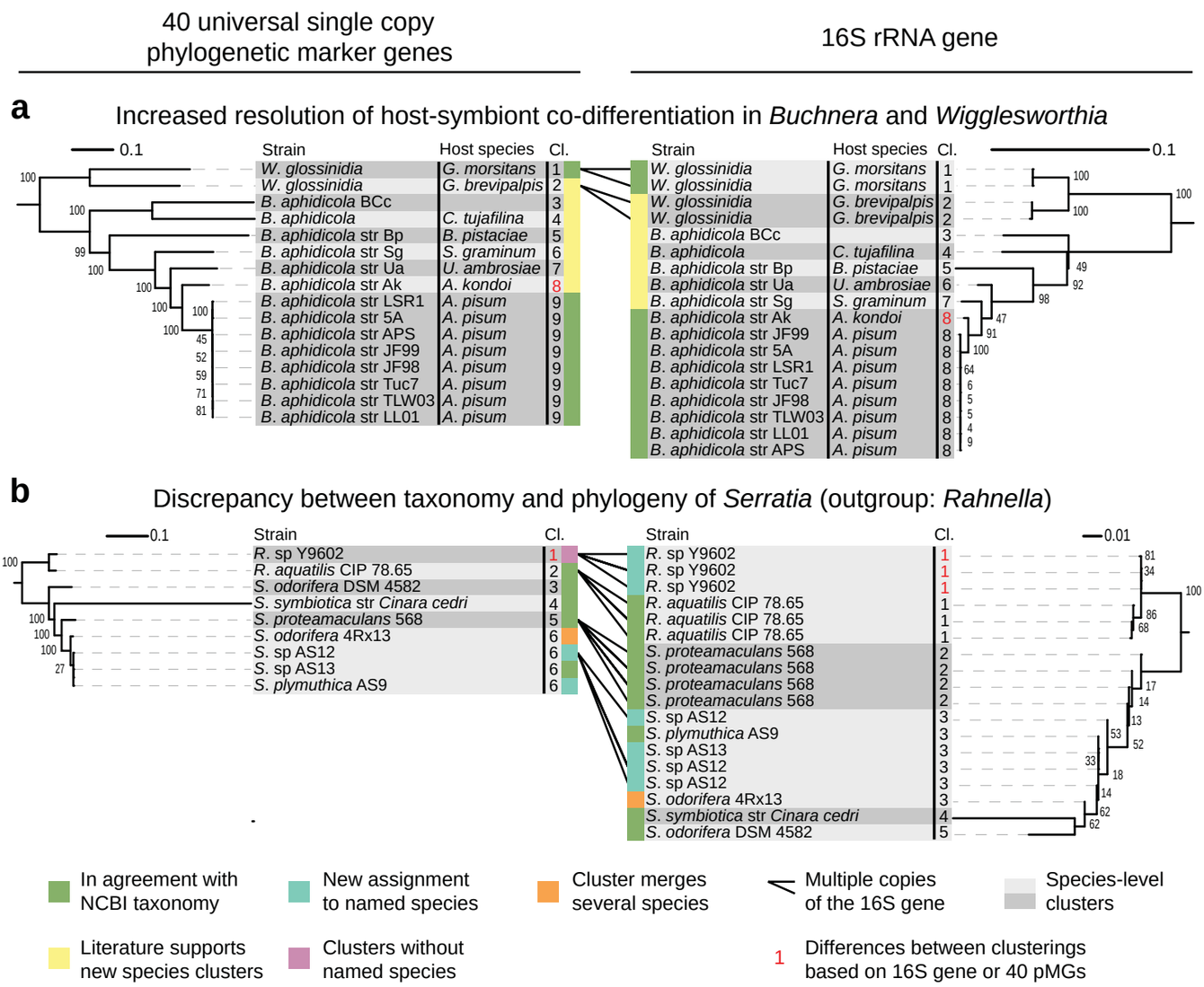
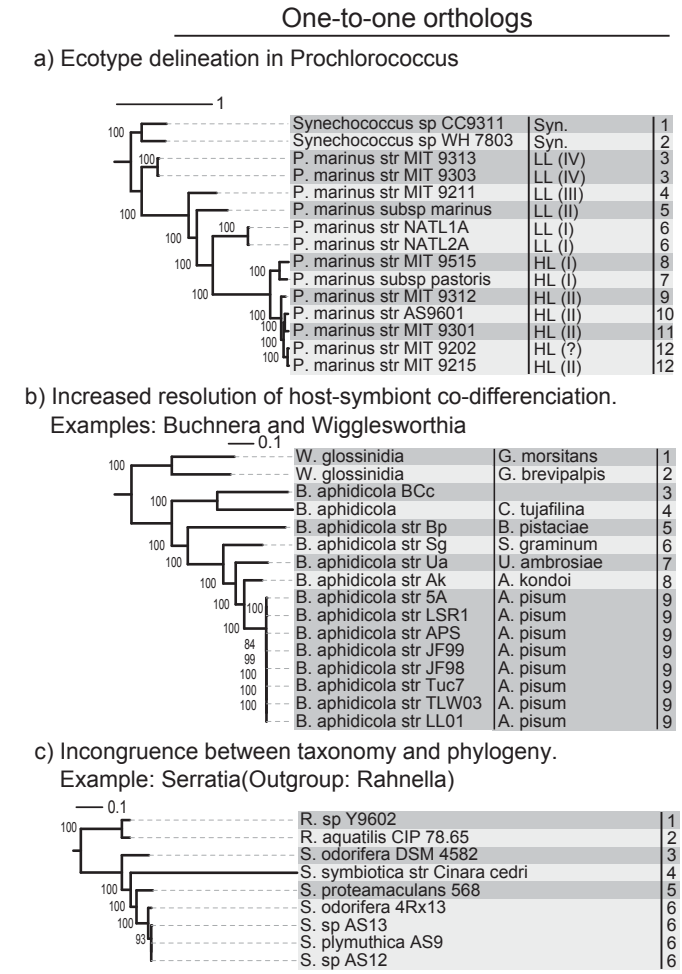
Supplementary Figure 5: Phylogenetic trees and species-level clustering of two additional clades displaying discrepancies to the NCBI taxonomy



