J Mol Evol (2007) 65:44–67
DOI: 10.1007/s00239-006-0235-3

JOURNAL OF MOLECULAR EVOLUTION

Intragenomic Variation and Evolution of the Internal Transcribed Spacer of the rRNA Operon in Bacteria

Frank J. Stewart, Colleen M. Cavanaugh

Department of Organismic and Evolutionary Biology, Harvard University, The Biological Laboratories, 16 Divinity Avenue, Cambridge, MA 02138, USA

Received: 20 October 2006 / Accepted: 13 March 2007 [*Reviewing Editor:* Dr. Margaret Riley]

Abstract. Variation in the internal transcribed spacer (ITS) of the rRNA (*rrn*) operon is increasingly used to infer population-level diversity in bacterial communities. However, intragenomic ITS variation may skew diversity estimates that do not correct for multiple *rrn* operons within a genome. This study characterizes variation in ITS length, tRNA composition, and intragenomic nucleotide divergence across 155 Bacteria genomes. On average, these genomes encode 4.8 *rrn* operons (range: 2–15) and contain 2.4 unique ITS length variants (range: 1–12) and 2.8 unique sequence variants (range: 1–12). ITS variation stems primarily from differences in tRNA gene composition, with ITS regions containing tRNA-Ala + tRNA-Ile (48% of sequences), tRNA-Ala or tRNA-Ile (10%), tRNA-Glu (11%), other tRNAs (3%), or no tRNA genes (27%). Intragenomic divergence among paralogous ITS sequences grouped by tRNA composition ranges from 0% to 12.11% (mean: 0.94%). Low divergence values indicate extensive homogenization among ITS copies. In 78% of alignments, divergence is $<1\%$, with 54% showing zero variation and 81% containing at least two identical sequences. ITS homogenization occurs over relatively long sequence tracts, frequently spanning the entire ITS, and is largely independent of the distance (basepairs) between operons. This study underscores the potential contribution of interoperon ITS variation to bacterial microdiversity studies, as well as unequivocally demonstrates the pervasiveness of concerted evolution in the *rrn* gene family.

Correspondence to: Colleen M. Cavanaugh; *email:* cavanaug@fas.harvard.edu

Key words: Bacterial diversity — Intergenic spacer — Concerted evolution — Gene conversion — Ribosomal RNA (*rrn*) operon — Copy number — Multigene family

Introduction

The internal transcribed spacer (ITS) region separating the bacterial 16S and 23S rRNA genes is increasingly used to assess microheterogeneity in the growing field of bacterial population genetics (e.g., Di Meo et al. 2000; Boyer et al. 2002; Hurtado et al. 2003; Vogel et al. 2003; Brown and Fuhrman 2005; DeChaine et al. 2006). Relative to molecular markers used in bacterial phylogenetics (e.g., 16S rRNA, *rpoB*), the ITS region experiences low selective constraint, evolves rapidly, and provides a high-resolution estimate of gene flow and genetic structuring at the population scale (Gürtler and Stanisich 1996; Antón et al. 1998; Schloter et al. 2000; Rocap et al. 2002, 2003; Brown and Fuhrman 2005). However, use of the ITS marker for studies of environmental samples is problematic due to the common occurrence of multiple ribosomal RNA (*rrn*) operons within a genome and to the possibility of intragenomic variation in ITS sequence and length. Given that strain-level genetic diversity, as measured by ITS sequence and length variation, has been conclusively linked to diversity in bacterial ecotypes (e.g., Rocap et al. 2002, 2003; Jaspers and Overmann 2004; Hahn and Pöckl 2005), intragenomic ITS variation may

lead to an overestimation of the number of functionally distinct bacteria in environmental samples. In particular, intragenomic heterogeneity in ITS length has the potential to skew estimates of diversity that are based solely on DNA fingerprinting techniques (e.g., ARISA; Fisher and Triplett 1999; Ranjard et al. 2000; Crosby and Criddle 2003). Accurately assessing bacterial microdiversity using ITS-based techniques therefore requires an understanding of how this region evolves in distinct bacterial taxa.

Several studies directed at specific bacterial taxa have provided important insight into ITS variation and evolution. These studies primarily describe ITS variation at the interstrain and interspecies levels, particularly with respect to medically relevant organisms (Chun et al. 1999; Osorio et al. 2005; González-Escalona et al. 2006), for which knowledge of ITS composition is valuable for differentiating and tracking pathogenic genetic variants. Several of these studies also directly or indirectly characterize variation among ITS regions within the same genome (Graham et al. 1997; Antón et al. 1998; Luz et al. 1998; Boyer et al. 2001; Gianninò et al. 2003; Milyutina et al. 2004). These show the potential both for substantial intragenomic heterogeneity in ITS sequence and length and for high levels of interoperon sequence conservation within certain regions of the ITS (Gürtler and Stanisich 1996; Antón et al. 1998; Nagpal et al. 1998; Rocap et al. 2003). This pattern has been attributed to homologous recombination that rearranges tRNA genes and other sequence blocks within the ITS region, often generating ITS regions characterized by a modular composition of alternating variable and conserved sequence blocks (Gürtler and Stanisich 1996; Antón et al. 1998; Lan and Reeves 1998; Liao 2000; Wenner et al. 2002; Osorio et al. 2005). Indeed, the nonreciprocal transfer of such sequence blocks between paralogous *rrn* operons (i.e., gene conversion) has been proposed as a mechanism that may ultimately homogenize parts or all of the *rrn* operon across all copies within a genome (Liao 2000). However, prior studies have focused on one or a few bacterial taxa or on variation in ITS length only. For many taxa, the extent to which recombination either homogenizes ITS regions or generates new combinations of sequence blocks within the ITS is unclear. In addition, the relative contribution of point mutations to intragenomic ITS divergence, though potentially large (Antón et al. 1998; Liao 2000) and of direct relevance to accurately interpreting population-level genetic processes, has been characterized for only a few bacteria (e.g., Antón et al. 1998; Boyer et al. 2001). A comprehensive analysis of ITS composition in all *rrn* operons in a genome and across diverse bacterial groups is needed to characterize the relative roles of recombination and point mutation in ITS evolution.

The recent wealth of whole-genome sequence data allows comprehensive assessment of ITS structure, variation, and evolution across distinct bacterial lineages. Using genomic data from 155 taxa representing all major bacterial groups, this study quantifies genetic variation in the ITS region among multiple *rrn* operons within a genome (paralogues). Our analyses identify clear differences in intragenomic ITS length, composition, and divergence among bacterial groups and underscore the surprising extent to which genetic content is homogenized across distinct *rrn* operons within a genome. These data greatly increase our understanding of the evolution of the ITS region in Bacteria and provide a framework for integrating this knowledge into studies that use the ITS for bacterial strain typing, microdiversity estimation, and population genetics.

Materials and Methods

Sequence Data

ITS and 16S rRNA gene sequences from 155 complete Bacteria genomes with multiple *rrn* operons were obtained from the National Center for Biotechnology Information (NCBI) Microbial Genome project in June 2006. Accession numbers for these genomes are listed as Supplemental Material. (Complete genomes containing only a single *rrn* operon [representing ~60 distinct species at the time of the analysis] were excluded from this study.) Genomes were chosen in order to span the phylogenetic diversity of Bacteria and include representatives from 12 of the major phylogenetic groups represented in the NCBI database (Table 1). However, given the current taxonomic coverage of the database, Bacteria from the *γ-Proteobacteria* and *Firmicutes* divisions are most abundant in our data set, constituting 35% and 20% of the genomes, respectively (Tables 1 and 2). In several instances multiple strains from the same species were included in the analysis to evaluate variation in ITS length and composition among closely related organisms. For each *rrn* operon within a genome, the base-pair coordinates corresponding to the 3' terminus of the 16S rRNA gene and the 5' start of the 23S rRNA gene were entered into the sequence retrieval function on the NCBI database and used to extract the intervening ITS sequence. Also, the 16S rRNA gene sequence corresponding to each ITS was downloaded from each genome's structural RNA table (in NCBI). Archaea, which commonly contain only one *rrn* operon (Acinas et al. 2004b) and which have been the focus of relatively few ITS-based diversity studies, were excluded from this analysis. Nonetheless, increasing use of the ITS for Archaea strain typing may warrant a systematic examination of ITS evolution in this domain.

Intragenomic Sequence Analysis

Each ITS sequence was categorized based on the presence of distinct tRNA genes as belonging to one of five classes, containing tRNA-alanine + tRNA-isoleucine (tRNA-Ala + Ile), tRNA-Ala or tRNA-Ile singly (tRNA-Ala/Ile), tRNA-glutamate (tRNA-Glu), other tRNAs (tRNA-other), or no tRNA genes (tRNA-none). The presence or absence of distinct tRNA genes within an ITS was determined based on each genome's annotation. For each genome, all ITS sequences belonging to the same

Table 1. Intragenomic sequence divergence among internal transcribed spacer (ITS) regions and 16S rRNA genes in 155 Bacteria with multiple ribosomal RNA (*rrn*) operons

Bacteria[a]	ITS tRNA class[b]	No. in class[c]	ITS length (bp)[d]	Variants per genome[e]	ITS % div[f]	16S % div[f]	No. of indels[g]	% var sites due to indels[h]
α-Proteobacteria								
Bartonella henselae str. Houston-1	Ala + Ile	2	1,255	1	0	0	0	0
Bartonella quintana str. Toulouse	Ala + Ile	2	1,205	1	0	0	0	0
Brucella abortus biovar 1 str. 9-941	Ala + Ile	3	789	1	0	0	0	0
Brucella melitensis 16M	Ala + Ile	3	821	1	0	0	0	0
Brucella melitensis biovar Abortus 2308	Ala + Ile	3	789, 790	1,2	0	0.05	1	100
Caulobacter crescentus CB15	Ala + Ile	2	706	1	0	0	0	0
Gluconobacter oxydans 621H	Ala + Ile	4	662	3	0.15	0.03	0	0
Magnetospirillum magneticum AMB-1	None	2	554	1	0	0	0	0
Novosphingobium aromaticivorans DSM 12444	Ala + Ile	3	671	1	0	0	0	0
Rhizobium etli CFN 42	Ala + Ile	3	1110, 1114	2	0.24	0	2	33
Rhodobacter sphaeroides 2.4.1	Ala + Ile	3	665	1	0	0.05	0	0
Rhodopseudomonas palustris CGA009	Ala + Ile	2	747	1	0	0	0	0
Rhodospirillum rubrum ATCC 11170	Ala + Ile	4	762, 765	2	1.27	0	2	10
Silicibacter pomeroyi DSS-3	Ala + Ile	3	890	1	0	0	0	0
Sinorhizobium meliloti 1021	Ala + Ile	3	1155	1	0	0	0	0
Zymomonas mobilis subsp. *mobilis* ZM4	Ala + Ile	3	606	1	0	0.05	0	0
β-Proteobacteria								
Azoarcus sp. EbN1	Ala + Ile	4	514	1	0	0	0	0
Bordetella bronchiseptica RB50	Ala + Ile	3	592	1	0	0	0	0
Bordetella parapertussis 12822	Ala + Ile	3	592	1	0	0	0	0
Bordetella pertussis Tohama I	Ala + Ile	3	584	1	0	0	0	0
Burkholderia pseudomallei 1710b	Ala + Ile	4	622, 661	4	0.22	0.03	0	0
Burkholderia pseudomallei K96243	Ala + Ile	4	633	1	0	0.03	0	0
Burkholderia sp. 383	Ala + Ile	5	543, 549, 550, 600	5	3.50	0.16	5	12
	None	1	393		NA	NA	NA	NA
Burkholderia thailandensis E264	Ala + Ile	4	593	2	0.09	0	0	0
*Chromobacterium violaceum** ATCC 12472	Ala + Ile	8*	507, 680, 681	3	0.26	0	1	33
Dechloromonas aromatica RCB	Ala + Ile	4	435	1	0	0	0	0
Methylobacillus flagellatus KT	Ala + Ile	2	684	1	0	0	0	0
Neisseria gonorrhoeae FA 1090	Ala + Ile	4	590, 595	3	0.17	0	5	71
Ralstonia eutropha JMP134	Ala + Ile	4	520, 524, 525	4,6	5.15	0.30	5	10
	None	2	327		0	0	1	100
Rhodoferax ferrireducens DSM 15236	Ala + Ile	2	624	1	0	0	0	0
Thiobacillus denitrificans ATCC 25259	Ala + Ile	2	802, 836	1	0	0	0	0
γ-Proteobacteria								
*Acinetobacter** sp. ADP1	Ala + Ile	7*	593, 594, 683, 684	2	1.06	0.02	0	0
Baumannia cicadellinicola str. Hc	Glu	2	270	1	0	0	0	0
*Chromohalobacter salexigens** DSM 3043	Ala + Ile	5*	573, 587, 675, 689, 707	5	1.10	0	0	0
*Colwellia psychrerythraea** 34H	Ala	7	609, 612, 629	6	1.40	0.04	4	13
	Ile	2*	627, 640		6.88	0.07	6	13

