Comparative genomics of prokaryotes with special reference to horizontal gene transfer, and its evolutionary implication

Yoji Nakamura

DOCTOR OF PHILOSOPHY

Department of Genetics
School of Life Science
The Graduate University for Advanced Studies

2002

Contents

Abstract

Until several years ago, it had been believed that horizontal gene transfer in the prokaryotic kingdom was a rare or restricted event in the evolutionary history. However, the progress of genome projects on a worldwide scale revealed that prokaryotic genomes have frequently undergone massive horizontal gene transfer as well as extensive genome rearrangements. Now, the paradigm has been drastically shifted, where horizontal gene transfer has been recognized as a major factor of prokaryotic evolution. In fact, horizontal gene transfer is more prompt procedure for prokaryotic organisms to acquire some metabolic traits rather than mutation of preexistent genes in the organisms. The purpose of this study is to reveal the evolutionary process in prokaryotic genomes, focusing mainly on horizontal gene transfer. I performed computational analyses using a large amount of complete genome sequences.

First, I developed a novel method for effectively detecting horizontally transferred genes. My method is based on Bayes' estimation and training models (Markov models), and the principle is to evaluate the posterior probabilities that query gene sequences are intrinsic in the genome. Using this method, I estimated that about 12% of all genes in 84 prokaryotic complete genomes examined may have been acquired by the recent gene transfer. I have successfully detected 867 clusters of transferred genes including 61 possible pathogenicity islands in 16 genomes. Interestingly, the genome comparisons between two different strains of *Neisseria meningitidis* and between two different species of *Xanthomonas* suggested that horizontal transfer of the large clusters were associated with genome rearrangement such as inversion in the genome. I have quantitatively shown that the functions of the transferred genes are mainly related to mobile elements, pathogenicity, cell surface structure, and some regulatory functions. Acquisition of cell surface structure genes may contribute to the cell defense against harmful chemical substances in the environment. Since genes of regulatory function include genes regulating transcriptions possibly by binding DNA, the acquisition of these genes may be able to alter gene expression network for adaptation under a variety of conditions. Moreover, the present method has shown a remarkable advantage in which donor species of transferred genes can be identified. As for the performance of this method, I compared the sensitivity and

specificity with those of Karlin's method, and the result has shown that this method is better. I have developed a database for horizontal gene transfer (HGT database) in collaboration with system engineers of Fujitsu Co., Ltd.

Second, as another approach to detect horizontally transferred genes, I conducted phylogenetic analysis on the following six taxonomic groups: (i) *Bacillus-Staphylococcus* group, (ii) *Lactococcus-Streptococcus* group, (iii) Gram-positive high GC% bacteria group, (iv) *Chlamydia* group, (v) Enterobacteria and its relatives group, (vi) *Rhizobium* group. For each group, I estimated the proportion of horizontal gene transfer by verifying the possible three topologies of four-OTU (Operational Taxonomic Unit) trees among four species. Phylogenetic trees for conserved genes among the all four species have shown the signature of inter-species gene exchange in (i) *Bacillus-Staphylococcus* group and (vi) *Rhizobium* group, but not in the other four taxonomic groups. In *Bacillus-Staphylococcus* group, the transposon-rich genome of *B. halodurans* might have enhanced the mobility of genes to and from other species. In *Rhizobium* group, self-transmittable plasmids might have made gene transfer easier. On the other hand, it was suggested that species-specific genes not conserved among the four species were frequently derived from distantly related species by horizontal transfer. The results suggested that inter-species gene exchange is caused not by homologous recombination of the organisms but by extra-chromosomal elements such as transposons that are often located on plasmids.

Lastly, I analyzed a complete genome of an amino acid producing bacterium, *Corynebacterium efficiens*, which is originally kept by Ajinomoto Co.,Ltd. and newly sequenced by National Institute of Technology and Evaluation (NITE). The approach based on Bayes' estimation was also useful for detecting horizontally transferred genes in the *C. efficiens* genome. *C. efficiens* is closely related to the other amino acid producing bacterium, *C. glutamicum*, but *C. efficiens* can produce amino acids at higher temperature than *C. glutamicum,* meaning that the enzymes required for metabolic reaction are thermostable in *C. efficiens*. Moreover, *C. efficiens* has a higher GC content than *C. glutamicum* and another close relative *C. diphtheriae*. I found that the thermostability of *C. efficiens* is due to biased codon usage depending on the change of GC content in *C. efficiens*. I proposed that the loss of a mutator gene in *C. efficiens* is one of the factors that have affected the increase in the GC content in the species. In addition to that, I conducted comparisons of genome structures among the closely related species, and have found that *Corynebacterium* species have exceptionally a stable genome structure with regard to the order of orthologous genes. The comparison of the *Corynebacterium* genomes with the *Mycobacterium* one has implied that recombinational repair system is involved in the rarity of genome rearrangements in *Corynebacterium* species.

The results and discussion reported here will provide a stimulating implication about the evolution of prokaryotic genomes. My approaches are quite useful for a large amount of genome sequences.

1.Introduction

1.1 Genome sequencing project of prokaryotic species

In 1995, the complete genome sequence of *Haemophilus influenzae*, a respiratory pathogen infecting children and classified into gamma-proteobacteria, was determined by the Institute of Genome Research (TIGR) (**Fleischmann *et al.* 1995**). This is the first endeavour to sequence the whole genome sequence of an organism in the living world of the prokaryote and eukaryote. In the same year, Fraser et al. (**Fraser *et al.* 1995**) sequenced *Mycoplasma genitalium* genome, which is the smallest genome in the published genomes known at present. Currently (as of Dec.1, 2002), 98 complete genome sequences including those of redundant species have been published in the public database, DDBJ/EMBL/Genbank (**Table 1.1**). The target species of genome projects are mainly (1) model organisms, (2) pathogens and (3) industrially useful bacterial strains. For example, two model organisms, *Escherichia coli* and *Bacillus subtilis*, were sequenced in 1997 (**Blattner *et al.*1997; Kunst *et al.*1997**). *Synechocystis* sp., a model organism for studying photosynthesis, was sequenced in 1996 (**Kaneko *et al.* 1996**). Most of sequenced species are pathogens as targets of disease treatment. It is said that the genome projects of about 350 species are currently ongoing in the world (**Supplemental table 1**). Apparently, the rate of sequencing completion is kept on accelerating.

1.2 The impact of genome sequencing

Until several years ago, it had been believed that genome structure was stable and that horizontal gene transfer and genome rearrangement were rare or restricted events in the evolutionary history.

At present, the outcomes of genome sequencing have revealed that prokaryotic genomes

Table 1.1 Published prokaryote genomes (as of Dec.1, 2002)

Species name*	Domain**	Genome size (Kb)	Institution	Date***	Publication	Authors
Corynebacterium efficiens YS-314T	B	3140	NITE, Ajinomoto Co., Inc	2002.11.15	Unpublished	------
Mycoplasma penetrans HF-2	B	1358	NIH-NET	2002.10.30	NAR, in press	Sasaki ***et al.***
Bifidobacterium longum NCC2705	B	2256	Nestle, Univ of Georgia	2002.10.29	PNAS, 99,14422-14427	Schell ***et al.***
Streptococcus mutans UA159	B	2030	Univ of Oklahoma, Ohio State Univ	2002.10.29	PNAS, 99,14434-14439	Ajdic ***et al.***
Wigglesworthia glossinidia	B	697	Yale Univ, Kitasato Univ, RIKEN	2002.10.24	Nature Genetics, 32,402-407	Akman ***et al.***
Leptospira interrogans serovar *lai* 56601	B	4691	Chinese National Human Genome Center at Shanghai	2002.10.21	Unpublished	------
Shigella flexneri 2a	B	4607	Microbial Genome Center, Beijing	2002.10.16	NAR, 30, 4432-4441	Jin ***et al.***
Shewanella oneidensis MR-1 ATCC700550	B	4969	TIGR	2002.10.7	Nature Biotechnology, 20, 1118-1123	Heidelberg ***et al.***
Brucella melitensis biovar *suis* 1330	B	3310	TIGR	2002.10.1	PNAS, 99, 13148-13153	Paulsen ***et al.***
Streptococcus agalactiae NEM316	B	2211	Institut Pasteur	2002.9.30	Mol Microbiol, 45, 1499-513	Glaser ***et al.***
Oceanobacillus iheyensis HTE831	B	3630	JAMSTEC	2002.9.7	NAR, 30, 3927-3935	Takami ***et al.***
Streptococcus agalactiae 2603V/R	B	2160	TIGR	2002.8.28	PNAS, 99, 12391-12396	Tettelin ***et al.***
Thermosynechococcus elongatus BP-1	B	2600	Kazusa DNA Research Institute	2002.8.19	DNA Res, 9, 123-30	Nakamura ***et al.***
Yersinia pestis KIM5 P12	B	4600	Univ of Wisconsin	2002.7.29	J. Bacteriol, 184, 4601-4611	Deng ***et al.***
Streptococcus pyogenes **MGAS315**	B	1900	RML-NIAID, Univ of Minnesota	2002.7.16	PNAS, 99, 10078-10083	Beres ***et al.***
Methanosarcina mazei **Goe1**	A	4096	Gottingen Genomics Laboratory, Integrated Genomics Inc	2002.7.10	J. Mol. Micro. Biotechnol., 4, 453-461	Deppenmeier ***et al.***
Chlorobium tepidum **TLS**	B	2154	TIGR	2002.7.9	PNAS, 99, 9509-9514	Eisen ***et al.***
Buchnera aphidicola **SG**	B	641	Univ of Uppsala	2002.6.28	Science, 296, 2376-2379	Tamas ***et al.***
Staphylococcus aureus **subsp.** ***aureus*** **MW2**	B	2820	NITE, Juntendo Univ	2002.5.25	Lancet, 359, 1819-1827	Baba ***et al.***
Xanthomonas axonopodis **pv.** ***citri*** **306**	B	5273	FAPESP, Univ of Sao Paulo, Univ of Campinas	2002.5.23	Nature, 417, 459-463	da Silva ***et al.***
Xanthomonas campestris **pv.*campestris*** **ATCC 33913**	B	5076	FAPESP, Univ of Sao Paulo	2002.5.23	Nature, 417, 459-463	da Silva ***et al.***
Streptomyces coelicolor **A3(2)**	B	8667	Sanger Institute, John Innes Centre, IGF	2002.5.9	Nature, 417, 141-147	Bentley ***et al.***
Thermoanaerobacter tengcongensis **MB4T**	B	2689	Beijing Genomics Institute, The Institute of Microbiology	2002.5.7	Genome Res., 5, 689-700	Bao ***et al.***
Fusobacterium nucleatum **ATCC 25586**	B	2170	Integrated Genomics Inc	2002.4.10	J Bacteriol, 184, 2005-2018	Kapatral ***et al.***
Methanosarcina acetivorans **C2A**	A	5751	Whitehead Inst, Univ of Illinois at Urbana-Champaign	2002.4.10	Genome Res., 12, 532-542	Galagan ***et al.***
Streptococcus pyogenes **MGAS8232**	B	1895	RML-NIAID, Univ of Minnesota	2002.4.2	PNAS, 99, 4668-4673	Smoot ***et al.***
Methanopyrus kandleri **AV19**	A	1694	Fidelity Systems, Inc	2002.4.2	PNAS, 99, 4644-4649	Slesarev ***et al.***
Corynebacterium glutamicum **ATCC-13032**	B	3309	Kyowa Hakko	2002.3.12	Unpublished	------
Pyrococcus abyssi **GE5**	A	1765	Genoscope	2002.2.13	Unpublished	------
Pyrococcus furiosus **DSM 3638**	A	1908	Univ of Utah, Univ of Maryland	2002.2.12	Meth. Enzymol., 330:134-57	Robb ***et al.***
Ralstonia solanacearum **GMI1000**	B	5810	Genoscope, INRA, CNRS	2002.1.31	Nature, 415,497-502	Salanoubat ***et al.***
Clostridium perfringens **13**	B	3031	Univ of Tsukuba, Kyushu Univ, Kitasato Univ	2002.1.22	PNAS, 99, 996-1001	Shimizu ***et al.***
Pyrobaculum aerophilum **IM2**	A	2222	CalTech, UCLA	2002.1.22	PNAS, 99,984-989	Fitz-Gibbon ***et al.***
Brucella melitensis **16M**	B	3294	Univ of Scranton, Integrated Genomics Inc	2002.1.8	PNAS, 99,443-448	DelVecchio ***et al.***
Agrobacterium tumefaciens **C58-DuPont**	B	4915	Univ of Washington, DuPont, Univ of Campinas	2001.12.14	Science, 294,2317-2323	Wood ***et al.***
Agrobacterium tumefaciens **C58-Cereon**	B	4915	Cereon Genomics, Univ of Richmond, Monsanto	2001.12.14	Science, 294,2323-2328	Goodner ***et al.***
Nostoc (Anabaena) **sp. PCC 7120**	B	6413	Kazusa DNA Research Institute, Michigan State Univ	2001.10.31	DNA Res., 8,205-213	Kaneko ***et al.***

Listeria monocytogenes EGD-e	B	2944	EC Concortium	2001.10.26	Science, 294,849-852	Glaser ***et al.***
Listeria innocua Clip11262	B	3011	Institut Pasteur	2001.10.26	Science, 294,849-852	Glaser ***et al.***
Salmonella typhimurium LT2 SGSC1412	B	4857	Washington Univ	2001.10.25	Nature, 413,852-856	McClelland ***et al.***
Salmonella typhi CT18	B	4809	Sanger Institute, Imperial College	2001.10.25	Nature, 413,848-852	Parkhill ***et al.***
Streptococcus pneumoniae R6	B	2038	Eli Lilly	2001.10.10	J Bacteriol., 183,5709-5717	Hoskins ***et al.***
Yersinia pestis CO-92 (Biovar Orientalis)	B	4653	Sanger Institute, MDS, Imperial College, DSTL	2001.10.4	Nature, 413,523-527	Parkhill ***et al.***
Mycobacterium tuberculosis CDC1551	B	4403	TIGR	2001.10.2	J Bacteriol, 184, 5479-90	Fleischmann ***et al.***
Rickettsia conorii Malish 7	B	1268	Genoscope	2001.9.14	Science, 293,2093-2098	Ogata ***et al.***
Sulfolobus tokodaii 7	A	2694	NITE	2001.8.31	DNA Res., 8,123-40	Kawarabayasi ***et al.***
Clostridium acetobutylicum ATCC 824D	B	4100	Genome Therapeutics	2001.8.10	J.Bacteriol., 183,4823-4838	Nolling ***et al.***
Sinorhizobium meliloti 1021	B	6690	European Union, Stanford Univ	2001.7.27	Science, 293,668-672	Galibert ***et al.***
Streptococcus pneumoniae TIGR4 ATCC-BAA-334	B	2160	TIGR	2001.7.20	Science, 293,498-506	Tettelin ***et al.***
Sulfolobus solfataricus P2	A	2992	European Union, Canadian Bioinformatics Resource	2001.7.3	PNAS, 98,7835-7840	She ***et al.***
Caulobacter crescentus CB15	B	4016	TIGR	2001.5.22	PNAS, 98,4136-4141	Nierman ***et al.***
Mycoplasma pulmonis UAB CTIP	B	963	Genoscope	2001.5.15	NAR, 29, 2145-2153	Chambaud ***et al.***
Lactococcus lactis subsp. ***lactis*** IL1403	B	2365	Genoscope, INRA	2001.5.10	Genome Research, 11, 731-753	Bolotin ***et al.***
Staphylococcus aureus Mu50 (VRSA)	B	2878	NITE, Juntendo Univ, Univ of Tsukuba, et al.	2001.4.21	The Lancet, 357, 1225-1240	Kuroda ***et al.***
Staphylococcus aureus N315 (MRSA)	B	2813	NITE, Juntendo Univ, Univ of Tsukuba, et al.	2001.4.21	The Lancet, 357,1225-1240	Kuroda ***et al.***
Streptococcus pyogenes SF370 (M1)	B	1852	Univ of Oklahoma	2001.4.10	PNAS, 98,4658-4663	Ferretti ***et al.***
Pasteurella multocida Pm70	B	2250	Univ of Minnesota	2001.3.13	PNAS, 98, 3460-3465	May ***et al.***
Escherichia col O157:H7. RIMD 0509952	B	5594	Japanese Consortium	2001.2.27	DNA Research, 8, 11-22	Hayashi ***et al.***
Mycobacterium leprae TN	B	3268	Sanger Institute, Institut Pasteur	2001.2.22	Nature, 409, 1007-1011	Cole ***et al.***
Escherichia coli O157:H7 EDL933	B	4100	Univ of Wisconsin	2001.1.25	Nature, 409,529-533	Perna ***et al.***
Thermoplasma volcanium GSS1	A	1584	AIST	2000.12.19	PNAS, 97, 14257-14262	Kawashima ***et al.***
Mesorhizobium loti MAFF303099	B	7596	Kazusa DNA Research Institute	2000.12.10	DNA Research, 7, 331-338	Kaneko ***et al.***
Halobacterium sp. NRC-1	A	2014	Univ of Washington- Seattle, Univ of Massachusetts	2000.10.24	PNAS, 97, 12176-12181	Ng ***et al.***
Ureaplasma urealyticum serovar 3	B	751	Univ of Alabama, Eli Lilly, Perkin-Elmer	2000.10.12	Nature, 407, 757-762	Glass ***et al.***
Pseudomonas aeruginosa PAO1	B	6264	Chiron, Univ of Washington- Seattle	2000.9.30	Nature, 406,959-964	Stover ***et al.***
Thermoplasma acidophilum DSM 1728	A	1564	Max-Planck-Institute for Biochemistry, Medigenomix	2000.9.28	Nature, 407, 508-513	Ruepp ***et al.***
Buchnera aphidicola AP	B	640	Univ of Tokyo, RIKEN	2000.9.7	Nature, 407, 81-86	Shigenobu ***et al.***
Vibrio cholerae serotype O1, strain N16961	B	4000	TIGR	2000.8.3	Nature, 406,477-483	Heidelberg ***et al.***
Xylella fastidiosa CVC 8.1.b clone 9.a.5.c	B	2679	ONSA	2000.7.13	Nature, 406,151-157	Simpson ***et al.***
Chlamydophila pneumoniae J138	B	1228	Yamaguchi Univ, Kyushu Univ	2000.6.15	NAR, 28,2311-2314	Shirai ***et al.***
Bacillus halodurans C-125	B	4202	Japan Marine Science and Technology Center	2000.4.10	Extremophiles, 4, 99-108	Takami ***et al.***
Neisseria meningitidis Z2491 (serogroup A)	B	2184	Sanger Institute, Univ of Oxford, Max-Planck-Berlin	2000.3.30	Nature, 404,502-506	Parkhill ***et al.***
Chlamydia trachomatis MoPn / Nigg	B	1069	TIGR, Univ of Manitoba	2000.3.15	NAR, 28,1397-1406	Read ***et al.***
Chlamydia pneumoniae AR39	B	1229	TIGR, Univ of Manitoba	2000.3.15	NAR, 28,1397-1406	Read ***et al.***
Neisseria meningitidis MC58 (serogroup B)	B	2272	TIGR	2000.3.10	Science, 287,1809-1815	Tettelin ***et al.***
Campylobacter jejuni subsp. ***jejuni*** NCTC 11168	B	1641	Sanger Institute, LSHTM, Univ of Leicester	2000.2.10	Nature, 403,665-668	Parkhill ***et al.***
Deinococcus radiodurans R1	B	3284	TIGR	1999.11.19	Science, 286,1571-1577	White ***et al.***
Thermotoga maritima MSB8	B	1860	TIGR	1999.5.27	Nature, 399,323-329	Nelson ***et al.***
Aeropyrum pernix K1	A	1669	NITE	1999.4.30	DNA Research, 6,83-101	Kawarabayasi ***et al.***
Chlamydophila pneumoniae CWL029	B	1230	Stanford Univ, Univ of California, Berkeley	1999.4.10	Nat Genet, 21,385-389	Kalman ***et al.***
Helicobacter pylori J99	B	1643	Genome Therapeutics, Astra	1999.1.14	Nature, 397,176-180	Alm ***et al.***
Rickettsia prowazekii Madrid E	B	1111	Univ of Uppsala	1998.11.12	Nature, 396,133-140	Andersson ***et al.***

Chlamydia trachomatis D/UW-3/CX (serovar D)	B	1042	Stanford Univ, Univ of California, Berkeley	1998.10.23	Science, 282,754-759	Stephens *et al.*
Treponema pallidum subsp. ***pallidum*** Nichols	B	1138	Univ of Texas, TIGR, Baylor College of Medicine	1998.7.17	Science, 281,375-388	Fraser *et al.*
Mycobacterium tuberculosis H37Rv	B	4411	Sanger Institute	1998.6.11	Nature, 393,537-544	Cole *et al.*
Pyrococcus horikoshii OT3	A	1738	Univ of Tokyo, NITE	1998.4.30	DNA Research, 5,55-76	Kawarabayasi *et al.*
Aquifex aeolicus VF5	B	1551	Univ of Illinois at Urbana-Champaign, Diversa	1998.3.26	Nature, 392,353-358	Deckert *et al.*
Borrelia burgdorferi B31	B	1230	Brookhaven Natl Lab, TIGR	1997.12.11	Nature, 390,580-586	Fraser *et al.*
Archaeoglobus fulgidus DSM4304	A	2178	Univ of Illinois at Urbana-Champaign, TIGR	1997.11.27	Nature, 390,364-370	Klenk *et al.*
Bacillus subtilis 168	B	4214	European Consortium, Japanese Consortium	1997.11.20	Nature, 390,249-256	Kunst *et al.*
Methanobacterium thermoautotrophicum delta H	A	1751	Genome Therapeutics, Ohio State Univ	1997.11.10	J.Bacteriology, 179,7135-7155	Smith *et al.*
Escherichia coli K12 MG1655	B	4639	Univ of Wisconsin	1997.9.5	Science, 277,1453-1474	Blattner *et al.*
Helicobacter pylori 26695	B	1667	TIGR	1997.8.7	Nature, 388,539-547	Tomb *et al.*
Mycoplasma pneumoniae M129	B	816	Univ of Heidelberg	1996.11.15	NAR, 24,4420-4449	Himmelreich *et al.*
Methanococcus jannaschii DSM 2661	A	1664	TIGR, Univ of Illinois at Urbana-Champaign	1996.9.28	Science, 273,1058-1073	Bult *et al.*
Synechocystis sp. PCC6803	B	3573	Kazusa DNA Research Institute	1996.6.30	DNA Res, 3,109-136	Kaneko *et al.*
Mycoplasma genitalium G-37	B	580	TIGR	1995.10.20	Science, 270,397-403	Fraser *et al.*
Haemophilus influenzae KW20	B	1830	TIGR	1995.7.28	Science, 269,496-512	Fleischmann *et al.*

* Species in bold are used in this study.

**B: Bacteria, A: Archaea

***Date when the sequecne was published in databases.

contain signatures of extrinsic genes that were possibly introgressed by horizontal transfer (HT genes). It has been recognized that horizontal gene transfer has frequently occurred in the evolutionary history. For example, when the complete genome sequence of *Thermotoga maritima*, a thermophilic eubacteria, was determined, the gene annotation of this genome sequence revealed that about 24% of genes in the genome were very similar to archaebacterial genes, implying that horizontal gene transfer took place between organisms possibly sharing niches, thermophilic eubacteria and archaebacteria (**Nelson *et al.* 1999; Nesbo *et al.* 2001**). More drastic result was published when two genome sequences of *E.coli* strains, O157 and K12, were compared. An O157 strain had a thousand more genes than a strain K12, implying that the genome structure was recently changed by horizontal gene transfer as well as gene duplication or gene loss (**Perna *et al.* 2001**).

On the other hand, signatures of genome rearrangement have been observed from the genome comparisons between closely related species. When the comparison was made between two strains of *Helicobacter pylori*, a gastric pathogen that is thought to cause a stomach ulcer (**Alm *et al.* 1999a**), we were surprised at the finding that large regions of the genome were rearranged between the two strains. In general, gene order is conserved between closely related species (**Huynen & Bork 1998; Tamames 2001**), but chromosomal segments were frequently rearranged even if they are the same species, as observed in *H.pylori* strains. In fact, it has been reported that in the same genus, such as *Bacillus*, *Chlamydia*, *Mycobacterium*, and *Pyrococcus*, the genome segments are considerably shuffled (**Takami *et al.* 2000; Read *et al.* 2000; Tillier & Collins 2000; Maeder *et al.* 1999**).

Interestingly, horizontal transfer and genome rearrangement are sometimes associated to each other. In the case of *H. pylori* mentioned above, it has been shown that this species has evolutionarily unstable regions termed "plasticity zone" (**Alm *et al.* 1999b**), where frequent genome rearrangements had occurred. These regions show lower GC contents than the rest of the genome, implying that the regions have been originated from other species by horizontal transfer (**see the next section**). For another example, Denamur et al. (**Denamur *et al.* 2000**) have indicated that mismatch repair genes such as *mutS* and *mutL* have horizontally transferred among strains in *E.coli* population. They argued that losses and acquisition of mismatch repair genes must have affected recombination rate in the genome, because mismatch repair genes controlled the accuracy of pairing between homologous sequences. These examples suggest that horizontal gene transfer can become a trigger for genome rearrangement and affect the stability of genome structure.

1.3 Purpose of the present study

1.3.1How prokaryotes have evolved?

It is now believed that horizontal gene transfer is one of the main factors to produce inter- and intra-species diversity (**Cruz & Davies 2000; Ochman *et al.* 2000; Lawrence 2002**). In addition, it has been recognized that a genome structure is unstable in the evolutionary history because of frequent genome rearrangements, and this instability is often related to horizontal gene transfer. Therefore, the purpose of my study is to illuminate the evolutionary process of prokaryote genomes, including genome rearrangement, from the viewpoint of horizontal gene transfer

The primary question about evaluating the significance of horizontal gene transfer is what amount of genes in the genome were heterogeneous origins. Now there are several methods for identifying horizontally transferred genes (HT genes) in a nucleotide sequence on the basis of the information about GC contents and codon usage bias. The principle underlying these methods is to find out nucleotide fragments that possess atypical features of base composition against the genome sequence. Since an extrinsic DNA segment recently inserted into a new host genome tends to keep the features that the original genome maintains, the segment is, in principle, distinguishable from the host genome sequences. For example, based on this principle, Lawrence and Ochman (**Lawrence & Ochman 1998**) estimated that the proportion of horizontally transferred genes in *E.coli* K-12 strain is about 17%. Karlin and his colleagues also detected HT genes in a variety of bacteria (**Mrazek & Karlin 1999; Mrazek *et al.* 2001; Karlin & Mrazek 2001; Karlin 2001**)

The second question is what kinds of genes are actually subject to horizontal transfer between prokaryotic taxa and to what degree of contribution to the functional differentiation in prokaryotic genomes the acquisition of novel genes has made. In general, it is thought that genes responsible for antibiotics-resistance, virulence, and some metabolic activity have undergone horizontal gene transfer (**Ochman *et al.* 2000**). Moreover, it is proposed that genes involved in transcription and translation are rarely transferred possibly because of their functional constraints (**Rivera *et al.* 1998; Jain *et al.* 1999**). Nesbo et al. (**Nesbo *et al.* 2001**) called this idea "core hypothesis" in which "core" means a set of nontransferrable genes. With relation to pathogenic bacteria, it is of great interest to note that a number of pathogenicity genes were acquired as large clusters possibly by horizontal gene transfer. These gene clusters (HT cluster), which often extend tens kb and exhibit atypical GC content, were termed "pathogenicity islands." (**Hacker & Kaper 2000**) Detection and elucidation of interspecific gene flux are thus thought to be important for drug discovery against frequent emergence of bacterial illness.

The third question is when and from where the HT gene came into the present species. It is difficult for the traditional methods to answer these questions. The most advantageous and established solution may be to conduct molecular phylogenetic analysis, but this analysis is limited to the case that homologous sequences are available enough to obtain reliable alignments.

1.3.2 Comparative analysis with special reference to horizontal gene transfer

In order to answer the questions raised in the previous section, we developed a novel and concise method for sensitively detecting HT genes as well as its possible donor(s) (**see Section 2.3**). I applied a simple statistics based on the Bayes' estimation from the training models: I just computed the posterior probability that query gene sequences are intrinsic in a genome sequence. To be more precise, for each gene sequence, I obtained the average probability by the window analysis, which I called "HT index" of the gene. Thus, genes having significantly lower indices were detected as the candidates of extrinsic HT genes. The training model was constructed according to the Markov chain model for each genome, and the statistical significance of HT index computed was evaluated by Monte Carlo simulation using parameters of the model. I estimated the proportion of HT genes and the donor species in 84 complete genome sequences published in the databases (**Table 1.1; The phylogenetic relationship is shown in Figure1.1.**). Subsequently, I inferred the functions of the HT genes extensively, and quantitatively estimated the proportion of HT genes in each functional category. Moreover, my method has shown the advantage that the donor species of transferred genes can be identified even though phylogenetic information cannot be obtained. I will discuss the usefulness of my method and the significance of horizontal gene transfer among the prokaryotic species. In addition, I constructed a horizontal gene transfer database (HGT database). I will also report an outline of this database.

1.3.3 Extensive phylogenetic analysis

As mentioned in **Section 1.3.1**, molecular phylogenetic analysis is also one of the most powerful methods to detect horizontally transferred genes and its possible donors. In principle, when a phylogenetic tree for a gene is inconsistent with a species tree, we can infer that horizontal gene transfer has occurred somewhere in the tree. The performance of this method is depending on enough homologous sequences that are assumed to have shared a common ancestor and on an alignment of good quality made from the sequences. By extensively parsing phylogenetic trees of orthologous genes, I estimated the proportion of horizontal gene transfer among closely related species.

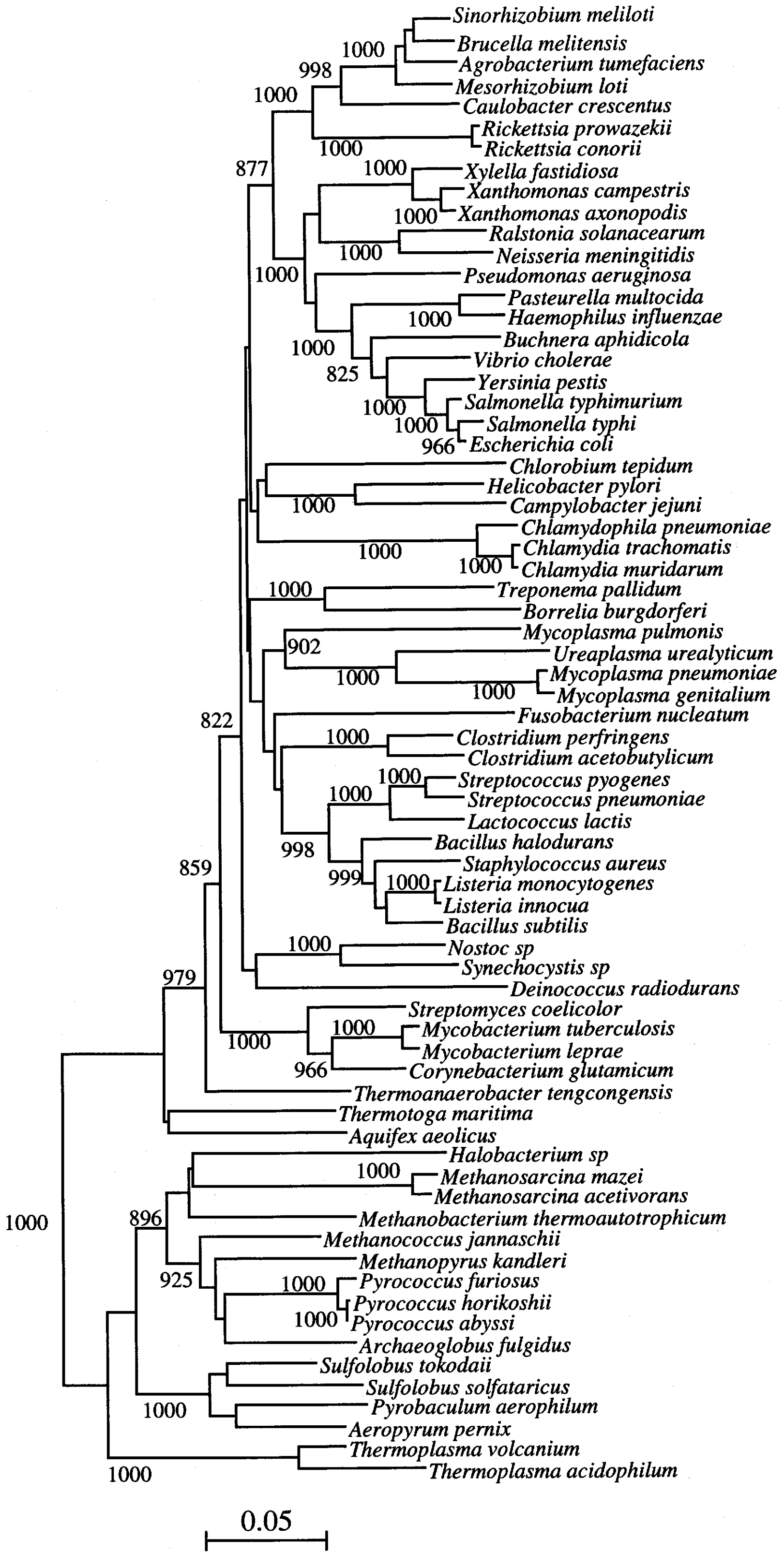


Figure 1.1 Phylogenetic tree for sequenced species using 16SrRNA sequences. The tree was reconstructed by neighbor-joining method. Numbers indicate 800 or more bootstrap values for 1000 replicates. Information of the alignment were obtained from European ribosomal RNA database (http://oberon.rug.ac.be:8080/rRNA/index.html)

1.3.4 Comparative genome analysis between *Corynebacterium* species

Recently Ajinomoto company has determined a complete genome sequence of *Corynebacterium efficiens*, a species of *Corynebacterium* genus (**in preparation**). The *Corynebacterium* species is a rod-shaped bacterium having a high GC content and classified into *Actinomyces*, an order of gram-positive bacteria, containing *Mycobacterium tuberculosis* and *Streptomyces coelicolor* (**Collins *et al.* 1986; Liebl 1991**). **Figure 1.2** shows the phylogenetic relationship among corynebacteria using mycobacteria as an outgroup. *C. efficiens* is very close to *Corynebacterium glutamicum*, sequenced by Kyowa Hakko (**Table 1.1**) and *Corynebacterium diphtheriae*, a causative agent of diphtheria, sequenced by the Sanger Institute (**Supplemental table 1**). In particular, both *C. efficiens* and *C. glutamicum* are industrial bacteria that produce a variety of amino acids such as glutamate and lysine by fermentation. However, there are some differences between these two species. A remarkable one is that *C. efficiens* can grow and produce glutamate at a higher temperature (40°C) than *C. glutamicum* (30°C) (**Fudou *et al.* 2002**). The thermostability of *C. efficiens* is a useful trait that can decrease the cost for cooling down the heat generated in amino acid fermentation. Another difference is that *C. efficiens* has a higher GC content than *C. glutamicum*. The difference in GC content was estimated to be 5% (**Fudou *et al.* 2002**).

These differences give us the motivation to demonstrate the mechanism of genomic evolution with regard to thermostabilisation, nucleotide substitutions and the relationship between both. Thus, I comparatively analyzed the *C. efficiens*, *C. glutamicum, C. diphtheriae,* and *M. tuberculosis* genomes. I will discuss the genome evolution of *Corynebacterium* species from the view points of nucleotide substitution, genome rearrangement and horizontal gene transfer.

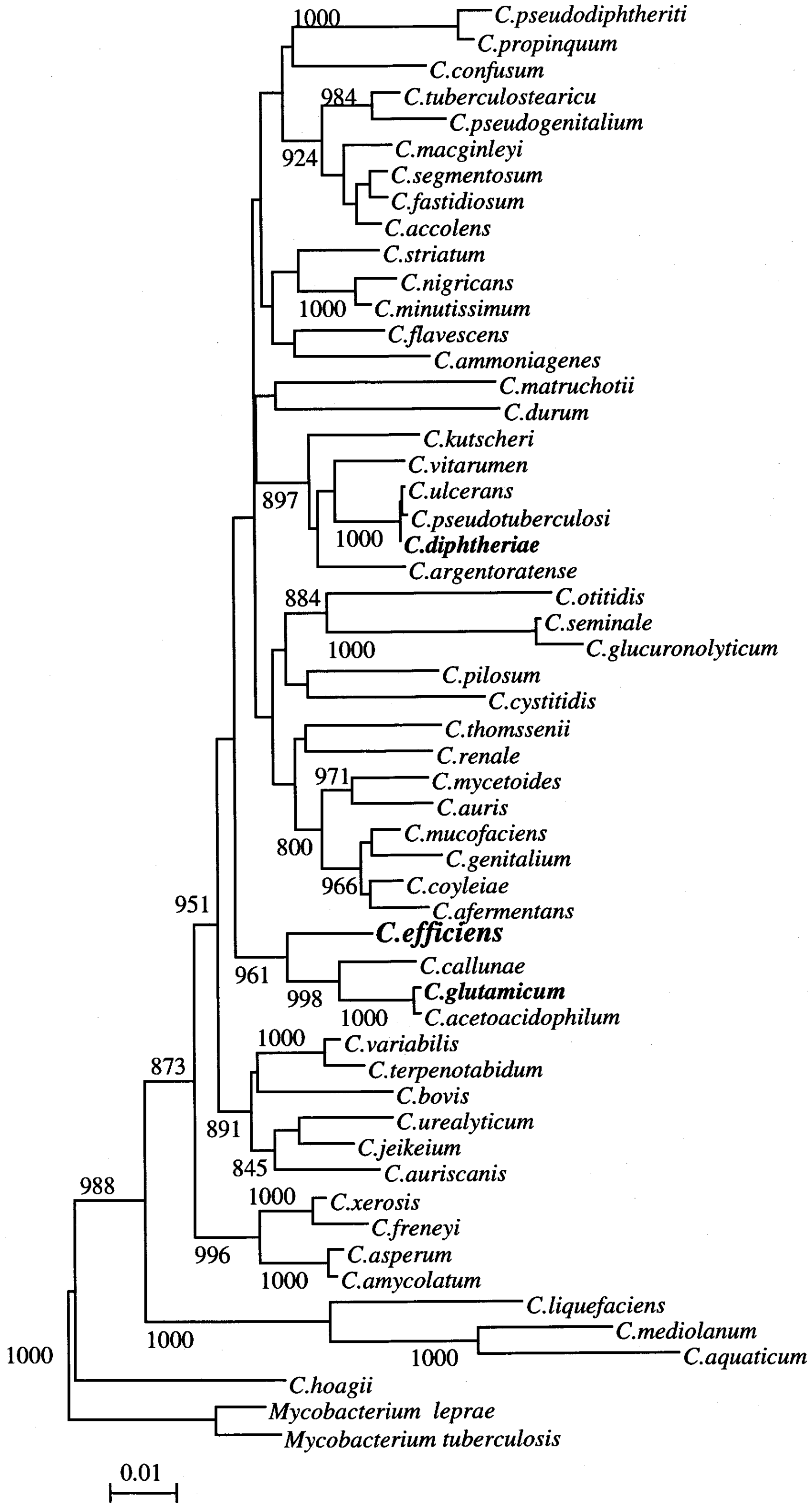


Figure 1.2 Phylogenetic tree of *Corynebacterium* 16SrRNA reconstructed by neighbor-joining method. Numbers indicate 800 or more bootstrap values for 1000 replicates. Information ofthe alignment were obtained from European ribosomal RNA database (http://oberon.rug.ac.be:8080/rRNA/index.html)

2. Materials and Methods

2.1 Overview

2.1.1 Markov chain model

Each nucleotide in a genome sequence is positioned non-randomly. In particular, a nucleotide sequence in a coding region is biased in the nucleotide composition, because a triplet of nucleotides encodes a amino acid. Therefore, I trained the nucleotide composition in coding regions and non-coding regions separately in a complete genome sequence by the Markov chain model. The parameters of this Markov model represent a species-specific nucleotide composition, and are useful for detecting genes that are not species-specific and thus regarded as being horizontally transferred.

2.1.2 Molecular phylogeny

Molecular phylogenetics by using DNA or protein sequences conserved among taxa is a reliable method for detecting horizontally transferred genes, when optimal alignments are obtained for homologous sequences. We can infer the occurrence of horizontal gene transfer by detecting a gene tree that is inconsistent with the species tree.

In particular, I firstly considered a simple situation in that I examine phylogenetic relationships among four genes, or, four operational taxonomy units (OTUs). There are three possible unrooted trees for four OTUs (**Figure 2.1**). Here one tree is correct and the other two are incorrect. Thus I estimate the proportion of horizontal gene transfer among four closely related species.

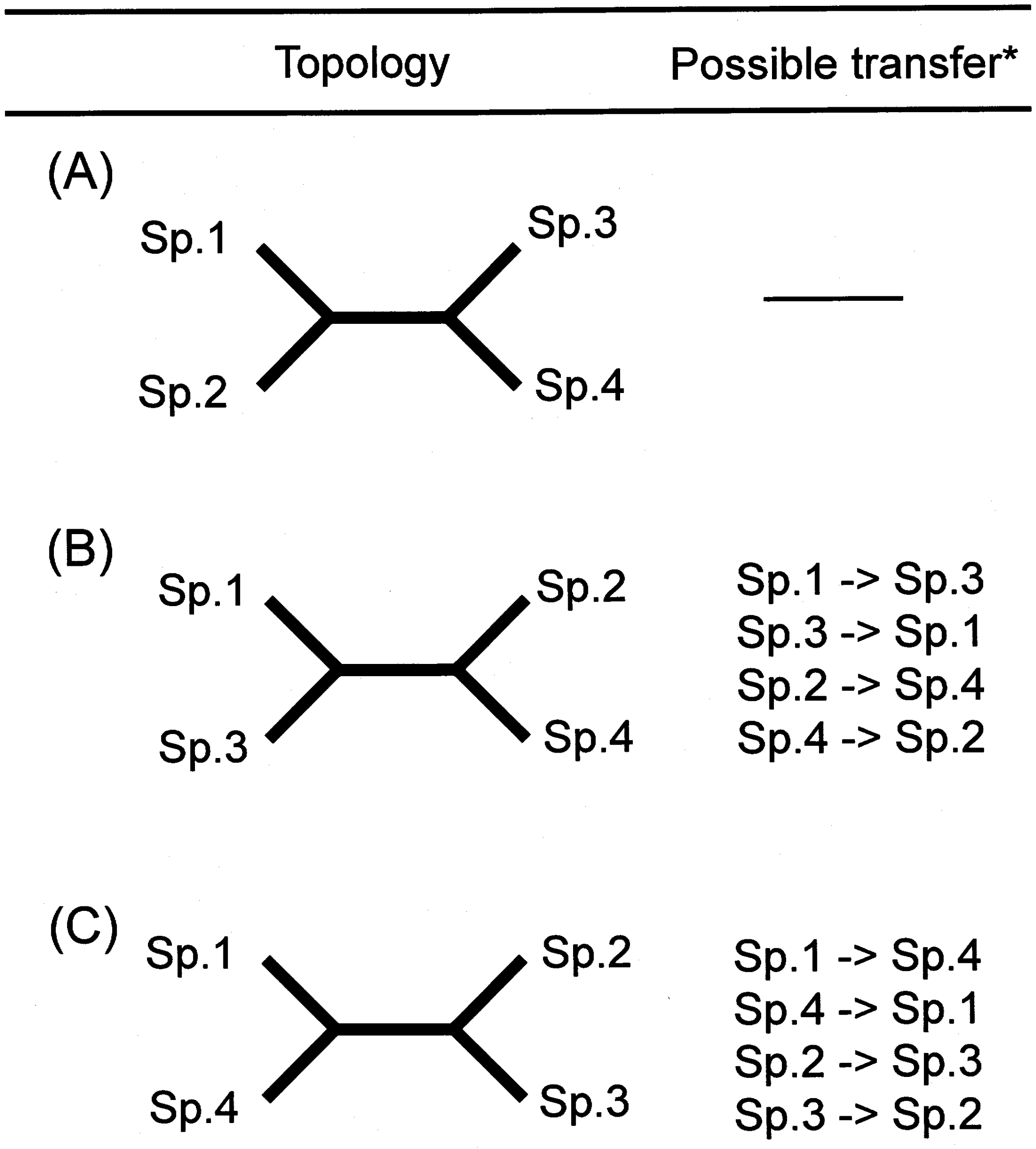


Figure 2.1 Three possible unrooted trees for four OTUs.

* Possible directions of horizontal gene transfer when (A) is a correct tree. Here I assumed that horizontal gene transfer has occured only once.

2.1.3 Genome comparison

Here, I analyzed GC contents, GC skews, codon usage patterns, orthologous gene orders, gene gains and losses by using the complete genome sequences of *Corynebacterium efficiens* and its relatives. The results from these analyses are useful for the elucidation of speciation mechanism. In particular, the change in GC content, GC skew, and codon usage pattern will explain mutational pressures operating on a genomes. Conservation of the orthologous gene order, pattern of gene gain and loss will reveal the stability of genome structures.

2.2 Materials

2.2.1 Genome sequences, plasmids, and bacteriophages

I retrieved the complete sequences of 84 prokaryote genomes, 284 plasmids and 110 bacteriophages from the DDBJ /EMBL/Genbank databases as of August 1, 2002. See **Table 1.1** and the following URLs:

http://gib.genes.nig.ac.jp/,

http://www.ebi.ac.uk/genomes/,

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome.

2.2.2 Functions of prokaryotic genes

I obtained the annotated gene sets in 79 prokaryotic species from the microbial database of TIGR (**Peterson 2001**), although I examined 84 species for horizontal gene transfer. The genes of the residual 5 species are not yet annotated by TIGR. In TIGR, they categorized the gene functions into the main roles and sub roles depending upon the classification adapted from Monica Riley (**Riley 1993**). One of the main role categories "Other category" is composed of three subrole categories "Plasmid functions", "Prophage functions" and "Transposon functions". On the other hand, there is a main role "Viral functions" in the database and this seems redundant with "Prophage functions", a sub role of "Other category". Therefore, here I conveniently united these "Other category" and "Viral functions" and redefined them as the category "Plasmid, phage, transposon functions". Furthermore, three main roles "Hypothetical proteins", "Unclassified", and "Unknown function" is summarized as "Unknown proteins". I excluded minor main-roles containing only less than 100 genes from my study. Finally, I obtained 16 main roles, 114 sub roles as shown in **Table 2.1**

Table 2.1 Functional gene categories based on the annotations of TIGR

Main Role*	Sub Role*
Amino acid biosynthesis	Aromatic amino acid family
	Aspartate family
	Glutamate family
	Histidine family
	Other
	Pyruvate family
	Serine family
Biosynthesis of cofactors, prosthetic groups, and carriers	Biotin
	Chlorophyll
	Folic acid
	Glutathione
	Heme, porphyrin, and cobalamin
	Lipoate
	Menaquinone and ubiquinone
	Molybdopterin
	Other
	Pantothenate and coenzyme A
	Pyridine nucleotides
	Pyridoxine
	Riboflavin, FMN, and FAD
	Thiamine
Cell envelope	Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides
	Biosynthesis of murein sacculus and peptidoglycan
	Other
	Surface structures
Cellular processes	Adaptations to atypical conditions
	Cell adhesion
	Cell division
	Chemotaxis and motility
	Conjugation
	DNA transformation
	Detoxification
	Other
	Pathogenesis
	Toxin production and resistance
Central intermediary metabolism	Amino sugars
	Nitrogen fixation
	Nitrogen metabolism
	One-carbon metabolism
	Other
	Phosphorus compounds
	Polyamine biosynthesis
	Sulfur metabolism
DNA metabolism	Chromosome-associated proteins
	DNA replication, recombination, and repair
	Degradation of DNA
	Other
	Restriction/modification
Energy metabolism	ATP-proton motive force interconversion
	Aerobic
	Amino acids and amines
	Anaerobic
	Biosynthesis and degradation of polysaccharides
	Chemoautotrophy
	Electron transport
	Entner-Doudoroff
	Fermentation
	Glycolysis/gluconeogenesis
	Methanogenesis
	Other
	Pentose phosphate pathway
	Photosynthesis
	Pyruvate dehydrogenase
	Sugars
	TCA cycle
Fatty acid and phospholipid metabolism	Biosynthesis
	Degradation
	Other

(continued)

Protein fate	Degradation of proteins, peptides, and glycopeptides
	Other
	Protein and peptide secretion and trafficking
	Protein folding and stabilization
	Protein modification and repair
Protein synthesis	Nucleoproteins
	Other
	Ribosomal proteins: synthesis and modification
	Translation factors
	tRNA aminoacylation
	tRNA and rRNA base modification
Purines, pyrimidines, nucleosides, and nucleotides	2'-Deoxyribonucleotide metabolism
	Nucleotide and nucleoside interconversions
	Other
	Purine ribonucleotide biosynthesis
	Pyrimidine ribonucleotide biosynthesis
	Salvage of nucleosides and nucleotides
	Sugar-nucleotide biosynthesis and conversions
Regulatory functions	DNA interactions
	Other
	Protein interactions
	RNA interactions
	Small molecule interactions
Transcription	DNA-dependent RNA polymerase
	Degradation of RNA
	Other
	RNA processing
	Transcription factors
Transport and binding proteins	Amino acids, peptides and amines
	Anions
	Carbohydrates, organic alcohols, and acids
	Cations
	Nucleosides, purines and pyrimidines
	Other
	Porins
	Unknown substrate
Plasmid, phage, transposon functions	Plasmid functions
	Prophage functions
	Transposon functions
	General
Unknown Proteins	Conserved
	Domain
	Not Conserved
	Role category not yet assigned
	Enzymes of unknown specificity
	General

* Role names are listed in alphabetical order.

2.2.3 *C.efficiens* and its relative genomes

The genome sequence of *C. efficiens* JCM 44549 (strain YS-314), which is a strain held by Ajinomoto Co.Ltd., was sequenced by National Institute of Technology and Evaluation (NITE). *C. glutamicum* ATCC 13032, which has been already sequenced by Kyowa Hakko, is the same as that used in detection of horizontal gene transfer (**Table 1.1**). *C. diphtheriae* NCTC 13129 was sequenced by the Sanger Institute, but the annotation of the genome is now ongoing (**Supplemental table 1**).

2.3 Methods 1 (Bayes' estimation)

2.3.1 Detection of horizontally transferred genes based on Bayes' estimation

First, I consider a nucleotide fragment denoted as F in the genome of a species. The posterior probability that F appears in the coding regions of a genome can be given by Bayes' theorem as follows:

$$\begin{aligned} \mathrm{P(coding|F)} &= \frac{\mathrm{P(F|coding)P(coding)}}{\mathrm{P(F)}} \\ &= \frac{\mathrm{P(F|coding)P(coding)}}{\mathrm{P(F|coding)P(coding) + P(F|non\text{-}coding)P(non\text{-}coding)}} \quad , \end{aligned} \tag{1}$$

where P(F|coding) and P(F|non-coding) are the probabilities that F is given from the coding and non-coding regions in a genome, respectively. The probability can be computed by constructing the training model of a species that represents the features of coding or non-coding sequences (**see below**). P(coding), P(non-coding) and P(F) are prior probabilities of coding sequences, non-coding sequences, and F, respectively. The denominator is obtained from the assumption that F is a coding or a non-coding fragment of the genome ($\mathrm{P(F) = P(F \cap coding) + P(F \cap non\text{-}coding)}$).