Phylogenetic trees and species-level clustering of two additional clades displaying discrepancies to the NCBI taxonomy. Trees were independently built either from a concatenation of 40 universal single-copy phylogenetic marker genes (pMGs) or from the 16S rRNA gene.
a) Host-symbiont co-speciation in Buchnera and Wigglesworthia is consistent with the highly resolved species clustering based on the 40 concatenated pMGs and in part also supported by the 16S rRNA gene.
b) Discrepancy between taxonomy and reconstructed phylogeny of Serratia. Both the combined MGs and the 16S rRNA gene support a reclassification of strain S. odorifera 4Rx13 forming one species together with S. plymuthica AS9 (and the two unnamed genomes sp AS12 and sp AS13). Please also note the long branch found for S. symbiotica str Cinara cedri, that is representative of a change of lifestyle from facultative to obligatory symbionts (Lamelas et al. PLoS Genetics, 2011).
Color key: clustering in agreement with the NCBI taxonomy (blue); clustering discrepant to the NCBI taxonomy, but supported by literature information (yellow); assignment of an unnamed species to a cluster formed of a named species (pink); species clusters that do not contain any named species (dark violet); clusters which merge different species and thus have multiple species names (orange); genomes of species that are split among different clusters (red boxes);discrepancies between the clusterings of 40 pMGs and the 16S rRNA gene (red numbers, other are in black print); multiple copies of the 16S rRNA gene in one genome (black lines; identical copies were removed beforehand).

Nature Methods: doi:10.1038/nmeth.2575

Supplementary Figure 6: Core genome based phylogenetic trees and species level clustering of three clades displaying discrepancies to the NCBI taxonomy



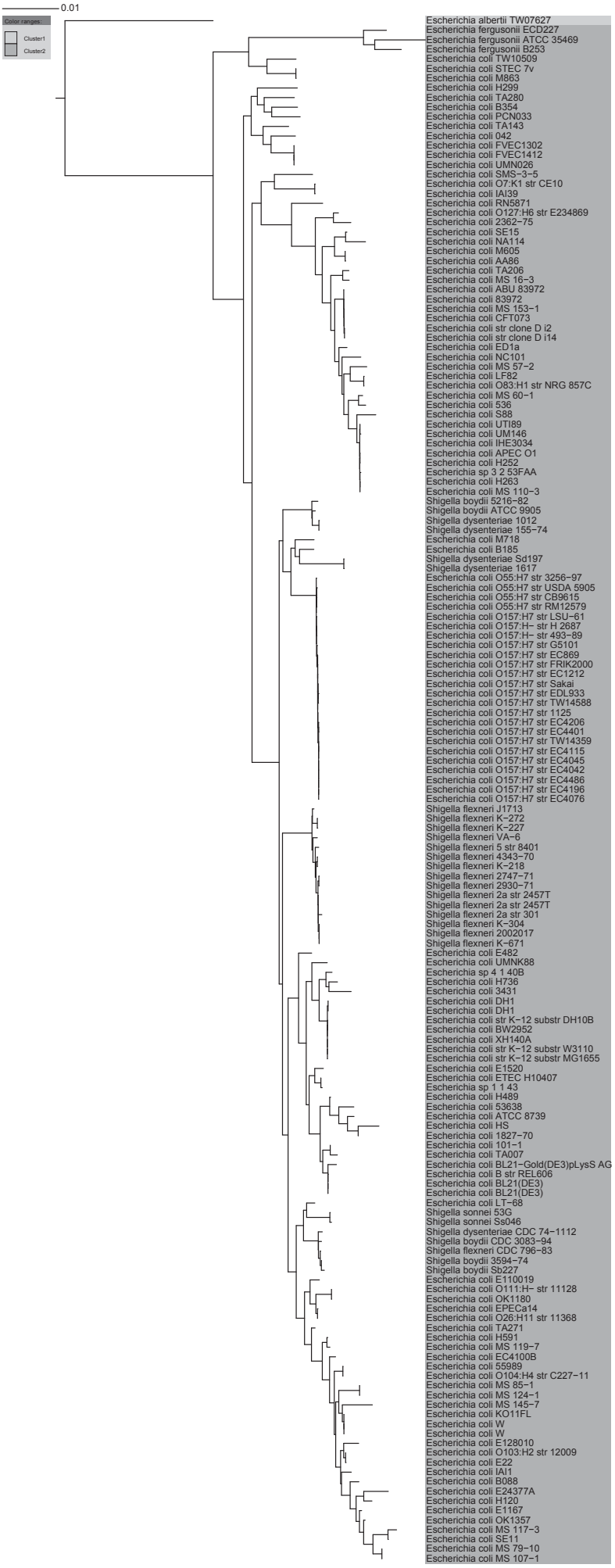
Phylogenetic trees and species level clustering of three clades displaying discrepancies to the NCBI taxonomy. Trees were independently built from a concatenation of all one-to-one orthologs found in the core-genomes used to build the trees.

a) Species-level clusterings and phylogenetic trees of the all one-to-one orthologs support speciation of so-called ecotypes in Prochlorococcus marinus, supporting the finer species resolution of the 40 combined pMGs resolving species in comparison to the 16S rRNA gene.

b) Discrepancy between taxonomy and reconstructed phylogeny of Serratia, especially strain S. odorifera 4Rx13 is also supported by trees built from all one-to-one orthologs

c) Host-symbiont co-differentiation in Buchnera and Wigglesworthia visible when using the 40 combined pMGs (Supplementary Fig. 5a) is supported by phylogenetic trees built from all one-to-one orthologs.

