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.

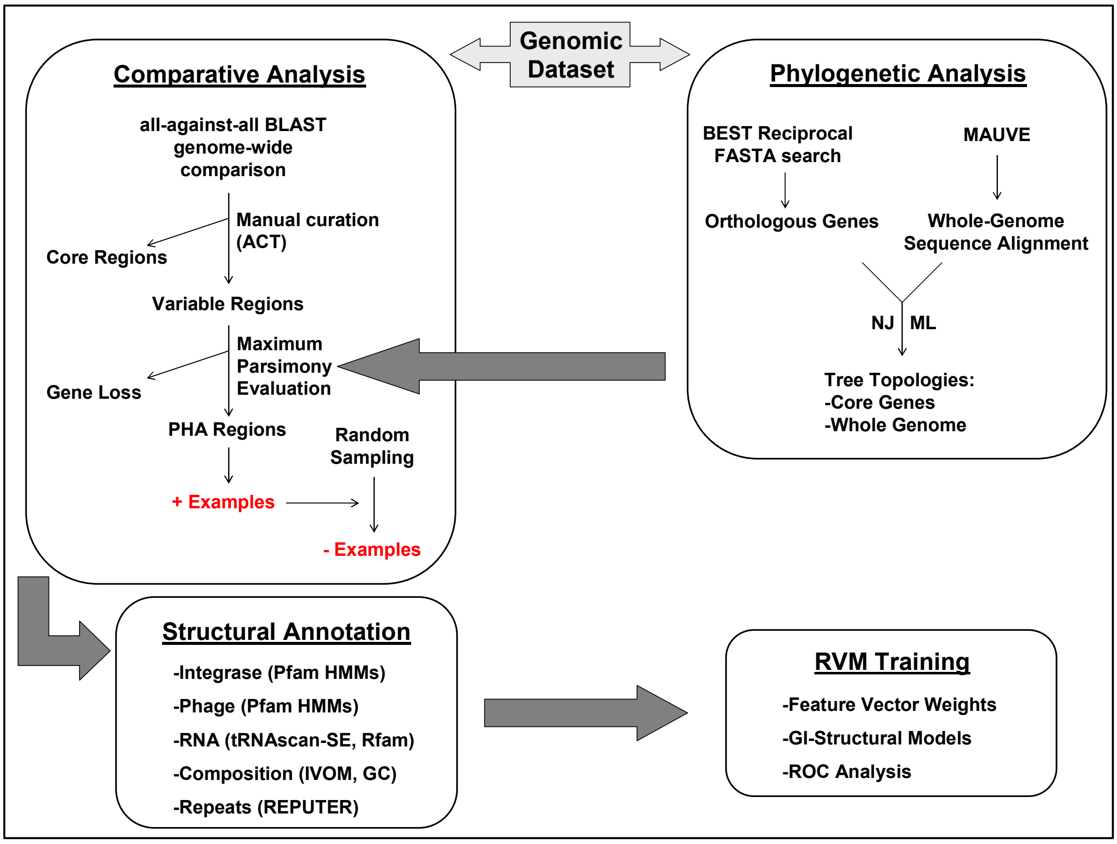


Figure 4.1: Flowchart summarizing the major steps in the methodology followed throughout this analysis: A phylogenetic analysis using both whole-genome sequence (if applicable) and the amino acid sequence of the core gene products was carried out enabling the construction of the reference tree topology 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

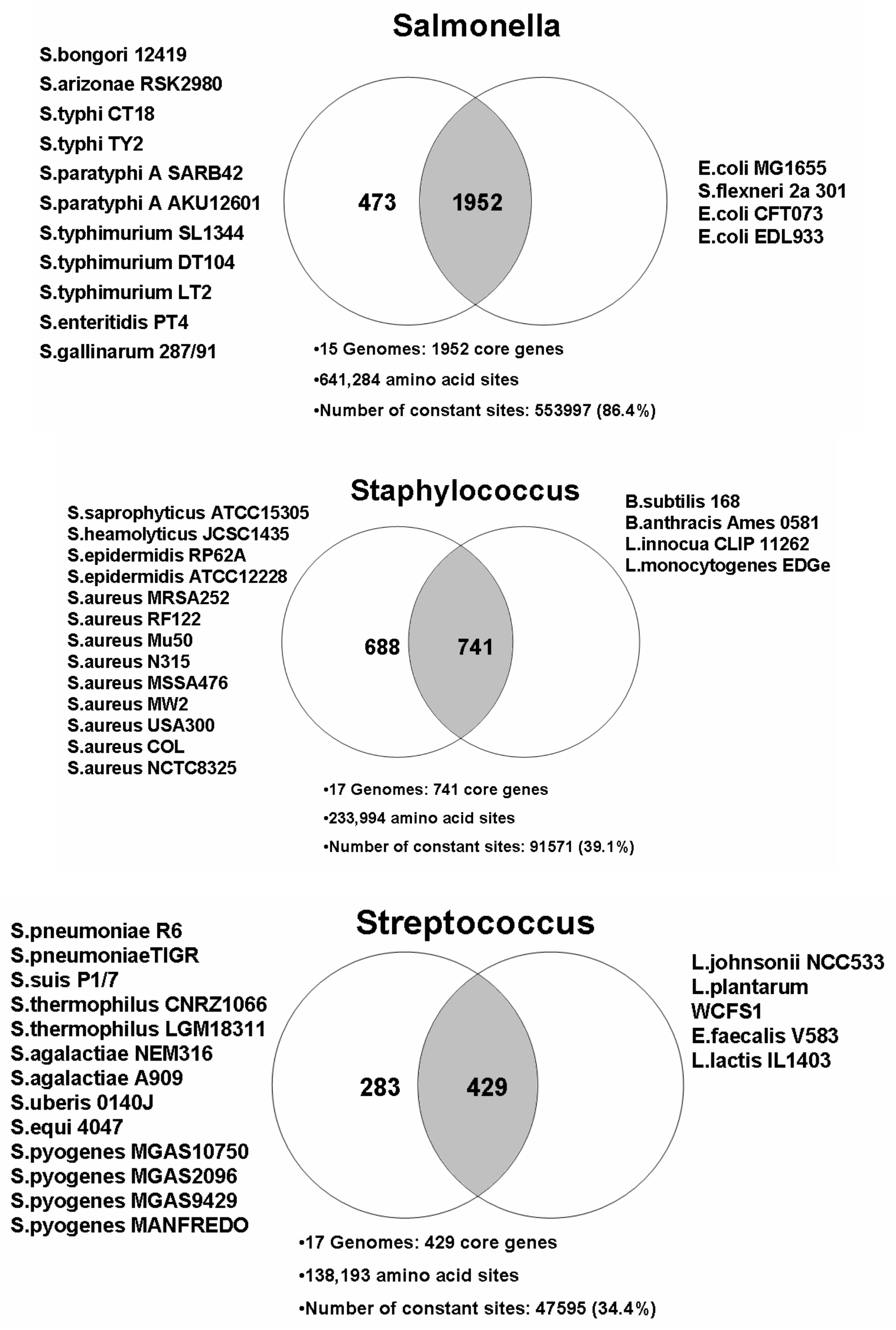


Figure 4.2: Venn diagram illustrating the orthologous genes shared between each of the three reference genera and the corresponding outgroup strains: 473 *Salmonella*-specific and 1952 core genes (genes shared between the *Salmonella* and the four outgroup strains) (top), 688 *Staphylococcus*-specific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

