Supplementary material for:

Identification and analysis of integrons and cassette arrays in bacterial genomes.

Jean Cury[1,2], Thomas Jové[3], Marie Touchon[1,2], Bertrand Néron[4], Eduardo PC Rocha[1,2]

[1] Microbial Evolutionary Genomics, Institut Pasteur, 28, rue Dr Roux, Paris, 75015, France

[2] CNRS, UMR3525, 28, rue Dr Roux, Paris, 75015, France

[3] Univ. Limoges, INSERM, CHU Limoges, UMR_S 1092, F-87000 Limoges, France.

[4] Centre d'Informatique pour la Biologie, C3BI, Institut Pasteur, Paris, France

Table of contents

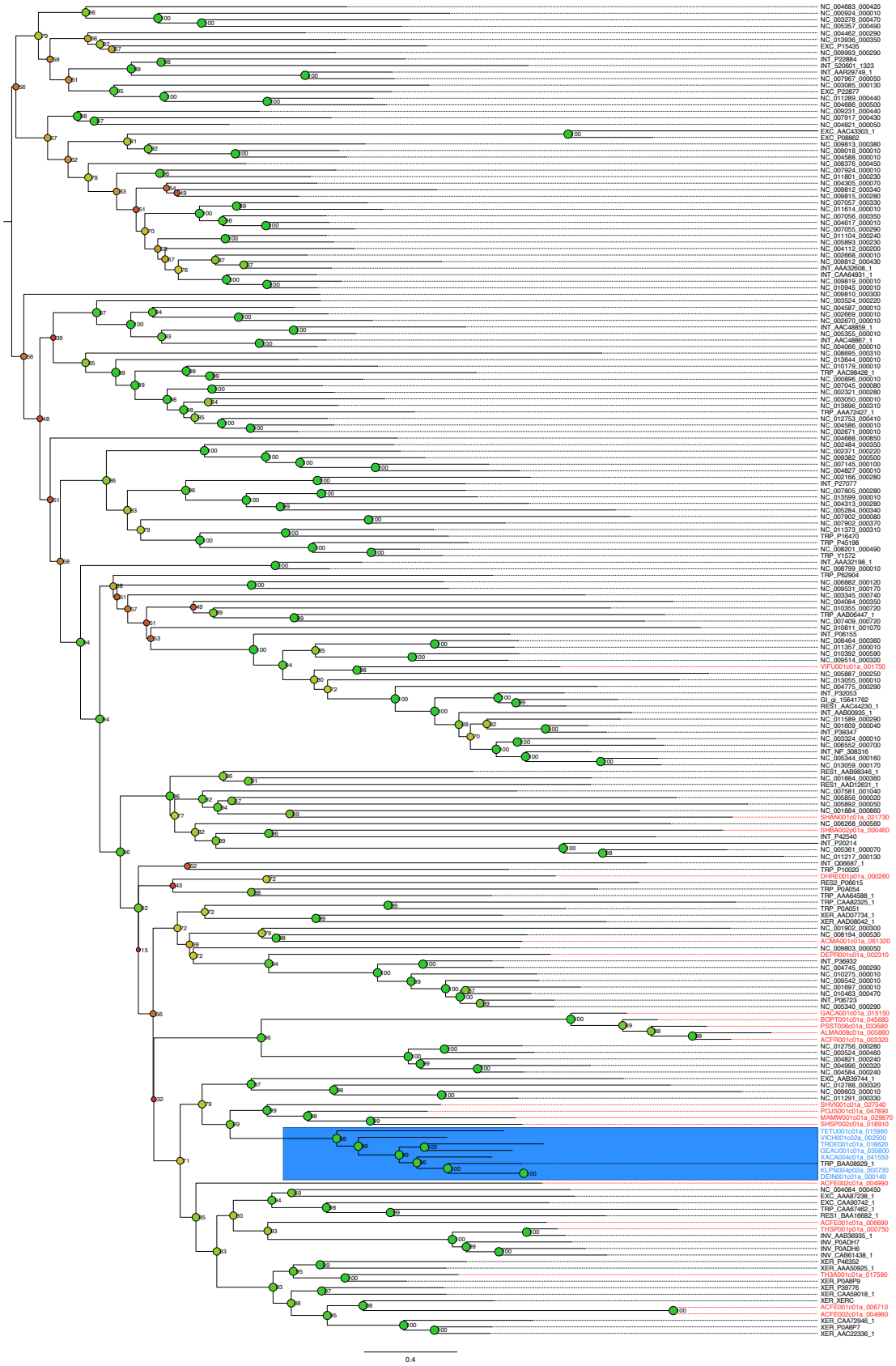

Figure S1 – Phylogenetic tree of tyrosine recombinases including the 21 proteins, which match the profile PF00589 but not intI_Cterm (red) and IntI from complete integron (blue).