Table 1. Continued

Bacteria[a]	ITS tRNA class[b]	No. in class[c]	ITS length (bp)[d]	Variants per genome[e]	ITS % div[f]	16S % div[f]	No. of indels[g]	% var sites due to indels[h]
Erwinia carotovora subsp. *atroseptica* SCRI1043	Glu	4	445	5	0.45	0.10	0	0
	Ala + Ile	3	474, 485		2.79	0	3	14
Escherichia coli CFT073	Glu	5	354, 355, 431, 440, 449	5	5.50	0.54	5	12
	Ala + Ile	1	437		NA	NA	NA	NA
	Other (Glu + Ala)	1	432		NA	NA	NA	NA
Escherichia coli K12	Glu	4	354, 431, 440	7	5.90	0.39	4	10
	Ala + Ile	3	437, 446		1.69	0.87	1	9
Escherichia coli O157:H7	Glu	4	354	7	0.90	0.37	0	0
	Ala + Ile	3	446		0.60	0.09	0	0
Francisella tularensis subsp. tularensis Schu 4	Ala + Ile	3	335	1	0	0.18	0	0
Haemophilus ducreyi 35000HP	Ala + Ile	3	571	2	0	0	0	0
	Glu	3	351		0	0	0	0
Haemophilus influenzae 86-028NP	Ala + Ile	3	716, 717, 723	3	0.19	0	1	33
	Glu	3	477		0	0.22	0	0
Haemophilus influenzae Rd KW20	Ala + Ile	3	723	2	0	0	0	0
	Glu	3	478		0	0	0	0
Hahella chejuensis KCTC 2396	Ala + Ile	5	631	1	0	0	0	0
Idiomarina loihiensis L2TR	Ala + Ile	4	590, 734, 736	4	1.97	0.11	5	20
Legionella pneumophila str. Lens	Ala	2	345	2	0	0.07	0	0
	Ile	1	359		NA	NA	NA	NA
Legionella pneumophila str. Paris	Ala	2	345	2	0	0.07	0	0
	Ile	1	360		NA	NA	NA	NA
Legionella pneumophila str. Philadelphia 1	Ala	2	383	2	0	0	0	0
	Ile	1	397		NA	NA	NA	NA
Mannheimia succiniciproducens MBEL55E	Ala + Ile	3	545	3	0	0.04	0	0
	Glu	3	400, 402		0.33	0.09	2	50
Methylococcus capsulatus str. Bath	Ala + Ile	2	567	1	0	0	0	0
Nitrosococcus oceani ATCC 19707	Ala + Ile	1	714	2	NA	NA	NA	NA
	None	1	224		NA	NA	NA	NA
*Photobacterium profundum** SS9	None	5	338, 371, 372	12	2.69	0.14	5	19
	Other (Glu + Lys + Val)	4*	748, 762, 786		3.24	0.84	10	19
	Ala + Ile	3	576, 578, 696		3.67	1.05	9	23
	Other (Glu + Lys-Val + Ala)	2*	856, 927		7.55	1.64	16	22
	Other (Glu + Val + Ala)	1	801		NA	NA	NA	NA
Photorhabdus luminescens subsp. *laumondii* TTO1	Glu	4	345	6	0.49	0.26	0	0
	Ala + Ile	3	512, 516		7.57	0.48	3	5
Pseudoalteromonas haloplanktis TAC125	None	7	299, 300, 301, 302	6	1.32	0.16	4	31
	Ala + Ile	2	647		0	0.66	0	0
Pseudomonas entomophila L48	Ala + Ile	7	485, 512	4	2.59	0.02	6	15
Pseudomonas fluorescens PfO-1	Ala + Ile	6	509	2	1.32	0.04	3	21
Pseudomonas syringae pv. phaseolicola 1448A	Ala + Ile	5	552	1	0	0	0	0
Pseudomonas syringae pv. syringae B728a	Ala + Ile	5	533, 549	3	1.56	0.09	3	19

Table 1. Continued

Bacteria[a]	ITS tRNA class[b]	No. in class[c]	ITS length (bp)[d]	Variants per genome[e]	ITS % div[f]	16S % div[f]	No. of indels[g]	% var sites due to indels[h]
Pseudomonas syringae pv. tomato str. DC3000	Ala + Ile	5	543	1	0	0	0	0
Psychrobacter arcticus 273-4	Ala + Ile	4	568, 593	1	0	0	0	0
Psychrobacter cryohalolentis K5	Ala + Ile	4	623	1	0	0	0	0
Saccharophagus degradans 2-40	None[i]	2	679, 2880	2	0	0	0	0
Salmonella enterica sv. Choleraesuis str. SC-B67	Glu	4	352, 353	3,4	0	0.30	5	100
	Ala + Ile	3	511, 512		0.13	0.17	1	50
Salmonella enterica sv. Paratyphi A str. ATCC 9150	Glu	4	351, 354, 355	3	0.14	0.03	0	0
	Ala + Ile	3	509, 513		0	0	0	0
Salmonella typhimurium LT2	Glu	4	352, 353, 391	5	3.00	0.23	1	6
	Ala + Ile	3	511, 513, 515		5.40	0.31	2	5
Shewanella oneidensis MR-1	None	6	315, 316	4	1.26	0.12	1	10
	Ala + Ile	3	640		0	0.04	0	0
Shigella boydii Sb227	Ala + Ile	5	446	3	0	0	0	0
	Glu	2	354		1.43	0.59	0	0
Shigella dysenteriae Sd197	Glu	4	346	2	0	0	0	0
	Ala + Ile	3	437		0	0	0	0
Shigella flexneri 2a str. 2457T	Glu	4	354	3	0	0.21	0	0
	Ala + Ile	3	446		0.22	0.33	0	0
Shigella sonnei Ss046	Glu	5	354	4	0.28	0.20	0	0
	Ala + Ile	2	446		0	0.39	0	0
Sodalis glossinidius str. 'morsitans'	Ala + Ile	4	671	2	0	0	0	0
	Glu	3	492		0	0.04	0	0
Thiomicrospira crunogena XCL-2	Ala + Ile	3	833, 834	2	0.65	0	1	11
Vibrio cholerae O1 biovar eltor str. N16961	Ala + Ile	3	511, 512	7	0.53	0.61	2	33
	Glu	2	432		0	0.20	0	0
	Ala	1	426		NA	NA	NA	NA
	Other (Ala + Val + Lys)	1	713		NA	NA	NA	NA
	Other (Val + Lys + Glu)	1	687		NA	NA	NA	NA
Vibrio fischeri ES114	Ala + Ile	3	527, 532, 637	12	4.65	0.09	6	15
	Other (Val + Lys + Glu)	3	624, 792		7.13	0	15	20
	Glu	2	418, 492		4.21	0.85	4	19
	None	2	325, 360		5.79	0	3	14
	Ala	1	428		NA	NA	NA	NA
	Other (Val + Ala + Lys + Glu)	1	738		NA	NA	NA	NA
Vibrio parahaemolyticus RIMD 2210633	None	3	357	8	0	0.09	0	0
	Ala + Ile	2	603		0.83	0.14	0	0
	Other (Glu + Lys + Val)	2	748		0	0	0	0
	Glu	1	510		NA	NA	NA	NA
	Other (Ala + Glu)	1	606		NA	NA	NA	NA
	Other (Glu + Lys + Ala)	1	780		NA	NA	NA	NA
	Other (Val + Ala + Lys)	1	785		NA	NA	NA	NA

Table 1. Continued

Bacteria[a]	ITS tRNA class[b]	No. in class[c]	ITS length (bp)[d]	Variants per genome[e]	ITS % div[f]	16S % div[f]	No. of indels[g]	% var sites due to indels[h]
Vibrio vulnificus YJ016	Ala + Ile	2	505, 507	6	0.60	0.07	2	40
	Glu	2	420		0	0.33	0	0
	Other (Glu + Lys-Ala-Val)	2	741		0	0.07	0	0
	Other (Val + Lys + Glu)	2	601		0	0.07	0	0
	Other (Glu + Lys-Val)	1	664		NA	NA	NA	NA
Wigglesworthia glossinidia	Glu	2	270	1	0	0.13	0	0
Xanthomonas campestris pv. campestris str. 8004	Ala + Ile	2	494	1	0	0	0	0
Xanthomonas campestris pv. vesicatoria str. 85-10	Ala + Ile	2	472	1	0	0	0	0
Xanthomonas oryzae pv. oryzae KACC10331	Ala + Ile	2	492	1	0	0	0	0
Xanthomonas oryzae pv. oryzae MAFF 311018	Ala + Ile	2	492	1	0	0.07	0	0
Xylella fastidiosa 9a5c	Ala + Ile	2	465	1	0	0	0	0
Yersinia pestis biovar Medievalis str. 91001	Ala + Ile	4	551	4	0.18	0.03	0	0
	Glu	3	491		0	0.05	0	0
Yersinia pestis CO92	Ala + Ile	3	551	5	0.24	0.09	0	0
	Glu	3	491, 508		0.14	0.09	0	0
Yersinia pseudotuberculosis IP 32953	Glu	4	508	5	0.30	0.24	0	0
	Ala + Ile	3	568		0.24	0.60	0	0
δ-Proteobacteria								
Anaeromyxobacter dehalogenans 2CP-C	Ala + Ile	2	565	1	0	0	0	0
Desulfotalea psychrophila LSv54	Ala + Ile	2	671, 688	5	3.23	1.22	4	16
	None	5	442, 467		2.18	0.04	2	13
Desulfovibrio desulfuricans G20	Ala + Ile	4	442, 460, 461, 472	4	0.85	0.13	0	0
Desulfovibrio vulgaris str. Hildenborough	Ala + Ile	5	415	2	0.19	0.13	0	0
Geobacter metallireducens GS-15	Ala + Ile	2	429	1	0	0	0	0
Geobacter sulfurreducens PCA	Ala + Ile	2	453	1	0	0	0	0
Myxococcus xanthus DK 1622	Ala + Ile	2	704	3	0	0	0	0
	Ile	2	609		0.99	0	0	0
Pelobacter carbinolicus DSM 2380	Ala + Ile	2	475	1	0	0.14	0	0
ε-Proteobacteria								
Campylobacter jejuni RM1221	Ala + Ile	3	905	1	0	0	0	0
Campylobacter jejuni subsp. jejuni NCTC 11168	Ala + Ile	3	806	1	0	0	0	0
Thiomicrospira denitrificans ATCC 33889	Ala	3	546	2	0	0	0	0
	Ile	1	531		NA	NA	NA	NA
Wolinella succinogenes DSM 1740	Ala + Ile	3	559	1	0	0	0	0
Cyanobacteria								
Anabaena variabilis ATCC 29413	Ala + Ile	2	501, 503	3	0	0	0	0
	None[j]	1	503		NA	NA	NA	NA
	None	1	285		NA	NA	NA	NA
Nostoc sp. PCC 7120	Ala + Ile	3	513	2	0	0	0	0
	None	1	286		NA	NA	NA	NA
Prochlorococcus marinus str. MIT 9313	Ala + Ile	2	829	1	0	0	0	0

Table 1. Continued

Bacteria[a]	ITS tRNA class[b]	No. in class[c]	ITS length (bp)[d]	Variants per genome[e]	ITS % div[f]	16S % div[f]	No. of indels[g]	% var sites due to indels[h]
Synechococcus elongatus PCC 6301	Ala + Ile	2	545	1	0	0	0	0
Synechococcus sp. CC9605	Ala + Ile	2	795	1	0	0	0	0
Synechococcus sp. CC9902	Ala + Ile	2	776, 777	1	0	0	0	0
Synechocystis sp. PCC 6803	Ile	2	465	1	0	0	0	0
***Firmicutes* (low G + C gram positive)**								
Aster yellows witches'-broom phytoplasma AYWB	Ile	2	245	2	2.07	0.13	0	0
Bacillus anthracis str. Ames	None	9	171, 178	3,4	0.89	0.16	1	20
	Ala + Ile	2	405		0	0.13	0	0
Bacillus cereus ATCC 10987	None	10	136, 175, 176	4,5	0.24	0.04	1	33
	Ala + Ile	2	403		0	0	0	0
Bacillus cereus ATCC 14579	None	11	181, 182	5,6	0.38	0.07	1	25
	Ala + Ile	2	408		0	0.07	0	0
Bacillus licheniformis ATCC 14580	None	5	169, 172	2,3	0	0.16	3	100
	Ala + Ile	2	334		0	0.13	0	0
Bacillus subtilis subsp. *subtilis* str. 168	None	8	165, 167, 168	4,5	0.88	0.38	1	17
	Ala + Ile	2	345		0	0.06	0	0
Carboxydothermus hydrogenoformans Z-2901	None	3	276, 352, 355	4	12.11	0.52	8	16
	Ala + Ile	1	340		NA	NA	NA	NA
Clostridium acetobutylicum ATCC 824	None	11	177, 178	1,2	0	0.06	1	100
Clostridium perfringens str. 13	None	6	188	7	0.82	0.24	0	0
	Ala + Ile	4	420, 422		0.44	0.33	2	40
Clostridium tetani E88	None	3	264	3	0	0.09	0	0
	Ile	2	348		0	0.13	0	0
	Ala + Ile	1	547		NA	NA	NA	NA
Desulfitobacterium hafniense Y51	Ala	2	599	5	0	0.06	0	0
	None	2	524, 535		5.80	3.54	5	15
	Ala + Ile	1	783		NA	NA	NA	NA
	Other (Phe)	1	588		NA	NA	NA	NA
*Geobacillus kaustophilus** HTA426	None	7*	340, 386, 399, 540, 665, 676, 735	9	10.11	0.38	2	6
	Ala + Ile	2*	632, 794		5.94	0.32	5	17
Lactobacillus acidophilus NCFM	None	3	130, 132	3	0.52	0.56	0	0
	Ala + Ile	1	377		NA	NA	NA	NA
Lactobacillus delbrueckii subsp. *bulgaricus* ATCC 11842	None	5	226	6	0.27	0.26	0	0
	Ala + Ile	4	465, 466		0.47	0.31	1	20
Lactobacillus sakei subsp. *sakei* 23K	None	6	223	3	0.15	0.13	0	0
	Ala + Ile	1	433		NA	NA	NA	NA
Lactobacillus salivarius subsp. *salivarius* UCC118	None	4	13	3	0	0.03	0	0
	Ala + Ile	3	408		0.49	0.22	0	0
Lactococcus lactis subsp. *lactis* Il1403	Ala	6	304	1	0	0.02	0	0
Mesoplasma florum L1	None	2	207	1	0	0	0	0
Mycoplasma capricolum subsp. *capricolum* ATCC 27343	None	2	229, 231	2	2.69	0.20	1	14