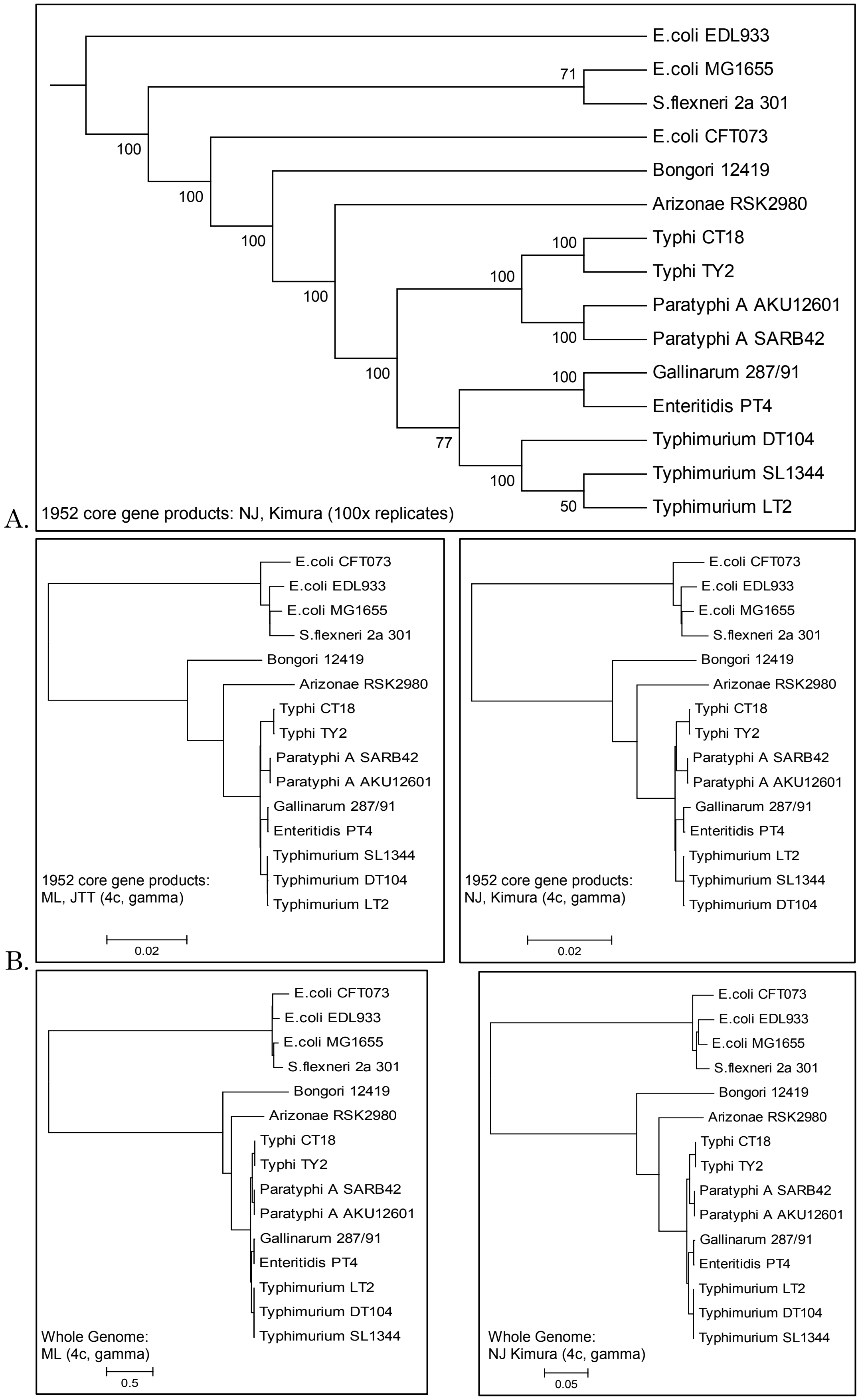
4.2.5 Comparative analysis

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


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

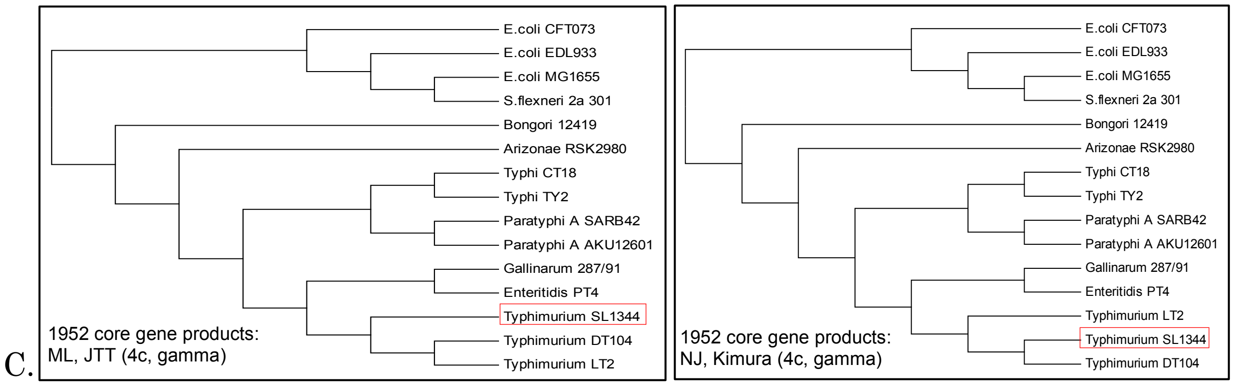
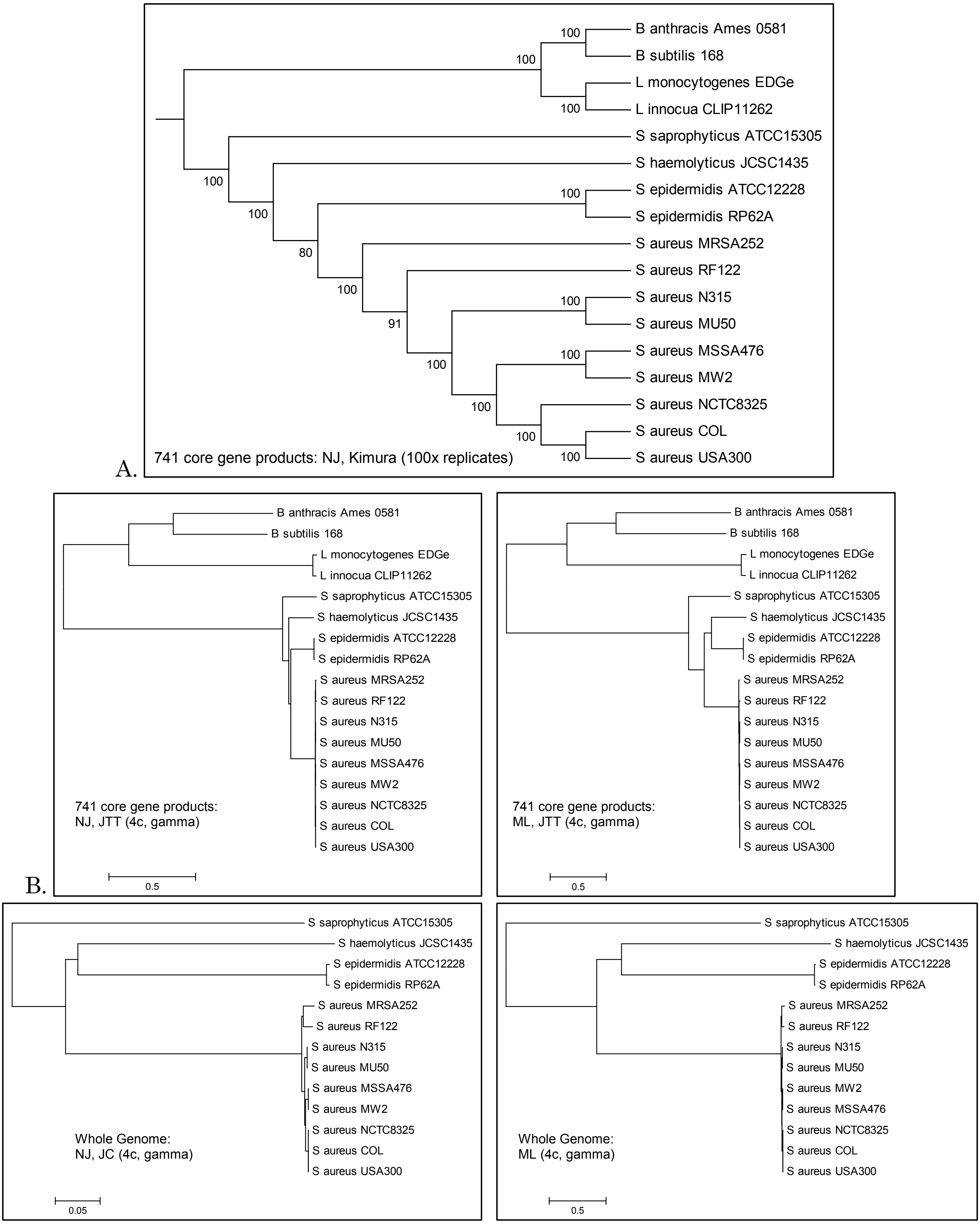
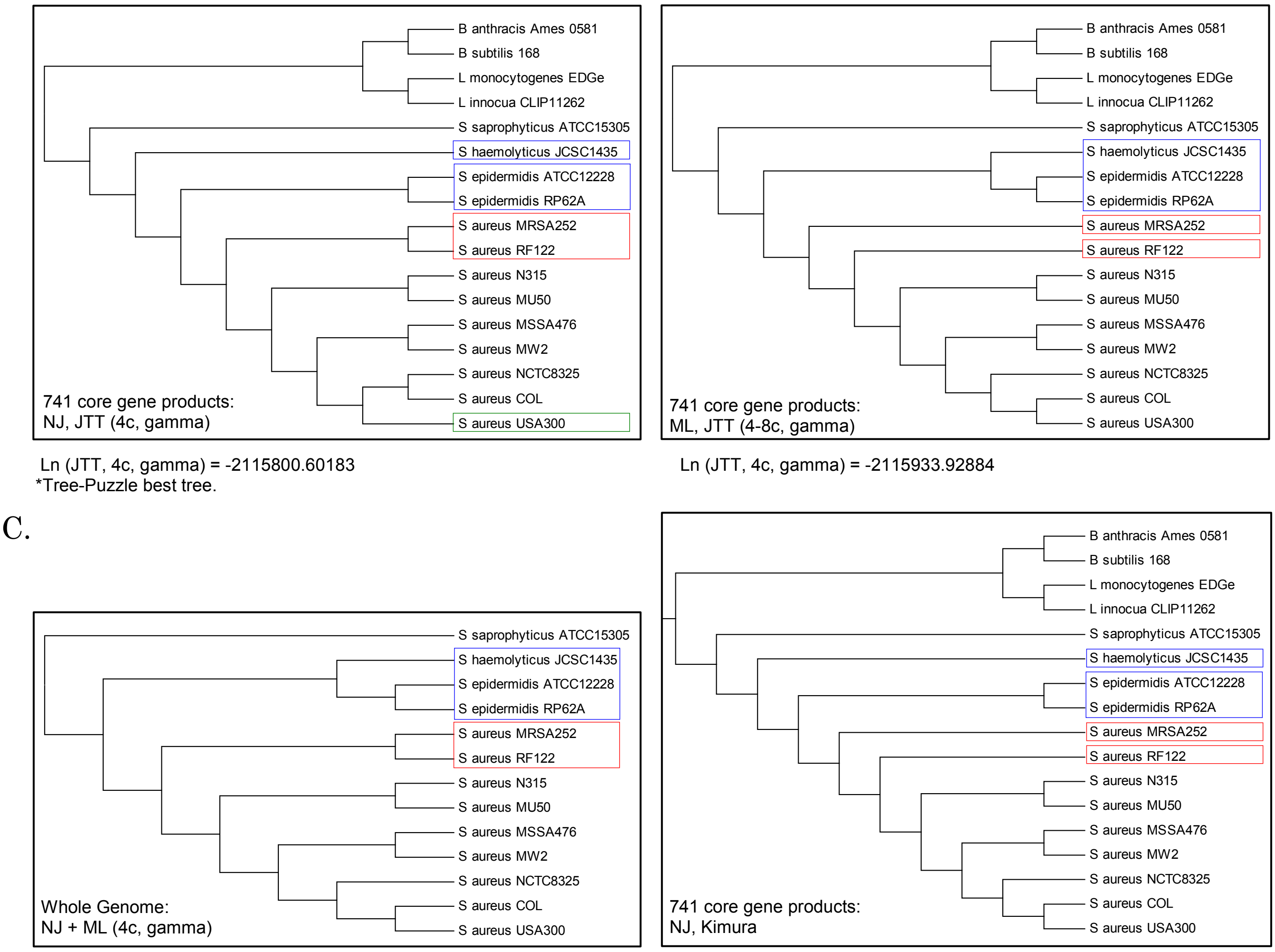


Figure 4.3: **A.** The phylogenetic relationship between the 11 *Salmonella* and the four outgroup genomes (ignoring branch length), is shown as cladogram. Bootstrap values (proportions out of 100) are given for each node. The tree topology is based on the amino acid sequence of 1952 core gene products 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.


