res*, **11**, 731-753.

Boucher, Y., Douady, C.J., Papke, R.T., Walsh, D.A., Boudreau, M.E., Nesbo, C.L., Case, R.J. and Doolittle, W.F. (2003) Lateral gene transfer and the origins of prokaryotic groups, *Annu Rev Genet*, **37**, 283-328.

Bourzac, K.M. and Guillemin, K. (2005) Helicobacter pylori-host cell interactions mediated by type IV secretion, *Cell Microbiol*, **7**, 911-919.

Boyd, D., Cloeckaert, A., Chaslus-Dancla, E. and Mulvey, M.R. (2002) Characterization of variant Salmonella genomic island 1 multidrug resistance regions from serovars Typhimurium DT104 and Agona, *Antimicrob Agents Chemother*, **46**, 1714-1722.

Boyd, D., Peters, G.A., Cloeckaert, A., Boumedine, K.S., Chaslus-Dancla, E., Imberechts, H. and Mulvey, M.R. (2001) Complete nucleotide sequence of a 43-kilobase genomic island associated with the multidrug resistance region of Salmonella enterica serovar Typhimurium DT104 and its identification in phage type DT120 and serovar Agona, *J Bacteriol*, **183**, 5725-5732.

Brochier, C., Philippe, H. and Moreira, D. (2000) The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome, *Trends Genet*, **16**, 529-533.

Broker, G. and Spellerberg, B. (2004) Surface proteins of Streptococcus agalactiae and horizontal gene transfer, *Int J Med Microbiol*, **294**, 169-175.

Broudy, T.B., Pancholi, V. and Fischetti, V.A. (2001) Induction of lysogenic bacteriophage and phage-associated toxin from group a streptococci during coculture with human pharyngeal cells, *Infect Immun*, **69**, 1440-1443.

Brussow, H. and Hendrix, R.W. (2002) Phage Genomics: Small Is Beautiful, *Cell*, **108**, 13-16.

Buchanan-Wollaston, V., Passiatore, J.E. and Cannon, F. (1987) The mob and oriT mobilization functions of a bacterial plasmid promote its transfer to plants, *Nature*, **328**, 172-175.

Buchrieser, C., Brosch, R., Bach, S., Guiyoule, A. and Carniel, E. (1998) The high-pathogenicity island of Yersinia pseudotuberculosis can be inserted into any of the three chromosomal asn tRNA genes, *Mol Microbiol*, **30**, 965-978.

Bueno, S.M., Santiviago, C.A., Murillo, A.A., Fuentes, J.A., Trombert, A.N., Rodas, P.I., Youderian, P. and Mora, G.C. (2004) Precise excision of the large pathogenicity island, SPI7, in Salmonella enterica serovar Typhi, *J Bacteriol*, **186**, 3202-3213.

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA, *J Mol Biol*, **268**, 78-94.

Burrus, V., Marrero, J. and Waldor, M.K. (2006) The current ICE age: biology and evolution of SXT-related integrating conjugative elements, *Plasmid*, **55**, 173-183.

Burrus, V., Pavlovic, G., Decaris, B. and Guedon, G. (2002) Conjugative transposons: the tip of the iceberg, *Mol Microbiol*, **46**, 601-610.

Burrus, V. and Waldor, M.K. (2003) Control of SXT integration and excision, *J Bacteriol*, **185**, 5045-5054.

Burrus, V. and Waldor, M.K. (2004) Shaping bacterial genomes with integrative and conjugative elements, *Res Microbiol*, **155**, 376-386.

Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M.L. and Brussow, H. (2003) Phage as agents of lateral gene transfer, *Curr Opin Microbiol*, **6**, 417-424.

Carattoli, A., Filetici, E., Villa, L., Dionisi, A.M., Ricci, A. and Luzzi, I. (2002) Antibiotic resistance genes and Salmonella genomic island 1 in Salmonella enterica serovar Typhimurium isolated in Italy, *Antimicrob Agents Chemother*, **46**, 2821-2828.

Carniel, E. (1999) The Yersinia high-pathogenicity island, *Int Microbiol*, **2**, 161-167.

Carniel, E. (2001) The Yersinia high-pathogenicity island: an iron-uptake island, *Microbes Infect*, **3**, 561-569.

Carniel, E., Guilvout, I. and Prentice, M. (1996) Characterization of a large chromosomal "high-pathogenicity island" in biotype 1B Yersinia enterocolitica, *J Bacteriol*, **178**, 6743-6751.

Carter, D. and Durbin, R. (2006) Vertebrate gene finding from multiple-species alignments using a two-level strategy, *Genome Biol*, **7 Suppl 1**, S6 1-12.

Carver, T.J., Rutherford, K.M., Berriman, M., Rajandream, M.A., Barrell, B.G. and Parkhill, J. (2005) ACT: the Artemis Comparison Tool, *Bioinformatics*, **21**, 3422-3423.

Censini, S., Lange, C., Xiang, Z., Crabtree, J.E., Ghiara, P., Borodovsky, M., Rappuoli, R. and Covacci, A. (1996) cag, a pathogenicity island of Helicobacter pylori, encodes type I-specific and disease-associated virulence factors, *Proc Natl Acad Sci U S A*, **93**, 14648-14653.

Chakravortty, D., Hansen-Wester, I. and Hensel, M. (2002) Salmonella pathogenicity island 2 mediates protection of intracellular Salmonella from reactive nitrogen intermediates, *J Exp Med*, **195**, 1155-1166.

Chen, I. and Dubnau, D. (2004) DNA uptake during bacterial transformation, *Nat Rev Microbiol*, **2**, 241-249.

Chiu, C.H., Tang, P., Chu, C., Hu, S., Bao, Q., Yu, J., Chou, Y.Y., Wang, H.S. and Lee, Y.S. (2005) The genome sequence of Salmonella enterica serovar Choleraesuis, a highly invasive and resistant zoonotic pathogen, *Nucleic Acids Res*, **33**, 1690-1698.

Choi, J., Shin, D. and Ryu, S. (2007) Implication of quorum sensing in Salmonella enterica serovar typhimurium virulence: the luxS gene is necessary for expression of genes in pathogenicity island 1, *Infect Immun*, **75**, 4885-4890.

Chua, K.Y., Pankhurst, C.E., Macdonald, P.E., Hopcroft, D.H., Jarvis, B.D. and Scott, D.B. (1985) Isolation and characterization of transposon

Tn5-induced symbiotic mutants of Rhizobium loti, *J Bacteriol*, **162**, 335-343.

Cloeckaert, A., Praud, K., Doublet, B., Demartin, M. and Weill, F.X. (2006) Variant Salmonella genomic island 1-L antibiotic resistance gene cluster in Salmonella enterica serovar Newport, *Antimicrob Agents Chemother*, **50**, 3944-3946.

Cloeckaert, A., Sidi Boumedine, K., Flaujac, G., Imberechts, H., D'Hooghe, I. and Chaslus-Dancla, E. (2000) Occurrence of a Salmonella enterica serovar typhimurium DT104-like antibiotic resistance gene cluster including the floR gene in S. enterica serovar agona, *Antimicrob Agents Chemother*, **44**, 1359-1361.

Coetzee, J.N., Datta, N. and Hedges, R.W. (1972) R factors from Proteus rettgeri, *J Gen Microbiol*, **72**, 543-552.

Couturier, M.R., Tasca, E., Montecucco, C. and Stein, M. (2006) Interaction with CagF is required for translocation of CagA into the host via the Helicobacter pylori type IV secretion system, *Infect Immun*, **74**, 273-281.

Crabtree, J.E., Covacci, A., Farmery, S.M., Xiang, Z., Tompkins, D.S., Perry, S., Lindley, I.J. and Rappuoli, R. (1995) Helicobacter pylori induced interleukin-8 expression in gastric epithelial cells is associated with CagA positive phenotype, *J Clin Pathol*, **48**, 41-45.

Darling, A.C., Mau, B., Blattner, F.R. and Perna, N.T. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements, *Genome Res*, **14**, 1394-1403.

Darwin, C. (1859) *On the origin of species by means of natural selection.* J. Murray, London.

Daubin, V., Moran, N.A. and Ochman, H. (2003) Phylogenetics and the cohesion of bacterial genomes, *Science*, **301**, 829-832.

Daubin, V. and Ochman, H. (2004) Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli, *Genome Res*, **14**, 1036-1042.

Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978) A model of evolutionary change in proteins. In Dayhoff, M.O. (ed), *Atlas of protein sequence and structure*. National Biomedical Research Foundation, Washington DC, 345-352.

Deng, W., Li, Y., Vallance, B.A. and Finlay, B.B. (2001) Locus of enterocyte effacement from Citrobacter rodentium: sequence analysis and evidence for horizontal transfer among attaching and effacing pathogens, *Infect Immun*, **69**, 6323-6335.

Deng, W., Liou, S.R., Plunkett, G., 3rd, Mayhew, G.F., Rose, D.J., Burland, V., Kodoyianni, V., Schwartz, D.C. and Blattner, F.R. (2003) Comparative genomics of Salmonella enterica serovar Typhi strains Ty2 and CT18, *J Bacteriol*, **185**, 2330-2337.

Denton, M. and Kerr, K.G. (1998) Microbiological and clinical aspects of infection associated with Stenotrophomonas maltophilia, *Clin Microbiol Rev*, **11**, 57-80.

Deppenmeier, U., Johann, A., Hartsch, T., Merkl, R., Schmitz, R.A., Martinez-Arias, R., Henne, A., Wiezer, A., Baumer, S., Jacobi, C., Bruggemann, H., Lienard, T., Christmann, A., Bomeke, M., Steckel, S., Bhattacharyya, A., Lykidis, A., Overbeek, R., Klenk, H.P., Gunsalus, R.P., Fritz, H.J. and Gottschalk, G. (2002) The genome of Methanosarcina mazei: evidence for lateral gene transfer between bacteria and archaea, *J Mol Microbiol Biotechnol*, **4**, 453-461.

Didelot, X., Achtman, M., Parkhill, J., Thomson, N.R. and Falush, D. (2007) A bimodal pattern of relatedness between the Salmonella Paratyphi A and Typhi genomes: convergence or divergence by homologous recombination?, *Genome Res*, **17**, 61-68.

Diep, B.A., Gill, S.R., Chang, R.F., Phan, T.H., Chen, J.H., Davidson, M.G., Lin, F., Lin, J., Carleton, H.A., Mongodin, E.F., Sensabaugh, G.F. and Perdreau-Remington, F. (2006) Complete genome sequence of

USA300, an epidemic clone of community-acquired meticillin-resistant Staphylococcus aureus, *Lancet*, **367**, 731-739.

Dimopoulou, I.D., Kraak, W.A., Anderson, E.C., Nichols, W.W., Slack, M.P. and Crook, D.W. (1992) Molecular epidemiology of unrelated clusters of multiresistant strains of Haemophilus influenzae, *J Infect Dis*, **165**, 1069-1075.

Dirac, P.A.M. (1958) *The Principles of Quantum Mechanics*. Oxford University Press, New York.

Donnenberg, M.S., Tzipori, S., McKee, M.L., O'Brien, A.D., Alroy, J. and Kaper, J.B. (1993) The role of the eae gene of enterohemorrhagic Escherichia coli in intimate attachment in vitro and in a porcine model, *J Clin Invest*, **92**, 1418-1424.

Doolittle, R.F., Feng, D.F., Anderson, K.L. and Alberro, M.R. (1990) A naturally occurring horizontal gene transfer from a eukaryote to a prokaryote, *J Mol Evol*, **31**, 383-388.

Doolittle, R.F., Feng, D.F., Tsang, S., Cho, G. and Little, E. (1996) Determining divergence times of the major kingdoms of living organisms with a protein clock, *Science*, **271**, 470-477.

Doolittle, W.F. (1999) Lateral genomics, *Trends Cell Biol*, **9**, M5-8.

Doolittle, W.F. (1999) Phylogenetic classification and the universal tree, *Science*, **284**, 2124-2129.

Doolittle, W.F. and Papke, R.T. (2006) Genomics and the bacterial species problem, *Genome Biol*, **7**, 116.

Doublet, B., Boyd, D., Mulvey, M.R. and Cloeckaert, A. (2005) The Salmonella genomic island 1 is an integrative mobilizable element, *Mol Microbiol*, **55**, 1911-1924.

Doublet, B., Lailler, R., Meunier, D., Brisabois, A., Boyd, D., Mulvey, M.R., Chaslus-Dancla, E. and Cloeckaert, A. (2003) Variant Salmonella genomic island 1 antibiotic resistance gene cluster in Salmonella enterica serovar Albany, *Emerg Infect Dis*, **9**, 585-591.

Down, T., Leong, B. and Hubbard, T.J. (2006) A machine learning strategy to identify candidate binding sites in human protein-coding sequence, *BMC Bioinformatics*, **7**, 419.

Down, T.A. and Hubbard, T.J. (2002) Computational detection and location of transcription start sites in mammalian genomic DNA, *Genome Res*, **12**, 458-461.

Down, T.A. and Hubbard, T.J. (2003) Relevance vector machines for classifying points and regions in biological sequences, *Quantitative Biology Archive*, [http://arxiv.org/abs/q-bio.GN/0312006].

Dubnau, D. (1999) DNA uptake in bacteria, *Annu Rev Microbiol*, **53**, 217-244.

Dunn, B.E., Cohen, H. and Blaser, M.J. (1997) Helicobacter pylori, *Clin Microbiol Rev*, **10**, 720-741.

Durbin, R., Eddy, S.R., Krogh, A. and Mitchison, G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.

Dykhuizen, D.E. and Green, L. (1991) Recombination in Escherichia coli and the definition of biological species, *J Bacteriol*, **173**, 7257-7268.