Table 1. Continued

Bacteria[a]	ITS tRNA class[b]	No. in class[c]	ITS length (bp)[d]	Variants per genome[e]	ITS % div[f]	16S % div[f]	No. of indels[g]	% var sites due to indels[h]
Mycoplasma synoviae 53	None	2	258, 259	2	1.18	0.13	1	25
Oceanobacillus iheyensis HTE831	None	6	233, 271, 332[k]	4	1.46	0.15	3	27
	Ala + Ile	1	533		NA	NA	NA	NA
*Staphylococcus aureus** subsp. *aureus* MRSA252	None	3*	364, 423, 439	5	1.36	0.30	1	14
	Ala + Ile	1	473		NA	NA	NA	NA
	Ile	1	458		NA	NA	NA	NA
Staphylococcus epidermidis ATCC 12228	None	3	261, 263	5	0.26	0.47	2	67
	Ala + Ile	1	450		NA	NA	NA	NA
	Ile	1	355		NA	NA	NA	NA
*Staphylococcus haemolyticus** JCSC1435	None	3*	336, 337, 438	5	2.04	0.13	1	10
	Ala + Ile	1	451		NA	NA	NA	NA
	Ile	1	358		NA	NA	NA	NA
Staphylococcus saprophyticus subsp. *saprophyticus*	None	4	282, 283	3	0	0.11	0	0
	Ala + Ile	1	465		NA	NA	NA	NA
	Ile	1	362		NA	NA	NA	NA
Streptococcus agalactiae A909	Ala	7	316, 318	1,2	0	0.03	1	100
Streptococcus mutans UA159	Ala	5	388	1	0	0.12	0	0
Streptococcus pyogenes MGAS5005	Ala	6	420, 422	1	0	0	0	0
Streptococcus pyogenes MGAS6180	None	4	124	3	0	0	0	0
	Ala	2	425, 427		0.71	0	1	25
Streptococcus pyogenes MGAS8232	Ala	6	532	1	0	0	0	0
Thermoanaerobacter tengcongensis MB4	None	3	147, 152	4	10.82	4.35	4	16
	Ala + Ile	1	343		NA	NA	NA	NA
Actinobacteria								
Bifidobacterium longum NCC2705	None	4	422, 428, 429, 430	2,3	0.16	0	1	50
Corynebacterium diphtheriae NCTC 13129	None	5	459	1	0	0.03	0	0
Corynebacterium glutamicum ATCC 13032	None	6	388, 390	4	2.18	0.15	3	16
Corynebacterium jeikeium K411	None	3	407, 412	2	0.83	0	3	38
Frankia sp. CcI3	None	2	412	1	0	0	0	0
Propionibacterium acnes KPA171202	None	3	384, 393	2	0.35	0.04	1	25
Streptomyces avermitilis MA-4680	None	6	306, 310	2,3	0.37	0	1	33
*Streptomyces coelicolor** A3(2)	None	6*	196, 276, 280	1,3	0	0.07	0	0
Thermobifida fusca YX	None	4	425, 522, 549	1,4	0	0.07	1	100
Bacteroides/Chlorobium								
Bacteroides fragilis NCTC 9343	Ala + Ile	6	477, 483	2	0.86	0	5	29
*Bacteroides fragilis** YCH46	Ala + Ile	6*	480, 482, 487, 556, 560	5	2.16	0.16	6	21
Bacteroides thetaiotaomicron VPI-5482	Ala + Ile	5	589	1	0	0.57	0	0
Chlorobium tepidum TLS	Ala + Ile	2	525	1	0	0	0	0
Pelodictyon luteolum DSM 273	Ala + Ile	2	576	1	0	0	0	0
Porphyromonas gingivalis W83	Ala + Ile	4	834	2	0.06	0	0	0
Chlamydiae								
Candidatus Protochlamydia amoebophila UWE25	None	2	340	2	0.59	0	0	0
Chlamydia trachomatis A/HAR-13	None	2	315	1	0	0	0	0

Table 1. Continued

Bacteria[a]	ITS tRNA class[b]	No. in class[c]	ITS length (bp)[d]	Variants per genome[e]	ITS % div[f]	16S % div[f]	No. of indels[g]	% var sites due to indels[h]
Spirochetes								
Treponema denticola ATCC 35405	Ile	1	384	2	NA	NA	NA	NA
	Ala	1	401		NA	NA	NA	NA
Treponema pallidum subsp. *pallidum* str. Nichols	Ala	1	293	2	NA	NA	NA	NA
	Ile	1	293		NA	NA	NA	NA
Aquificales								
Aquifex aeolicus VF5	Ala + Ile	2	314	1	0	0	0	0

[a]Sequence data from complete genomes in the NCBI Microbial Genome database; taxonomic groupings reflect NCBI designations.
[b]ITS class defined by presence or absence of distinct tRNA genes: ITS with tRNA-alanine + tRNA-isoleucine (Ala + Ile), tRNA-Ala or tRNA-Ile only (Ala or Ile), tRNA-glutamate (Glu), other tRNAs (other), or no tRNA genes (none).
[c]Number of ITS sequences/*rrn* operons within each ITS class.
[d]Multiple numbers indicate length variants within an ITS class within a genome.
[e]Number of unique (nonidentical) ITS sequence variants per genome; sequences are unique if they differ by as few as 1 bp (within alignable region) from paralogues within the same ITS class; when two numbers are given, the second number indicates the count when sequence variants generated by internal insertion-deletion events (indels; ≥1 bp) are included; single values reflect instances in which including internal indels does not increase the number of variants.
[f]% div equals average pairwise nucleotide divergence as a percentage of the alignable sequence length; divergence values were corrected for multiple hits using a Kimura two-parameter model of nucleotide substitution, with the transition/transversion ratio set to 2.0; indels were excluded from calculations.
[g]Number of insertion-deletion events (gaps ≥1 bp) within an ITS alignment; overlapping indels are counted as a single event.
[h]Percentage of variable sites (number of segregating sites due to base substitution + number of indels) due to indels.
[i]ITS of operon 1 is 2880 bp long and contains two protein-coding genes but no tRNA genes; the terminal 270 bp align perfectly with the terminal 270 bp of the second ITS.
[j]Does not align with other ITS regions in the genome.
[k]ITS of operon 7 (332 bp) cannot be aligned with others and is excluded from diversity estimates.
*ITS alignments contained nonhomologous regions, which were excluded from pairwise diversity (% div) estimates and indel counts.

Table 2. Summary data showing variation in internal transcribed spacer (ITS) length, intragenomic nucleotide divergence, and class among 155 genomes from major taxonomic groups of Bacteria

	Genomes (n)	Genome size (mean Mbp)	*rrn* copies[a] (mean)	ITS seqs. (n)	ITS length (mean bp)	ITS length range (bp)	ITS seq. Variants per genome[b] (mean)	Nucleotide divergence (mean)				Alignments with zero div[e]	ITS classes per genome[f] (mean)	ITS seqs. of class tRNA:[g]				
								ITS		16S								
								(% div)[c]	(% S)[d]	(% div)[c]	(% S)[d]			Ala + Ile	Ala/Ile	Glu	Other	None
Protobacteria																		
α	16	3.55	2.8	45	825	554–1255	1.25, 1.31	0.10	0.20	0.01	0.02	13/16	1.0	43	0	0	0	2
β	15	5.13	3.9	59	580	327–836	2.00, 2.13	0.59	1.09	0.03	0.06	10/16	1.1	56	0	0	0	3
γ	54	4.21	5.7	309	511	224–2880	3.44, 3.46	1.15	1.72	0.18	0.26	42/87	1.9	156	20	84	24	25
δ	8	4.56	3.5	28	500	415–704	2.25	0.75	0.92	0.17	0.20	5/10	1.3	21	2	0	0	5
ϵ	4	1.93	3.3	13	691	531–905	1.25	0	0	0	0	4/4	1.3	9	4	0	0	0
Cyanobacteria	7	3.74	2.6	18	580	285–829	1.43	0	0	0	0	7/7	1.3	14	2	0	0	2
Firmicutes	31	2.77	6.5	202	299	13–794	3.32, 3.55	1.44	2.22	0.34	0.51	19/43	1.9	34	42	0	1	125
Actinobacteria	9	4.42	4.3	39	383	196–549	1.78, 2.56	0.43	0.76	0.04	0.09	4/9	1.0	0	0	0	0	39
Bacteroides	6	3.93	4.2	25	576	477–834	2.00	0.51	1.26	0.12	0.25	3/6	1.0	25	0	0	0	0
Other divisions[h]	5	1.80	2.2	10	331	293–401	1.60	0.20	0.20	0.03	0.03	2/3	1.4	2	4	0	0	4
Total	**155**	**3.81**	**4.8**	**748**	**476**	**13–2880**	**2.63, 2.75**	**0.94**	**1.44**	**0.17**	**0.25**	**109/201**	**1.6**	**360**	**74**	**84**	**25**	**205**
	(123)[i]	(3.67)	(4.7)	(581)	(480)	(13–2880)	(2.63, 2.75)	(1.05)	(1.57)	(0.18)	(0.27)	(84/158)	(1.6)	(275)	(51)	(51)	(24)	(180)

[a]Number of ribosomal RNA (*rrn*) operons per genome.

[b]Number of unique (nonidentical) ITS sequence variants per genome; sequences are unique if they differ by as few as 1 bp (within alignable region) from paralogues within the same ITS class; when two numbers are given, the second number indicates the count when sequence variants generated by internal insertion-deletion events (indels; ≥1 bp) are included; single values reflect instances in which including internal indels does not increase the number of variants.

[c]% div equals average pairwise nucleotide divergence as a percentage of the alignable sequence length; divergence values were corrected for multiple hits using a Kimura two-parameter model of nucleotide substitution, with the transition/transversion ratio set to 2.0; indels were excluded from calculations.

[d]% S equals nonpairwise nucleotide divergence calculated as the number of segregating sites as a percentage of the total alignable sequence length.

[e]Number of ITS alignments (as fraction of total) in which nucleotide divergence equals zero.

[f]Number of ITS classes (as defined by tRNA gene composition) per genome.

[g]Reflects subdivisions of ITS regions into classes: ITS with tRNA-alanine + tRNA-isoleucine (Ala + Ile), tRNA-Ala or tRNA-Ile only (Ala/Ile), tRNA-glutamate (Glu), other tRNAs, or no tRNA genes (none).

[h]"Other divisions" includes the *Chlamydiae, Spirochetes*, and *Aquificales*.

[i]To estimate bias due to the inclusion of multiple strains from the same species, summary statistics (in parentheses) were also calculated based on a data set from which replicates/strains of the same species were excluded.

ITS class were aligned either manually in MacClade 4.0 (Maddison and Maddison 2000) or automatically in CLUSTAL X (Chenna et al. 2003). Due to high levels of interclass variation in length and sequence and an obvious lack of positional homology among sequences, ITS sequences representing distinct classes were not aligned. Rather, nucleotide diversity was calculated only for alignments of ITS sequences belonging to the same class. For 12 genomes in this study (see Table 1, footnote *) CLUSTAL misaligned regions of ITS sequences (from the same class) that, upon manual inspection, were clearly identified as nonhomologous. These regions, while representing sequence divergence in the form of the recombination (insertion or deletion) of sequence blocks, erroneously inflate estimates of base substitution and therefore were excluded from calculations of nucleotide divergence. For each alignment, the average pairwise nucleotide divergence as a percentage of the alignable sequence length (% div) was calculated using the program dnadist in the PHYLIP 3.66 software package (Felsenstein 2005). Divergence values were corrected using a Kimura two-parameter model of nucleotide substitution, with the expected transition to transversion ratio set to 2.0. Uncorrected, nonpairwise nucleotide divergence (the number of segregating sites [S] as a percentage of the total alignable sequence length: % S) was also calculated for comparison using the program DnaSP V.4.0 (Rozas et al. 2003). Also, intragenomic divergence (% div and % S) in corresponding 16S rRNA gene sequences was calculated to assess the relationship between ITS and 16S rRNA gene divergence.

The insertion of gaps into alignments of sequences that differ in length may introduce error in genetic analyses due to the difficulty of unequivocally determining positional homology among nucleotides and, therefore, the uncertainty over gap placement (Lutzoni et al. 2000). Consequently, insertion-deletion events (indels) within ITS regions were not extensively examined in this paper and were excluded from calculations of average pairwise divergence. Nonetheless, the contribution of indels to intragenomic ITS divergence was estimated separately by counting the total number and length of internal gaps (≥1 bp) within each alignment. As with base substitutions, indels falling within misaligned regions were not included in total counts. For genomes in which this occurred (see Table 1, footnote *), the contribution of indels to ITS divergence is underestimated. For all other genomes in which the regions flanking gaps were clearly homologous, each internal indel, regardless of its length or whether it was composed of smaller overlapping indels, was counted as a single evolutionary event. The number of indels per alignment is expressed as a proportion of the total number of variable sites (number of indels + number of segregating sites due to base substitution; Table 1). All alignments generated in this study are available from the authors upon request.

rrn Operon Location

To assess the relationship between intragenomic variation and the spatial distribution of ITS-containing *rrn* operons in the bacterial genome, operon locations (bps relative to the origin of replication) were displayed graphically along the bacterial chromosome (shown as linear). Linear representations of the chromosome showing operon positions coded by ITS class were then displayed on a phylogenetic tree constructed for each bacterial group. Phylogenies were estimated in PAUP 4.0 (Swofford 2003) using maximum parsimony analysis of 16S rRNA gene sequences taken from the first *rrn* operon of each genome (based on bp numbering along the coding strand relative to the origin of replication). Analyses were conducted under the heuristic search option. For the few genomes (12 of 155) composed of more than one *rrn* operon-containing chromosome, only the chromosome containing the greatest number of *rrn* operons was displayed (Fig. 4). In addition, for all pairwise comparisons among ITS sequences of the same class within a genome, the genetic distance between sequences was plotted as a function of the physical distance (bp) separating these regions on the chromosome.