A.
741 core gene products: NJ, Kimura (100x replicates)
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
100
100
100
100
80
100
91
100
100
100
100
100
100
100
B.
741 core gene products:
NJ, JTT (4c, gamma)
0.5
741 core gene products:
ML, JTT (4c, gamma)
0.5
Whole Genome:
NJ, JC (4c, gamma)
0.05
Whole Genome:
ML (4c, gamma)
0.5



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

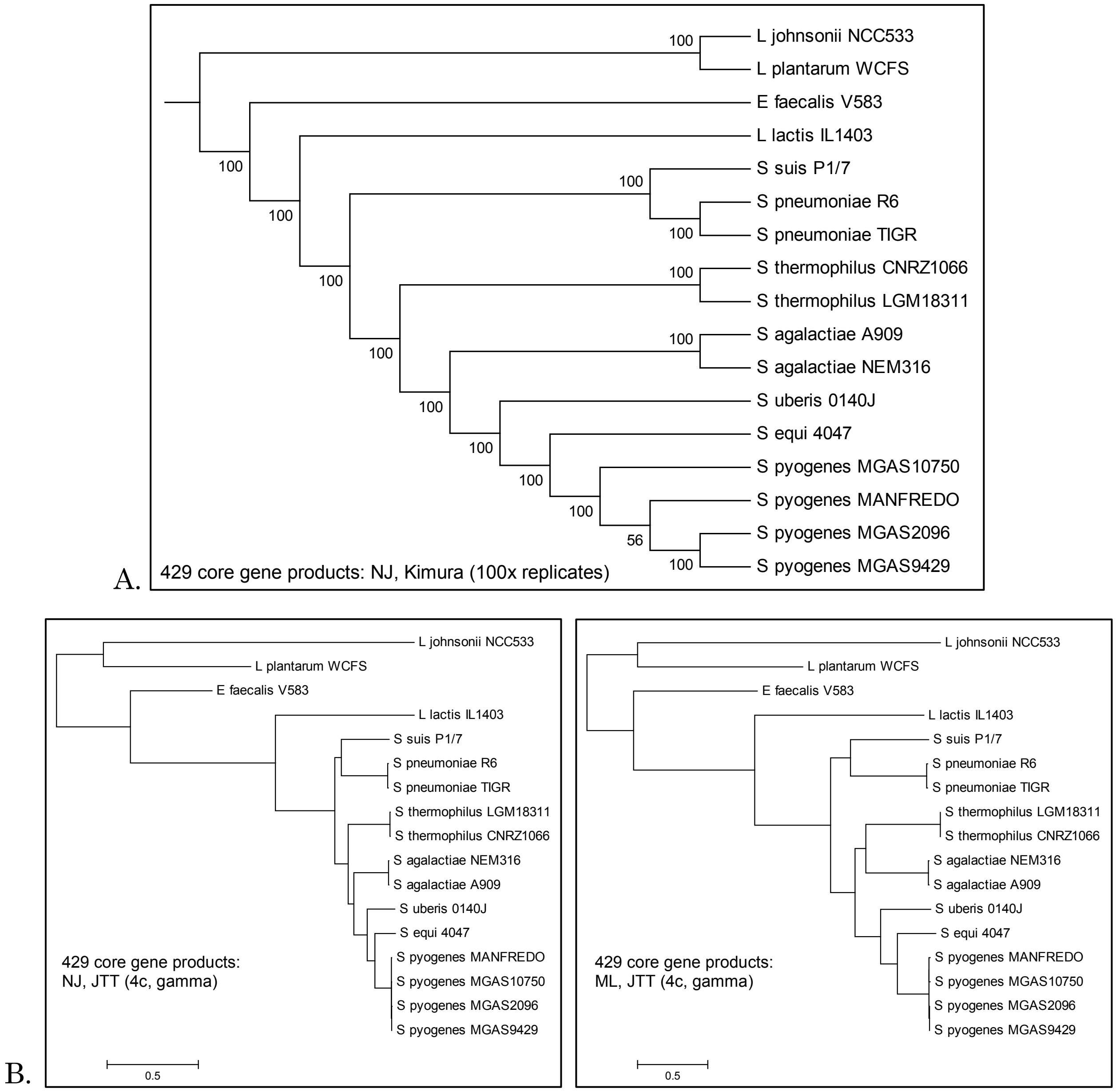
Ln (JTT, 4c, gamma) = -2115933.92884

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
100
56
100
B.
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
429 core gene products:
NJ, JTT (4c, gamma)
0.5
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
429 core gene products:
ML, JTT (4c, gamma)
0.5

C.

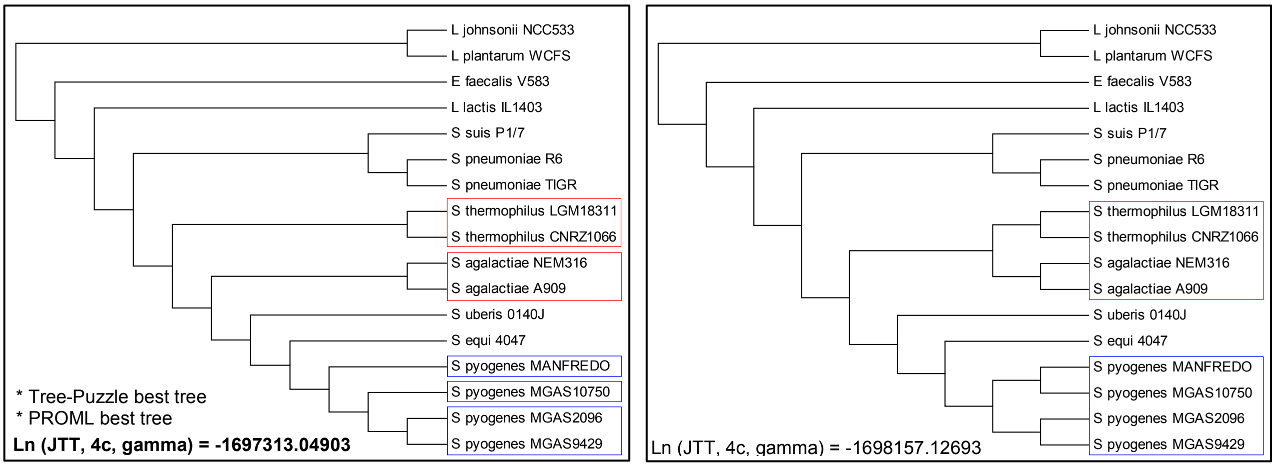
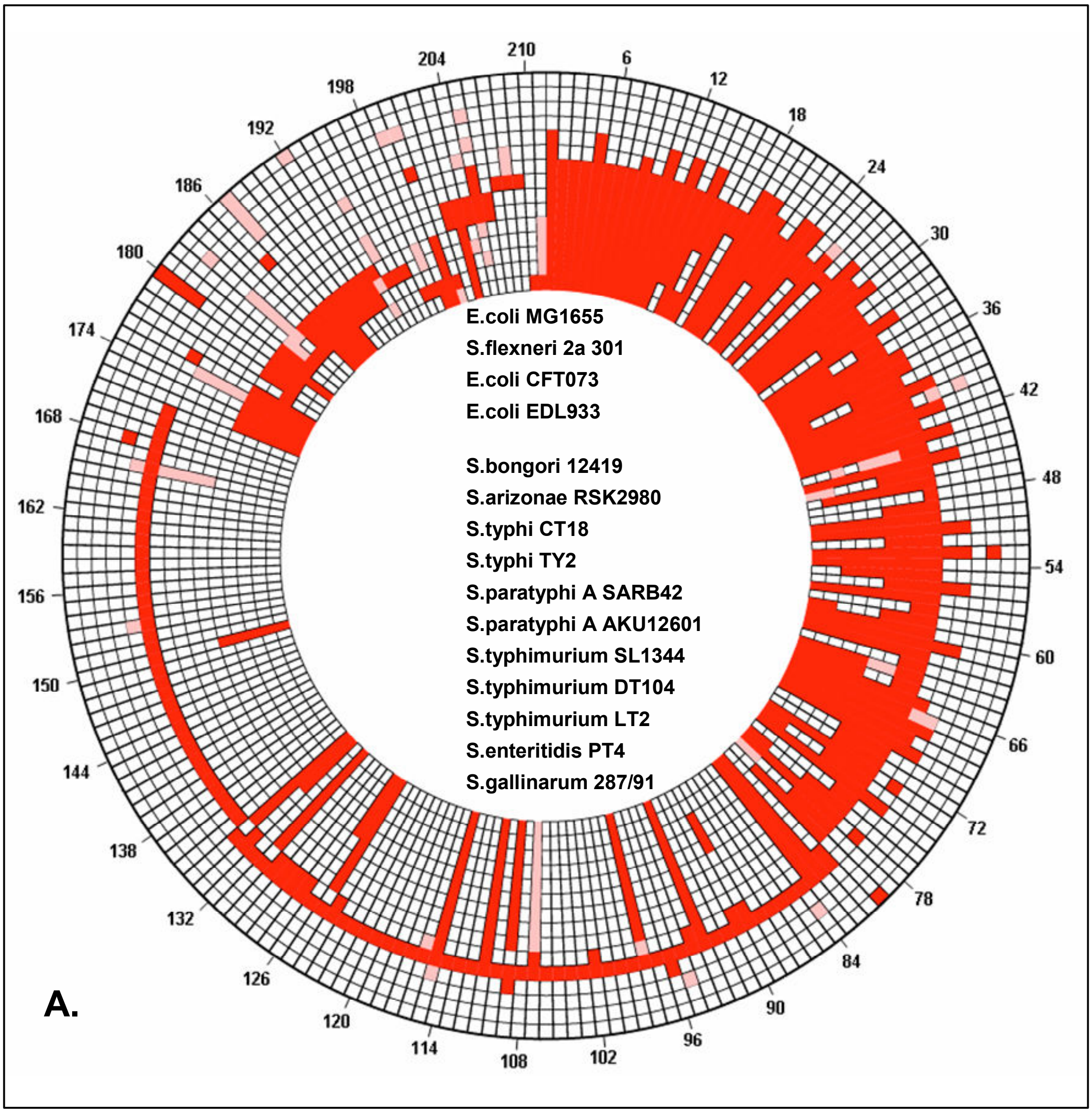
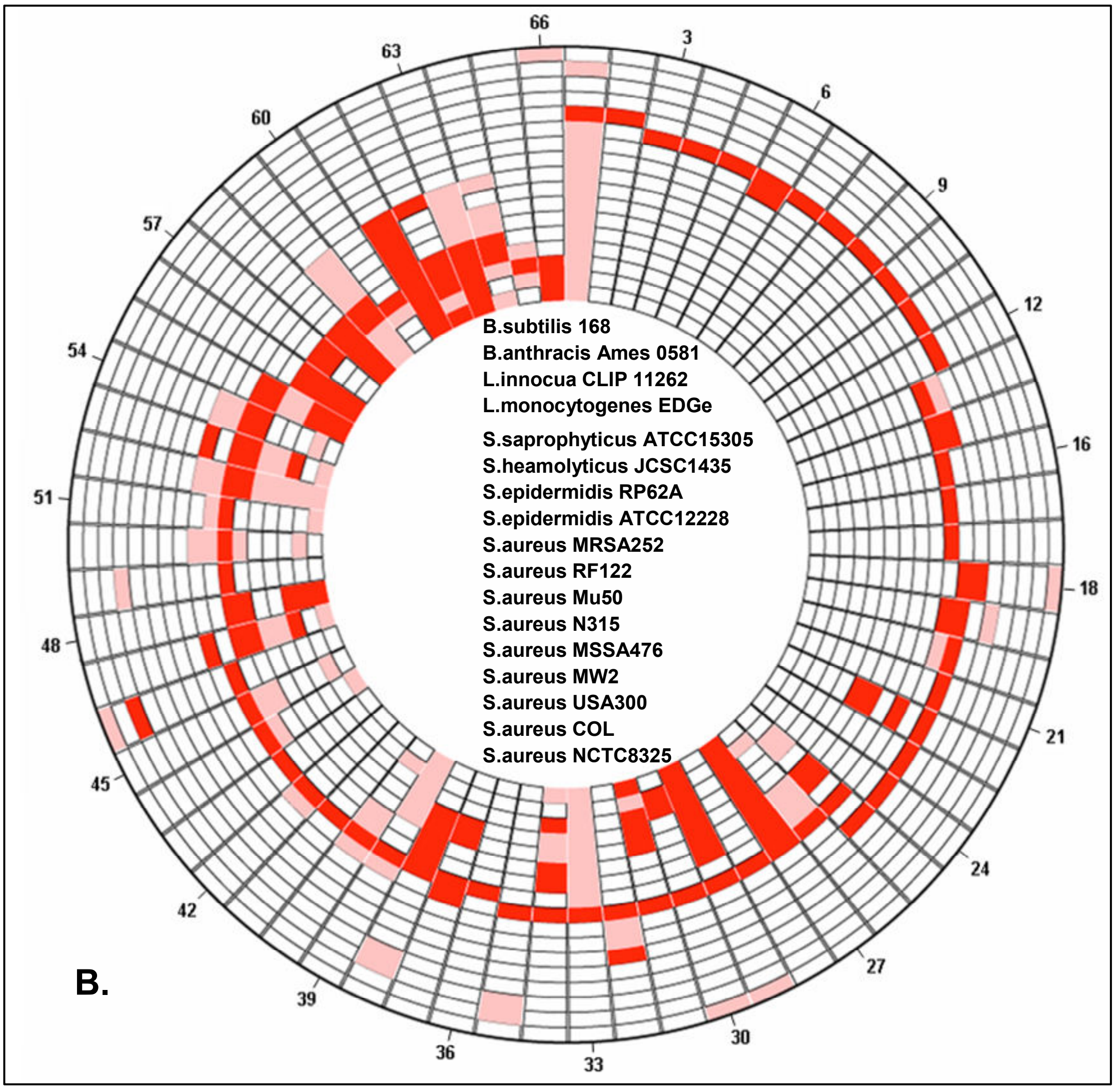