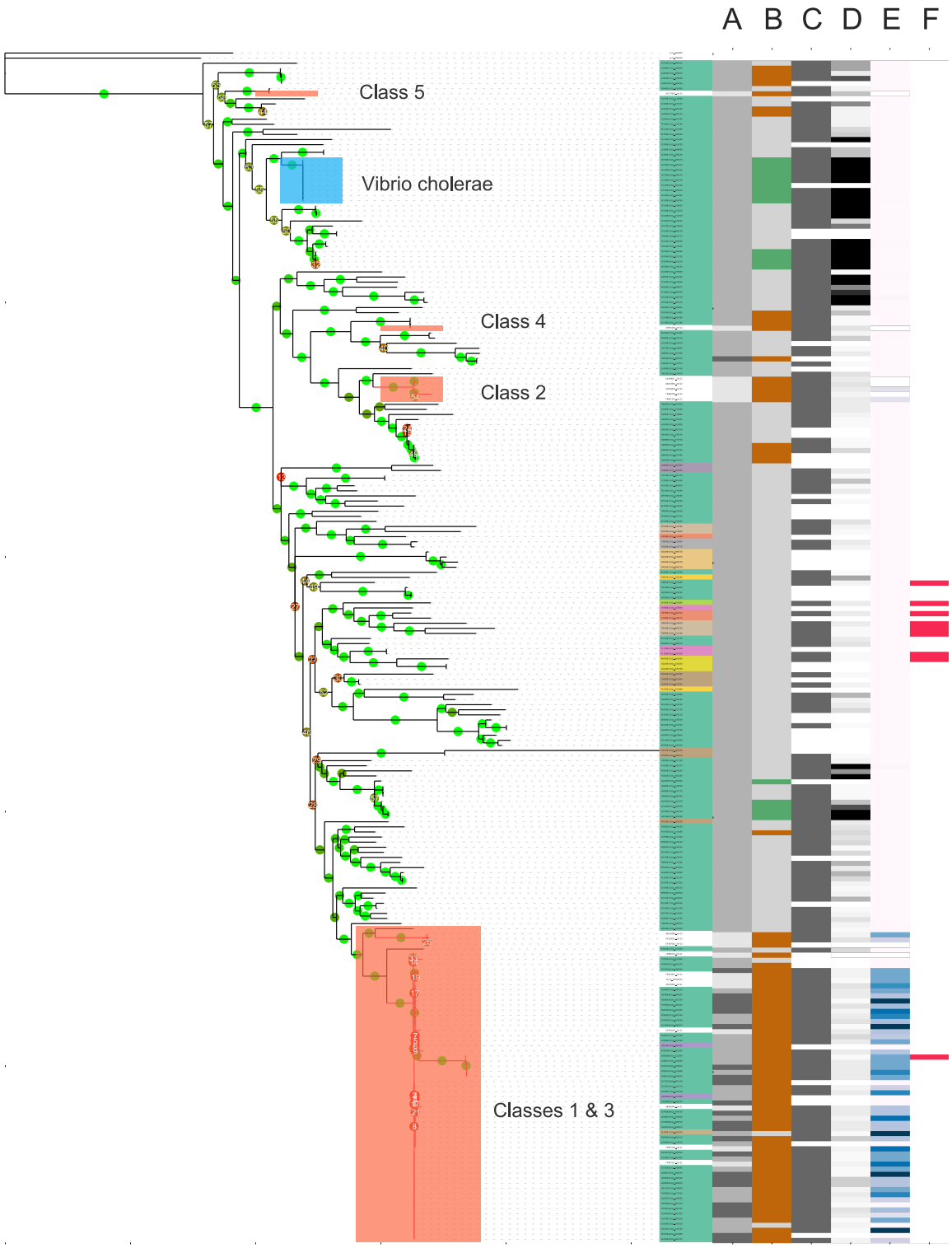

Figure S2 - Phylogenetic tree of IntI. The tree was made using IQ-Tree on the multiple alignments of the integrases carrying the two profiles (intI_Cterm and PF00589) using XerC and XerD as outgroups (see Methods for details). Bootstrap values are indicated if lower than 60%. The matrix on the right represents traits associated with the integron. Column (A) represents whether the corresponding