Results

We analyzed a total of 748 ITS sequences from 155 Bacteria genomes representing 12 major taxonomic divisions (Table 1). In all instances the number of ITS sequences per genome corresponds to the genome's *rrn* operon copy number; there are no instances in which an *rrn* operon (with full-length rRNA genes) does not contain a spacer separating the 16S and 23S rRNA genes. The number of *rrn* operons per genome ranges from 2 to 15 (maximum in *Photobacterium profundum*; Table 1), with a mean of 4.8 (SD, 2.7; median, 4.0). On average, bacteria in the gram-positive division *Firmicutes* contain the highest number of *rrn* operons per genome (mean, 6.5; SD, 3.0). (Note: To estimate bias due to the inclusion of multiple strains of the same species in the analysis, summary statistics were also calculated based on a smaller data set from which replicates/strains of the same species were excluded (Table 2, footnote i). These results differ insignificantly from those based on the entire data set.)

ITS Length

ITS length varies substantially among bacteria, ranging from 13 bp in *Lactobacillus salivarius* (*Firmicutes*) to 2880 bp in *Saccharophagus degradans* (*γ-Proteobacteria*; Tables 1 and 2), with an average of 476 bp (SD: 213). While ITS length differs considerably among bacterial groups, with the smallest mean ITS size in the gram-positive *Firmicutes*, there is considerable overlap in ITS length among taxa (Fig. 1). Of the 277 unique ITS lengths represented in this data set, 69 are shared by two or more genomes (range: two to seven genomes per ITS length). Conversely, ITS length can also vary substantially among *rrn* operons within a genome. Indeed, 59% of the genomes contain ITS regions that differ in length among operons. Interoperon ITS length variation ranges from 1 to 2201 bp, with the number of ITS length variants per genome ranging from 1 to 12 (mean, 2.4; maximum in *Photobacterium profundum*; Table 1).

ITS Class—tRNA Composition

Based on tRNA gene composition, ITS regions are separated into five primary classes (Table 1): ITS with tRNA-alanine + tRNA-isoleucine (tRNA-Ala + Ile; 48% of sequences), ITS with tRNA-Ala or tRNA-Ile (tRNA-Ala/Ile; 10%), ITS with tRNA-

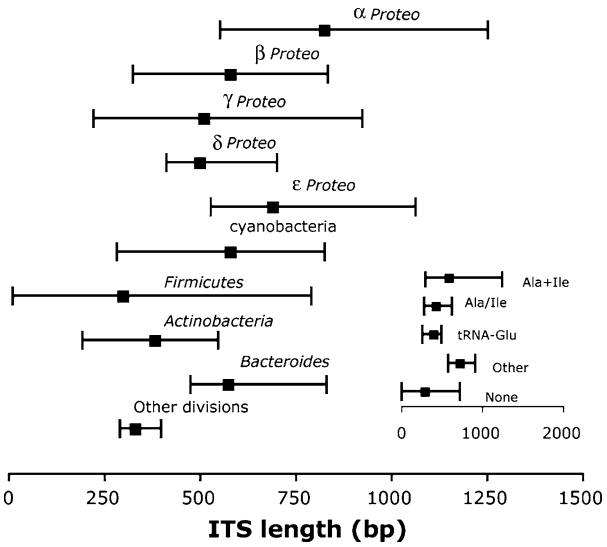

Fig. 1. Variation in mean ITS length among bacterial groups and ITS classes (inset). Bars show range in values. The anomalously high ITS length of 2880 bp in the genome of the γ-proteobacterium *Saccharophagus degradans* is excluded from the display. ITS classes are defined by the composition of tRNA genes within the ITS. "Other divisions" includes Bacteria belonging to the divisions *Chlamydiae, Spirochetes*, and *Aquificales*.

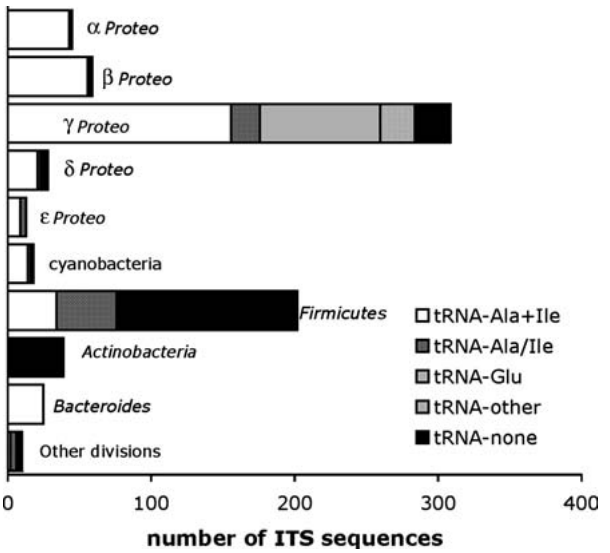

Fig. 2. Total numbers of ITS sequences analyzed for each group of Bacteria included in this study. Total numbers are subdivided to show proportions occupied by ITS sequences representing distinct ITS classes. ITS classes are defined by the composition of tRNA genes within the ITS. "Other divisions" includes bacteria belonging to the divisions *Chlamydiae, Spirochetes*, and *Aquificales*.

glutamate (tRNA-Glu; 11%), ITS with other tRNAs present (tRNA-other; 3%), and ITS with no tRNA genes (tRNA-none; 27%). ITS regions within the class tRNA-other are further subdivided according to their exact tRNA composition, as specified in Table 1 (column 2). The number of ITS classes per genome ranges from one to seven, with 41% of the genomes analyzed containing ITS sequences representing two or more ITS classes (Table 1). Multiple ITS classes occur most commonly in enteric bacteria and *Vibrio* species within the *γ-Proteobacteria*, as well as in gram-positive *Firmicutes* bacteria (Tables 1 and 2). In all cases, differences in ITS class among operons within a genome also reflect differences in ITS length. There is considerable variation in the relative abundances of different ITS classes among major groups of Bacteria (Fig. 2). Sixty-two percent (125 of 202) and 100% (39 of 39) of ITS sequences from the *Firmicutes* and the gram-positive *Actinobacteria*, respectively, contain no tRNA genes; this contrasts with 8% (25 of 309) in *γ-Proteobacteria*. The greater percentage of ITS sequences with no tRNA genes in *Firmicutes* and *Actinobacteria* likely corresponds to the typically shorter ITS length observed in these taxa. The presence of the gene for tRNA-Glu in the ITS may be specific to the *γ-Proteobacteria*, occurring in 27% (84 of 309) of ITS sequences from this division but not appearing in sequences from any other group. These data indicate conservation of ITS tRNA composition corresponding to phylogenetic affiliation, as suggested previously by Boyer et al. (2001).

Instances in which an ITS contains genes other than those for tRNA-Ala, tRNA-Ile, or tRNA-Glu are relatively rare and are confined to bacteria of the genus *Vibrio* (*γ-Proteobacteria*) and to the related bacterium *Photobacterium profundum* (Table 1). ITS regions in these organisms also may contain genes for tRNA-lysine (Lys) and tRNA-valine (Val), which often occur in combination with tRNA-Ala and tRNA-Glu in the same ITS region (Table 1). Also, in one species (*Desulfitobacterium hafniense* (*Firmicutes*)), an ITS contains the gene encoding the tRNA for phenylalanine (Phe). Interestingly, the first ITS (2880 bp long) in the genome of the γ-proteobacterium *Saccharophagus degradans* contains genes encoding a response regulator receiver domain protein and a putative lipopolysaccharide heptosyltransferase-1. These genes and the sequences flanking them are absent from the second ITS (679 bp) in the genome. However, the terminal 270 bp of each ITS is conserved with 100% identity between the two operons (Table 1). A similar pattern has been documented for *Mycoplasma imitans*, which has an exceptionally long ITS (2488 bp) containing a putative transposase gene (Harasawa et al. 2004). How the insertion of these genes into the ITS affects the cotranscription of this region with adjacent ribosomal rRNA genes or the potential role of the ITS in the processing of rRNA transcripts is unclear.

Intragenomic Nucleotide Divergence

To provide an initial estimate of intragenomic ITS sequence diversity, we counted the total number of unique ITS sequence variants per genome (Tables 1 and 2). These estimates reflect sequence variation due

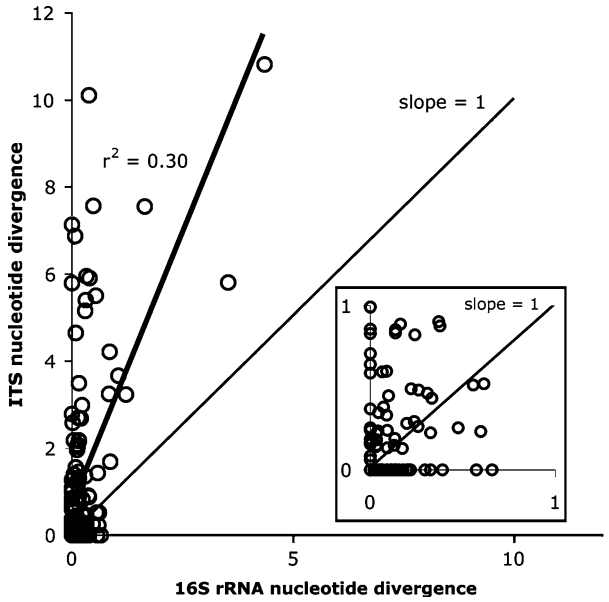

Fig. 3. Relationship between intragenomic ITS divergence and divergence in corresponding 16S rRNA genes. The best-fit line (bold) is shown. Inset shows only values below 1% divergence. A line of slope = 1 is shown to highlight points in which 16S rRNA divergence exceeds ITS divergence, all of which are restricted to instances when divergence is relatively low. Divergence values are corrected for multiple hits using a Kimura two-parameter model of nucleotide substitution.

to differences in ITS class (Supplementary Fig. 1A) as well as variation (≥1-bp difference) among sequences of the same class. Bacterial genomes in this data set contain, on average, 2.63 unique ITS sequence variants (range, 1–12; Tables 1 and 2). This value increases only slightly (to 2.75) when sequence variants generated by internal insertion-deletion events (indels; ≥1 bp) are included in the counts. As anticipated, the number of sequence variants per genome increases with *rrn* operon copy number (Supplementary Fig. 1B) and is highest in genomes with multiple ITS classes, reaching a peak of 12 variants in *Vibrio fisheri* and *Photobacterium profundum* (*γ-Proteobacteria)*, whose ITS sequences represent 6 and 7 distinct ITS classes, respectively.

Within an ITS class, intragenomic nucleotide divergence (% div) is surprisingly low. Across all genomes mean divergence is 0.94% (median, zero; range, 0%–12.11%; SD, 2.01), with 78% of all ITS alignments (157 of 201) showing <1% divergence (Supplementary Fig. 2). Indeed, 54% of all ITS alignments show zero sequence variation within each genome (i.e., no base substitutions; Table 1), and 81% contain at least two identical sequences (Fig. 4). Instances of zero variation occur in every major Bacteria taxa (Table 1, Fig. 4) and are not restricted to alignments with few sequences or to alignments in which functional constraint on sequence change may be imposed by the presence of tRNA genes. For example, the gram-positive bacterium *Clostridium acetobutylicum* possesses 11 ITS sequences, none of which encodes a tRNA gene and all of which are identical in sequence (Table 1). Furthermore, sequence homogenization is not limited only to short ITS regions; the length of ITS sequences showing zero intragenomic variation averages 546 bp and ranges from 13 to 1255 bp (Table 1). Interestingly, in seven genomes in which multiple *rrn* operon-containing chromosomes are present (*Brucella abortus* biovar 1 str. 9-941, *Brucella melitensis* 16M, *Brucella melitensis* biovar Abortus 2308, *Burkholderia pseudomallei* K96243, *Ralstonia eutropha* JMP134, *Rhodobacter sphaeroides* 2.4.1, *Vibrio vulnificus* YJ016), identical ITS sequences are found on separate chromosomes (see legend to Fig. 4; only the chromosome containing the greatest number of *rrn* operons is displayed in Fig. 4).

Intragenomic sequence divergence does differ among taxonomic groups (Table 2), however, ranging from zero in *ϵ-Proteobacteria* ($n=4$ genomes) and cyanobacteria ($n=7$) to an average of 1.44% in *Firmicutes* ($n=31$). On average, ITS divergence is highest in regions that do not contain a tRNA gene, suggesting that functional constraint on tRNA sequence change may impact the rate of base substitution in the ITS. Across all genomes analyzed, the greatest nucleotide divergence occurs in the *Firmicutes* bacteria *Geobacillus kaustophilus* (10.1% div), *Thermoanaerobacter tengcongensis* (10.8%), and *Carboxydothermus hydrogenoformans* (12.1%), all of which inhabit hydrothermal environments.