Ebner, P., Garner, K. and Mathew, A. (2004) Class 1 integrons in various Salmonella enterica serovars isolated from animals and identification of genomic island SGI1 in Salmonella enterica var. Meleagridis, *J Antimicrob Chemother*, **53**, 1004-1009.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucl. Acids Res.*, **32**, 1792-1797.

Eisen, J.A., Heidelberg, J.F., White, O. and Salzberg, S.L. (2000) Evidence for symmetric chromosomal inversions around the replication origin in bacteria, *Genome Biol*, **1**, RESEARCH0011.

Elliott, S.J., Wainwright, L.A., McDaniel, T.K., Jarvis, K.G., Deng, Y.K., Lai, L.C., McNamara, B.P., Donnenberg, M.S. and Kaper, J.B. (1998) The complete sequence of the locus of enterocyte effacement (LEE) from enteropathogenic Escherichia coli E2348/69, *Mol Microbiol*, **28**, 1-4.

Feil, E.J., Holmes, E.C., Bessen, D.E., Chan, M.S., Day, N.P., Enright, M.C., Goldstein, R., Hood, D.W., Kalia, A., Moore, C.E., Zhou, J. and Spratt, B.G. (2001) Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences, *Proc Natl Acad Sci U S A*, **98**, 182-187.

Felsenstein, J. (1978) The Number of Evolutionary Trees, *Systematic Zoology*, **27**, 27-33.

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach, *J Mol Evol*, **17**, 368-376.

Felsenstein, J. (1989) PHYLIP - Phylogeny Inference Package (version 3.2), *Cladistics*, **5**, 164-166.

Felsenstein, J. and Churchill, G.A. (1996) A Hidden Markov Model approach to variation among sites in rate of evolution, *Mol Biol Evol*, **13**, 93-104.

Fetherston, J.D., Schuetze, P. and Perry, R.D. (1992) Loss of the pigmentation phenotype in Yersinia pestis is due to the spontaneous deletion of 102 kb of chromosomal DNA which is flanked by a repetitive element, *Mol Microbiol*, **6**, 2693-2704.

Fischetti, V.A. (2007) In vivo acquisition of prophage in Streptococcus pyogenes, *Trends Microbiol*, **15**, 297-300.

Fitch, W.M. (1971) Toward defining the course of evolution: Minimum change for a specified tree topology, *Systematic Zoology*, **20**, 406-416.

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

Fitzgerald, J.R., Sturdevant, D.E., Mackie, S.M., Gill, S.R. and Musser, J.M. (2001) Evolutionary genomics of Staphylococcus aureus: insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic, *Proc Natl Acad Sci U S A*, **98**, 8821-8826.

Forsdyke, D.R. (1996) Different biological species "broadcast" their DNAs at different (G+C)% "wavelengths", *J Theor Biol*, **178**, 405-417.

Freire-Picos, M.A., Gonzalez-Siso, M.I., Rodriguez-Belmonte, E., Rodriguez-Torres, A.M., Ramil, E. and Cerdan, M.E. (1994) Codon usage in Kluyveromyces lactis and in yeast cytochrome c-encoding genes, *Gene*, **139**, 43-49.

Friebel, A., Ilchmann, H., Aepfelbacher, M., Ehrbar, K., Machleidt, W. and Hardt, W.D. (2001) SopE and SopE2 from Salmonella typhimurium activate different sets of RhoGTPases of the host cell, *J Biol Chem*, **276**, 34035-34040.

Fujita, J., Yamadori, I., Xu, G., Hojo, S., Negayama, K., Miyawaki, H., Yamaji, Y. and Takahara, J. (1996) Clinical features of Stenotrophomonas maltophilia pneumonia in immunocompromised patients, *Respir Med*, **90**, 35-38.

Fuqua, W.C. and Winans, S.C. (1994) A LuxR-LuxI type regulatory system activates Agrobacterium Ti plasmid conjugal transfer in the presence of a plant tumor metabolite, *J Bacteriol*, **176**, 2796-2806.

Fuxelius, H.H., Darby, A.C., Cho, N.H. and Andersson, S.G. (2008) Visualization of pseudogenes in intracellular bacteria reveals the different tracks to gene destruction, *Genome Biol*, **9**, R42.

Gal-Mor, O. and Finlay, B.B. (2006) Pathogenicity islands: a molecular toolbox for bacterial virulence, *Cell Microbiol*, **8**, 1707-1719.

Galan, J.E. (1996) Molecular genetic bases of Salmonella entry into host cells, *Mol Microbiol*, **20**, 263-271.

Garcia-Vallve, S., Guzman, E., Montero, M.A. and Romeu, A. (2003) HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes., *Nucleic Acids Res.*, **31**, 187-189.

Gerlach, R.G., Jackel, D., Geymeier, N. and Hensel, M. (2007) Salmonella pathogenicity island 4-mediated adhesion is coregulated with invasion genes in Salmonella enterica, *Infect Immun*, **75**, 4697-4709.

Gevers, D., Cohan, F.M., Lawrence, J.G., Spratt, B.G., Coenye, T., Feil, E.J., Stackebrandt, E., Van de Peer, Y., Vandamme, P., Thompson, F.L. and Swings, J. (2005) Opinion: Re-evaluating prokaryotic species, *Nat Rev Microbiol*, **3**, 733-739.

Ginocchio, C.C., Rahn, K., Clarke, R.C. and Galan, J.E. (1997) Naturally occurring deletions in the centisome 63 pathogenicity island of environmental isolates of Salmonella spp, *Infect Immun*, **65**, 1267-1272.

Glaser, P., Frangeul, L., Buchrieser, C., Rusniok, C., Amend, A., Baquero, F., Berche, P., Bloecker, H., Brandt, P., Chakraborty, T., Charbit, A., Chetouani, F., Couve, E., de Daruvar, A., Dehoux, P., Domann, E., Dominguez-Bernal, G., Duchaud, E., Durant, L., Dussurget, O., Entian, K.D., Fsihi, H., Garcia-del Portillo, F., Garrido, P., Gautier, L., Goebel, W., Gomez-Lopez, N., Hain, T., Hauf, J., Jackson, D., Jones, L.M., Kaerst, U., Kreft, J., Kuhn, M., Kunst, F., Kurapkat, G., Madueno, E., Maitournam, A., Vicente, J.M., Ng, E., Nedjari, H., Nordsiek, G., Novella, S., de Pablos, B., Perez-Diaz, J.C., Purcell, R., Remmel, B., Rose, M., Schlueter, T., Simoes, N., Tierrez, A., Vazquez-Boland, J.A., Voss, H., Wehland, J. and Cossart, P. (2001) Comparative genomics of Listeria species, *Science*, **294**, 849-852.

Glaser, P., Rusniok, C., Buchrieser, C., Chevalier, F., Frangeul, L., Msadek, T., Zouine, M., Couve, E., Lalioui, L., Poyart, C., Trieu-Cuot, P. and Kunst, F. (2002) Genome sequence of Streptococcus agalactiae, a pathogen causing invasive neonatal disease, *Mol Microbiol*, **45**, 1499-1513.

Gogarten, J.P. and Townsend, J.P. (2005) Horizontal gene transfer, genome innovation and evolution, *Nat Rev Microbiol*, **3**, 679-687.

Goldman, B.S. and Kranz, R.G. (1998) Evolution and horizontal transfer of an entire biosynthetic pathway for cytochrome c biogenesis: Helicobacter, Deinococcus, Archae and more, *Mol Microbiol*, **27**, 871-873.

Gomez-Valero, L., Rocha, E.P., Latorre, A. and Silva, F.J. (2007) Reconstructing the ancestor of Mycobacterium leprae: the dynamics of gene loss and genome reduction, *Genome Res*, **17**, 1178-1185.

Goodman, S.D. and Scocca, J.J. (1988) Identification and arrangement of the DNA sequence recognized in specific transformation of Neisseria gonorrhoeae, *Proc Natl Acad Sci U S A*, **85**, 6982-6986.

Gould, V.C., Okazaki, A. and Avison, M.B. (2006) Beta-lactam resistance and beta-lactamase expression in clinical Stenotrophomonas maltophilia isolates having defined phylogenetic relationships, *J Antimicrob Chemother*, **57**, 199-203.

Gould, V.C., Okazaki, A., Howe, R.A. and Avison, M.B. (2004) Analysis of sequence variation among smeDEF multi drug efflux pump genes and flanking DNA from defined 16S rRNA subgroups of clinical Stenotrophomonas maltophilia isolates, *J Antimicrob Chemother*, **54**, 348-353.

Gouy, M. and Gautier, C. (1982) Codon usage in bacteria: correlation with gene expressivity, *Nucleic Acids Res*, **10**, 7055-7074.

Gribaldo, S., Lumia, V., Creti, R., de Macario, E.C., Sanangelantoni, A. and Cammarano, P. (1999) Discontinuous occurrence of the hsp70 (dnaK) gene among Archaea and sequence features of HSP70 suggest a novel outlook on phylogenies inferred from this protein, *J Bacteriol*, **181**, 434-443.

Griffith, F. (1928) The significance of pneumococcal types, *J Hyg*, **27**, 113-159.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. and Eddy, S.R. (2003) Rfam: an RNA family database, *Nucleic Acids Res*, **31**, 439-441.

Groisman, E.A. (2001) The pleiotropic two-component regulatory system PhoP-PhoQ, *J Bacteriol*, **183**, 1835-1842.

Groisman, E.A. and Ochman, H. (1996) Pathogenicity islands: bacterial evolution in quantum leaps, *Cell*, **87**, 791-794.

Guan, S., Bastin, D.A. and Verma, N.K. (1999) Functional analysis of the O antigen glucosylation gene cluster of Shigella flexneri bacteriophage SfX, *Microbiology*, **145 (Pt 5)**, 1263-1273.

Hacker, J. (1990) Genetic determinants coding for fimbriae and adhesins of extraintestinal Escherichia coli, *Curr Top Microbiol Immunol*, **151**, 1-27.

Hacker, J., Bender, L., Ott, M., Wingender, J., Lund, B., Marre, R. and Goebel, W. (1990) Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal Escherichia coli isolates, *Microb Pathog*, **8**, 213-225.

Hacker, J., Blum-Oehler, G., Muhldorfer, I. and Tschape, H. (1997) Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution, *Mol Microbiol*, **23**, 1089-1097.

Hacker, J. and Carniel, E. (2001) Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes, *EMBO Rep*, **2**, 376-381.

Hacker, J. and Kaper, J.B. (2002) *Pathogenicity Islands and the Evolution of Pathogenic Microbes*. Springer-Verlag, Berlin.

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

Hardt, W.D., Urlaub, H. and Galan, J.E. (1998) A substrate of the centisome 63 type III protein secretion system of Salmonella typhimurium is encoded by a cryptic bacteriophage, *Proc Natl Acad Sci U S A*, **95**, 2574-2579.

Hasegawa, M., Kishino, H. and Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA, *J Mol Evol*, **22**, 160-174.

Hashimoto, Y., Li, N., Yokoyama, H. and Ezaki, T. (1993) Complete nucleotide sequence and molecular characterization of ViaB region encoding Vi antigen in Salmonella typhi, *J Bacteriol*, **175**, 4456-4465.

Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, **57**, 97-109.

He, J., Baldini, R.L., Deziel, E., Saucier, M., Zhang, Q., Liberati, N.T., Lee, D., Urbach, J., Goodman, H.M. and Rahme, L.G. (2004) The broad host range pathogen Pseudomonas aeruginosa strain PA14 carries two pathogenicity islands harboring plant and animal virulence genes, *Proc Natl Acad Sci U S A*, **101**, 2530-2535.

Henderson, I.R., Navarro-Garcia, F. and Nataro, J.P. (1998) The great escape: structure and function of the autotransporter proteins, *Trends Microbiol*, **6**, 370-378.

Hensel, M. (2004) Evolution of pathogenicity islands of Salmonella enterica, *Int J Med Microbiol*, **294**, 95-102.

Hensel, M., Hinsley, A.P., Nikolaus, T., Sawers, G. and Berks, B.C. (1999) The genetic basis of tetrathionate respiration in Salmonella typhimurium, *Mol Microbiol*, **32**, 275-287.

Hensel, M., Nikolaus, T. and Egelseer, C. (1999) Molecular and functional analysis indicates a mosaic structure of Salmonella pathogenicity island 2, *Mol Microbiol*, **31**, 489-498.

Hensel, M., Shea, J.E., Baumler, A.J., Gleeson, C., Blattner, F. and Holden, D.W. (1997) Analysis of the boundaries of Salmonella pathogenicity island 2 and the corresponding chromosomal region of Escherichia coli K-12, *J Bacteriol*, **179**, 1105-1111.

Hensel, M., Shea, J.E., Raupach, B., Monack, D., Falkow, S., Gleeson, C., Kubo, T. and Holden, D.W. (1997) Functional analysis of ssaJ and the ssaK/U operon, 13 genes encoding components of the type III secretion apparatus of Salmonella Pathogenicity Island 2, *Mol Microbiol*, **24**, 155-167.

Hensel, M., Shea, J.E., Waterman, S.R., Mundy, R., Nikolaus, T., Banks, G., Vazquez-Torres, A., Gleeson, C., Fang, F.C. and Holden, D.W. (1998) Genes encoding putative effector proteins of the type III secretion system of Salmonella pathogenicity island 2 are required for bacterial virulence and proliferation in macrophages, *Mol Microbiol*, **30**, 163-174.

Hentschel, U. and Hacker, J. (2001) Pathogenicity islands: the tip of the iceberg., *Microbes Infect*, **3**, 545-548.

Herron-Olson, L., Fitzgerald, J.R., Musser, J.M. and Kapur, V. (2007) Molecular Correlates of Host Specialization in Staphylococcus aureus, *PLoS ONE*, **2**, e1120.