In the present study, I used P(coding|F) as an indicator of horizontal transfer for F, because this probability well qualifies that F is intrinsic in the species.

2.3.2 Construction of a training model based on Markov chains

In order to construct a training model, I primarily extracted coding and non-coding sequences from the complete genome sequence according to the database annotations (**Figure 2.2** (A)). I excluded tRNA and rRNA genes and annotated pseudogenes from my analysis. By

By using these coding and non-coding sequences, I prepared the training model of the species composed of two Markov chain models, separately for both regions, where the parameters (initiation/transition probabilities) were obtained by the maximum likelihood estimation of nucleotide frequencies in each region (**Figure 2.2 (B)**). To be more precise, since there are six possible reading frames for a coding sequence, the Markov chain for coding sequences is composed of six parameter sets (**Pcm (m=1,2,3,4,5,6) in Figure 2.2 (C)**, where the case of m=1 is assumed to be the true reading frame). Thus, P(F|coding)P(coding) in the equation **(1)** are rewritten as P(F|COD1)P(COD1) in the numerator and the sum of P(F|CODm)P(CODm) (m=1,2,3,4,5,6) in the denominator, respectively.

2.3.3 Computation of posterior probability and HT index

Finally, the equation (1) is rewritten as follows :

$$P(\mathrm{COD}_1 \mid F) = \frac{P(F \mid \mathrm{COD}_1)\, P(\mathrm{COD}_1)}{\sum_{m=1}^{6} P(F \mid \mathrm{COD}_m)\, P(\mathrm{COD}_m) + P(F \mid \mathrm{NON})\, P(\mathrm{NON})}$$

(m = 1, 2, 3, 4, 5, 6).

(2)

Here, P(CODm|F) is the posterior probability that F is the coding sequence of the m-th reading frame. The conditional probabilities, P(F|CODm) and P(F|NON), are those that F is given from the m-th frame coding region and the non-coding region, respectively. The prior probabilities, P(CODm) and P(NON), are assumed to be 1/12 and 1/2, respectively (P(COD1) + ...+ P(COD6) + P(NON) = 1). This algorithm is based on the study by Borodovsky and McIninch (**Borodovsky & McIninch 1993**).

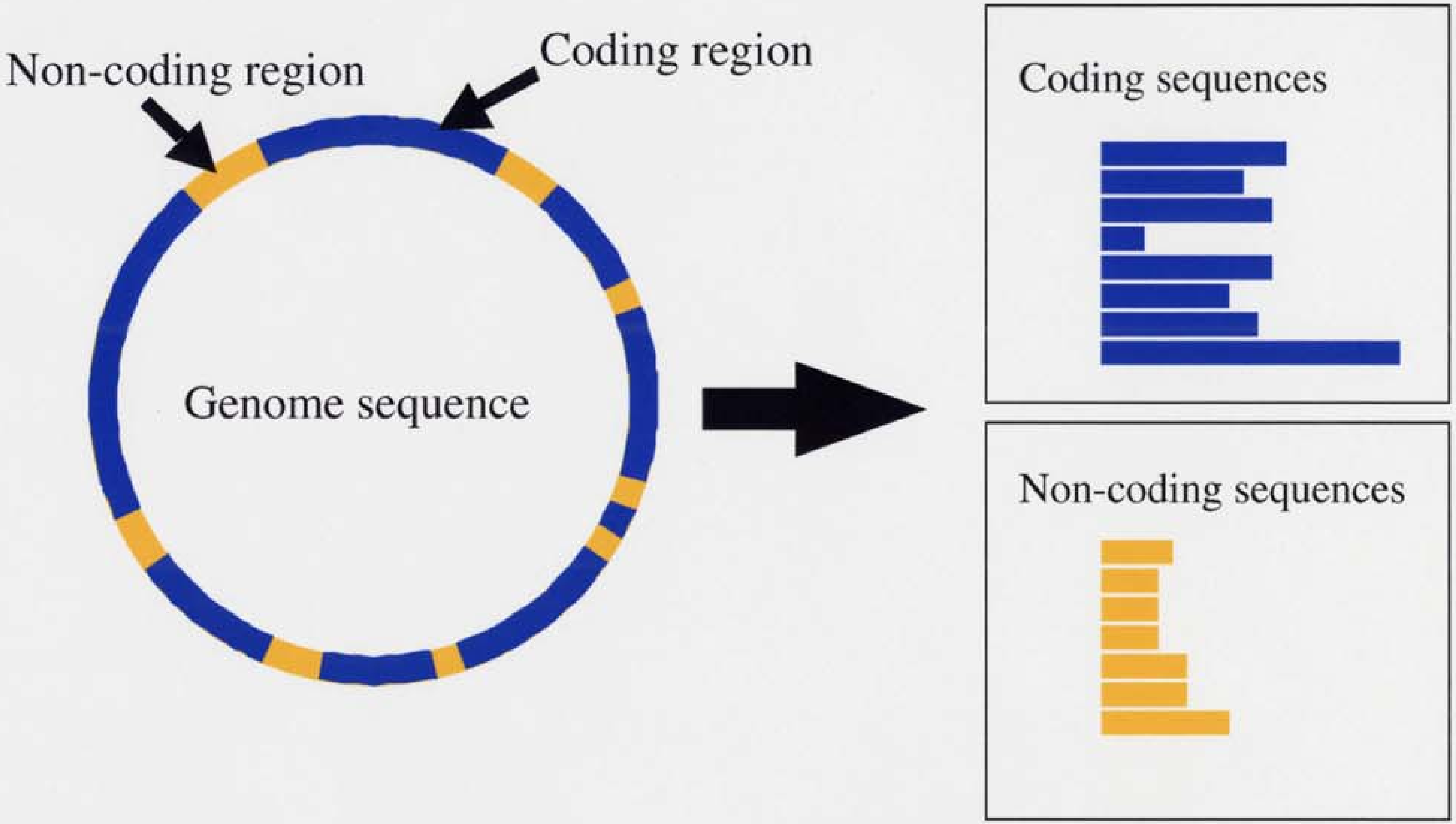
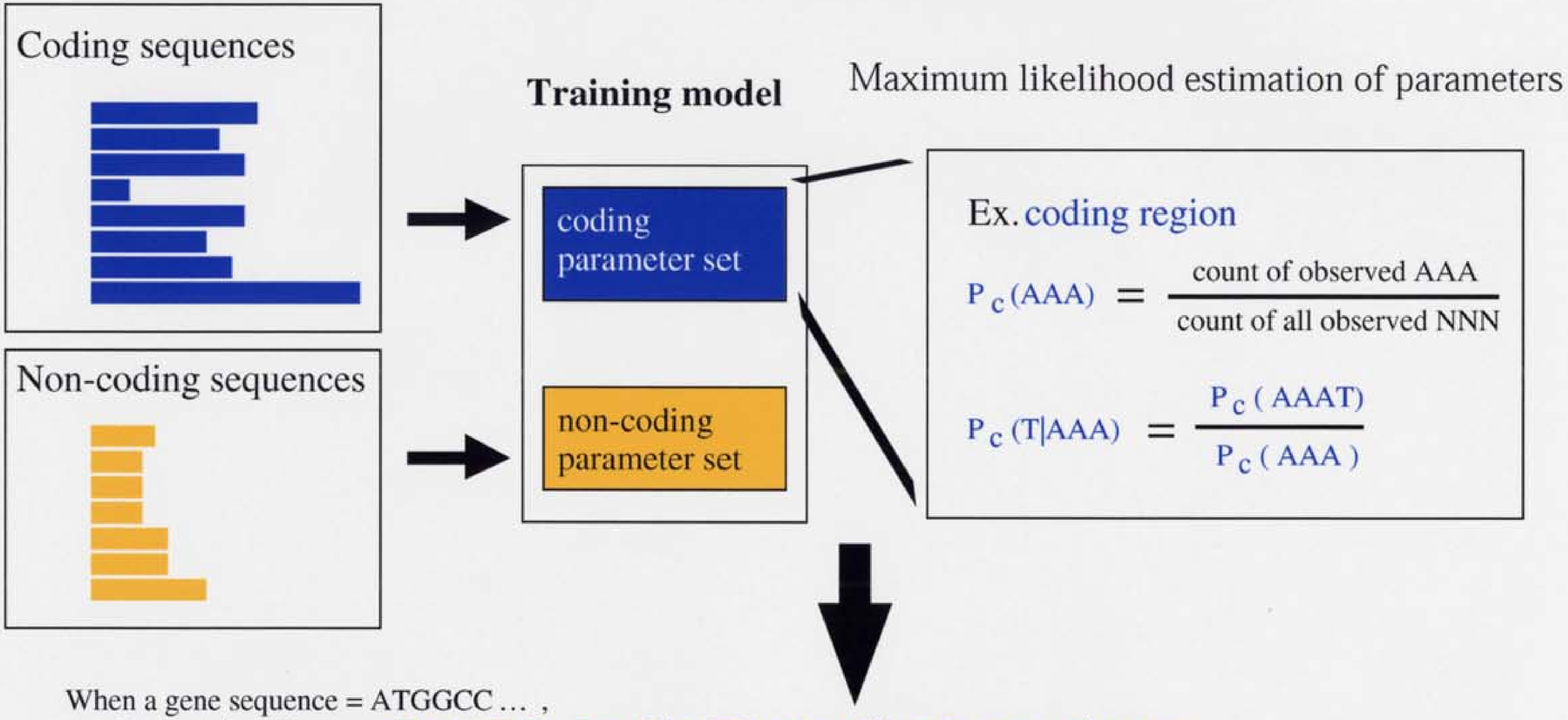


Figure 2.2 (A) Construction of training model (1) (Extraction of coding and non-coding sequences)



Coding sequences

Non-coding sequences

Training model

coding parameter set

non-coding parameter set

Maximum likelihood estimation of parameters

Ex. coding region

$$P_c(\text{AAA}) = \frac{\text{count of observed AAA}}{\text{count of all observed NNN}}$$

$$P_c(\text{T}|\text{AAA}) = \frac{P_c(\text{AAAT})}{P_c(\text{AAA})}$$

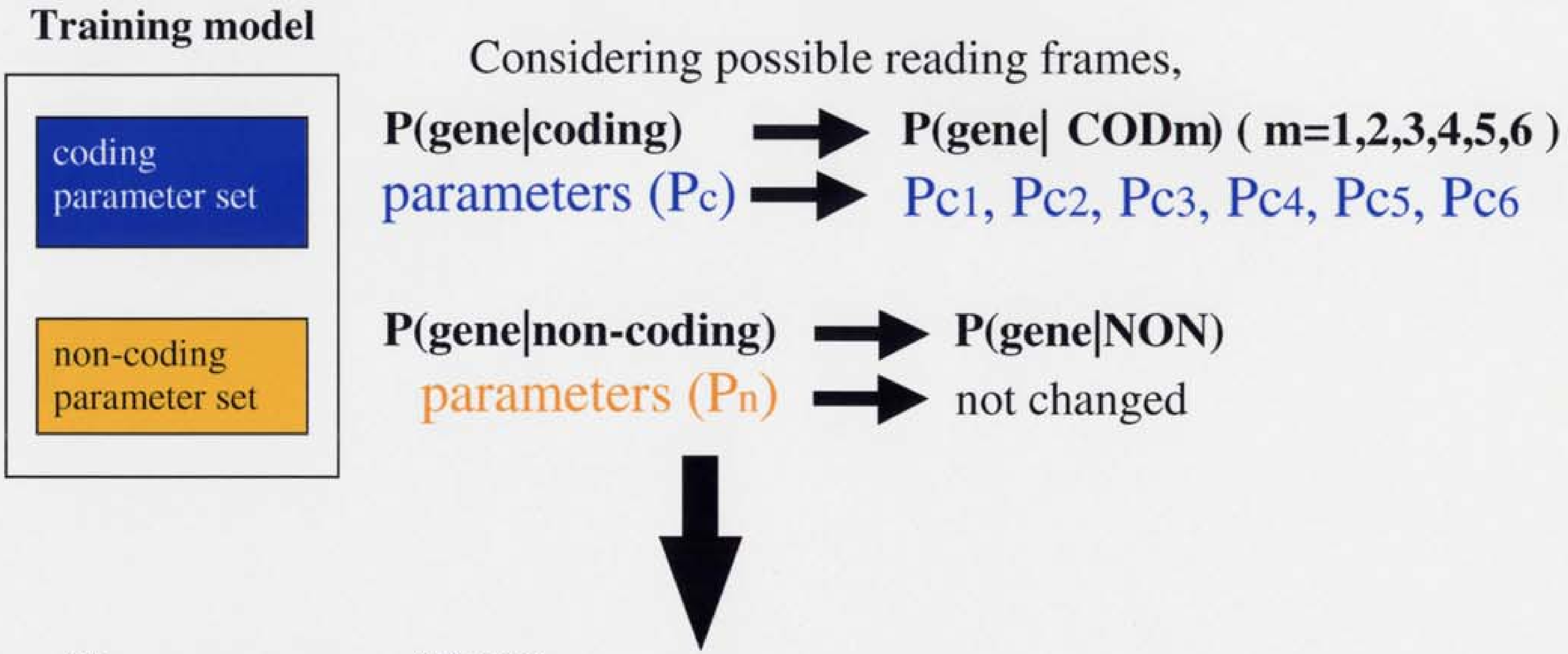
When a gene sequence = ATGGCC … ,

$$P(\text{gene}|\text{coding}) = Pc\,(\text{ATG})\ \ Pc\,(\text{G}|\text{ATG})\ \ Pc\,(\text{C}|\text{TGG})\ \ Pc\,(\text{C}|\text{GGC}) \ldots$$

$$P(\text{gene}|\text{non-coding}) = Pn\,(\text{ATG})\ \ Pn\,(\text{G}|\text{ATG})\ \ Pn\,(\text{C}|\text{TGG})\ \ Pn\,(\text{C}|\text{GGC}) \ldots$$

(3 rd order Markov chain)

Figure 2.2 (B) Construction of training model (2) (Computation of P(gene|coding) and P(gene|non-coding))



Training model

coding parameter set

non-coding parameter set

Considering possible reading frames,

P(gene|coding) → **P(gene| CODm) (m=1,2,3,4,5,6)**

parameters (P_c) → P_{c1}, P_{c2}, P_{c3}, P_{c4}, P_{c5}, P_{c6}

P(gene|non-coding) → **P(gene|NON)**

parameters (P_n) → not changed

When a gene sequence = ATGGCC … ,

$P(F|COD1) = P_{c1}(ATG)\, P_{c1}(G|ATG) P_{c2}(C|TGG) P_{c3}(C|GGC) \ldots$ (defined as a true reading frame)

$P(F|COD2) = P_{c2}(ATG)\, P_{c2}(G|ATG) P_{c3}(C|TGG) P_{c1}(C|GGC) \ldots$

$P(F|COD3) = P_{c3}(ATG)\, P_{c3}(G|ATG) P_{c1}(C|TGG) P_{c2}(C|GGC) \ldots$

$P(F|COD4) = P_{c4}(ATG)\, P_{c4}(G|ATG) P_{c5}(C|TGG) P_{c6}(C|GGC) \ldots$

$P(F|COD5) = P_{c5}(ATG)\, P_{c5}(G|ATG) P_{c6}(C|TGG) P_{c4}(C|GGC) \ldots$

$P(F|COD6) = P_{c6}(ATG)\, P_{c6}(G|ATG) P_{c4}(C|TGG) P_{c5}(C|GGC) \ldots$

$P(F|NON) = P_n(ATG)\, P_n(G|ATG) P_n(C|TGG) P_n(C|GGC) \ldots$ **(3 rd order Markov chain)**

Figure 2.2 (C) Construction of training model (3) (Redefinition of P(gene|coding) by P(gene|CODm))

Finally, for each gene in the genome, I computed an index defined as the average of P(COD1|F) by the window analysis (here F is a window sequence of the query gene) and I call this "HT index" of the gene. The window size was of 96 bp and slid on the gene sequence by a step of 12 bp. The order of Markov chains was set to 5 to avoid overfitting parameters (**Borodovsky *et al.* 1995**). In computation of the HT index, the parameters of the training model contain nucleotide frequencies of a query gene itself. Therefore, in order to cancel the contribution of the gene, I computed the HT index using the parameters without the nucleotide frequencies of the gene.

2.3.4 Evaluation by Monte Carlo simulation

In order to test the statistical significance of the HT index, I performed the Monte-Carlo simulation. I generated artificial coding fragments at random based on the parameters in the computation of P(F|COD1). When the total number of genes in a given genome is T, I computed the probabilities of 100 x T artificial fragments and obtained the expected parent population for one-tailed test. The length of each of the 100 fragments corresponds to that of a real gene.

2.3.5 Correction by highly expressed genes (HT gene criteria)

Since ribosomal protein genes, elongation factor genes, chaperone genes often have abnormal base compositions or codon usage biases under the selective pressures for keeping high expression efficiency (**Karlin *et al.* 1998; Sharp & Li 1987**), these genes might be false-positive in the detection. Therefore, I prepared the referential model for detecting highly expressed genes. Likewise, the model was composed of two Markov models, which were constructed using ribosomal protein gene regions (coding and neighboring non-coding sequences). I also performed Monte-Carlo simulations using the model. The order of Markov chains was set to 3, because the sequences of ribosomal protein gene regions required for the training model are limited (50~60 genes in a genome).

Finally, I defined the genes satisfying the following two criteria as HT genes:

(i) genes having significantly low indices with the training model of the species at one percent level ($p < 0.01$),
(ii) genes not having significantly high indices with the referential model at five percent level ($p < 0.05$).

2.3.6 Window analysis

Extrinsic regions such as pathogenicity islands were often inserted as a large cluster into the genome (**Hacker *et al.* 1997; Hacker & Kaper 2000**). The large clusters should possess horizontally transferred genes detected in limited locations of the genome. In order to detect such clusters, I counted the number of HT candidates in a window of 10 genes slid by 1 gene over the genome, and then obtained the regions in which the proportions of HT candidates are larger than 40%. Next, I manually corrected the boundaries or joined the regions that seem to be consecutive in the genome.

2.3.7 Donor identification (HT donor index)

The HT index of a gene in a species computed with the training models of other species can be used as an indicator of donor species, where the HT index of the gene derived from the original model is significantly low and the HT index from the donor's model is higher.

In particular, when the donor species is estimated from other information in advance, I used another indicator, which I defined as "HT donor index", using both models of a recipient and a probable donor. This is represented by the following formula:

$$\text{HT donor index} = \frac{P(F|COD_1 \text{ of donor})P(COD_1 \text{ of donor})}{P(F|COD_1 \text{ of recipient})P(COD_1 \text{ of recipient}) + P(F|COD_1 \text{ of donor})P(COD_1 \text{ of donor})}$$

$$= \frac{P(F|COD_1 \text{ of donor})}{P(F|COD_1 \text{ of recipient}) + P(F|COD_1 \text{ of donor})} . \qquad \textbf{(3)}$$

Here, I compared the probability $P(F|COD_1)$ between training models of a recipient and a donor, and I assumed $P(COD_1 \text{ of recipient}) = P(COD_1 \text{ of donor}) = 1/2$. This is useful as the case of HT index, and I actually computed HT origin index as the average of indices by the window.

2.3.8 Homology search and phylogenetic analysis

The FASTA program (**Pearson & Lipman 1988**) was used for searching homologues of HT genes from protein databases [SWISS-PROT(ver.40), PIR(ver.72)], and as for genes having enough homologues ($E < 10^{-8}$) I constructed a phylogenetic tree by using the alignment program CLUSTAL W (**Thompson *et al.* 1994**) and the neighbor-joining method (**Saitou & Nei 1987**).

2.3.9 Comparison with other methods

Once Mrazek and Karlin (**Mrazek & Karlin 1999**) developed a powerful method for detecting horizontally transferred genes, which they called "alien genes", in the complete genome sequence. In order to evaluate the performance of my method, I obtained the "alien gene" list for 18 species out of 19 species surveyed by Karlin (**Karlin 2001**), and compared between the ratios of truth-positives and false-positives in detection. The remaining one species was *B. burgdoriferi*, which I did not use the genome for comparison, because Karlin *et al.* used *B.*

burgdoriferi plasmid genes in detection of HT genes in the genomes. I excluded plasmid data from the training model. Here, I assumed that the following genes were recent HT genes; all of genes encoding transposases and integrases and genes located in probable prophage regions of the genome. At the same time, I assumed the genes encoding ribosomal proteins were not HT genes because of functional constraint, although there are a number of exceptions reported so far (**Brochier *et al.* 2000; Hansmann & Martin 2000**). I finally prepared 367 genes as mobile element genes and 686 genes as ribosomal protein genes from the 18 species examined. I compared between the truth-positives and false-positives for the genes of at least 300 bp or longer, which is the same condition as in Karlin's method (**Karlin 2001**).

2.4 Methods 2 (phylogenetic analysis)

2.4.1 Orthologous gene set (four-orthologue condition)

At first, I constructed a database with all of the protein sequences encoded in the 84 complete genomes as shown in **Table 1.1**. Then, I conducted BLAST searches against the database (E value cut-off 10^{-8}), using each protein sequence in the genomes as a query. I used the program "blastpgp", because this program is applicable for gapped alignments of protein sequences (**Altschul *et al.* 1997**). I obtained the search result where each gene in any species has no hit or the most similar gene (=best hit) in each of the other 83 genomes. The best hit was defined as the gene having the highest BLAST score among hits in the genome in question.

Here I defined an orthologous gene pair between two species as the pair in which two genes between two species must be mutually the best hits. For an orthologous gene set among four species I expanded the definition, that is, when all of possible 6 gene pairs among four species satisfy the condition of the orthologous pair mentioned above, I defined the four-gene set as an orthologous gene set (**Figure 2.3**).

2.4.2 Taxonomic groups

In order to evaluate the proportion of horizontal gene transfer among closely related species, I focused on 6 lineages in 84 sequenced species based on the taxonomic groups and the distance among 16S rRNA sequences ($d < 0.12$) as given below (**Table 2.2**):

(i) *Bacillus-Staphylococcus* group

(ii) *Lactococcus-Streptococcus* group

(iii) Gram-positive high GC% bacteria group

(iv) *Chlamydia* group

(v) Enterobacteria and its relatives group

(vi) *Rhizobium* group.

As for *Escherichia coli*, *Staphylococcus aureus*, *Chlamydophila pneumoniae*, *Mycobacterium tuberculosis*, *Streptococcus pneumoniae*, *Streptococcus pyogens*, *Agrobacterium tumefaciens*, and *Salmonella enterica* (serovar *typhi, typhimurium*), the genomes of the same species but different strains (2 ~ 3 strains) have been determined.

2.4.3 Alignment and tree construction, and its evaluation

For all of possible combinations of four species in each group in **Table 2.2**, I prepared orthologous gene sets of the four species as mentioned in **Section 2.4.1**. Then I extensively constructed alignments and phylogenetic trees using four protein sequences in the orthologous gene sets. Phylogenetic trees were constructed by the neighbor-joining method (**Saitou & Nei 1987**). A program CLUSTALW for alignments and tree constructions was used (**Thompson *et al.* 1994**).

Here I assumed that the phylogenetic topology based on 16S rRNA sequences was correct one (**Figure 2.4**). Therefore, the other two topologies show that at least one gene has been originated by horizontal transfer (**Figure 2.1**). The significance of each tree was evaluated by the bootstrap value and the threshold was specified to 900 in 1000 trials (90%).

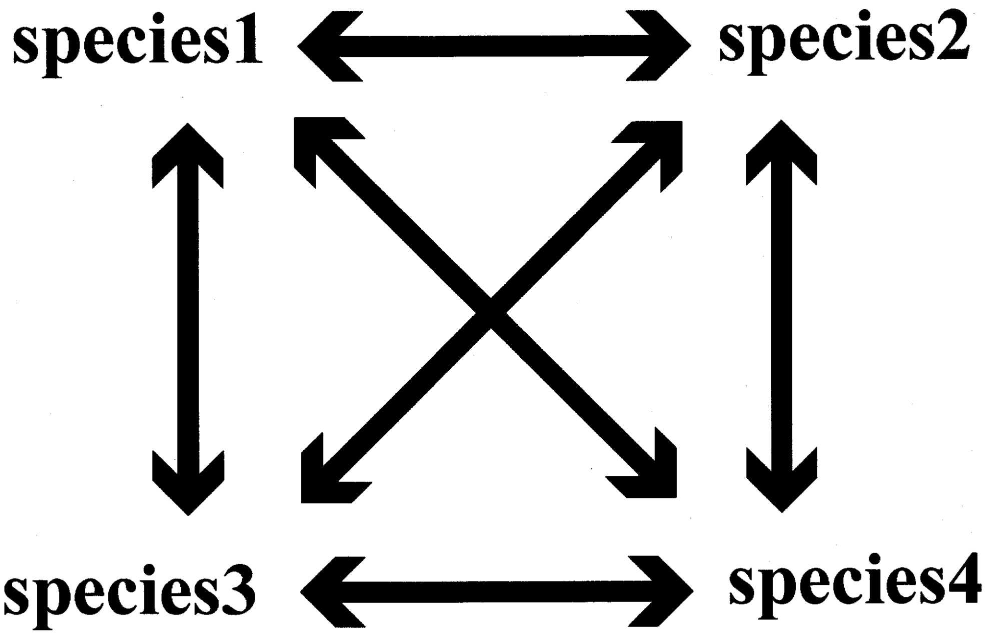


Figure 2.3 Orthologous genes in four species (four-orthologue condition)

An arrow indicates the orthologous relationship by reciprocal best hits between two species. If four genes in these species satisfy all of the six relationships, the four genes are orthologous genes each other.

Table 2.2 Examined taxonomic groups

Examined taxonomic group	**Species list***	**Taxonomic classification****	
		Order	**Family**
(i) *Bacillus-Staphylococcus* group	*Bacillus halodurans* *Bacillus subtilis* *Listeria innocua* *Listeria monocytegenes* *Staphylococus aureus* (3)	*Bacillales*	*Bacillaceae,Listeriaceae,* *Staphylococcaceae*
(ii) *Streptococcus* group	*Lactococcus lactis* *Streptococus pneumoniae* (2) *Streptococus pyogenes* (3)	*Lactobacillales*	*Streptococcaceae*
(iii) Gram-positive high GC% group	*Corynebacterium glutamicum* *Mycobacterium leprae* *Mycobacterium tuberculosis* (2) *Streptomyces coelicolor*	*Actinomycetales*	*Corynebacteriaceae,* *Mycobacteriaceae,* *Streptomycetaceae*
(iv) *Chlamydia* group	*Chlamydophila pneumoniae* (3) *Chlamydia trachomatis* *Chlamydia muridarum*	*Chlamydiales*	*Chlamydiaceae*
(v) Enterobacteria and its realtives group	*Escherichia coli* (3) *Salmonella typhimurium* *Salmonella typhi* *Yersinia pestis* *Vibrio cholerae*	*Enterobacteriales* or *Vibrionales*	*Enterobacteriaceae,* *Vibrionaceae*
(vi) *Rhizobium* group	*Agrobacterium tumefaciens* (2) *Brucella melitensis* *Mesorhizoboium loti* *Sinorhizobium meliloti*	*Rhizobiales*	*Rhizobiaceae,* *Brucellaceae*

*A number in each parenthesis is the number of same species sequenced.

** These are based on *Bergey's Manual of Systematic Bacteriology,* 2nd Edition.

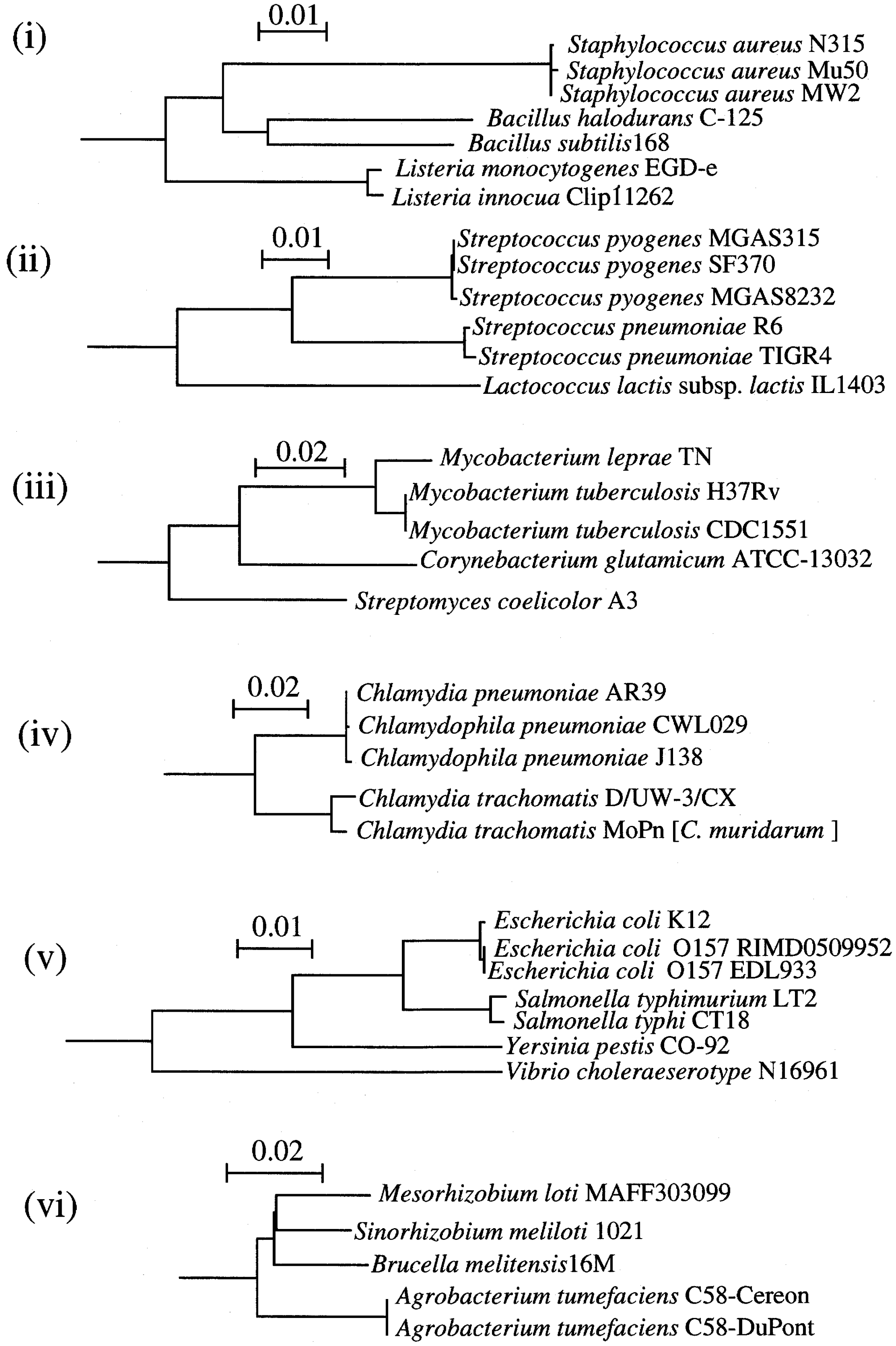

(i)
0.01
Staphylococcus aureus N315
Staphylococcus aureus Mu50
Staphylococcus aureus MW2
Bacillus halodurans C-125
Bacillus subtilis168
Listeria monocytogenes EGD-e
Listeria innocua Clip11262
(ii)
0.01
Streptococcus pyogenes MGAS315
Streptococcus pyogenes SF370
Streptococcus pyogenes MGAS8232
Streptococcus pneumoniae R6
Streptococcus pneumoniae TIGR4
Lactococcus lactis subsp. lactis IL1403
(iii)
0.02
Mycobacterium leprae TN
Mycobacterium tuberculosis H37Rv
Mycobacterium tuberculosis CDC1551
Corynebacterium glutamicum ATCC-13032
Streptomyces coelicolor A3
(iv)
0.02
Chlamydia pneumoniae AR39
Chlamydophila pneumoniae CWL029
Chlamydophila pneumoniae J138
Chlamydia trachomatis D/UW-3/CX
Chlamydia trachomatis MoPn [C. muridarum]
(v)
0.01
Escherichia coli K12
Escherichia coli O157 RIMD0509952
Escherichia coli O157 EDL933
Salmonella typhimurium LT2
Salmonella typhi CT18
Yersinia pestis CO-92
Vibrio choleraeserotype N16961
(vi)
0.02
Mesorhizobium loti MAFF303099
Sinorhizobium meliloti 1021
Brucella melitensis16M
Agrobacterium tumefaciens C58-Cereon
Agrobacterium tumefaciens C58-DuPont

Figure 2.4 Phylogenetic trees of 16SrRNA sequences in (i) ˜ (vi) groups.

2.5 Methods 3 (genome comparison)

2.5.1 Assignment of coding region and function

Assignment of coding regions and their functions in *C. efficiens* genome were conducted by the following criteria:

(i) Potential protein coding regions were assigned by software, Glimmer 2.0. This program worked under default conditions.
(ii) The Shine-Dalgarno sequence, 5' -AAAGAGG -3' , was used for assignment of the start point of translation.
(iii) The BLASTP similarity searches for each coding region were performed against a non-redundant protein database.
(iv) A potential protein coding region shorter than 50 bps and having no similarity to others was removed from the list of the potential protein coding regions

I was a member of the annotation team of *C. efficiens*, and checked the performance of Glimmer 2.0 in (i), and conducted the final assignment in (iv) about 500 genes out of the whole candidates. In the case of *C. diphtheriae*, the potential protein-coding regions were obtained by the Glimmer prediction only.

2.5.2 Window analysis of GC% and GC-skew

A Guanine (G) + Cytosine (C) ratio in the whole sequence was computed by the window of 20 kilobases (kb) and the step of 1kb. The GC skew on one strand of the genome sequence, which is represented by the equation (G-C)/(G+C) was computed in the same window and step size.

2.5.3 Orthologous gene pairs between two closely related species

I defined an orthologous gene pair between two species (*C. efficiens* - *C. glutamicum, C. efficiens* - *C. diphtheriae* and *C. glutamicum* - *C. diphtheriae*) as mentioned in **Section 2.4.1**. Here I used the FASTA program for searching the best hits (**Pearson & Lipman 1988**); the best hit was defined as the gene having the highest z-score in the subject genes.

2.5.4 Amino acid substitution matrix, and estimation of synonymous nucleotide substitution

Firstly, I constructed pairwise alignments by using amino acid sequences encoded by orthologous genes between *C. efficiens* and *C. glutamicum*, respectively. Next, I replaced all of matched sites in each alignment back with the original triplets of nucleotides, and then computed the number of substitutions in each codon between two species. Here, I excluded the first triplet of both sequences, so-called "initiation codon", from the computation. The reason is that the initiation codon as firstly loading methionine (Met) in translation is often an irregular triplet such as GTG originally translated to valine (Val). The number of synonymous substitutions per site between an orthlogous gene pair was estimated by Nei & Gojobori's method (**Nei & Gojobori 1986**).

2.5.5 Horizontal gene transfer in *C.efficiens* genome

The algorithm and criteria for detecting horizontally transferred genes in *C. efficiens* genome are the same as those described in **Section 2.3**.

3. Results and Discussion

3.1 Detection of horizontally transferred genes in the complete genome sequences by Bayes' estimation

3.1.1 Estimation of horizontally transferred genes in prokaryotic genomes

Figure 3.1 shows the distribution of HT genes that were detected by my method in *Escherichia coli* O157:H7 strain RIMD 0509952 genome (Genbank accession:BA000007). *Escherichia coli* O157:H7 is a pathogen and has many prophage regions some of which contain pathogenicity islands (**Hayashi *et al.* 2001**). All of prophages or prophage-like elements well match the regions where HT genes were located, and apparently the HT index is more sensitive indicator for detecting HT gene than GC content at the third positions of codons.

Subsequently, I applied my method to all of the 84 prokaryotic complete genome sequences and estimated the average for the proportions of HT genes over all the prokaryotic genomes examined (**Supplemental figures**). The results have shown that about 12% of all prokaryotic genes in the complete genomes may have been derived by horizontal gene transfer. Moreover, the proportion varies depending heavily upon prokaryotic lineage: It ranges from 0.5% to 25%, namely *de facto* zero to about one forth of the total genes in a genome (**Table 3.1**). *Buchnera* sp. APS, which has the second smallest genome size (~0.6Mb) in the sequenced genomes (**Shigenobu *et al.* 2000**), has the smallest proportion of HT gene (0.5%), and the largest proportion was obtained for euryarchaeota *Methanosarcina acetivorans* (25%). In fact, since 0.5% of *Buchnera* sp. APS is smaller than the statistically significant level (P = 0.01 = 1%) (**see Section 2.3.5, HT gene criteria (i)**), these transferred genes are probably false-positives by chance.

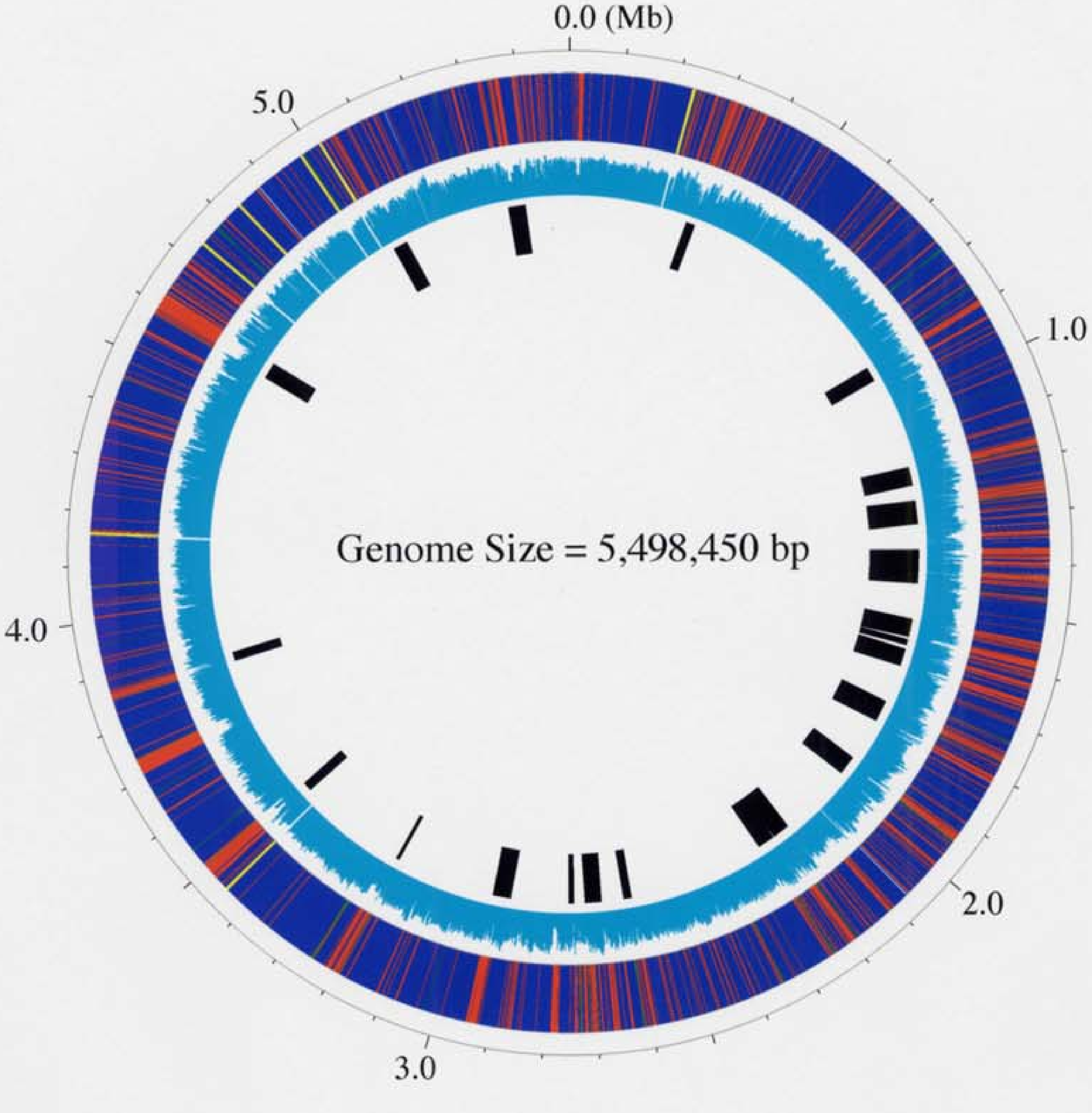


Figure 3.1 Circular map of *Escherichia coli* O157 : H7 (strain RIMD 0509952) genome.

Figure 3.1

Circular map of *Escherichia coli* O157 : H7 genome. From the outside to the inside, the first circle is the scale in megabases. The second shows genes or RNA located on both strands: blue, genes not detected as HT genes; red, detected HT genes; yellow, ribosomal RNA; green, transfer RNA. The third shows GC contents at the third positions of codons in genes. Black bars inside it correspond to the regions of prophages or probable prophages.

Table 3.1 Proportion of horizontally transferred genes in complete genomes

Species name*	Domain**	No. of analysed genes	No. of HT genes	Parcent (%)	No. of Cluster
Methanosarcina acetivorans C2A	A	4527	1143	**25.2**	**53**
Chlorobium tepidum TLS	B	2226	536	**24.1**	**19**
Neisseria meningitidis MC58 (serogroup B)	B	2013	440	**21.9**	**19**
Aeropyrum pernix K1	A	1839	392	**21.3**	**13**
Mesorhizobium loti MAFF303099	B	6744	1428	**21.2**	**29**
Xylella fastidiosa CVC 8.1.b clone 9.a.5.c	B	2747	569	**20.7**	**18**
Xanthomonas axonopodis pv. ***citri*** 306	B	4311	865	**20.1**	**25**
Escherichia col O157:H7. RIMD 0509952	B	5347	1071	**20.0**	**46**
Xanthomonas campestris pv.***campestris*** ATCC 33913	B	4174	829	**19.9**	**26**
Methanosarcina mazei Goe1	A	3368	636	**18.9**	**26**
Escherichia coli O157:H7 EDL933	B	5303	999	**18.8**	**46**
Corynebacterium glutamicum ATCC-13032	B	3099	571	**18.4**	**15**
Streptococcus pneumoniae TIGR4 ATCC-BAA-334	B	2066	370	**17.9**	**16**
Streptococcus pyogenes MGAS8232	B	1845	329	**17.8**	**14**
Neisseria meningitidis Z2491 (serogroup A)	B	2054	358	**17.4**	**12**
Salmonella typhi CT18	B	4380	752	**17.2**	**29**
Escherichia coli K12 MG1655	B	4278	721	**16.9**	**31**
Streptococcus pyogenes MGAS315	B	1865	315	**16.9**	**18**
Streptomyces coelicolor A3(2)	B	7499	1260	**16.8**	**37**
Vibrio cholerae serotype O1, strain N16961	B	3790	633	**16.7**	**13**
Agrobacterium tumefaciens C58-DuPont	B	4660	769	**16.5**	**13**
Streptococcus pneumoniae R6	B	2037	336	**16.5**	**12**
Salmonella typhimurium LT2 SGSC1412	B	4440	732	**16.5**	**37**
Caulobacter crescentus CB15	B	3733	601	**16.1**	**6**
Ralstonia solanacearum GMI1000	B	3436	547	**15.9**	**25**
Yersinia pestis CO-92 (Biovar Orientalis)	B	3881	603	**15.5**	**26**
Brucella melitensis 16M	B	3198	493	**15.4**	**14**
Thermoanaerobacter tengcongensis MB4T	B	2588	396	**15.3**	**11**
Agrobacterium tumefaciens C58-Cereon	B	4549	663	**14.6**	**16**
Staphylococcus aureus subsp. ***aureus*** MW2	B	2617	381	**14.6**	**10**
Deinococcus radiodurans R1	B	2937	424	**14.4**	**3**
Methanobacterium thermoautotrophicum delta H	A	1869	243	**13.0**	**6**
Synechocystis sp. PCC6803	B	3160	411	**13.0**	**9**
Lactococcus lactis subsp. ***lactis*** IL1403	B	2265	288	**12.7**	**7**
Staphylococcus aureus Mu50 (VRSA)	B	2688	335	**12.5**	**8**
Mycobacterium tuberculosis CDC1551	B	4178	510	**12.2**	**9**
Thermoplasma acidophilum DSM 1728	A	1478	181	**12.2**	**4**
Methanopyrus kandleri AV19	A	1681	203	**12.1**	**1**
Mycoplasma pneumoniae M129	B	675	82	**12.1**	**2**
Pyrococcus horikoshii OT3	A	1800	214	**11.9**	**3**
Streptococcus pyogenes SF370 (M1)	B	1695	194	**11.4**	**6**
Mycobacterium tuberculosis H37Rv	B	3903	442	**11.3**	**15**
Staphylococcus aureus N315 (MRSA)	B	2584	292	**11.3**	**5**
Bacillus subtilis 168	B	4092	451	**11.0**	**14**
Pyrococcus abyssi GE5	A	1768	192	**10.9**	**4**
Pseudomonas aeruginosa PAO1	B	5562	597	**10.7**	**15**
Sinorhizobium meliloti1021	B	3341	356	**10.7**	**6**
Pasteurella multocida Pm70	B	2014	214	**10.6**	**6**
Archaeoglobus fulgidus DSM4304	A	2401	253	**10.5**	**9**
Halobacterium sp. NRC-1	A	2056	215	**10.5**	**4**
Listeria innocua Clip11262	B	2968	301	**10.1**	**7**

(continued)

Haemophilus influenzae KW20	B	1708	169	**9.9**	**2**
Clostridium acetobutylicum ATCC 824D	B	3670	361	**9.8**	**0**
Listeria monocytogenes EGD-e	B	2845	273	**9.6**	**6**
Nostoc (Anabaena) sp. PCC 7120	B	5365	509	**9.5**	**1**
Sulfolobus tokodaii 7	A	2826	269	**9.5**	**2**
Mycobacterium leprae TN	B	1605	149	**9.3**	**2**
Sulfolobus solfataricus P2	A	2977	258	**8.7**	**1**
Helicobacter pylori 26695	B	1542	132	**8.6**	**4**
Thermoplasma volcanium GSS1	A	1525	128	**8.4**	**4**
Aquifex aeolicus VF5	B	1521	125	**8.2**	**4**
Helicobacter pylori J99	B	1487	120	**8.1**	**3**
Thermotoga maritima MSB8	B	1837	143	**7.8**	**4**
Pyrococcus furiosus DSM 3638	A	2062	156	**7.6**	**1**
Treponema pallidum subsp. ***pallidum*** Nichols	B	1028	76	**7.4**	**1**
Bacillus halodurans C-125	B	4028	295	**7.3**	**10**
Chlamydia pneumoniae AR39	B	1104	81	**7.3**	**0**
Pyrobaculum aerophilum IM2	A	2579	188	**7.3**	**3**
Chlamydia trachomatis MoPn / Nigg	B	818	57	**7.0**	**0**
Fusobacterium nucleatum ATCC 25586	B	2058	138	**6.7**	**1**
Chlamydophila pneumoniae J138	B	1069	67	**6.3**	**0**
Mycoplasma pulmonis UAB CTIP	B	782	48	**6.1**	**0**
Clostridium perfringens 13	B	2660	159	**6.0**	**0**
Methanococcus jannaschii DSM 2661	A	1714	100	**5.8**	**0**
Chlamydophila pneumoniae CWL029	B	1052	59	**5.6**	**0**
Rickettsia prowazekii Madrid E	B	834	41	**4.9**	**0**
Chlamydia trachomatis D/UW-3/CX (serovar D)	B	894	42	**4.7**	**0**
Borrelia burgdorferi B31	B	842	36	**4.3**	**0**
Rickettsia conorii Malish 7	B	1374	58	**4.2**	**0**
Campylobacter jejuni subsp. ***jejuni*** NCTC 11168	B	1630	51	**3.1**	**0**
Buchnera aphidicola SG	B	545	16	**2.9**	**0**
Mycoplasma genitalium G-37	B	480	9	**1.9**	**0**
Ureaplasma urealyticum serovar 3	B	611	10	**1.6**	**0**
Buchnera aphidicola AP	B	564	3	**0.5**	**0**
Average: 84 species (Archaea=16|Bacteria=68)		2660.14	363.06	12.0	10.3
		Total number of clusters			867

* Species are listed in descending order with regared to proportion of HT genes.
**B: Bacteria, A: Archaea

3.1.2 Distribution pattern of horizontally transferred genes among taxa

Obligate parasite bacteria, such as *Borrelia burgdorferi* and bacteria belonging to *Mycoplasma*, *Rickettsia*, and *Chlamydia* genera also have relatively less HT gene candidates except for *Mycoplasma pneumoniae* (13.0%). These proportions are not different from the order of magnitude of the significant level (**see Section 2.3.5, HT gene criteria (i)**), suggesting that such bacteria have little extrinsic genes in the genome. In general, there is a positive correlation between the total coding sequence in a genome and the proportion of horizontally transferred coding sequence (**Figure 3.2**). This means that species having a larger amount of genes contain more signatures of horizontal gene transfer, and supports the two hypotheses for the change of genome size in prokaryotic lineages: (1) the genome expansion was caused by gene acquisition other than gene duplication and (2) the genome shrinking was caused by the loss of genes.

Campylobacter jejuni, a causative agent of food-borne diarrheal disease, has low proportion of HT genes (3.1%), meaning that this species ahs a highly stable genome. The rarity of horizontal gene transfer in *C. jejuni* may be due to the fact that this genome has neither prophage nor insertion sequence homologs (**Parkhill *et al.* 2000a**).