Figure 4.5: **A.** The phylogenetic relationship between the 13 *Streptococcus* and the four outgroup genomes (ignoring branch length), is shown as cladogram. Bootstrap values are given for each node. The tree topology is based on the amino acid sequence of 429 core gene products shared by the 17 genomes. **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.


B.
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

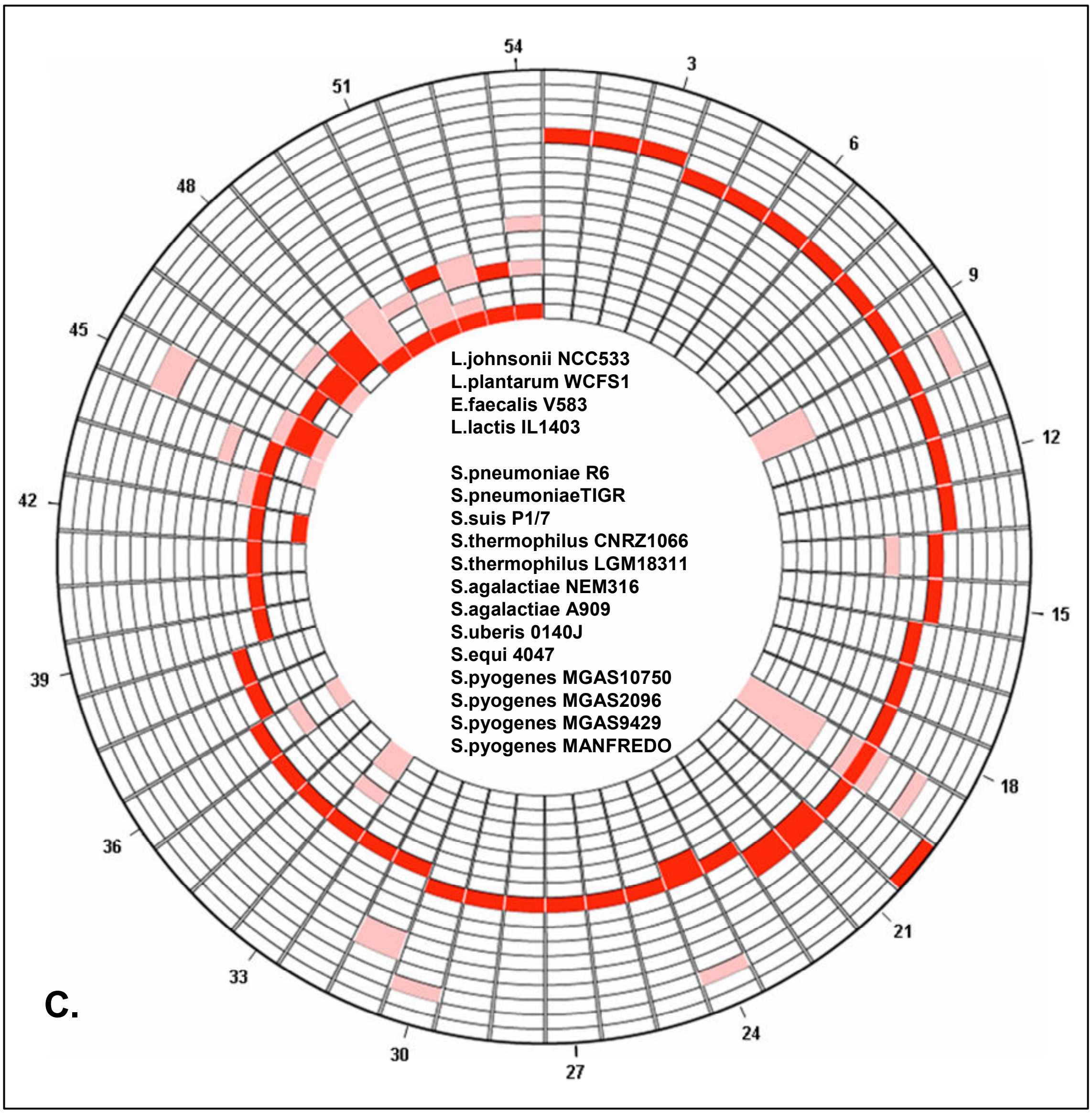


Figure 4.6: Circular map of the *Salmonella* (A), *Staphylococcus* (B) and *Streptococcus* (C) "mobilome", illustrating the phylogenetic distribution of the putative GIs identified in the three reference lineages (red: presence, pink: partial presence, white: absence). The list of strains (outwards-inwards 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).