integron is on a plasmid (dark grey) or on a chromosome (grey). Lighter grey corresponds to integrons of classes 1 to 5, but whose replicon types are unknown (unavailable genomes). The column (B) represents whether the corresponding integron is a sedentary chromosomal integron (green) or a mobile integron (dark orange), or other (not determinable) (grey). Column (C) represents the integrases associated with at least one attC site (grey). The column (D) represents the number of *attC* sites with a gradient of grey (the darker the more *attC* sites, with saturation from 20 *attC* sites). The column (E) represents the frequency of resistance genes among cassettes (darker blue indicates higher values). The last column (F) represents the integron with inverted integrase (red). The leaves of the tree are colored according to their clade. For the names corresponding to the colors, see Figure 5 and Figure S6. See Table S3 for full data.

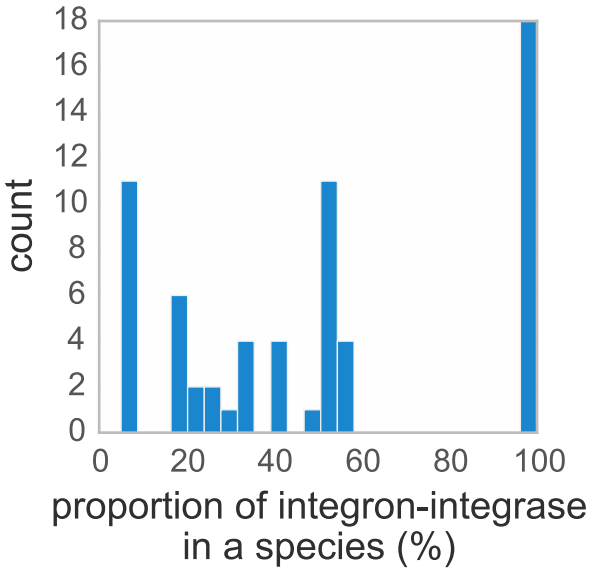

Figure S3 – Histogram of the proportion of genomes in a species carrying orthologs for a given integron-integrase**.** Only species with more than 4 genomes were used for the analysis.

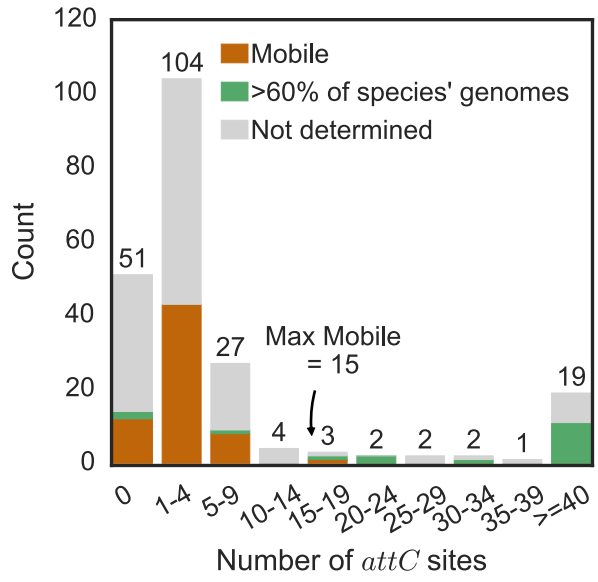

Figure S4. Histogram of the distribution of the number of *attC* sites per integron. Mobile integrons are depicted in dark orange; sedentary chromosomal integrons (present in more than 60% of the genomes of a species, actually 100%) are depicted in green; undetermined elements are depicted in grey. The largest mobile integron has 15 *attC* sites.