High, N.J., Hales, B.A., Jann, K. and Boulnois, G.J. (1988) A block of urovirulence genes encoding multiple fimbriae and hemolysin in Escherichia coli O4:K12:H, *Infect Immun*, **56**, 513-517.

Hilario, E. and Gogarten, J.P. (1993) Horizontal transfer of ATPase genes--the tree of life becomes a net of life, *Biosystems*, **31**, 111-119.

Hochhut, B., Beaber, J.W., Woodgate, R. and Waldor, M.K. (2001) Formation of chromosomal tandem arrays of the SXT element and R391, two conjugative chromosomally integrating elements that share an attachment site, *J Bacteriol*, **183**, 1124-1132.

Hochhut, B. and Waldor, M.K. (1999) Site-specific integration of the conjugal Vibrio cholerae SXT element into prfC, *Mol Microbiol*, **32**, 99-110.

Hofreuter, D., Odenbreit, S. and Haas, R. (2001) Natural transformation competence in Helicobacter pylori is mediated by the basic components of a type IV secretion system, *Mol Microbiol*, **41**, 379-391.

Holden, M.T., Feil, E.J., Lindsay, J.A., Peacock, S.J., Day, N.P., Enright, M.C., Foster, T.J., Moore, C.E., Hurst, L., Atkin, R., Barron, A., Bason, N., Bentley, S.D., Chillingworth, C., Chillingworth, T., Churcher, C., Clark, L., Corton, C., Cronin, A., Doggett, J., Dowd, L., Feltwell, T., Hance, Z., Harris, B., Hauser, H., Holroyd, S., Jagels, K., James, K.D., Lennard, N., Line, A., Mayes, R., Moule, S., Mungall, K., Ormond, D., Quail, M.A., Rabbinowitsch, E., Rutherford, K., Sanders, M., Sharp, S., Simmonds, M., Stevens, K., Whitehead, S., Barrell, B.G., Spratt, B.G. and Parkhill, J. (2004) Complete genomes of two clinical Staphylococcus aureus strains: evidence for the rapid evolution of virulence and drug resistance, *Proc Natl Acad Sci U S A*, **101**, 9786-9791.

Holder, M. and Lewis, P.O. (2003) Phylogeny estimation: traditional and Bayesian approaches, *Nat Rev Genet*, **4**, 275-284.

Hornick, R.B., Greisman, S.E., Woodward, T.E., DuPont, H.L., Dawkins, A.T. and Snyder, M.J. (1970) Typhoid fever: pathogenesis and immunologic control, *N Engl J Med*, **283**, 686-691.

Hoskins, J., Alborn, W.E., Jr., Arnold, J., Blaszczak, L.C., Burgett, S., DeHoff, B.S., Estrem, S.T., Fritz, L., Fu, D.J., Fuller, W., Geringer, C., Gilmour, R., Glass, J.S., Khoja, H., Kraft, A.R., Lagace, R.E., LeBlanc, D.J., Lee, L.N., Lefkowitz, E.J., Lu, J., Matsushima, P., McAhren, S.M., McHenney, M., McLeaster, K., Mundy, C.W., Nicas, T.I., Norris, F.H., O'Gara, M., Peery, R.B., Robertson, G.T., Rockey, P., Sun, P.M., Winkler, M.E., Yang, Y., Young-Bellido, M., Zhao, G., Zook, C.A., Baltz, R.H., Jaskunas, S.R., Rosteck, P.R., Jr., Skatrud, P.L. and Glass, J.I. (2001) Genome of the bacterium Streptococcus pneumoniae strain R6, *J Bacteriol*, **183**, 5709-5717.

Hotopp, J.C., Clark, M.E., Oliveira, D.C., Foster, J.M., Fischer, P., Torres, M.C., Giebel, J.D., Kumar, N., Ishmael, N., Wang, S., Ingram, J., Nene, R.V., Shepard, J., Tomkins, J., Richards, S., Spiro, D.J., Ghedin, E., Slatko, B.E., Tettelin, H. and Werren, J.H. (2007) Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes, *Science*, **317**, 1753-1756.

Hou, Y.M. (1999) Transfer RNAs and pathogenicity islands, *Trends Biochem Sci*, **24**, 295-298.

Hsiao, W., Wan, I., Jones, S.J. and Brinkman, F.S. (2003) IslandPath: aiding detection of genomic islands in prokaryotes, *Bioinformatics*, **19**, 418-420.

Huelsenbeck, J.P. (1995) The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining, *Mol Biol Evol*, **12**, 843-849.

Huelsenbeck, J.P., Ronquist, F., Nielsen, R. and Bollback, J.P. (2001) Bayesian inference of phylogeny and its impact on evolutionary biology, *Science*, **294**, 2310-2314.

Huson, D.H. and Bryant, D. (2006) Application of phylogenetic networks in evolutionary studies, *Mol Biol Evol*, **23**, 254-267.

Ikemura, T. (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes, *J Mol Biol*, **146**, 1-21.

Jain, R., Rivera, M.C. and Lake, J.A. (1999) Horizontal gene transfer among genomes: the complexity hypothesis., *Proc Natl Acad Sci U S A*, **96**, 3801-3806.

Jerse, A.E., Yu, J., Tall, B.D. and Kaper, J.B. (1990) A genetic locus of enteropathogenic Escherichia coli necessary for the production of attaching and effacing lesions on tissue culture cells, *Proc Natl Acad Sci U S A*, **87**, 7839-7843.

Jin, Q., Yuan, Z., Xu, J., Wang, Y., Shen, Y., Lu, W., Wang, J., Liu, H., Yang, J., Yang, F., Zhang, X., Zhang, J., Yang, G., Wu, H., Qu, D., Dong, J., Sun, L., Xue, Y., Zhao, A., Gao, Y., Zhu, J., Kan, B., Ding, K., Chen, S., Cheng, H., Yao, Z., He, B., Chen, R., Ma, D., Qiang, B., Wen, Y., Hou, Y. and Yu, J. (2002) Genome sequence of Shigella flexneri 2a: insights into pathogenicity through comparison with genomes of Escherichia coli K12 and O157, *Nucleic Acids Res*, **30**, 4432-4441.

Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992) The rapid generation of mutation data matrices from protein sequences, *Comput Appl Biosci*, **8**, 275-282.

Jukes, T.H. and Cantor, C.R. (1969) Evolution of protein molecules. In, *Mammalian protein metabolism*. Academic Press, New York, 21-132.

Kaper, J.B. and Hacker, J. (1999) *Pathogenicity Islands and Other Mobile Virulence Elements.* American Society for Microbiology Press, Washington DC.

Karlin, S. (1998) Global dinucleotide signatures and analysis of genomic heterogeneity, *Curr Opin Microbiol*, **1**, 598-610.

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

Karlin, S. and Mrazek, J. (2000) Predicted highly expressed genes of diverse prokaryotic genomes, *J Bacteriol*, **182**, 5238-5250.

Karlin, S., Mrazek, J. and Campbell, A.M. (1998) Codon usages in different gene classes of the Escherichia coli genome, *Mol Microbiol*, **29**, 1341-1355.

Kenny, B., DeVinney, R., Stein, M., Reinscheid, D.J., Frey, E.A. and Finlay, B.B. (1997) Enteropathogenic E. coli (EPEC) transfers its receptor for intimate adherence into mammalian cells, *Cell*, **91**, 511-520.

Kent, W.J. (2002) BLAT--the BLAST-like alignment tool, *Genome Res*, **12**, 656-664.

Khan, S.A. (1997) Rolling-circle replication of bacterial plasmids, *Microbiol Mol Biol Rev*, **61**, 442-455.

Kidgell, C., Reichard, U., Wain, J., Linz, B., Torpdahl, M., Dougan, G. and Achtman, M. (2002) Salmonella typhi, the causative agent of typhoid fever, is approximately 50,000 years old, *Infect Genet Evol*, **2**, 39-45.

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences, *J Mol Evol*, **16**, 111-120.

Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Kishino, H. and Hasegawa, M. (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea, *J Mol Evol*, **29**, 170-179.

Kleerebezem, M., Boekhorst, J., van Kranenburg, R., Molenaar, D., Kuipers, O.P., Leer, R., Tarchini, R., Peters, S.A., Sandbrink, H.M., Fiers, M.W., Stiekema, W., Lankhorst, R.M., Bron, P.A., Hoffer, S.M., Groot, M.N., Kerkhoven, R., de Vries, M., Ursing, B., de Vos, W.M. and Siezen, R.J. (2003) Complete genome sequence of Lactobacillus plantarum WCFS1, *Proc Natl Acad Sci U S A*, **100**, 1990-1995.

Klockgether, J., Reva, O., Larbig, K. and Tummler, B. (2004) Sequence analysis of the mobile genome island pKLC102 of Pseudomonas aeruginosa C, *J Bacteriol*, **186**, 518-534.

Knapp, S., Hacker, J., Jarchau, T. and Goebel, W. (1986) Large, unstable inserts in the chromosome affect virulence properties of uropathogenic Escherichia coli O6 strain 536, *J Bacteriol*, **168**, 22-30.

Kuhle, V., Abrahams, G.L. and Hensel, M. (2006) Intracellular Salmonella enterica redirect exocytic transport processes in a Salmonella pathogenicity island 2-dependent manner, *Traffic*, **7**, 716-730.

Kunin, V., Goldovsky, L., Darzentas, N. and Ouzounis, C.A. (2005) The net of life: reconstructing the microbial phylogenetic network, *Genome Res*, **15**, 954-959.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S.C., Bron, S., Brouillet, S., Bruschi, C.V., Caldwell, B., Capuano, V., Carter, N.M., Choi, S.K., Codani, J.J., Connerton, I.F., Danchin, A. and et al. (1997) The complete genome sequence of the gram-positive bacterium Bacillus subtilis, *Nature*, **390**, 249-256.

Kurland, C.G. (2000) Something for everyone. Horizontal gene transfer in evolution, *EMBO Rep*, **1**, 92-95.

Kurland, C.G., Canback, B. and Berg, O.G. (2003) Horizontal gene transfer: a critical view, *Proc Natl Acad Sci U S A*, **100**, 9658-9662.

Kurtz, S. and Schleiermacher, C. (1999) REPuter: fast computation of maximal repeats in complete genomes, *Bioinformatics*, **15**, 426-427.

Kutsukake, K., Ohya, Y. and Iino, T. (1990) Transcriptional analysis of the flagellar regulon of Salmonella typhimurium, *J Bacteriol*, **172**, 741-747.

Lawrence, J.G. and Ochman, H. (1997) Amelioration of bacterial genomes: rates of change and exchange, *J Mol Evol*, **44**, 383-397.

Lawrence, J.G. (1999) Gene transfer, speciation, and the evolution of bacterial genomes, *Curr Opin Microbiol*, **2**, 519-523.

Lawrence, J.G. (2001) Catalyzing bacterial speciation: correlating lateral transfer with genetic headroom, *Syst Biol*, **50**, 479-496.

Lawrence, J.G. (2002) Gene transfer in bacteria: speciation without species?, *Theor Popul Biol*, **61**, 449-460.

Lawrence, J.G. and Hendrickson, H. (2003) Lateral gene transfer: when will adolescence end?, *Mol Microbiol*, **50**, 739-749.

Lawrence, J.G., Hendrix, R.W. and Casjens, S. (2001) Where are the pseudogenes in bacterial genomes?, *Trends Microbiol*, **9**, 535-540.

Lawrence, J.G. and Ochman, H. (1998) Molecular archaeology of the Escherichia coli genome, *Proc Natl Acad Sci U S A*, **95**, 9413-9417.

Lawrence, J.G. and Roth, J.R. (1995) The cobalamin (coenzyme B12) biosynthetic genes of Escherichia coli, *J Bacteriol*, **177**, 6371-6380.

Lawrence, J.G. and Roth, J.R. (1996) Evolution of coenzyme B12 synthesis among enteric bacteria: evidence for loss and reacquisition of a multigene complex, *Genetics*, **142**, 11-24.

Lederberg, J. and Tatum, E.L. (1946) Gene recombination in E. coli, *Nature*, **158**, 558.

Lerat, E., Daubin, V., Ochman, H. and Moran, N.A. (2005) Evolutionary origins of genomic repertoires in bacteria, *PLoS Biol*, **3**, e130.

Lessl, M. and Lanka, E. (1994) Common mechanisms in bacterial conjugation and Ti-mediated T-DNA transfer to plant cells, *Cell*, **77**, 321-324.

Levings, R.S., Lightfoot, D., Partridge, S.R., Hall, R.M. and Djordjevic, S.P. (2005) The genomic island SGI1, containing the multiple antibiotic resistance region of Salmonella enterica serovar Typhimurium DT104 or variants of it, is widely distributed in other S. enterica serovars, *J Bacteriol*, **187**, 4401-4409.

Levings, R.S., Partridge, S.R., Djordjevic, S.P. and Hall, R.M. (2007) SGI1-K, a variant of the SGI1 genomic island carrying a mercury resistance region, in Salmonella enterica serovar Kentucky, *Antimicrob Agents Chemother*, **51**, 317-323.

Lewenza, S., Conway, B., Greenberg, E.P. and Sokol, P.A. (1999) Quorum sensing in Burkholderia cepacia: identification of the LuxRI homologs CepRI, *J Bacteriol*, **181**, 748-756.

Libanore, M., Bicocchi, R., Pantaleoni, M. and Ghinelli, F. (2004) Community-acquired infection due to Stenotrophomonas maltophilia: a rare cause of meningitis, *Int J Infect Dis*, **8**, 317-319.