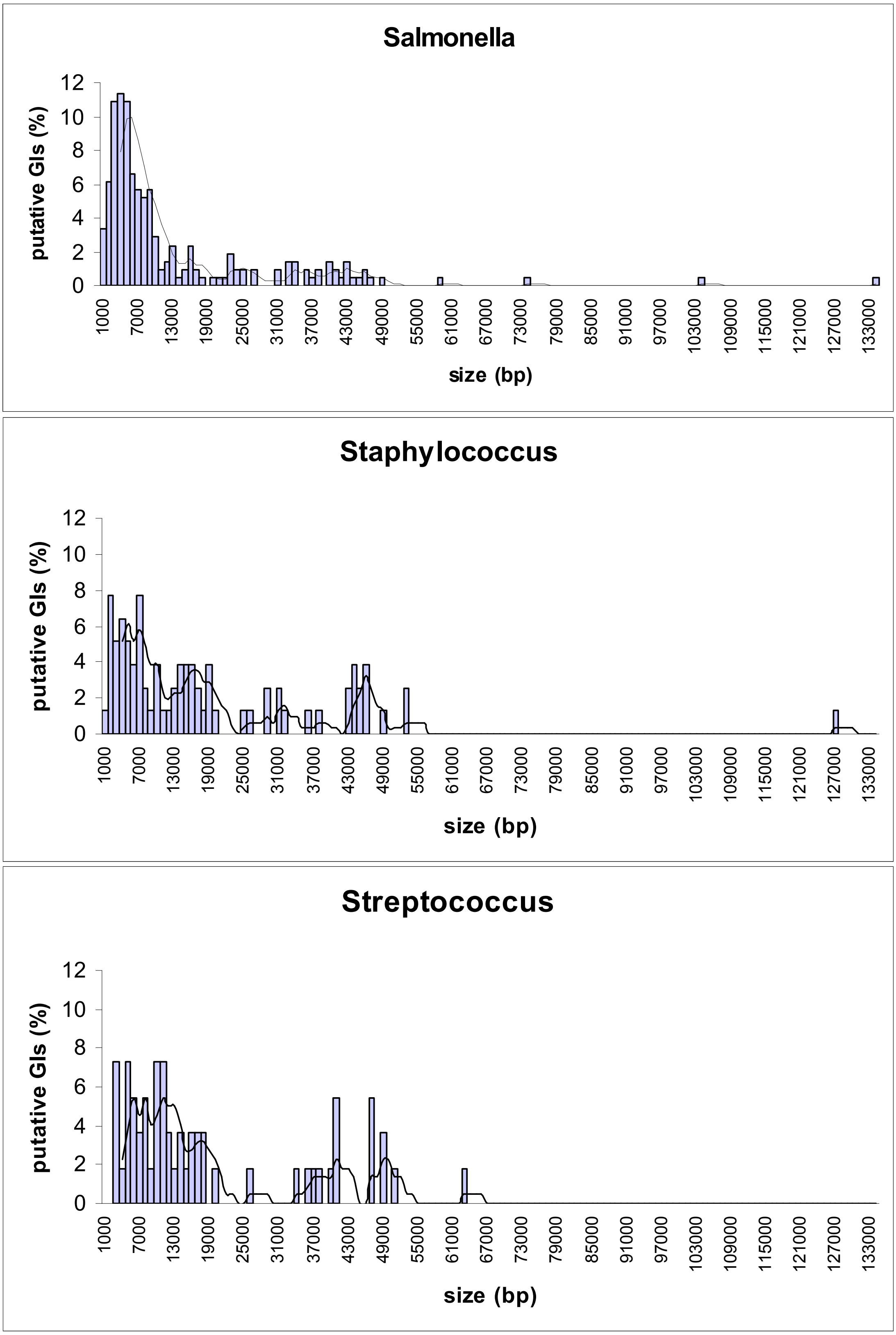


Figure 4.7: Size distribution of the putative Genomic Islands identified in this analysis for the three reference genera.

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.

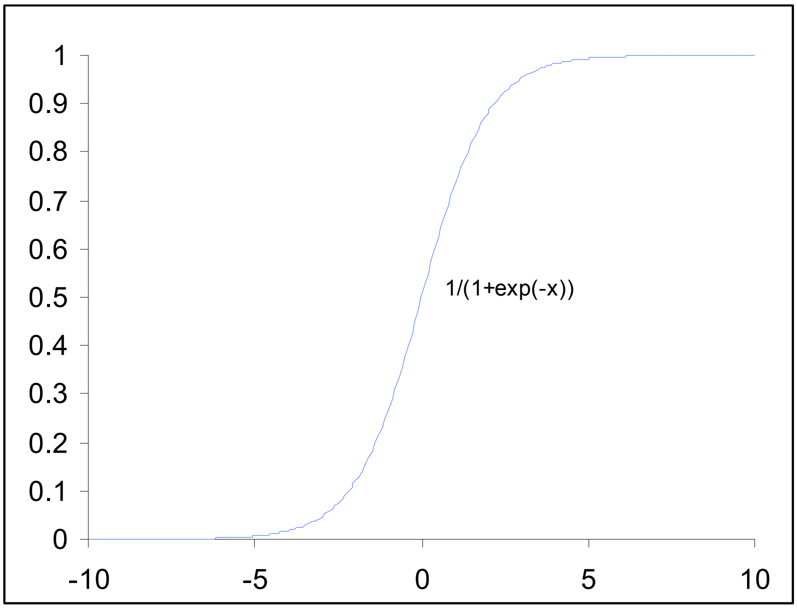


Figure 4.8: The logistic function $f(x) = 1/(1+e^{-x})$.

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*.

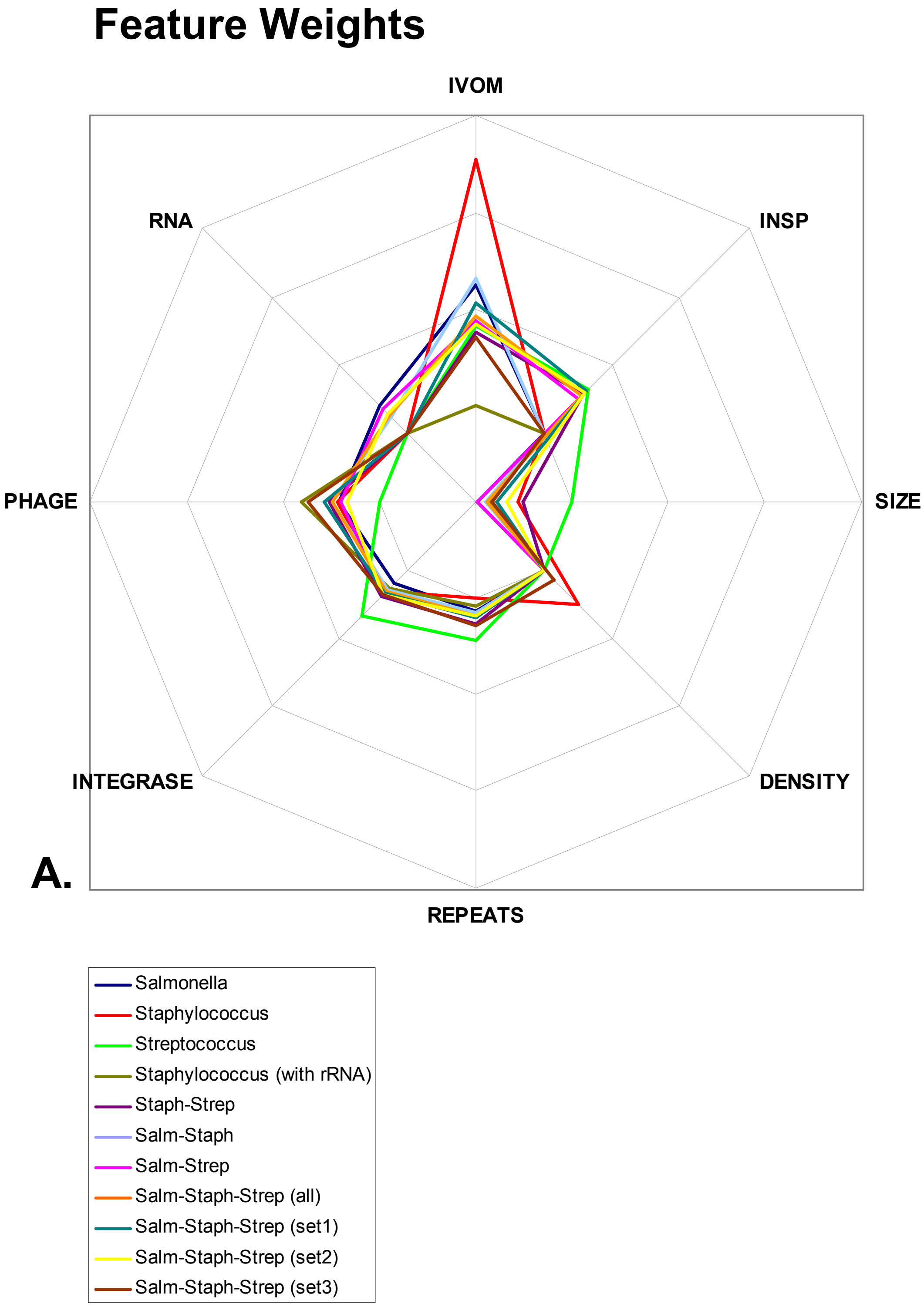
4.3.1.1 Genus-specific

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)

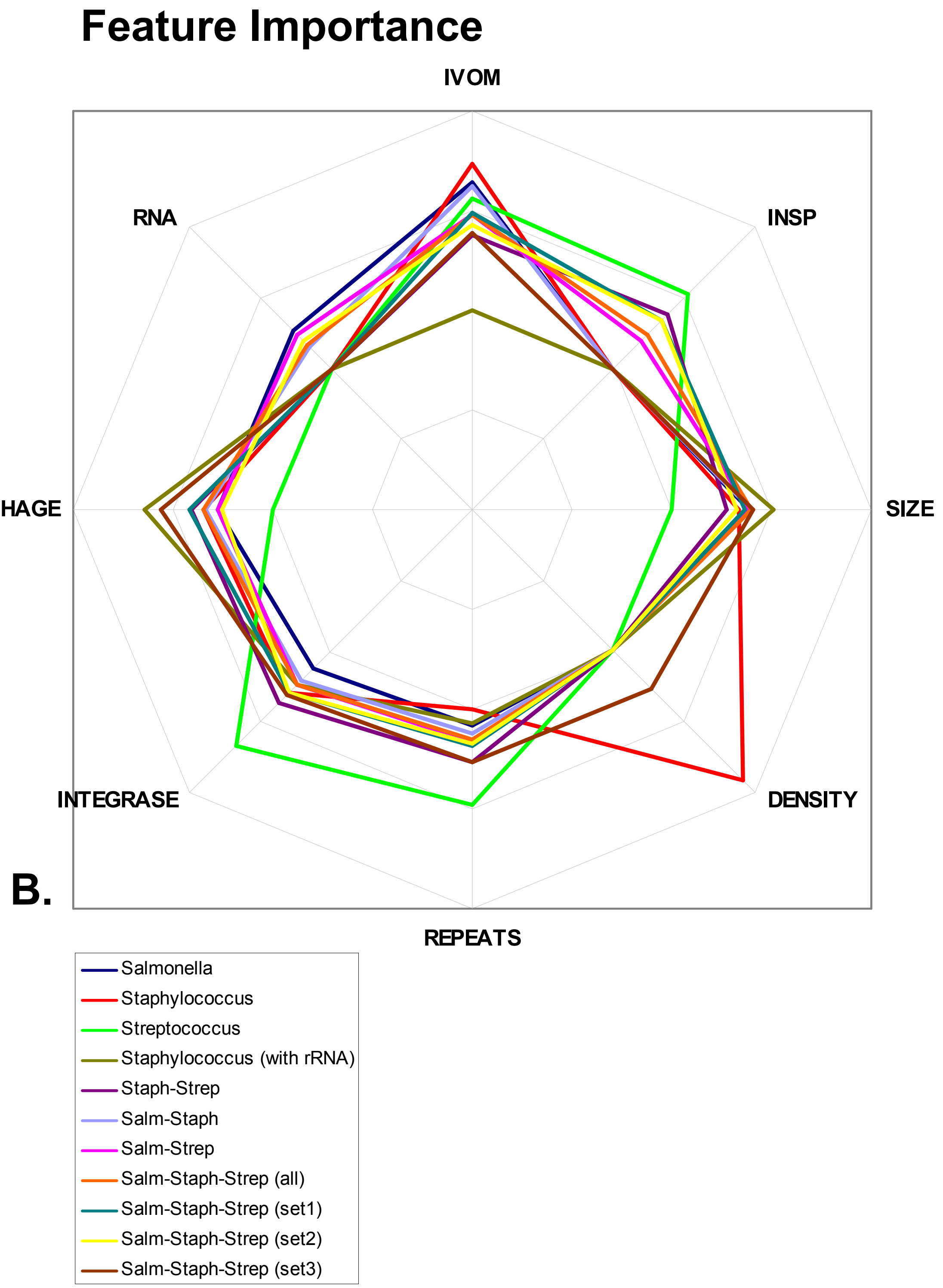


Figure 4.9: Radar diagram illustrating the feature weight (A) and "importance" (B) of the eight structural features under different GI models, based on 11 training datasets. Features: IVOM (feature composition), INSP (insertion point), SIZE (the size of each region), DENSITY (gene density), REPEATS (repeats 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).