As expected, intragenomic divergence is greater for the ITS region (mean: 0.94% div) than for the 16S rRNA gene (mean: 0.17%; Table 2). Interestingly, however, in 24% (48 of 201) of the alignments, 16S rRNA divergence exceeds that in the corresponding ITS region (Table 1, Fig. 3), and the percentage of alignments showing zero divergence is lower for the 16S rRNA gene (42%) than for the ITS (54%). However, the extent to which 16S rRNA divergence exceeds ITS divergence is relatively small, averaging 0.13% ± 0.14 (SD) and ranging from 0.02% to 0.66%. The r^2 for the comparison between ITS and 16S divergence shows that 16S variation accounts for less than one-third of the intragenomic variation in the ITS (Fig. 3), suggesting that the processes generating intragenomic diversity, or the distribution of nucleotide polymorphisms among operons, may differ between ITS and 16S regions.

Counts of the number of internal gaps within alignments of ITS sequences from the same class show that insertion-deletion events (indels; ≥1bp) on average represent 16% of the total number of variable sites (number of indels + number of segregating sites due to base substitution; alignments with no variation excluded from averaging). Of the total number of indels, 48% are a single base pair

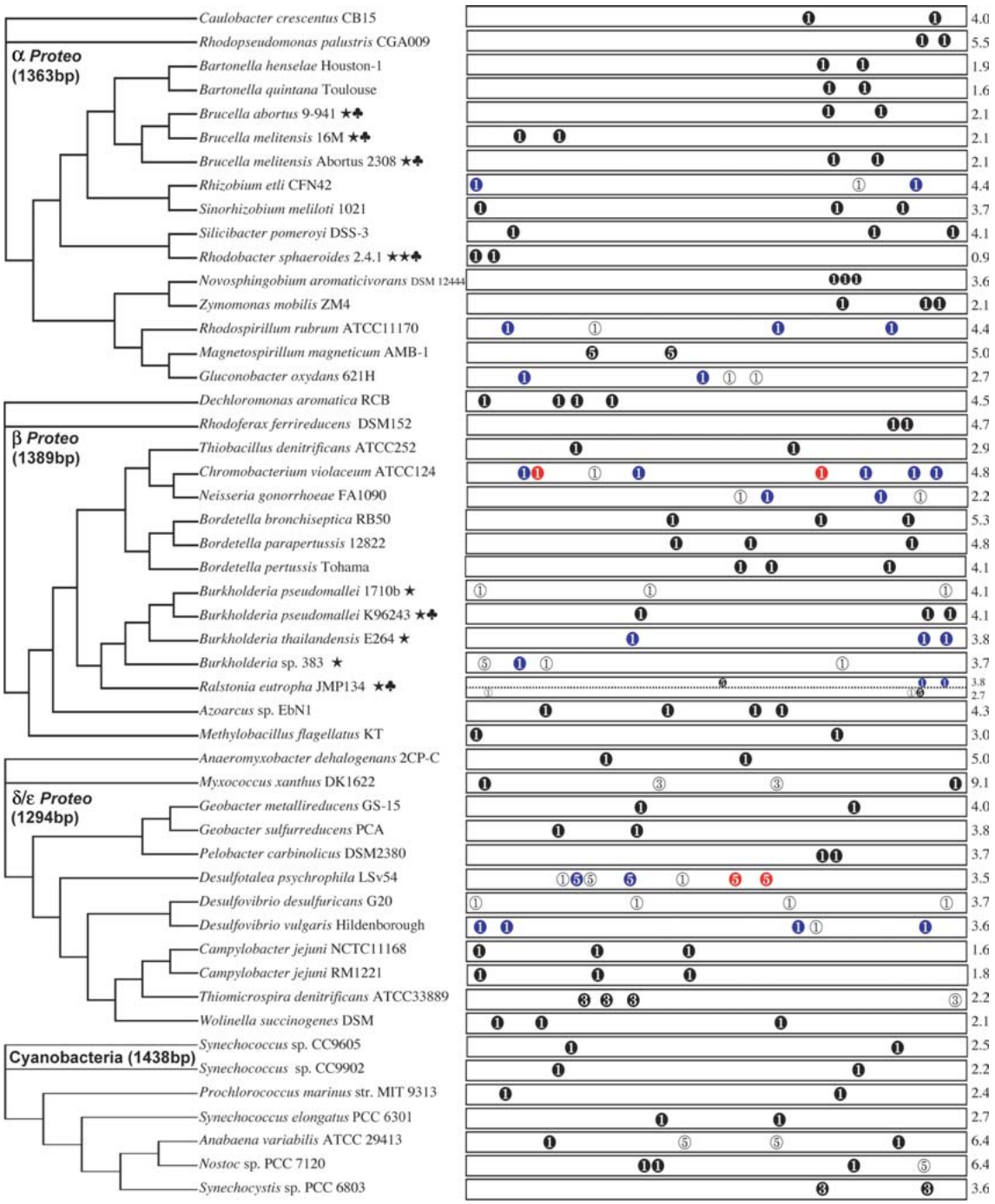

Fig. 4. The positions of ITS-containing *rrn* operons within the genomes of Bacteria included in this analysis. Bars show relative positions (standardized to chromosome length) of *rrn* operons along the chromosome. Numbers to the right of bars reflect chromosome size in Mbp. Numbered symbols correspond to ITS classes: ① = tRNA-Ala + Ile, ② = tRNA-Glu, ③ = tRNA-Ala/Ile, ④ = tRNA-other, ⑤ = tRNA-none. Note that ITS regions within class tRNA-other are further subdivided (using subscripts) according to their exact tRNA composition; this pertains only to *γ-Proteobacteria* in the genera *Vibrio* and *Photobacterium*. Filled symbols indicate identical ITS sequences (along the alignable length of the sequence and excluding indels). Black fill highlights instances in which all sequences within a class are identical. Colored fill identifies instances in which some but not all sequences are identical, with identical sequences grouped by the same color. For genomes in which multiple *rrn* operons are arrayed in close proximity to one another (e.g., *Bacillus* sp.), symbols were made smaller to more accurately display their position on the chromosome. ★ = genome contains more than one *rrn* operon-containing chromosome; only the chromosome containing the greatest number of *rrn* operons is displayed. ★★ = *Rhodobacter sphaeroides* chromosome II is shown; chromosome I is larger (3.2 Mbp) but contains only one *rrn* operon. ♣ = identical ITS sequences occur on separate chromosomes. "Other divisions" includes bacteria belonging to the divisions *Actinobacteria, Bacteroides, Chlamydiae, Spirochetes*, and *Aquificales*. Taxa are organized according to phylogenetic affiliation. Trees were generated via maximum parsimony analysis of 16S rRNA gene sequences from complete genomes using the heuristic search option in PAUP 4.0. Numbers by taxonomic group names indicate the length of 16S sequences used in each analysis; indels were excluded.

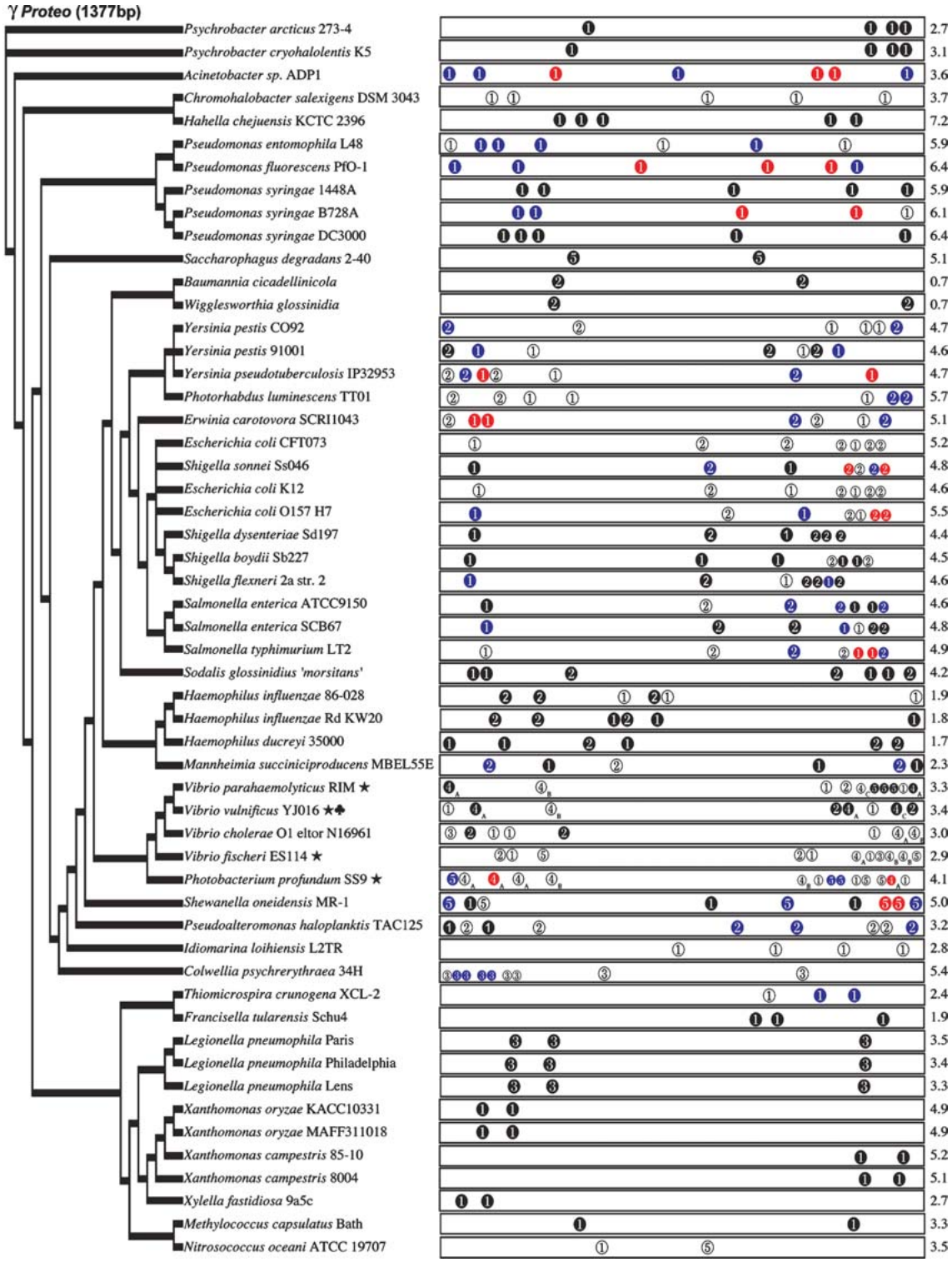

Fig. 4. Continued.

in length; when these singletons are excluded, the percentage of variable sites due to indels decreases to 9%. However, for some genomes, extensive recombination of sequence blocks (i.e., multiple indel events) within regions of the ITS precludes accurately determining the placement of gaps and establishing positional homology among nucleotides. Such ambiguously aligned regions were excluded from diversity analyses. In the 12 genomes in which this occurred the contribution of indels to intragenomic ITS variability is underestimated (see Table 1, footnote *).

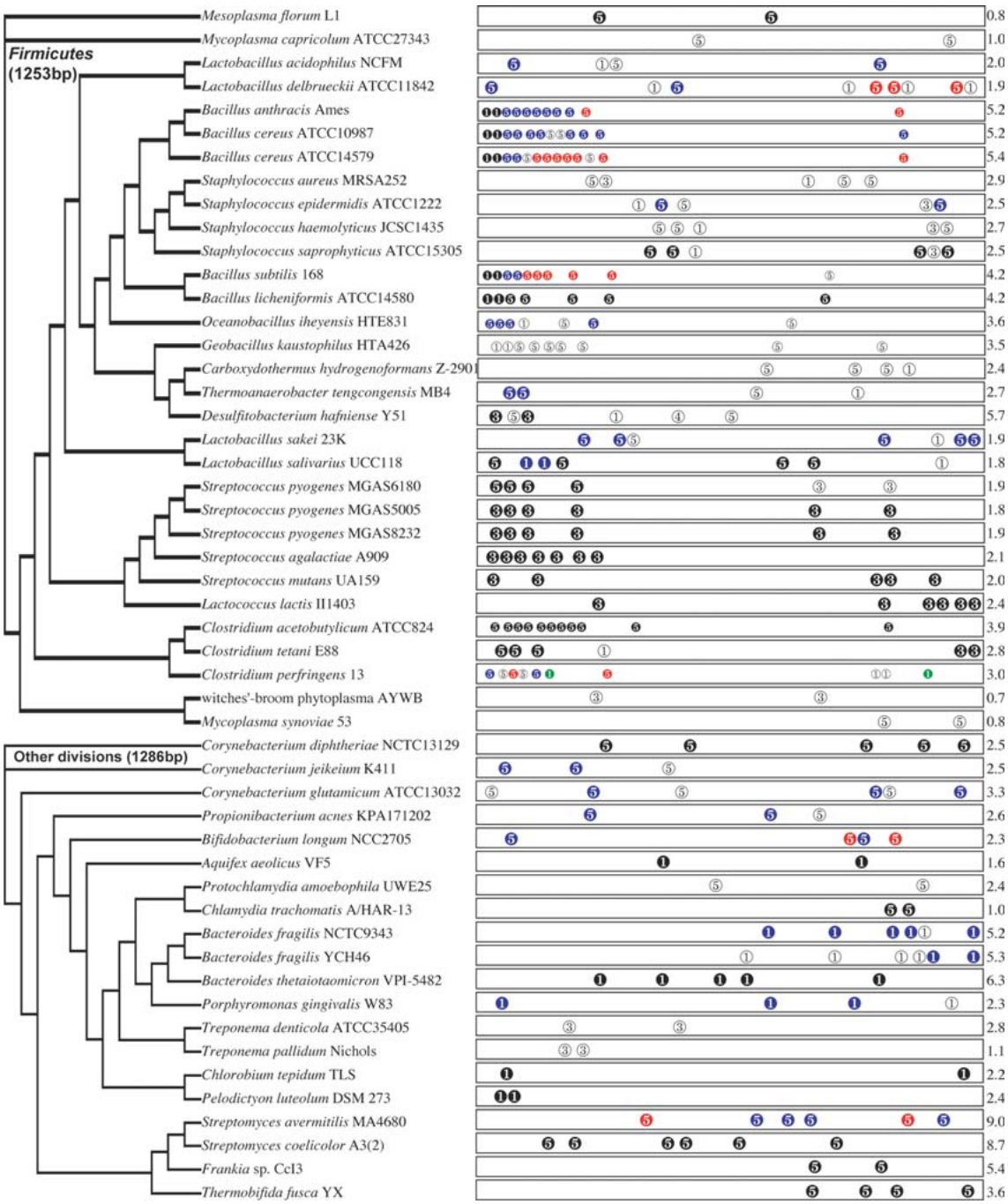

Fig. 4. Continued.