3.1.3 Gene clusters including possible pathogenicity islands, and its implication to genome rearrangement

Although a single gene may display a low HT index purely by chance, a large cluster of the consecutive or closely linked genes where the indices are uniformly low, strongly suggests that all of those genes were introgressed together as a unit. In order to detect such clusters, I computed the local densities of transferred genes using the simple window analysis of HT indices in the genome scale (**see Section 2.3.6**), and detected the regions where horizontally transferred genes are densely located in the genome. As a result, I found 867 possible clusters

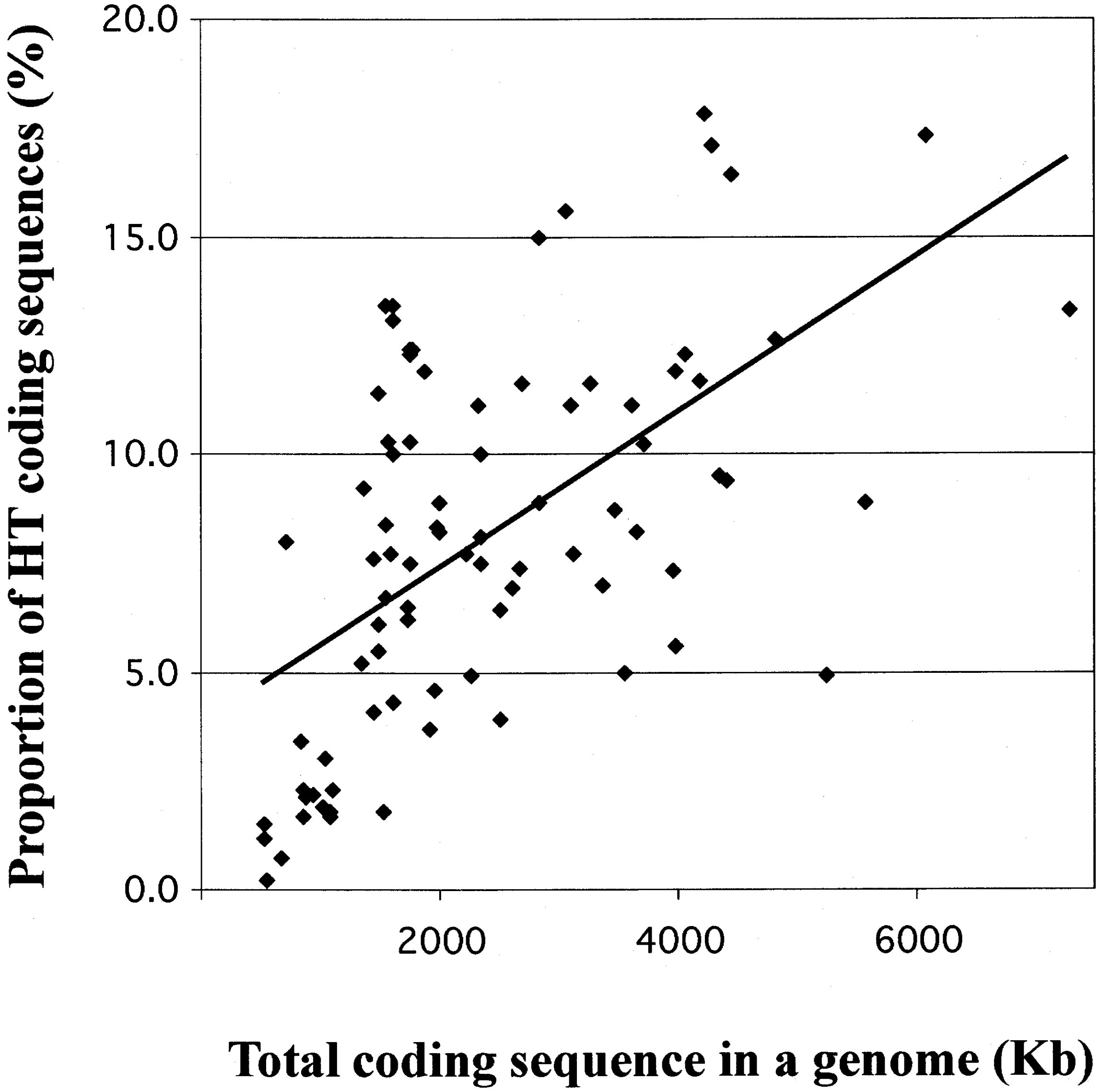


Figure 3.2 Relationship between total coding sequence and proportion of HT coding sequences in a genome.

in the complete genome sequences (**Table 3.1**). Many of these clusters correspond to parts of mobile element such as prophages, previously known as pathogenicity islands. Moreover, I surveyed possible pathogenicity islands in which putative virulence genes such as adhesin, haemolysin were encoded, and newly found 61 candidates from 16 pathogens infecting animals or plants (**Table 3.2**).

Interestingly enough, when I investigated the relationships between the orthologous gene order and the location of horizontally transferred clusters in two closely related species, I found that, in two *Neisseria meningitidis* genomes, horizontally transferred clusters were frequently located on or beside the syntenic break points where the genome inversions must have occurred (**Figure 3.3(A)**). I observed such a correlation in the comparison between two *Xanthomonas* species, *X. axonopodis* and *X. campestris* (**Figure 3.3(B)**). These results strongly suggest that large transferred regions are evolutionary unstable in the host genome, and often cause genome rearrangement, as observed in *H. pylori.*

3.1.4 Functional categorization of horizontally transferred genes

I assigned the candidates of HT genes to the biological roles of the TIGR microbial database. I have found that mainly four categories, "plasmid, phage, and transposon functions", "cell envelope", "regulatory function" and "cellular process" genes show higher (>10%) percentages of HT genes than other categories (**Figure 3.4**). The frequent gene transfer of "plasmid, phage, and transposon functions" is quite reasonable, because this category contains genes related to mobile elements as the name represents. Acquisition of "cell envelope" genes may contribute to cell defense against harmful chemical substances in the environment. Of "cellular process" genes, genes obviously related to pathogen, toxin-production/detoxification including antibiotics synthesis are frequently transferred (**Figure 3.5 (A),(B)**). These results quantitatively revealed, for the first time, that pathogenicity or antibiotics related genes are often subject to horizontal transfer among species. Interestingly enough, "regulatory function" genes

Table 3.2 Possible pathogenicity islands (PAIs) detected in this study

Species name	Detected cluster	Possible PAI	Genomic region	Kb	Gene	Main genes	Mobile element*	tRNA locus**	Putative function***
Brucella melitensis **16M**	14	3	BMEI1393 - BMEI1424	21.5	32		tra	Met	O-antigen
			BMEI1674 - BMEI1706	21.6	33		tra	Phe	virulence-associated protein E
			BMEII0709 - BMEII0729	15.4	21		tra	Ser	hemagglutinin
Pasteurella multocida **Pm70**	6	2	phyB - PM0777	11.1	6	phyAB	---	---	capsule biosynthesis
			PM0842 - PM0850	8.5	9	tad	---	---	adherence
Escherichia col **O157:H7** **(RIMD 0509952)**	46	9	ECs0324 - ECs0356	29.4	33		---	---	putative invasin, adhesin
			ECs1160 - ECs1220	32	61		int	3 tRNA	shiga toxin
			ECs1267 - ECs1284	26.8	18		---	---	hemagglutinin/hemolysin-related protein
			ECs1357 - ECs1394	25.5	38		tra	---	Iha adhesin
			ECs2102 - ECs2114	14.2	13		---	---	putatice adhesin
			ECs2831 - ECs2845	13.7	15		---	---	H/O-antigen
			ECs2971 - ECs3013	22	43		int	---	shiga toxin
			ECs3702 - ECs3737	28.9	36		---	Gly	type III secretion system
			ECs3843 - ECs3865	18.9	23		int, tra	Phe	virulence-related membrane protein/adherence factor
Escherichia coli **O157:H7** **(EDL933)**	43	8	ykgK - Z0397	30.1	34		---	---	putative adhesin
			intW - Z1503	61.1	68		int	3 tRNA	shiga-like toxin
			ydeK - Z2211	21.1	16		---	---	putative adhesin
			ydcE - Z2264	9.8	11		---	---	H-antigen
			wbdR - wbdN	13.7	12		---	---	O-antigen
			Z3334 - intV	25.5	39		int	---	shiga-like toxin
			Z4165 - Z4201	28.9	36		---	Gly	type III secretion
			Z4313 - Z4333	17.1	17		int, tra	Phe	putative enterotoxin/cytotoxin
Helicobacter pylori **26695**	4	1	HP0431 - HP0459	28.8	29	virB4	tra	---	virulence
Helicobacter pylori **J99**	3	1	jhp0914 - jhp0924	10.2	11	virB4	---	---	virulence

(continued)

Neisseria meningitidis MC58 (serogroup B)	15	7	NMB0363 - NMB0376	8.8	13	mafAB,frpC	---	---	mafA(adhesin)
			NMB0491 - NMB0521	35.7	30		---	---	hemagglutinin/hemolysin-related protein
			NMB0643 - NMB0660	11.7	17	mafA	---	Pro	mafA(adhesin)
			NMB1208 - NMB1215	9.2	8		---	---	toxin-activating protein,hemagglutinin/hemolysin-related protein
			NMB1397 - NMB1410	13	13		tra	---	FrpA/C-related protein
			NMB1746 - NMB1785	44.8	36		tra	---	hemolysin activation protein
			NMB2105 - NMB2125	10.7	16	mafB	---	---	mafB(adhesin)
Neisseria meningitidis Z2491 (serogroup A)	12	3	NMA0307 - mafA	8.8	19	mafAB	---	---	adhesin
			mafB3 - NMA0858	3.7	6		---	Pro	adhesin
			mafA2 - NMA2124	8.6	12	mafAB	---	---	adhesin
Mycobacterium tuberculosis H37Rv	15	2	Rv1904 - Rv1913	9	10	furA	---	---	
			drrA - Rv2962c	42.9	27	mas,fadD28	---	---	acyl-CoA synthase
Salmonella typhimurium LT2	37	7	STM0274A - STM0307	35.3	34	saf	int, tra	---	shiga-like toxin, VirG
			STM1239 - pagC	6.7	7	msgA,pagDC	---	Other	macrophage survival gene, virulence protein
			STM1265 - STM1276	6.9	12		---	---	putative hemolysin
			STM1667 - STM1673	7.5	7		---	---	homology to invasin C of Yersinia
			prpA - STM1872	17.6	21	sopE2	int, tra	---	toxin
			STM2230 - STM2245	15	16	msgA	phr	Pro	virulence protein MsgA
			STM2761 - mig-14	29	21	iro,virK	int, tra	---	virulence gene
Salmonella typhi CT18	29	3	STY1391 - STY1397	7.5	7		---	---	invasin-like protein
			STY1877 - STY1893	9.9	14	pagCD etc.	phr	Arg	putative virulence proteins, toxin-like protein
			STY4521 - int	133	144	vex	int	Phe	Vi polysaccharide biosynthesis
Streptococcus pneumoniae TIGR4	16	5	SP0343 - SP0359	17.9	16	Cps genes	tra	---	capsular polysaccharide biosynthesis
			SP1030 - SP1065	27.8	34		tra	---	iron-compound ABC transporter
			SP1760 - SP1775	32.5	14		---	---	glycosyl transferase
			SP1818 - SP1836	15.4	17		---	---	UDP-glucose 4-epimerase
			SP1924 - SP1938	8.3	15		tra	---	autolysin

(continued)

Streptococcus pyogenes **SF370**	6	2	SPy0431 - SPy0437	4	6	speJ	---	---	exotoxin
			sagA - SPy0746	8.3	9	sagA	---	---	streptolysin S
Vibrio cholerae **N16961**	13	1	VC1443 - VC1465	34.6	23		---	---	RTX toxin, cholera enterotoxin
Xanthomonas axonopodis	25	4	XAC1489 - int	23.7	22	xrvA	int, tra	---	virulence regulator
			XAC1911 - XAC1925	13.4	15	XAC1918	tra	---	hemolysin related protein
			XAC2174 - intS	127	113	hlyBD,pilL	int, tra	---	hemolysin secretion protein, PilL
			XAC2604 - XAC2622	16.9	19	virB1~4,6,8~11	tra	Val	virulence genes
Xanthomonas campestris	26	3	aglA - XCC2482	11.6	12	virB1~4,6,8~11	tra	Val	virulence genes
			XCC3114 - intS	40.1	33	virB6	int, tra	Gly	virulence genes
			XCC3293 - XCC3311	18.5	19	virB6	tra	---	virulence genes

Total: 16 genomes

* int: integrase, tra: transposase, phr: phage remnant
** This indicates tRNA loci within or adjacent to detected PAI.
*** Functional annotations are according to Genbank annotations.

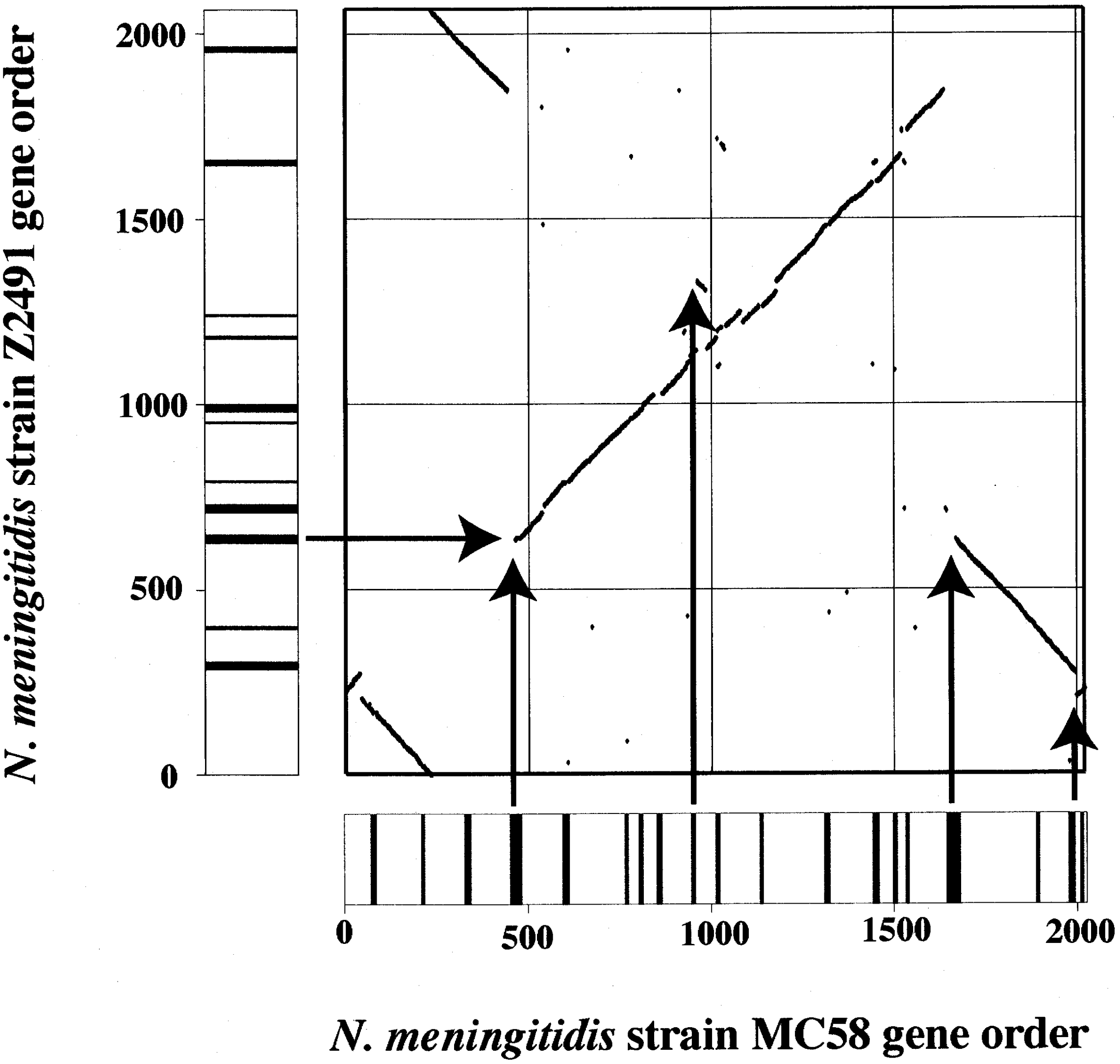


Figure 3.3 (A) Orthologous gene order between two *Neisseria meningitidis* genomes, and horizontally transferred gene clusters in both genomes. Arrows indicate genome inversion points near by horizontally transferred gene clusters. Orthologous genes between both species were defined as mentioned in Section 2.4.1.

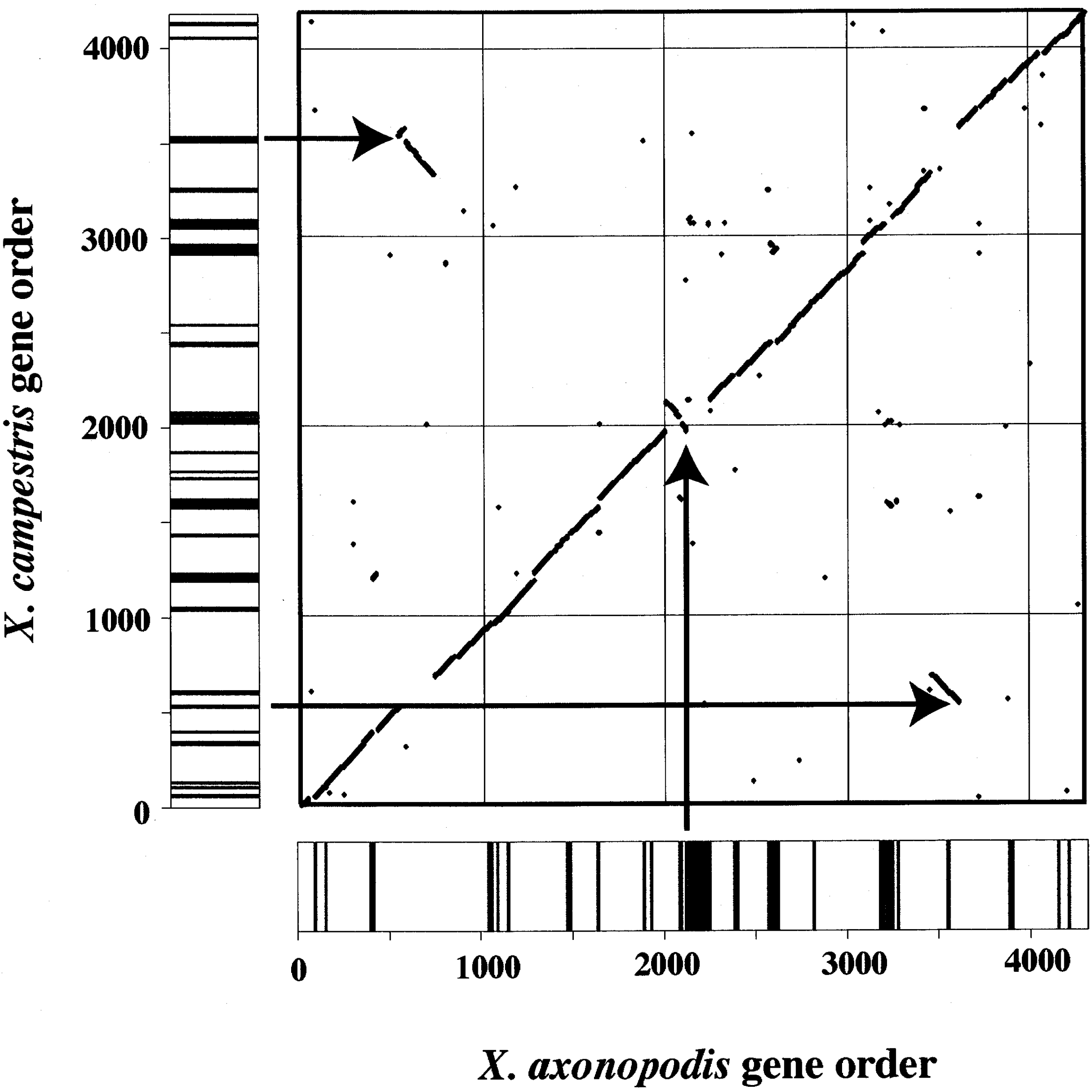


Figure 3.3 (B) Orthologous gene order between *Xanthomonas axonopodis* and *Xanthomonas campestris* genomes, and horizontally transferred gene clusters in both genomes. Arrows indicate genome inversion points near by horizontally transferred gene clusters. Orthologous genes between both species were defined as mentioned in Section 2.4.1.

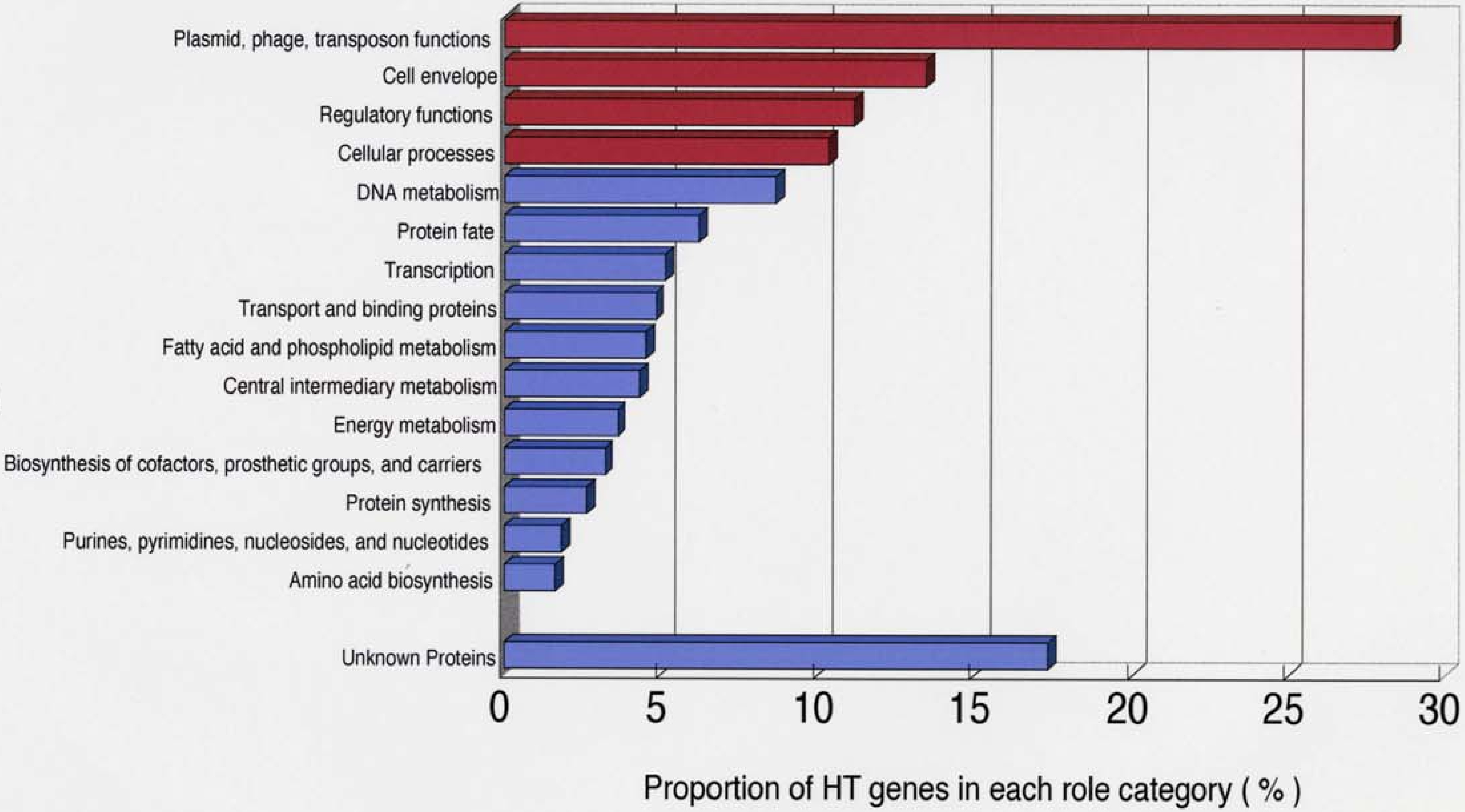


Figure 3.4 Proportions of HT genes in each functional category (main role)

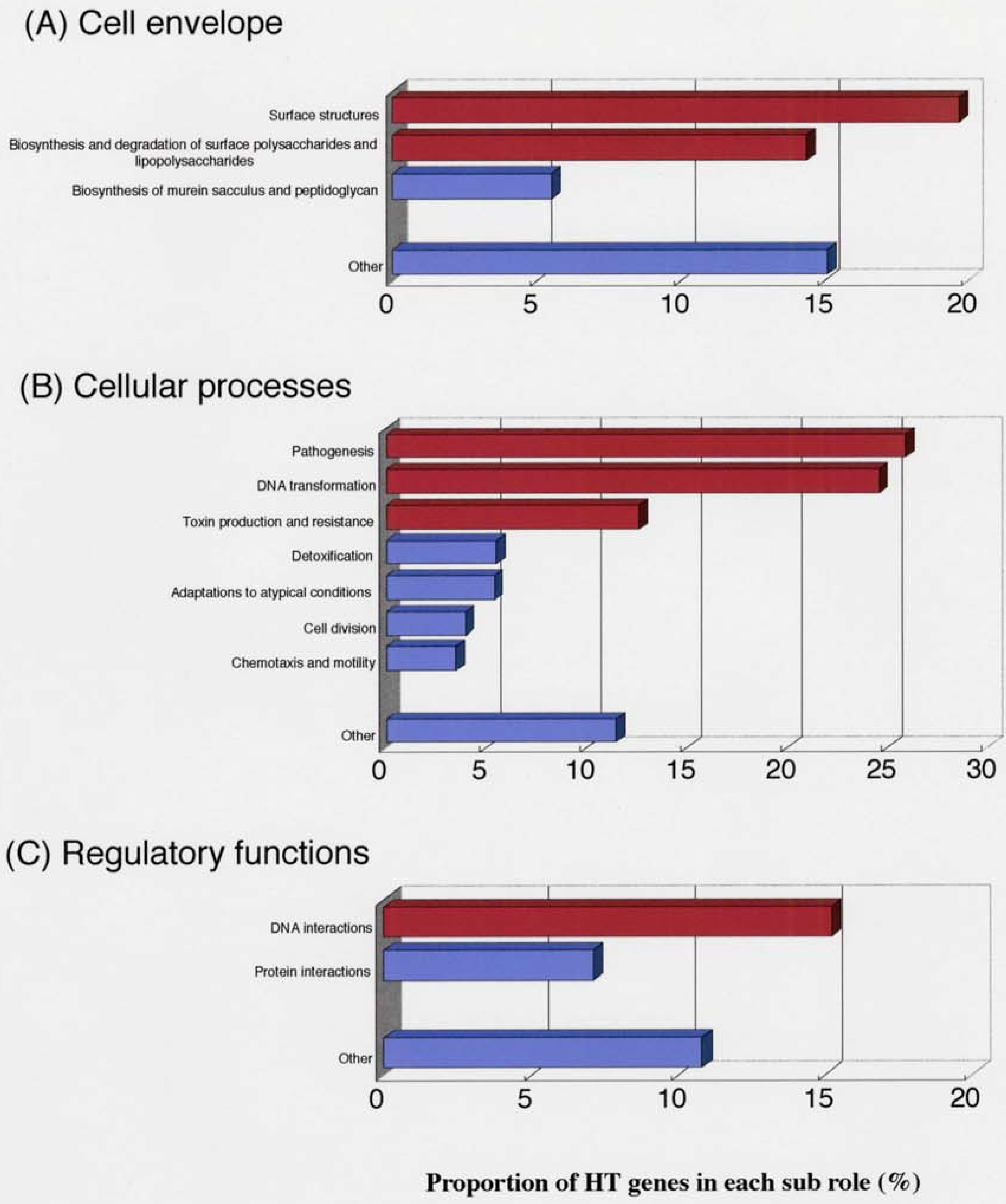


Figure 3.5 Proportions of HT genes in three main roles

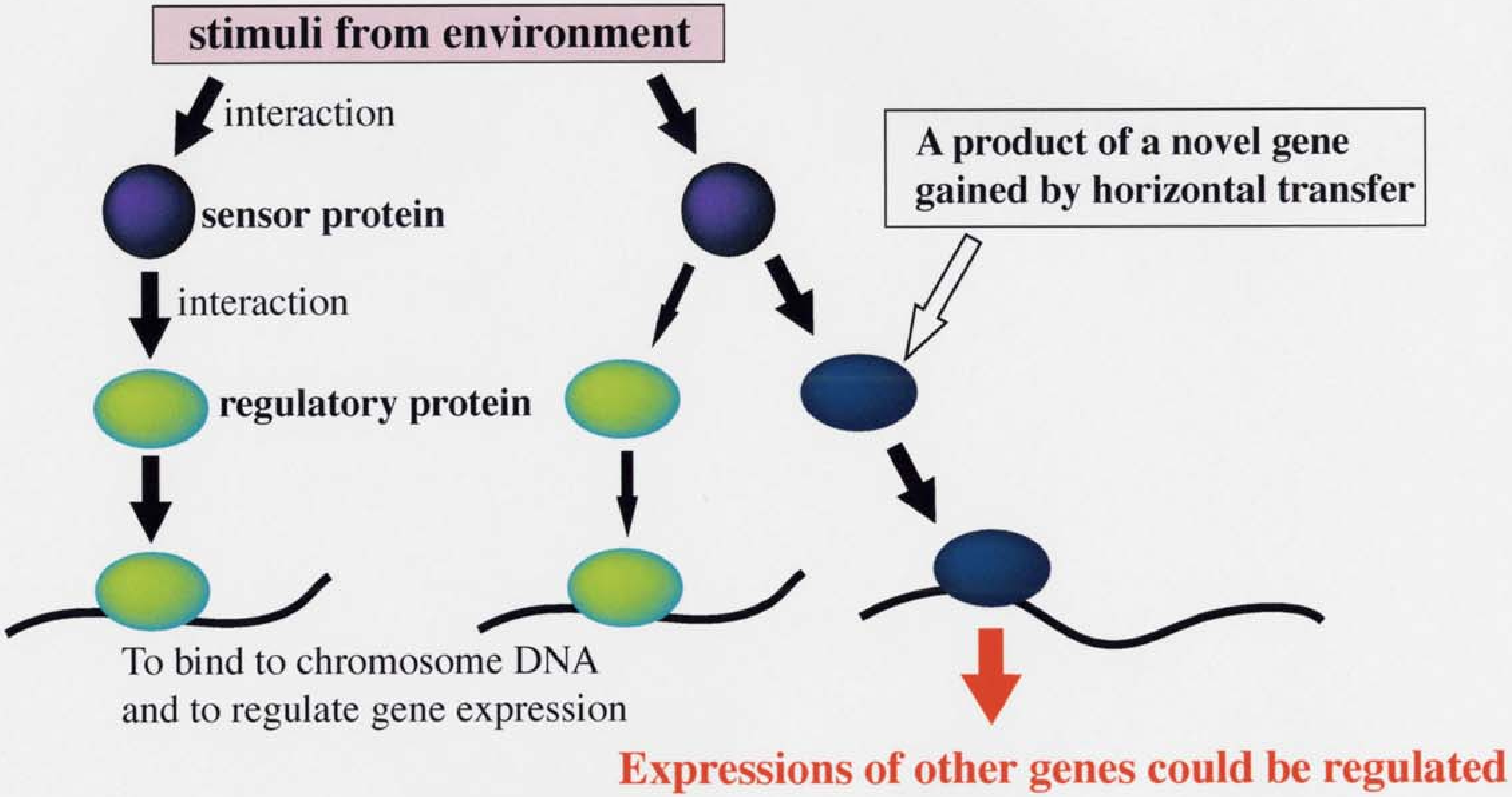

stimuli from environment
interaction
sensor protein
interaction
regulatory protein
A product of a novel gene gained by horizontal transfer
To bind to chromosome DNA and to regulate gene expression
Expressions of other genes could be regulated

Figure 3.6 Horizontal transfer of DNA interaction genes. A product of "DNA interaction" gene may have an effect on the transcription pattern of genes, and may change the gene network in recipient cells.

also frequently detected as transferred genes (**Figure 3.5 (C)**). Since "regulatory function" genes include those regulating transcription possibly by binding DNA, the acquisition of these genes may be able to alter gene expression network for adaptation under a variety of conditions (**Figure 3.6**).

3.1.5 Identification of donor species

Figure 3.7 shows that HT donor indices of *E.coli* genomic regions using the *Streptococcus pneumoniae* model are quite high. In fact, a number of genes are homologous to *Streptococcus* genes, although I could not obtain reliable information about a phylogenetic tree. Moreover, quite remarkable outcomes were obtained in survey of *Neisseria meningitidis* genome. Although the horizontal transfer between genera *Neisseria* and *Haemophilus* is previously reported (**Kroll *et al.* 1998; Davis *et al.* 2001**), in the present study, I newly identified extrinsic genes originated from *Staphylococcus* and *Streptococcus* lineages as well as those of *Haemophilus* origins, which were also independently supported by phylogenetic analysis (**Figure 3.8**). My method strongly suggests that *Neisseria meningitidis* genome has a so-called "mosaic structure" composed of genes that were derived from multiple origins.

Figure 3.9 shows that HT indices in a pathogenicity island of a cholera pathogen *Vibrio cholerae,* termed "TCP (toxin coregulated pilus) island", are higher on the *Campylobacter jejuni* model than on the original *V. cholerae* one. Although HT donor indices did not show a clear pattern as the cases mentioned above, a number of genes in the TCP island was weakly homologous to those of *C. jejuni*. This observation may imply that the TCP island was derived from a species that was closely related to *C. jejuni* with the nucleotide compositions similar to those of it. Since both *Campylobacter* and *Vibrio* species live in the intestine of animals, it is possible that horizontal gene transfer had occurred between the two species there.

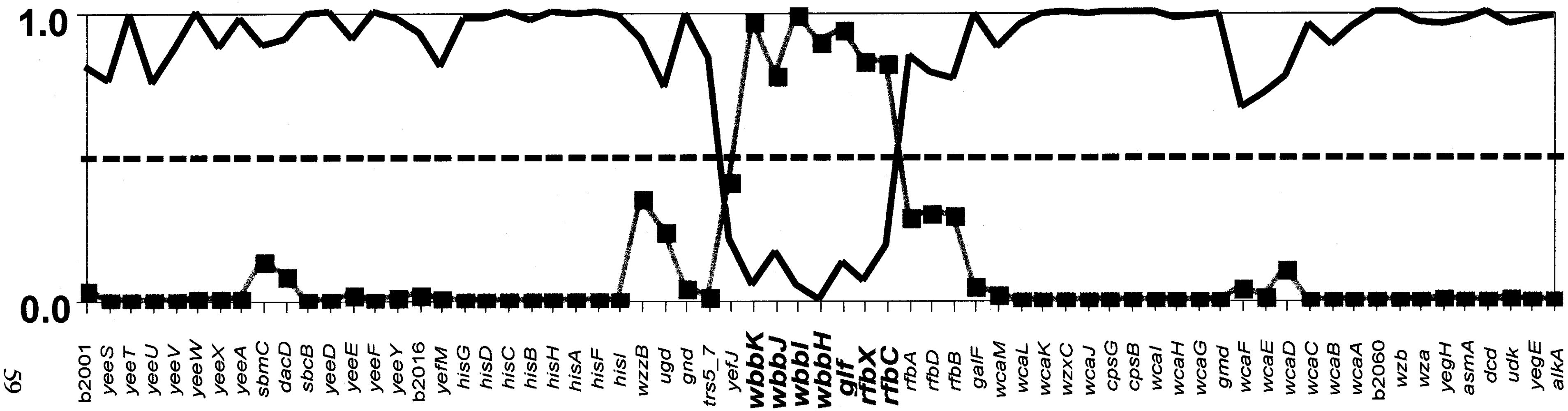


Figure 3.7 HT indices in *Escherichia coli* K12 and HT donor indices using *Streptococcus pneumoniae* model. Black line shows HT indices of genes (b2001 – *alkA*) by *E.coli* model itself and dotted line (all 0.5) and yellow line show HT donor indices by *E.coli* model itself and *S. pneumoniae* model, respectively. Bold genes (*wbbK* - *rfbC*) indicate that these genes are possiblly transferred from *S. pneumoniae*. Both flanking 30 genes from *wbbK* and *rfbC* are also shown in the figure and these genes considered to be intrinsic.

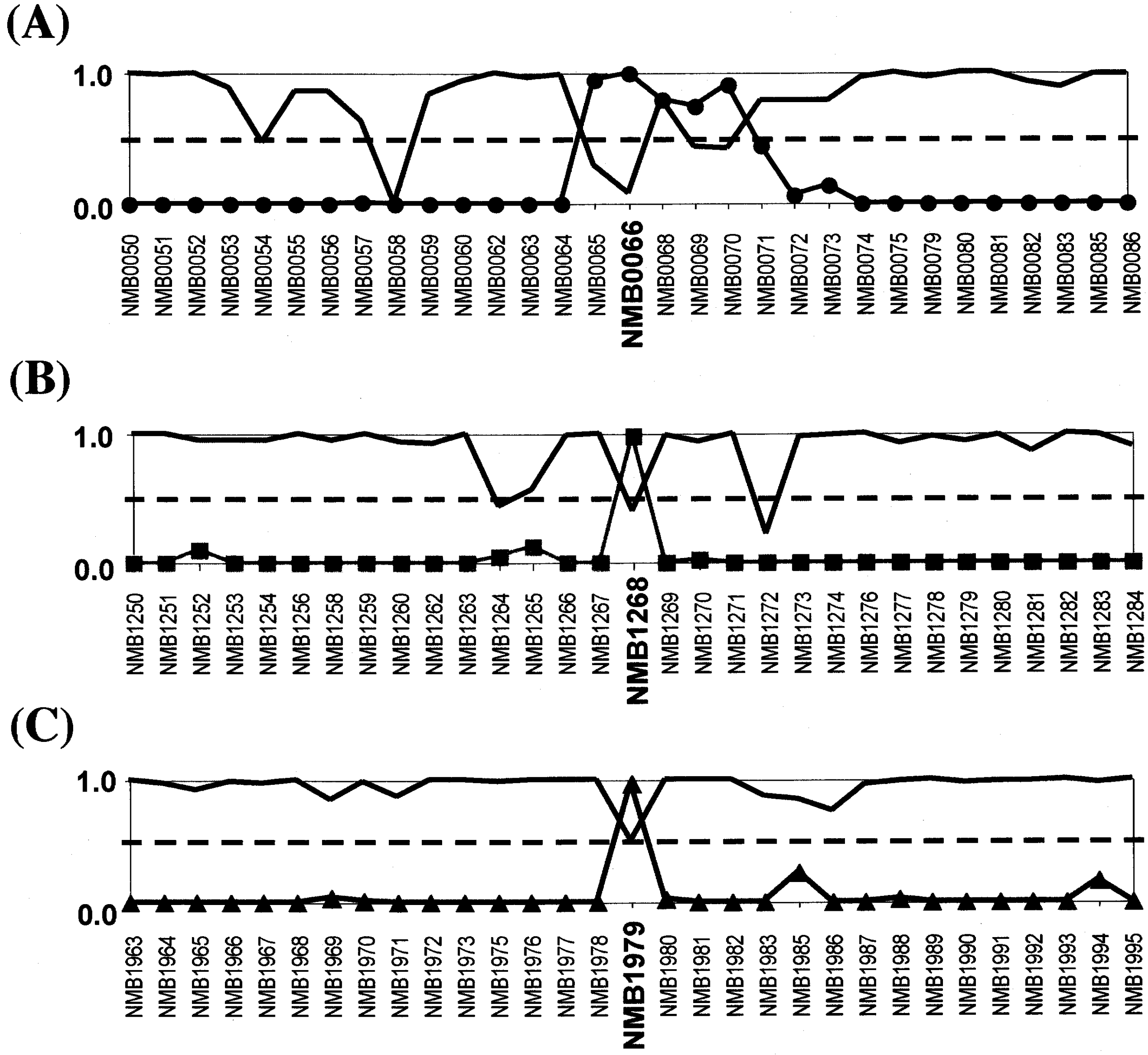


Figure 3.8 HT indices in *Neisseria meningitidis* MC58 and HT donor indices using *Staphylococcus aureus* (A), *Streptococcus pneumoniae* (B) and *Haemophilus influenzae* (C) models. Black lines are HT indices of genes by *N. meningitidis* model itself and dotted lines are HT donor indices (all 0.5) by *N. meningitidis* itself. Colored lines show HT donor indices by the models of reference species, respectively. Horizontal transfers of bold genes from these species are supported by molecular phylogenetic analysis. Flanking 15 genes from these genes to both directions are shown in the figures.

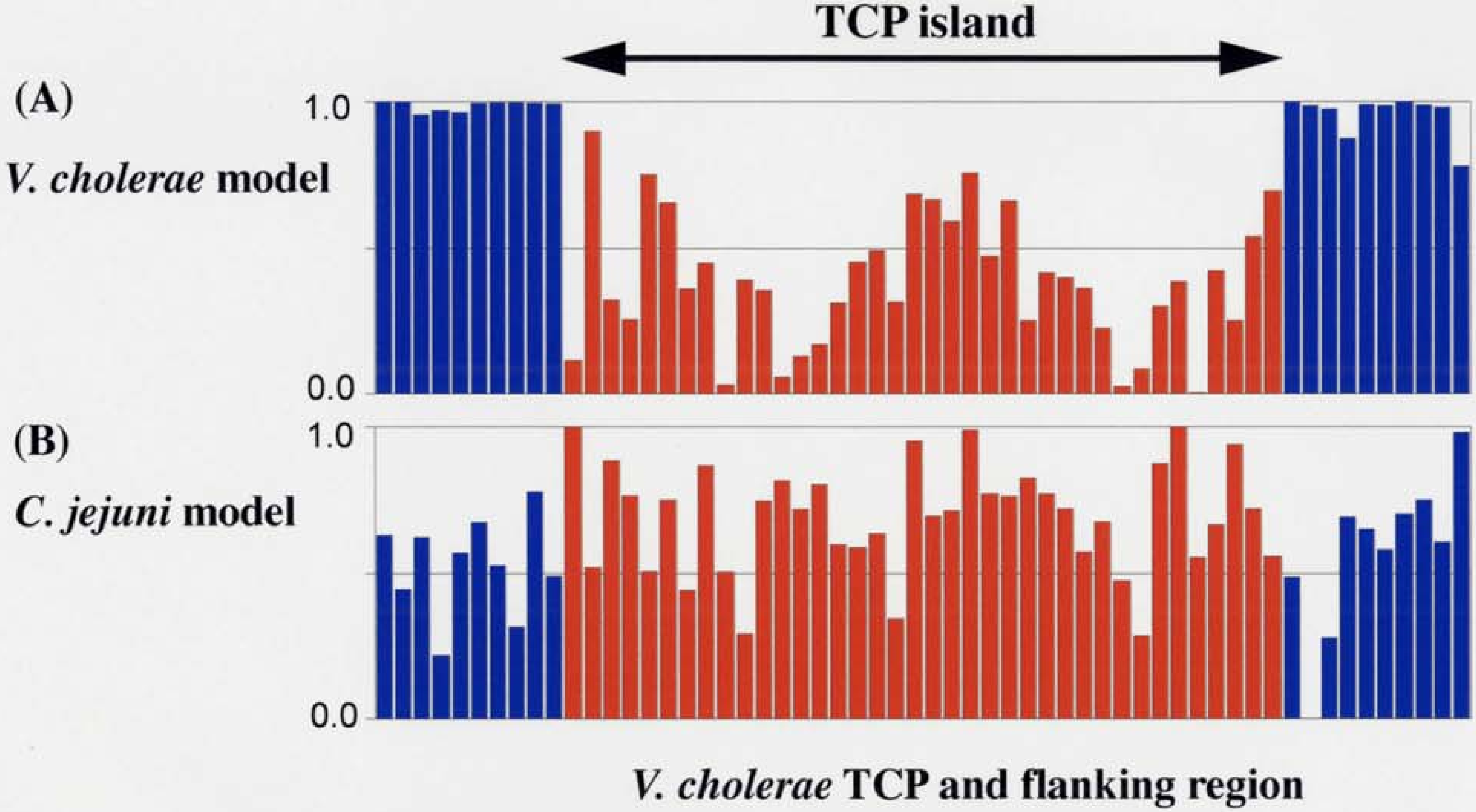


Figure 3.9 HT indices in *Vibrio cholerae* TCP island using *Vibrio cholerae* model (A) and *Campylobacter jejuni* model (B). Red bars show the HT indices of genes in TCP island. Flanking 10 genes from these genes to both directions are shown in the figures (blue bars).

3.1.6 Relationship between horizontally transferred genes and their possible vectors (plasmids, bacteriophages)

In the previous section, I implicitly assumed that genes of HT clusters originated from those encoded in the chromosomal genomes. However, it is possible that HT genes had been located on plasmids or bacteriophages for a long time. Even if not, foreign DNA sequences are thought to be transferred mainly by means of plasmids or bacteriophages. Hence, my method may detect a plasmid or bacteriophage origin of HT genes. To detect these HT genes, I first split a complete genome sequence into horizontally transferred regions and non-transferred regions, and constructed two separate training models (HT model, and non-HT model). I then computed and compared HT indices of genes encoded in plasmid and bacteriophage genomes using both models (**Table 3.3**). In most species, the HT model was able to predict plasmid or phage genes more effectively than the non-HT model. These observations are consistent with the argument that transferred genes were acquired by plasmids or bacteriophages. Exceptionally, in the case of *Borrelia burgdoriferi* plasmids, all of HT indices are higher with the non-HT model than with the HT model. This implies that these plasmids have stayed in the cell for a long time, and that their nucleotide compositions have become similar to those of the chromosomes.

A symbiotic bacterium *Mesorhizobium loti* has a giant region in its chromosome, termed "symbiotic island", required for the symbiosis with leguminous plants. Some of genes in this region are similar to those in two large plasmids (mega-plasmids) of *M. loti*, named pMLa and pMLb both of which are larger than 100kb (**Kaneko *et al.* 2000**). Since these plasmids have enough coding or non-coding regions to construct their training models, I computed HT donor indices of the *M. loti* symbiotic island with the training models of these plasmids. As expected, the symbiotic island was preferentially detected by pMLa model (**Figure 3.10**) as well as by pMLb model (data not shown).

Table 3.3 HT indices of plasmid/bacteriophage genes using HT model and non-HT model

	Host organism (species for the trainig model)	The number of plasmids/phages examined[1]	$HI_{HT} > HI_{non\text{-}HT}$[2]
Plasmid	*Bacillus subtilis*	6	6
	Corynebacterium glutamuicum	5	2
	Escherichia coli *	19	19
	Helicobacter pylori *	4	4
	Lactococcus lactis	12	12
	Nostoc sp. PCC	6	5
	Salmonella enterica *	7	7
	Staphylocpoccus aureus ^	15	11,12,15
	Yersinia pestis	9	8
	Borrelia burgdoriferi	21	0
Phage	*Escherichia coli* *	6	6
	Lactococcus lactis	9	9
	Pseudomonas aeruginosa	4	4
	Staphylocpoccus aureus *	4	4
	Vibrio cholerae	4	3

1 I used plasmid/phage genomes having 4 or more genes.

2 HI_{HT} and $HI_{non\text{-}HT}$ show the averages of HT indices computed by HT gene model and non-HT gene model, respectively.

* Same results are obtained when the models of different strains are used.

^ Different results are obtained when the models of different strains are used. The strains are N315, MW2, Mu50 from the right to the left.

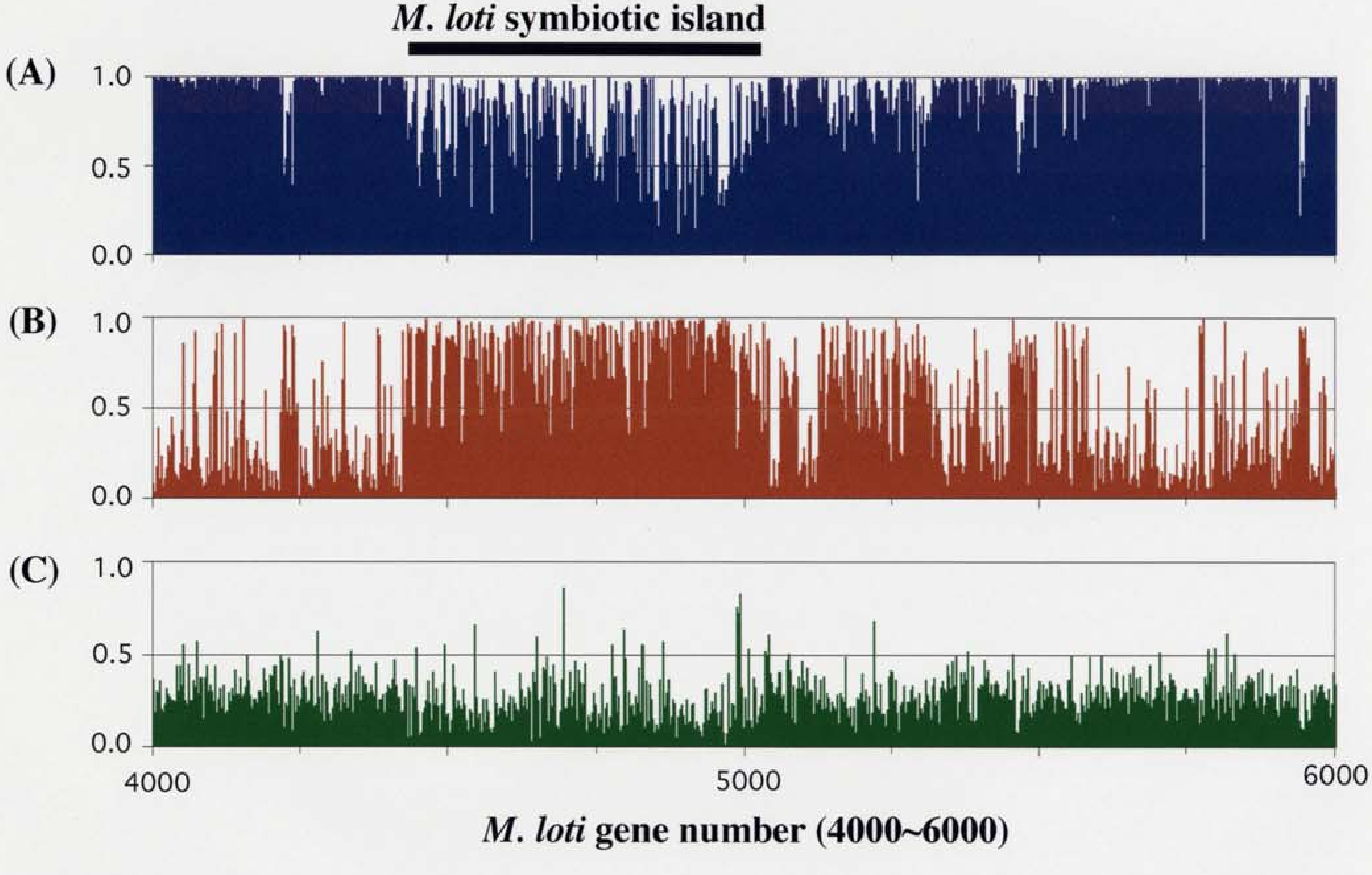


Figure 3.10 HT indices and HT donor indices in the symbiotic island of *Mesorhizobium loti*. (A) HT indices of genes by *M. loti* model itself. (B) HT donor indices by a *M. loti* megaplasmid pMLa model. (C) HT indices by a megaplasmid model of a distantly related species, *Ralstonia solanacearum*.

3.1.7 Comparison of performance with other methods

Table 3.4 shows the proportions of detected HT genes between my and Karlin's methods for the 18 species that Karlin and his colleagues previously surveyed (**Karlin 2001**). My method could detect more mobile element genes than Karlin's method under the same condition (all genes >= 300bp), which indicates that truth-positive ratio is better in my method than Karlin's one. One may think that this is due to the difference in the total numbers of detected genes between my and Karlin's methods (3149 and 1495, respectively). However, the false positive ratio represented by the number of detected ribosomal protein genes is not largely different between the two methods. These results together indicate that my method performed better than Karlin's method.