Lindsay, J.A., Ruzin, A., Ross, H.F., Kurepina, N. and Novick, R.P. (1998) The gene for toxic shock toxin is carried by a family of mobile pathogenicity islands in Staphylococcus aureus, *Mol Microbiol*, **29**, 527-543.

Linehan, S.A., Rytkonen, A., Yu, X.J., Liu, M. and Holden, D.W. (2005) SlyA regulates function of Salmonella pathogenicity island 2 (SPI-2) and expression of SPI-2-associated genes., *Infect Immun*, **73**, 4354-4362.

Lio, P. and Vannucci, M. (2000) Finding pathogenicity islands and gene transfer events in genome data, *Bioinformatics*, **16**, 932-940.

Liu, S.L. and Sanderson, K.E. (1995) Rearrangements in the genome of the bacterium Salmonella typhi, *Proc Natl Acad Sci U S A*, **92**, 1018-1022.

Lober, S., Jackel, D., Kaiser, N. and Hensel, M. (2006) Regulation of Salmonella pathogenicity island 2 genes by independent environmental signals, *Int J Med Microbiol*, **296**, 435-447.

Lorenz, M.G. and Wackernagel, W. (1994) Bacterial gene transfer by natural genetic transformation in the environment, *Microbiol Rev*, **58**, 563-602.

Low, D., David, V., Lark, D., Schoolnik, G. and Falkow, S. (1984) Gene clusters governing the production of hemolysin and mannose-resistant hemagglutination are closely linked in Escherichia coli serotype O4 and O6 isolates from urinary tract infections, *Infect Immun*, **43**, 353-358.

MacQueen, J.B. (1967) Some Methods for classification and Analysis of Multivariate Observations. In, *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, 281-297.

Maeda, S., Akanuma, M., Mitsuno, Y., Hirata, Y., Ogura, K., Yoshida, H., Shiratori, Y. and Omata, M. (2001) Distinct mechanism of Helicobacter pylori-mediated NF-kappa B activation between gastric cancer cells and monocytic cells, *J Biol Chem*, **276**, 44856-44864.

Mahenthiralingam, E., Simpson, D.A. and Speert, D.P. (1997) Identification and characterization of a novel DNA marker associated with epidemic Burkholderia cepacia strains recovered from patients with cystic fibrosis, *J Clin Microbiol*, **35**, 808-816.

Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M. and Spratt, B.G. (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms, *Proc Natl Acad Sci U S A*, **95**, 3140-3145.

Maione, D., Margarit, I., Rinaudo, C.D., Masignani, V., Mora, M., Scarselli, M., Tettelin, H., Brettoni, C., Iacobini, E.T., Rosini, R., D'Agostino, N., Miorin, L., Buccato, S., Mariani, M., Galli, G., Nogarotto, R., Nardi Dei, V., Vegni, F., Fraser, C., Mancuso, G., Teti, G., Madoff, L.C., Paoletti, L.C., Rappuoli, R., Kasper, D.L., Telford, J.L. and Grandi, G. (2005) Identification of a universal Group B streptococcus vaccine by multiple genome screen, *Science*, **309**, 148-150.

Maiques, E., Ubeda, C., Tormo, M.A., Ferrer, M.D., Lasa, I., Novick, R.P. and Penades, J.R. (2007) Role of staphylococcal phage and SaPI integrase in intra- and interspecies SaPI transfer, *J Bacteriol*, **189**, 5608-5616.

Mantri, Y. and Williams, K.P. (2004) Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities, *Nucleic Acids Res*, **32**, D55-58.

Martin, W. (1999) Mosaic bacterial chromosomes: a challenge en route to a tree of genomes, *Bioessays*, **21**, 99-104.

Mavris, M., Manning, P.A. and Morona, R. (1997) Mechanism of bacteriophage SfII-mediated serotype conversion in Shigella flexneri, *Mol Microbiol*, **26**, 939-950.

Mayr, E. (1942) *Systematics and the Origin of Species*. Columbia University Press, New York.

McClelland, M., Sanderson, K.E., Clifton, S.W., Latreille, P., Porwollik, S., Sabo, A., Meyer, R., Bieri, T., Ozersky, P., McLellan, M., Harkins, C.R., Wang, C., Nguyen, C., Berghoff, A., Elliott, G., Kohlberg, S., Strong, C., Du, F., Carter, J., Kremizki, C., Layman, D., Leonard, S., Sun, H., Fulton, L., Nash, W., Miner, T., Minx, P., Delehaunty, K., Fronick, C., Magrini, V., Nhan, M., Warren, W., Florea, L., Spieth, J. and Wilson, R.K. (2004) Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of Salmonella enterica that cause typhoid, *Nat Genet*, **36**, 1268-1274.

McClelland, M., Sanderson, K.E., Spieth, J., Clifton, S.W., Latreille, P., Courtney, L., Porwollik, S., Ali, J., Dante, M., Du, F., Hou, S., Layman, D., Leonard, S., Nguyen, C., Scott, K., Holmes, A., Grewal, N., Mulvaney, E., Ryan, E., Sun, H., Florea, L., Miller, W., Stoneking, T., Nhan, M., Waterston, R. and Wilson, R.K. (2001) Complete genome sequence of Salmonella enterica serovar Typhimurium LT2, *Nature*, **413**, 852-856.

McCullagh, P. and Nelder, J.A. (1989) *Generalized linear models*. Chapman and Hall, London.

McDaniel, T.K., Jarvis, K.G., Donnenberg, M.S. and Kaper, J.B. (1995) A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens, *Proc Natl Acad Sci U S A*, **92**, 1664-1668.

McDaniel, T.K. and Kaper, J.B. (1997) A cloned pathogenicity island from enteropathogenic Escherichia coli confers the attaching and effacing phenotype on E. coli K-12, *Mol Microbiol*, **23**, 399-407.

McGillivary, G., Tomaras, A.P., Rhodes, E.R. and Actis, L.A. (2005) Cloning and sequencing of a genomic island found in the Brazilian purpuric fever clone of Haemophilus influenzae biogroup aegyptius, *Infect Immun*, **73**, 1927-1938.

Medini, D., Donati, C., Tettelin, H., Masignani, V. and Rappuoli, R. (2005) The microbial pan-genome, *Curr Opin Genet Dev*, **15**, 589-594.

Medini, D., Serruto, D., Parkhil, J., Relman, D.A., Donati, C., Moxon, R., Falkow, S. and Rappuoli, R. (in press) Microbiology in the post-genomic era, *Nat Rev Microbiol*.

Mehta, N.J., Khan, I.A., Mehta, R.N. and Gulati, A. (2000) Stenotrophomonas maltophilia endocarditis of prosthetic aortic valve: report of a case and review of literature, *Heart Lung*, **29**, 351-355.

Merkl, R. (2004) SIGI: score-based identification of genomic islands, *BMC Bioinformatics*, **5**, 22.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953) Equation of State Calculations by Fast Computing Machines, *The Journal of Chemical Physics*, **21**, 1087-1092.

Meunier, D., Boyd, D., Mulvey, M.R., Baucheron, S., Mammina, C., Nastasi, A., Chaslus-Dancla, E. and Cloeckaert, A. (2002) Salmonella enterica serotype Typhimurium DT 104 antibiotic resistance genomic island I in serotype paratyphi B, *Emerg Infect Dis*, **8**, 430-433.

Michener, C.D. and Sokal, R.R. (1957) A Quantitative Approach to a Problem in Classification, *Evolution*, **11**, 130-162.

Mietzner, T.A. and Morse, S.A. (1994) The role of iron-binding proteins in the survival of pathogenic bacteria, *Annu Rev Nutr*, **14**, 471-493.

Mirkin, B.G., Fenner, T.I., Galperin, M.Y. and Koonin, E.V. (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes, *BMC Evol Biol*, **3**, 2.

Mirold, S., Rabsch, W., Rohde, M., Stender, S., Tschape, H., Russmann, H., Igwe, E. and Hardt, W.D. (1999) Isolation of a temperate bacteriophage encoding the type III effector protein SopE from an epidemic Salmonella typhimurium strain, *Proc Natl Acad Sci U S A*, **96**, 9845-9850.

Mohd-Zain, Z., Turner, S.L., Cerdeno-Tarraga, A.M., Lilley, A.K., Inzana, T.J., Duncan, A.J., Harding, R.M., Hood, D.W., Peto, T.E. and Crook, D.W. (2004) Transferable antibiotic resistance elements in Haemophilus influenzae share a common evolutionary origin with a diverse family of syntenic genomic islands, *J Bacteriol*, **186**, 8114-8122.

Morgan, E., Bowen, A.J., Carnell, S.C., Wallis, T.S. and Stevens, M.P. (2007) SiiE is secreted by the Salmonella enterica serovar Typhimurium pathogenicity island 4-encoded secretion system and contributes to intestinal colonization in cattle, *Infect Immun*, **75**, 1524-1533.

Morse, M.L., Lederberg, E.M. and Lederberg, J. (1956) Transduction in Escherichia Coli K-12, *Genetics*, **41**, 142-156.

Muder, R.R., Harris, A.P., Muller, S., Edmond, M., Chow, J.W., Papadakis, K., Wagener, M.W., Bodey, G.P. and Steckelberg, J.M. (1996) Bacteremia due to Stenotrophomonas (Xanthomonas) maltophilia: a prospective, multicenter study of 91 episodes, *Clin Infect Dis*, **22**, 508-512.

Mulvey, M.R., Boyd, D., Cloeckaert, A., Ahmed, R. and Ng, L.K. (2004) Emergence of multidrug-resistant Salmonella Paratyphi B dT+, Canada, *Emerg Infect Dis*, **10**, 1307-1310.

Mulvey, M.R., Boyd, D.A., Olson, A.B., Doublet, B. and Cloeckaert, A. (2006) The genetics of Salmonella genomic island 1, *Microbes Infect*, **8**, 1915-1922.

Muto, A. and Osawa, S. (1987) The guanine and cytosine content of genomic DNA and bacterial evolution, *Proc Natl Acad Sci U S A*, **84**, 166-169.

Nair, S., Alokam, S., Kothapalli, S., Porwollik, S., Proctor, E., Choy, C., McClelland, M., Liu, S.L. and Sanderson, K.E. (2004) Salmonella enterica serovar Typhi strains from which SPI7, a 134-kilobase island with genes for Vi exopolysaccharide and other functions, has been deleted, *J Bacteriol*, **186**, 3214-3223.

Navarre, W.W., Porwollik, S., Wang, Y., McClelland, M., Rosen, H., Libby, S.J. and Fang, F.C. (2006) Selective silencing of foreign DNA with low GC content by the H-NS protein in Salmonella, *Science*, **313**, 236-238.

Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Nelson, W.C., Ketchum, K.A., McDonald, L., Utterback, T.R., Malek, J.A., Linher, K.D., Garrett, M.M., Stewart, A.M., Cotton, M.D., Pratt, M.S., Phillips, C.A., Richardson, D., Heidelberg, J., Sutton, G.G., Fleischmann, R.D., Eisen, J.A., White, O., Salzberg, S.L., Smith, H.O., Venter, J.C. and Fraser, C.M. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima, *Nature*, **399**, 323-329.

Nomura, M. (1999) Engineering of bacterial ribosomes: replacement of all seven Escherichia coli rRNA operons by a single plasmid-encoded operon, *Proc Natl Acad Sci U S A*, **96**, 1820-1822.

Novick, R.P. (2003) Mobile genetic elements and bacterial toxinoses: the superantigen-encoding pathogenicity islands of Staphylococcus aureus, *Plasmid*, **49**, 93-105.

Novick, R.P., Schlievert, P. and Ruzin, A. (2001) Pathogenicity and resistance islands of staphylococci, *Microbes Infect*, **3**, 585-594.

Novick, R.P. and Subedi, A. (2007) The SaPIs: Mobile Pathogenicity Islands of Staphylococcus, *Chem Immunol Allergy*, **93**, 42-57.

Ochman, H. and Groisman, E.A. (1996) Distribution of pathogenicity islands in Salmonella spp, *Infect Immun*, **64**, 5410-5412.

Ochman, H., Lawrence, J.G. and Groisman, E.A. (2000) Lateral gene transfer and the nature of bacterial innovation, *Nature*, **405**, 299-304.

Ochman, H. and Wilson, A.C. (1987) Evolution in bacteria: evidence for a universal substitution rate in cellular genomes, *J Mol Evol*, **26**, 74-86.

Ohnishi, M., Kurokawa, K. and Hayashi, T. (2001) Diversification of Escherichia coli genomes: are bacteriophages the major contributors?, *Trends Microbiol*, **9**, 481-485.

Osborn, A.M. and Boltner, D. (2002) When phage, plasmids, and transposons collide: genomic islands, and conjugative- and mobilizable-transposons as a mosaic continuum, *Plasmid*, **48**, 202-212.

Ou, H.Y., He, X., Harrison, E.M., Kulasekara, B.R., Thani, A.B., Kadioglu, A., Lory, S., Hinton, J.C., Barer, M.R., Deng, Z. and Rajakumar, K. (2007) MobilomeFINDER: web-based tools for in silico and experimental discovery of bacterial genomic islands, *Nucleic Acids Res*, **35**, W97-W104.

Parkhill, J., Achtman, M., James, K.D., Bentley, S.D., Churcher, C., Klee, S.R., Morelli, G., Basham, D., Brown, D., Chillingworth, T., Davies, R.M., Davis, P., Devlin, K., Feltwell, T., Hamlin, N., Holroyd, S., Jagels, K., Leather, S., Moule, S., Mungall, K., Quail, M.A., Rajandream, M.A., Rutherford, K.M., Simmonds, M., Skelton, J., Whitehead, S., Spratt, B.G. and Barrell, B.G. (2000) Complete DNA sequence of a serogroup A strain of Neisseria meningitidis Z2491, *Nature*, **404**, 502-506.