Intragenomic Divergence and rrn *Operon Position*

To examine the relationship between chromosome position and ITS structure and interoperon divergence, we mapped the linear distributions of *rrn* operons (coded by ITS class) onto the bacterial chromosome and displayed these distributions on the phylogenies of the taxa examined in this study (Fig. 4). Our results show that the spatial distribution of *rrn* operons on the chromosome is generally conserved among closely related taxa. However, there is not a clear relationship between intragenomic ITS divergence and *rrn* operon position on the chromosome; instances of zero divergence between operons can occur regardless of whether operons are clustered or dispersed throughout the genome (Fig. 4). Furthermore, across all genomes, pairwise nucleotide divergence between ITS sequences belonging to the same class within a genome did not significantly covary with pairwise distance (bp) between operons (regression slope, 2×10^{-7}; $R^2 = 0.01$; data not shown).

Discussion

To accurately assess intragenomic ITS divergence in Bacteria, we confined our analyses to sequence data from completed genome projects present in the NCBI database. Focusing solely on genomes with multiple *rrn* operons, we examined ITS sequences from 155 taxa representing 12 major phylogenetic divisions of Bacteria. The breadth of this analysis ensures a comprehensive description of ITS sequence variation within and across diverse bacterial groups. These data are of direct relevance both to microbial ecologists who use ITS variation as a proxy for microbial diversity, as well as to evolutionary biologists and geneticists seeking to understand recombination processes within the *rrn* gene family of Bacteria.

ITS Length Variation

As reported by other authors (Gürtler and Stanisich 1996; Fisher and Triplett 1999; Ranjard et al. 2000; Boyer et al. 2001), our data set shows considerable overlap in ITS length range among bacterial groups (Table 2). This raises the possibility that multiple phylotypes are represented by single gel bands or electropherogram peaks in ITS length-based diversity analyses (e.g., automated ribosomal intergenic spacer analysis [ARISA] [Crosby and Criddle 2003]). Such occurrences would result in underestimates of bacterial diversity. Of the 277 unique ITS lengths represented in this data set, 69 are shared by two or more genomes (range: 2–7 genomes per ITS length). Assuming an optimal resolution of 1 bp, if the diversity of organisms used in this analysis was assessed using the total number of unique ITS length variants as a proxy for richness, 25% of gel bands (or electropherogram peaks in ARISA) would be represented by multiple organisms. If resolution drops to 2 or 3 bp, this percentage increases to 43% or 54%, respectively. However, for this data set, the number of genomes sharing ITS sequences of identical length is biased upward by the inclusion of multiple strains from the same bacterial species. Indeed, of the 13 instances in which multiple strains of the same species occur in the data set (Table 1), 6 (46%) show conservation of identical ITS length variants among strains.

The potential masking of distinct strain-level genetic variants by ITS length-based diversity analyses should not be overlooked. Strain-level genetic diversity is increasingly recognized as an important component of natural microbial communities (Schloter et al. 2000; Acinas et al. 2004a; Coleman et al. 2006). Indeed, strain-level variants in ITS or 16S rRNA gene sequence have been linked to genomes that differ significantly in allelic composition and size (up to 1.1 Mbp size variation [Thompson et al. 2005]). Furthermore, strain-level variation in co-existing bacteria has been definitively linked to variation in metabolic and ecological capabilities (e.g., substrate use, growth rate, motility [Rocap et al. 2002; Jaspers and Overmann 2004; Hahn and Pöckl 2005]). If strains of the same species commonly possess ITS regions of identical length, as suggested here, analyses that use the total number of ITS length variants as a proxy for diversity may exclude functionally divergent, and ecologically important, genetic variants from diversity estimates.

Conversely, community fingerprinting techniques such as ARISA, unless coupled to sequence-based assessments of intragenomic ITS variation (e.g., clone library analysis [García-Martínez et al. 1999; Brown et al. 2005; Kent et al. 2006]), may overestimate microbial diversity when bacteria with multiple *rrn* operons are present. Indeed, 59% of the genomes analyzed here contain ITS regions that differ in length among operons (range in interoperon length variation: 1–2201 bp), and the total number of unique ITS length variants exceeds the true richness of the data set by 79%. However, for some environments (e.g., oligotrophic marine waters), the contribution of bacteria with multiple *rrn* operons to estimates of diversity has been considered relatively minor (Brown et al. 2005). This is partly based on genome sequences showing that some common marine bacteria (e.g., *Prochlorococcus*, *Synechococcus*, *Silicibacter pomeroyi*) typically possess few *rrn* operons and that when multiple copies are present little to no ITS length variation occurs (Rocap et al. 2003; Palenik et al. 2003; Moran et al. 2004; Brown et al. 2005). Indeed, intragenomic ITS length homogeneity is confirmed in this study for several bacterial clades, including the cyanobacteria (Table 1). However, significant variation in ITS length occurs in other groups. For example, members of the γ-proteobacterial genus *Vibrio* and their relative *Photobacterium profundum* possess between 6 and 12 ITS length variants per genome (Table 1). *Vibrio* species are often of importance to both ecological processes and human health in natural systems (e.g., Colwell 1996). For these systems, ITS-based estimates of microbial diversity that do not account for interoperon variation will likely misrepresent the richness of the bacterial community. The same may be true for nutrient-rich or temporally or environmentally heterogeneous habitats, such as soil, which may select for bacteria with multiple *rrn* operons (Klappenbach et al. 2000; Weider et al. 2005). Determining the ratio between total *rrn* number and phylotype richness for ecologically distinct habitats will help us better understand how intragenomic ITS variation may bias diversity estimates.

Intragenomic ITS Variation—Nucleotide Divergence

Intragenomic ITS variation may also be evaluated directly at the level of sequence divergence. On average, a bacterium in this study contains 2.75 unique ITS sequence variants per genome, with the number of variants per genome increasing by 0.61 per additional *rrn* operon copy (Supplementary Fig. 1B). This suggests that multicopy (paralogous) *rrn* operons within a genome may represent a substantial source of bias in studies where ITS sequence richness is used as a proxy for taxonomic richness (e.g., clone library analysis). Indeed, for this data set, if distinct ITS sequences were to be interpreted as distinct taxa, taxonomic richness would be overestimated by 175% (mean number of sequence variants per taxa = 2.75). However, this is a maximum value given the possibility that distinct bacterial species share identical ITS sequences. Indeed, while the conservation of ITS sequences across distinct species was not examined, the sharing of an identical ITS sequence among bacteria that differ at the strain level occurs in 3 of the 13 instances in which multiple strains of the same species were included in the data set (Table 1). The conservation of ITS sequence among functionally divergent bacteria that differ genetically at the strain level in natural samples therefore warrants closer examination. However, given prior reports of high variability in ITS length and sequence among even closely related taxa (Gürtler and Stanisich 1996; García-Martínez et al. 1999), ITS sequences are unlikely to be conserved across more distantly related bacteria (i.e., species level and greater). Rather, the primary error in diversity analyses is likely an overestimation of taxonomic richness due to erroneous interpretation of distinct paralogue sequences as distinct taxa.

The increase in ITS sequence richness due to increasing *rrn* operon copy number is commonly manifested as differences in the presence or absence of distinct sequence blocks that define ITS classes (e.g., tRNA genes). Indeed, 58% of ITS richness within a genome is explained by differences in ITS class (tRNA composition) among operons (mean number of ITS classes per genome, 1.6; Supplementary Fig. 1A); the remainder is due to sequence variation within a class. Alignments of ITS sequences grouped by class within a genome allowed direct quantification of the role of base substitutions in generating intragenomic ITS divergence.

Nucleotide divergence due to base substitution occurs in 46% of alignments of paralogous ITS sequences, with values ranging from 0% to 12.11% (mean: 0.94%). While low, these values are certainly within the range of those resulting from interpopulation, or strain-level, divergence between closely related bacteria (e.g., Brown and Fuhrman 2005; Osorio et al. 2005; Dechaine et al. 2006). This raises the possibility that interoperon ITS divergence may spuriously be interpreted as genetic variation resulting from population-level processes (e.g., divergence along biogeographic or environmental gradients). This is possible if preferential amplification of the ITS from a specific operon occurs in some samples but not in others (Boyer et al. 2001). Avoiding this bias would likely require the use of operon-specific primers for PCR amplification (Antón et al. 1998; Boyer et al. 2001). This restricts the use of the ITS for population-level genetic analyses to organisms for which *rrn* operon variation within the genome has been quantified.

Interestingly, intragenomic divergence is highest in the genomes of three thermophiles, the gram positive bacteria *Geobacillus kaustophilus*, *Thermoanaerobacter tengcongensis*, and *Carboxydothermus hydrogenoformans*. Indeed, for *G. kaustophilus*, seven of its nine ITS sequences, despite belonging to the same ITS class (tRNA-none), exhibit such high length (386−735 bp) and sequence heterogeneity that they cannot be effectively aligned, aside from a sequence block of 138 bp that is conserved across all seven ITS regions. Even within this conserved block, 24% of nucleotide sites are polymorphic. ITS variability in these thermophiles is consistent with a prior study showing that the highest interoperon divergence in 16S rRNA sequences occurs in thermophilic Bacteria and Archaea, suggesting a connection between *rrn* divergence and environmental conditions (Acinas et al. 2004b).

While our estimates of sequence divergence do not reflect insertion-deletion events (indels), prior studies suggest that indels contribute substantially to intragenomic ITS variation in some organisms. Several studies focusing on one or few bacterial taxa have analyzed ITS sequences across all operons within a genome, showing interoperon variation in the form of distinct combinations of conserved sequence blocks (Gürtler and Stanisich 1996; Antón et al. 1998; Lan and Reeves 1998; Wenner et al. 2002). These blocks include both tRNA genes and non-tRNA elements. Such non-tRNA blocks may include short terminal sequences (5' and 3' ends of ITS) that participate in RNA processing (Gürtler 1999), as well as other functionally significant elements, such as the Box A and B sequences, which putatively function as antitermination sites during RNA transcription and have been shown to be highly conserved across *rrn* operons within a genome and even across bacterial species (Berg et al. 1989; Iteman et al. 2000; Osorio et al. 2005). Such sequence blocks may occur in a mosaic pattern, being inserted into some ITS regions and deleted from others, or arranged in distinct combinations in different operons. Indeed, this mosaic composition has been cited as evidence that the ITS

region undergoes frequent homologous recombination that serves both to create new ITS variants and to homogenize sequences from distinct operons (Lan and Reeves 1998; Privitera et al. 1998; García-Martínez et al. 1999; Gürtler 1999; Liao 2000; Gianninò et al. 2003; Milyutina et al. 2004; Osorio et al. 2005).

Though indels resulting from recombination undoubtedly contribute to intragenomic diversification of ITS sequences, our data suggest that they do so primarily at the level of ITS class. In most genomes ITS class designation (tRNA composition) is a useful indicator of homology (Osorio et al. 2005; this study). Though not always the case (see below), sequences belonging to the same class can typically be aligned along the full length of the sequence, indicating that they share the same combination of sequence blocks. This suggests that recombination involving short sequence blocks, while potentially important in generating distinct ITS classes, plays a smaller role in generating sequence variants within a class. Indeed, our calculations show that if indels are considered when distinguishing ITS sequence variants, the average number of variants per genome increases only slightly, from 2.63 to 2.75. Furthermore, on average, indels >1 bp represent only $\sim$10% of the total number of variable sites within an ITS alignment, suggesting that ITS divergence among sequences of the same class within a genome is due primarily to base substitution.