3.1.8 Database of horizontally transferred genes in complete genomes

In order to visualize the flow of horizontally transferred genes in all of complete genomes, I developed a database of horizontal gene transfer (HGT database) in collaboration with software engineers of Fujitsu Co.Ltd. Examples of the database contents are shown in **Figure 3.11 (A),(B)**.

An upper window in **Figure 3.11(A)** shows the circular genome map of *Vibrio cholerae*. In this circle, each blue bar shows a protein-encoding gene and the height from inside to outside indicates the HT indices (0 ~ 1). For example, the region where HT genes are densely located is present like a valley composed of lower blue peaks (**arrows**). A number of black bars expanding to inside show genes not encoding proteins, that is tRNA and rRNA genes and annotated pseudogenes. Next, by clicking a region on the chromosomal circle, it can display the region on a linear scale (**Figure 3.11(A): lower window**). Genes in red are candidates of horizontally transferred genes having significantly lower HT indices. Moreover, one can retrieve the annotation of the gene clicked on the linear map (**Figure 3.11(B)**).

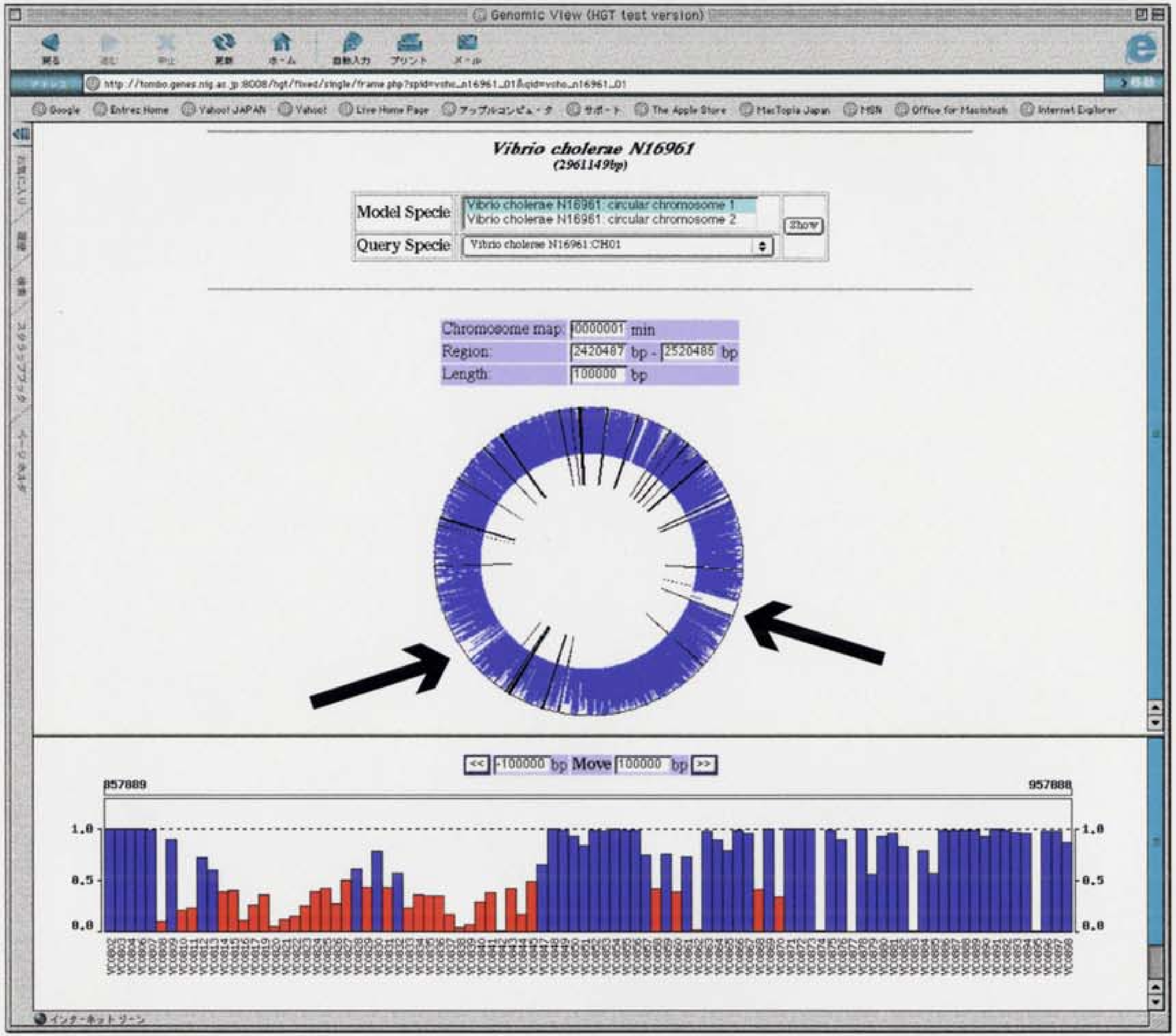
In addition to the database and interface, we developed a tool for inferring donor species of horizontally transferred genes (**Figure 3.12: a part of whole view**). Each column represents a gene in the species in red in rows. A color in each cell indicates the HT index that is computed with the model of any species.

Table 3.4 Sensitivities and specificities of my and Karlin's methods*

18 species	Our method	Karlin's method	Shared with both methods
Total detected genes	3149	1495	1065

	Our method	Karlin's method	Total
Mobile-element genes	124 (33.8%)	87 (23.7%)	367
Ribosomal protein genes	4 (0.58%)	3 (0.44%)	686

* All genes are >=300 bp long.



Genomic View (HGT test version)

http://tombo.genes.nig.ac.jp:8008/hgt/fixed/single/frame.php?spid=vcho_n16961_01&qid=vcho_n16961_01

Vibrio cholerae N16961

(2961149bp)

Model Specie: Vibrio cholerae N16961: circular chromosome 1 / Vibrio cholerae N16961: circular chromosome 2

Query Specie: Vibrio cholerae N16961:CH01

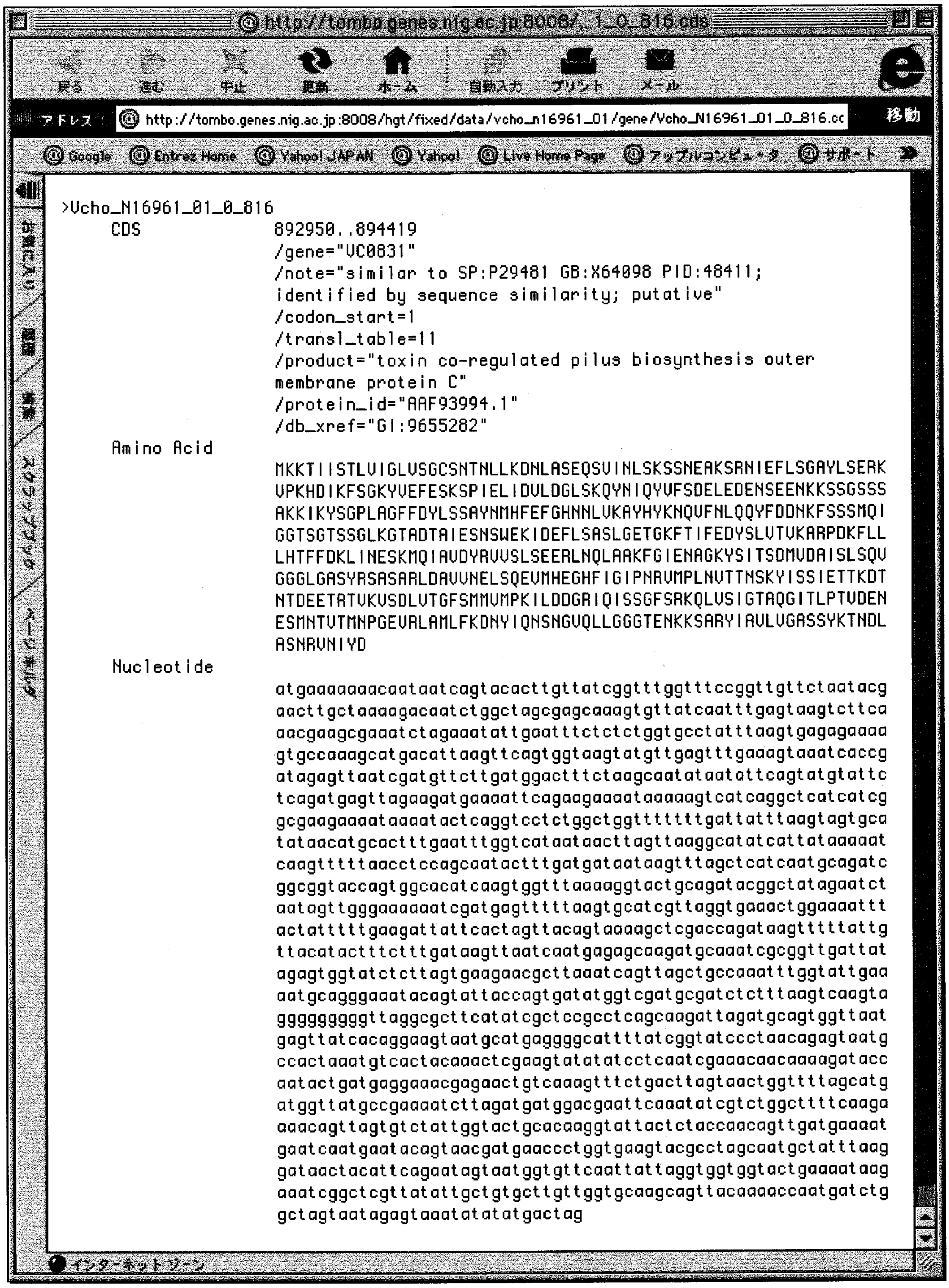
Chromosome map: 0000001 min

Region: 2420487 bp - 2520486 bp

Length: 100000 bp

<< -100000 bp **Move** 100000 bp >>

Figure 3.11(A) An appearance of HGT database



アドレス : http://tombo.genes.nig.ac.jp:8008/hgt/fixed/data/vcho_n16961_01/gene/Vcho_N16961_01_0_816.cc 移動

```
>Vcho_N16961_01_0_816
     CDS              892950..894419
                      /gene="VC0831"
                      /note="similar to SP:P29481 GB:X64098 PID:48411;
                      identified by sequence similarity; putative"
                      /codon_start=1
                      /transl_table=11
                      /product="toxin co-regulated pilus biosynthesis outer
                      membrane protein C"
                      /protein_id="AAF93994.1"
                      /db_xref="GI:9655282"
     Amino Acid
                      MKKTIISTLVIGLVSGCSNTNLLKDNLASEQSVINLSKSSNEAKSRNIEFLSGAYLSERK
                      VPKHDIKFSGKYVEFESKSPIELIDVLDGLSKQYNIQYVFSDELEDENSEENKKSSGSSS
                      AKKIKYSGPLAGFFDYLSSAYNMHFEFGHNNLVKAYHYKNQVFNLQQYFDDNKFSSSMQI
                      GGTSGTSSGLKGTADTAIESNSWEKIDEFLSASLGETGKFTIFEDYSLVTVKARPDKFLL
                      LHTFFDKLINESKMQIAVDYRVVSLSEERLNQLAAKFGIENAGKYSITSDMVDAISLSQV
                      GGGLGASYRSASARLDAVVNELSQEVMHEGHFIGIPNRVMPLNVTTNSKYISSIETTKDT
                      NTDEETATVKVSDLVTGFSMMVMPKILDDGRIQISSGFSRKQLVSIGTAQGITLPTVDEN
                      ESMNTVTMNPGEVALAMLFKDNYIQNSNGVQLLGGGTENKKSARYIAVLVGASSYKTNDL
                      ASNRVNIYD
     Nucleotide
                      atgaaaaaaacaataatcagtacacttgttatcggtttggtttccggttgttctaatacg
                      aacttgctaaaagacaatctggctagcgagcaaagtgttatcaatttgagtaagtcttca
                      aacgaagcgaaatctagaaatattgaatttctctctggtgcctatttaagtgagagaaaa
                      gtgccaaagcatgacattaagttcagtggtaagtatgttgagtttgaaagtaaatcaccg
                      atagagttaatcgatgttcttgatggactttctaagcaatataatattcagtatgtattc
                      tcagatgagttagaagatgaaaattcagaagaaaataaaaagtcatcaggctcatcatcg
                      gcgaagaaaataaaatactcaggtcctctggctggtttttttgattatttaagtagtgca
                      tataacatgcactttgaatttggtcataataacttagttaaggcatatcattataaaaat
                      caagttttaacctccagcaatactttgatgataataagtttagctcatcaatgcagatc
                      ggcggtaccagtggcacatcaagtggtttaaaaggtactgcagatacggctatagaatct
                      aatagttgggaaaaaatcgatgagtttttaagtgcatcgttaggtgaaactggaaaattt
                      actatttttgaagattattcactagttacagtaaaagctcgaccagataagttttattg
                      ttacatactttctttgataagttaatcaatgagagcaagatgcaaatcgcggttgattat
                      agagtggtatctcttagtgaagaacgcttaaatcagttagctgccaaatttggtattgaa
                      aatgcagggaaatacagtattaccagtgatatggtcgatgcgatctctttaagtcaagta
                      gggggggggttaggcgcttcatatcgctccgcctcagcaagattagatgcagtggttaat
                      gagttatcacaggaagtaatgcatgaggggcattttatcggtatccctaacagagtaatg
                      ccactaaatgtcactacaaactcgaagtatatatcctcaatcgaaacaacaaaagatacc
                      aatactgatgaggaaacgagaactgtcaaagtttctgacttagtaactggttttagcatg
                      atggttatgccgaaaatcttagatgatggacgaattcaaatatcgtctggcttttcaaga
                      aaacagttagtgtctattggtactgcacaaggtattactctaccaacagttgatgaaaat
                      gaatcaatgaatacagtaacgatgaaccctggtgaagtacgcctagcaatgctatttaag
                      gataactacattcagaatagtaatggtgttcaattattaggtggtggtactgaaaataag
                      aaatcggctcgttatattgctgtgcttgttggtgcaagcagttacaaaaccaatgatctg
                      gctagtaatagagtaaatatatatgactag
```

Figure 3.11(B) An appearance of HGT database (HT gene annotation)

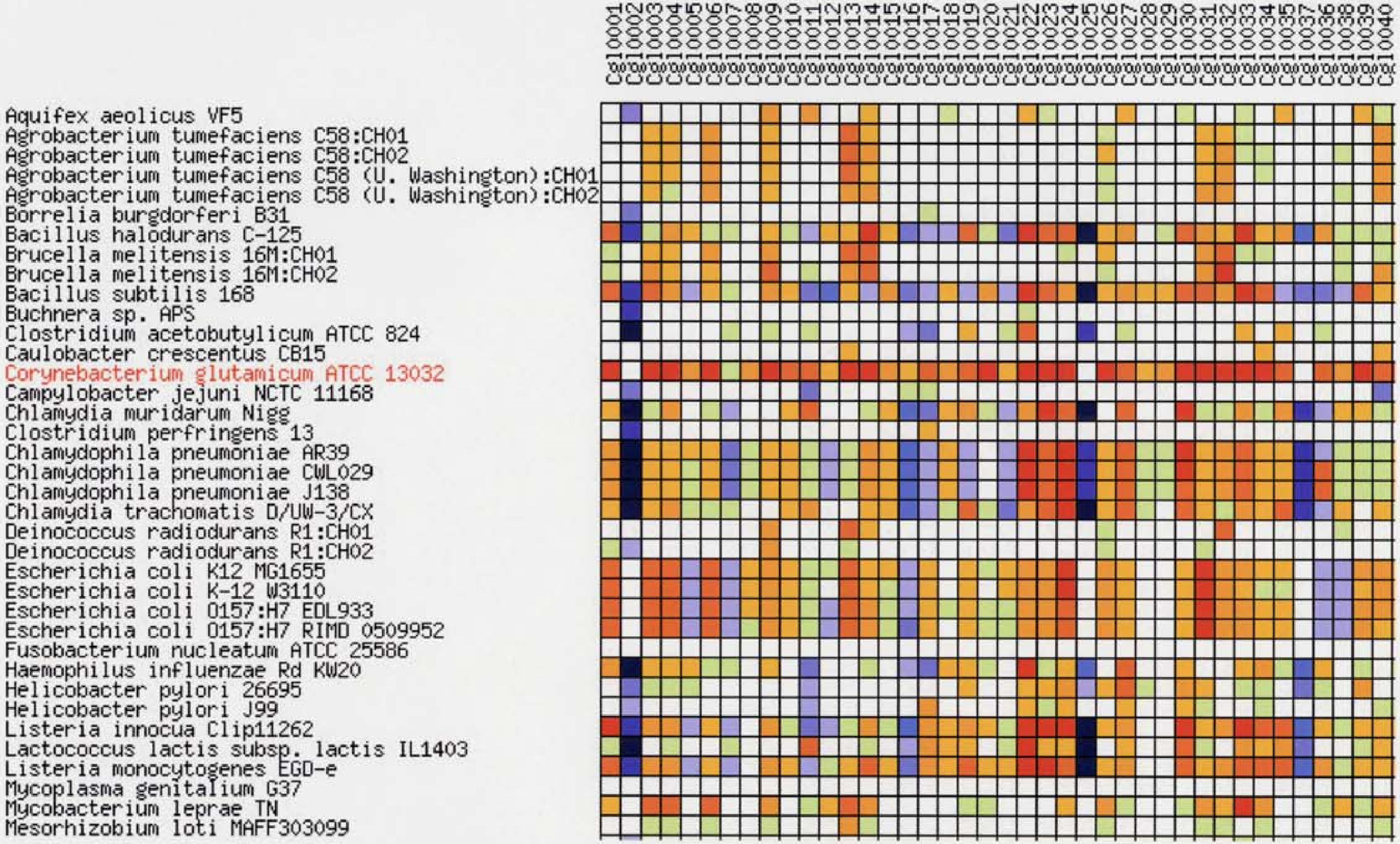


Figure 3.12 Estimation of donor species using multiple training models in HGT database (partial).

Here, possible donor(s) of *Corynebacterium glutamicum* genes (column) are estimated against multiple training models (row). Colors of cells in a row of *C. glitamicum* indicate HT indices using *C. glutamicum* model itself (gradation: white ~ red -> small HI ~ large HI*), those of heterogeneous species indicate HT indices using the model of the species (gradation: white ~ red -> small ~ large HT, light blue ~ deep blue -> small dHI ~ large dHI**).

* HI = HT index.

** dHI = (HI using heterogeneous species) - (HI using the original species)

3.2 Detection of horizontally transferred genes by extensive analysis of molecular phylogeny

3.2.1 Robustness of 4 species trees

Table 3.5(i) ~ (vi) shows the proportions of trees supported in the six groups with the bootstrap probability larger than 90%. These tables are representative parts of whole results, which are shown in a **Supplemental table 2-(i) ~ 2-(vi)**. As a tendency among all phylogenetic groups, the correct trees were preferred among almost all of possible species combinations, meaning that genes conserved among species were rarely exchanged (**see Supplemental table 2-(i) ~ 2-(vi)**). This result is due to two possibilities: one is that conserved genes may have functionally important roles, some of which may be defined as nontransferrable "core" genes in "core hypothesis" (**Nesbo *et al.* 2001**). These are involved in characteristic or essential traits for defining the taxonomic groups. Another possibility is that the sequence similarity in the examined orthologous genes is insufficient for the gene exchange by homologous recombination. Although it was reported that intra-species gene exchanges by recombination occurred in the population of several species (**Graham & Istock 1979; Smith *et al.* 1993; Feil *et al.* 2001**), recombination efficiency between homologous sequences was reduced exponentially, as sequence homology decreased (**Zawadzki *et al.* 1995; Vulic *et al.* 1997; Majewski *et al.* 2000**). Therefore, gene exchange by homologous recombination among closely related species has rarely occurred. In fact, it has been reported that even pathogenicity genes that are often subject to horizontal transfer have rarely undergone gene exchange among pathovers within a species (**Sawada *et al.* 1999**).

However, for specific combinations in *Bacillus-Staphylococcus* group (*B. halodurans, B. subtilis,* any *Listeria,* and any *Staphylococcus*), incorrect trees caused by possibly intra-species gene exchanges were frequently observed (**Table 3.5(i); Supplemental table 2-(i)**). Since *B. halodurans* has an IS- and transposon-rich genome (**Takami *et al.* 2000; Takami *et al.* 2001**), this species may be able to easily retrieve foreign DNA sequences mediated by transposons. In the case of *Rhizobium* group, a small portion of intra-species exchange was detected (**Table**

3.5(vi); Supplemental table 2-(vi)). Since *Rhizibium* species have self-transmissible plasmids (**Goodner *et al.* 2001; Wood *et al.* 2001; Kaneko *et al.* 2000; Galibert *et al.* 2001**) some of which are required for symbiosis, these plasmids may have mediated the transfer among closely related species.

Incorrect trees were observed in the case where three out of the four OTUs were of the same species (*Staphylococcus aureus, Chlamydophila pneumoniae, Escherichia coli,* and *Streptococcus pyogenes*) (**Supplemental table 2-(i), 2-(ii), 2-(iv) and 2-(v)**). In this case, "uncertain" trees, which means that the length of internal branch is zero, were frequently observed, indicating that these species have hardly diverged. In the case of *E. coli*, where two O157 strains seem to be diverged from a K12 strain, a number of abnormal trees may have been caused by inter-species gene exchange.

3.2.2 Relationship with results from Bayes' estimation

In the previous section, I detected horizontally transferred genes in the prokaryote genomes independently by Bayes' estimation. I then investigated the correspondence between genes detected by Bayes' estimation and those examined in the four OTU analysis. In this study, a gene in a species examined should have zero, one, two, or three orthologues with the other three species. Here, to have zero orthologue in the other species means that the gene is present only in that species, so called a species-specific gene, although incompletely similar homologs or possible paralogues might be present. To have three orthologues means that the gene is conserved among the all four species, although the four-orthologue condition (**see Section 2.4.1**) may not be completely satisfied.

I have found that in all lineages HT genes are detected as species-specific genes more than as multiple orthologues (**Figure 3.13**). The result strongly suggests that horizontal gene transfer is involved in the acquisition of novel genes that are not conserved among closely related species.

Table 3.5 (i) *Bacillus - Staphylococcus* group

Abbreviations:

Bha : ***Bacillus halodurans***
Bsu : ***Bacillus subtilis***
Lin : ***Listeria innocua***
Lmo : ***Listeria monocytogenes***
SauN : ***Staphylococcus aureus*** N315

(Bacillus halodurans, Bacillus subtilis,
Listeria innocua, Staphylococcus aureus* N315*)

Topology*	Tree**	%	BP>=90%***	%
(Bha, Bsu) - (Lin, SauN)	**576**	**68.8**	**292**	**85.9**
(Bha, Lin) - (Bsu, SauN)	99	11.8	13	3.8
(Bha, SauN) - (Bsu, Lin)	162	19.4	35	10.3
Uncertain****	0	0	0	0
Total	837		340	

(Bacillus halodurans, Bacillus subtilis,
Listeria monocytogenes, Staphylococcus aureus N315)

Topology	Tree	%	BP>=90%	%
(Bha, Bsu) - (Lmo, SauN)	**590**	**70.4**	**294**	**85.7**
(Bha, Lmo) - (Bsu, SauN)	93	11.1	14	4.1
(Bha, SauN) - (Bsu, Lmo)	155	18.5	35	10.2
Uncertain	0	0	0	0
Total	838		343	

(continued)

(Bacillus halodurans, Bacillus subtilis, Listeria innocua, Listeria monocytogenes)

Topology	Tree	%	BP >= 90%	%
(Bha, Bsu) - (Lin, Lmo)	**1096**	**99.9**	**1096**	**99.9**
(Bha, Lin) - (Bsu, Lmo)	0	0	0	0
(Bha, Lmo) - (Bsu, Lin)	1	0.091	1	0.091
Uncertain	0	0	0	0
Total	1097		1097	

(Bacillus subtilis, Listeria innocua, Listeria monocytogenes, Staphylococcus aureus **N315)**

Toplogy	Tree	%	BP>=90%	%
(Bsu, SauN) - (Lin, Lmo)	**983**	**100**	**982**	**100**
(Bsu, Lin) - (Lmo, SauN)	0	0	0	0
(Bsu, Lmo) - (Lin, SauN)	0	0	0	0
Uncertain	0	0	0	0
Total	983		982	

* A topology in bold is a correct tree based on 16SrRNA tree.
** Total number of obtained trees
*** Bootstrap probability >= 90% for 1000 replicates
**** Internal branch length = 0

Table 3.5 (ii) *Streptococcus* group

Abbreviation:

Lla : ***Lactococcus lactis***
Spn : ***Streptococcus pneumoniae***
SpnR : ***Streptococcus pneumoniae*** R6
Spy : ***Streptococus pyogens*** SF370
SpyM : ***Streptococus pyogens*** MGAS315

(Lactococcus lactis, Streptococcus pneumoniae,
Streptococus pyogens **SF370,** ***Streptococus pyogens*** **MGAS315)**

Topology*	Tree**	%	BP>=90%***	%
(Lla, Spn) - (Spy, SpyM)	**828**	**99.8**	**825**	**99.9**
(Lla, Spy) - (Spn, SpyM)	0	0.0	0	0.0
(Lla, SpyM) - (Spn, Spy)	2	0.2	1	0.1
Uncertain****	0	0.0	0	0.0
Total	830		826	

(Lactococcus lactis, Streptococcus pneumoniae,
Streptococus pneumoniae **R6,** ***Streptococus pyogens*** **SF370)**

Topology	Tree	%	BP>=90%	%
(Lla, Spy) - (Spn, SpnR)	**827**	**99.9**	**825**	**99.9**
(Lla, Spn) - (SpnR, Spy)	1	0.1	1	0.1
(Lla, Spy) - (Spn, SpnR)	0	0.0	0	0.0
Uncertain	0	0.0	0	0.0
Total	828		826	

* A topology in bold is a correct tree based on 16SrRNA tree.
** Total number of obtained trees
*** Bootstrap probability >= 90% for 1000 replicates
**** Internal branch length = 0

Table 3.5 (iii) Gram-positive high GC % group

Abbreviation:

Cgl : ***Corynebacterium glutamicum***
Mle : ***Mycobacterium leprae***
Mtu : ***Mycobacterium tuberculosis*** H37Rv
Sco : ***Streptomyces coelicolor***

(***Corynebacterium glutamicum, Mycobacterium leprae, Mycobacterium tuberculosis*** H37Rv, ***Streptomyces coelicolor***)

Topology*	Tree**	%	BP>=90%***	%
(Cgl, Sco) - (Mle, Mtu)	**706**	**97.9**	**698**	**98.9**
(Cgl, Mle) - (Mtu, Sco)	4	0.55	2	0.28
(Cgl, Mtu) - (Mle, Sco)	11	1.53	6	0.85
Uncertain****	0	0	0	0
Total	721		706	

* A topology in bold is a correct tree based on 16SrRNA tree.
** Total number of obtained trees
*** Bootstrap probability >= 90% for 1000 replicate
**** Internal branch length = 0

Table 3.5 (iv) *Chlamydia* group

Abbreviation:

Cpn : ***Chlamydophila pneumoniae*** CWL029
CpnA : ***Chlamydophila pneumoniae*** AR39
Ctr : ***Chlamydia trachomatis***
Cmu : ***Chlamydia muridarum***

(***Chlamydophila pneumoniae*** CWL029, ***Chlamydophila pneumoniae*** AR39, ***Chlamydia trachomatis, Chlamydia muridarum***)

Topology*	Tree**	%	BP>=90%***	%
(Cpn, CpnA) - (Ctr, Cmu)	**692**	**100**	**692**	**100**
(Cpn, Ctr) - (CpnA, Cmu)	0	0	0	0
(Cpn, Cmu) - (CpnA, Ctr)	0	0	0	0
Uncertain****	0	0	0	0
Total	692		692	

* A topology in bold is a correct tree based on 16SrRNA tree.
** Total number of obtained trees
*** Bootstrap probability >= 90% for 1000 replicate:
**** Internal branch length = 0

Table 3.5 (v) Enterobacteria and its relatives group

Abbreviation:

Eco : ***Escherichia coli*** K12
EcoO : ***Escherichia coli*** O157
Sty : ***Salmonella typhi***
Stym : ***Salmonella typhimurium***
Vch : ***Vibrio cholerae***
Ype : ***Yersinia pestis***

(***Escherichia coli*** **K12,** ***Salmonella typhimurium, Vibrio cholerae, Yersinia pestis***)

Topology*	Tree**	%	BP>=90%***	%
(Eco, Stym) - (Vch, Ype)	**1509**	**99.3**	**1466**	**99.9**
(Eco, Vch) - (Stym, Ype)	6	0.4	1	0.07
(Eco, Ype) - (Stym, Vch)	4	0.3	1	0.07
Uncertain****	0	0.0	0	0.00
Total	1519		1468	

(***Escherichia coli*** **K12,** ***Escherichia coli*** **O157,** ***Salmonella typhi, Salmonella typhimurium***)

Topology	Tree	%	BP>=90%	%
(Eco, EcoO) - (Sty, Stym)	**2883**	**99.6**	**2879**	**99.7**
(Eco, Sty) - (EcoO, Stym)	8	0.28	7	0.24
(Eco, Stym) - (EcoO, Sty)	1	0.03	0	0.00
Uncertain	2	0.07	2	0.07
Total	2894		2888	

* A topology in bold is a correct tree based on 16SrRNA tree.
** Total number of obtained trees
*** Bootstrap probability >= 90% for 1000 replicates
**** Internal branch length = 0

Table 3.5 (vi) *Rhizobium* group

Abbreviation:

Atu : ***Agrobacterium tumefaciens***
Bme : ***Brucella melitensis***
Mlo : ***Mesorhizobium loti***
Sme : ***Sinorhizobium meliloti***

(Agrobacterium tumefaciens, Brucella melitensis, Mesorhizobium loti, Sinorhizobium meliloti)

Topology*	Tree**	%	BP>=90%***	%
(Atu, Sme) - (Bme , Mlo)	**1353**	**89.1**	**1165**	**95.1**
(Atu, Bme) - (Mlo, Sme)	92	6.06	40	3.27
(Atu, Mlo) - (Bme , Sme)	74	4.87	20	1.63
Uncertain****	0	0	0	0
Total	1519		1225	

* A topology in bold is a correct tree based on 16SrRNA tree.
** Total number of obtained trees
*** Bootstrap probability >= 90% for 1000 replicates
**** Internal branch length = 0

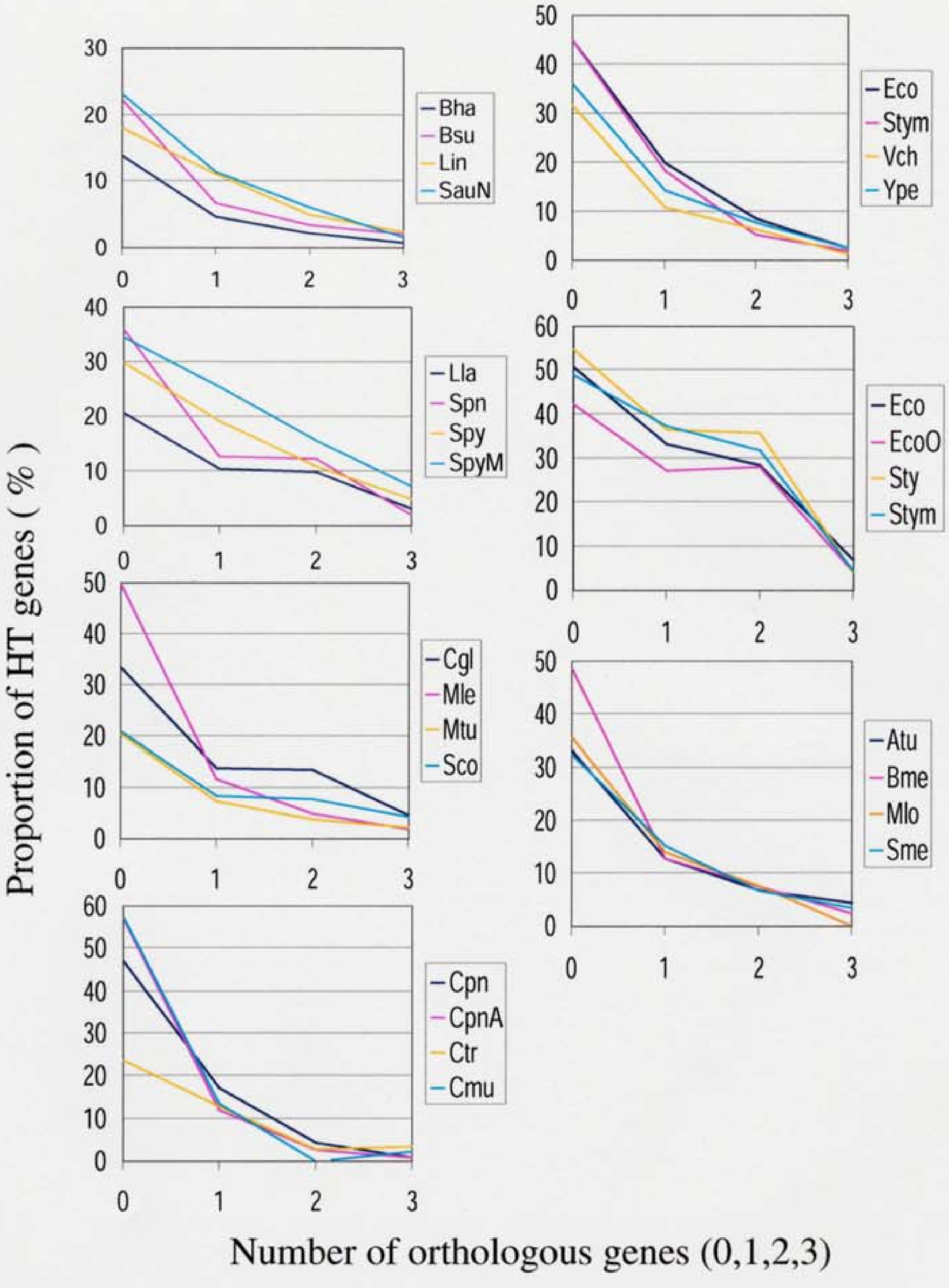


Figure 3.13 Relationship between the number of orthologous genes and the proportion of HT genes. Used species are the same as in **Table 3.5 (i)~(vi)**.

3.3 Genome comparison between *Corynebacterium* species

3.3.1 The features of three *Corynebacterium* genomes

In **Table 3.6**, I summarized the general features of three *Corynebacterium* genomes. It was previously reported that the GC content difference between *C. efficiens* and *C. glutamicum* was 5% (**Fudou *et al.* 2002**), but I have shown that it is actually about 10% (63.14 – 53.81 = 9.33%). I have identified 2,101 orthologous genes between *C. efficiens* and *C. glutamicum*, and found that 849 genes are *C. efficiens* specific and 998 genes are *C. glutamicum* specific. In the same way, I have identified 1,552 orthologous genes between *C. efficiens* and *C. diphtheriae*, and 1,587 genes between *C. glutamicum* and *C. diphtheriae.* I have then detected 580 candidates of horizontally transferred genes in the *C.efficiens* genome, and the proportion (19.7%) is similar to that of *C. glutamicum* (571 candidates: 18.4%). However, 415 out of 580 genes (71.6%) detected as horizontally transferred genes in *C. efficiens* were not present in *C. glutamicum,* and this occupied about a half of *C. efficiens*-specific genes (849 genes). These results suggest that a substantial number of species-specific genes have been acquired by horizontal transfer, as shown in the previous section (**Section 3.2**).

3.3.2 Amino acid substitution between *C. efficiens* and *C. glutamicum*

I compared the codon usage pattern between *C. efficiens* and *C. glutamicum*, and have found that lysine in *C. glutamicum* was frequently substituted to arginine in *C. efficiens* (**Table 3.7**). This substitution is known to increase protein stability because of the resonance effect of arginine (**Vieille & Zeikus 2001**). Thus, biased substitutions to arginine in *C. efficiens* can explain very well its thermostability. Furthermore, I examined the number of synonymous substitutions that had occurred in the orthologous genes between *C. efficiens* and *C. glutamicum.*

Table 3.6 Genomic features of *C.efficiens* , *C.glutamicum* , *C.diphtheriae*

	C.efficiens	***C.glutamicum***	***C.diphtheriae***
genome size (bp)	3,147,090	3,309,401	2,488,635*
G + C content (%)	63.14	53.81	53.48
number of coding regions	2,950	3,099	2,757**
horizontally transferred gene***	580(19.7%)	571(18.4%)	----

* Last updated 01-Nov-2001.

** Result by only glimmer predcition.

*** Result by Bayes' estimation.

Table 3.7 Biased amino acid substitutions in the orthologous genes between *C. glutamicum* and *C. efficiens*

C. glutamicum	*C. efficiens*	Point
Lys	Arg	1356.5
Ser	Arg	695.5
Ile	Val	593.0
Ser	Thr	591.5
Gln	Arg	406.5
Ile	Leu	406.0
Asn	Asp	374.0
Ser	Gly	312.5
Ser	Pro	255.0
Lys	Thr	250.5

Note: Point is defined as the difference between the number of amino acidsubstitutions from C. glutamicum to C. efficiens and that in the opposite direction, devided by 2.

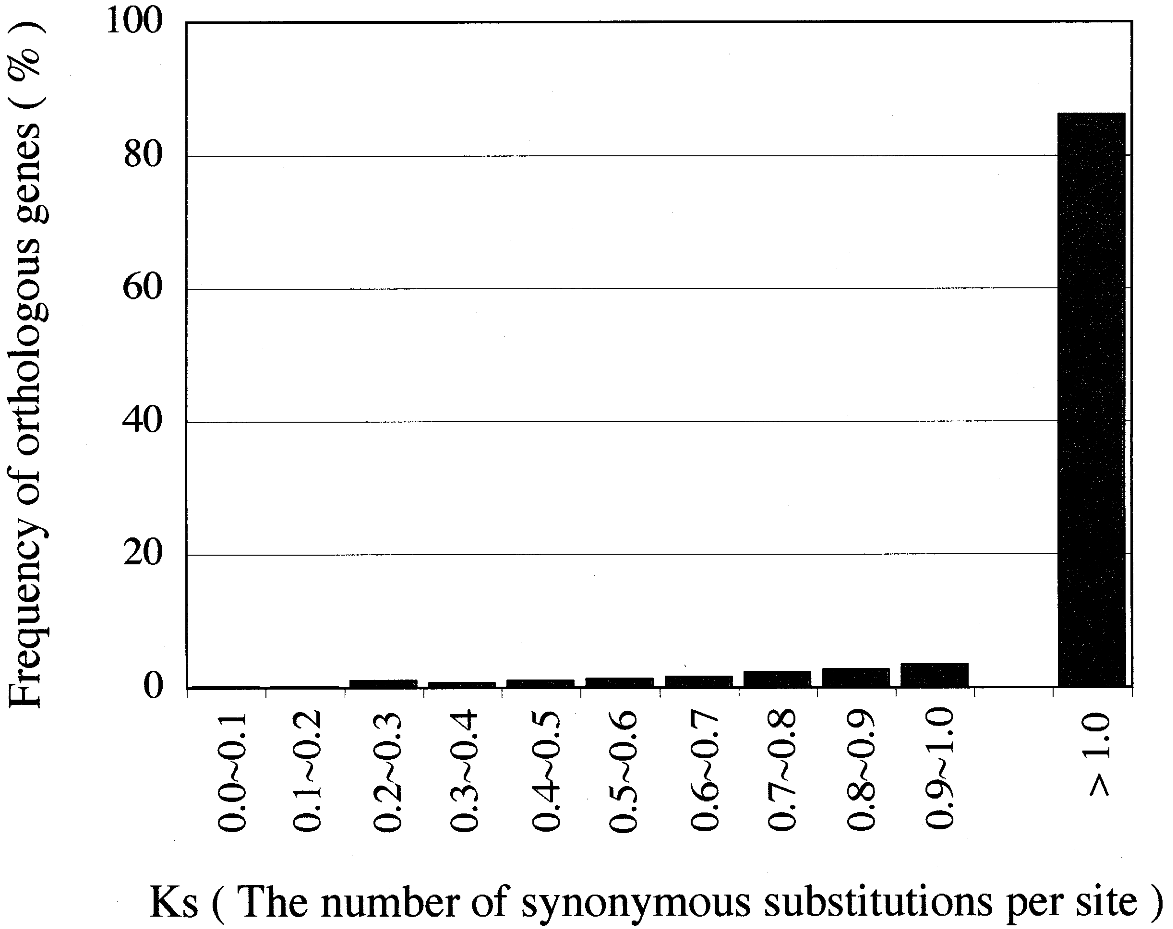


Figure 3.14 Ks between *C. efficiens* and *C. glutamicum* orthologues.

The result shows that in most of the orthologous genes (86.1%), synonymous substitution has occurred more than 1 time per site (**Figure 3.14**). This suggests that these two species have diverged distantly enough to have multiple substitutions in the orthologous genes.

3.3.3 GC contents and GC skew on a whole genome scale

I analyzed the GC contents and GC skews in *C. efficiens, C. glutamicum,* and *C. diphtheriae* genomes on the whole scale (**Figure 3.15; Figure 3.16 (A),(B)andC)**). Apparently, *C. efficiens* has shown a higher GC content, and the GC skew is not so clearly observed like those of the other two genomes, implying that GC skew of *C. efficiens* is being eliminated.

In 1967, Cox and Yanofsky (**Cox & Yanofsky 1967**) revealed that *E. coli* mutator gene named *mutT* increased the frequency of transversion from an AT to a CG pair, and as a result increased the GC content of *E. coli*. This means that GC contents of prokaryotic genomes can easily be affected by mutator genes such as *mutT*. Therefore, I surveyed known mutator genes which can preferentially change the GC content (A or T ⇔ G or C) (**Horst *et al.* 1999**) among corynebacteria.

Interestingly enough, *C. efficiens* was lacking the very *mutT* gene in spite of the presence in the other two corynebacteria (**Table 3.8**). Since a defective *mutT* allele increases GC content in a genome, my observation is consistent with that *C. efficiens* has a higher GC content than *C. glutamicum* and *C. diphtheriae. C. diphtheriae* does not possess a *mutT* homolog that is conserved among *C. efficiens*, *C. glutamicum,* and *M. tuberculosis*. Since the *mutT* homolog is not similar to the *E.coli mutT* whose function was experimentally examined (**Table 3.8**), its function remains to be examined.

As for the distribution of *mutT* among the three corynebacteria, there are two possibilities: (1) a common ancestor of *Corynebacterium* genus had *mutT* and only *C. efficiens* has recently lost the *mutT*, (2) a common ancestor of *Corynebacterium* genus did not have *mutT*, and *C. glutamicum* and *C. diphtheriae* have gained *mutT* independently. At present, possibility (1) is

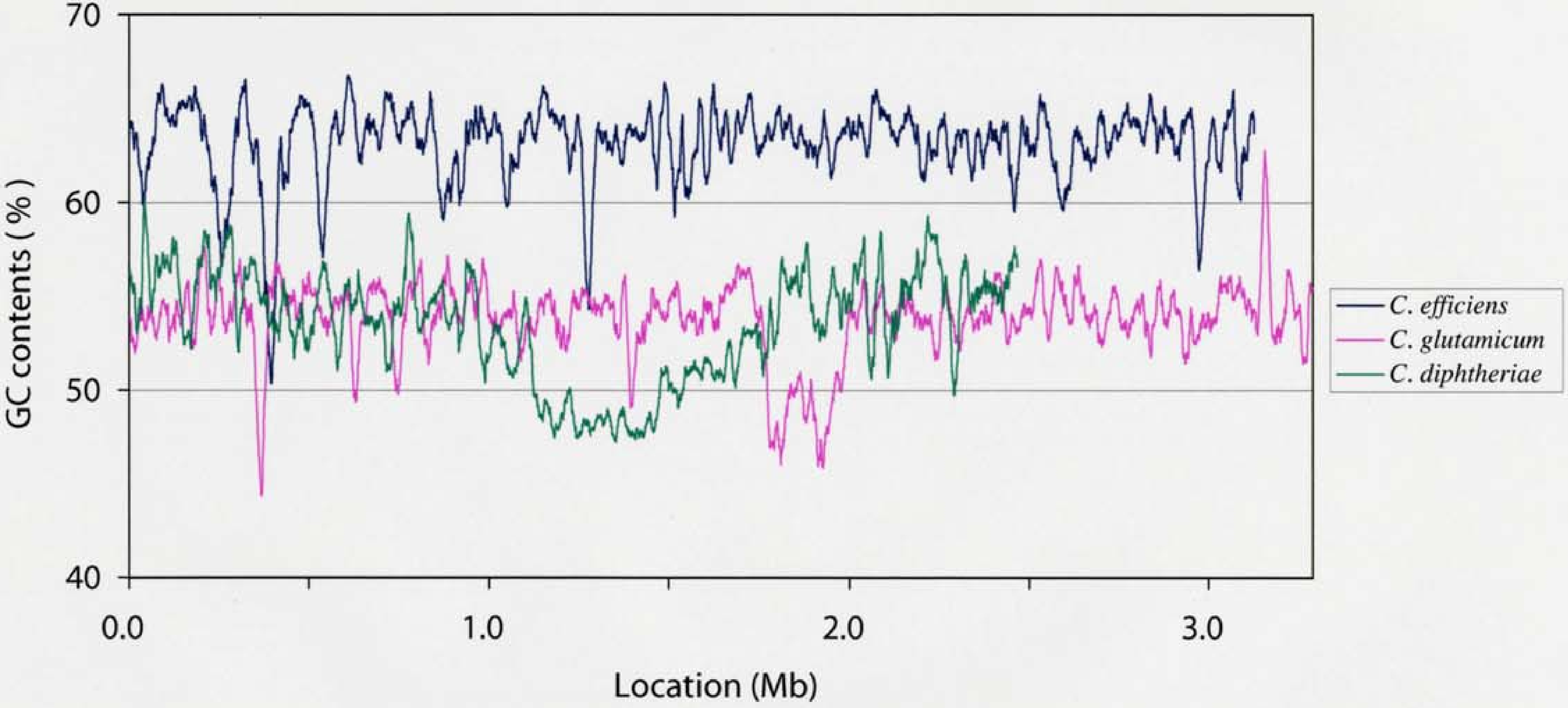


Figure 3.15 GC contents in corynebacteria genomes

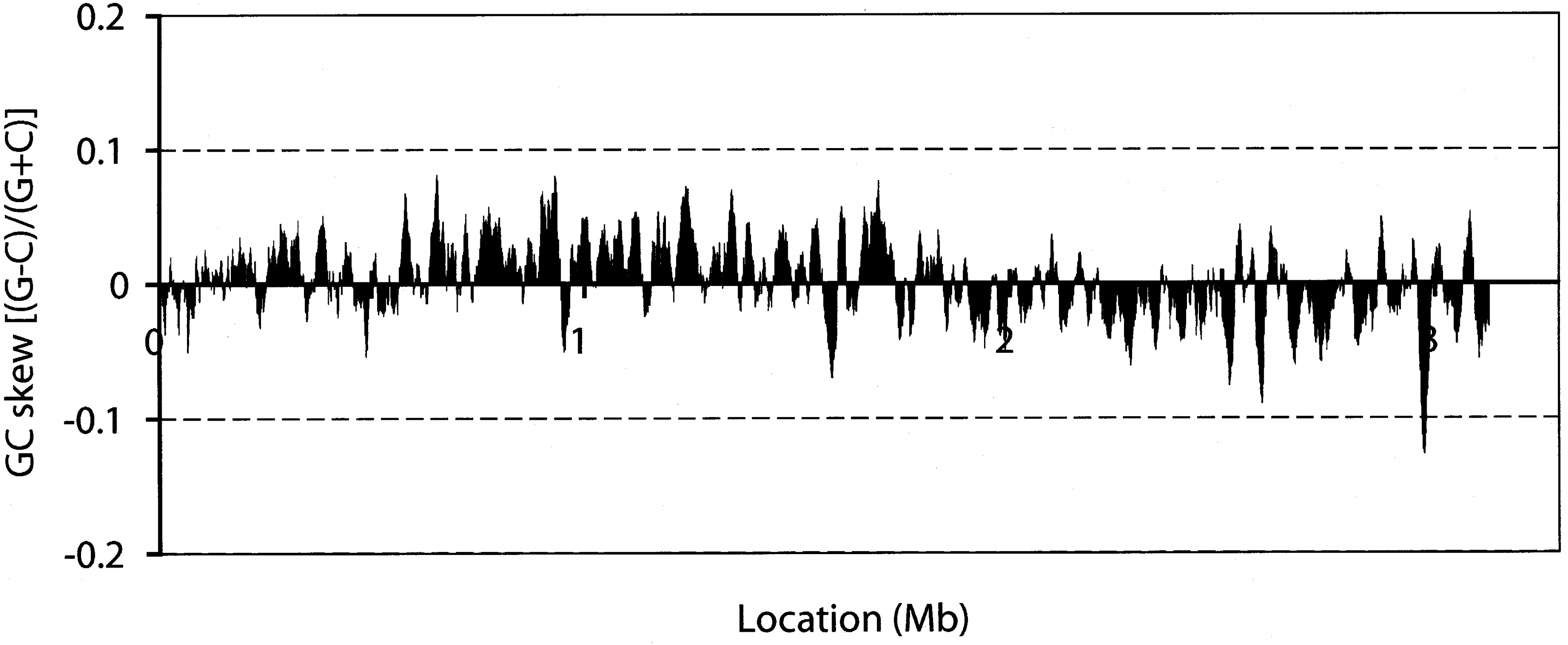


Figure 3.16 (A) GC skew in *C. efficiens* genome

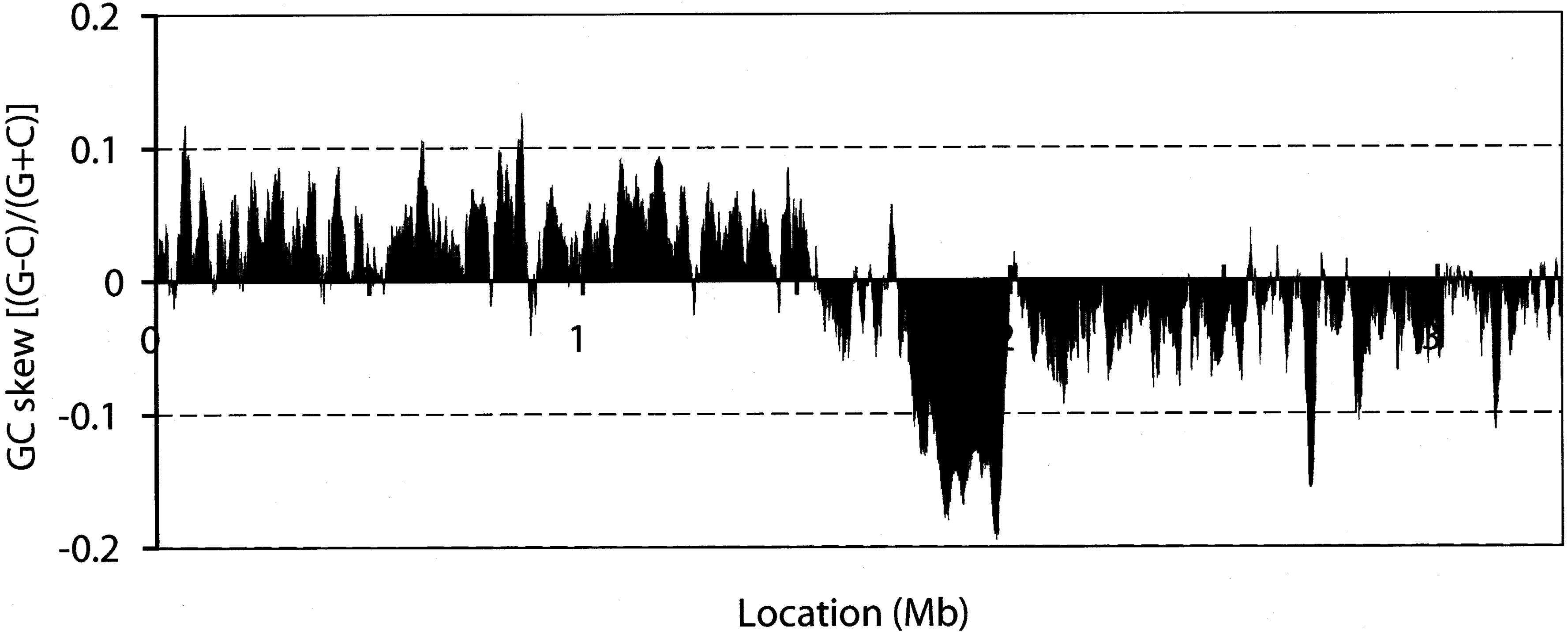


Figure 3.16 (B) GC skew in *C. glutamicum* genome

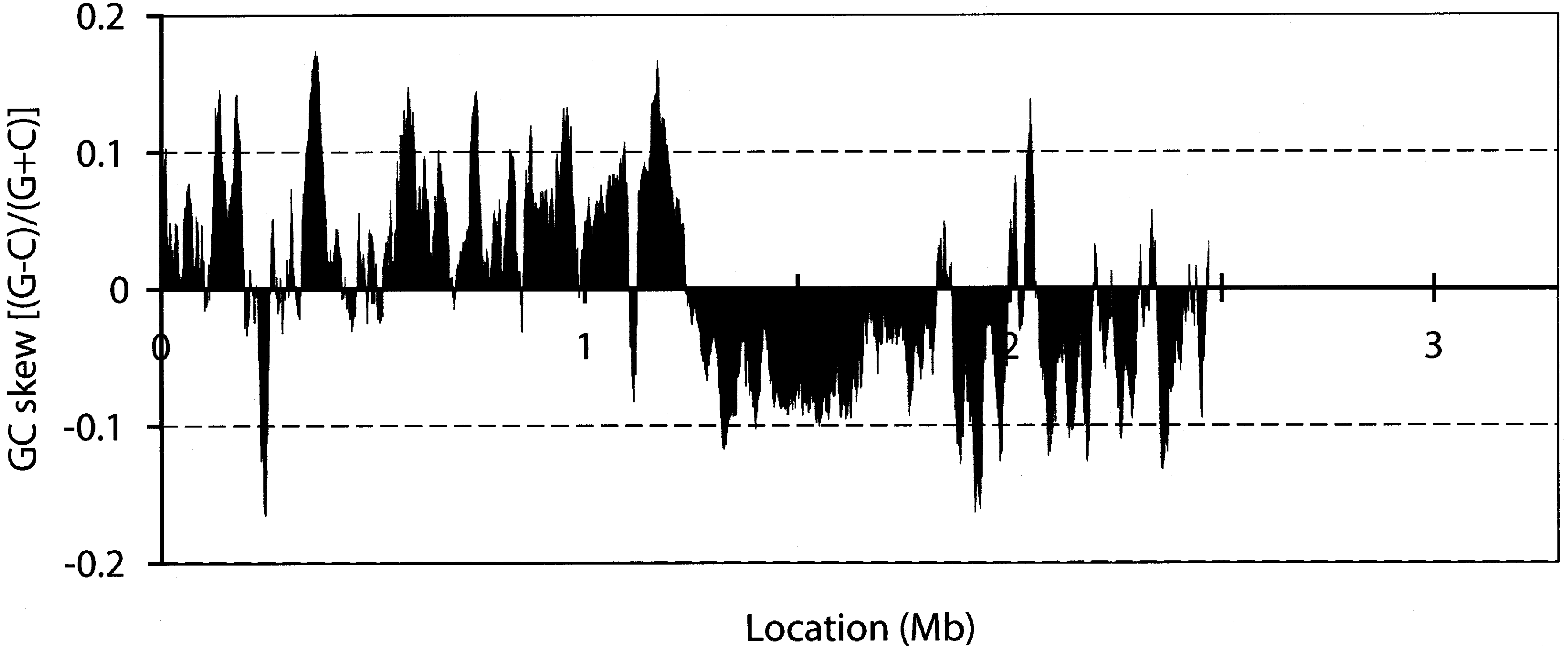


Figure 3.16 (C) GC skew in *C. diphtheriae* genome

Table 3.8 Mismatch repair (MMR) genes in corynebacreria

Gene	C.eff	C.glu	C.diph	M.tub	E.coli	Specificity	Product
ada	+	+	+	-	+	GC->AT	O6-methylguanine-DNA methyltransferase
miaA	+	+	+	+	+	GC->TA	delta(2)-isopentenylpyrophosphate tRNA-adenosinetransferase
mutM	+	+	+	+	+	GC->TA	formamidopyrimidine DNA glycosylase
mutT	-	+	+	+	+	AT->CG	7,8-dihydro-8-oxoguanine-triphosphatase
mutT -like	+	+	+	+	-		
mutT -like	+	+	-	+	-		
mutY	+	+	+	+	+	GC->TA	adenine glycosylase
nei	+	+	+	+	+	GC->AT	endonuclease VIII and DNA N-glycosylase with anAP lyase activity
nth	+	+	+	+	+	GC->AT	endonuclease III; specific for apurinic and/orapyrimidinic sites
ogt	-	-	-	+	+	GC->AT	O-6-alkylguanine-DNA/cysteine-proteinmethyltransferase
recA	+	+	+	+	+	GC->TA, AT->TA	DNA-dependent ATPase, DNA- and ATP-dependent coprotease
ung	+	+	+	+	+	GC->AT	uracil-DNA-glycosylase
vsr	-	-	-	-	+	GC->AT	DNA mismatch endonuclease, patch repair protein

Abbreviation:

C.eff = ***C.efficiens***
C.glu = ***C.glutamicum***
C.diph = ***C.diphtheriae***
M.tub = ***M.tuberculosis***

more likely than (2) in parsimony, although only three species were examined. Moreover, the recent loss of *mutT* in *C. efficiens* may give an explanation on the elimination of the GC skew in the genome (**Figure 3.16 (A)**), meaning that the GC skew is now changing gradually after the recent loss of *mutT*. Thus, I propose that the loss of *mutT* in *C. efficiens* has contributed to the increase in its GC content to some extent.