Parkhill, J., Dougan, G., James, K.D., Thomson, N.R., Pickard, D., Wain, J., Churcher, C., Mungall, K.L., Bentley, S.D., Holden, M.T., Sebaihia, M., Baker, S., Basham, D., Brooks, K., Chillingworth, T., Connerton, P., Cronin, A., Davis, P., Davies, R.M., Dowd, L., White, N., Farrar, J., Feltwell, T., Hamlin, N., Haque, A., Hien, T.T., Holroyd, S., Jagels, K., Krogh, A., Larsen, T.S., Leather, S., Moule, S., O'Gaora, P., Parry, C., Quail, M., Rutherford, K., Simmonds, M., Skelton, J., Stevens, K., Whitehead, S. and Barrell, B.G. (2001) Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18, *Nature*, **413**, 848-852.

Parkhill, J., Wren, B.W., Thomson, N.R., Titball, R.W., Holden, M.T., Prentice, M.B., Sebaihia, M., James, K.D., Churcher, C., Mungall, K.L., Baker, S., Basham, D., Bentley, S.D., Brooks, K., Cerdeno-Tarraga, A.M., Chillingworth, T., Cronin, A., Davies, R.M., Davis, P., Dougan, G., Feltwell, T., Hamlin, N., Holroyd, S., Jagels, K., Karlyshev, A.V., Leather, S., Moule, S., Oyston, P.C., Quail, M., Rutherford, K., Simmonds, M., Skelton, J., Stevens, K., Whitehead, S. and Barrell, B.G. (2001) Genome sequence of Yersinia pestis, the causative agent of plague, *Nature*, **413**, 523-527.

Parry, C.M., Hien, T.T., Dougan, G., White, N.J. and Farrar, J.J. (2002) Typhoid fever, *N Engl J Med*, **347**, 1770-1782.

Patel, J.C. and Galan, J.E. (2005) Manipulation of the host actin cytoskeleton by Salmonella--all in the name of entry, *Curr Opin Microbiol*, **8**, 10-15.

Paulsen, I.T., Banerjei, L., Myers, G.S., Nelson, K.E., Seshadri, R., Read, T.D., Fouts, D.E., Eisen, J.A., Gill, S.R., Heidelberg, J.F., Tettelin, H., Dodson, R.J., Umayam, L., Brinkac, L., Beanan, M., Daugherty, S., DeBoy, R.T., Durkin, S., Kolonay, J., Madupu, R., Nelson, W., Vamathevan, J., Tran, B., Upton, J., Hansen, T., Shetty, J., Khouri, H., Utterback, T., Radune, D., Ketchum, K.A., Dougherty, B.A. and Fraser, C.M. (2003) Role of mobile DNA in the evolution of vancomycin-resistant Enterococcus faecalis, *Science*, **299**, 2071-2074.

Pearson, W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA, *Methods Enzymol*, **183**, 63-98.

Perna, N.T., Mayhew, G.F., Posfai, G., Elliott, S., Donnenberg, M.S., Kaper, J.B. and Blattner, F.R. (1998) Molecular evolution of a pathogenicity island from enterohemorrhagic Escherichia coli O157:H7, *Infect Immun*, **66**, 3810-3817.

Perna, N.T., Plunkett, G., 3rd, Burland, V., Mau, B., Glasner, J.D., Rose, D.J., Mayhew, G.F., Evans, P.S., Gregor, J., Kirkpatrick, H.A., Posfai, G., Hackett, J., Klink, S., Boutin, A., Shao, Y., Miller, L., Grotbeck, E.J., Davis, N.W., Lim, A., Dimalanta, E.T., Potamousis, K.D., Apodaca, J., Anantharaman, T.S., Lin, J., Yen, G., Schwartz, D.C., Welch, R.A. and Blattner, F.R. (2001) Genome sequence of enterohaemorrhagic Escherichia coli O157:H7, *Nature*, **409**, 529-533.

Perry, R.D. and Fetherston, J.D. (1997) Yersinia pestis--etiologic agent of plague, *Clin Microbiol Rev*, **10**, 35-66.

Pickard, D., Wain, J., Baker, S., Line, A., Chohan, S., Fookes, M., Barron, A., Gaora, P.O., Chabalgoity, J.A., Thanky, N., Scholes, C., Thomson, N., Quail, M., Parkhill, J. and Dougan, G. (2003) Composition, acquisition, and distribution of the Vi exopolysaccharide-encoding Salmonella enterica pathogenicity island SPI-7, *J Bacteriol*, **185**, 5055-5065.

Pizza, M., Scarlato, V., Masignani, V., Giuliani, M.M., Arico, B., Comanducci, M., Jennings, G.T., Baldi, L., Bartolini, E., Capecchi, B., Galeotti, C.L., Luzzi, E., Manetti, R., Marchetti, E., Mora, M., Nuti, S., Ratti, G., Santini, L., Savino, S., Scarselli, M., Storni, E., Zuo, P., Broeker, M., Hundt, E., Knapp, B., Blair, E., Mason, T., Tettelin, H., Hood, D.W., Jeffries, A.C., Saunders, N.J., Granoff, D.M., Venter, J.C., Moxon, E.R., Grandi, G. and Rappuoli, R. (2000) Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing, *Science*, **287**, 1816-1820.

Pridmore, R.D., Berger, B., Desiere, F., Vilanova, D., Barretto, C., Pittet, A.C., Zwahlen, M.C., Rouvet, M., Altermann, E., Barrangou, R., Mollet, B., Mercenier, A., Klaenhammer, T., Arigoni, F. and Schell, M.A. (2004) The genome sequence of the probiotic intestinal bacterium Lactobacillus johnsonii NCC 533, *Proc Natl Acad Sci U S A*, **101**, 2512-2517.

Rabiner, L.R. (1989) A tutorial on hidden Markov models and selected applications inspeech recognition, *Proceedings of the IEEE*, **77**, 257-286.

Rakin, A. and Heesemann, J. (1995) Virulence-associated fyuA/irp2 gene cluster of Yersinia enterocolitica biotype 1B carries a novel insertion sequence IS1328, *FEMS Microbiol Lett*, **129**, 287-292.

Ramsay, J.P., Sullivan, J.T., Stuart, G.S., Lamont, I.L. and Ronson, C.W. (2006) Excision and transfer of the Mesorhizobium loti R7A symbiosis island requires an integrase IntS, a novel recombination directionality factor RdfS, and a putative relaxase RlxS, *Mol Microbiol*, **62**, 723-734.

Ramsden, A.E., Mota, L.J., Munter, S., Shorte, S.L. and Holden, D.W. (2007) The SPI-2 type III secretion system restricts motility of Salmonella-containing vacuoles, *Cell Microbiol*, **9**, 2517-2529.

Raphael, C. (1999) Automatic Segmentation of Acoustic Musical Signals Using Hidden Markov Models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **21**, 360-370.

Rappuoli, R. (2000) Reverse vaccinology, *Curr Opin Microbiol*, **3**, 445-450.

Rappuoli, R. and Covacci, A. (2003) Reverse vaccinology and genomics, *Science*, **302**, 602.

Ravatn, R., Studer, S., Zehnder, A.J. and van der Meer, J.R. (1998) Int-B13, an unusual site-specific recombinase of the bacteriophage P4 integrase family, is responsible for chromosomal insertion of the 105-kilobase clc element of Pseudomonas sp. Strain B13, *J Bacteriol*, **180**, 5505-5514.

Recchia, G.D. and Hall, R.M. (1995) Gene cassettes: a new class of mobile element, *Microbiology*, **141 (Pt 12)**, 3015-3027.

Reiter, W.D., Palm, P. and Yeats, S. (1989) Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements, *Nucleic Acids Res*, **17**, 1907-1914.

Ritter, A., Blum, G., Emody, L., Kerenyi, M., Bock, A., Neuhierl, B., Rabsch, W., Scheutz, F. and Hacker, J. (1995) tRNA genes and pathogenicity islands: influence on virulence and metabolic properties of uropathogenic Escherichia coli, *Mol Microbiol*, **17**, 109-121.

Ritter, A., Gally, D.L., Olsen, P.B., Dobrindt, U., Friedrich, A., Klemm, P. and Hacker, J. (1997) The Pai-associated leuX specific tRNA5(Leu) affects type 1 fimbriation in pathogenic Escherichia coli by control of FimB recombinase expression, *Mol Microbiol*, **25**, 871-882.

Robbins, J.D. and Robbins, J.B. (1984) Reexamination of the protective role of the capsular polysaccharide (Vi antigen) of Salmonella typhi, *J Infect Dis*, **150**, 436-449.

Rosini, R., Rinaudo, C.D., Soriani, M., Lauer, P., Mora, M., Maione, D., Taddei, A., Santi, I., Ghezzo, C., Brettoni, C., Buccato, S., Margarit, I., Grandi, G. and Telford, J.L. (2006) Identification of novel genomic islands coding for antigenic pilus-like structures in Streptococcus agalactiae, *Mol Microbiol*, **61**, 126-141.

Roumagnac, P., Weill, F.X., Dolecek, C., Baker, S., Brisse, S., Chinh, N.T., Le, T.A., Acosta, C.J., Farrar, J., Dougan, G. and Achtman, M. (2006) Evolutionary history of Salmonella typhi, *Science*, **314**, 1301-1304.

Rozen, S. and Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers, *Methods Mol Biol*, **132**, 365-386.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A. and Barrell, B. (2000) Artemis: sequence visualization and annotation, *Bioinformatics*, **16**, 944-945.

Rzhetsky, A. and Nei, M. (1993) Theoretical foundation of the minimum-evolution method of phylogenetic inference, *Mol Biol Evol*, **10**, 1073-1095.

Sacchi, C.T., Whitney, A.M., Reeves, M.W., Mayer, L.W. and Popovic, T. (2002) Sequence diversity of Neisseria meningitidis 16S rRNA genes and use of 16S rRNA gene sequencing as a molecular subtyping tool, *J Clin Microbiol*, **40**, 4520-4527.

Saino, Y., Kobayashi, F., Inoue, M. and Mitsuhashi, S. (1982) Purification and properties of inducible penicillin beta-lactamase isolated from Pseudomonas maltophilia, *Antimicrob Agents Chemother*, **22**, 564-570.

Saitou, N. and Imanishi, T. (1989) Relative Efficiencies of the Fitch-Margoliash, Maximum-Parsimony, Maximum-Likelihood, Minimum-Evolution, and Neighbor-joining Methods of Phylogenetic Tree, *Mol Biol Evol*, **6**, 514-525.

Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Mol Biol Evol*, **4**, 406-425.

Salzberg, S.L., Delcher, A.L., Kasif, S. and White, O. (1998) Microbial gene identification using interpolated Markov models, *Nucleic Acids Res*, **26**, 544-548.

Sanchez, P., Alonso, A. and Martinez, J.L. (2002) Cloning and characterization of SmeT, a repressor of the Stenotrophomonas maltophilia multidrug efflux pump SmeDEF, *Antimicrob Agents Chemother*, **46**, 3386-3393.

Sandberg, R., Winberg, G., Branden, C.I., Kaske, A., Ernberg, I. and Coster, J. (2001) Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier., *Genome Res*, **11**, 1404-1409.

Schmidt, H. and Hensel, M. (2004) Pathogenicity islands in bacterial pathogenesis, *Clin Microbiol Rev*, **17**, 14-56.

Schmidt, H.A., Strimmer, K., Vingron, M. and von Haeseler, A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing, *Bioinformatics*, **18**, 502-504.

Schölkopf, B., Burges, C.J.C. and Smola, A.J. (eds) (1999) *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge, Massachusetts.

Schubert, S., Rakin, A., Karch, H., Carniel, E. and Heesemann, J. (1998) Prevalence of the "high-pathogenicity island" of Yersinia species among Escherichia coli strains that are pathogenic to humans, *Infect Immun*, **66**, 480-485.

Segal, E.D., Cha, J., Lo, J., Falkow, S. and Tompkins, L.S. (1999) Altered states: involvement of phosphorylated CagA in the induction of host cellular growth changes by Helicobacter pylori, *Proc Natl Acad Sci U S A*, **96**, 14559-14564.

Shah, D.H., Lee, M.J., Park, J.H., Lee, J.H., Eo, S.K., Kwon, J.T. and Chae, J.S. (2005) Identification of Salmonella gallinarum virulence genes in a chicken infection model using PCR-based signature-tagged mutagenesis, *Microbiology*, **151**, 3957-3968.

Sharma, S.A., Tummuru, M.K., Blaser, M.J. and Kerr, L.D. (1998) Activation of IL-8 gene expression by Helicobacter pylori is regulated by transcription factor nuclear factor-kappa B in gastric epithelial cells, *J Immunol*, **160**, 2401-2407.

Sharp, P.M. and Li, W.H. (1986) Codon usage in regulatory genes in Escherichia coli does not reflect selection for 'rare' codons, *Nucleic Acids Res*, **14**, 7737-7749.

Sharp, P.M. and Li, W.H. (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications, *Nucleic Acids Res*, **15**, 1281-1295.

Sherburne, C.K., Lawley, T.D., Gilmour, M.W., Blattner, F.R., Burland, V., Grotbeck, E., Rose, D.J. and Taylor, D.E. (2000) The complete DNA sequence and analysis of R27, a large IncHI plasmid from Salmonella typhi that is temperature sensitive for transfer, *Nucleic Acids Res*, **28**, 2177-2186.

Shields, D.C. and Sharp, P.M. (1987) Synonymous codon usage in Bacillus subtilis reflects both translational selection and mutational biases, *Nucleic Acids Res*, **15**, 8023-8040.