Nonetheless, while indels are a minor fraction of sequence variation and are typically absent from alignments showing zero nucleotide divergence, in a few genomes intragenomic divergence within an ITS class is clearly attributable to indel events. These generated ITS variants differing in length and the presence or absence of distinct sequence blocks. For example, the β-proteobacterium *Chromobacterium violaceum* possesses eight ITS sequences, all of which encode tRNA-Ala and tRNA-Ile. These sequences consist of two distinct ITS variants: a short (507-bp) and a long (680- to 681-bp) variant present in two and six *rrn* operons, respectively. Comparisons of sequences corresponding to each variant show that the two short sequences are identical and the six longer sequences contain only one polymorphic site. However, alignment of all eight sequences shows two conserved regions (393-bp total; 0.25% divergence) flanking an unalignable central region containing long indels that define each ITS variant. Similarly, the γ-proteobacterium *Chromohalobacter salexigens* contains five ITS regions that belong to the same class (tRNA-Ale + Ile) but differ in length (range: 573–707 bp) due to insertions and deletions of distinct sequence blocks. These indels, which prevent alignment along $\sim$40% of the sequence and contribute considerably to interoperon ITS variation, are not represented in our estimates of sequence divergence, which account only for nucleotide substitutions within alignable regions. This pattern also occurs in the ITS sequences of the actinobacterium *Streptomyces coelicolor* A3(2), which show no nucleotide substitutions in alignable regions but nonetheless vary considerably in length (196–280 bp) and indel composition. These examples illustrate that in some genomes insertion-deletion events operate within an ITS class to create a mosaic of alignable and unalignable sequence blocks, and that this diversity is potentially masked by estimates of divergence that do not account for indel events. However, the insertion of gaps into alignments of sequences potentially undergoing rapid evolution, such as the ITS, is complicated by the difficulty of accurately determining gap placement (Simmons and Ochoterena 2000; Pearce 2006). Until this problem can be resolved, hypervariable regions in some ITS alignments should be treated as possible sources of error in diversity estimates.

Implications for Concerted Evolution of the ITS

Our analysis provides important insight into the evolution of the ITS in Bacteria. Most notably, we show that, despite the potential for multiple ITS variants per genome, ITS regions undergo extensive sequence homogenization among *rrn* operons. Fifty-four percent of all ITS alignments show zero nucleotide polymorphisms, and 81% contain at least two identical sequences. This pattern could arise if *rrn* operons duplicated recently, leaving little time for divergence to occur. However, given the elevated rate of sequence evolution in the ITS (Gürtler and Stanisich 1996; Antón et al. 1998; Schloter et al. 2000; Rocap et al. 2002, 2003; Brown and Fuhrman 2005), operon duplication would have had to occur very recently over a diverse range of bacteria to explain the high level of within-genome ITS homogeneity observed here. This seems unlikely, though it is impossible to rule out recent duplication as a potential contributor to low ITS divergence in some genomes. A more likely hypothesis is that low ITS divergence is driven by concerted evolution.

Concerted evolution, the process by which genetic content is homogenized among paralogs in a multigene family, is extensively studied in eukaryotes. This process has been invoked as the primary mode of evolution in the eukaryotic *rrn* gene family (Hillis et al. 1991; Nei and Rooney 2005), which in some species consists of hundreds to thousands of identical operons arrayed in tandem along the chromosome (Brown et al. 1972; Nei and Rooney 2005). While *rrn* gene families of bacteria contain considerably fewer operons, which are often dispersed throughout the genome rather than tandemly arrayed (e.g., Fig. 4),

prior studies show that bacterial *rrn* operons also evolve in concert (Mattatall and Sanderson 1996; Antón et al. 1998; Liao 2000; González-Escalona et al. 2005; Santoyo and Romero 2005). Concerted evolution of bacterial *rrn* operons has been demonstrated in part by comparisons of ITS regions within and among genomes of related taxa (Gurtler and Mayall 1999). Specifically, if each operon evolves independently, ITS variation within a genome should equal variation between genomes (assuming that operons duplicated prior to the divergence of the genomes being compared, as evidenced by conservation in the number and chromosomal position of operons; Fig. 4). Conversely, if operons evolve in concert, between-genome variation should exceed within-genome variation. Unfortunately, multiple between-genome comparisons are beyond the scope of this study, owing partly to the large number of taxa included and to the inability to unambiguously align ITS regions from all but the most closely related taxa. Nonetheless, other authors focusing on fewer taxa or more specifically on rRNA have demonstrated *rrn* operon homogenization within species but little sequence similarity between species (Gürtler and Stanisich 1996; Anton et al. 1998), providing strong evidence of concerted evolution in this gene family.

Prior studies suggest that homogenization of sequence tracts among *rrn* operons (concerted evolution) likely occurs via multiple gene conversion events (nonreciprocal DNA transfer between homologous sequences; Gürtler 1999; Liao 2000; González-Escalona et al. 2005; Santoyo and Romero 2005). Such events rearrange or delete relatively short sequence blocks, which are generally less than 500 bp but more typically less than 120 bp (e.g., tRNA genes, Box A,B elements [Privitera et al. 1998; Gürtler 1999]). As discussed above, these events are likely responsible for generating ITS variants of distinct length and composition, each of which may be present in multiple identical copies within a genome (Fig. 4) (Gürtler 1999; Gürtler and Mayall 1999; Liao 2000). However, gene conversion may also involve longer stretches of nucleotides. Interoperon gene conversion involving the entire ITS region (~400 bp) was shown experimentally in *Escherichia coli*, in which conversion events between 16S rRNA genes caused the replacement of an ITS lacking tRNA-Ala and tRNA-Ile genes with one containing them (Hashimoto et al. 2003). Also, conversion tracts of a similar size range (150 to 800 bp) were detected in a nitrogenase structural gene in the α-proteobacterium *Rhizobium etli* (Santoyo et al. 2005). In our study, the length of ITS regions showing zero intragenomic variation averages 546 bp and ranges from 13 to 1255 bp (Table 1). These data confirm that processes homogenizing the *rrn* operon can act over large (> 500-bp) sequence tracts, which often include the entire ITS region. The prevalence of homogenized ITS sequences observed in this study raises the possibility that in some genomes concerted evolution in the *rrn* multigene family involves the gross replacement of one operon with another. This would constitute a refinement of the model presented by Liao (2000), which, based on analysis of 12 Bacteria and Archaea genomes, suggests that sequence conversion is patchy and occurs in short, discontinuous tracts throughout the *rrn* operon. However, sequence analysis along the full length of the *rrn* operon is needed to definitively demonstrate whole-operon conversion events.

In contrast to a strict model of concerted evolution in which all gene copies in an *rrn* multigene family are homogenized (e.g., Nei and Rooney 2005; Rooney and Ward 2005), our data indicate that concerted evolution homogenizes operons within distinct subgroups (defined here by ITS class or tRNA composition). Thus, while concerted evolution (potentially via gene conversion) occurs commonly among *rrn* operons in Bacteria, it often does not result in the sweep of a single operon type and its fixation throughout a genome. Distinct ITS classes are maintained in 41% of genomes (often with homogenization occurring within a class), suggesting selection for multiple ITS types within a genome. In eukaryotes, selection has been shown to optimize the length of the ITS, with length variants differing in the number of enhancer and promoter sites and therefore in the rate of transcription of the *rrn* operon (Weider et al. 2005). In contrast, for ITS regions in Bacteria, in which enhancer and promoter sites are absent, sequence elements under selection may include double-stranded processing stems involved in maturation of adjacent rRNAs, antitermination sites, tRNA genes, or other sequence blocks whose function has not yet been identified (Gürtler 1999). Quantifying a selective benefit, if any, of maintaining multiple ITS variants within a genome will involve examining the conservation of different combinations of sequence blocks among operons and across genomes, as well as determining the function of uncharacterized ITS sequence blocks. Selection studies also may require quantifying the fitness effect of experimentally altering the relative abundance of distinct ITS types within a genome. Such studies, focused on both ITS regions and rRNA genes, will help determine the role strong purifying selection may play in homogenizing the *rrn* gene family. To date, the balance between purifying selection and concerted evolution in the homogenization of *rrn* operons has been largely understudied (Nei and Rooney 2005).

Sequence divergence in the ITS appears not to be tightly linked to divergence in the corresponding 16S rRNA gene. Indeed, intragenomic variation in 16S genes explains less than one-third of the variation in

the ITS (Fig. 3), suggesting that these two regions may not evolve in concert across all operons. Furthermore, though intragenomic ITS variation typically exceeds variation in adjacent 16S genes, in approximately one-quarter of the alignments, the opposite occurs. This seems counter to the assumption that the 16S gene experiences significantly greater functional constraint on sequence change relative to the ITS (e.g., García-Martínez et al. 1999; Liao 2000). However, 16S variation exceeds ITS variation only in instances when both values are low (typically $< 0.5\%$). Such instances may reflect recent divergence among operons (i.e., recent gene conversion or duplication events). Given short interoperon divergence times and the stochastic nature in which mutations accumulate, 16S divergence may be expected to exceed ITS divergence by chance alone, even under strong constraint.

Alternatively, these regions may differ in the rate at which they are homogenized. Liao (2000) suggests that homogenization (via gene conversion) occurs primarily in genic regions of the *rrn* operon, citing as evidence the extensive sequence heterogeneity in the cotranscribed ITS regions of 12 Bacteria and Archaea genomes. However, additional evidence is needed to rule out the possibility that higher heterogeneity in the ITS might be due to reduced selective constraint in this region. Furthermore, direct measurements of gene conversion rates will be necessary to definitively conclude that recombination dynamics differ between 16S and ITS regions. Unfortunately, estimating intragenomic recombination rates is complicated by the low levels of genetic diversity observed among 16S and ITS paralogues. Most estimators of recombination rate perform poorly at such low levels of diversity (Wall 2000; Posada et al. 2002) and would likewise be inhibited by the relatively low number of sequences per alignment (mean: $\sim$5). Though beyond the scope of this study, empirically measuring recombination rates for distinct regions of the *rrn* operon would provide important insight into the concerted evolution of the operon as a whole.

ITS Homogenization and rrn *Operon Location*

This study also examined the possibility that sequence homogenization among paralogous ITS regions may depend on the physical location of *rrn* operons on the bacterial chromosome (Fig. 4). In contrast to eukaryotic genomes, in which multiple *rrn* operons are typically clustered in a tandem array along a chromosome (Brown et al. 1972; Nei and Rooney 2005), bacterial genomes do not exhibit a consistent operon distribution. In some genomes (e.g., *Bacillus* sp.), *rrn* operons are clustered near the origin of replication, presumably to accommodate the greater need for newly synthesized proteins at the beginning of cell division (García-Martínez et al. 1999). However, this pattern is far from universal; *rrn* operons also are frequently dispersed throughout the genome, as in several species of β-, δ-, and ϵ-*Proteobacteria* (Fig. 4). But among closely related genera and species (e.g., the enteric bacteria *Escherichia, Shigella, Salmonella* sp.) and certainly among strains (e.g., strains of *Legionella pneumophila*, *Campylobacter jejuni*), the spatial distribution of *rrn* operons may be conserved (Fig. 4). However, there is not a clear relationship between intragenomic ITS homogenization and *rrn* operon position on the chromosome. Based on analysis of 14 Bacteria and Archaea genomes, Hashimoto et al. (2003) finds a significant positive correlation between the pairwise genetic distance between *rrn* operons and the physical distance between them on the chromosome. This suggests an inverse relationship between operon proximity and the rate of gene conversion, with higher conversion rates leading to higher levels of sequence homogenization. Indeed, the relationship between gene proximity and recombination rate has been demonstrated experimentally for both eukaryotes and prokaryotes (Dvorak et al. 1987; Lovett et al. 1994; Segall and Roth 1994). Interestingly, however, our data indicate that in many genomes ITS homogenization, potentially via gene conversion (Liao 2000), is not dependent on operon proximity. In contrast to the work of Hashimoto et al. (2003), our analysis does not reveal a significant positive relationship between genetic distance and physical distance (bp) between operons (regression slope, 2×10^{-7}; $R^2 = 0.01$, data not shown). Figure 4 shows that homogenization (zero divergence) can occur regardless of whether operons are clustered (e.g., 11 operons of *Clostridium acetobutylicum* (*Firmicutes*)) or dispersed throughout the genome (e.g., 5 operons of *Corynebacterium diptheriae* (*Actinobacteria*)). Furthermore, in seven genomes analyzed here (Fig. 4) identical ITS sequences occur on separate chromosomes. This provides further evidence that sequence homogenization also occurs across chromosomes in bacteria, as demonstrated previously for gene conversion events involving 16S rRNA genes in the soil bacterium *Ochrobactrum intermedium* (Teyssier et al. 2003). Large physical distances between paralogous *rrn* operons therefore may not be enough to allow an ITS region to avoid gene conversion and begin to diverge. Together, our data indicate that extensive sequence homogenization among paralogues (concerted evolution) is the dominant feature of ITS evolution in Bacteria.

Summary

This study used sequence data from 155 complete Bacteria genomes to systematically assess the structure, intragenomic divergence, and evolution of the

ITS region. We show that large intragenomic variation may occur, but primarily at the level of ITS class. In contrast, sequence divergence within classes is surprisingly low, and most within-class ITS alignments show no variation. Our results underscore the pervasiveness of concerted evolution in the *rrn* gene family, showing that in many instances distinct ITS classes are maintained within a genome and that homogenization of sequence occurs within a class. Knowledge of how the ITS region evolves across diverse bacterial groups helps microbial ecologists and population geneticists assess the efficacy of using this marker for studies of strain-level variation and gene flow in natural populations. Furthermore, this work reveals the extent to which different mechanisms (e.g., functional constraint on sequence evolution, gene conversion) purge or maintain sequence variation in the *rrn* gene family of Bacteria.

Acknowledgments. We thank Rob Young, Scott Edwards, and members of the Cavanaugh lab for their critical comments and support during the preparation of the manuscript. This work was supported by National Science Foundation Grants EF-0412205 and OCE-0453901 awarded to C. Cavanaugh and by the Genetics and Genomics Training Program (GGT) at Harvard University.