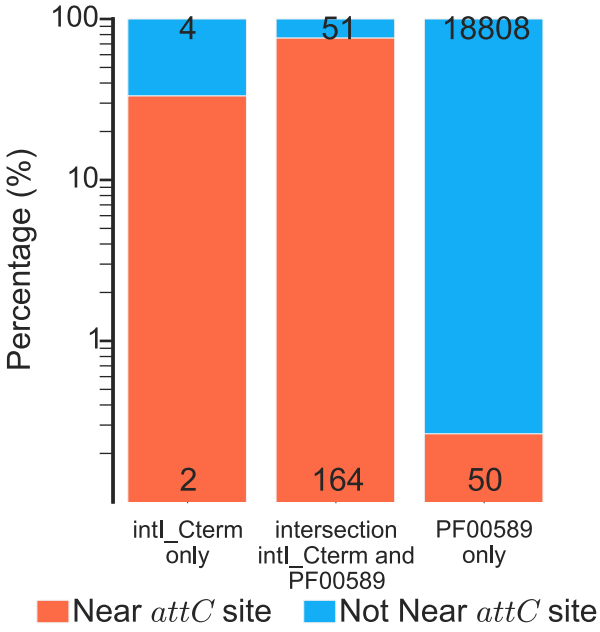

Figure S5 - Distribution of hits of PF00589 and intI_Cterm in function of each other and the existence of neighboring *attC* sites. The y-axis represents percentage in a log10 scale. Numbers on the bar represent the actual quantity for the underlying bar.

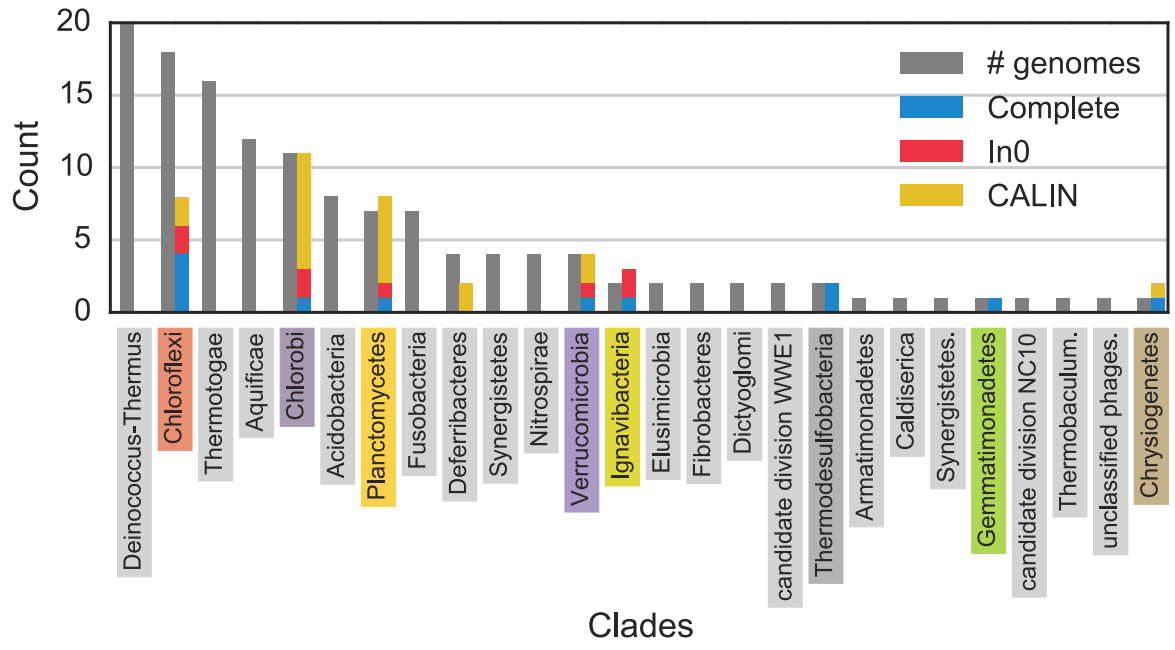

Figure S6 – Taxonomic distribution of integrons in clades with less than 50 genomes available in our dataset. The grey bar represents the number of genome sequenced for a given clade. The blue bar represents the number of complete integron, the red bar number of In0 and the yellow bar the number of CALIN.