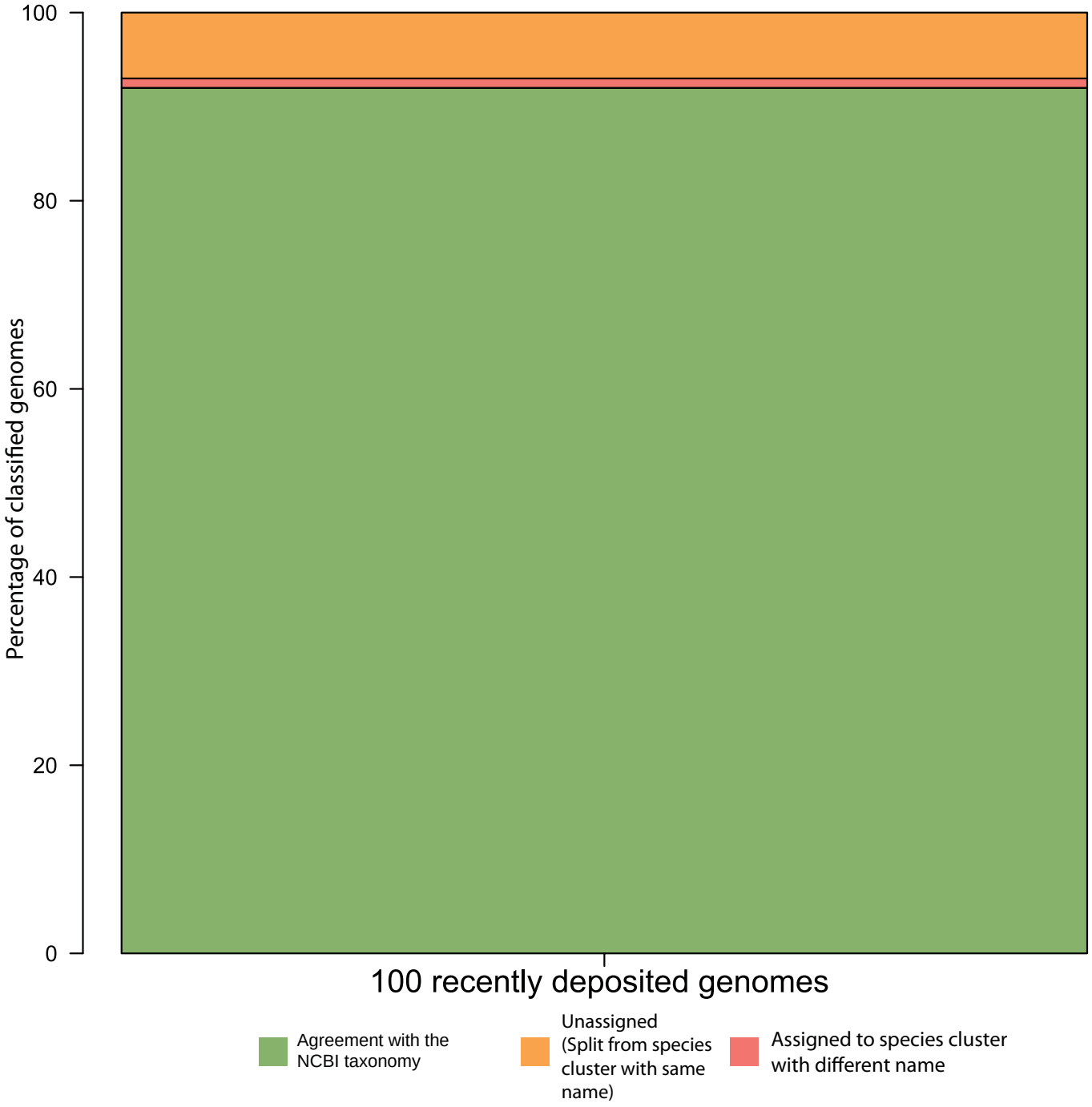
Supplementary Figure 7: Phylogenetic tree of the Escherichia/Shigella clade based on the 40 universal single copy marker genes (MGs).



Phylogenetic tree of the *Escherichia*/*Shigella* clade based on the 40 universal single copy phylogenetic marker genes (pMGs). The polyphyly of *Shigella* reported earlier is clearly visible. Additionally, the tree shows that the phylogenetic distance between *Escherichia* fergusoniiand the closest *Escherichia coli* genome are phylogenetically closer related than some of the *Escherichia coli* genomes, explaining why *Escherichia fergusonii* and *Escherichia coli* are found to be in the same cluster by our method.

Nature Methods: doi:10.1038/nmeth.2575

Supplementary Figure 8: Evaluation of the accuracy of genomic placements for specI



Evaluation of the accuracy of genomic placements for specI based on 100 recently sequenced genomes (also see Supplementary Table 19 for details on the discrepancies).