References

Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL, Polz MF (2004a) Fine-scale phylogenetic architecture of a complex bacterial community. Nature 430:551–5554

Acinas SG, Marcelino LA, Klepac-Ceraj V, Polz MF (2004b) Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. J Bacteriol 186:2628–2635

Antón AI, Martínez-Murcia AJ, Rodríguez-Valera F (1998) Sequence diversity in the 16S-23S intergenic spacer region (ISR) of the rRNA operons in representatives of the *Escherichia coli*. ECOR collection J Mol Evol 47:62–72

Berg KL, Squires C, Squires CL (1989) Ribosomal RNA operon anti-termination—function of leader and spacer region box B-box A sequences and their conservation in diverse microorganisms. J Mol Biol 209:345–358

Boyer SL, Flechtner VR, Johansen JR (2001) Is the 16S-23S rRNA internal transcribed spacer region a good tool for use in molecular systematics and population genetics? A case study in cyanobacteria. Mol Biol Evol 18:1057–1069

Boyer SL, Johansen JR, Flechtner VR, Howard GL (2002) Phylogeny and genetic variance in terrestrial Microcoleus (Cyanophyceae) species based on sequence analysis of the 16S rRNA gene and associated 16S-23S ITS region. J Phycol 38:1222–1235

Brown MV, Fuhrman JA (2005) Marine bacterial microdiversity as revealed by internal transcribed spacer analysis. Aquat Microb Ecol 41:15–23

Brown MV, Schwalbach MS, Hewson I, Fuhrman JA (2005) Coupling 16S-ITS rDNA clone libraries and automated ribosomal intergenic spacer analysis to show marine microbial diversity: development and application to a time series. Environ Microbiol 7:1466–1479

Brown DD, Wensink PC, Jordan E (1972) *Xenopus laevis* and *Xenopus mulleri*: the evolution of tandem genes. J Mol Biol 63:57–73

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD (2003) Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 31:3497–3500

Chun J, Huq A, Colwell RR (1999) Analysis of 16S-23S rRNA intergenic spacer regions of *Vibrio cholerae* and *Vibrio mimicus*. Appl Environ Microbiol 65:2202–2208

Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, DeLong EF, Chisholm SW (2006) Genomic islands and the ecology and evolution of Prochlorococcus. Science 311:1768–1770

Colwell RR (1996) Global climate and infectious disease: the cholera paradigm. Science 274:2025–2031

Crosby LD, Criddle CS (2003) Understanding bias in microbial community analysis techniques due to *rrn* operon copy number heterogeneity. Biotechniques 34:790–802

DeChaine EG, Bates AE, Shank TM, Cavanaugh CM (2006) Off-axis symbiosis found: characterization and biogeography of bacterial symbionts of *Bathymodiolus* mussels from Lost City hydrothermal vents. Environ Microbiol 8:1902–1912

D'Auria G, Pushker R, Rodriguez-Valera F (2006) IwoCS: analyzing ribosomal intergenic transcribed spacers configuration and taxonomic relationships. Bioinformatics 22:527–531

Di Meo CA, Wilbur AE, Holben WE, Feldman R, Vrijenhoek RC, Cary SC (2000) Genetic variation among endosymbionts of widely distributed vestimentiferan tubeworms. Appl Environ Microbiol 66:651–658

Dvorak JD, Jue D, Lassner M (1987) Homogenization of tandemly repeated nucleotide sequences by distant-dependent nucleotide sequence conversion. Genetics 116:487–498

Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle

Fisher MM, Triplett EW (1999) Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. Appl Environ Microbiol 65:4630–4636

García-Martínez J, Acinas SG, Antón AI, Rodríguez-Valera F (1999) Use of the 16S-23S ribosomal genes spacer in studies of prokaryotic diversity. J Microbiol Meth 36:55–64

Gianninò V, Santagati M, Guardo G, Cascone C, Rappazzo G, Stefani S (2003) Conservation of the mosaic structure of the four internal transcribed spacers and localization of the *rrn* operons on the *Streptococcus pneumoniae* genome. FEMS Microbiol Lett 223:245–252

González-Escalona N, Romero J, Espejo RT (2005) Polymorphism and gene conversion of the 16S rRNA genes in the multiple rRNA operons of *Vibrio parahaemolyticus*. FEMS Microbiol Lett 246:213–219

González-Escalona N, Romero J, Guzmán CA, Espejo RT (2006) Variation in the 16S-23S rRNA intergenic spacer regions in *Vibrio parahaemolyticus* strains are due to indels nearby their $tRNA^{Glu}$. FEMS Microbiol Lett 256:38–43

Graham TA, Golsteyn-Thomas EJ, Thomas JE, Gannon VP (1997) Inter- and intraspecies comparison of the 16S-23S rRNA operon intergenic spacer regions of six *Listeria* spp. Int J Syst Bacteriol 47:863–869

Gürtler V (1999) The role of recombination and mutation in 16S-23S rDNA spacer rearrangements. Gene 238:241–252

Gürtler V, Stanisich V (1996) New approaches to typing and identification of bacteria using the 16S-23S rDNA spacer region. Microbiology 142:3–16

Gürtler V, Mayall BC (1999) rDNA spacer rearrangements and concerted evolution. Microbiology 145:2–3

Hahn MW, Pöckl M (2005) Ecotypes of planktonic Actinobacteria with identical 16S rRNA genes adapted to thermal niches in temperate, subtropical, and tropical freshwater habitats. Appl Environ Microbiol 71:766–773

Harasawa R, Pitcher DG, Ramírez AS, Bradbury JM (2004) A putative transposase gene in the 16S-23S rRNA intergenic spacer region of *Mycoplasma imitans*. Microbiology 150:1023–1029

Hashimoto JG, Stevenson BS, Schmidt TM (2003) Rates and consequences of recombination between rRNA operons. J Bacteriol 185:966–972

Hillis DM, Moritz C, Porter CA, Baker RJ (1991) Evidence for biased gene conversion in concerted evolution of ribosomal DNA. Science 251:308–310

Hurtado LA, Mateos M, Lutz RA, Vrijenhoek RC (2003) Coupling of bacterial endosymbiont and host mitochondrial genomes in the hydrothermal vent clam *Calyptogena magnifica*. Appl Environ Microbiol 69:2058–2064

Iteman I, Rippka R, de Tandeau Marsac N, Herdman M (2000) Comparison of conserved structural and regulatory domains within divergent 16S rRNA-23S rRNA spacer sequences of cyanobacteria. Microbiology 146:1275–1286

Jaspers E, Overmann K (2004) Ecological significance of microdiversity: identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. Appl Environ Microbiol 70:4831–4839

Kent AD, Jones SE, Lauster GH, Graham JM, Newton RJ, McMahon KD (2006) Experimental manipulations of microbial food web interactions in a humic lake: shifting biological drivers of bacterial community structure. Environ Microbiol 8:1448–1459

Klappenbach JA, Dunbar JM, Schmidt TM (2000) rRNA operon copy number reflects ecological strategies of bacteria. Appl Environ Microbiol 66:1328–1333

Lan R, Reeves PR (1998) Recombination between rRNA operons created most of the ribotype variation observed in the seventh pandemic clone of *Vibrio cholerae*. Microbiology 144: 1213–1221

Liao D (2000) Gene conversion drives within genic sequences: concerted evolution of ribosomal RNA genes in Bacteria and Archaea. J Mol Evol 51:305–317

Lovett ST, Gluckman TJ, Simon PJ, Sutera VA Jr, Drapkin PT (1994) Recombination between repeats in *Escherichia coli* by a *recA*-independent, proximity-sensitive mechanism. Mol Gen Genet 245:294–300

Lutzoni F, Wagner P, Reeb V, Zoller S (2000) Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses without violating positional homology. Syst Biol 49:628–651

Luz SP, Rodriguez-Valera F, Lan R, Reeves PS (1998) Variation of the ribosomal operon 16S-23S gene spacer region in representatives of *Salmonella enterica* subspecies. J Bacteriol 180: 2144–2151

Maddison DR, Maddison WP (2000) MacClade 4: Analysis of phylogeny and character evolution. Sinauer Associates, Sunderland, MA

Mattatall NR, Sanderson KE (1996) *Salmonella typhimurium* LT2 possesses three distinct 23S rRNA intervening sequences. J Bacteriol 178:2272–2278

Milyutina IA, Bobrova VK, Matveeva EV, Schaad NW, Troitsky AV (2004) Intragenomic heterogeneity of the 16S rRNA-23S rRNA internal transcribed spacer among *Pseudomonas syringae* and *Pseudomonas fluorescens* strains. FEMS Microbiol Lett 239:17–23

Moran MA, Buchan A, Gonzalez JM, Heidelberg JF, Whitman WB, Kiene RP, Henriksen JR, King GM, Belas R, Fuqua C, Brinkac L, Lewis M, Johri S, Weaver B, Pai G, Eisen JA, Rahe E, Sheldon WM, Ye WY, Miller TR, Carlton J, Rasko DA, Paulsen IT, Ren QH, Daugherty SC, Deboy RT, Dodson RJ, Durkin AS, Madupu R, Nelson WC, Sullivan SA, Rosovitz MJ, Haft DH, Selengut J, Ward N (2004) Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. Nature 432:910–913

Nagpal ML, Fox KF, Fox A (1998) Utility of 16S-23S rRNA spacer region methodology: How similar are interspace regions within a genome and between strains for closely related organisms? J Microbiol Meth 33:211–219

Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. Annu Rev Genet 39:121–152

Osorio CR, Collins MD, Romalde JL, Toranzo AE (2005) Variation in 16S-23S rRNA intergenic spacer regions in *Photobacterium damselae*: a mosaic-like structure. Appl Environ Microbiol 71:636–645

Palenik B, Brahamsha B, Larimer FW, Land M, Hauser L, Chain P, Lamerdin J, Regala W, Allen EE, McCarren J, Paulsen I, Dufresne A, Partensky F, Webb EA, Waterbury J (2003) The genome of a motile marine *Synechococcus*. Nature 424:1037–1042

Pearce JM (2006) Minding the gap: frequency of indels in mtDNA control region sequence data and influence on population genetic analyses. Mol Ecol 15:333–341

Posada D, Crandall KA, Holmes EC (2002) Recombination in evolutionary genomics. Annu Rev Genet 36:75–97

Privitera A, Rappazzo G, Sangari P, Giannino V, Licciardello L, Stefani S (1998) Cloning and sequencing of a 16S/23S ribosomal spacer from *Haemophilus parainfluenzae* reveals an invariant, mosaic-like organization of sequence blocks. FEMS Microbiol Lett 164:289–294

Ranjard L, Brothier E, Nazaret S (2000) Sequencing bands of ribosomal intergenic spacer analysis fingerprints for characterization and microscale distribution of soil bacterium populations responding to mercury spiking. Appl Environ Microbiol 66:5334–5339

Rocap G, Distel DL, Waterbury JB, Chisholm SW (2002) Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. Appl Environ Microbiol 68:1180–1191

Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman M, Hauser L, Hess WR, Johnson ZI, Land M, Lindell D, Post AF, Regala W, Shah M, Shaw SL, Steglich C, Sullivan MB, Ting CS, Tolonen A, Webb EA, Zinser ER, Chisholm SW (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424:1042–1047

Rooney AP, Ward TJ (2005) Evolution of a large ribosomal RNA multigene family in filamentous fungi: birth and death of a concerted evolution paradigm. Proc Natl Acad Sci USA 102:5084–5089

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496–2497

Santoyo G, Romero D (2005) Gene conversion and concerted evolution in bacterial genomes. FEMS Microbiol Rev 29:169–183

Santoyo G, Martínez-Salazar JM, Rodríguez C, Romero D (2005) Gene conversion tracts associated with crossovers in *Rhizobium etli*. J Bacteriol 187:4116–4126

Schloter M, Lebuhn M, Heulin T, Hartmann A (2000) Ecology and evolution of bacterial microdiversity. FEMS Microbiol Rev 24:647–660

Segall AM, Roth JR (1994) Approaches to half-tetrad analysis in bacteria: recombination between repeated, inverse-order chromosomal sequences. Genetics 136:27–39

Simmons MP, Ochoterena H (2000) Gaps as characters in sequence-based phylogenetic analyses. Syst Biol 49:369–381

Swofford DL (2003) PAUP* Phylogenetic Analysis Using Parsimony (*and other methods). Sinauer Associates, Sunderland, MA

Teyssier C, Marchandin H, Siméon De Buochberg M, Ramuz M, Jumas-Bilak E (2003) Atypical 16S rRNA gene copies in *Ochrobactrum intermedium* strains reveal a large genomic rearrangement by recombination between *rrn* copies. J Bacteriol 185:2901–2909

Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, Benoit J, Sarma-Rupavtarm R, Distel DL, Polz MF (2005)

Genotypic diversity within a natural coastal bacterioplankton population. Science 307:1311–1313

Vogel J, Normand P, Thiolouse J, Nesme X, Grundmann GL (2003) Relationship between spatial and genetic distance in *Agrobacterium* spp in 1 cubic centimeter of soil. Appl Environ Microbiol 69:1482–1487

Wall JD (2000) A comparison of estimators of the population recombination rate. Mol Biol Evol 17:156–163

Walsh JB (1987) Sequence-dependent gene conversion: Can duplicated genes diverge fast enough to escape conversion? Genetics 117:543–557

Weider LJ, Elser JJ, Crease TJ, Mateos M, Cotner JB, Markow TA (2005) The functional significance of ribosomal (r)DNA variation: impacts on the evolutionary ecology of organisms. Annu Rev Ecol Evol Syst 36:219–242

Wenner T, Roth V, Decaris B, Leblond P (2002) Intragenomic and intraspecific polymorphism of the 16S-23S rRNA internally transcribed sequences of *Streptomyces ambofaciens*. Microbiology 148:633–642

Zar JH (1999) Biostatistical Analysis, 4th ed. Prentice Hall, Upper Saddle River, NJ