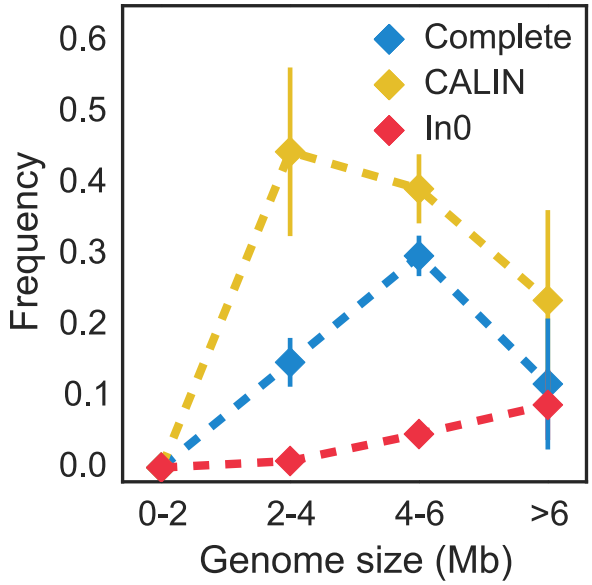

Figure S7 - Frequency of integrons and related elements as a function of the genome size when the analysis is restricted to Gammaproteobacteria. Vertical bar

represents standard error of the mean. The sample sizes of each bin are: 68 [0-2], 108 [2-4], 339 [4-6], and 34 [>6].

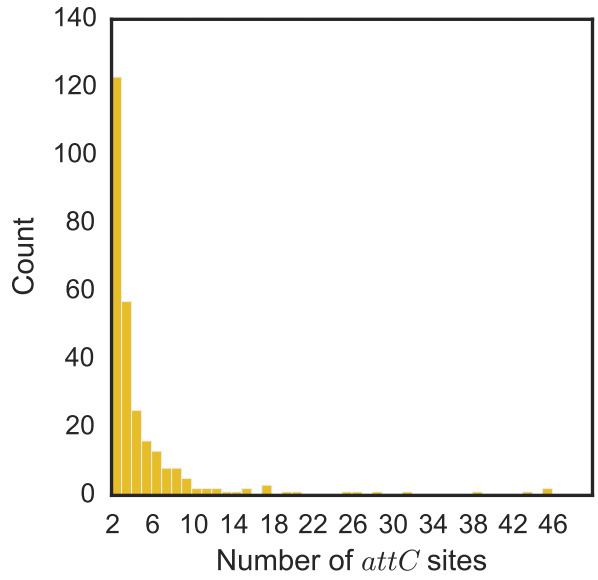

Figure S8 - Histogram of the distribution of the number of *attC* sites in CALIN elements. One point was a clear outlier and is not represented on the figure for sake of clarity (its value was 114).