Smith, J.M., Smith, N.H., O'Rourke, M. and Spratt, B.G. (1993) How clonal are bacteria?, *Proc Natl Acad Sci U S A*, **90**, 4384-4388.

Snavely, M.D., Miller, C.G. and Maguire, M.E. (1991) The mgtB Mg2+ transport locus of Salmonella typhimurium encodes a P-type ATPase, *J Biol Chem*, **266**, 815-823.

Snel, B., Bork, P. and Huynen, M.A. (1999) Genome phylogeny based on gene content, *Nat Genet*, **21**, 108-110.

Sokol, P.A., Sajjan, U., Visser, M.B., Gingues, S., Forstner, J. and Kooi, C. (2003) The CepIR quorum-sensing system contributes to the virulence of Burkholderia cenocepacia respiratory infections, *Microbiology*, **149**, 3649-3658.

Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A. and Durbin, R. (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains, *Nucleic Acids Res*, **26**, 320-322.

Southern, E.M. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis, *J Mol Biol*, **98**, 503-517.

Stanhope, M.J., Lupas, A., Italia, M.J., Koretke, K.K., Volker, C. and Brown, J.R. (2001) Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates, *Nature*, **411**, 940-944.

Strimmer, K. and von Haeseler, A. (1996) Quartet Puzzling: A Quartet Maximum-Likelihood Method for Reconstructing Tree Topologies, *Mol Biol Evol*, **13**, 964-969.

Suerbaum, S. and Achtman, M. (1999) Evolution of Helicobacter pylori: the role of recombination, *Trends Microbiol*, **7**, 182-184.

Sullivan, J.T., Patrick, H.N., Lowther, W.L., Scott, D.B. and Ronson, C.W. (1995) Nodulating strains of Rhizobium loti arise through chromosomal symbiotic gene transfer in the environment, *Proc Natl Acad Sci U S A*, **92**, 8985-8989.

Sullivan, J.T. and Ronson, C.W. (1998) Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene, *Proc Natl Acad Sci U S A*, **95**, 5145-5149.

Syvanen, M. (1985) Cross-species gene transfer; implications for a new theory of evolution, *J Theor Biol*, **112**, 333-343.

Takeuchi, F., Watanabe, S., Baba, T., Yuzawa, H., Ito, T., Morimoto, Y., Kuroda, M., Cui, L., Takahashi, M., Ankai, A., Baba, S., Fukui, S., Lee, J.C. and Hiramatsu, K. (2005) Whole-genome sequencing of staphylococcus haemolyticus uncovers the extreme plasticity of its genome and the evolution of human-colonizing staphylococcal species, *J Bacteriol*, **187**, 7292-7308.

Tamas, I., Klasson, L., Canback, B., Naslund, A.K., Eriksson, A.S., Wernegreen, J.J., Sandstrom, J.P., Moran, N.A. and Andersson, S.G. (2002) 50 million years of genomic stasis in endosymbiotic bacteria, *Science*, **296**, 2376-2379.

Tateno, Y., Takezaki, N. and Nei, M. (1994) Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site, *Mol Biol Evol*, **11**, 261-277.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D. and Koonin, E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes, *Nucleic Acids Res*, **29**, 22-28.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., Deboy, R.T., Davidsen, T.M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J.D., Hauser, C.R., Sundaram, J.P., Nelson, W.C., Madupu, R., Brinkac, L.M., Dodson, R.J., Rosovitz, M.J., Sullivan, S.A., Daugherty, S.C., Haft, D.H., Selengut, J., Gwinn, M.L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K.J., Smith, S., Utterback, T.R., White, O., Rubens, C.E., Grandi, G., Madoff, L.C., Kasper, D.L., Telford, J.L., Wessels, M.R., Rappuoli, R. and Fraser, C.M. (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome", *Proc Natl Acad Sci U S A*, **102**, 13950-13955.

Tettelin, H., Nelson, K.E., Paulsen, I.T., Eisen, J.A., Read, T.D., Peterson, S., Heidelberg, J., DeBoy, R.T., Haft, D.H., Dodson, R.J., Durkin, A.S., Gwinn, M., Kolonay, J.F., Nelson, W.C., Peterson, J.D., Umayam, L.A., White, O., Salzberg, S.L., Lewis, M.R., Radune, D., Holtzapple, E., Khouri, H., Wolf, A.M., Utterback, T.R., Hansen, C.L., McDonald, L.A., Feldblyum, T.V., Angiuoli, S., Dickinson, T., Hickey, E.K., Holt, I.E., Loftus, B.J., Yang, F., Smith, H.O., Venter, J.C., Dougherty, B.A., Morrison, D.A., Hollingshead, S.K. and Fraser, C.M. (2001) Complete genome sequence of a virulent isolate of Streptococcus pneumoniae, *Science*, **293**, 498-506.

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

Thomson, N., Baker, S., Pickard, D., Fookes, M., Anjum, M., Hamlin, N., Wain, J., House, D., Bhutta, Z., Chan, K., Falkow, S., Parkhill, J., Woodward, M., Ivens, A. and Dougan, G. (2004) The role of prophage-like elements in the diversity of Salmonella enterica serovars, *J Mol Biol*, **339**, 279-300.

Thomson, N.R., Holden, M.T., Carder, C., Lennard, N., Lockey, S.J., Marsh, P., Skipp, P., O'Connor, C.D., Goodhead, I., Norbertzcak, H., Harris, B., Ormond, D., Rance, R., Quail, M.A., Parkhill, J., Stephens, R.S. and Clarke, I.N. (2008) Chlamydia trachomatis: genome sequence analysis of lymphogranuloma venereum isolates, *Genome Res*, **18**, 161-171.

Thomson, N.R., Howard, S., Wren, B.W., Holden, M.T., Crossman, L., Challis, G.L., Churcher, C., Mungall, K., Brooks, K., Chillingworth, T., Feltwell, T., Abdellah, Z., Hauser, H., Jagels, K., Maddison, M., Moule, S., Sanders, M., Whitehead, S., Quail, M.A., Dougan, G., Parkhill, J. and Prentice, M.B. (2006) The Complete Genome Sequence and Comparative Genome Analysis of the High Pathogenicity Yersinia enterocolitica Strain 8081, *PLoS Genet*, **2**, e206.

Threlfall, E.J. (2000) Epidemic salmonella typhimurium DT 104--a truly international multiresistant clone, *J Antimicrob Chemother*, **46**, 7-10.

Tillier, E.R. and Collins, R.A. (2000) Genome rearrangement by replication-directed translocation, *Nat Genet*, **26**, 195-197.

Tipping, M.E. (2001) Sparse Bayesian Learning and the Relevance Vector Machine, *Journal of Machine Learning Research*, **1**, 211-244.

Toussaint, A. and Merlin, C. (2002) Mobile elements as a combination of functional modules, *Plasmid*, **47**, 26-35.

Towers, R.J., Gal, D., McMillan, D., Sriprakash, K.S., Currie, B.J., Walker, M.J., Chhatwal, G.S. and Fagan, P.K. (2004) Fibronectin-binding protein gene recombination and horizontal transfer between group A and G streptococci, *J Clin Microbiol*, **42**, 5357-5361.

Tsirigos, A. and Rigoutsos, I. (2005) A new computational method for the detection of horizontal gene transfer events, *Nucleic Acids Res*, **33**, 922-933.

Tsirigos, A. and Rigoutsos, I. (2005) A sensitive, support-vector-machine method for the detection of horizontal gene transfers in viral, archaeal and bacterial genomes, *Nucleic Acids Res*, **33**, 3699-3707.

Tummuru, M.K., Sharma, S.A. and Blaser, M.J. (1995) Helicobacter pylori picB, a homologue of the Bordetella pertussis toxin secretion protein, is required for induction of IL-8 in gastric epithelial cells, *Mol Microbiol*, **18**, 867-876.

Ubeda, C., Maiques, E., Knecht, E., Lasa, I., Novick, R.P. and Penades, J.R. (2005) Antibiotic-induced SOS response promotes horizontal dissemination of pathogenicity island-encoded virulence factors in staphylococci, *Mol Microbiol*, **56**, 836-844.

Uchiya, K., Barbieri, M.A., Funato, K., Shah, A.H., Stahl, P.D. and Groisman, E.A. (1999) A Salmonella virulence protein that inhibits cellular trafficking, *Embo J*, **18**, 3924-3933.

Vazquez-Torres, A., Xu, Y., Jones-Carson, J., Holden, D.W., Lucia, S.M., Dinauer, M.C., Mastroeni, P. and Fang, F.C. (2000) Salmonella pathogenicity island 2-dependent evasion of the phagocyte NADPH oxidase, *Science*, **287**, 1655-1658.

Vernikos, G.S. (2008) Genome watch: Overtake in reverse gear, *Nat Rev Microbiol*, **advanced online publication**.

Vernikos, G.S. and Parkhill, J. (2006) Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands, *Bioinformatics*, **22**, 2196-2203.

Vernikos, G.S. and Parkhill, J. (2008) Resolving the structural features of genomic islands: A machine learning approach, *Genome Res*, **18**, 331-342.

Vernikos, G.S., Thomson, N.R. and Parkhill, J. (2007) Genetic flux over time in the Salmonella lineage, *Genome Biol*, **8**, R100.

Viterbi, A.J. (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, *IEEE Transactions on Information Theory*, **IT-13**, 260-269.

Wales, A.D., Woodward, M.J. and Pearson, G.R. (2005) Attaching-effacing bacteria in animals, *J Comp Pathol*, **132**, 1-26.

Walker, J.E., Saraste, M., Runswick, M.J. and Gay, N.J. (1982) Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold, *Embo J*, **1**, 945-951.

Ward, J.E., Akiyoshi, D.E., Regier, D., Datta, A., Gordon, M.P. and Nester, E.W. (1988) Characterization of the virB operon from an Agrobacterium tumefaciens Ti plasmid, *J Biol Chem*, **263**, 5804-5814.

Waterhouse, J.C. and Russell, R.R. (2006) Dispensable genes and foreign DNA in Streptococcus mutans, *Microbiology*, **152**, 1777-1788.

Weiss, A.A., Johnson, F.D. and Burns, D.L. (1993) Molecular characterization of an operon required for pertussis toxin secretion, *Proc Natl Acad Sci U S A*, **90**, 2970-2974.

Welch, R.A., Burland, V., Plunkett, G., 3rd, Redford, P., Roesch, P., Rasko, D., Buckles, E.L., Liou, S.R., Boutin, A., Hackett, J., Stroud, D., Mayhew, G.F., Rose, D.J., Zhou, S., Schwartz, D.C., Perna, N.T., Mobley, H.L., Donnenberg, M.S. and Blattner, F.R. (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli, *Proc Natl Acad Sci U S A*, **99**, 17020-17024.

Williams, K.P. (2003) Traffic at the tmRNA gene, *J Bacteriol*, **185**, 1059-1070.

Williams, K.P. (2002) Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies, *Nucleic Acids Res*, **30**, 866-875.

Winans, S.C., Burns, D.L. and Christie, P.J. (1996) Adaptation of a conjugal transfer system for the export of pathogenic macromolecules, *Trends Microbiol*, **4**, 64-68.

Woese, C.R. (1987) Bacterial evolution, *Microbiol Rev*, **51**, 221-271.

Woese, C.R. (2000) Interpreting the universal phylogenetic tree, *Proc Natl Acad Sci U S A*, **97**, 8392-8396.

Wolf, Y.I., Aravind, L., Grishin, N.V. and Koonin, E.V. (1999) Evolution of aminoacyl-tRNA synthetases--analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events, *Genome Res*, **9**, 689-710.

Wong, K.K., McClelland, M., Stillwell, L.C., Sisk, E.C., Thurston, S.J. and Saffer, J.D. (1998) Identification and sequence analysis of a 27-kilobase chromosomal fragment containing a Salmonella pathogenicity island located at 92 minutes on the chromosome map of Salmonella enterica serovar typhimurium LT2, *Infect Immun*, **66**, 3365-3371.

Wood, M.W., Jones, M.A., Watson, P.R., Hedges, S., Wallis, T.S. and Galyov, E.E. (1998) Identification of a pathogenicity island required for Salmonella enteropathogenicity, *Mol Microbiol*, **29**, 883-891.

Wood, M.W., Rosqvist, R., Mullan, P.B., Edwards, M.H. and Galyov, E.E. (1996) SopE, a secreted protein of Salmonella dublin, is translocated into the target eukaryotic cell via a sip-dependent mechanism and promotes bacterial entry, *Mol Microbiol*, **22**, 327-338.

Worley, M.J., Ching, K.H. and Heffron, F. (2000) Salmonella SsrB activates a global regulon of horizontally acquired genes, *Mol Microbiol*, **36**, 749-761.

Wright, F. (1990) The 'effective number of codons' used in a gene, *Gene*, **87**, 23-29.

Wu, T.J., Huang, Y.H. and Li, L.A. (2005) Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences, *Bioinformatics*, **21**, 4125-4132.

Wyborn, N.R., Stapleton, M.R., Norte, V.A., Roberts, R.E., Grafton, J. and Green, J. (2004) Regulation of Escherichia coli hemolysin E expression by H-NS and Salmonella SlyA, *J Bacteriol*, **186**, 1620-1628.

Yang, Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods, *J Mol Evol*, **39**, 306-314.

Yang, Z. (1996) Statistical properties of a DNA sample under the finite-sites model, *Genetics*, **144**, 1941-1950.

Yap, W.H., Zhang, Z. and Wang, Y. (1999) Distinct types of rRNA operons exist in the genome of the actinomycete Thermomonospora chromogena and evidence for horizontal transfer of an entire rRNA operon, *J Bacteriol*, **181**, 5201-5209.