Supplementary Table 1

An example of high intra-species 16S rRNA gene divergences. All pair-wise divergences between different 16S rRNA gene copies of *Desulfitobacterium hafniense* Y51.
All values above 3% are marked in red as this represents the classical 97% sequence identity cutoff used to delineate species.

Nucleotide divergence between 16S rRNA gene copies	138119.DSY_16SrRNA1	138119.DSY_16SrRNA2	138119.DSY_16SrRNA3	138119.DSY_16SrRNA4	138119.DSY_16SrRNA5	138119.DSY_16SrRNA6
138119.DSY_16SrRNA1	0.00%	0.38%	0.06%	4.65%	4.76%	5.20%
138119.DSY_16SrRNA2	0.38%	0.00%	0.32%	4.78%	4.64%	4.95%
138119.DSY_16SrRNA3	0.06%	0.32%	0.00%	4.71%	4.70%	5.26%
138119.DSY_16SrRNA4	4.65%	4.78%	4.71%	0.00%	4.42%	0.56%
138119.DSY_16SrRNA5	4.76%	4.64%	4.70%	4.42%	0.00%	4.59%
138119.DSY_16SrRNA6	5.20%	4.95%	5.26%	0.56%	4.59%	0.00%

Desulfitobacterium hafniense Y51 (NCBI Taxonomy ID: 138119)

Supplementary Table 2

An example of extremely low inter-species 16S rRNA gene divergences.
The rows represent different copies of the 16S rRNA gene in Aeromonas salmonicida subsp. salmonicida A449 while the columns represent different copies of the 16S rRNA gene in Aeromonas hydrophila subsp. hydrophila ATCC 7966

Nucleotide divergence between 16S rRNA gene copies	380703.AHA_0077	380703.AHA_0145	380703.AHA_0198	380703.AHA_0335	380703.AHA_0752	380703.AHA_0865	380703.AHA_3516	380703.AHA_3944	380703.AHA_4016	380703.AHA_4246
382245.ASA_0079	1.23%	1.23%	1.30%	1.23%	1.23%	1.23%	1.30%	1.23%	1.23%	1.23%
382245.ASA_0295	1.09%	1.09%	1.16%	1.09%	1.09%	1.09%	1.16%	1.09%	1.09%	1.09%
382245.ASA_0801	0.96%	0.96%	1.02%	0.96%	0.96%	0.96%	1.02%	0.96%	0.96%	0.96%
382245.ASA_3425	1.09%	1.09%	1.16%	1.09%	1.09%	1.09%	1.16%	1.09%	1.09%	1.09%
382245.ASA_3599	1.09%	1.09%	1.16%	1.09%	1.09%	1.09%	1.16%	1.09%	1.09%	1.09%
382245.ASA_4060	1.09%	1.09%	1.16%	1.09%	1.09%	1.09%	1.16%	1.09%	1.09%	1.09%
382245.ASA_4192	1.23%	1.23%	1.30%	1.23%	1.23%	1.23%	1.30%	1.23%	1.23%	1.23%
382245.ASA_4244	1.09%	1.09%	1.16%	1.09%	1.09%	1.09%	1.16%	1.09%	1.09%	1.09%
382245.ASA_4317	1.16%	1.16%	1.23%	1.16%	1.16%	1.16%	1.23%	1.16%	1.16%	1.16%

Aeromonas salmonicida subsp. salmonicida A449 (NCBI Taxonomy ID: 382245)
Aeromonas hydrophila subsp. hydrophila ATCC 7966 (NCBI Taxonomy ID: 380703)

Supplementary Table 3

List of the 40 universal single copy phylogenetic marker genes used in this study.

COG0012	Predicted GTPase, probable translation factor
COG0016	Phenylalanyl-tRNA synthetase alpha subunit
COG0018	Arginyl-tRNA synthetase
COG0048	Ribosomal protein S12
COG0049	Ribosomal protein S7
COG0052	Ribosomal protein S2
COG0080	Ribosomal protein L11
COG0081	Ribosomal protein L1
COG0085	DNA-directed RNA polymerase, beta subunit/140 kD subunit
COG0087	Ribosomal protein L3
COG0088	Ribosomal protein L4
COG0090	Ribosomal protein L2
COG0091	Ribosomal protein L22
COG0092	Ribosomal protein S3
COG0093	Ribosomal protein L14
COG0094	Ribosomal protein L5
COG0096	Ribosomal protein S8
COG0097	Ribosomal protein L6P/L9E
COG0098	Ribosomal protein S5
COG0099	Ribosomal protein S13
COG0100	Ribosomal protein S11
COG0102	Ribosomal protein L13
COG0103	Ribosomal protein S9
COG0124	Histidyl-tRNA synthetase
COG0172	Seryl-tRNA synthetase
COG0184	Ribosomal protein S15P/S13E
COG0185	Ribosomal protein S19
COG0186	Ribosomal protein S17
COG0197	Ribosomal protein L16/L10E
COG0200	Ribosomal protein L15
COG0201	Preprotein translocase subunit SecY
COG0202	DNA-directed RNA polymerase, alpha subunit/40 kD subunit
COG0215	Cysteinyl-tRNA synthetase
COG0256	Ribosomal protein L18
COG0495	Leucyl-tRNA synthetase
COG0522	Ribosomal protein S4 and related proteins
COG0525	Valyl-tRNA synthetase
COG0533	Metal-dependent proteases with possible chaperone activity
COG0541	Signal recognition particle GTPase
COG0552	Signal recognition particle GTPase