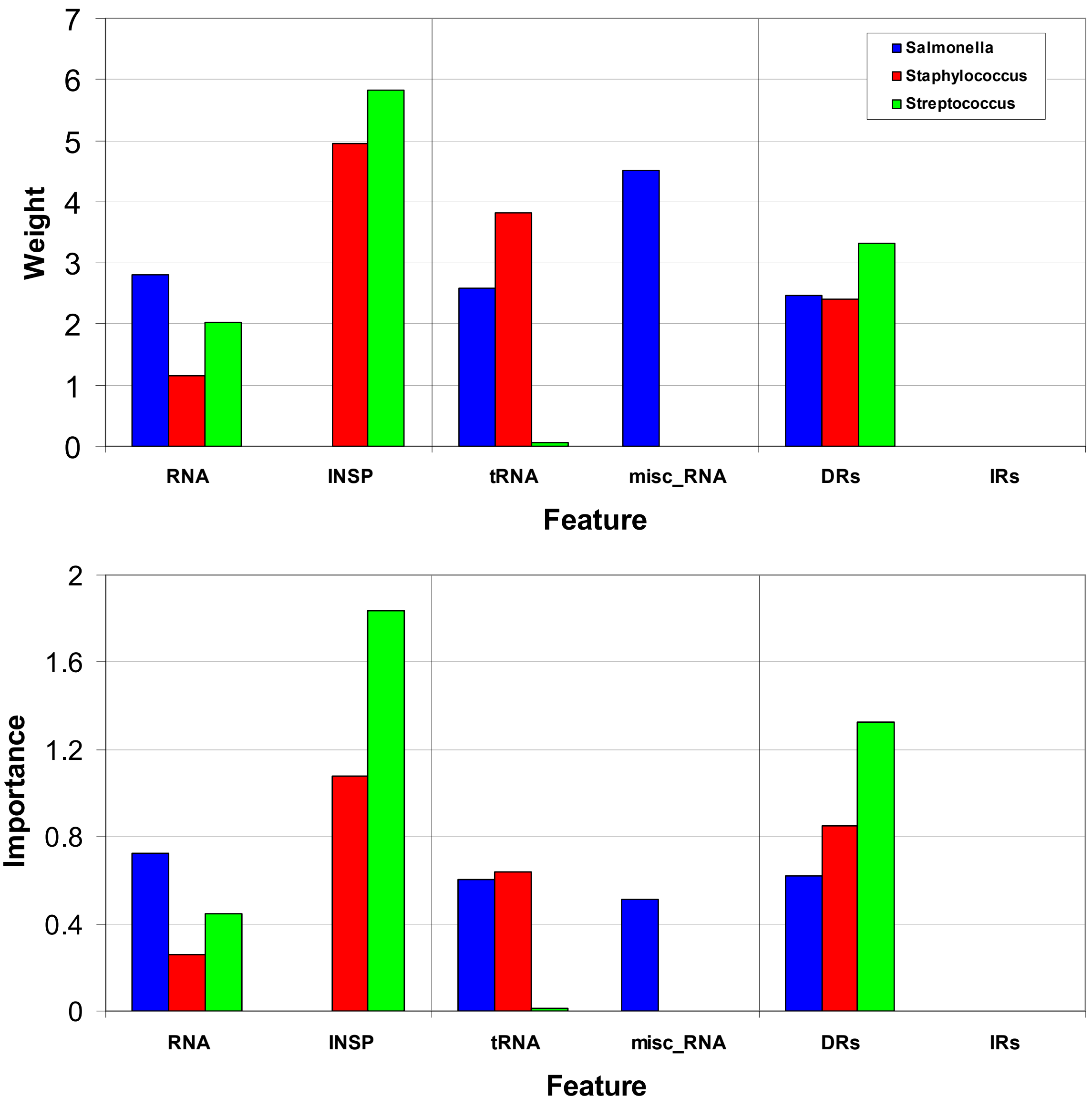


Figure 4.10: Bar chart illustrating the feature weight (top) and "importance" (bottom) of six structural features (evaluated pair-wise), under three different dual-featured GI models, trained on: *Salmonella*, *Staphylococcus* and *Streptococcus*-specific regions respectively. Features: [RNA, INSP],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

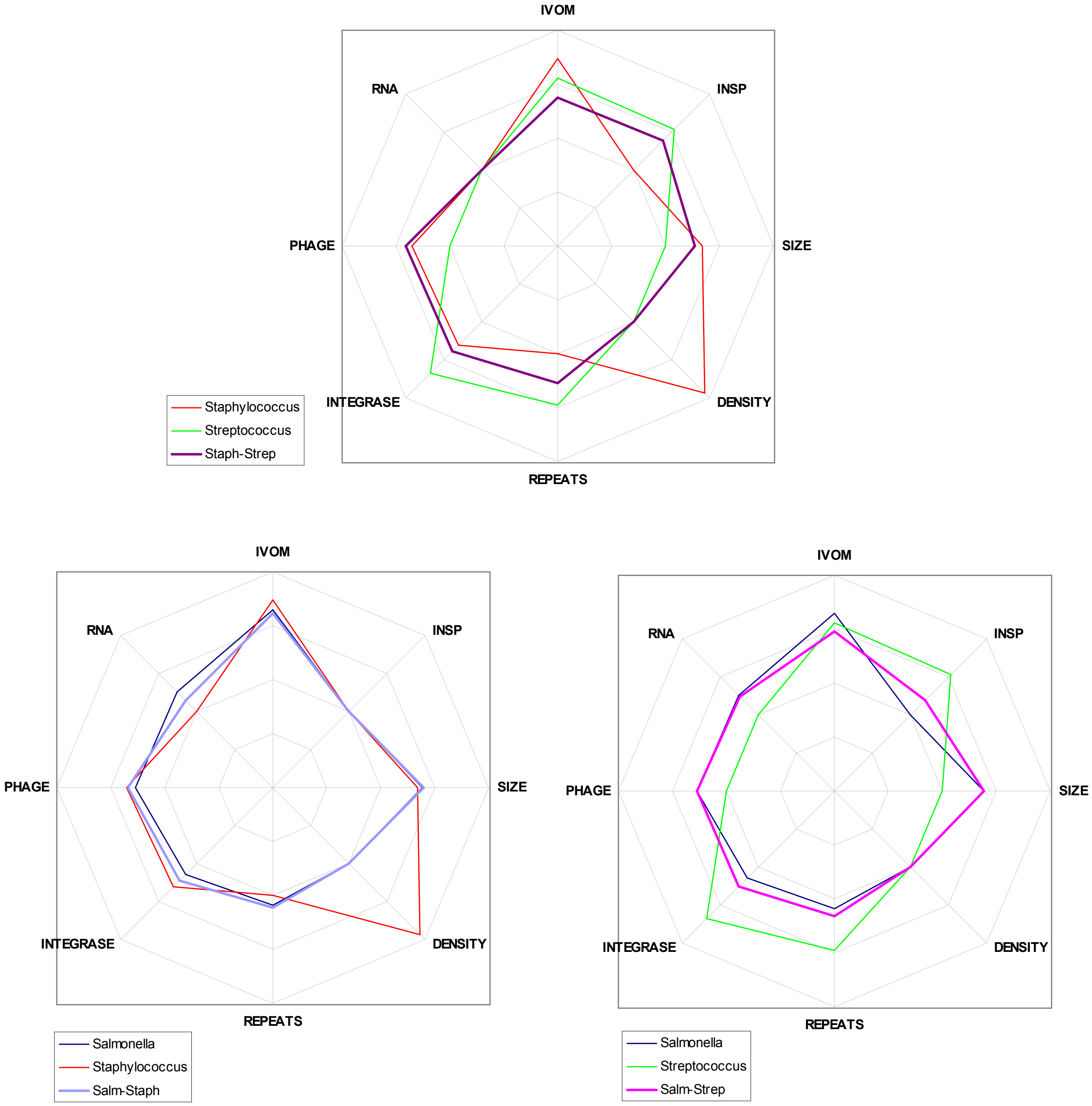


Figure 4.11: Radar diagram illustrating the "importance" of eight structural features under different genus-specific and cross-genus (2 genera) GI models: *Staphylococcus-Streptococcus* (top), *Salmonella-Staphylococcus* (bottom-left) and *Salmonella-Streptococcus* (bottom-right).

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).