3.3.4 Rare genome rearrangement among corynebacteria

I compared the order of 2,101 orthologous genes between *C. efficiens* and *C. glutamicum*. Interestingly, the order of the orthologous genes is highly conserved between the two genomes (**Figure 3.17 (A)**). Discontinuities in large regions are associated with the presence of genes related to transposons or bacteriophages in the regions, strongly suggesting that the regions have been acquired from other organisms. This synteny was also observed in the orthologous genes between *C. efficiens* and *C. diphtheriae* as well as *C. glutamicum* and *C. diphtheriae* (**Figure 3.17 (B) and (C)**).

These results strongly suggest that *Corynebacterium* species have hardly undergone genome rearrangement on a large scale, although horizontal gene transfer has occurred many times. The genome stability may be one of the features of *Corynebacterium* genus, because frequent rearrangements were observed between *C. efficiens* and its outgroup *Mycobacterium tuberculosis* (**Figure 3.17 (D)**) as well as between *M. tuberculosis* and *M. leprae* (**Tillier & Collins 2000**).

Since genome rearrangement in an organism represents exchanges and shuffles of DNA segments in the chromosome(s), it is considered to have some connection with recombinational repair systems in a species. Therefore, I examined the presence or absence of genes related to chromosomal recombinations in *C. efficiens, C. glutamicum* and *C. diphtheriae* genomes. I have then found that the distribution patterns of recombinational repair genes are different between the three corynebacteria and *M. tuberculosis* (**Table 3.9**). A remarkable difference is that *M. tuberculosis* has *recBCD* required for RecBCD pathway, but the three corynebacteria do not have them. Mahan and Roth (**Mahan & Roth 1989**) examined the functions of *recBC* in *E. coli* and suggested that these proteins stimulated chromosomal inversions. Therefore, it is possible that lacking of *recBCD* enhanced the genome stability in corynebacterium species, alternatively, the acquisition of these genes reduced the genome stability in mycobacterium species.

Another difference is that *M. tuberculosis* does not have any *recQ* homolog encoding DNA helicase, but the corynebacteria do. The experiments using *E. coli* suggested that *recQ* was involved in RecFOR pathway for homologous recombination and that the mutation of this gene caused chromosomal instability because of illegitimate recombination (**Hanada *et al.* 1997**). The DNA helicase genes identified in three corynebacteria were distantly related to *E. coli recQ*, but still homologous to RecQ family genes. In general, RecQ family genes in eukaryotes such as human and drosophila also have some important roles in chromosomal recombination repair (**Bjergbaek *et al.* 2002; Cobb *et al.* 2002; Wu & Hickson 2002**). Thus, a RecQ family gene in corynebacteria may have affected their genome stability compared with mycobacteria.

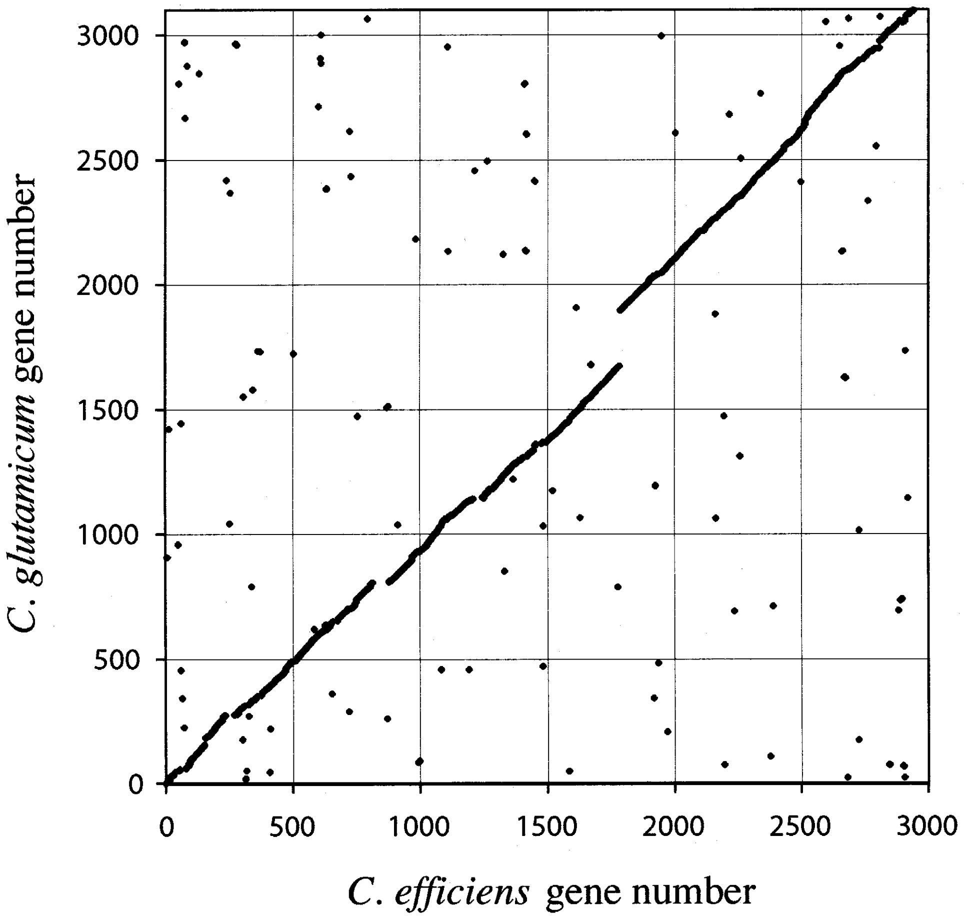


Figure 3.17 (A) Orthologous genes between *C. efficiens* and *C. glutamicum*

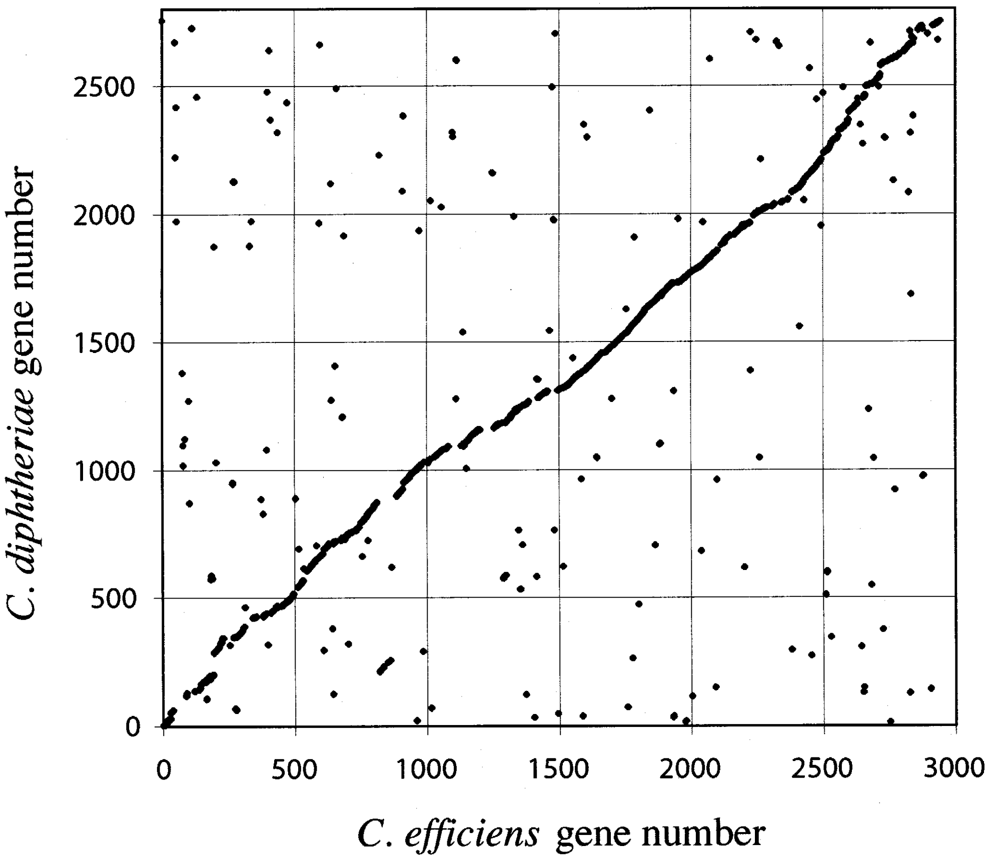


Figure 3.17 (B) Orthologous genes between *C. efficiens* and *C. diphtheriae*

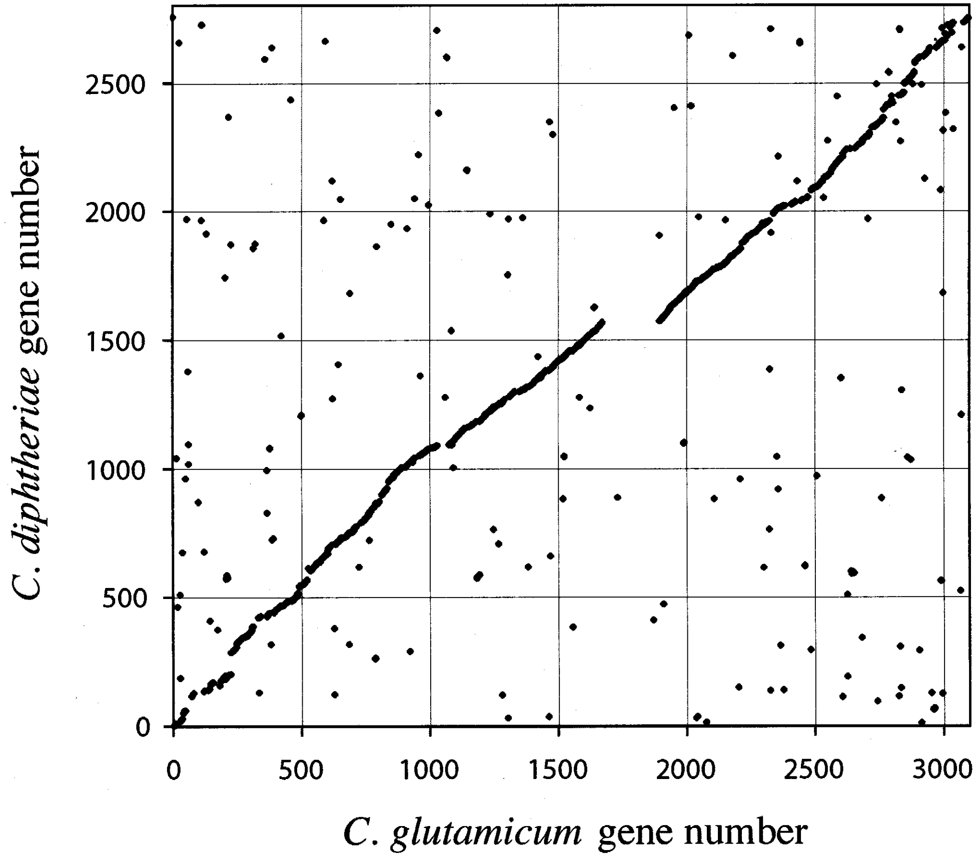


Figure 3.17 (C) Orthologous genes between *C. glutamicum* and *C. diphtheriae*

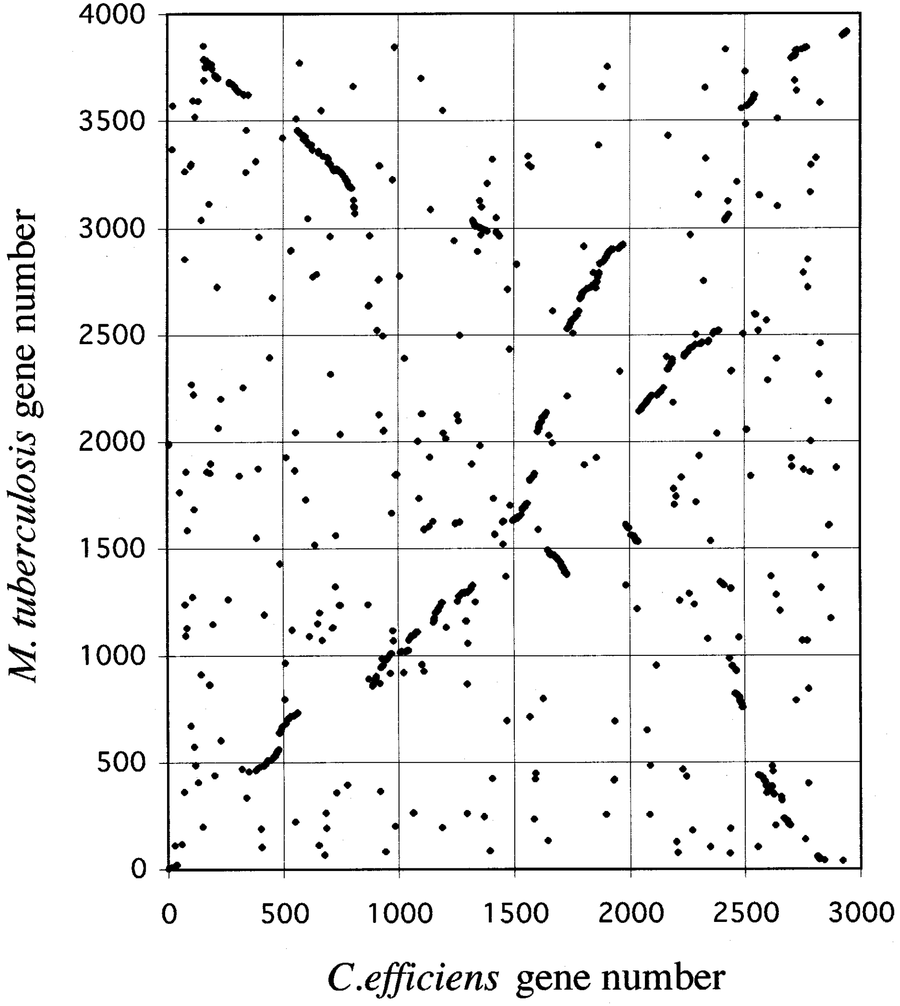


Figure 3.17 (D) Orthologous genes between *C. efficiens* and *M. tuberculosis*

Table 3.9 Recombinational repair genes in corynebacteria

Pathway and genes	C.eff	C.glu	C.diph	M.tub	E.coli	Product
RecBCD pathway						
recB	-	-	-	+	+	DNA helicase, ATP-dependent dsDNA/ssDNAexonuclease V subunit, ssDNA endonuclease
recC	-	-	-	+	+	DNA helicase, ATP-dependent dsDNA/ssDNAexonuclease V subunit, ssDNA endonuclease
recD	-	-	-	+	+	DNA helicase, ATP-dependent dsDNA/ssDNAexonuclease V subunit, ssDNA endonuclease
RecF pathway						
recF	+	+	+	+	+	ssDNA and dsDNA binding, ATP binding
recJ	-	+	-	-	+	ssDNA exonuclease, 5' --> 3' specific
recN(radB)	+	+	+	+	+	protein used in recombination and DNA repair
recO	+	+	+	+	+	protein interacts with RecR and possibly RecFproteins
recQ (family)	+*	+*	+*	-	+	ATP-dependent DNA helicase
recR	+	+	+	+	+	recombination and repair
RecE pathway						
recE	-	-	-	-	+	exonuclease VIII, ds DNA exonuclease, 5' --> 3'specific
recT	-	-	-	-	+	recombinase
SbcBCD pathway						
sbcB	-	-	-	-	+	exonuclease I, 3' --> 5' specific;deoxyribophosphodiesterase
sbcC	-	-	-	-	+	ATP-dependent dsDNA exonuclease
sbcD	-	-	-	-	+	ATP-dependent dsDNA exonuclease
AddAB pathway						
addA	-	-	-	-	-	ATP-dependent deoxyribonuclease (subunit A) alternate gene name: recE5
addB	-	-	-	-	-	ATP-dependent deoxyribonuclease (subunit B)

(continued)

Branch migration / resolution						
recG	+	+	+	+	+	DNA helicase, resolution of Holliday junctions,branch migration
rus	-	-	-	-	+	endodeoxyribonuclease RUS (Holliday junction resolvase)
ruvA	+	+	+	+	+	Holliday junction helicase subunit B; branch migration; repair
ruvB	+	+	+	+	+	Holliday junction helicase subunit A; branch migration; repair
ruvC	+	+	+	+	+	Holliday junction nuclease; resolution of structures; repair
Recombinase						
recA	+	+	+	+	+	DNA-dependent ATPase, DNA- and ATP-dependent coprotease
Site-specific recombination						
xerC	+	+	+	+	+	site-specific recombinase
xerD	+	+	+	+	+	site-specific recombinase
Other related genes						
lexA	+	+	+	+	+	LexA repressor
lig	+	+	+	+	+	DNA ligase
oraA(recX)	+	+	+	+	+	putative regulator, recA inhibitor
polA	+	+	+	+	+	DNA polymerase I, 3' --> 5' polymerase, 5' -->3' and 3' --> 5' exonuclease
priA	+	+	+	+	+	primosomal protein N'(= factor Y) (putative helicase)
radA(sms)	+	+	+	+	+	probable ATP-dependent protease
radC	-	-	-	-	+	DNA repair protein
ssb	+	+	+	+	+	ssDNA-binding protein

Abbreviation:

C.eff = ***C.efficiens***
C.glu = ***C.glutamicum***
C.diph = ***C.diphtheriae***
M.tub = ***M.tuberculosis***

* A homolog is found, but the similarity is weak and phylogenetically distant from ***E. coli recQ*** .

4. Summary

4.1 Recent gene transfer revealed by Bayes' estimation

Until now several *in silico* methods for detecting transferred genes have been developed, but these methods are often lacking the information for modeling, based on ambiguous statistics or complicated algorithms. Since my method is based on a plain and clear statistics and accommodates more precise information (5th-order Markov chains of non-coding regions and six frames of coding regions; see Methods) than other *in silico* methods based on GC content or codon usage, it is expected that my method would be much more effective. In fact, I have successfully found novel candidates of horizontally acquired clusters. The truth-positive ratio in detection in my method is better than other *in silico* methods with the false-positive ratio unaffected, although the assumptions required for evaluation are somewhat hypothetical. I have also shown that many of transferred genes have important roles such as pathogenesis, antibiotics-resistance, cell surface, gene network, and adaptation to the environment. Moreover, my method has an advantage in that possible donor species can be identified. I have not been able to clearly identify donors for all of HT genes. One of the reasons for this is limitation of the database at present. Another possibility is that HT genes might have rapidly diverged after the introgression into new hosts. In particular, the latter case may be related to the observation that HT clusters are often located on evolutionarily unstable regions where frequent genome rearrangements had occurred. Actually, a number of HT clusters detected in two *Neisseria meningitidis* and in two *Xanthomonas* species are located in such rearranged regions. Thus, my approach will give the basis for understanding the evolution of bacterial genomes form the view of horizontal gene transfer.

4.2 Gene transfer revealed by phylogenetic analysis

As another approach to detect horizontal gene transfer, I analyzed phylogenetic trees for closely related species, particularly using four orthologous gene trees in six taxonomic groups. The result has shown that the orthologous genes conserved among all of the four species are rarely subject to gene exchange among the species. One possible reason is that conserved genes among species are essential to some degree for their life. The mobility of such genes may thus cause the instability of the biological activity in the cells and be selected against. Another possibility is that once the species have diverged, recombinational barrier prevents gene exchange even if the species are phylogenetically close to each other.

On the other hand, the horizontally transferred genes detected by Bayes' estimation are frequently species-specific genes that are not conserved among species. This strongly indicates that horizontal gene transfer is involved in the acquisition of novel genes absent in the lineage. Such gene-gain events may have accelerated the differentiation of species. Moreover, the mechanism enhancing horizontal gene transfer among distantly related species is apparently different from that of homologous recombination that is responsible for inter-species gene exchanges. My hypothesis for explaining the difference is that transferred genes have rarely been originated immediately from the donor species by the direct conjugation, which would need homologous recombination between the donor and recipient DNA segments. Transferring genes may have retained on extra-chromosomal replicons such as plasmids for a long time and may have occasionally inserted into genomes on the coattails of transposons often present in plasmids.

4.3 Genome rearrangement

I analyzed the *C. efficiens* genome for investing the evolutionary mechanism of prokaryotic genomes. Since its closely related species, *C. glutamicum, C. diphtheriae*, and *M. tuberculosis* have been sequenced, I used the information of these complete genomes as reference.

First, I estimated the proportion of horizontally transferred genes in *C. efficiens*, and compared it with that of *C. glutamicum*. The proportions are similar to each other, but most of the horizontally transferred genes in *C. efficiens* were different from those in *C. glutamicum*. This indicates that horizontal gene transfer is involved in the acquisition of novel genes, as mentioned in the previous section (**Section 4.2**).

Second, nucleotide compositions, such as GC content and GC skew, were different between *C. efficiens* and both of *C. glutamicum* and *C. diphtheriae*. In comparison of amino acid composition, I have detected biased patterns of amino acid substitutions between *C. efficiens* and *C. glutamicum*, suggesting the enhanced thermostability of the *C. efficiens* genome. In particular, the fact that a copy of mutT is lacking only in *C. efficiens* may well explain the increase in the GC content and the elimination of the GC skew in *C. efficiens*.

Third, I have found that *Corynebacterium* species have stable genomes with respect to the order of orthologous genes compared with *Mycobacterium* species, belonging to the same order *Actinomyces*. As for the genome stability from the view of gene order, Suyama and Bork (**Suyama & Bork 2001**) have surveyed 21 pairs of closely related species. They found that the degree of gene order disruption showed a positive and almost linear correlation with the divergence time. *Mycoplasma* and *Chlamydia*, both obligate parasites, are the exceptions against the tendency, and they argued that loss of genes required for DNA replication and repair, such as *recG*, was involved in genome rearrangement. Their idea seems reasonable, because *Mycoplasma* and *Chlamydia* have relatively small genomes and might have undergone genome reduction by losing genes including *recG*.

I have found for the first time that the gene order is highly conserved among free-living bacteria such as corynebacteria. Furthermore, the three corynebacteria examined here possess considerable number of genes containing *recG*, meaning that other explanation than by *recG* is required for genome rearrangement. Fortunately, the *Mycobacterium* genome with drastic disorders of orthologous genes was available as reference genome. Therefore, the direct comparison between the corynebacterium and mycobacterium genomes might reveal the mechanism of genome rearrangement in *Actinomyces*. Here, I have proposed that *recBC* genes and RecQ family genes are responsible for genome stability, where the former is involved in chromosomal inversions, while the latter may have affected the homologous or illegitimate recombinations.

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor, Dr. Takashi Gojobori for his guidance during my Ph.D. work. Next, I would like to thank Dr. Yoshio Tateno for his advice to my study. I also express my gratitude to Drs. Hideaki Sugawara, Naruya Saitou, Hideo Matsuda, and Akiko Nishimura. I would like to thank Dr. Takeshi Itoh for his guidance.

I would like to thank two assistant professors in the laboratory for DNA Data Analysis, Dr. Kazuho Ikeo and Dr. Yoshiyuki Suzuki, who gave precise advice to my study and presentation. I thank Hisakazu Iwama, who kindly discussed basic mathematics and statistics with me. I also thank Yousuke Nishio of Ajinomoto Co.,Ltd., Naoki Nishinomiya and system engineers of DDBJ for their collaboration and help. I thank Dr. Kinya Ota for his friendship. I express special thanks to Kousuke Hanada and Atsushi Ogura for friendly discussion. I am grateful for help by everyone in the laboratory.

Finally, I would like to dedicate this thesis to my parents.

References

Ajdic D *et al.* (2002) Genome sequence of *Streptococcus mutans* UA159, a cariogenic dental pathogen. Proc Natl Acad Sci U S A. 99, 14434-14439.

Akman L *et al.* (2002) Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia.* Nat Genet. 32, 402-407.

Alm RA *et al.* (1999a) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori.* Nature. 397, 176-180.

Alm RA *et al.* (1999b) Analysis of the genetic diversity of *Helicobacter pylori*: the tale of two genomes. J Mol Med. 77, 834-846.

Altschul SF *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.

Andersson SG *et al.* (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature. 396, 133-140.

Baba T *et al.* (2002) Genome and virulence determinants of high virulence community-acquired MRSA. Lancet. 359, 1819-1827.

Bao Q *et al.* (2002) A complete sequence of the *T. tengcongensis* genome. Genome Res. 12, 689-700.

Bentley SD *et al.* (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature. 417, 141-147.

Beres SB *et al.* (2002) Genome sequence of a serotype M3 strain of group A *Streptococcus*: phage-encoded toxins, the high-virulence phenotype, and clone emergence. Proc Natl Acad Sci U S A. 99, 10078-10083.

Bjergbaek L *et al.* (2002) RecQ helicases and genome stability: lessons from model organisms and human disease. Swiss Med Wkly. 132, 433-442.

Blattner FR *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. Science. 277, 1453-1474.

Borodovsky, M. & McIninch, J.D. (1993) GENMARK: Parallel gene recognition for both DNA strands.
*Computers Chem.*17,123-133.

Borodovsky,M. et al. (1995) Detection of new genes in a bacterial genome using Markov models for three gene classes. *Nucleic Acid Res.* 23, 3554-3562

Bolotin A *et al.* (2001) The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. *lactis* IL1403. Genome Res. 11, 731-753.

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

Bult CJ *et al.* (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii.* Science. 273, 1058-1073.

Chambaud I *et al.* (2001) The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. Nucleic Acids Res. 29, 2145-2153.

Cobb JA *et al.* (2002) RecQ helicases: at the heart of genetic stability. FEBS Lett. 529, 43-48.

Cole ST *et al.* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature. 393, 537-544.

Cole ST *et al.* (2001) Massive gene decay in the leprosy bacillus. Nature. 409, 1007-1011.

Collins MD *et al.* (1986) Genus *Corynebacterium.* Bergey's Manual of Systematic Bacteriology. 2, 1266-1276.

Cox EC *et al.* (1967) Altered base ratios in the DNA of an *Escherichia coli* mutator strain. Proc Natl Acad Sci U S A. 58, 1895-1902.

da Silva AC *et al.* (2002) Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. Nature. 417, 459-463.

Davis J *et al.* (2001) Evolution of an autotransporter: domain shuffling and lateral transfer from pathogenic *Haemophilus* to *Neisseria.* J Bacteriol. 183, 4626-4635.

de la Cruz F *et al.* (2000) Horizontal gene transfer and the origin of species: lessons from bacteria. Trends Microbiol. 8, 128-133.

Deckert G *et al.* (1998) The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. Nature. 392, 353-358.

DelVecchio VG *et al.* (2002) The genome sequence of the facultative intracellular pathogen *Brucella melitensis*. Proc Natl Acad Sci U S A. 99, 443-448.

Denamur E *et al.* (2000) Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. Cell. 103, 711-721.

Deng W *et al.* (2002) Genome sequence of *Yersinia pestis* KIM. J Bacteriol. 184, 4601-4611.

Deppenmeier U *et al.* (2002) The genome of *Methanosarcina mazei*: evidence for lateral gene transfer between bacteria and archaea. J Mol Microbiol Biotechnol. 4, 453-461.

Eisen JA *et al.* (2002) The complete genome sequence of *Chlorobium tepidum* TLS, a photosynthetic, anaerobic, green-sulfur bacterium. Proc Natl Acad Sci U S A. 99, 9509-9514.

Feil EJ *et al.* (2001) Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc Natl Acad Sci U S A. 98, 182-187.

Ferretti JJ *et al.* (2001) Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. Proc Natl Acad Sci U S A. 98, 4658-4663.

Fitz-Gibbon ST *et al.* (2002) Genome sequence of the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum.* Proc Natl Acad Sci U S A. 99, 984-989.

Fleischmann RD *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science. 269, 496-512.

Fleischmann RD *et al.* (2002) Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. J Bacteriol. 184, 5479-5490.

Fraser CM *et al.* (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi.* Nature. 390, 580-586.

Fraser CM *et al.* (1995) The minimal gene complement of *Mycoplasma genitalium.* Science. 270, 397-403.

Fraser CM *et al.* (1998) Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. Science. 281, 375-388.

Fudou R *et al.* (2002) *Corynebacterium efficiens* sp. nov., a glutamic-acid-producing species from soil and vegetables. Int J Syst Evol Microbiol. 52, 1127-1131.

Galagan JE *et al.* (2002) The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity. Genome Res. 12, 532-542.

Galibert F *et al.* (2001) The composite genome of the legume symbiont *Sinorhizobium meliloti.* Science. 293, 668-672.

Glaser P *et al.* (2001) Comparative genomics of *Listeria* species. Science. 294, 849-852.

Glaser P *et al.* (2002) Genome sequence of *Streptococcus agalactiae*, a pathogen causing invasive neonatal disease. Mol Microbiol. 45, 1499-1513.

Glass JI *et al.* (2000) The complete sequence of the mucosal pathogen *Ureaplasma urealyticum.* Nature. 407, 757-762.

Goodner B *et al.* (2001) Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. Science. 294, 2323-2328.

Graham JP *et al.* (1979) Gene exchange and natural selection cause *Bacillus subtilis* to evolve in soil culture. Science. 204, 637-639.

Hacker J *et al.* (1997) Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol Microbiol. 23, 1089-1097.

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

Hanada K *et al.* (1997) RecQ DNA helicase is a suppressor of illegitimate recombination in *Escherichia coli.* Proc Natl Acad Sci U S A. 94, 3860-3865.

Hansmann S *et al.* (2000) Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. Int J Syst Evol Microbiol. 50 Pt 4, 1655-1663.

Hayashi T *et al.* (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res. 8, 11-22.

Heidelberg JF *et al.* (2000) DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae.* Nature. 406, 477-483.

Heidelberg JF *et al.* (2002) Genome sequence of the dissimilatory metal ion-reducing bacterium *Shewanella oneidensis.* Nat Biotechnol. 20, 1118-1123.

Himmelreich R *et al.* (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae.* Nucleic Acids Res. 24, 4420-4449.

Horst JP *et al.* (1999) *Escherichia coli* mutator genes. Trends Microbiol. 7, 29-36.

Hoskins J *et al.* (2001) Genome of the bacterium *Streptococcus pneumoniae* strain R6. J Bacteriol. 183, 5709-5717.

Huynen MA *et al.* 1998. Measuring genome evolution. Proc Natl Acad Sci U S A. 95, 5849-5856.

Jain R *et al.* (1999) Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A. 96, 3801-3806.

Jin Q *et al.* (2002) Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. Nucleic Acids Res. 30, 4432-4441.

Kalman S *et al.* (1999) Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis.* Nat Genet. 21, 385-389.

Kaneko T *et al.* (2000) Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti.* DNA Res. 7, 331-338.

Kaneko T *et al.* (2001) Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. DNA Res. 8, 205-13; 227.

Kaneko T *et al.* (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. 3, 109-136.

Kapatral V *et al.* (2002) Genome sequence and analysis of the oral bacterium *Fusobacterium nucleatum* strain ATCC 25586. J Bacteriol. 184, 2005-2018.

Karlin S *et al.* (1998) Codon usages in different gene classes of the *Escherichia coli* genome. Mol Microbiol. 29, 1341-1355.

Karlin S *et al.* (2001) Predicted highly expressed and putative alien genes of *Deinococcus radiodurans* and implications for resistance to ionizing radiation damage. Proc Natl Acad Sci U S A. 98, 5240-5245.

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

Kawarabayasi Y *et al.* (2001) Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* strain7. DNA Res. 8, 123-140.

Kawarabayasi Y *et al.* (1999) Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. DNA Res. 6, 83-101, 145.

Kawarabayasi Y *et al.* (1998) Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. DNA Res. 5, 55-76.

Kawashima T *et al.* (2000) Archaeal adaptation to higher temperatures revealed by genomic sequence of *Thermoplasma volcanium.* Proc Natl Acad Sci U S A. 97, 14257-14262.

Klenk HP *et al.* (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature. 390, 364-370.

Kroll JS *et al.* (1998) Natural genetic exchange between *Haemophilus* and *Neisseria*: intergeneric transfer of chromosomal genes between major human pathogens. Proc Natl Acad Sci U S A. 95, 12381-12385.

Kunst F *et al.* (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature. 390, 249-256.

Kuroda M *et al.* (2001) Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. Lancet. 357, 1225-1240.

Lawrence JG *et al.* (1998) Molecular archaeology of the *Escherichia coli* genome. Proc Natl Acad Sci U S A. 95, 9413-9417.

Lawrence JG. (2002) Gene transfer in bacteria: speciation without species? Theor Popul Biol. 61, 449-460.

Liebl W. (1991) The Genus *Corynebacterium*-Nonmedical. The Prokaryotes 2nd edition. 2, 1157-1171.

Maeder DL *et al*. (1999) Divergence of the hyperthermophilic archaea *Pyrococcus furiosus* and *P. horikoshii* inferred from complete genomic sequences. Genetics. 152, 1299-1305.

Mahan MJ *et al*. (1989) Role of *recBC* function in formation of chromosomal rearrangements: a two-step model for recombination. Genetics. 121, 433-443.

Majewski J *et al.* (2000) Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation. J Bacteriol. 182, 1016-1023.

May BJ *et al.* (2001) Complete genomic sequence of *Pasteurella multocida*, Pm70. Proc Natl Acad Sci U S A. 98, 3460-3465.

McClelland M *et al.* (2001) Complete genome sequence of *Salmonella enterica* serovar *Typhimurium* LT2. Nature. 413, 852-856.

Mrazek J *et al*. (2001) Highly expressed and alien genes of the *Synechocystis* genome. Nucleic Acids Res. 29, 1590-1601.

Mrazek J *et al.* (1999) Detecting alien genes in bacterial genomes. Ann N Y Acad Sci. 870, 314-329.

Nakamura Y *et al.* (2002) Complete genome structure of the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1. DNA Res. 9, 123-130.

Nei M *et al.* (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3, 418-426.

Nelson KE *et al.* (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima.* Nature. 399, 323-329.

Nesbo CL *et al.* (2001) Defining the core of nontransferable prokaryotic genes: the euryarchaeal core. J Mol Evol. 53, 340-350.

Ng WV *et al.* (2000) Genome sequence of *Halobacterium* species NRC-1. Proc Natl Acad Sci U S A. 97, 12176-12181.

Nierman WC *et al.* (2001) Complete genome sequence of *Caulobacter crescentus*. Proc Natl Acad Sci U S A. 98, 4136-4141.

Nolling J *et al.* (2001) Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum.* J Bacteriol. 183, 4823-4838.

Ochman H *et al.* (2000) Lateral gene transfer and the nature of bacterial innovation. Nature. 405, 299-304.

Ogata H *et al.* (2001) Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. Science.

293, 2093-2098.

Parkhill J *et al.* (2001a) Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar *Typhi* CT18. Nature. 413, 848-852.

Parkhill J *et al.* (2001b) Genome sequence of *Yersinia pestis*, the causative agent of plague. Nature. 413, 523-527.

Parkhill J *et al.* (2000a) The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. Nature. 403, 665-668.

Parkhill J *et al.* (2000b) Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. Nature. 404, 502-506.

Paulsen IT *et al.* (2002) The *Brucella suis* genome reveals fundamental similarities between animal and plant pathogens and symbionts. Proc Natl Acad Sci U S A. 99, 13148-13153.

Pearson WR *et al.* (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 85, 2444-2448.

Perna NT *et al.* (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature. 409, 529-533.

Peterson JD *et al.* (2001) The Comprehensive Microbial Resource. Nucleic Acids Res. 29, 123-125.

Read TD *et al.* (2000) Genome sequences of *Chlamydia trachomatis* MoPn and Chlamydia pneumoniae AR39. Nucleic Acids Res. 28, 1397-1406.

Riley M. (1993) Functions of the gene products of *Escherichia coli.* Microbiol Rev. 57, 862-952.

Rivera MC *et al.* (1998) Genomic evidence for two functionally distinct gene classes. Proc Natl Acad Sci U S A. 95, 6239-6244.

Robb FT *et al.* (2001) Genomic sequence of hyperthermophile, *Pyrococcus furiosus*: implications for physiology and enzymology. Methods Enzymol. 330, 134-157.

Ruepp A *et al.* (2000) The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum.* Nature. 407, 508-513.

Saitou N *et al.* (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4, 406-425.

Salanoubat M *et al.* (2002) Genome sequence of the plant pathogen *Ralstonia solanacearum.* Nature. 415, 497-502.

Sasaki Y *et al.* (2002) The complete genomic sequence of *Mycoplasma penetrans*, an intracellular bacterial pathogen in humans. Nucleic Acids Res. 30, 5293-5300.

Sawada H *et al.* (1999) Phylogenetic analysis of *Pseudomonas syringae* pathovars suggests the horizontal gene transfer of *argK* and the evolutionary stability of *hrp* gene cluster. J Mol Evol. 49, 627-644.

Schell MA *et al.* (2002) The genome sequence of *Bifidobacterium longum* reflects its adaptation to the human gastrointestinal tract. Proc Natl Acad Sci U S A. 99, 14422-14427.

Sharp, P.M. & Li,W.H. (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications.

Nucleic Acid Res. 15, 1281-1295.

She Q *et al.* (2001) The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. Proc Natl Acad Sci U S A. 98, 7835-7840.

Shigenobu S *et al.* (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. Nature. 407, 81-86.

Shimizu T *et al.* (2002) Complete genome sequence of *Clostridium perfringens*, an anaerobic flesh-eater. Proc Natl Acad Sci U S A. 99, 996-1001.

Shirai M *et al.* (2000) Comparison of whole genome sequences of *Chlamydia pneumoniae* J138 from Japan and CWL029 from USA. Nucleic Acids Res. 28, 2311-2314.

Simpson AJ *et al.* (2000) The genome sequence of the plant pathogen *Xylella fastidiosa.* The *Xylella fastidiosa* Consortium of the Organization for Nucleotide Sequencing and Analysis. Nature. 406, 151-157.

Slesarev AI *et al.* (2002) The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens. Proc Natl Acad Sci U S A. 99, 4644-4649.

Smith DR. *et al.* (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. J Bacteriol 1997 Nov;179(22):7135-55.,.

Smith JM *et al.* (1993) How clonal are bacteria? Proc Natl Acad Sci U S A. 90, 4384-4388.

Smoot JC *et al.* (2002) Genome sequence and comparative microarray analysis of serotype M18 group A *Streptococcus* strains associated with acute rheumatic fever outbreaks. Proc Natl Acad

Sci U S A. 99, 4668-4673.

Stephens RS *et al.* (1998) Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. Science. 282, 754-759.

Stover CK *et al.* (2000) Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. Nature. 406, 959-964.

Suyama M *et al.* (2001) Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends Genet. 17, 10-13.

Takami H. *et al.* (2002) Genome sequence of *Oceanobacillus iheyensis* isolated from the Iheya Ridge and its unexpected adaptive capabilities to extreme environments. Nucleic Acids Res. 30(18):3927-35. , .

Takami H *et al.* (2000) Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis*. Nucleic Acids Res. 28, 4317-4331.

Tamames J. (2001) Evolution of gene order conservation in prokaryotes. Genome Biol. 2, RESEARCH0020-RESEARCH0020.

Tamas I *et al.* (2002) 50 million years of genomic stasis in endosymbiotic bacteria. Science. 296, 2376-2379.

Tettelin H *et al.* (2002) Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. Proc Natl Acad Sci U S A. 99, 12391-12396.

Tettelin H *et al.* (2001) Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science. 293, 498-506.

Tettelin H *et al.* (2000) Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. Science. 287, 1809-1815.

Thompson JD *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673-4680.

Tillier ER *et al.* (2000) Genome rearrangement by replication-directed translocation. Nat Genet. 26, 195-197.

Tomb JF *et al.* (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori.* Nature. 388, 539-547.

Vieille C *et al.* (2001) Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. Microbiol Mol Biol Rev. 65, 1-43.

Vulic M *et al.* (1997) Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proc Natl Acad Sci U S A. 94, 9763-9767.

White O *et al.* (1999) Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science. 286, 1571-1577.

Wood DW *et al.* (2001) The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. Science. 294, 2317-2323.

Wu L *et al.* (2002) RecQ helicases and cellular responses to DNA damage. Mutat Res. 509, 35-

47.

Zawadzki P *et al.* (1995) The log-linear relationship between sexual isolation and sequence divergence in *Bacillus* transformation is robust. Genetics. 140, 917-932.