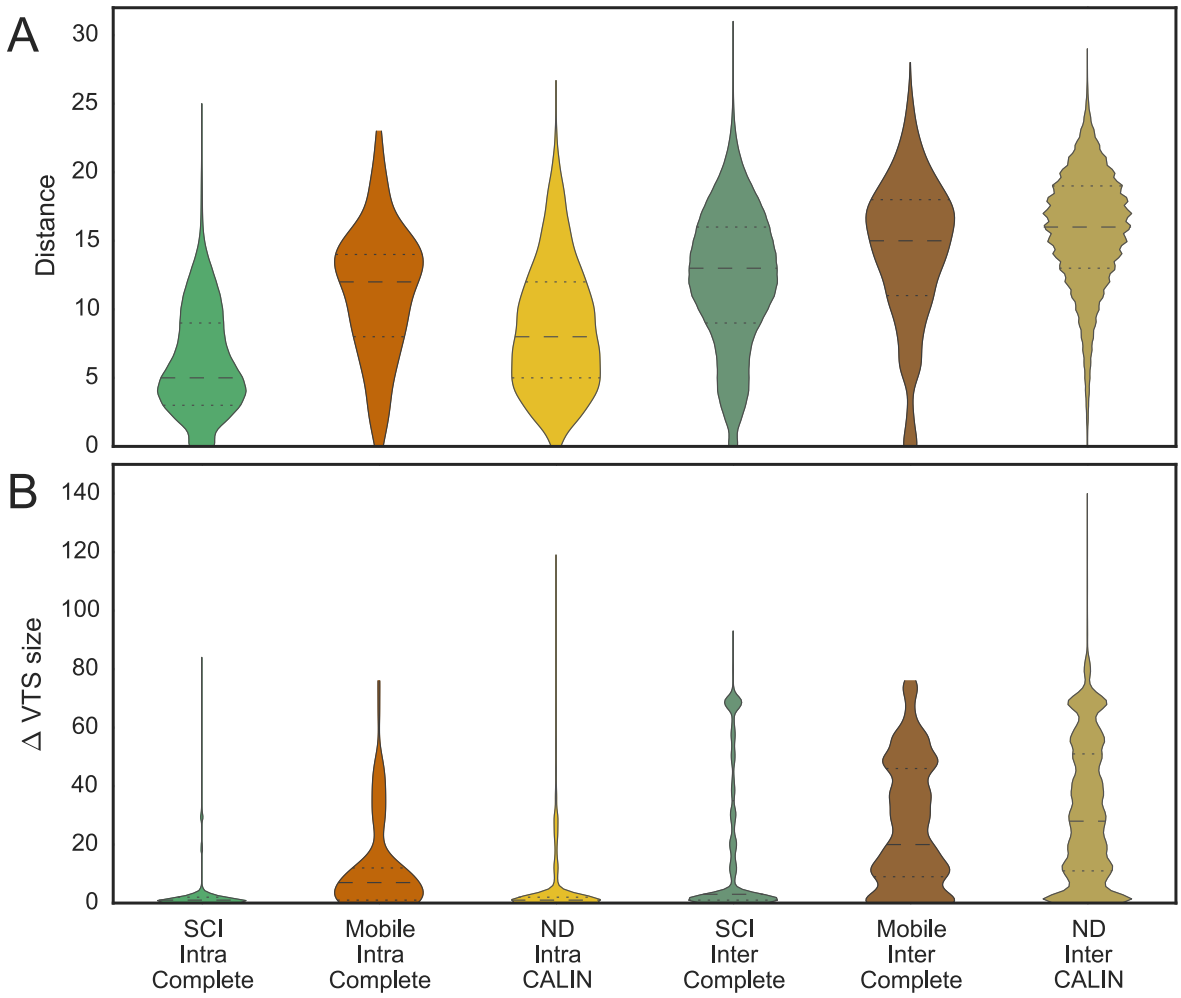

Figure S9 Comparison of *attC* sites. A column represents comparisons of *attC* sites between (inter) or within (intra) element(s) (Complete integron or CALIN) depending on the type of element (mobile or sedentary chromosomal integron (SCI)). (A) Distribution of the sequence distance between two R-UCS-L boxes of two *attC* sites. (B) Distribution of the difference in VTS size. All distributions are significantly different from each other in any given panel (Mann-Whitney rank test, all p-values < 10^{-3})

Table S3 – Sequence of the promoter of the integrase (Pint), the promoter of the cassette (Pc), and of the attI site in 3 classes of integron when available. Sequences were provided by INTEGRALL. Square brackets indicate that it can be one of the letters at that position. Curly brackets indicate that the square brackets before is repeated a number of times comprised between a minimum and a maximum.

Class	P_{int}	P_c	*attI*
1	TTGCTGCTTGGATGCCCGAGG CATAGACTGTACA	T[GT]G[AG][CT]ATAAGCCTGTT CGGTT[CG]GT[AG]A[AG]CTGTA ATCGCA	TGATGTTATGGAGCAGCAACG ATGTTACGCAGCAGGGCAGTC GCCCTAAAACAAAGTT
		TTGTTATGACTGTTTTTTT[G-]{1,4}[GT]ACA[GCA][AT]	
2	ND	ND	TTAATTAACGGTAAGCATCAGC GGGTGACAAAACGAGCATGCT TACTAATAAAATGTT
3	ND	TAGACATAAGCTTTCTCGGTCT GTAGG[CA]TGTAATG	CTTTGTTTAACGACCACGGTTG TGGGTATCCGGTGTTTGGTCA GATAAACCACAAGTT

Table S5 – List of genome used to create the 6 pseudo-genomes with different GC% background composition.

Genome	Size (pb)	%GC
Mycoplasma hyorhinis GDL-1	837480	25.91
Anaerococcus prevotii DSM 20548	1883067	36.07
Bacilllus subtilis subsp. subtilis str. 168	4215606	43.51
Escherichia coli str. K-12 substr. MG1655	4639675	50.79
Arthrobacter aurescens TC1	4597686	62.34
Clavibacter michiganensis subsp. michiganensis NCPPB 382	3297891	72.66