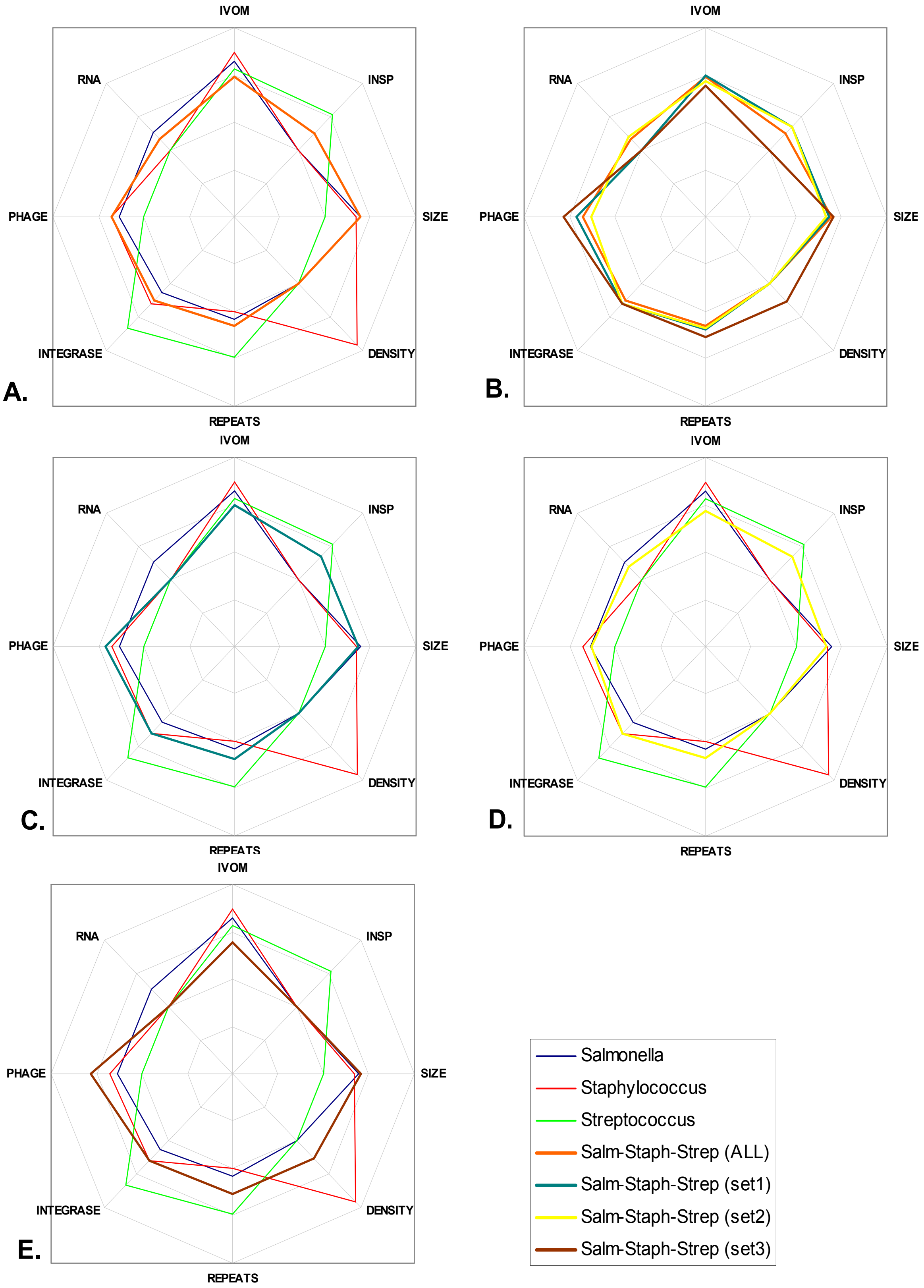


Figure 4.12: Radar diagram illustrating the "importance" of eight structural features under different genus-specific and cross-genus (3 genera) GI models: The *Salmonella* complete dataset (A), set1 (C), set2 (D) and set3 (E) are combined with the complete *Staphylococcus* and *Streptococcus* training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

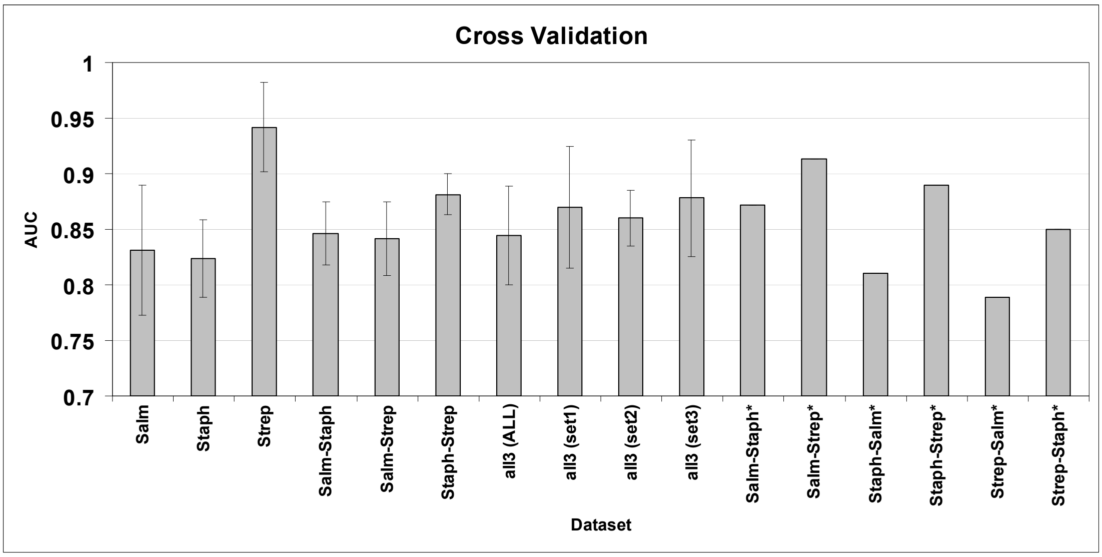


Figure 4.13: A Bar chart illustrating the average performance of the RVM classifier, under different training and test datasets. Each dataset is split into five subsets of approximately equal size; four of the five subsets are used to train an RVM model while the omitted subset is used to test the performance 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).

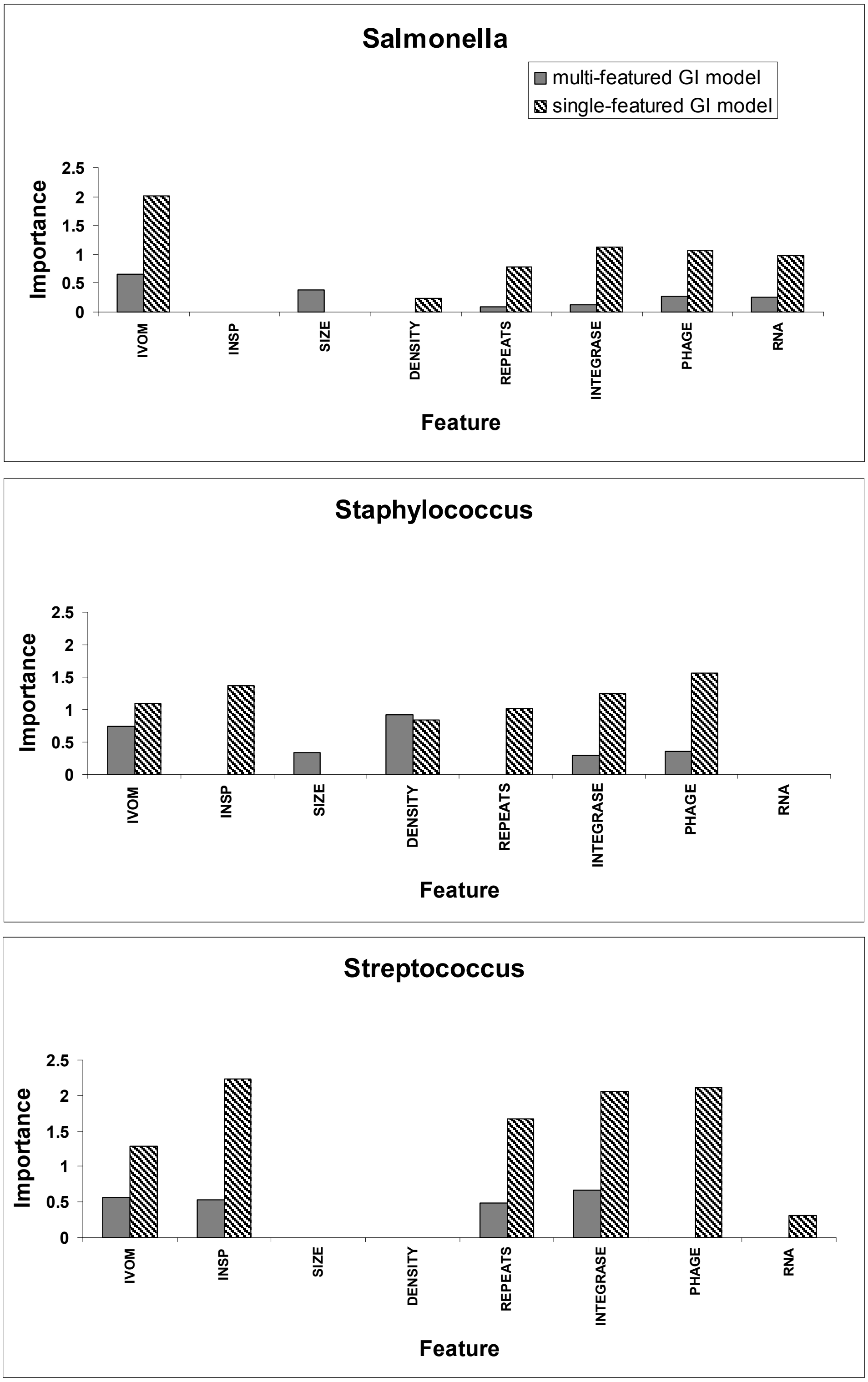


Figure 4.14: Bar chart illustrating the "importance" of eight structural features under a *Salmonella*, *Staphylococcus* and *Streptococcus* GI model. Grey-coloured bars show the "importance" of every feature, in a (multi-featured) GI model in which all eight features are taken into account (relative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