Supplementary Table 5

Discrepancies found between the species clustering and the type strain taxonomy. All species for which a sequenced type strain could be found were included. Please note that some of the discrepancies could later be curated using literature information (see Supplementary Table 16 for details)

Type of Discepancy	NCBI Taxonomy ID	Genome Project ID	NCBI Taxonomy Name	Cluster Majority Taxonomy Name	Cluster ID
Split	491916	59115	Rhizobium etli	Rhizobium etli	682
Merged	590168	42777	Thermotoga naphthophila	Thermotoga petrophila	1224
Split	682634	42253	Serratia odorifera	Serratia odorifera	373
Split	525919	59219	Anaerococcus prevotii	Anaerococcus prevotii	1501
Split	1037409	67463	Bradyrhizobium japonicum	Bradyrhizobium japonicum	699
Merged	888063	67395	Enterobacter hormaechei	Enterobacter cloacae	384
Split	718254	45967	Enterobacter cloacae	Enterobacter cloacae	386
Split	701347	59969	Enterobacter cloacae	Enterobacter cloacae	386
Split	716541	48363	Enterobacter cloacae	Enterobacter cloacae	386
Split	717608	45855	Clostridium saccharolyticum	Clostridium saccharolyticum	1602
Merged	583358	49481	Thermoanaerobacter mathranii	Thermoanaerobacter italicus	1524
Merged	509193	55639	Thermoanaerobacter brockii	Thermoanaerobacter pseudethanolicus	1525
Split	866774	51527	Atopobium vaginae	Atopobium vaginae	810
Split	706437	61277	Streptococcus anginosus	Streptococcus anginosus	1393
Split	521097	59197	Capnocytophaga ochracea	Capnocytophaga ochracea	1121
Merged	997348	70617	Neisseria macacae	Neisseria sicca	437
Merged	435590	58253	Bacteroides vulgatus	Bacteroides dorei	1104
Merged	702446	47771	Bacteroides vulgatus	Bacteroides dorei	1104
Split	883	59089	Desulfovibrio vulgaris	Desulfovibrio vulgaris	754
Split	349102	58451	Rhodobacter sphaeroides	Rhodobacter sphaeroides	601
Split	391616	54699	Octadecabacter antarcticus	Octadecabacter antarcticus	635
Split	452638	58967	Polynucleobacter necessarius	Polynucleobacter necessarius	466
Merged	632516	60575	Caldicellulosiruptor lactoaceticus	Caldicellulosiruptor kristjanssonii	1511
Split	393480	54419	Fusobacterium nucleatum	Fusobacterium nucleatum	1479
Merged	566552	55369	Bifidobacterium catenulatum	Bifidobacterium pseudocatenulatum	973
Split	331678	58131	Chlorobium phaeobacteroides	Chlorobium phaeobacteroides	1046
Merged	868595	67317	Desulfotomaculum carboxydivorans	Desulfotomaculum nigrificans	1680
Split	637911	55947	Actinobacillus minor	Actinobacillus minor	310
Merged	526997	55207	Bacillus mycoides	Bacillus cereus	1337
Split	553190	43211	Gardnerella vaginalis	Gardnerella vaginalis	965
Split	682147	49423	Gardnerella vaginalis	Gardnerella vaginalis	965
Split	682148	49673	Gardnerella vaginalis	Gardnerella vaginalis	965
Split	457424	55575	Bacteroides fragilis	Bacteroides fragilis	1090
Split	927666	65449	Streptococcus oralis	Streptococcus oralis	1410
Split	871237	59469	Streptococcus infantis	Streptococcus infantis	1415
Split	1005705	67809	Streptococcus infantis	Streptococcus infantis	1415
Split	1035189	67191	Streptococcus infantis	Streptococcus infantis	1415
Merged	340047	58525	Mycoplasma capricolum	Mycoplasma leachii	1252
Merged	440085	58933	Methylobacterium chloromethanicum	Methylobacterium extorquens	710
Split	1048834	68273	Alicyclobacillus acidocaldarius	Alicyclobacillus acidocaldarius	1663
Merged	281309	58089	Bacillus thuringiensis	Bacillus cereus	1338
Merged	412694	58795	Bacillus thuringiensis	Bacillus cereus	1338
Merged	527019	55239	Bacillus thuringiensis	Bacillus cereus	1338
Merged	527021	55223	Bacillus thuringiensis	Bacillus cereus	1338
Merged	527022	55215	Bacillus thuringiensis	Bacillus cereus	1338
Merged	527024	55219	Bacillus thuringiensis	Bacillus cereus	1338
Merged	527025	55225	Bacillus thuringiensis	Bacillus cereus	1338
Merged	527028	55237	Bacillus thuringiensis	Bacillus cereus	1338
Merged	527029	55233	Bacillus thuringiensis	Bacillus cereus	1338
Merged	527030	55235	Bacillus thuringiensis	Bacillus cereus	1338
Merged	527031	55229	Bacillus thuringiensis	Bacillus cereus	1338
Merged	527032	55231	Bacillus thuringiensis	Bacillus cereus	1338
Merged	541229	43737	Bacillus thuringiensis	Bacillus cereus	1338
Merged	714359	49135	Bacillus thuringiensis	Bacillus cereus	1338
Merged	930170	60447	Bacillus thuringiensis	Bacillus cereus	1338
Split	699184	50375	Bacillus thuringiensis	Bacillus cereus	1338
Split	526972	55161	Bacillus thuringiensis	Bacillus cereus	1338
Split	526990	55197	Bacillus thuringiensis	Bacillus cereus	1338
Split	526976	55169	Bacillus thuringiensis	Bacillus cereus	1338
Split	526988	55193	Bacillus thuringiensis	Bacillus cereus	1338
Split	579138	68445	Zymomonas mobilis	Zymomonas mobilis	580
Split	864569	52359	Streptococcus equinus	Streptococcus equinus	1381
Split	451756	54851	Clostridium perfringens	Clostridium perfringens	1541
Split	451757	54853	Clostridium perfringens	Clostridium perfringens	1541
Merged	632348	60491	Caldicellulosiruptor kronotskyensis	Caldicellulosiruptor bescii	1510
Split	246199	47357	Ruminococcus albus	Ruminococcus albus	1566
Merged	1032845	70839	Rickettsia heilongjiangensis	Rickettsia rickettsii	1747
Merged	272951	54113	Rickettsia sibirica	Rickettsia rickettsii	1747
Merged	652620	73963	Rickettsia japonica	Rickettsia rickettsii	1747
Merged	698757	42375	Pyrobaculum oguniense	Pyrobaculum arsenaticum	40
Split	562981	66131	Gemella haemolysans	Gemella haemolysans	1299