Supplemental table 1 Ongoing prokaryote genome projects (as of Dec.1, 2002)

Species name*	Domain**	Genome size (Kb)	Institution	Sequencing state***
Acidianus (Sulfolobus) brierleyi	A	unknown	Univ of Copenhagen	incomplete
Acidithiobacillus ferrooxidans ATCC 23270	B	2611	Integrated Genomics Inc	incomplete
Acidithiobacillus ferrooxidans ATCC-23270	B	2900	TIGR	incomplete
Acidobacterium capsulatum	B	unknown	TIGR	incomplete
Acinetobacter baumannii	B	unknown	Genome Therapeutics	complete
Acinetobacter calcoaceticus ADP1	B	3800	Genoscope	complete
Acinetobacter sp. ADP1	B	3583	Integrated Genomics Inc	complete
Actinobacillus actinomycetemcomitans HK1651	B	2105	Univ of Oklahoma	complete
Actinobacillus pleuropneumoniae serovar 1	B	2128	Univ of Oklahoma	incomplete
Actinobacillus pleuropneumoniae serovar 5	B	2159	Univ of Oklahoma	incomplete
Actinomyces naeslundii MG1	B	3000	TIGR	incomplete
Anabaena variabilis ATCC29413	B	8000	Joint Genome Institute, Univ of Missouri	incomplete
Anaplasma marginale	B	unknown	Amplicon Express Inc, ADRU	incomplete
Anaplasma (Ehrlichia sp. 'HGE agent'***) phagocytophilum*** HZ	B	1500	Ohio State Univ, TIGR	incomplete
Atopobium minutum ATCC 33267	B	1965	Integrated Genomics Inc	complete
Azotobacter vinelandii AvOP	B	4500	Univ of Arizona, Joint Genome Institute	incomplete
Bacillus anthracis Ames	B	5227	TIGR	incomplete
Bacillus anthracis Krugger B	B	5363	TIGR	incomplete
Bacillus anthracis WesternNA	B	unknown	TIGR	incomplete
Bacillus anthracis (Florida isolate) A2012	B	5093	TIGR	complete
Bacillus brevis	B	unknown	NITE, Tokyo Univ of Agriculture	incomplete
Bacillus cereus ATCC 10987	B	5200	TIGR	incomplete
Bacillus cereus ATCC 14579	B	5458	Integrated Genomics Inc, INRA	complete
Bacillus stearothermophilus 10	B	4250	Univ of Oklahoma	incomplete
Bacillus thuringiensis israelensis ATCC 35646	B	unknown	Integrated Genomics Inc	incomplete
Bacteroides fragilis	B	unknown	Genome Therapeutics	complete
Bacteroides fragilis NCTC9343 (+638R)	B	5200	Sanger Institute, Queen's Univ of Belfast, Univ of Wales, et al.	complete
Bacteroides thetaiotaomicron	B	unknown	Washington Univ, AstraZeneca	incomplete
Bacteroides (Tannerella) forsythus (forsythensis)	B	unknown	TIGR	incomplete
Bartonella henselae Houston 1	B	2000	Uppsala Univ	complete
Bartonella quintana Toulouse	B	1600	Uppsala Univ	complete
Bifidobacterium breve NCIMB8807	B	unknown	Univ College, Cork	incomplete
Bifidobacterium longum DJO10A	B	2100	Univ of Minnesota, Joint Genome Institute	incomplete
Bordetella bronchiseptica RB50 NCTC-13252	B	5340	Sanger Institute, Univ of Cambridge	complete
Bordetella parapertussis 12822 NCTC-13253	B	4770	Sanger Institute, Univ of Cambridge	complete
Bordetella pertussis Tohama I NCTC-13251	B	4090	Sanger Institute, Univ of Cambridge	complete
Borrelia hermsii	B	unknown	Univ of Minnesota	incomplete
Bradyrhizobium japonicum	B	unknown	Clemson Univ	incomplete

Bradyrhizobium japonicum USDA 110	B	10231	Integrated Genomics Inc	complete
Brevibacterium linens BL2	B	3000	Joint Genome Institute	incomplete
Brucella abortus	B	3287	IIB-UNSAM, Uppsala Univ, Univ of Minnesota	complete
Brucella melitensis 16M	B	unknown	Univ Notre-Dame De La Paix	incomplete
Burkholderia fungorum LB400	B	8000	Joint Genome Institute	incomplete
Burkholderia mallei ATCC 23344	B	6000	TIGR, USAMRIID	incomplete
Burkholderia pseudomallei K96243	B	7240	Sanger Institute, Porton Down	complete
Burkholderia vietnamiensis CF	B	9000	Joint Genome Institute, Michigan State Univ	incomplete
Burkholderia vietnamiensis G4	B	9000	Joint Genome Institute, Michigan State Univ	incomplete
Burkholderia vietnamiensis rhizosphere colonizer	B	9000	Joint Genome Institute, Michigan State Univ	incomplete
Burkholderia (Pseudomonas) cepacia J2315	B	7600	Sanger Institute, Cardiff Univ, Univ of Edinburgh, Univ Gent	incomplete
Campylobacter fetus	B	1500	IIB-UNSAM	complete
Campylobacter jejuni	B	unknown	Genome Therapeutics	complete
Campylobacter jejuni RM1221	B	1809	TIGR	incomplete
Carboxydothermus hydrogenoformans	B	2100	TIGR, COMB	incomplete
Cenarchaeum symbiosum	A	2500	MBARI	incomplete
Cenarchaeum symbiosum	A	2500	Diversa	incomplete
Chlamydia pneumoniae	B	1230	GENSET	complete
Chlamydia pneumoniae TW183	B	unknown	Gene Alliance, GPC-AG	complete
Chlamydia trachomatis L2	B	1038	GENSET	complete
Chlamydophila abortus	B	1144	Sanger Institute, Scottish Crop Res Inst, Moredun Res Inst	complete
Chlamydophila caviae GPIC	B	1200	TIGR	incomplete
Chlamydophila psittaci	B	unknown	TIGR	incomplete
Chloroflexus aurantiacus J-10-fl	B	3000	Joint Genome Institute	incomplete
Chromobacterium violaceum CCT 3496/ JMC 3496	B	4600	Brazilian Genome	incomplete
Clavibacter michiganensis subsp. ***sepedonicus*** ATCC 33113	B	2500	Sanger Institute, Colorado State Univ, Ohio State Univ	incomplete
Clostridium botulinum Hall strain A	B	4000	Sanger Institute, Univ of Reading, Institute of Food Research	incomplete
Clostridium difficile 630 (epidemic type X)	B	4400	Sanger Institute, LSHTM, Imperial College, et al.	incomplete
Clostridium perfringens ATCC 13124	B	unknown	TIGR	incomplete
Clostridium sp. BC1 ATCC 53464	B	3815	Brookhaven Natl Lab	incomplete
Clostridium tetani Massachusetts	B	4100	Gottingen Genomics Laboratory	incomplete
Clostridium thermocellum ATCC 27405	B	4500	Joint Genome Institute	incomplete
Colwellia sp. 34H	B	5300	TIGR	incomplete
Corynebacterium diphtheriae NCTC13129	B	2488	Sanger Institute, PHLS, Degussa, Bielefeld Univ	complete
Corynebacterium glutamicum ATCC 13032	B	3100	LION Bioscience AG, Degussa, IIT GmbH	incomplete
Corynebacterium glutamicum	B	unknown	Integrated Genomics Inc	complete
Corynebacterium thermoaminogenes FERM9246	B	unknown	NITE, Ajinomoto Co., Inc	incomplete
Coxiella burnetii Nine Mile (RSA 493)	B	2100	TIGR	incomplete
Crocosphaera watsonii WH8501	B	4000	Joint Genome Institute, Woods Hole Oceanographic Institute	incomplete
Cytophaga hutchinsonii ATCC 33406	B	4000	Joint Genome Institute	incomplete
Dechloromonas sp. RCB	B	2000	Joint Genome Institute	incomplete
Dehalococcoides ethenogenes	B	1500	TIGR	incomplete
Desulfitobacterium hafniense DCB-2	B	4600	Joint Genome Institute	incomplete
Desulfobacterium autotrophicum HRM2	B	unknown	Gottingen Genomics Laboratory, REGX	incomplete
Desulfotalea psychrophila LSv54	B	3660	Epidauros Biotechnologie AG, REGX	complete
Desulfovibrio desulfuricans G20	B	3900	Joint Genome Institute	incomplete

Desulfovibrio vulgaris Hildenborough	B	3200	TIGR	incomplete
Desulfuromonas acetoxidans	B	4100	Joint Genome Institute	incomplete
Dichelobacter nodosus VCS1703A	B	1600	Univ or Arizona, TIGR	incomplete
Ehrlichia canis jake	B	1000	Joint Genome Institute	incomplete
Ehrlichia chaffeensis Arkansas	B	1200	Ohio State Univ, TIGR	incomplete
Ehrlichia chaffeensis sapulpa	B	1000	Joint Genome Institute	incomplete
Ehrlichia sennetsu Miyayama	B	900	Ohio State Univ, TIGR	incomplete
Ehrlichia (Cowdria) ruminantium	B	1576	Sanger Institute, Utrecht Univ, ARC-OVI	incomplete
Enterobacter cloacae	B	unknown	Genome Therapeutics	complete
Enterococcus faecalis	B	unknown	Genome Therapeutics	complete
Enterococcus faecalis V583	B	3209	TIGR	complete
Enterococcus faecium	B	unknown	Genome Therapeutics	complete
Enterococcus faecium ATCC 35667	B	2092	Integrated Genomics Inc	complete
Enterococcus faecium DO	B	2980	Joint Genome Institute, Baylor College of Medicine	complete
Escherichia coli DH10B	B	unknown	Baylor College of Medicine	incomplete
Escherichia coli EAEC-042	B	unknown	Sanger Institute, John Radcliffe Hospital, Univ of Birmingham, et al.	incomplete
Escherichia coli EPEC-E2348/69	B	unknown	Sanger Institute, Univ of Oxford, Univ of Birmingham, et al.	incomplete
Escherichia coli ETEC-H10407	B	unknown	Univ of Wisconsin	incomplete
Escherichia coli K12 W3110	B	unknown	NAIST	complete
Escherichia coli O18ac:H7:K1 RS218	B	unknown	Univ of Wisconsin	incomplete
Escherichia coli UPEC-CFT073	B	5230	Univ of Wisconsin	complete
Escherichia coli non-K1 invasive clinical isolate	B	unknown	Sanger Institute, Univ of Oxford, Univ of Birmingham, et al.	incomplete
Exiguobacterium sp. 255-15	B	3000	Joint Genome Institute	incomplete
Ferroplasma acidarmanus	A	2000	Joint Genome Institute	incomplete
Fibrobacter succinogenes S85	B	3600	TIGR, NACRB	incomplete
Francisella tularensis schu 4	B	2000	Univ of Uppsala, WRAIR, MDS	complete
Fusobacterium nucleatum polymorphum ATCC 10953	B	2400	Baylor College of Medicine, UCLA	incomplete
Fusobacterium nucleatum vincentii ATCC 49256	B	unknown	Integrated Genomics Inc	incomplete
Gemmata obscuriglobus UQM 2246	B	9000	TIGR	incomplete
Gemmata sp. Wa1-1	B	unknown	Integrated Genomics Inc	incomplete
Geobacillus (Bacillus) kaustophilus HTA426	B	3500	JAMSTEC	incomplete
Geobacter metallireducens	B	6800	Joint Genome Institute	incomplete
Geobacter sulfurreducens	B	2500	TIGR, Univ of Massachusetts, Amherst & Exxon Corporation	incomplete
Gloeobacter violaceus PCC 7421	B	4600	Kazusa DNA Research Institute	incomplete
Gluconacetobacter diazotrophicus	B	unknown	UFRJ, LNCC/MCT, AGROBIOLOGIA, UENF, UERJ	incomplete
Gluconacetobacter diazotrophicus PAI5 (ATCC 49037)	B	2700	Univ of Wisconsin	incomplete
Gluconobacter oxydans	B	unknown	Julich GmbH, Georg-August-Univ Gottingen, et al.	incomplete
Haemophilus ducreyi 35000HP	B	1800	The Institute for Systems Biology, CRI, Ohio State Univ	complete
Haemophilus influenzae NTHi 3224A	B	unknown	Ohio State Univ	incomplete
Haemophilus influenzae NTHi 86028	B	unknown	Ohio State Univ	incomplete
Haemophilus somnus 129PT	B	2500	Joint Genome Institute	incomplete
Haemophilus somnus 2336	B	2500	Ohio State Univ, Virginia Polytechnic Institute and State Univ	incomplete
Haloarcula marismortui ATCC 43049	A	2700	UMBI, Institute for Systems Biology	incomplete
Halobacterium salinarium ATCC 19700	A	4000	Max-Planck-Institute for Biochemistry	incomplete
Haloferax volcanii DS2 ATCC 29605	A	4200	Univ of Scranton, Integrated Genomics Inc	incomplete
Helicobacter hepaticus ATCC51449	B	1800	MWG-Biotech, Univ of Wuerzburg, MIT, GeneData	complete

Heliobacillus mobilis	B	4200	Integrated Genomics Inc	complete
Hyperthermus butylicus	A	1900	Epidauros Biotechnologie AG, Univ of Copenhagen	incomplete
Hyphomonas neptunium	B	2700	Univ of Georgia, TIGR	incomplete
Kineococcus radiotolerans	B	4400	Joint Genome Institute, Savannah River Site	incomplete
Klebsiella pneumoniae	B	unknown	Genome Therapeutics	complete
Klebsiella pneumoniae MGH78578	B	unknown	Washington Univ	incomplete
Lactobacillus acidophilus ATCC 700396	B	1900	California Polytechnic State Univ, Environm Biotech Inst	incomplete
Lactobacillus brevis ATCC367	B	2000	Joint Genome Institute	incomplete
Lactobacillus casei ATCC334	B	2500	Joint Genome Institute	incomplete
Lactobacillus delbrueckii subsp. ***bulgaricus***	B	2300	INRA, Genoscope	complete
Lactobacillus delbrueckii subsp. ***bulgaricus*** ATCCBAA-365	B	2300	Joint Genome Institute	incomplete
Lactobacillus gasseri ATCC 33323	B	1800	Joint Genome Institute	incomplete
Lactobacillus helveticus CNRZ32	B	2300	Univ of Wisconsin, Utah St Univ	incomplete
Lactobacillus johnsonii NCC533	B	unknown	Nestle	incomplete
Lactobacillus plantarum WCFS1	B	3300	Wageningen Centre for Food Sciences, Greenomics	complete
Lactobacillus rhamnosus HN001	B	unknown	New Zealand Dairy Board	incomplete
Lactobacillus sakei	B	unknown	INRA	incomplete
Lactococcus lactis MG1363	B	unknown	Univ of Groningen	incomplete
Lactococcus lactis subsp. ***cremoris*** SK11	B	2300	Joint Genome Institute	incomplete
Lawsonia intracellularis	B	4200	Univ of Minnesota	incomplete
Legionella pneumophila Philadelphia-1	B	4100	Columbia Univ	complete
Leifsonia xyli subsp. ***xyli***	B	3000	Univ of Campinas	incomplete
Leuconostoc mesenteroides ATCC 8293	B	2000	Joint Genome Institute	incomplete
Listeria ivanovii PAM55	B	3000	Institut Pasteur, Competence Center Pathogenomik Wuerzburg, et al.	incomplete
Listeria monocytogenes 4b	B	2900	TIGR	incomplete
Listeria welshimeri	B	3000	University of Giessen, Integrated Genomics-GmbH	incomplete
Magnetococcus MC-1	B	4500	Joint Genome Institute	incomplete
Magnetospirillum magnetotacticum MS-1, ATCC 31632	B	4500	Joint Genome Institute	incomplete
Mannheimia (Pasteurella) haemolytica 2.4	B	2400	LION Bioscience AG, Intervet GmbH	complete
Mannheimia (Pasteurella) haemolytica PHL213 (ST1)	B	2700	Baylor College of Medicine	incomplete
Mesorhizobium sp. BNC1	B	5000	Joint Genome Institute	incomplete
Methanococcoides burtonii DSM6242	A	3000	Joint Genome Institute	incomplete
Methanococcus maripaludis LL	A	1660	Univ of Washington- Seattle	incomplete
Methanococcus thermolithotrophicus	A	unknown	Molecular Dynamics, Integrated Genomics Inc	incomplete
Methanococcus voltae	A	unknown	Molecular Dynamics, Integrated Genomics Inc	incomplete
Methanogenium frigidum	A	unknown	Univ of New S. Wales, Australian Genome Research Facility, et al.	incomplete
Methanopyrus kandleri	A	unknown	Molecular Dynamics, Univ of Illinois at Urbana-Champaign, et al.	incomplete
Methanosarcina acetivorans	A	unknown	Gottingen Genomics Laboratory	incomplete
Methanosarcina barkeri	A	unknown	Gottingen Genomics Laboratory	incomplete
Methanosarcina barkeri Fusaro	A	2580	Joint Genome Institute	incomplete
Methanosarcina thermophila	A	unknown	Gottingen Genomics Laboratory	incomplete
Methylobacillus flagellatus KT	B	2884	Integrated Genomics Inc	complete
Methylobacillus flagellatus KT	B	3100	Joint Genome Institute, Univ of Washington- Seattle	incomplete
Methylobacterium extorquens AM1	B	6000	Univ of Washington- Seattle, Integrated Genomics Inc	incomplete
Methylococcus capsulatus Bath	B	4600	TIGR, Univ of Bergen, Norway	incomplete
Methylomonas 16a	B	4000	DuPont	incomplete

Methylophaga thalassica S1	B	unknown	Integrated Genomics Inc	incomplete
Microbulbifer degradans 2-40	B	6000	Joint Genome Institute	incomplete
Microcystis aeruginosa PCC 7806	B	4800	Institut Pasteur	incomplete
Moorella (Clostridium) thermoacetica ATCC39073	B	unknown	Joint Genome Institute, Univ of Nebraska	incomplete
Moraxella catarrhalis	B	unknown	Genome Therapeutics	complete
Mycobacterium avium 104	B	4700	TIGR	incomplete
Mycobacterium avium subsp.***paratuberculosis*** K-10	B	4200	Univ of Minnesota	complete
Mycobacterium bovis AF2122/97(spoligotype 9)	B	4400	Sanger Institute, Institut Pasteur	complete
Mycobacterium bovis BCG, Pasteur 1173P2	B	4400	Institut Pasteur	incomplete
Mycobacterium marinum M	B	6000	Sanger Institute, University of Washington, Institut Pasteur, et al.	incomplete
Mycobacterium smegmatis MC2155	B	7000	TIGR	incomplete
Mycobacterium tuberculosis CSU#93	B	4447	TIGR	complete
Mycobacterium ulcerans	B	4400	Institut Pasteur	incomplete
Mycoplasma capricolum	B	1200	George Mason Univ	incomplete
Mycoplasma fermentans	B	1100	Yang-Ming Univ	incomplete
Mycoplasma hyopneumoniae 232	B	890	Iowa State Univ, Univ of Washington	complete
Mycoplasma mycoides subsp.***mycoides*** SC PG1	B	1280	Royal Institute of Technology, Stockholm, et al.	incomplete
Mycoplasma orale	B	675	Yang-Ming Univ	incomplete
Mycoplasma synoviae	B	unknown	Brazilian Genome	incomplete
Myxococcus xanthus	B	unknown	TIGR	incomplete
Myxococcus xanthus.	B	9450	Cereon Genomics, Stanford Univ	incomplete
Nanoarchaeum equitans	A	500	Diversa, Celera	complete
Neisseria gonorrhoeae FA 1090, ATCC 700825	B	2153	Univ of Oklahoma, Ohio State Univ	complete
Neisseria meningitidis	B	unknown	Genome Therapeutics	complete
Neisseria meningitidis serogroup C, 8013	B	2100	Institut Pasteur	incomplete
Neisseria meningitidis serogroup C, FAM18	B	2190	Sanger Institute, Max-Planck-Berlin	complete
Neorickettsia (Ehrlichia) sennetsu	B	900	Ohio State Univ, TIGR	incomplete
Nitrosomonas europaea ATCC 25978	B	2980	Joint Genome Institute	complete
Nostoc punctiforme ATCC 29133	B	9760	Joint Genome Institute	incomplete
Ochrobactrum anthropi	B	4800	Clemson Univ	incomplete
Oenococcus (Leuconostoc) oeni	B	unknown	Genome Express, INRA	incomplete
Oenococcus (Leuconostoc) oeni PSU-1	B	1800	Joint Genome Institute	incomplete
Parachlamydia sp. UWE25	B	1600	Technische Univ - Munchen	incomplete
Pectobacterium (Erwinia) carotovora subsp. ***atroseptica***	B	4000	Sanger Institute, Scottish Crop Res Inst, Univ of Cambridge	incomplete
Pectobacterium (Erwinia) chrysanthemi 3937	B	3700	Univ of Wisconsin, TIGR	incomplete
Pediococcus pentosaceus ATCC25745	B	2000	Joint Genome Institute	incomplete
Persephonella marina	B	unknown	Portland State Univ, TIGR	incomplete
Persephonella marina AZ-Fu1	B	unknown	Portland State Univ, TIGR	incomplete
Petrotoga miotherma ATCC 51224	B	2177	Integrated Genomics Inc	incomplete
Photobacterium profundum	B	unknown	Padova Univ	incomplete
Photorhabdus luminescens TT01	B	5680	Institut Pasteur	complete
Picrophilus torridus	A	unknown	TU Hamburg-Harburg, Georg-August-Univ Gottingen	complete
Pirellula sp.1	B	7150	REGX	complete
Polaribacter filamentus	B	4184	Integrated Genomics Inc	complete
Porphyromonas gingivalis W83	B	2200	TIGR, Forsyth Dental Center	incomplete
Prevotella intermedia 17	B	2800	TIGR	incomplete

Prevotella ruminicola	B	unknown	Ohio State Univ, TIGR	incomplete
Prochlorococcus marinus MIT9313	B	2400	Joint Genome Institute	complete
Prochlorococcus marinus NATL2A	B	2000	Joint Genome Institute, MIT	incomplete
Prochlorococcus marinus SS120	B	1800	Genoscope	incomplete
Prochlorococcus marinus subsp. ***pastoris*** CCMP1378 (MED4)	B	1670	Joint Genome Institute	complete
Prosthecobacter dejongeii ATCC 27091	B	4554	Integrated Genomics Inc	complete
Proteus mirabilis	B	unknown	Genome Therapeutics	complete
Pseudomonas anaerooleophila HD-1	B	4500	Takara Bio Inc, Kyoto Univ	complete
Pseudomonas fluorescens Pf-5	B	6500	Oregon State Univ, TIGR	incomplete
Pseudomonas fluorescens Pf0-1	B	3500	Joint Genome Institute	incomplete
Pseudomonas fluorescens SBW25	B	6600	Sanger Institute, Univ of Oxford, Univ of Birmingham	incomplete
Pseudomonas putida KT2440	B	6100	German Consortium, TIGR	incomplete
Pseudomonas putida PRS1	B	6100	TIGR	incomplete
Pseudomonas syringae pv. ***syringae*** B728a	B	5600	Joint Genome Institute	incomplete
Pseudomonas syringae pv. ***tomato*** DC3000	B	6100	Cornell Univ, Univ of Nebraska, Univ of Missouri, TIGR, et al.	incomplete
Psychrobacter sp. 273-4	B	2500	Joint Genome Institute	incomplete
Pyrolobus fumarii	A	1850	Diversa, Celera	complete
Ralstonia eutropha	B	unknown	Humboldt Univ, Berlin, Georg-August-Univ Gottingen, et al.	incomplete
Ralstonia metallidurans (eutropha) CH34	B	5000	Joint Genome Institute, Brookhaven Natl Lab	incomplete
Rhizobium etli CFN42	B	unknown	Univ Nacional Autonoma de Mexico	incomplete
Rhizobium leguminosarum bv. ***viciae*** 3841	B	5800	Sanger Institute, Univ of York, Univ of East Anglia	incomplete
Rhodobacter capsulatus SB1003	B	3700	Univ of Chicago, Institute of Mol Genetics, et al.	complete
Rhodobacter sphaeroides 2.4.1.	B	3900	Joint Genome Institute, Univ of Texas - Houston, et al.	incomplete
Rhodococcus sp. RHA1	B	unknown	Genome British Columbia	incomplete
Rhodococcus sp. I24	B	5487	Integrated Genomics Inc	complete
Rhodopseudomonas palustris CGA009	B	5460	Joint Genome Institute, Institute of Molecular Genetics	incomplete
Rhodospirillum rubrum ATCC 11170	B	3400	Joint Genome Institute	incomplete
Rickettsia rickettsii	B	unknown	The Institute for Systems Biology	incomplete
Rickettsia siberica	B	unknown	Univ of Maryland School of Medicine, CDC, Agencourt	complete
Rickettsia typhi Wilmington	B	1400	Univ of Texas, Baylor College of Medicine	incomplete
Roseobacter denitrificans Shiba O Ch 114	B	unknown	Integrated Genomics Inc	incomplete
Rubrobacter xylanophilus	B	2600	Joint Genome Institute, Louisiana State Univ	incomplete
Ruminococcus albus 8	B	4000	TIGR, NACRB	incomplete
Ruminococcus flavefaciens FD-1	B	4440	Univ of Illinois at Urbana-Champaign	incomplete
Salmonella bongori	B	4400	Sanger Institute, Univ of Glasgow, Univ of Cambridge, et al.	incomplete
Salmonella enterica serovar Choleraesuis	B	unknown	Univ of Illinois at Urbana-Champaign	incomplete
Salmonella enterica serovar Dublin	B	unknown	Univ of Illinois at Urbana-Champaign	incomplete
Salmonella enterica serovar Pullorum	B	unknown	Univ of Illinois at Urbana-Champaign	incomplete
Salmonella enteritidis LK5	B	4500	Univ of Illinois at Urbana-Champaign	incomplete
Salmonella enteritidis PT4	B	unknown	Sanger Institute, Univ of Glasgow, Univ of Cambridge, et al.	incomplete
Salmonella gallinarum 287/91	B	5000	Sanger Institute, Univ of Glasgow, Univ of Cambridge, et al.	incomplete
Salmonella paratyphi ATCC 9150D	B	4600	Washington Univ	incomplete
Salmonella typhi Ty2	B	unknown	Univ of Wisconsin	incomplete
Salmonella typhimurium DT104	B	5000	Sanger Institute, Univ of Glasgow, Univ of Cambridge, et al.	incomplete
Salmonella typhimurium SL1344	B	5000	Sanger Institute, Univ of Glasgow, Univ of Cambridge, et al.	incomplete
Salmonella typhimurium TR7095	B	4500	TIGR, Washington Univ	incomplete

Shewanella violacea DSS12	B	unknown	Kinki Univ, JAMSTEC	incomplete
Shigella dysenteriae M131	B	unknown	Sanger Institute, Univ of Oxford, Univ of Birmingham, et al.	incomplete
Shigella flexneri serotype 2a 2457T	B	unknown	Univ of Wisconsin	complete
Shigella sonnei 53G	B	unknown	Sanger Institute, Univ of Oxford, Univ of Birmingham, et al.	incomplete
Silicibacter pomeroyi DSS-3	B	4400	Univ of Georgia, TIGR	incomplete
Sphingomonas aromaticivorans DSM 12444	B	3800	Joint Genome Institute	incomplete
Sphingomonas aromaticivorans SMCC F199	B	3800	Joint Genome Institute	incomplete
Spiroplasma citri	B	unknown	Central Washington Univ	incomplete
Spiroplasma kunkelii CR2-3x	B	1600	Univ of Oklahoma	incomplete
Spirulina platensis	B	unknown	Human Genome Center, Beijing	incomplete
Staphylococcus aureus	B	unknown	Genome Therapeutics	complete
Staphylococcus aureus 930131	B	2564	Integrated Genomics Inc	complete
Staphylococcus aureus ATCC 29213	B	2621	Integrated Genomics Inc	complete
Staphylococcus aureus COL	B	2800	TIGR	incomplete
Staphylococcus aureus MRSA252	B	2902	Sanger Institute, Trinity College, Univ of Bath	complete
Staphylococcus aureus MSSA476	B	2804	Sanger Institute, Trinity College, Univ of Bath	complete
Staphylococcus aureus NCTC 8325	B	2800	Univ of Oklahoma, Ohio State Univ	complete
Staphylococcus aureus bovine	B	unknown	Univ of Minnesota	complete
Staphylococcus epidermidis	B	unknown	Genome Therapeutics	complete
Staphylococcus epidermidis ATCC 12228	B	2400	Chinese National Human Genome Center at Shanghai, et al.	incomplete
Staphylococcus epidermidis ATCC 14990	B	2377	Integrated Genomics Inc	complete
Staphylococcus epidermidis RP62A	B	2400	TIGR	incomplete
Staphylococcus haemolyticus	B	unknown	NITE, Juntendo Univ	incomplete
Stigmatella aurantiaca DW4/3-1	B	unknown	Integrated Genomics Inc	incomplete
Streptococcus agalactiae A909	B	2136	TIGR	incomplete
Streptococcus equi	B	2300	Sanger Institute, Univ of Newcastle, Univ of Cambridge	incomplete
Streptococcus gordonii Challis (NCTC7868)	B	4351	TIGR	incomplete
Streptococcus mitis NCTC 12261	B	2200	TIGR	incomplete
Streptococcus pneumoniae	B	unknown	Genome Therapeutics	complete
Streptococcus pneumoniae 23F (Spanish 23F-1)	B	2200	Sanger Institute, Univ of Glasgow, Univ of Leicester	incomplete
Streptococcus pneumoniae 670	B	unknown	TIGR	incomplete
Streptococcus pneumoniae serotype 6	B	unknown	Univ of Alabama	incomplete
Streptococcus pyogenes Manfredo (M5)	B	1840	Sanger Institute, Univ of Newcastle	complete
Streptococcus sanguinis SK36	B	unknown	Commonwealth Biotechnologies, Inc, Virginia Commonwealth Univ	incomplete
Streptococcus sobrinus 6715	B	2200	TIGR	incomplete
Streptococcus suis	B	1700	Sanger Institute, Univ of Cambridge, Univ of Newcastle, et al.	incomplete
Streptococcus suis 1591	B	2200	Joint Genome Institute	incomplete
Streptococcus thermophilus ATCC BAA-491	B	unknown	Joint Genome Institute	incomplete
Streptococcus thermophilus CNRZ 1066	B	1796	Integrated Genomics Inc, INRA	complete
Streptococcus thermophilus LMD-9	B	1800	Joint Genome Institute	incomplete
Streptococcus thermophilus LMG 18311	B	1800	Univ Catholique de Louvain, Belgium, INRA	complete
Streptococcus uberis	B	1700	Sanger Institute, Univ of Cambridge, Univ of Newcastle, et al.	incomplete
Streptomyces ambofaciens	B	8000	Genoscope	incomplete
Streptomyces avermitilis MA-4680	B	8700	Kitasato Univ, Univ of Tokyo, NITE	complete
Streptomyces diversa	B	unknown	Diversa, Celera	complete
Sulfolobus acidocaldarius DSM 639	A	1900	Epidauros Biotechnologie AG, Univ of Copenhagen	incomplete

Synechococcus elongatus PCC7942	B	2500	Joint Genome Institute, Texas A&M Univ	incomplete
Synechococcus sp. PCC 6301	B	2690	Nagoya Univ	incomplete
Synechococcus sp. PCC 7002	B	3200	Beijing Univ, Penn State Univ	incomplete
Synechococcus sp. PCC 7942	B	unknown	Texas A&M Univ	incomplete
Synechococcus sp. WH8102	B	2720	Joint Genome Institute	incomplete
Tannerella forsythensis ATCC 43037	B	3420	TIGR	incomplete
Thermobifida fusca YX	B	3600	Joint Genome Institute	incomplete
Thermochromatium tepidum MC ATCC 43061	B	3295	Integrated Genomics Inc	complete
Thermus flavus	B	unknown	Thermogene	incomplete
Thermus thermophilus HB27	B	1820	Gottingen Genomics Laboratory	incomplete
Treponema denticola 35405	B	2800	TIGR, Univ of Texas, Baylor College of Medicine	incomplete
Trichodesmium erythraeum IMS101	B	6500	Joint Genome Institute, Woods Hole Oceanographic Institution	incomplete
Tropheryma whippelii	B	925	Sanger Institute, Stanford Univ, Univ of Birmingham	complete
Tropheryma whipplei Twist	B	unknown	Genoscope	incomplete
Uncultivated Riftia endosymbiont	A	unknown	Molecular Dynamics, Scripps Institute of Oceanography, et al.	incomplete
Verrucomicrobium spinosum	B	unknown	TIGR	incomplete
Vibrio fischeri ES114	B	4136	Integrated Genomics Inc, Univ of Hawaii	complete
Vibrio vulnificus YJ016	B	5211	Yang-Ming Univ, Taiwan	complete
Wolbachia pipientis (Culex quinquefasciatus)	B	1500	Sanger Institute, Univ of Queensland, et al.	incomplete
Wolbachia sp. ***(Brugia malayi)***	B	956	New England Biolabs, Integrated Genomics Inc	incomplete
Wolbachia sp. ***(Dirofilaria immitis)***	B	unknown	Univ of Milano, Univ of Uppsala, Univ of Kopenhagen, et al.	incomplete
Wolbachia sp. ***(Drosophila and Brugia malayi)***	B	1400	TIGR, Yale Univ	incomplete
Wolbachia sp. ***(Onchocerca volvulus)***	B	1100	Sanger Institute, Univ of Edinburgh, et al.	incomplete
Wolbachia sp. wNo ***(D.simulans)***	B	1500	Univ of Uppsala, Univ of Kopenhagen, IMBB-FORTH	incomplete
Wolbachia sp. wUni ***(Muscidifurax uniraptor)***	B	1500	Univ of Uppsala, Univ of Kopenhagen, IMBB-FORTH	incomplete
Wolbachia sp. wVul ***(Armadillidium vulgare)***	B	1500	Univ of Uppsala, Univ of Kopenhagen, IMBB-FORTH	incomplete
Xanthomonas axonopodis pv. ***aurantifolii***	B	5000	FAPESP, Univ of Sao Paulo	incomplete
Xanthomonas campestris pv. ***campestris*** 8004	B	5000	Guangxi Univ, The Institute of Microbiology, et al.	complete
Xanthomonas citri	B	5000	FAPESP, Univ of Sao Paulo, UNICAMP	incomplete
Xylella fastidiosa Pierce's Disease Strain	B	2700	Univ of Campinas	incomplete
Xylella fastidiosa-almond dixon	B	2600	Joint Genome Institute	incomplete
Xylella fastidiosa-grape Temecula1	B	2600	AEG Brazilian Consortium	incomplete
Xylella fastidiosa-oleander ann1	B	2600	Joint Genome Institute	incomplete
Yersinia enterocolitica 8081	B	4620	Sanger Institute, St. Bartholomew's Hospital, Institut Pasteur	complete
Yersinia pseudotuberculosis IP32953	B	4300	LLNL	incomplete
Zymomonas mobilis	B	1833	Integrated Genomics Inc	complete
Zymomonas mobilis ZM4	B	2052	Macrogen	complete

* Species are listed in alphabetical order.

** B: Bacteria, A: Archaea

*** Here, "complete" means that the sequencing is finished, but the annotation is not yet.

Supplemental table 2 - (i) *Bacillus - Staphylococcus* group

Abbreviations:

Bha : ***Bacillus halodurans***
Bsu : ***Bacillus subtilis***
Lin : ***Listeria innocua***
Lmo : ***Listeria monocytogenes***
SauN : ***Staphylococcus aureus*** N315
SauM : ***Staphylococcus aureus*** Mu50
SauMW : ***Staphylococcus aureus*** MW2

Species list	Correct topology	Correct tree	%	Incorrect tree	%	Uncertain tree*	%	Significant tree**	Total tree	%
Bsu, SauN, SauW, SauM	**(Bsu, SauW) - (SauN, SauM)**	**690**	**70.3**	**40**	**4.1**	**251**	**25.6**	**981**	**1333**	**73.6**
Bha, SauN, SauW, SauM	**(Bha, SauW) - (SauN, SauM)**	**649**	**70.6**	**34**	**3.7**	**236**	**25.7**	**919**	**1234**	**74.5**
Lin, SauN, SauW, SauM	**(Lin, SauW) - (SauN, SauM)**	**631**	**71.5**	**31**	**3.6**	**221**	**25.0**	**883**	**1206**	**73.2**
Lmo, SauN, SauW, SauM	**(Lmo, SauW) - (SauN, SauM)**	**635**	**71.9**	**31**	**3.6**	**217**	**24.6**	**883**	**1215**	**72.7**
Bha, Bsu, Lmo, SauM	**(Bha, Bsu) - (Lmo, SauM)**	**290**	**85.5**	**49**	**14.4**	**0**	**0.0**	**339**	**819**	**41.4**
Bha, Bsu, Lmo, SauN	**(Bha, Bsu) - (Lmo, SauN)**	**294**	**85.7**	**49**	**14.3**	**0**	**0.0**	**343**	**838**	**40.9**
Bha, Bsu, Lin, SauM	**(Bha, Bsu) - (Lin, SauM)**	**289**	**85.8**	**48**	**14.3**	**0**	**0.0**	**337**	**819**	**41.1**
Bha, Bsu, Lmo, SauW	**(Bha, Bsu) - (Lmo, SauW)**	**295**	**85.8**	**49**	**14.3**	**0**	**0.0**	**344**	**838**	**41.1**
Bha, Bsu, Lin, SauN	**(Bha, Bsu) - (Lin, SauN)**	**292**	**85.9**	**48**	**14.1**	**0**	**0.0**	**340**	**837**	**40.6**
Bha, Bsu, Lin, SauW	**(Bha, Bsu) - (Lin, SauW)**	**293**	**86.4**	**46**	**13.5**	**0**	**0.0**	**339**	**837**	**40.5**
Bsu, Lmo, SauW, SauM	(Bsu, Lmo) - (SauW, SauM)	969	99.8	2	0.2	0	0.0	971	972	99.9
Bha, Bsu, Lin, Lmo	(Bha, Bsu) - (Lin, Lmo)	1096	99.9	1	0.1	0	0.0	1097	1097	100.0
Bsu, Lin, SauN, SauW	(Bsu, Lin) - (SauN, SauW)	982	99.9	1	0.1	0	0.0	983	984	99.9
Bsu, Lin, SauN, SauM	(Bsu, Lin) - (SauN, SauM)	967	99.9	1	0.1	0	0.0	968	969	99.9
Bsu, Lin, SauW, SauM	(Bsu, Lin) - (SauW, SauM)	960	99.9	1	0.1	0	0.0	961	963	99.8
Bsu, Lmo, SauN, SauW	(Bsu, Lmo) - (SauN, SauW)	992	99.9	1	0.1	0	0.0	993	994	99.9
Bsu, Lmo, SauN, SauM	(Bsu, Lmo) - (SauN, SauM)	973	99.9	1	0.1	0	0.0	974	975	99.9
Bha, Bsu, SauN, SauW	(Bha, Bsu) - (SauN, SauW)	1041	100.0	0	0.0	0	0.0	1041	1041	100.0
Bha, Bsu, SauN, SauM	(Bha, Bsu) - (SauN, SauM)	1019	100.0	0	0.0	0	0.0	1019	1020	99.9
Bha, Bsu, SauW, SauM	(Bha, Bsu) - (SauW, SauM)	1014	100.0	0	0.0	0	0.0	1014	1015	99.9
Bha, Lin, Lmo, SauN	(Bha, SauN) - (Lin, Lmo)	932	100.0	0	0.0	0	0.0	932	932	100.0
Bha, Lin, Lmo, SauW	(Bha, SauW) - (Lin, Lmo)	933	100.0	0	0.0	0	0.0	933	933	100.0
Bha, Lin, Lmo, SauM	(Bha, SauM) - (Lin, Lmo)	911	100.0	0	0.0	0	0.0	911	911	100.0
Bha, Lin, SauN, SauW	(Bha, Lin) - (SauN, SauW)	940	100.0	0	0.0	0	0.0	940	940	100.0
Bha, Lin, SauN, SauM	(Bha, Lin) - (SauN, SauM)	920	100.0	0	0.0	0	0.0	920	922	99.8
Bha, Lin, SauW, SauM	(Bha, Lin) - (SauW, SauM)	918	100.0	0	0.0	0	0.0	918	920	99.8
Bha, Lmo, SauN, SauW	(Bha, Lmo) - (SauN, SauW)	936	100.0	0	0.0	0	0.0	936	936	100.0
Bha, Lmo, SauN, SauM	(Bha, Lmo) - (SauN, SauM)	917	100.0	0	0.0	0	0.0	917	919	99.8
Bha, Lmo, SauW, SauM	(Bha, Lmo) - (SauW, SauM)	915	100.0	0	0.0	0	0.0	915	917	99.8
Bsu, Lin, Lmo, SauN	(Bsu, SauN) - (Lin, Lmo)	982	100.0	0	0.0	0	0.0	982	983	99.9
Bsu, Lin, Lmo, SauW	(Bsu, SauW) - (Lin, Lmo)	982	100.0	0	0.0	0	0.0	982	983	99.9
Bsu, Lin, Lmo, SauM	(Bsu, SauM) - (Lin, Lmo)	964	100.0	0	0.0	0	0.0	964	965	99.9
Lin, Lmo, SauN, SauW	(Lin, Lmo) - (SauN, SauW)	1202	100.0	0	0.0	0	0.0	1202	1202	100.0
Lin, Lmo, SauN, SauM	(Lin, Lmo) - (SauN, SauM)	1185	100.0	0	0.0	0	0.0	1185	1185	100.0
Lin, Lmo, SauW, SauM	(Lin, Lmo) - (SauW, SauM)	1179	100.0	0	0.0	0	0.0	1179	1179	100.0

NOTE1: Data that proportions of correct trees are lower than 95% are shown in bold.
NOTE2: In fact, correct trees including three ***Staphylococcus*** species were ambiguous by a 16S rRNS tree.

* The number of trees in that the branch lengths are zero.
** Bootstrap probability >= 90%

Supplemental table 2 - (ii) *Streptococcus* group

Abbreviation:

Lla : ***Lactococcus lactis***
Spn : ***Streptococcus pneumoniae***
SpnR : ***Streptococcus pneumoniae*** R6
Spy : ***Streptococus pyogens*** SF370
SpyM : ***Streptococus pyogens*** MGAS315
Spy8 : ***Streptococus pyogens*** MGAS8232

Species list	Correct topology	Correct tree	%	Incorrect tree	%	Uncertain tree*	%	Significant tree**	Total tree	%
Spn, Spy, SpyM, Spy8	**(Spn, Spy8) - (Spy, SpyM)**	**161**	**34.8**	**243**	**52.5**	**59**	**12.7**	**463**	**1044**	**44.3**
Lla, Spy, SpyM, Spy8	**(Lla, Spy8) - (Spy, SpyM)**	**148**	**35.4**	**216**	**51.7**	**54**	**12.9**	**418**	**947**	**44.1**
SpnR, Spy, SpyM, Spy8	**(SpnR, Spy8) - (Spy, SpyM)**	**161**	**35.7**	**234**	**51.8**	**56**	**12.4**	**451**	**1051**	**42.9**
Lla, SpnR, SpyM, Spy8	(Lla, SpnR) - (SpyM, Spy8)	846	99.5	4	0.4	0	0	850	852	99.8
Lla, Spn, SpyM, Spy8	(Lla, Spn) - (SpyM, Spy8)	834	99.6	3	0.3	0	0	837	840	99.6
Lla, SpnR, Spy, SpyM	(Lla, SpnR) - (Spy, SpyM)	837	99.8	2	0.2	0	0	839	841	99.8
Lla, Spn, SpnR, Spy	(Lla, Spy) - (Spn, SpnR)	824	99.9	1	0.1	0	0	825	828	99.6
Lla, Spn, SpnR, SpyM	(Lla, SpyM) - (Spn, SpnR)	836	99.9	1	0.1	0	0	837	839	99.8
Lla, Spn, SpnR, Spy8	(Lla, Spy8) - (Spn, SpnR)	837	99.9	1	0.1	0	0	838	840	99.8
Lla, Spn, Spy, SpyM	(Lla, Spn) - (Spy, SpyM)	825	99.9	1	0.1	0	0	826	830	99.5
Lla, Spn, Spy, Spy8	(Lla, Spn) - (Spy, Spy8)	823	99.9	1	0.1	0	0	824	828	99.5
Lla, SpnR, Spy, Spy8	(Lla, SpnR) - (Spy, Spy8)	835	99.9	1	0.1	0	0	836	839	99.6
Spn, SpnR, Spy, SpyM	(Spn, SpnR) - (Spy, SpyM)	1030	100	0	0	0	0	1030	1030	100
Spn, SpnR, Spy, Spy8	(Spn, SpnR) - (Spy, Spy8)	1029	100	0	0	0	0	1029	1029	100
Spn, SpnR, SpyM, Spy8	(Spn, SpnR) - (SpyM, Spy8)	1043	100	0	0	0	0	1043	1043	100

NOTE1: Data that proportions of correct trees are lower than 95% are shown in bold.
NOTE2: In fact, correct trees including three ***Streptococcus*** species were ambiguous by a 16S rRNS tree.

* The number of trees in that the branch lengths are zero.
** Bootstrap probability >= 90%

Supplemental table 2 - (iii) Gram-positive highGC% group

Abbreviation:

Cgl : ***Corynebacterium glutamicum***
Mle : ***Mycobacterium leprae***
Mtu : ***Mycobacterium tuberculosis*** H37Rv
MtuC : ***Mycobacterium tuberculosis*** CDC1551
Sco : ***Streptomyces coelicolor***

Species list	Correct topology	Correct tree	%	Incorrect tree	%	Uncertain tree*	%	Significant tree**	Total tree	%
Cgl, Mle, Mtu, Sco	(Cgl, Sco) - (Mle, Mtu)	698	98.9	8	1.1	0	0	706	721	97.9
Cgl, Mle, MtuC, Sco	(Cgl, Sco) - (Mle, MtuC)	699	99	7	1	0	0	706	722	97.8
Cgl, Mle, Mtu, MtuC	(Cgl, Mle) - (Mtu, MtuC)	911	100	0	0	0	0	911	915	99.6
Cgl, Mtu, MtuC, Sco	(Cgl, Sco) - (Mtu, MtuC)	954	100	0	0	0	0	954	954	100
Mle, Mtu, MtuC, Sco	(Mle, Sco) - (Mtu, MtuC)	888	100	0	0	0	0	888	891	99.7

* The number of trees in that the branch lengths are zero.
** Bootstrap probability >= 90%

Supplemental table 2 - (iv) *Chlamydia* group

Abbreviations:

Cpn : ***Chlamydophila pneumoniae*** CWL029
CpnA : ***Chlamydophila pneumoniae*** AR
CpnJ : ***Chlamydophila pneumoniae*** J138
Ctra : ***Chlamydia trachomatis***
Cmur : ***Chlamydia muridarum***

Species list	Correct topology	Correct tree	%	Incorrect tree	%	Uncertain tree*	%	Significant tree**	Total tree	%
Cpn, CpnA, CpnJ, Cmur	**(Cpn, Cmur) - (CpnA, CpnJ)**	**40**	**7.0**	**55**	**9.7**	**473**	**83.3**	**568**	**699**	**81.3**
Cpn, CpnA, CpnJ, Ctra	**(Cpn, Ctra) - (CpnA, CpnJ)**	**50**	**7.7**	**63**	**9.7**	**540**	**82.7**	**653**	**798**	**81.8**
CpnA, CpnJ, Ctra, Cmur	(CpnA, CpnJ) - (Ctra, Cmur)	694	100.0	0	0.0	0	0.0	694	694	100.0
Cpn, CpnA, Ctra, Cmur	(Cpn, CpnA) - (Ctra, Cmur)	692	100.0	0	0.0	0	0.0	692	692	100.0
Cpn, CpnJ, Ctra, Cmur	(Cpn, CpnJ) - (Ctra, Cmur)	696	100.0	0	0.0	0	0.0	696	696	100.0

NOTE1: Data that proportions of correct trees are lower than 95% are shown in bold.
NOTE2: In fact, correct trees including three ***Chlamydophila pneumoniae*** were ambiguous by a 16S rRNS tree.

* The number of trees in that the branch lengths are zero.
** Bootstrap probability >= 90%

Supplemental table 2 - (v) Enterobacteria and its relatives group

Abbreviation:

Eco : ***Escherichia coli*** K12
EcoO : ***Escherichia coli*** O157 EDL933
EcoOR : Escherichia coliO157 RIMD0509952
Sty : ***Salmonella typhi***
Stym : ***Salmonella typhimurium***
Vch : ***Vibrio cholerae***
Ype : ***Yersinia pestis***

Species list	Correct topology	Correct tree	%	Incorrect tree	%	Uncertain tree*	%	Significant tree**	Total tree	%
EcoOR, Eco, EcoO, Vch	**(EcoOR, EcoO) - (Eco, Vch)**	**1282**	**89**	**68**	**4.7**	**90**	**6.2**	**1440**	**1755**	**82.1**
EcoOR, Eco, EcoO, Ype	**(EcoOR, EcoO) - (Eco, Ype)**	**1702**	**91.5**	**51**	**2.8**	**107**	**5.8**	**1860**	**2273**	**81.8**
EcoOR, Eco, EcoO, Sty	**(EcoOR, EcoO) - (Eco, Sty)**	**2240**	**93.2**	**26**	**1.1**	**138**	**5.7**	**2404**	**2931**	**82**
EcoOR, Eco, EcoO, Stym	**(EcoOR, EcoO) - (Eco, Stym)**	**2315**	**93.3**	**27**	**1.1**	**139**	**5.6**	**2481**	**3028**	**81.9**
EcoOR, Stym, Vch, Ype	(EcoOR, Stym) - (Vch, Ype)	1461	99.3	11	0.7	0	0	1472	1524	96.6
EcoO, Stym, Vch, Ype	(EcoO, Stym) - (Vch, Ype)	1463	99.3	11	0.7	0	0	1474	1527	96.5
EcoOR, Sty, Stym, Ype	(EcoOR, Ype) - (Sty, Stym)	2031	99.5	9	0.4	1	0	2041	2101	97.1
EcoO, Sty, Stym, Ype	(EcoO, Ype) - (Sty, Stym)	2039	99.5	10	0.4	1	0	2050	2108	97.2
EcoOR, Sty, Vch, Ype	(EcoOR, Sty) - (Vch, Ype)	1440	99.6	6	0.4	0	0	1446	1502	96.3
Eco, EcoO, Sty, Vch	(Eco, EcoO) - (Sty, Vch)	1597	99.6	5	0.3	1	0.1	1603	1643	97.6
Eco, EcoO, Stym, Vch	(Eco, EcoO) - (Stym, Vch)	1625	99.6	5	0.3	1	0.1	1631	1669	97.7
Eco, Sty, Stym, Ype	(Eco, Ype) - (Sty, Stym)	2026	99.6	8	0.4	1	0	2035	2094	97.2
EcoO, Sty, Vch, Ype	(EcoO, Sty) - (Vch, Ype)	1442	99.6	6	0.4	0	0	1448	1505	96.2
EcoOR, Eco, Sty, Stym	(EcoOR, Eco) - (Sty, Stym)	2877	99.7	6	0.2	2	0.1	2885	2888	99.9
EcoOR, Eco, Sty, Vch	(EcoOR, Eco) - (Sty, Vch)	1601	99.7	4	0.2	1	0.1	1606	1641	97.9
EcoOR, Eco, Stym, Vch	(EcoOR, Eco) - (Stym, Vch)	1630	99.7	4	0.2	1	0.1	1635	1667	98.1
Eco, EcoO, Sty, Stym	(Eco, EcoO) - (Sty, Stym)	2879	99.7	7	0.2	2	0.1	2888	2894	99.8
EcoOR, Eco, Sty, Ype	(EcoOR, Eco) - (Sty, Ype)	2046	99.8	4	0.1	1	0	2051	2092	98
EcoOR, Eco, Stym, Ype	(EcoOR, Eco) - (Stym, Ype)	2071	99.8	3	0.1	1	0	2075	2120	97.9
EcoOR, EcoO, Sty, Vch	(EcoOR, EcoO) - (Sty, Vch)	1647	99.8	2	0.2	1	0.1	1650	1670	98.8
EcoOR, EcoO, Stym, Vch	(EcoOR, EcoO) - (Stym, Vch)	1680	99.8	2	0.2	1	0.1	1683	1701	98.9
EcoOR, Sty, Stym, Vch	(EcoOR, Vch) - (Sty, Stym)	1627	99.8	2	0.2	1	0.1	1630	1658	98.3
Eco, EcoO, Sty, Ype	(Eco, EcoO) - (Sty, Ype)	2041	99.8	4	0.1	1	0	2046	2093	97.8
Eco, EcoO, Stym, Ype	(Eco, EcoO) - (Stym, Ype)	2068	99.8	4	0.2	1	0	2073	2122	97.7
Eco, Sty, Stym, Vch	(Eco, Vch) - (Sty, Stym)	1615	99.8	2	0.2	1	0.1	1618	1650	98.1
EcoO, Sty, Stym, Vch	(EcoO, Vch) - (Sty, Stym)	1630	99.8	2	0.2	1	0.1	1633	1661	98.3
EcoOR, EcoO, Sty, Stym	(EcoOR, EcoO) - (Sty, Stym)	2995	99.9	1	0	2	0.1	2998	2999	100
Eco, Sty, Vch, Ype	(Eco, Sty) - (Vch, Ype)	1445	99.9	1	0.1	0	0	1446	1502	96.3
Eco, Stym, Vch, Ype	(Eco, Stym) - (Vch, Ype)	1466	99.9	2	0.2	0	0	1468	1519	96.6
EcoOR, Eco, Vch, Ype	(EcoOR, Eco) - (Vch, Ype)	1532	100	0	0	0	0	1532	1536	99.7
EcoOR, EcoO, Sty, Ype	(EcoOR, EcoO) - (Sty, Ype)	2097	100	0	0	1	0	2098	2124	98.8
EcoOR, EcoO, Stym, Ype	(EcoOR, EcoO) - (Stym, Ype)	2125	100	0	0	1	0	2126	2157	98.6
EcoOR, EcoO, Vch, Ype	(EcoOR, EcoO) - (Vch, Ype)	1574	100	0	0	0	0	1574	1578	99.7
Eco, EcoO, Vch, Ype	(Eco, EcoO) - (Vch, Ype)	1531	100	0	0	0	0	1531	1536	99.7
Sty, Stym, Vch, Ype	(Sty, Stym) - (Vch, Ype)	1540	100	0	0	0	0	1540	1542	99.9

NOTE1: Data that proportions of correct trees are lower than 95% are shown in bold.