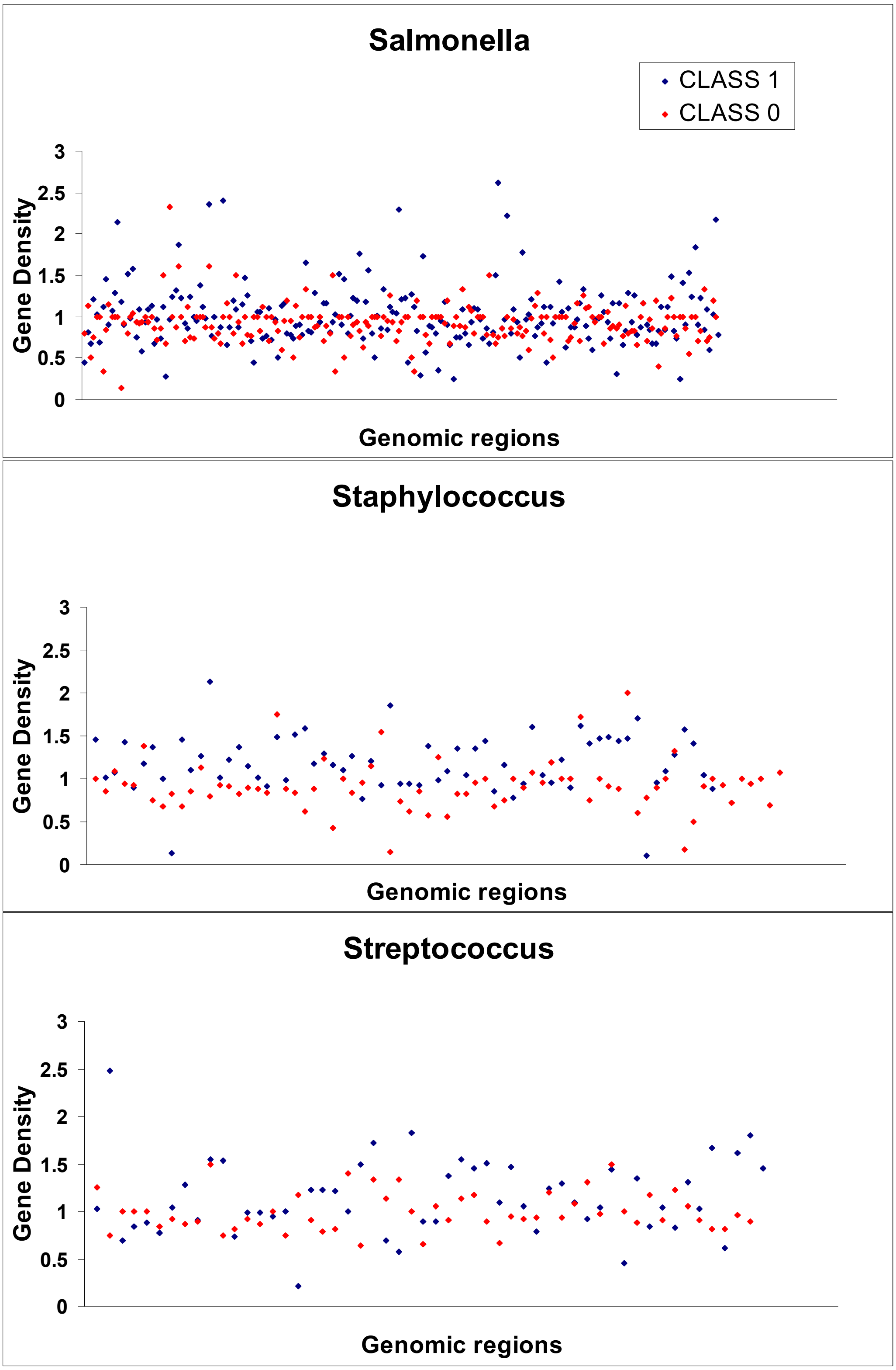


Figure 4.15: Gene Density of class "1" (GIs) and class "0" (randomly sampled) regions in *Salmonella* (top, *p*-value = 1.7 x 10^{-3}), *Staphylococcus* (middle, *p*-value = 1.4 x 10^{-6}) and *Streptococcus* (bottom, *p*-value = 1.3 x 10^{-2}) genera. The *p*-value has been calculated using a two-tailed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

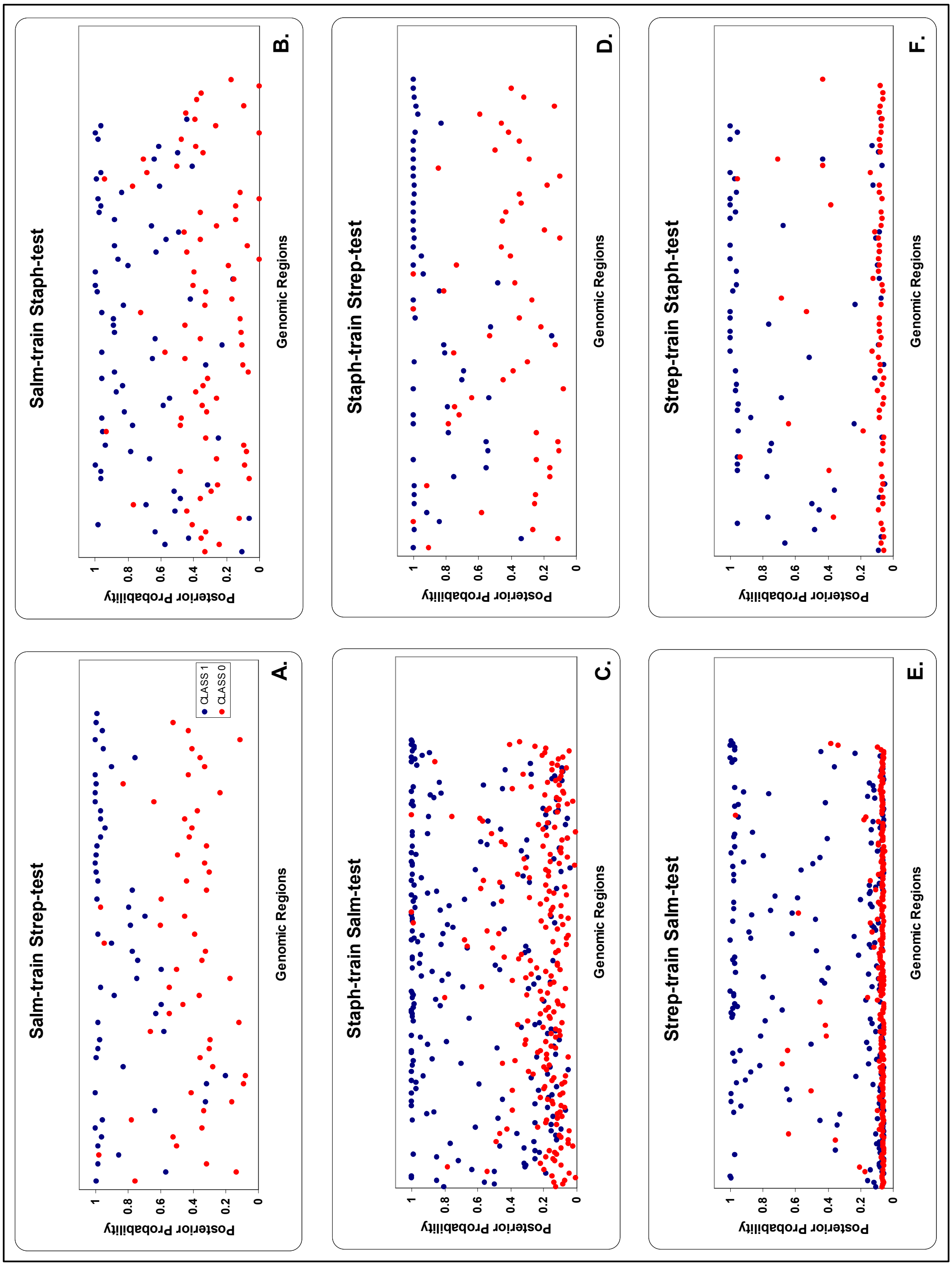


Figure 4.16: Scatter plot showing the posterior probability of a given region of being a true GI, given the model. Each genus specific dataset (e.g. Salm) is used to train an RVM model (e.g. Salm-train) that is then tested on the dataset of one of the other two genera (e.g. Strep-test). Each point in 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.

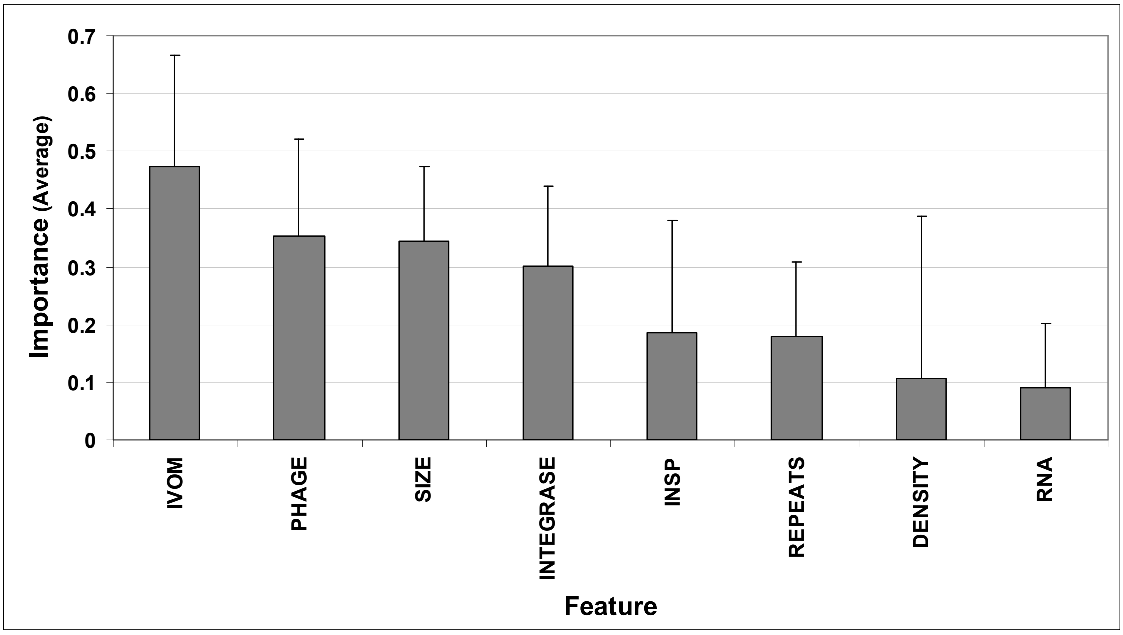


Figure 4.17: Bar chart illustrating the average “importance”, across 11 structural GI models, of the eight structural features evaluated in this analysis. The eight features have been sorted (in decreasing order) based on their average “importance”. Error bars show 1 SD.

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.