Supplementary Table 6

Genomes that were recently made available and are not part of the 3,496 genomes in the species clustering that were not classified by specI according to their NCBI species name (also see Supplementary Figure 8).

NCBI taxonomy ID	NCBI species name	specI assignment	Average pMG sequence identity to most closely related species cluster
1004785	Alteromonas macleodii	UNASSINGED	92.4038405
1123519	Pseudomonas stutzeri	UNASSINGED	89.36955966
1185652	Sinorhizobium fredii	UNASSINGED	94.00208992
1202538	Candidatus Carsonella ruddii	UNASSINGED	<80%
1224746	Gluconobacter oxydans	UNASSINGED	84.39660304
245012	butyrate-producing bacterium SM4/1	Clostridium saccharolyticum	99.67464208
670307	Hyphomicrobium denitrificans	UNASSINGED	88.82299904
710686	Mycobacterium smegmatis	UNASSINGED	86.28364602

Supplementary Table 7

Type strain co-clustering: In this table all co-clustering type strains are listed together with their ANI values as well as their combined 40 pMG percent nucleotide identity, showing that the vast mahority of incongruencies between the specI clustering and the type strain taxonomy are supported ANI.

Type of error	Cluster ID (also see table 1A)	NXBI Taxonomy ID type strain 1	NCBI Taxonomy species name type strain 1	NCBI Taxonomy name type strain 1	NCBI Taxonomy ID type strain 2	NCBI Taxonomy species name type strain 2	NCBI Taxonomy name type strain 2	ANIb	ANIm	combined 40 pMG % nucleotide identity
Merge	1224	590168	Thermotoga naphthophila	Thermotoga naphthophila RKU-10	390874	Thermotoga petrophila	Thermotoga petrophila RKU-1	96.39	97.13	0.987144705
Merge	1524	580331	Thermoanaerobacter italicus	Thermoanaerobacter italicus Ab9	583358	Thermoanaerobacter mathranii	Thermoanaerobacter mathranii subsp. mathranii str. A3	97.84	98.14	0.986785036
Merge	1525	340099	Thermoanaerobacter pseudethanolicus	Thermoanaerobacter pseudethanolicus ATCC 33223	509193	Thermoanaerobacter brockii	Thermoanaerobacter brockii subsp. finnii Ako-1	99.95	99.82	1
Merge (supported by literature)	437	997348	Neisseria macacae	Neisseria macacae ATCC 33926	547045	Neisseria sicca	Neisseria sicca ATCC 29256	96.65	96.71	0.965930679
Merge	1104	435590	Bacteroides vulgatus	Bacteroides vulgatus ATCC 8482	483217	Bacteroides dorei	Bacteroides dorei DSM 17855	94.58	95.85	0.972849828
Merge	1511	632516	Caldicellulosiruptor lactoaceticus	Caldicellulosiruptor lactoaceticus 6A	632335	Caldicellulosiruptor kristjanssonii	Caldicellulosiruptor kristjanssonii 177R1B	98.13	98.22	0.985169584
Merge	973	547043	Bifidobacterium pseudocatenulatum	Bifidobacterium pseudocatenulatum DSM 20438 = JCM 1200	566552	Bifidobacterium catenulatum	Bifidobacterium catenulatum DSM 16992 = JCM 1194	91.96	92.88	0.96692405
Merge	1680	696369	Desulfotomaculum nigrificans	Desulfotomaculum nigrificans DSM 574	868595	Desulfotomaculum carboxydivorans	Desulfotomaculum carboxydivorans CO-1-SRB	99.09	99.06	0.997402205
Merge (supported by literature)	1252	340047	Mycoplasma capricolum	Mycoplasma capricolum ATCC 27343 uid58525	880447	Mycoplasma leachii	Mycoplasma leachii PG50 uid60849	93.26	94.83	0.974388654
Merge	710	440085	Methylobacterium chloromethanicum	Methylobacterium extorquens DM4 uid61617	661410	Methylobacterium extorquens	Methylobacterium chloromethanicum CM4 uid58933	96.49	96.91	0.980051344
Merge (supported by literature)	1338	226900	Bacillus cereus	Bacillus cereus ATCC 14579	527031	Bacillus thuringiensis	Bacillus thuringiensis serovar berliner ATCC 10792	96.68	96.73	0.972805999
Merge	1516	521460	Caldicellulosiruptor bescii	Caldicellulosiruptor bescii DSM 6725	632348	Caldicellulosiruptor kronotskyensis	Caldicellulosiruptor kronotskyensis 2002	95.59	96.14	0.972684753
Merge (supported by literature)	40	340102	Pyrobaculum arsenaticum	Pyrobaculum arsenaticum DSM 13514	698757	Pyrobaculum oguniense	Pyrobaculum oguniense TE7	95.4	96.02	0.968539392
Merge (supported by literature)	1747	1032845	Rickettsia heilongjiangensis	Rickettsia heilongjiangensis 054	652620	Rickettsia japonica	Rickettsia japonica strain YH	99.16	99.04	0.996944196
Merge (supported by literature)	1747	1032845	Rickettsia heilongjiangensis	Rickettsia heilongjiangensis 054	272951	Rickettsia sibirica	Rickettsia sibirica 246	97.4	97.09	0.988646231