Yoon, S.H., Hur, C.G., Kang, H.Y., Kim, Y.H., Oh, T.K. and Kim, J.F. (2005) A computational approach for identifying pathogenicity islands in prokaryotic genomes, *BMC Bioinformatics*, **6**, 184.

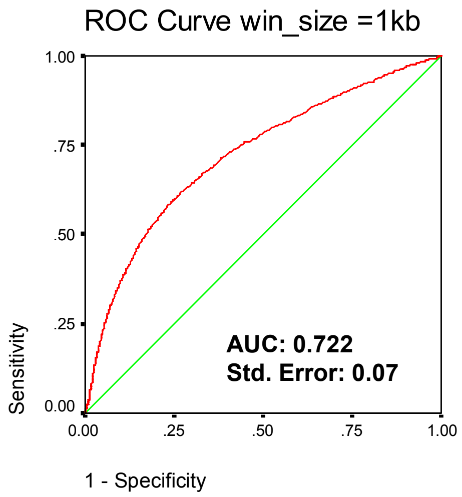
Yoon, S.H., Park, Y.K., Lee, S., Choi, D., Oh, T.K., Hur, C.G. and Kim, J.F. (2007) Towards pathogenomics: a web-based resource for pathogenicity islands, *Nucleic Acids Res*, **35**, D395-400.

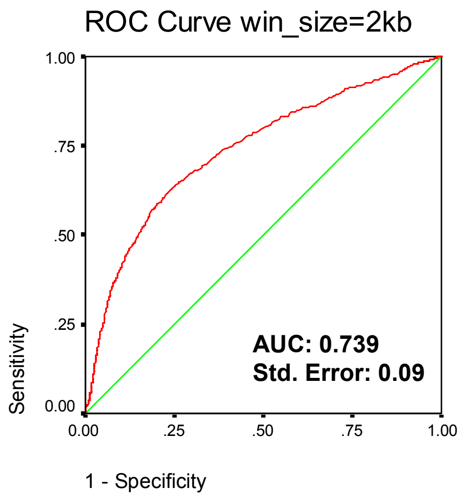
Zhang, L., Radziejewska-Lebrecht, J., Krajewska-Pietrasik, D., Toivanen, P. and Skurnik, M. (1997) Molecular and chemical characterization of the lipopolysaccharide O-antigen and its role in the virulence of Yersinia enterocolitica serotype O:8, *Mol Microbiol*, **23**, 63-76.

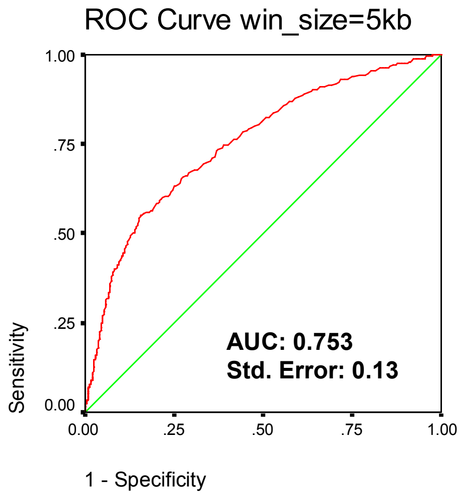
Zhang, S., Kingsley, R.A., Santos, R.L., Andrews-Polymenis, H., Raffatellu, M., Figueiredo, J., Nunes, J., Tsolis, R.M., Adams, L.G. and Baumler, A.J. (2003) Molecular pathogenesis of Salmonella enterica serotype typhimurium-induced diarrhea, *Infect Immun*, **71**, 1-12.

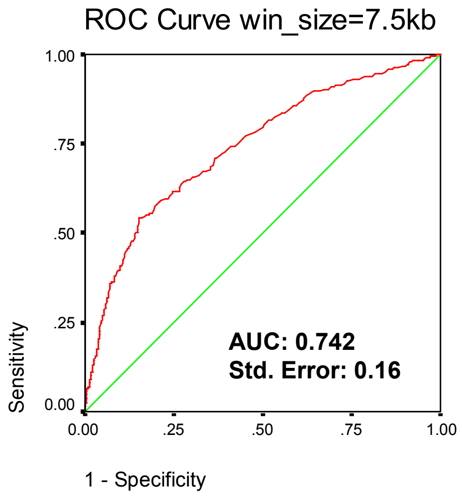
Zhang, X.L., Morris, C. and Hackett, J. (1997) Molecular cloning, nucleotide sequence, and function of a site-specific recombinase encoded in the major 'pathogenicity island' of Salmonella typhi, *Gene*, **202**, 139-146.

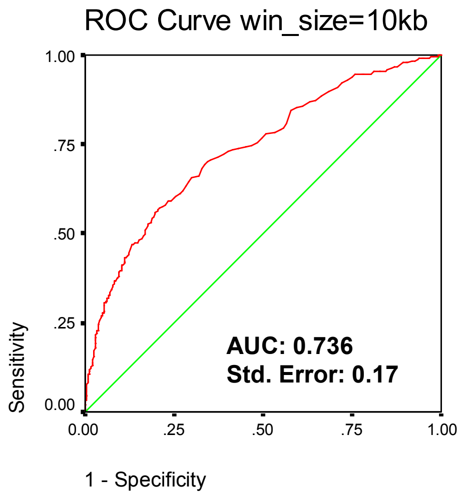
Zhang, X.L., Tsui, I.S., Yip, C.M., Fung, A.W., Wong, D.K., Dai, X., Yang, Y., Hackett, J. and Morris, C. (2000) Salmonella enterica serovar typhi uses type IVB pili to enter human intestinal epithelial cells, *Infect Immun*, **68**, 3067-3073.

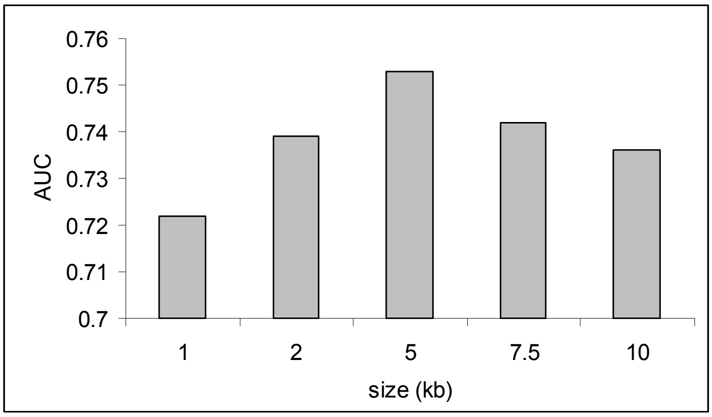

ROC Curve win_size =1kb
1.00
.75
.50
.25
0.00
Sensitivity
0.00
.25
.50
.75
1.00
1 - Specificity
AUC: 0.722
Std. Error: 0.07


ROC Curve win_size=2kb
1.00
.75
.50
.25
0.00
Sensitivity
0.00
.25
.50
.75
1.00
1 - Specificity
AUC: 0.739
Std. Error: 0.09


ROC Curve win_size=5kb
1.00
.75
.50
.25
0.00
Sensitivity
0.00
.25
.50
.75
1.00
1 - Specificity
AUC: 0.753
Std. Error: 0.13


ROC Curve win_size=7.5kb
1.00
.75
.50
.25
0.00
Sensitivity
0.00
.25
.50
.75
1.00
1 - Specificity
AUC: 0.742
Std. Error: 0.16


ROC Curve win_size=10kb
1.00
.75
.50
.25
0.00
Sensitivity
0.00
.25
.50
.75
1.00
1 - Specificity
AUC: 0.736
Std. Error: 0.17


0.76
0.75
0.74
0.73
0.72
0.71
0.7
AUC
1
2
5
7.5
10
size (kb)

Region 1

From	To	Product
60466	61656	putative integrase
62251	62523	putative phage DNA-binding protein
62520	62723	hypothetical protein
62733	63020	putative bacteriophage transcriptional regulator protein
63050	63526	hypothetical protein
63523	63738	conserved hypothetical protein
63735	64385	hypothetical protein
64372	65223	putative primase domain protein
65282	67132	conserved hypothetical protein
67166	67954	hypothetical protein
67990	68991	putative phage portal vertex protein
69087	69434 c	conserved hypothetical protein
69431	69772 c	conserved hypothetical protein
69882	70085 c	hypothetical protein

Region 4

299949	301127 c	putative phage integrase
301127	301351 c	conserved hypothetical protein
301699	302004 c	conserved hypothetical protein
302179	302604 c	conserved hypothetical protein
302681	302959 c	hypothetical protein
302956	305652 c	putative phage-related protein
306023	306484 c	hypothetical protein
306894	307100 c	conserved hypothetical protein
307169	307486 c	conserved hypothetical protein
307510	307842 c	conserved hypothetical protein
307912	308343	putative phage-related protein
308630	309400	conserved hypothetical protein
309586	310977	conserved hypothetical protein
310974	311966 c	putative phage-related protein
311963	312361 c	putative phage tail protein
312373	315237 c	putative phage tail protein
315263	315379 c	putative phage protein
315388	315714 c	putative phage tail protein
315770	316279 c	putative major tail tube protein
316300	317469 c	putative major tail sheath protein (protein fi)
317485	317835 c	putative baseplate assembly protein W (gpw)

317832	318380 c	putative baseplate assembly protein v (gpv)
318467	318748 c	
318758	319990 c	putative phage collar protein
319995	320546 c	putative tail protein I (gpi)
320539	321429 c	putative baseplate assembly protein J (gpj)
321516	321977 c	putative tail completion protein S (gps)
321974	322459 c	putative tail completion protein R (gpr)
322456	322980 c	conserved hypothetical protein
322980	323615 c	putative phage lytic enzyme
323617	323892 c	putative phage protein
323885	324238 c	putative phage protein
324241	324456 c	putative tail protein X (gpx)
324456	324923 c	putative head completion/stabilization protein (gpl)
325028	325735 c	putative terminase, endonuclease subunit (gpm)
325739	326755 c	putative capsid proteins precursor (gpn)
326789	327652 c	presumed capsid scaffolding protein (gpo)
327771	329564	putative phage terminase, ATPase subunit (gpp)
329564	330574	putative presumed portal vertex protein (gpq)
330746	331456	putative site-specific DNA-methyltransferase
332227	332523	conserved hypothetical protein
332777	333241	hypothetical protein
333316	334077 c	conserved hypothetical protein
334344	334940 c	putative DNA recombinase
334915	335202 c	hypothetical protein

Region 7

631290	632120	putative transmembrane ABC transporter
632110	633504	putative ABC transporter component, polysaccharide related
633501	634088	putative chloramphenicol acetyltransferase
634123	635145	putative LPS O-antigen biosynthesis protein
635293	636603	putative glycosyl transferase
636600	638300	putative glycosyltransferase, fusion protein
638464	640812	putative transmembrane protein
640842	641507 c	putative lipase
641819	643768	putative transmembrane protein
643808	644212 c	putative transmembrane GtrA-like cell surface polysaccharide biosynthesis protein
644327	644734 c	putative transmembrane cell surface polysaccharide biosynthesis protein
644731	646038 c	putative FMN amine oxidoreductase
646068	647012 c	putative transmembrane anchor NAD-dependent epimerase/dehydratase/dehydrogenase

647041	647712	c	conserved hypothetical protein
647750	648481	c	putative transmembrane anchor short-chain dehydrogenase
648484	649794	c	putative FAD-binding oxidoreductase
649791	651215	c	putative transmembrane UbiA prenyltransferase family protein
651385	653316	c	putative transmembrane protein
653313	654161	c	putative glycosyl transferase
654172	654531	c	putative transmembrane protein
654528	655358	c	putative xylose isomerase
655355	656137	c	conserved hypothetical protein
656224	657372	c	conserved hypothetical protein
657365	658375	c	putative undecaprenyl-phosphate 4-deoxy-4-formamido-l-arabinose transferase
658467	659435	c	putative transmembrane protein
659540	661255	c	putative transmembrane sulfatase protein

Region 12

1324492	1324737	c	conserved hypothetical protein
1324734	1325993	c	putative conjugal transfer protein
1325996	1326982	c	putative conjugal transfer protein
1326979	1327683	c	putative conjugal transfer protein
1327702	1329000	c	putative conjugal transfer protein
1329012	1329749	c	putative conjugal transfer protein
1329746	1332232	c	putative conjugal transfer protein
1332243	1332515	c	putative conjugal transfer protein
1332512	1332895	c	putative conjugal transfer protein
1332892	1333932	c	putative conjugal transfer protein
1333929	1334375	c	conserved hypothetical protein
1334372	1336372	c	putative conjugal transfer protein
1336606	1336866	c	conserved hypothetical protein
1336911	1337804	c	putative LysR family transcriptional regulator
1337815	1337940	c	putative regulator, pseudogene
1337956	1338126	c	putative esterase, pseudogene
1338123	1338425	c	conserved hypothetical protein
1338412	1339266	c	putative NAD(P)H dehydrogenase [quinone]
1339362	1340264		putative LysR family transcriptional regulator
1340359	1340619	c	conserved hypothetical protein
1340656	1341549	c	putative LysR family transcriptional regulator
1341582	1342424	c	conserved hypothetical protein
1342421	1342804	c	putative 4-carboxymuconolactone decarboxylase
1342737	1343207	c	putative MerR family transcriptional regulator