* The number of trees in that the branch lengths are zero.
** Bootstrap probability >= 90%

Supplemental table 2 - (vi) *Rhizobium* group

Abbreviation:

Atu : ***Agrobacterium tumefaciens*** C58 Cereon
AtuD : ***Agrobacterium tumefaciens*** C58 DuPont
Bme : ***Brucella melitensis***
Mlo : ***Mesorhizobium loti***
Sme : ***Sinorhizobium meliloti***

Species list	Correct topology	Correct tree	%	Incorrect tree	%	Uncertain tree*	%	Significant tree**	Total tree	%
AtuD, Bme, Mlo, Sme	(AtuD, Sme) - (Bme, Mlo)	1158	95	61	5	0	0	1219	1517	80.4
Atu, Bme, Mlo, Sme	(Atu, Sme) - (Bme, Mlo)	1165	95.1	60	4.9	0	0	1225	1519	80.6
Atu, AtuD, Bme, Mlo	(Atu, AtuD) - (Bme, Mlo)	1758	100	0	0	0	0	1758	1758	100
Atu, AtuD, Bme, Sme	(Atu, AtuD) - (Bme, Sme)	1613	100	0	0	0	0	1613	1613	100
Atu, AtuD, Mlo, Sme	(Atu, AtuD) - (Mlo, Sme)	1890	100	0	0	0	0	1890	1890	100

* The number of trees in that the branch lengths are zero.
** Bootstrap probability >= 90%

Supplemental figures

Genome maps of 84 species examined in this study

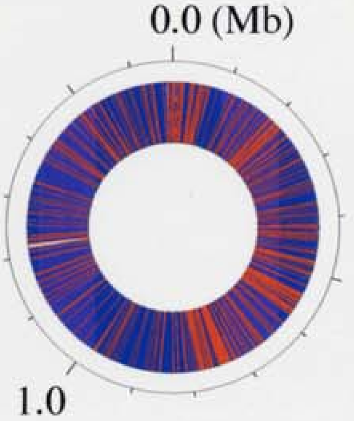
1. Archaea
2. Bacteria
(in alphabetical order)

Legends to figures :

Red bars show horizontally transferred (HT) genes.
Blue bars show non-HT genes.

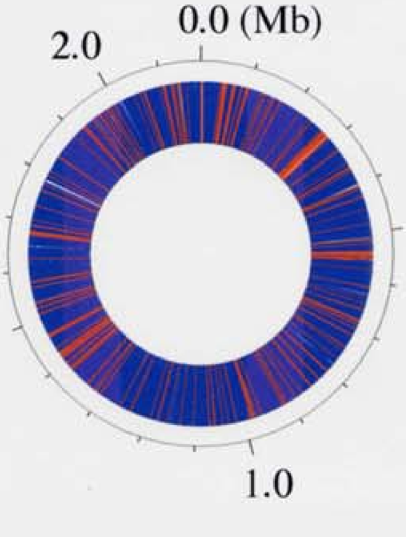
1. Archaea

Aeropyrum pernix (circular)



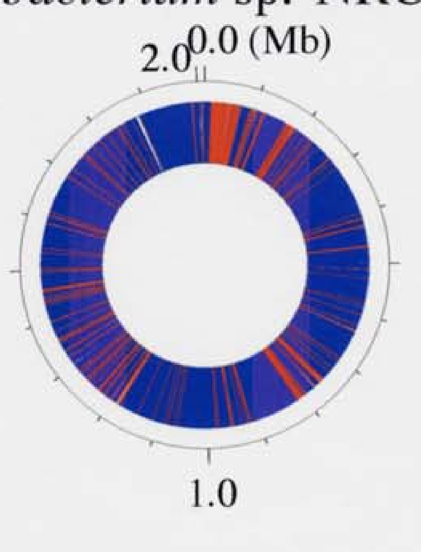
Genome Size = 1,669,695 bp

Archaeoglobus fulgidus (circular)



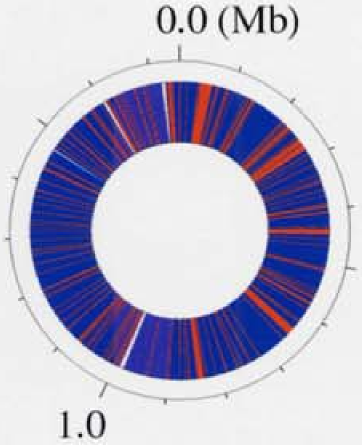
Genome Size = 2,178,400 bp

Halobacterium sp. NRC-1 (circular)



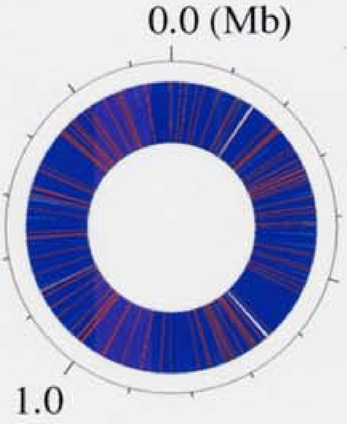
Genome Size = 2,014,239 bp

Methanobacterium thermoautotrophicum (circular)



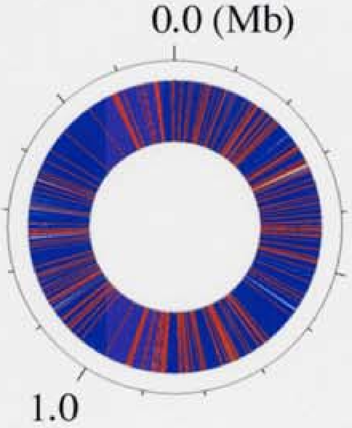
Genome Size = 1,751,377 bp

Methanococcus jannaschii (circular)



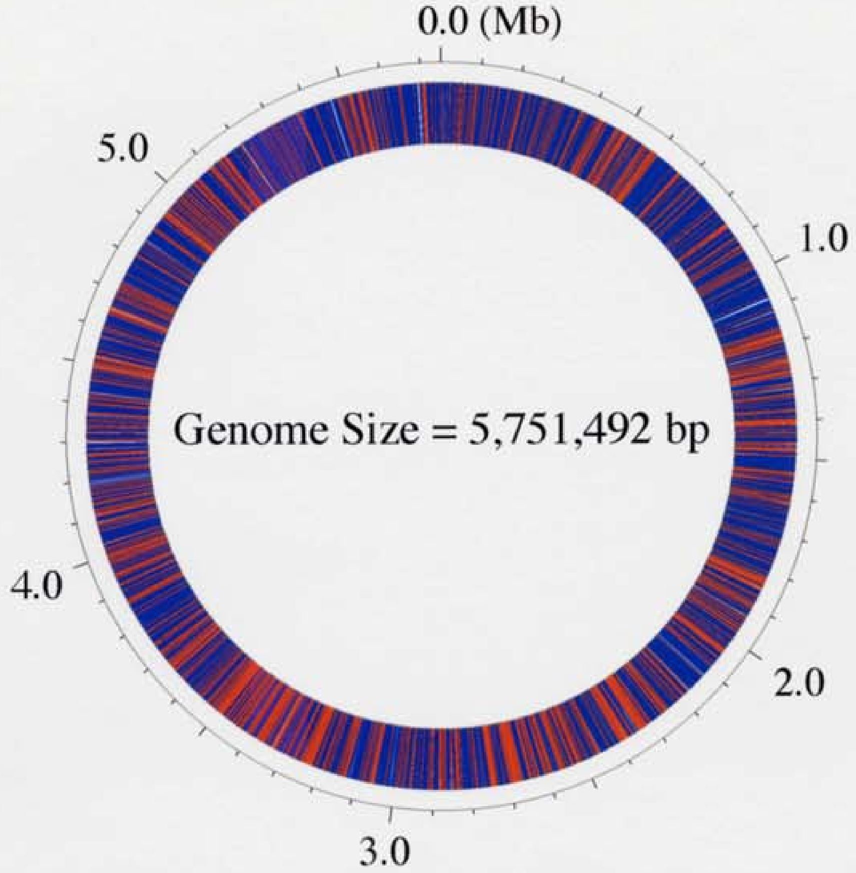
Genome Size = 1,664,970 bp

Methanopyrus kandleri AV19 (circular)

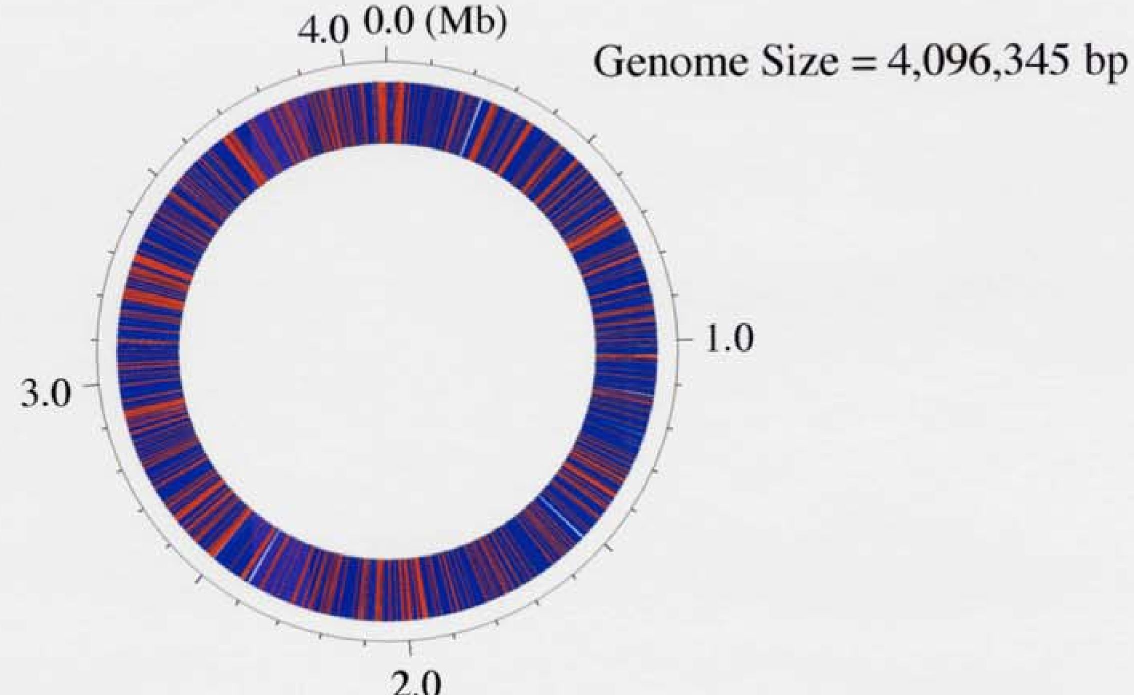


Genome Size = 1,694,969 bp

Methanosarcina acetivorans C2A (circular)

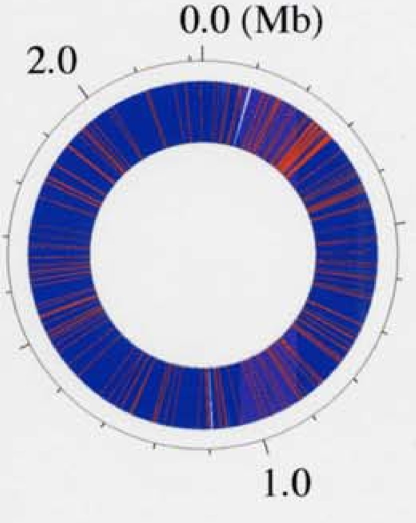


Methanosarcina mazei Goe1 (circular)



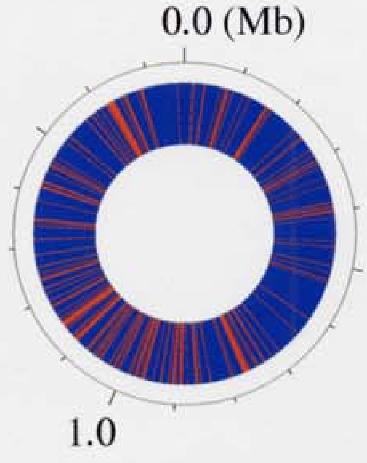
Pyrobaculum aerophilum
(circular)

Genome Size = 2,222,430 bp



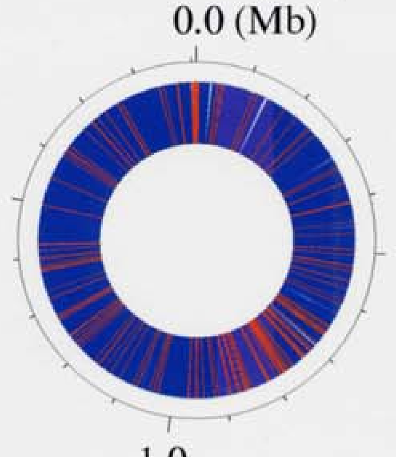
Pyrococcus abyssi
(circular)

Genome Size = 1,765,118 bp



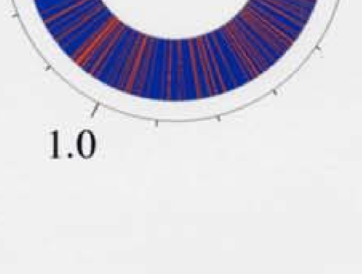
Pyrococcus furiosus DSM 3638
(circular)

Genome Size = 1,908,256 bp

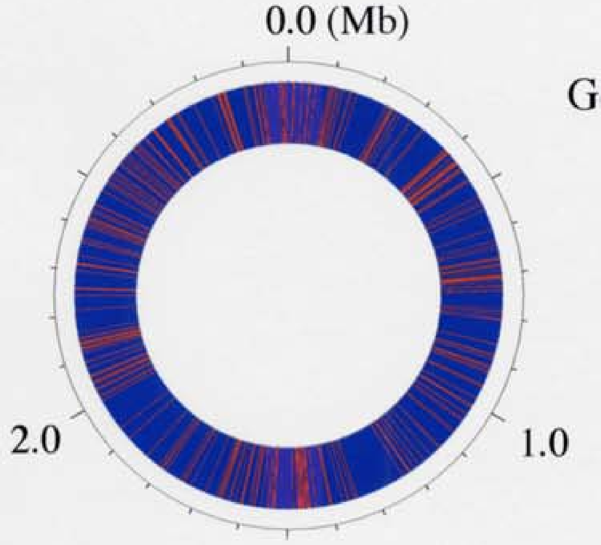


Pyrococcus horikoshii
(circular)

Genome Size = 1,738,505 bp

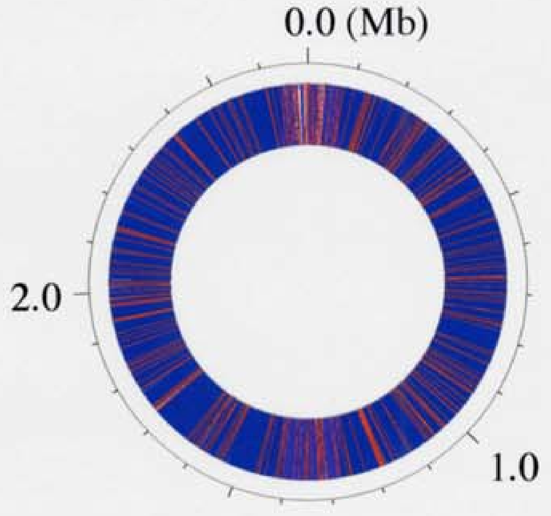


Sulfolobus solfataricus (circular)



Genome Size = 2,992,245 bp

Sulfolobus tokodaii (circular)

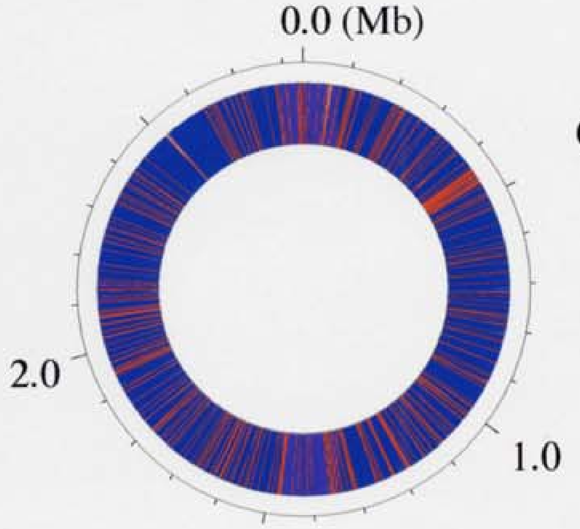


Genome Size = 2,694,765 bp

2. Bacteria

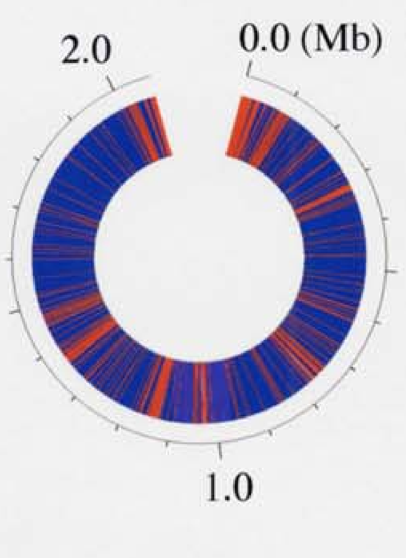
Agrobacterium tumefaciens C58

Chromosome 1 (circular)



Genome Size = 2,841,581 bp

Chromosome 2 (linear)

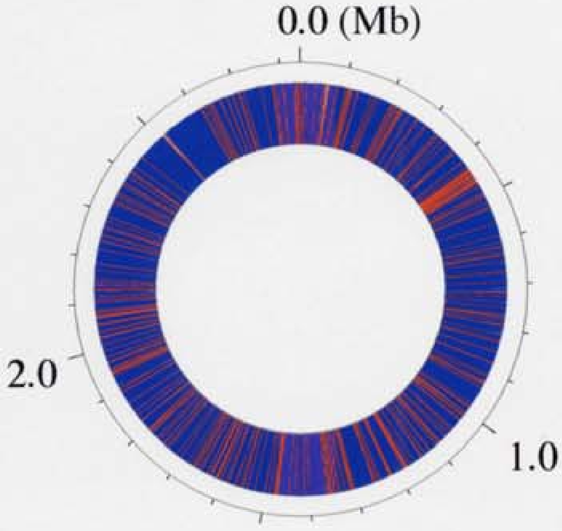


Genome Size = 2,074,782 bp

2. Bacteria

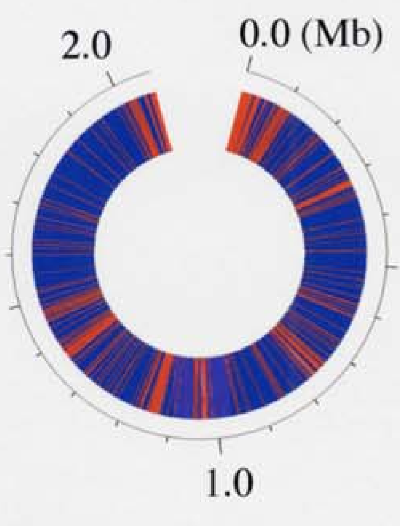
Agrobacterium tumefaciens C58

Chromosome 1 (circular)



Genome Size = 2,841,581 bp

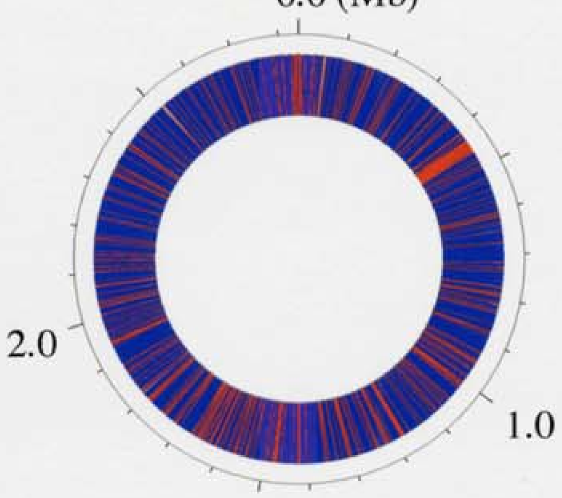
Chromosome 2 (linear)



Genome Size = 2,074,782 bp

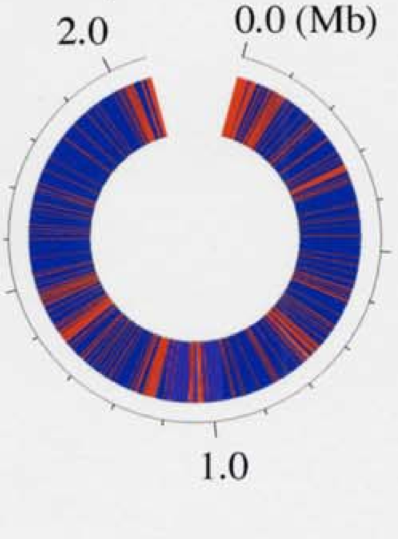
Agrobacterium tumefaciens C58 (Dupont)

Chromosome 1 (circular)



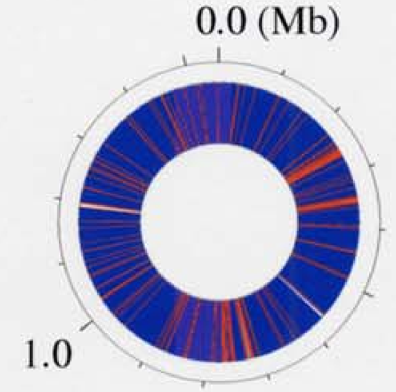
Genome Size = 2,841,490 bp

Chromosome 2 (linear)



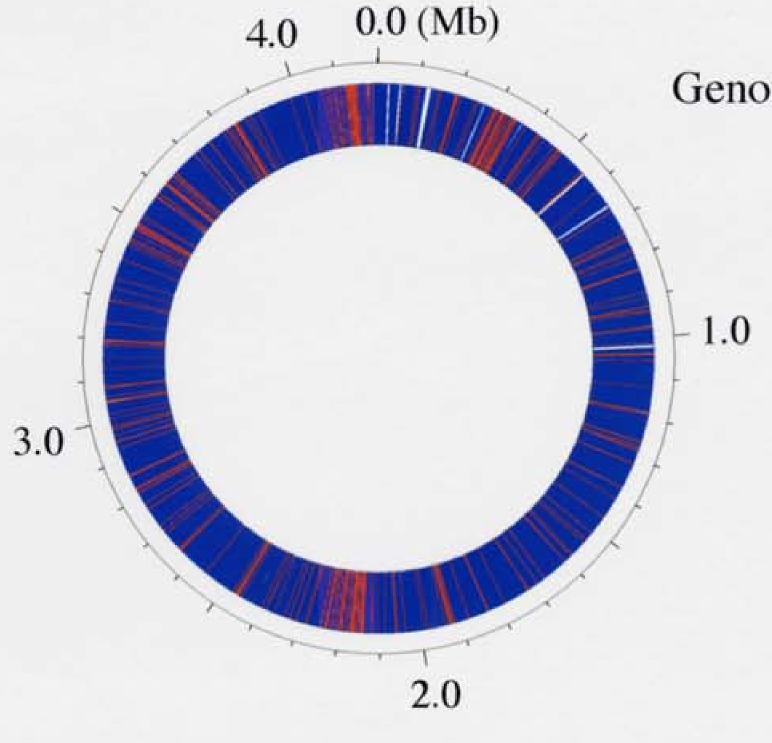
Genome Size = 2,074,782 bp

Aquifex aeolicus (circular)



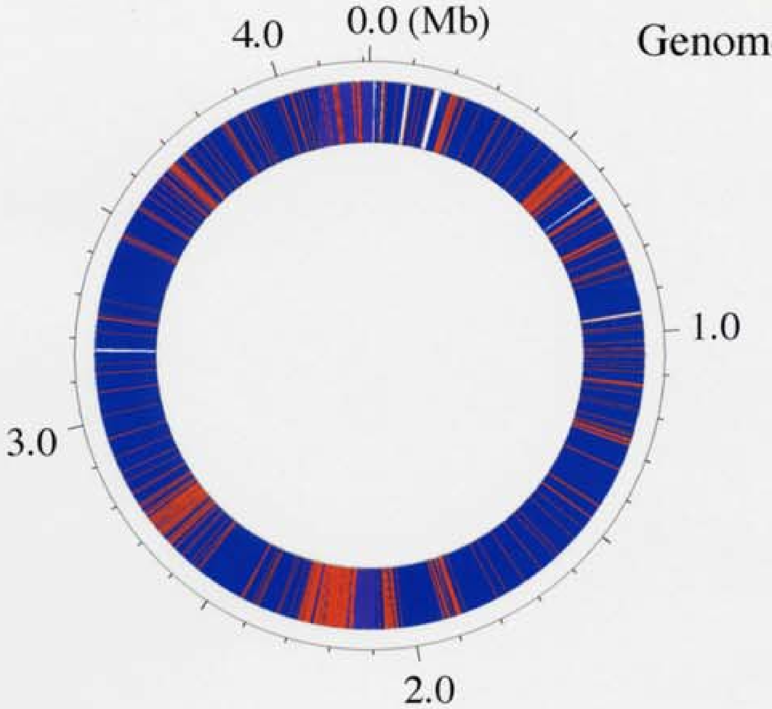
Genome Size = 1,551,335 bp

Bacillus halodurans (circular)



Genome Size = 4,202,353 bp

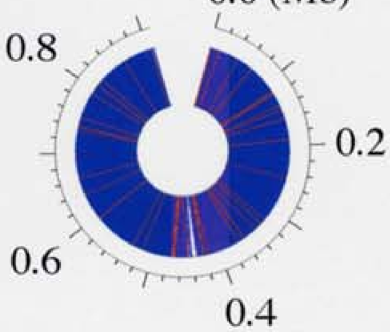
Bacillus subtilis (circular)



Genome Size = 4,214,814 bp

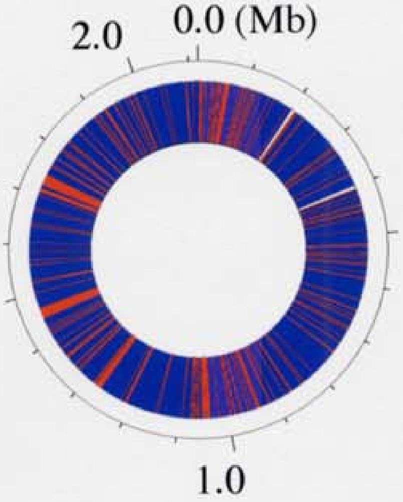
Borrelia burgdorferi (linear)

Genome Size = 910,724 bp



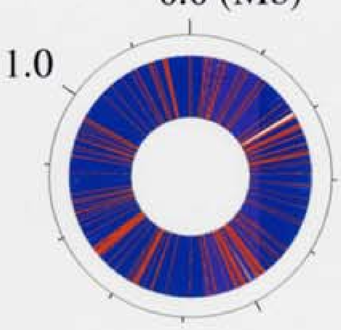
Brucella melitensis

Chromosome 1 (circular)



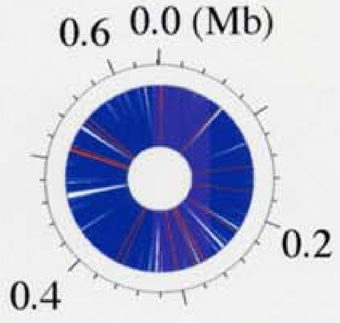
Genome Size = 2,117,144 bp

Chromosome 2 (circular)



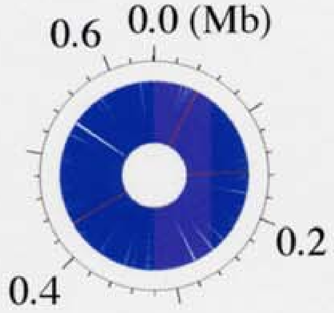
Genome Size = 1,177,787 bp

Buchnera aphidicola Sg (circular)



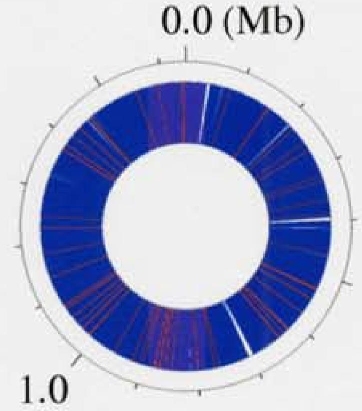
Genome Size = 641,454 bp

Buchnera sp. APS (circular)



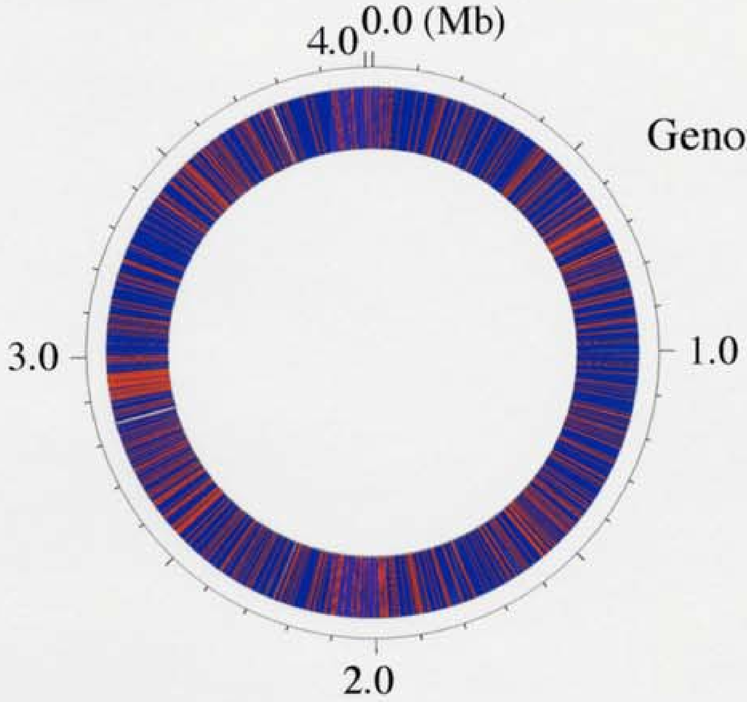
Genome Size = 640,681 bp

Campylobacter jejuni (circular)



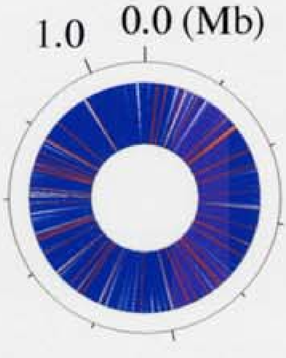
Genome Size = 1,641,481 bp

Caulobacter crescentus (circular)



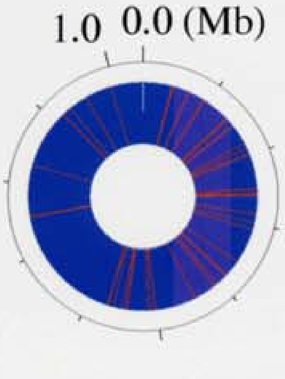
Genome Size = 4,016,947 bp

Chlamydia muridarum (circular)



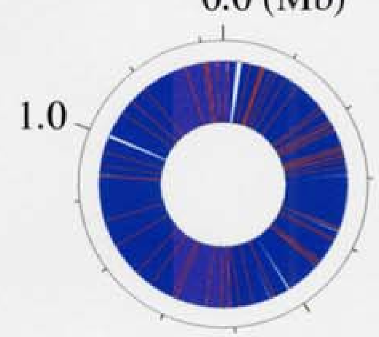
Genome Size = 1,069,412 bp

Chlamydia trachomatis (circular)



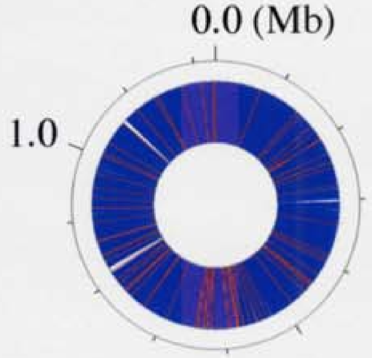
Genome Size = 1,042,519 bp

Chlamydophila pneumoniae CWL029 (circular)



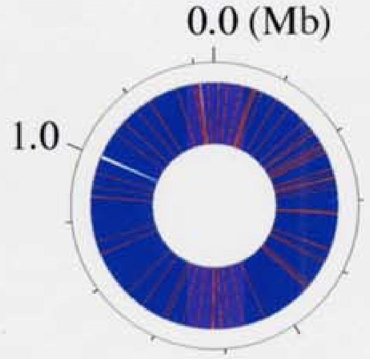
Genome Size = 1,230,230 bp

Chlamydophila pneumoniae AR39 (circular)



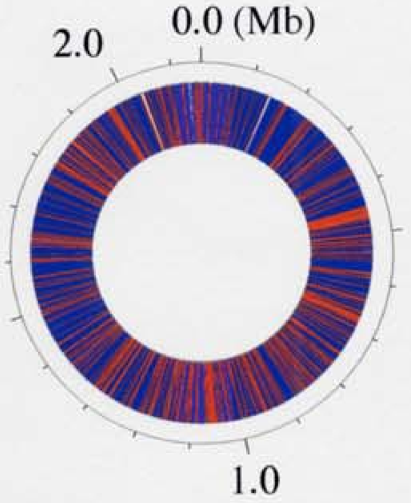
Genome Size = 1,229,853 bp

Chlamydophila pneumoniae J138 (circular)



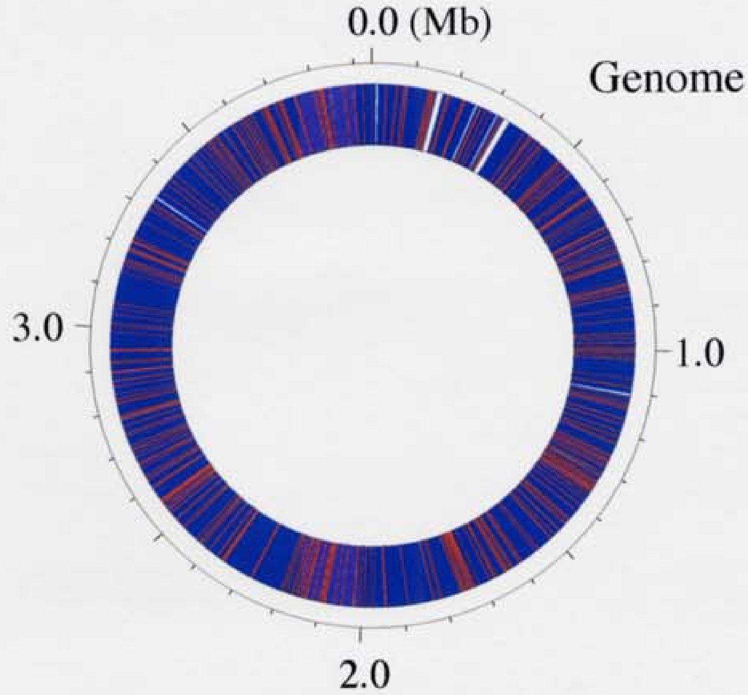
Genome Size = 1,228,267 bp

Chlorobium tepidum TLS (circular)



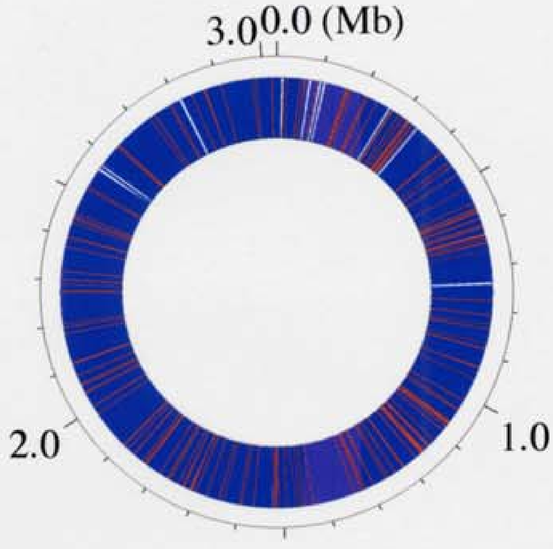
Genome Size = 2,154,946 bp

Clostridium acetobutylicum (circular)



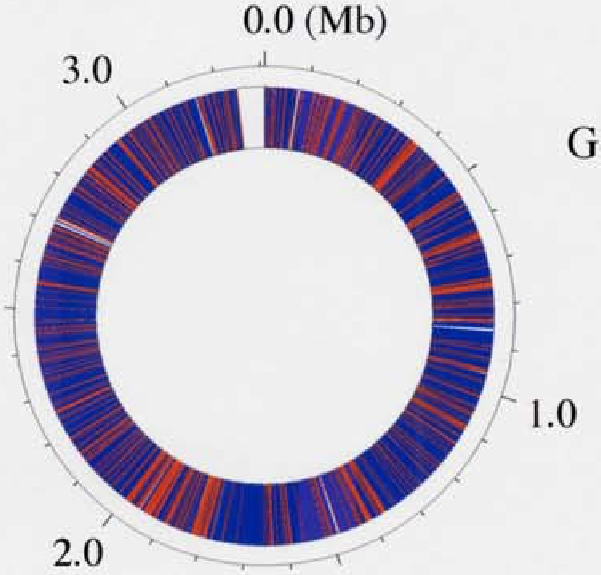
Genome Size = 3,940,880 bp

Clostridium perfringens (circular)



Genome Size = 3,031,430 bp

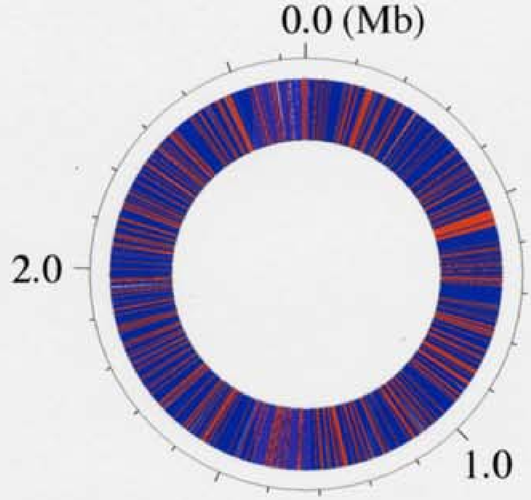
Corynebacterium glutamicum ATCC 13032 (circular)



Genome Size = 3,309,401 bp

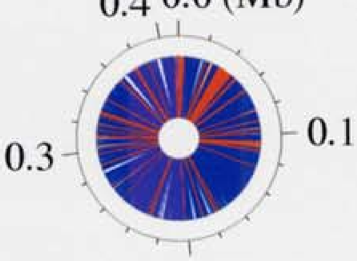
Deinococcus radiodurans

Chromosome 1 (circular)



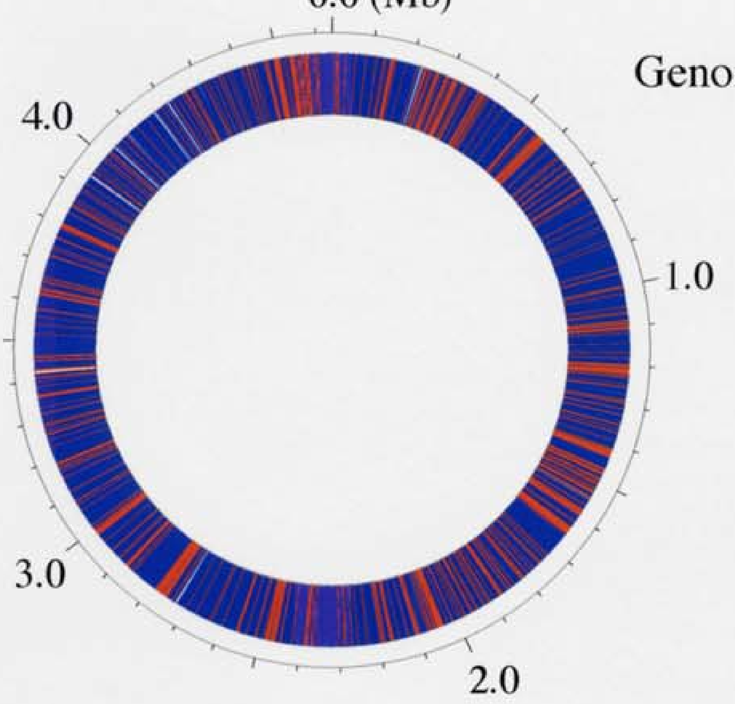
Genome Size = 2,648,638 bp

Chromosome 2 (circular)



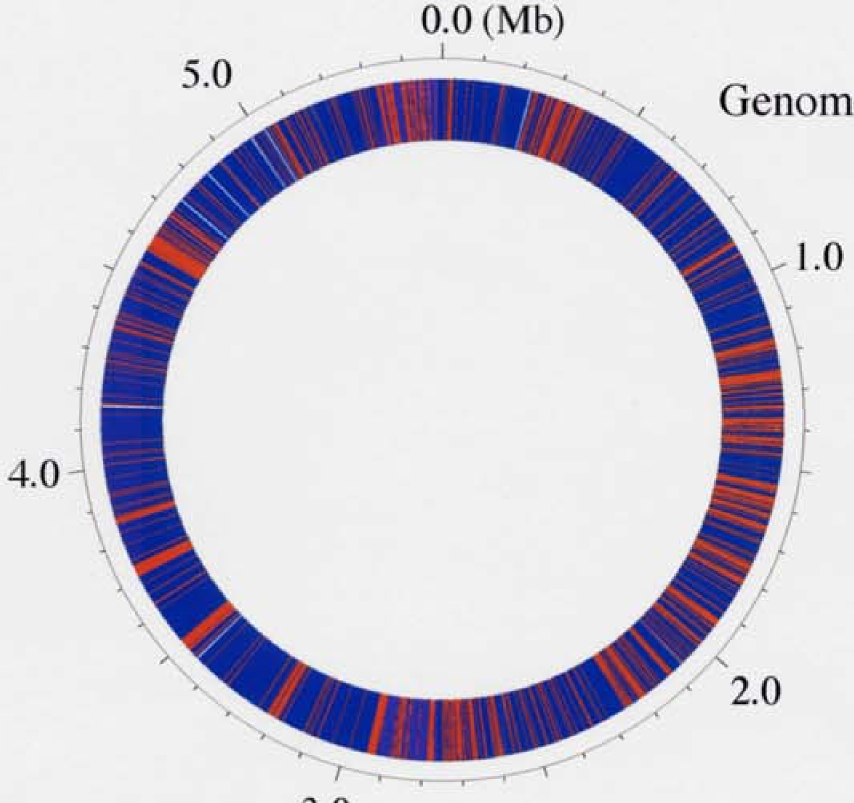
Genome Size = 412,348 bp

Escherichia coli K12 (circular)



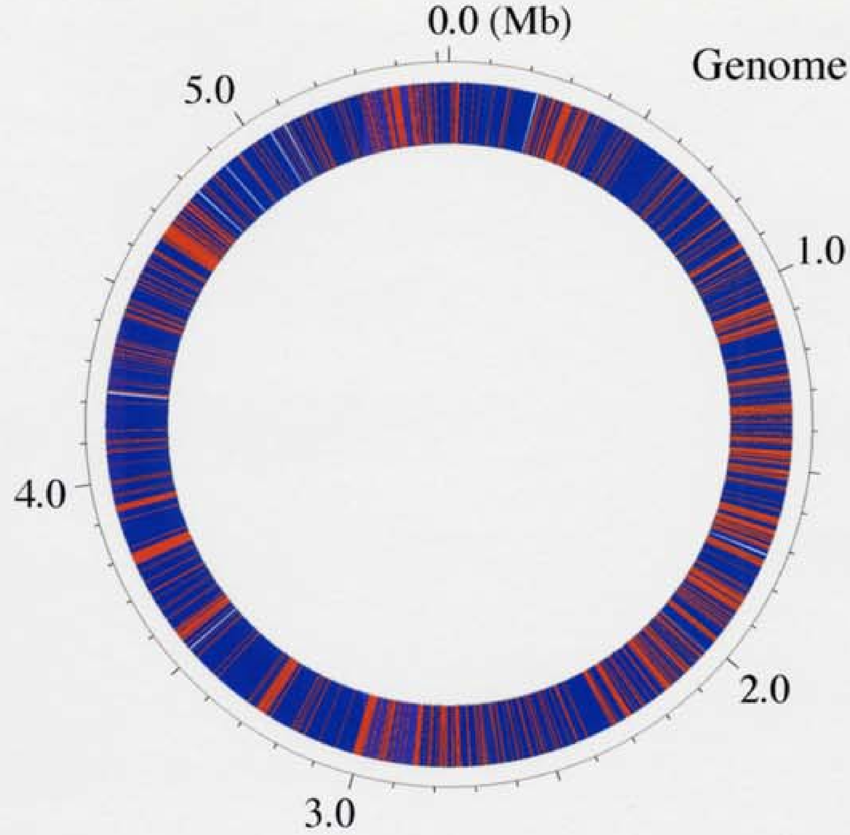
Genome Size = 4,639,221 bp

Escherichia coli O157:H7 RIMD 0509952 (circular)



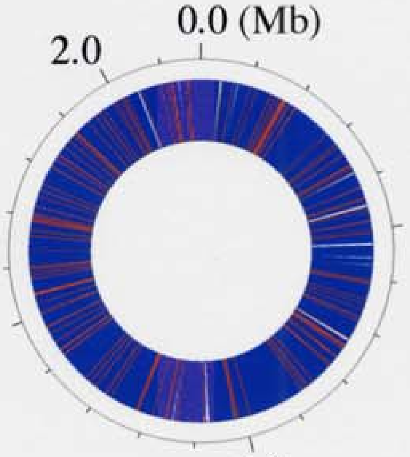
Genome Size = 5,498,450 bp

Escherichia coli O157:H7 EDL933 (circular)



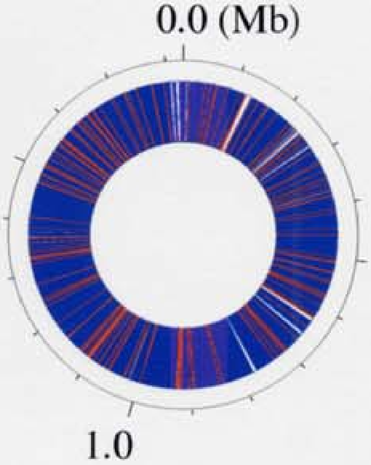
Genome Size = 5,528,445 bp

Fusobacterium nucleatum subsp. *nucleatum* ATCC 25586 (circular)



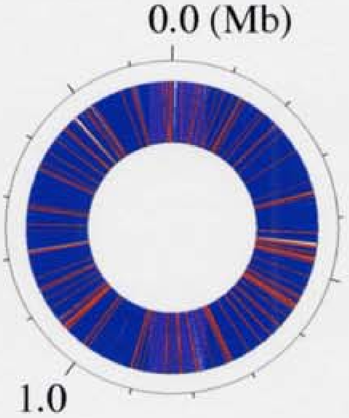
Genome Size = 2,174,500 bp

Haemophilus influenzae Rd (circular)



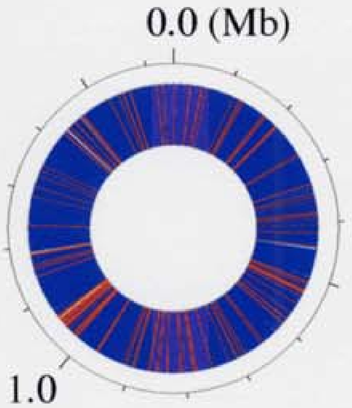
Genome Size = 1,830,138 bp

Helicobacter pylori 26695 (circular)



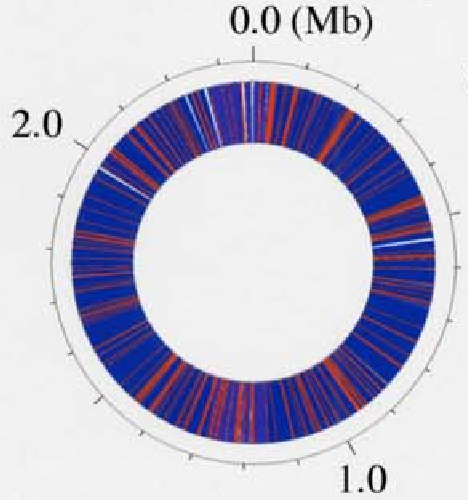
Genome Size = 1,667,867 bp

Helicobacter pylori J99 (circular)



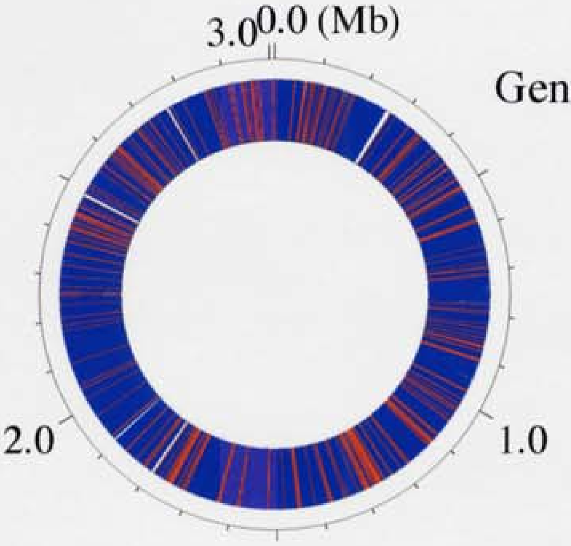
Genome Size = 1,643,831 bp

Lactococcus lactis subsp. *lactis* (circular)



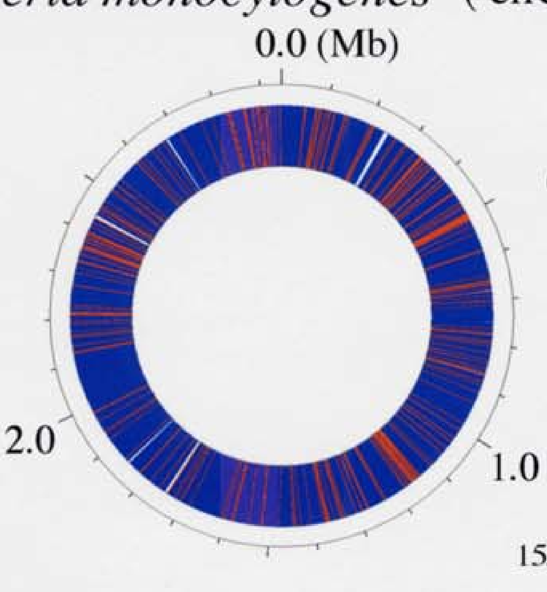
Genome Size = 2,365,589 bp

Listeria innocua (circular)