Merge (supported by literature)	1747	1032845	Rickettsia heilongjiangensis	Rickettsia heilongjiangensis 054	392021	Rickettsia rickettsii	Rickettsia rickettsii str. 'Sheila Smith'	97.17	96.88	0.986732071
Merge (supported by literature)	1747	652620	Rickettsia japonica	Rickettsia japonica strain YH	272951	Rickettsia sibirica	Rickettsia sibirica 246	97.54	97.18	0.988971202
Merge (supported by literature)	1747	652620	Rickettsia japonica	Rickettsia japonica strain YH	392021	Rickettsia rickettsii	Rickettsia rickettsii str. 'Sheila Smith'	97.31	96.94	0.986706217
Merge (supported by literature)	1747	272951	Rickettsia sibirica	Rickettsia sibirica 246	392021	Rickettsia rickettsii	Rickettsia rickettsii str. 'Sheila Smith'	98.18	97.88	0.990712154
Merge (supported by literature)	667	520456	Brucella neotomae	Brucella neotomae 5K33	204722	Brucella suis	Brucella suis 1330	99.75	99.71	0.998639572
Merge (supported by literature)	667	520456	Brucella neotomae	Brucella neotomae 5K33	444178	Brucella ovis	Brucella ovis ATCC 25840	99.58	99.6	0.997792326
Merge (supported by literature)	667	520456	Brucella neotomae	Brucella neotomae 5K33	483179	Brucella canis	Brucella canis ATCC 23365	99.73	99.69	0.998284046
Merge (supported by literature)	667	520456	Brucella neotomae	Brucella neotomae 5K33	224914	Brucella melitensis	Brucella melitensis bv. 1 str. 16M	99.66	99.62	0.997732276
Merge (supported by literature)	667	520456	Brucella neotomae	Brucella neotomae 5K33	568815	Brucella microti	Brucella microti CCM 4915	99.8	99.77	0.998870204
Merge (supported by literature)	667	204722	Brucella suis	Brucella suis 1330	444178	Brucella ovis	Brucella ovis ATCC 25840	99.63	99.6	0.99834215
Merge (supported by literature)	667	204722	Brucella suis	Brucella suis 1330	483179	Brucella canis	Brucella canis ATCC 23365	99.9	99.87	0.999077634
Merge (supported by literature)	667	204722	Brucella suis	Brucella suis 1330	224914	Brucella melitensis	Brucella melitensis bv. 1 str. 16M	99.68	99.62	0.998158548
Merge (supported by literature)	667	204722	Brucella suis	Brucella suis 1330	568815	Brucella microti	Brucella microti CCM 4915	99.8	99.77	0.999216598
Merge (supported by literature)	667	444178	Brucella ovis	Brucella ovis ATCC 25840	483179	Brucella canis	Brucella canis ATCC 23365	99.62	99.57	0.997922528
Merge (supported by literature)	667	444178	Brucella ovis	Brucella ovis ATCC 25840	224914	Brucella melitensis	Brucella melitensis bv. 1 str. 16M	99.56	99.51	0.997441788
Merge (supported by literature)	667	444178	Brucella ovis	Brucella ovis ATCC 25840	568815	Brucella microti	Brucella microti CCM 4915	99.67	99.65	0.998836836
Merge (supported by literature)	667	483179	Brucella canis	Brucella canis ATCC 23365	224914	Brucella melitensis	Brucella melitensis bv. 1 str. 16M	99.64	99.61	0.997777705
Merge (supported by literature)	667	483179	Brucella canis	Brucella canis ATCC 23365	568815	Brucella microti	Brucella microti CCM 4915	99.78	99.75	0.998828425
Merge (supported by literature)	667	224914	Brucella melitensis	Brucella melitensis bv. 1 str. 16M	568815	Brucella microti	Brucella microti CCM 4915	99.71	99.68	0.998279471

Supplementary Table 15

Summary of the number of pMGs and 16S rRNA gene copies found per genome.

Number of pMGs	Number of Genomes with this number of pMGs	Number of Genomes with at least this number of pMGs	Percentage of genomes with at least this number of pMGs
40	2860	2860	0.818077803
39	320	3180	0.909610984
38	118	3298	0.943363844
37	61	3359	0.960812357
36	28	3387	0.96882151
35	21	3408	0.974828375
34	18	3426	0.979977117
33	23	3449	0.986556064
32	14	3463	0.990560641
31	9	3472	0.993135011
30	10	3482	0.995995423
29	2	3484	0.996567506
27	2	3486	0.997139588
26	2	3488	0.99771167
24	1	3489	0.997997712
22	1	3490	0.998283753
21	1	3491	0.998569794
18	1	3492	0.998855835
17	1	3493	0.999141876
13	1	3494	0.999427918
11	1	3495	0.999713959
10	1	3496	1

Average Number of pMGs per genome	39.49141876
Genomes with full length 16S rRNA gene sequence	2899
Total full length 16S rRNA gene copies sequences found	9435
Average Number of 16S rRNA gene copies	3.254570542

Supplementary Table 17

Genome quality statistics and the number of found pMGs for all genomes from the clades used for the examples showing in Figure 2 and Supplementary Figures 5 and 6.