1343295	1344665	c	putative MFS family transmembrane transporter
1344784	1345635		putative LysR family transcriptional regulator
1345632	1347596	c	conserved hypothetical protein
1347593	1348030	c	conserved hypothetical protein
1348056	1348631	c	putative transmembrane anchor conjugal transfer protein
1348628	1349152	c	conserved hypothetical protein
1349149	1349412	c	putative parB partition protein
1349409	1350047	c	putative ParA/CobQ/CobB/MinD nucleotide binding domain protein
1350298	1351131	c	putative RepA-like replication protein
1351158	1351448	c	conserved hypothetical protein
1351532	1352335	c	conserved hypothetical protein
1352687	1353037	c	conserved hypothetical protein
1353340	1353636		putative HTH transcriptional regulator
1353993	1354307	c	conserved hypothetical protein
1354642	1355064	c	conserved hypothetical protein
1355155	1355490		hypothetical protein
1355532	1356575		putative transposase
1356572	1357087		conserved hypothetical protein
1357200	1357388	c	conserved hypothetical protein
1357451	1359481	c	putative ParB-like nuclease domain protein
1359562	1360392	c	conserved hypothetical protein
1360425	1360622	c	conserved hypothetical protein
1361053	1362183		conserved hypothetical protein
1362149	1363831		putative transmembrane protein
1363664	1364188	c	putative RadC DNA repair protein, pseudogene
1364225	1365508	c	conserved hypothetical protein
1365534	1365839	c	conserved hypothetical protein
1365820	1366203	c	conserved hypothetical protein
1366343	1367584	c	putative prophage integrase

Region 14

1720129	1720437		hypothetical protein
1720450	1720872	c	hypothetical protein
1720901	1721134		conserved hypothetical protein
1721405	1721854		putative modification methylase
1722115	1722405		conserved hypothetical protein
1722566	1723417	c	putative ISXac3 like transposase protein
1723435	1723713	c	putative ISXac3 like transposase
1723923	1724387		putative transmembrane protein

Region 15

1945505	1946305		hypothetical protein
1946295	1947431		putative phage-related protein
1947488	1948663	c	putative phage integrase protein
1948606	1948884	c	putative phage excisionase
1949118	1949345		hypothetical protein
1949409	1949687		hypothetical protein
1949779	1950516	c	hypothetical protein
1950503	1950811	c	hypothetical protein
1950814	1951029	c	conserved hypothetical protein
1951052	1951777	c	conserved hypothetical protein
1951781	1952488	c	conserved hypothetical protein
1952747	1953649	c	putative recombination-associated protein
1953804	1954067	c	hypothetical protein
1954064	1954582	c	hypothetical protein
1954703	1955122	c	putative transcriptional regulatory protein
1955122	1955622	c	hypothetical transmembrane anchored protein
1955683	1956141	c	hypothetical protein
1956143	1956526	c	conserved hypothetical protein
1956600	1956920		hypothetical protein
1957137	1957514	c	conserved hypothetical protein
1958018	1958398	c	hypothetical protein
1958628	1959035		hypothetical protein
1959004	1959528		hypothetical protein
1959546	1960592		conserved hypothetical protein
1960585	1961139		conserved hypothetical protein
1961136	1961675		conserved hypothetical protein
1961672	1962103		conserved hypothetical proteins
1962100	1962438		hypothetical protein
1962435	1963262		conserved hypothetical protein
1963259	1964230		hypothetical protein
1964366	1964863	c	conserved hypothetical protein
1964951	1965382		putative transmembrane protein
1965431	1965763		putative transmembrane protein
1966393	1966935		putative transmembrane phage lysozyme
1967303	1967632		hypothetical protein
1967637	1967933		hypothetical protein
1967963	1968247		conserved hypothetical protein

1968244	1968648		conserved hypothetical protein
1968648	1969196		conserved hypothetical protein
1969201	1969728		conserved hypothetical protein
1969933	1971006		putative phage tail fiber protein
1971470	1972105		putative phage terminase
1972109	1973662		putative DNA packaging protein gp17 (terminase)
1973659	1973937		hypothetical protein
1973939	1974127		hypothetical protein
1974124	1974597		hypothetical protein
1974597	1976768		putative phage protein
1976755	1977750		putative phage related protein
1977756	1977971		hypothetical protein
1978049	1979158		putative phage-related protein
1979230	1979676		hypothetical protein
1979731	1980294		putative phage-related protein
1980296	1980976		conserved hypothetical protein
1980978	1982291		conserved hypothetical protein
1982497	1983120		conserved hypothetical protein
1983120	1985021		hypothetical protein
1985030	1985503		conserved hypothetical protein
1985520	1985990		conserved hypothetical protein
1985992	1986729		conserved hypothetical protein
1986726	1988207		putative phage-related protein
1988210	1988542		hypothetical protein
1988556	1990445		conserved hypothetical protein
1990474	1991988		hypothetical protein
1991992	1999923		conserved hypothetical protein
2000022	2000375		conserved hypothetical protein, partial
2000458	2001390		conserved hypothetical protein
2001421	2001822		conserved hypothetical protein
2001926	2002540		conserved hypothetical protein

Region 16

3089388	3089654	c	hypothetical protein
3089651	3090445	c	putative peptidase
3090702	3092462		putative Hep Hag family adhesin
3093052	3093834	c	conserved hypothetical protein
3093974	3094429		hypothetical protein

3094419	3097199	c	putative conjugal transfer protein TraA
3097317	3098039	c	hypothetical protein
3098061	3099935	c	putative type IV secretory protein conjugation protein TraD
3099913	3100737	c	putative ankyrin repeat protein
3100776	3101039	c	hypothetical protein
3101231	3101434	c	putative transmembrane protein
3101572	3102096	c	hypothetical protein
3102183	3105164	c	conserved hypothetical protein
3105188	3105520	c	putative transmembrane protein
3105586	3106713	c	putative transmembrane protein
3106710	3107345	c	conserved hypothetical exported protein
3107374	3107775	c	hypothetical protein
3107997	3108554		hypothetical protein
3108541	3109632		hypothetical protein
3109638	3110534		conserved hypothetical protein
3110654	3110800		conserved hypothetical protein
3110838	3111677	c	conserved hypothetical protein
3111950	3112555	c	hypothetical protein
3112555	3113151	c	putative plasmid partitioning like protein
3113266	3113730	c	hypothetical protein
3114059	3115846		putative type IV pilus protein
3116009	3116770	c	conserved hypothetical protein
3116767	3117684	c	hypothetical protein
3117773	3118255	c	conserved hypothetical protein
3118477	3119181	c	conserved hypothetical protein
3119229	3119615		hypothetical protein
3119590	3120078	c	hypothetical protein
3120071	3120601	c	conserved hypothetical protein
3120704	3121537	c	hypothetical protein
3121497	3121979	c	putative permease ABC transporter protein
3122039	3122401	c	hypothetical exported protein
3122585	3123445	c	putative transposase
3123454	3123737	c	putative transposase, pseudogene
3124357	3124710	c	hypothetical protein
3124816	3125547	c	hypothetical protein
3125547	3127070	c	putative phage-related integrase

Region 20

3913248	3914057	c	putative transmembrane protein

3914061	3917930	c	putative autotransporter haemagglutinin-related protein
3918118	3918825	c	putative giant cable pilus chaperone protein
3918842	3919987	c	putative minor pilin and initiator protein
3919984	3922710	c	putative outer membrane usher (colonisation factor antigen I subunit c)
3922761	3923270	c	putative giant cable pilus fimbrial subunit
3923789	3924877	c	putative transmembrane protein
3925243	3926520	c	putative citrate synthase
3926817	3927059	c	putative 50S ribosomal protein L31
3927148	3928086	c	putative nucleoside hydrolase
3928471	3930582	c	putative ATP-dependent DNA helicase
3930590	3930976	c	putative endoribonuclease protein

```
CLUSTAL 2.0.3 multiple sequence alignment


R1-SK279a       CGCGGCGGACCATGCCCAGAGCCTGTTCGAACGCCTTCGGCGAAACCGCCGTGACCAGGG 60
R1-SK279(1)     CGCGGCGGACCATGCCCAGAGCCTGTTCGAAAGCCTTACGCGAAACCGCCATGACCAGGG 60
R1-S28          CGCGGCGGACCATGCCCAGAGCCTGTTCGAAAGCCTTAGGCGACACTGCAGTGACCACGG 60
                ******************************* *****  **** ** **  ****** **

R1-SK279a       CGCCGTGCGCACCGCCGATCTCCTTCTTCAGGAATGCGGCCGGATCGGTGGTACGCGCGT 120
R1-SK279(1)     CGCCTTGCGCACCGCCGATCTCCTTCTTCACGAATGCGGCCGGATCGGTGGTACGCGCGT 120
R1-S28          CGCCTTGCGCACCGCCAATCTCCTTCTTCACGAATGCGGCCCGATCGATGGTACGCACGT 120
                **** *********** ************* ********** ***** ******** ***

R1-SK279a       TCACGGTGATCTGCGCGCCGAGCTGCCGGGCCAATGCCAGCTTGTTGTCGTCCACGTCCA 180
R1-SK279(1)     TCACGGTGATCTGCGCGCCGAGCTGCCGGGCCAATGCCAGCTTGTTGTCGTCCACGTCCA 180
R1-S28          TCACTGTGATCTGCGCACCGAGCTGCCGGGCCAATGCCAGCTTGTTGTCGTCCACGTCCA 180
                **** *********** *******************************************

R1-SK279a       CCGCAGCCACATTCAGGCCCATCGCACGGGCGTACTGCACCGCCATGTGGCCCAGGCCAC 240
R1-SK279(1)     CCGCAGCCACATTCAGGCCCATCGCACGGGCGTACTGCACCGCCATGTGGCCCAGGCCAC 240
R1-S28          CCGCAGCCACGTTCAGGCCCATCACACGGGCGTACTGCACCGCCATGTGGCCCAGGCCAC 240
                ********** ************ ************************************

R1-SK279a       CGATGCCGGAAATCACCACCCAGTCCCCGGGCTTGGTGTCGGTCACCTTCAGGCCCTTGT 300
R1-SK279(1)     CGATGCCGGAAATCACCACCCAGTCCCCGGGCTTGGTGTCGGTCACCTTCAGGCCCTTGT 300
R1-S28          CGATGCCGGAAACCGCCACCCAGTCCCCGGGCTTGGTGTCGGTCACCTTCAGGCCCTTGT 300
                ************ * *********************************************

R1-SK279a       AGACGGTCACGCCGGCGCACAACACCGGCGCGATCTCGACGAAGCCCACTTCCTTCGGAA 360
R1-SK279(1)     AGACGGTCACGCCGGCGCACAACACCGGCGCGATCTCGACGAAGCCCACTTCCTTCGGAA 360
R1-S28          AGACGGTCACGCCGGCGCACAGCACCGGCGCGATCTCGACGAAGCCCACTTCCTTCGGAA 360
                ********************* **************************************

R1-SK279a       GCAGGCCGACATAGTTGGCATCGGCCAGTGCGTACTCGGCGAAGCCGCCGTTGACCGAGT 420
R1-SK279(1)     GCAGGCCGACATAGTTGGCATCGGCCAGTGCGTACTCGGCGAAGCCGCCGTTGACCGAGT 420
R1-S28          GCAGGCCGACATAGTTGGCATCGGCCAGTGCGTACTCGGCGAAGCCGCCGTTGACCGAGT 420
                ************************************************************

R1-SK279a       AACCGGTGTTGCGCTGCGTCTCGCACAGCGTTTCCCAGCCACCCAGGCAGTGTTCGCAAT 480
R1-SK279(1)     AACCGGTGTTGCGCTGCGTCTCGCACAGCGTTTCCCAGCCACCCAGGCAGTGTTCGCAAT 480
R1-S28          AACCGGTGTTGCGCTGCGTCTCGCACAGCGTTTCCCAGCCACCCAGGCAGTGTTCGCAAT 480
                ************************************************************

R1-SK279a       GGCCACACGCCGAGTACAACCAGGGGATGCCGACCCTGTCGCCTTCCTTGACGTGCCCTA 540
R1-SK279(1)     GGCCACACGCCGAGTACAACCAGGGGATGCCGACCCTGTCGCCTTCCTTGACGTGCCCTA 540
R1-S28          GGCCACACGCCGAGTACAACCAGGGGATGCCGACCCTGTCGCCTTCCTTGACGTGCCCTA 540
                ************************************************************

R1-SK279a       CCCCGCCTCCCACGGCCACGACGTGCCCCACGCCCTCGTGGCCGGGGATGAATGGCGGGT 600
R1-SK279(1)     CCCCGCCTCCCACGGCCACGACGTGCCCCACGCCCTCGTGGCCGGGGATGAATGGCGGGT 600
R1-S28          CCCCGCCTCCCACGGCCACGACGTGCCCCACGCCCTCGTGGCCGGGGATGAATGGCGGGT 600
                ************************************************************

R1-SK279a       TCGGTTTCACCGGCCAGTCGCCCTCGGCGGCGTGCAGGTCGGTGTGGCAGACGCCACAGG 660
R1-SK279(1)     TCGGTTTCACCGGCCAGTCGCCCTCGGCGGCGTGCAGGTCGGTGTGGCAGACGCCACAGG 660
R1-S28          TCGGTTTCACCGGCCAGTCGCCCTCGGCGGCGTGCAGGTCGGTGTGGCAGACGCCACAGG 660
                ************************************************************

R1-SK279a       GCCTCGATCCTGACCAGCACCTCGCCCGCCCCCGGGCGCGGTACCGAGACTTCCTCGATG 720
R1-SK279(1)     -CCTCGATCCTGACCAGCACCTCGCCCGCCCCCGGGCGCGGTACCGAGACTTCCTCGATG 719
R1-S28          -CCTCGATCCTGACCAGCACCTCGCACGCCCCCGGGCGCGGTACCGAGACTTCCTCGATG 719
                 ************************ **********************************

R1-SK279a       ACCAGCGGCT 730
R1-SK279(1)     ACCAGCGGCT 729
R1-S28          ACCAGCGGCT 729
                **********
```