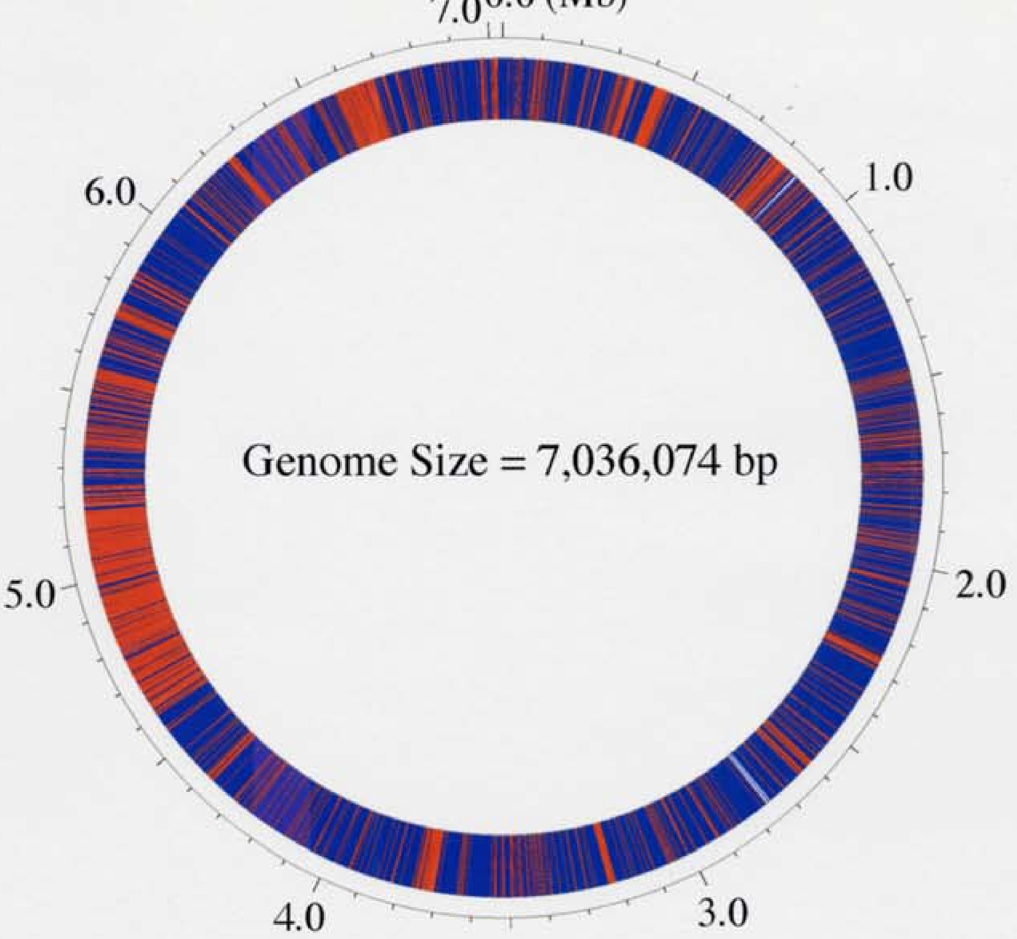
Genome Size = 3,011,208 bp

Listeria monocytogenes (circular)

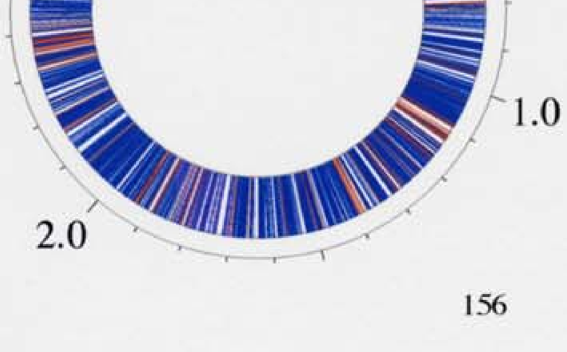


Genome Size = 2,944,528 bp

Mesorhizobium loti (circular)

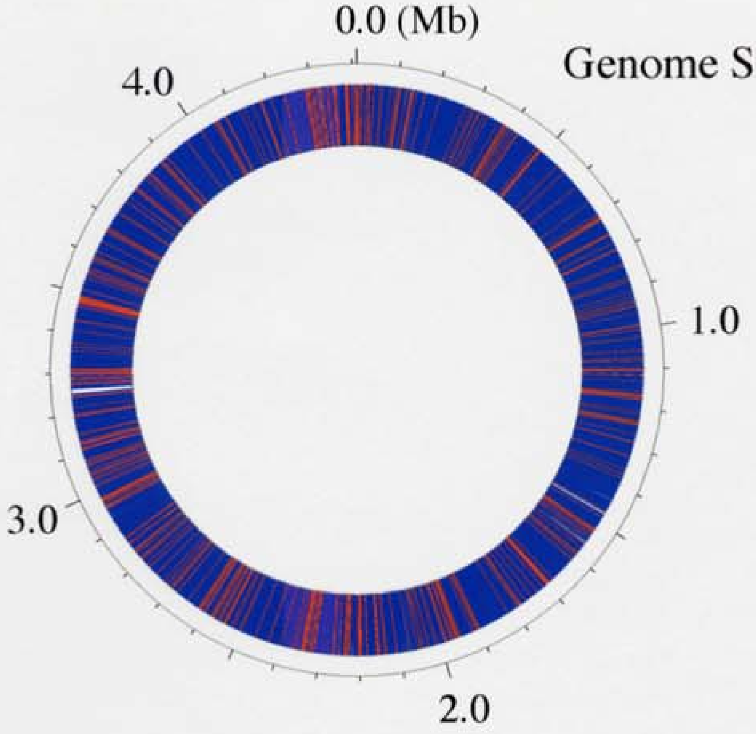


Mycobacterium leprae (circular)



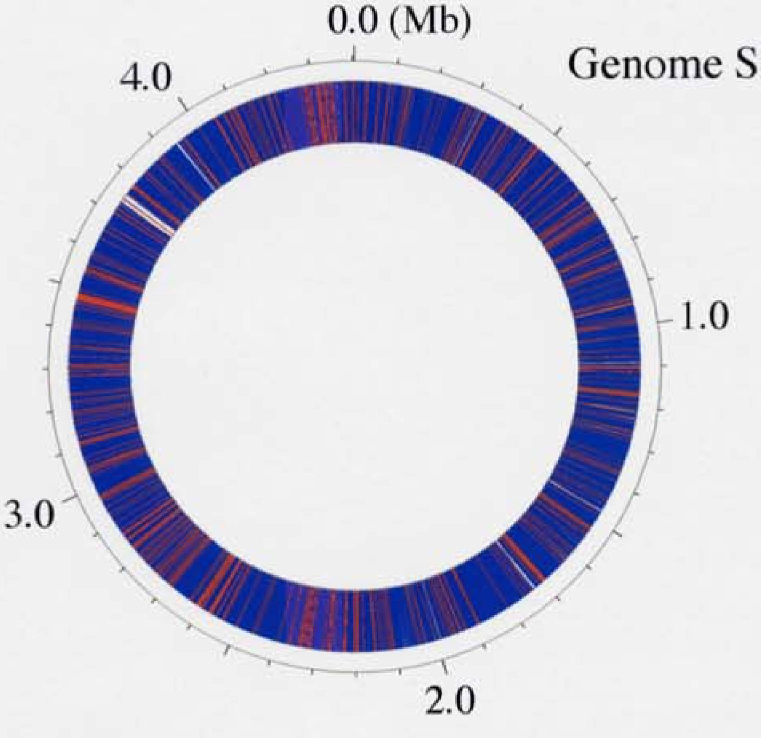
Genome Size = 3,268,203 bp

Mycobacterium tuberculosis (circular)



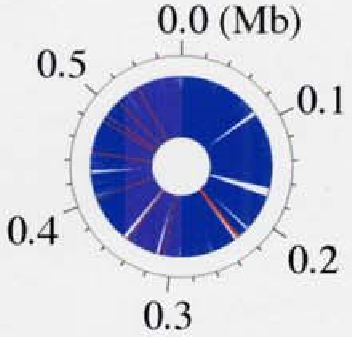
Genome Size = 4,411,529 bp

Mycobacterium tuberculosis CDC1551 (circular)



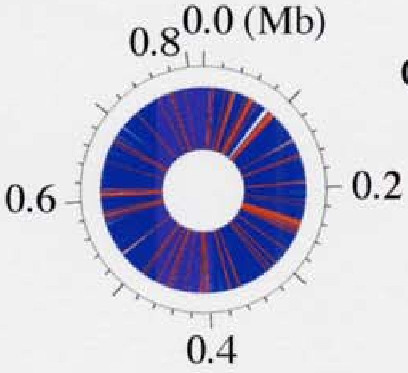
Genome Size = 4,403,836 bp

Mycoplasma genitalium (circular)



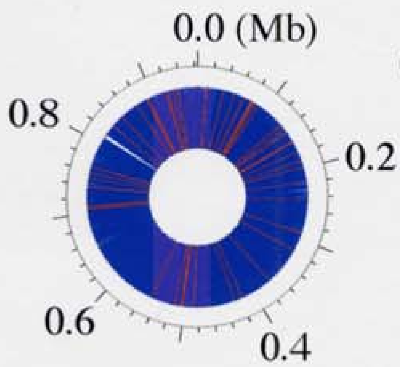
Genome Size = 580,074 bp

Mycoplasma pneumoniae (circular)



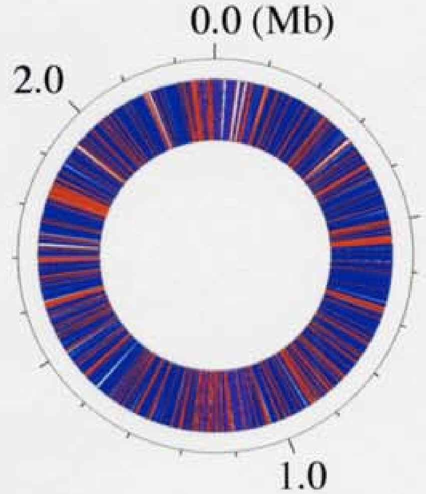
Genome Size = 816,394 bp

Mycoplasma pulmonis (circular)



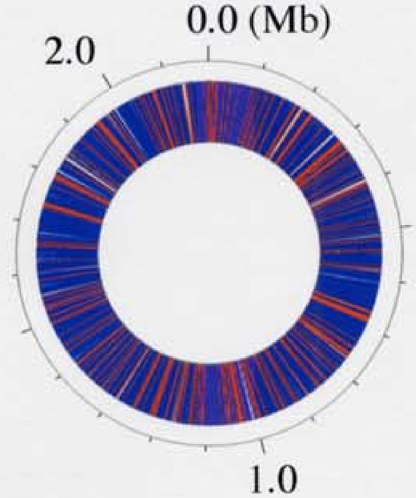
Genome Size = 963,879 bp

Neisseria meningitidis MC58 (serogroup B) (circular)



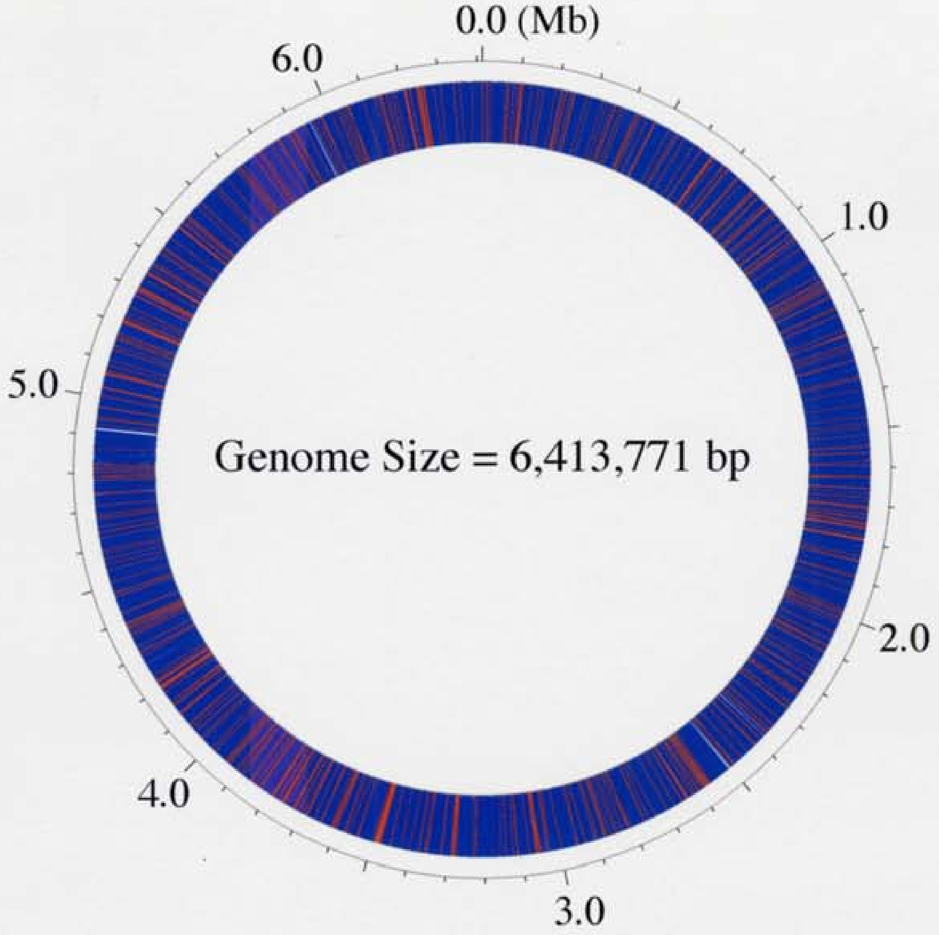
Genome Size = 2,272,351 bp

Neisseria meningitidis Z2491 (serogroup A) (circular)

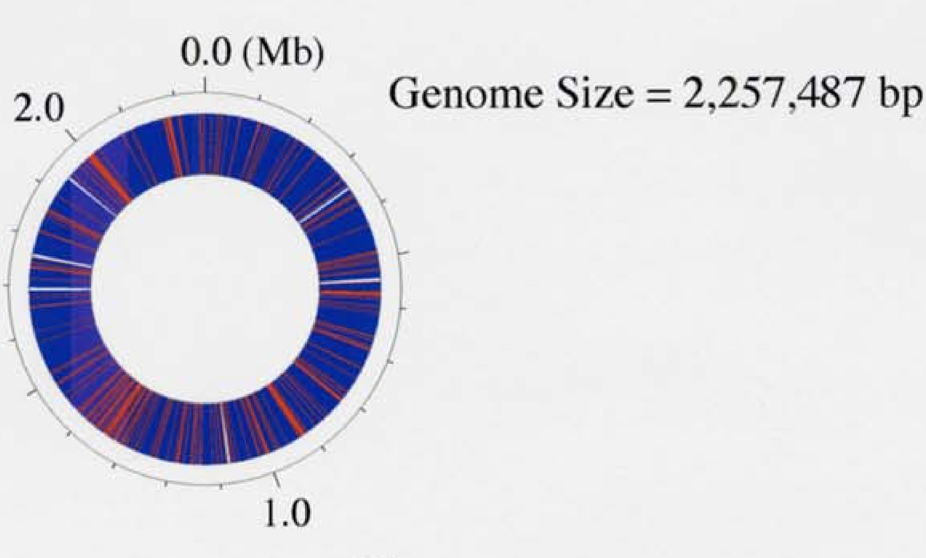


Genome Size = 2,184,406 bp

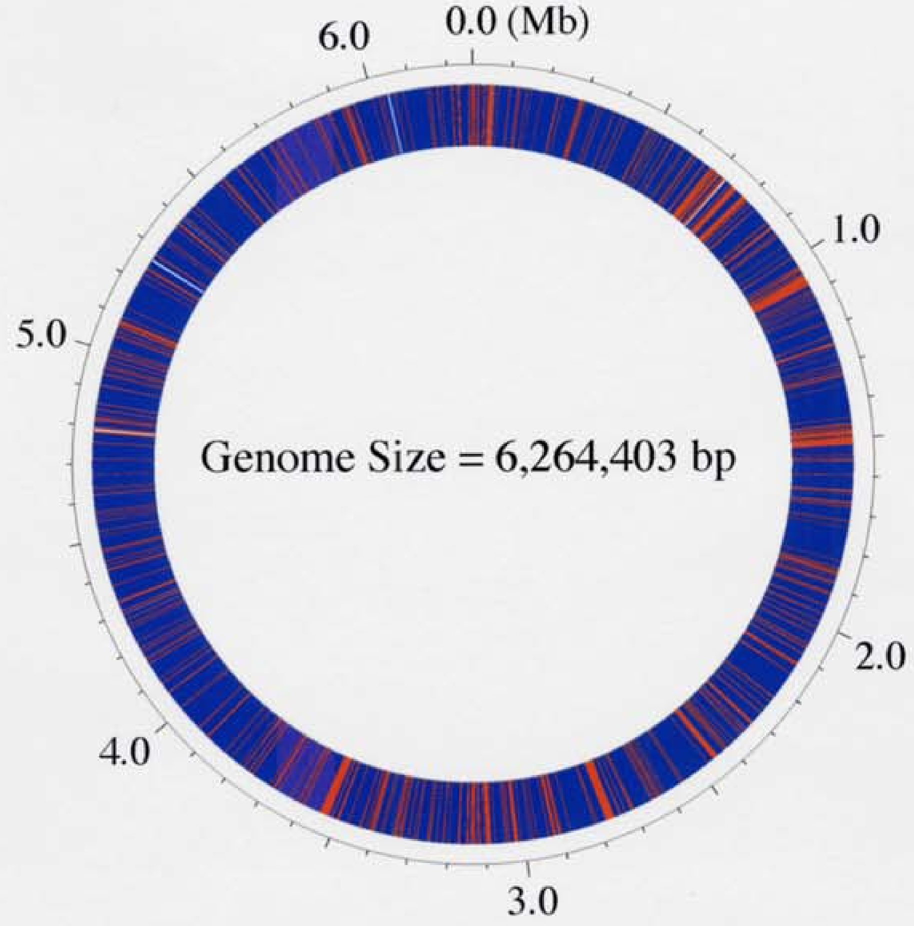
Nostoc sp. PCC 7120 (circular)



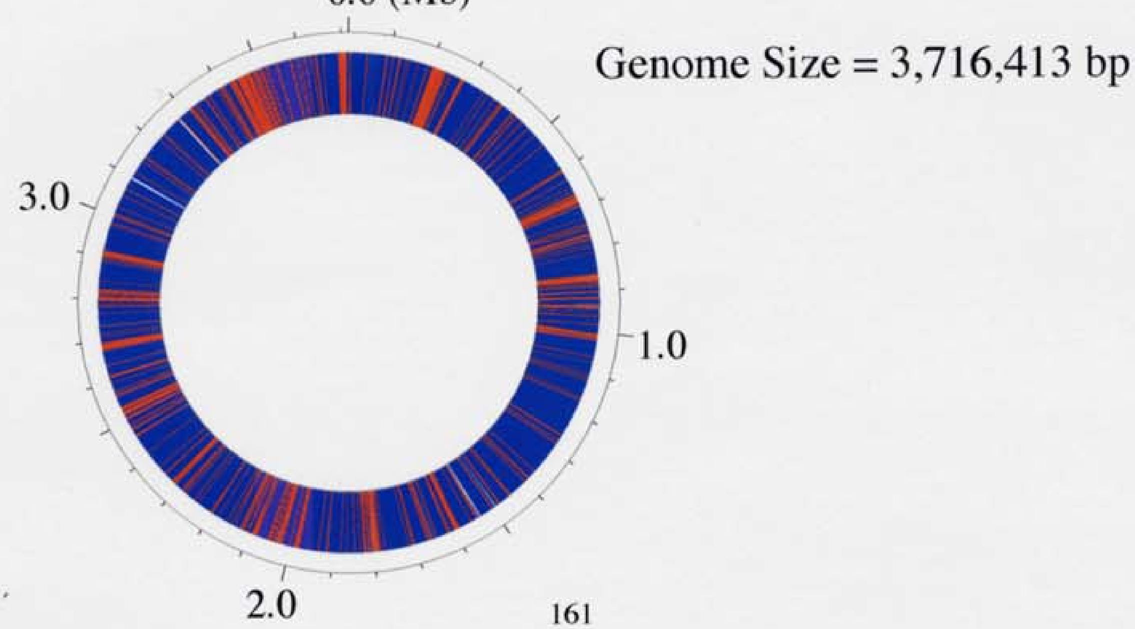
Pasteurella multocida (circular)



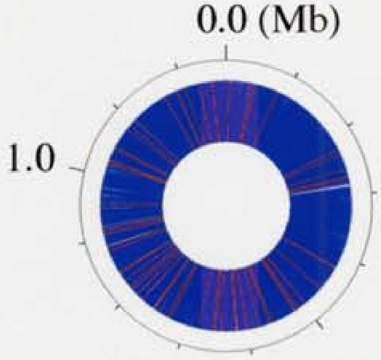
Pseudomonas aeruginosa (circular)



Ralstonia solanacearum (circular)

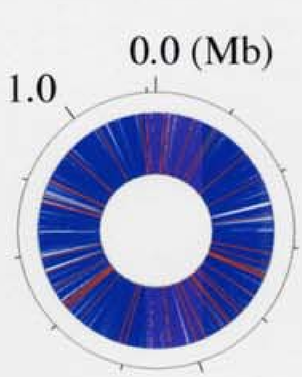


Rickettsia conorii (circular)



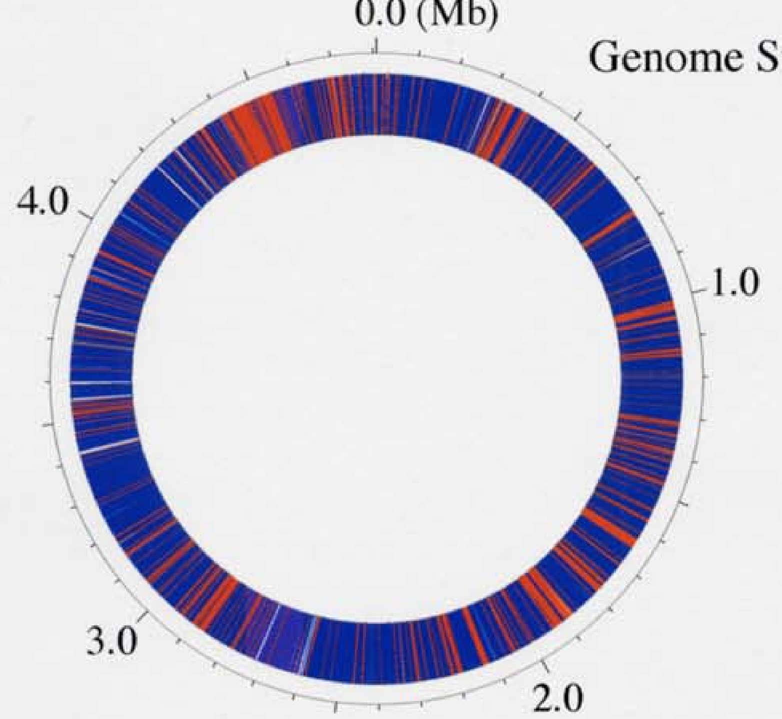
Genome Size = 1,268,755 bp

Rickettsia prowazekii (circular)



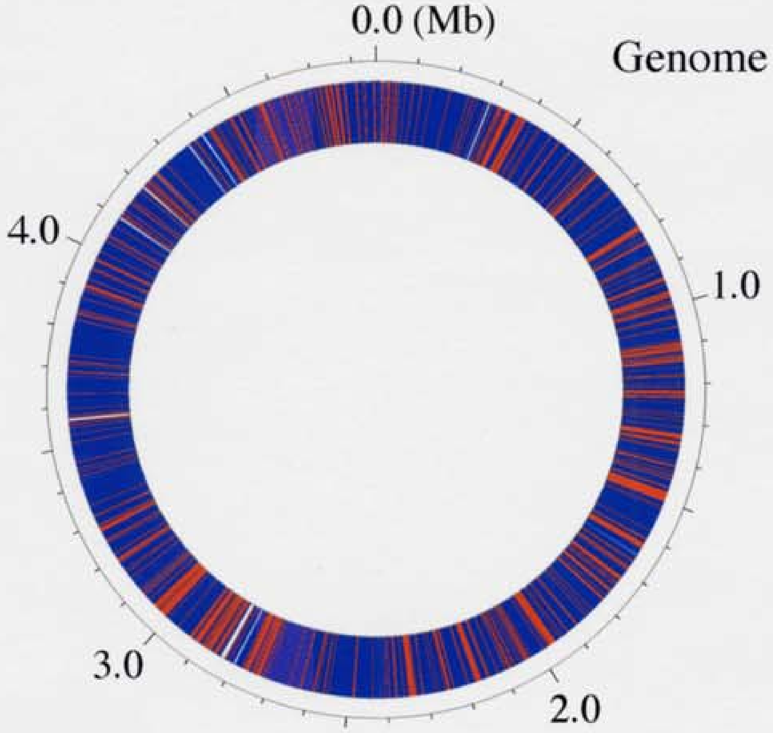
Genome Size = 1,111,523 bp

Salmonella enterica subsp. *enterica* serovar Typhi (circular)



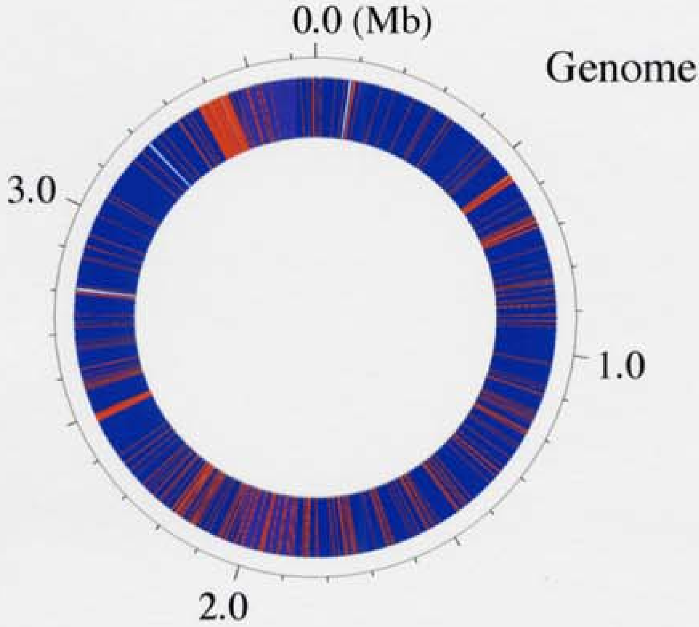
Genome Size = 4,809,037 bp

Salmonella typhimurium LT2 (circular)



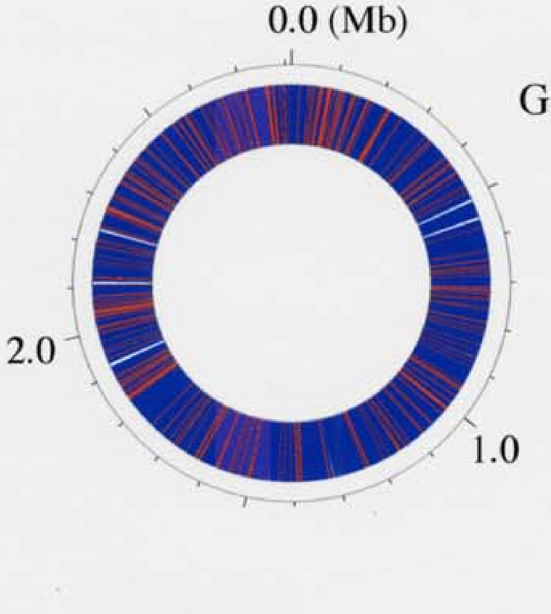
Genome Size = 4,857,432 bp

Sinorhizobium meliloti (circular)



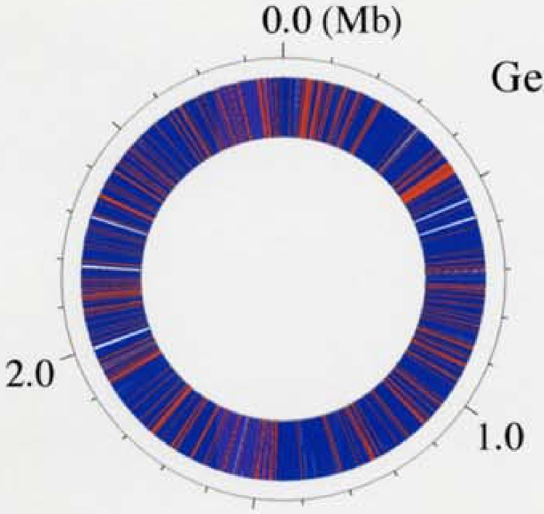
Genome Size = 3,654,135 bp

Staphylococcus aureus N315 (circular)



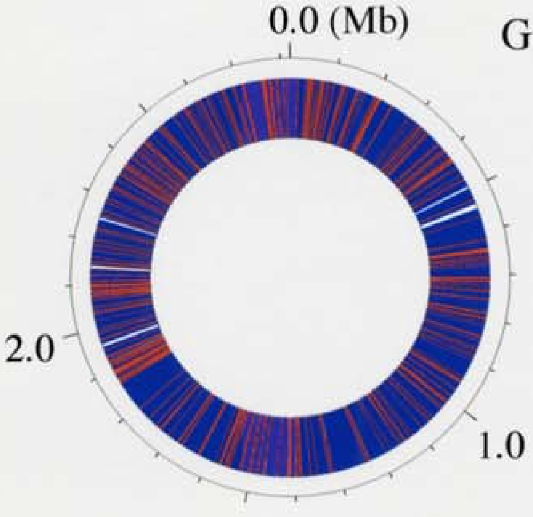
Genome Size = 2,813,641 bp

Staphylococcus aureus Mu50 (circular)



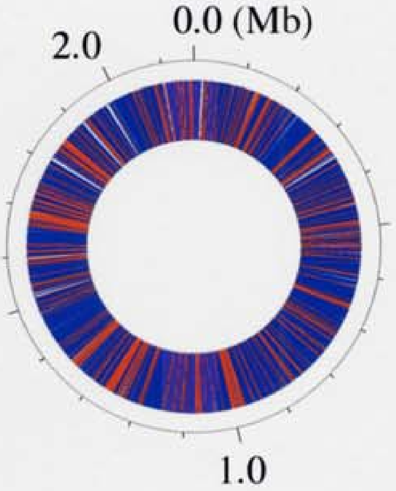
Genome Size = 2,878,134 bp

Staphylococcus aureus MW2 (circular)



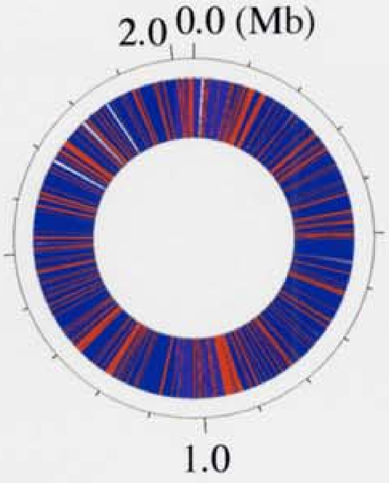
Genome Size = 2,820,462 bp

Streptococcus pneumoniae (circular)



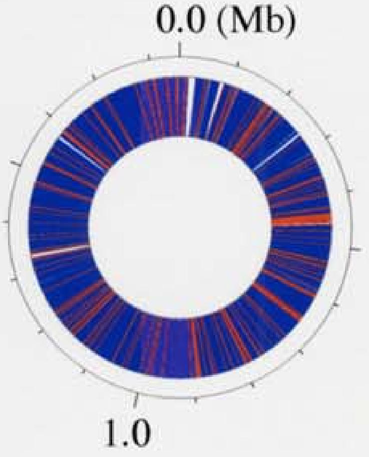
Genome Size = 2,160,837 bp

Streptococcus pneumoniae R6 (circular)



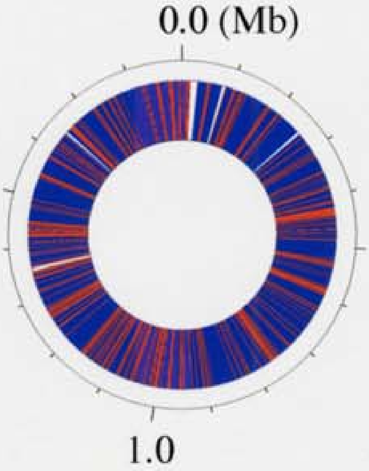
Genome Size = 2,038,615 bp

Streptococcus pyogenes SF370 (circular)



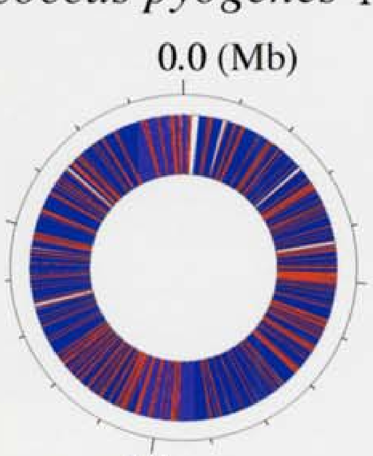
Genome Size = 1,852,441 bp

Streptococcus pyogenes MGAS315 (circular)



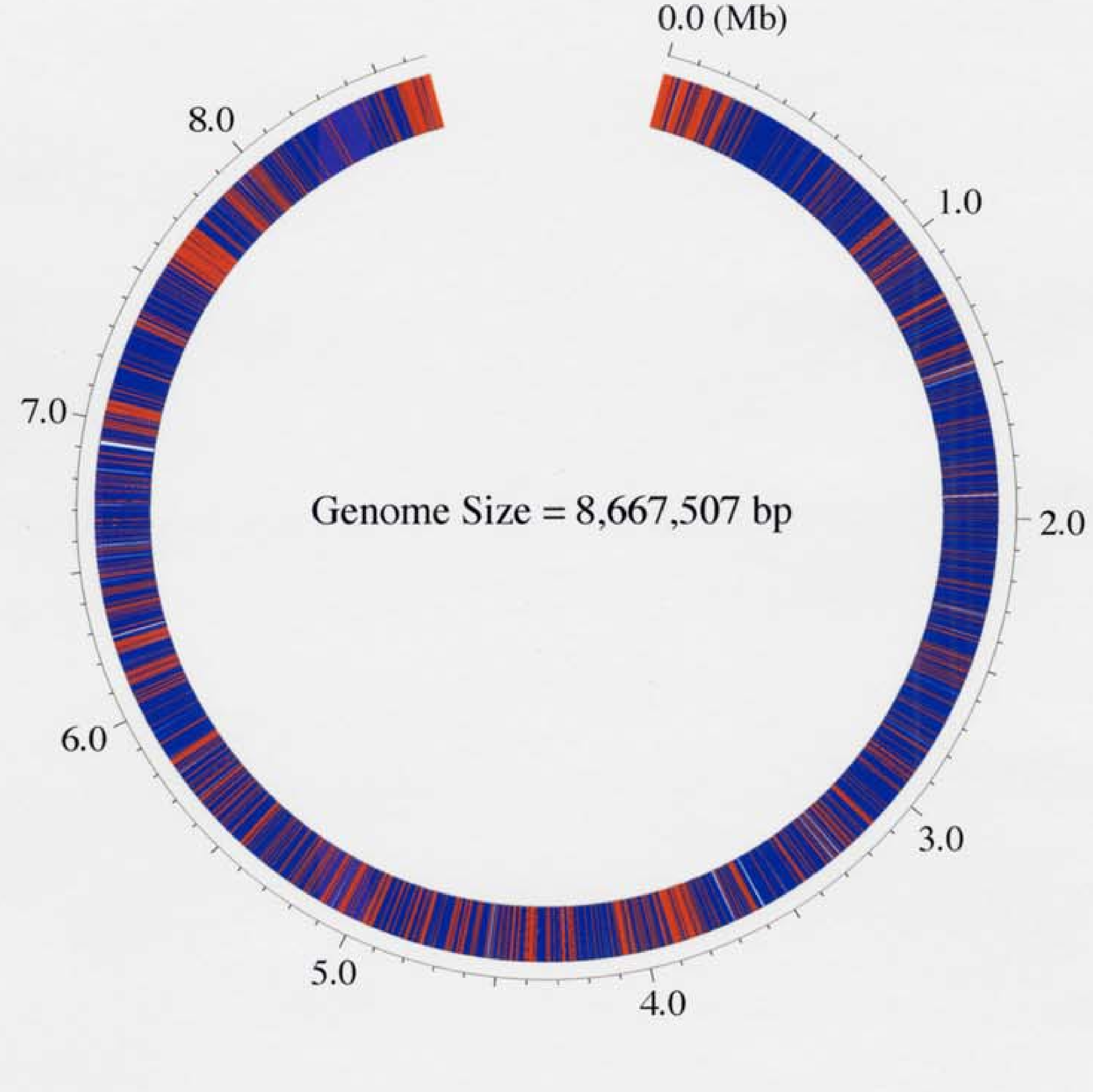
Genome Size = 1,900,521 bp

Streptococcus pyogenes MGAS8232 (circular)

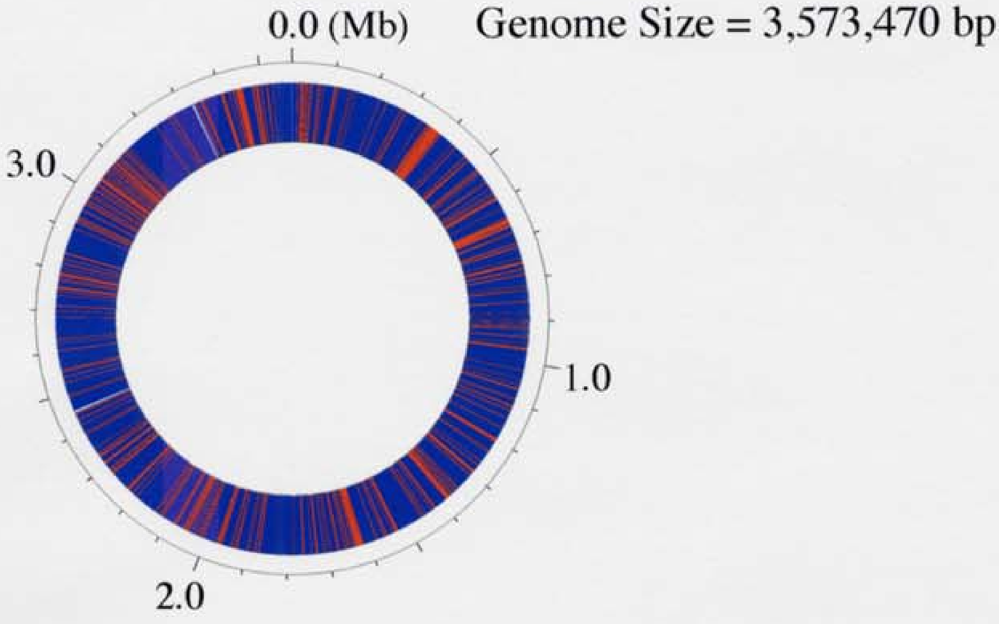


Genome Size = 1,895,017 bp

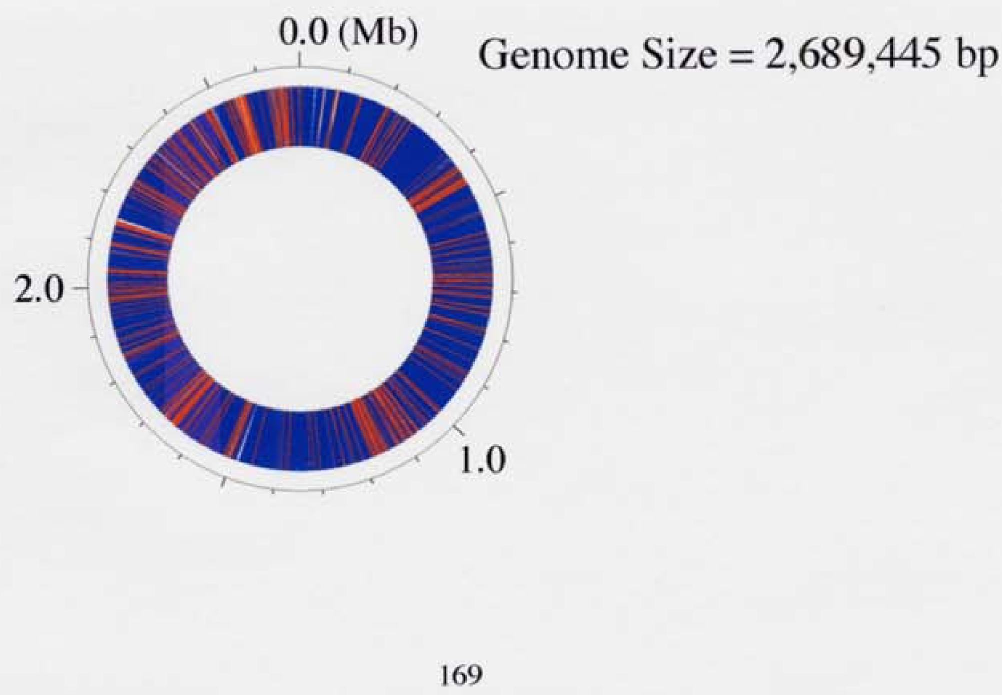
Streptomyces coelicolor (linear)



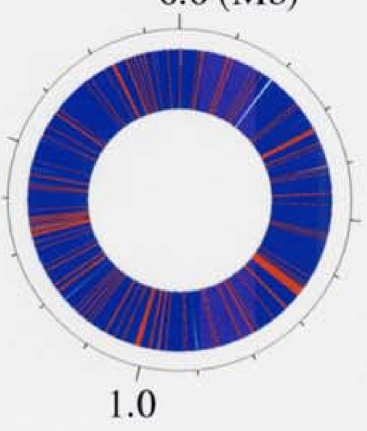
Synechocystis PCC6803 (circular)



Thermoanaerobacter tengcongensis (circular)

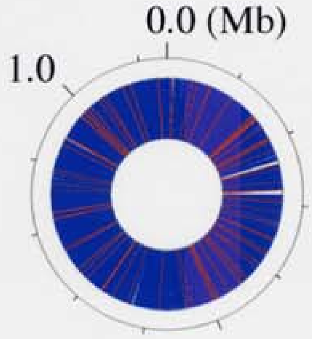


Thermotoga maritima (circular)



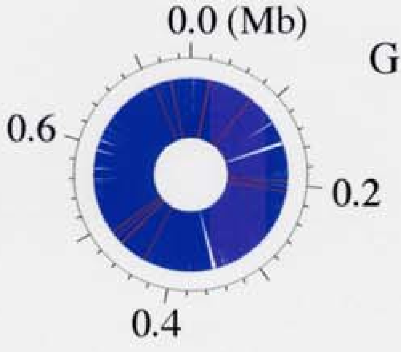
Genome Size = 1,860,725 bp

Treponema pallidum (circular)



Genome Size = 1,138,011 bp

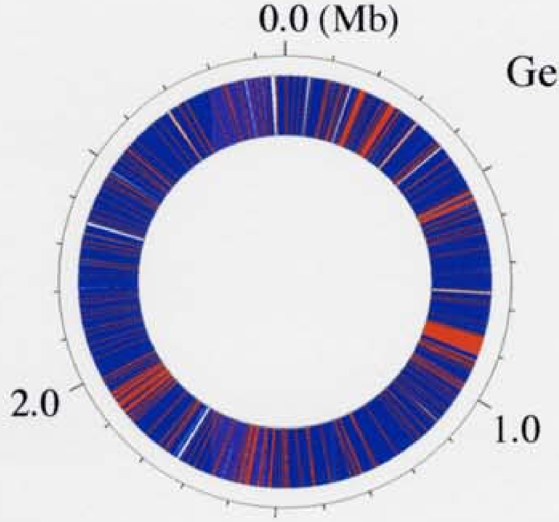
Ureaplasma urealyticum (circular)



Genome Size = 751,719 bp

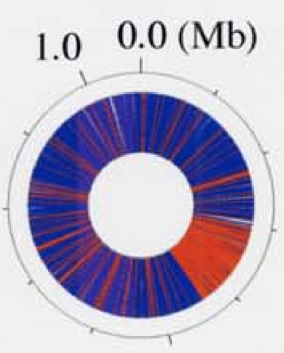
Vibrio cholerae

Chromosome 1 (circular)



Genome Size = 2,961,149 bp

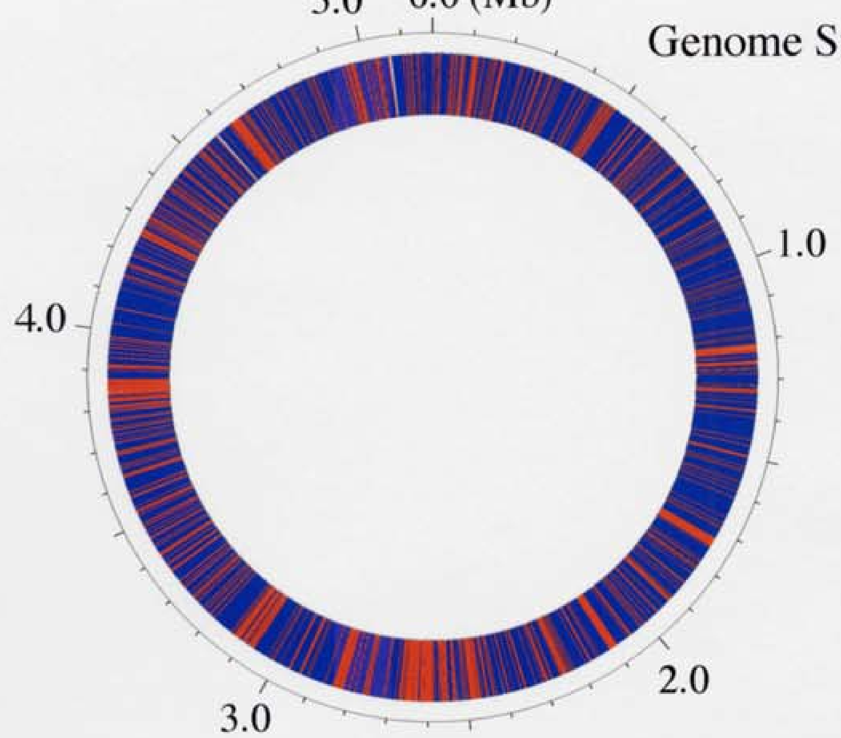
Chromosome 2 (circular)



Genome Size = 1,072,315 bp

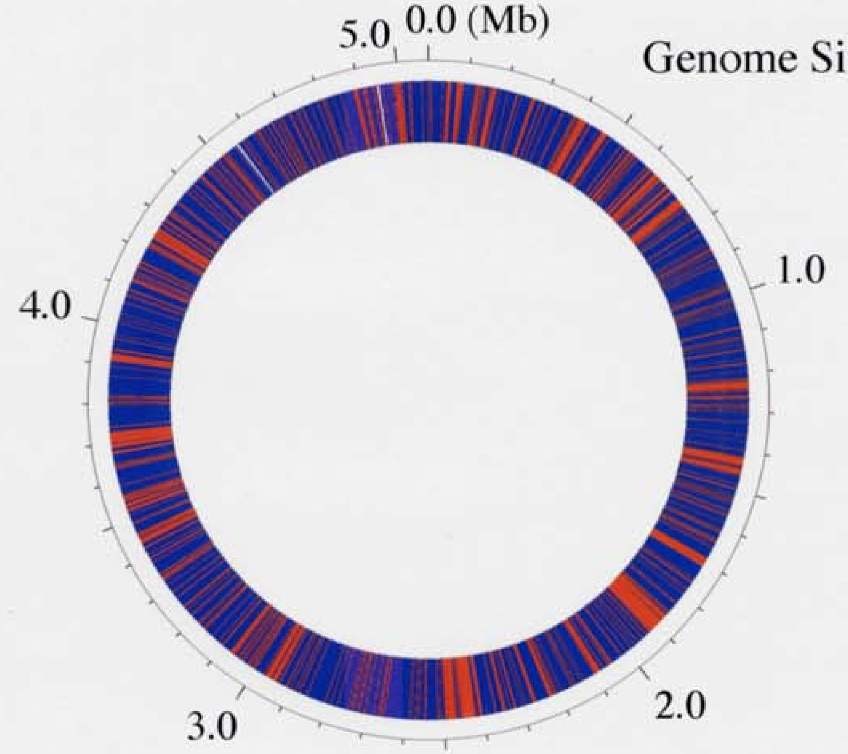
Xanthomonas axonopodis pv. citri (circular)

Genome Size = 5,175,554 bp

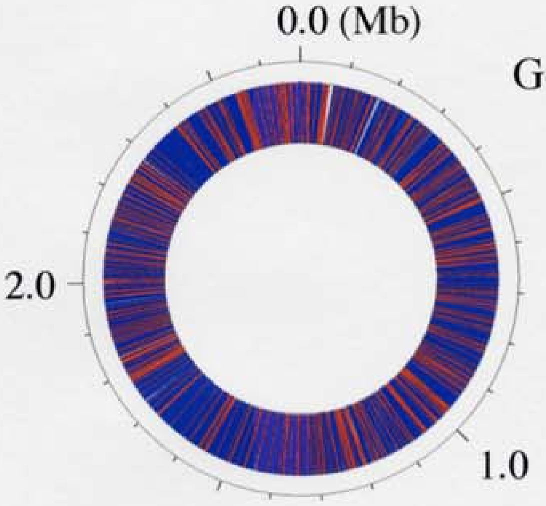


Xanthomonas campestris pv. campestris (circular)

Genome Size = 5,076,188 bp

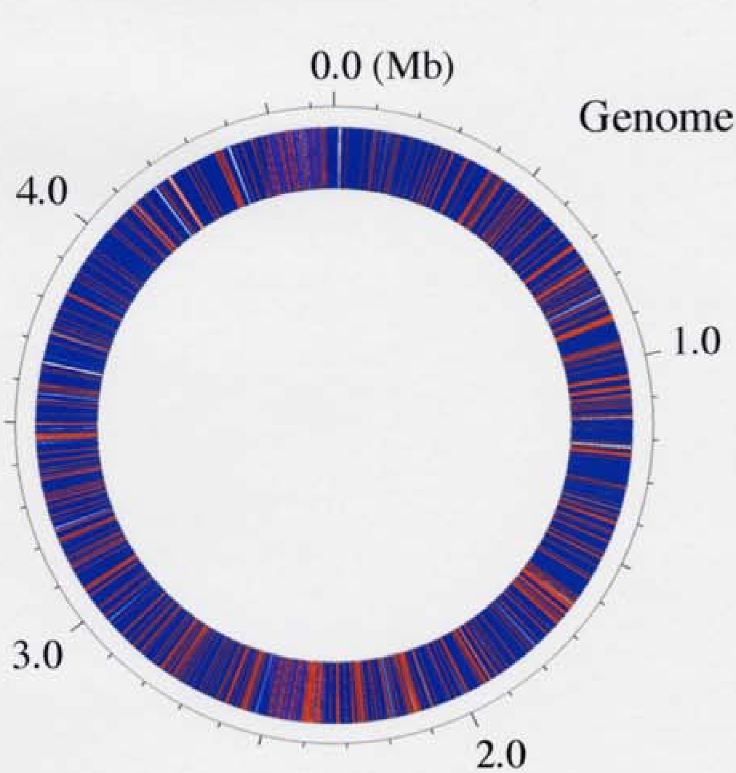


Xylella fastidiosa (circular)



Genome Size = 2,679,306 bp

Yersinia pestis CO92 (circular)



Genome Size = 4,653,728 bp