NCBI taxonomy name	NCBI Taxonomy ID	NCBI genome project ID	N50	Number of contigs	Number of pMGs found	Number of predicted genes	Comments
Prochlorococcus marinus str. MIT 9202	93058	54709	1691453	1	39	1927	
Prochlorococcus marinus subsp. pastoris str. CCMP1986	59919	57761	1657990	1	40	1760	
Prochlorococcus marinus str. MIT 9313	74547	57773	2410873	1	40	2324	
Prochlorococcus marinus subsp. marinus str. CCMP1375	167539	57995	1751080	1	40	1929	
Prochlorococcus marinus str. MIT 9303	59922	58305	2682675	1	40	3049	
Prochlorococcus marinus str. AS9601	146891	58307	1669886	1	40	1963	
Prochlorococcus marinus str. MIT 9211	93059	58309	1688963	1	40	1897	
Prochlorococcus marinus str. MIT 9515	167542	58313	1704176	1	40	1947	
Prochlorococcus marinus str. MIT 9312	74546	58357	1709204	1	40	1854	
Prochlorococcus marinus str. NATL2A	59920	58359	1842899	1	40	2205	
Prochlorococcus marinus str. NATL1A	167555	58423	1864731	1	40	2236	
Prochlorococcus marinus str. MIT 9301	167546	58437	1641879	1	40	1948	
Prochlorococcus marinus str. MIT 9215	93060	58819	1738790	1	40	2024	
Synechococcus sp. CC9311	64471	58123	2606748	1	40	2944	
Synechococcus sp. WH 7803	32051	61607	2366980	1	40	2586	
Serratia odorifera 4Rx13	682634	42253	457543	17	40	4741	
Serratia odorifera DSM 4582	667129	46991	170690	91	37	5194	
Serratia proteamaculans 568	399741	58725	5448853	1	40	5000	
Serratia sp. AS13	768493	60455	5442549	1	40	5060	
Serratia symbiotica str. Tucson	914128	62293	201972	389	0	2156	*
Serratia sp. AS9	768492	67313	5442880	1	40	5061	
Serratia sp. AS12	768490	67315	5443009	1	40	5061	
Serratia symbiotica str. 'Cinara cedri'	568817	82363	1762765	1	40	769	
Rahnella aquatilis CIP 78.65 = ATCC 33071	745277	60227	4861101	1	40	4442	
Rahnella sp. Y9602	741091	62715	4864217	1	40	4490	
Wigglesworthia glossinidia endosymbiont of Glossina brevipa	36870	57853	697724	1	40	651	
Wigglesworthia glossinidia endosymbiont of Glossina morsita	1142511	82191	719535	1	40	676	
Buchnera aphidicola str. LL01 (Acyrthosiphon pisum)	713603	43505	641799	1	39	612	
Buchnera aphidicola str. TLW03 (Acyrthosiphon pisum)	713602	43507	641770	1	40	608	
Buchnera aphidicola str. JF99 (Acyrthosiphon pisum)	713601	43509	641716	1	38	625	
Buchnera aphidicola str. JF98 (Acyrthosiphon pisum)	713600	43511	641771	1	32	512	
Buchnera aphidicola str. LSR1 (Acyrthosiphon pisum)	593275	55847	642011	1	40	629	
Buchnera aphidicola str. APS (Acyrthosiphon pisum)	107806	57805	640681	1	40	600	
Buchnera aphidicola str. Bp (Baizongia pistaciae)	224915	57827	615980	1	40	540	
Buchnera aphidicola str. Sg (Schizaphis graminum)	198804	57913	641454	1	40	582	
Buchnera aphidicola BCc	372461	58579	416380	1	40	393	
Buchnera aphidicola str. Tuc7 (Acyrthosiphon pisum)	561501	59283	641895	1	40	588	
Buchnera aphidicola str. 5A (Acyrthosiphon pisum)	563178	59285	642122	1	40	590	
Buchnera aphidicola str. Ua (Uroleucon ambrosiae)	1005057	65479	615380	1	40	564	
Buchnera aphidicola str. Ak (Acyrthosiphon kondoi)	1005090	65481	641794	1	40	594	
Buchnera aphidicola (Cinara tujafilina)	261317	68101	444925	1	40	394	

* = Was exluded before the number of marker genes was computed as it failed the initial genome quality control

Supplementary Table 19

Clusters of size two or more containing only not fully identified genomes and a list of the genomes falling into these clusters.

Cluster	NCBI Taxonomy ID	NCBI Taxonomy ID NCBI Project ID	NCBI Taxonomy Name of sequenced organi
1	395962	395962.5914	Cyanothece sp. PCC 8802
1	41431	41431.59027	Cyanothece sp. PCC 8801
2	1148	1148.57659	Synechocystis sp. PCC 6803
2	1148	1148.67081	Synechocystis sp. PCC 6803
3	60480	60480.58345	Shewanella sp. MR-4
3	60481	60481.58343	Shewanella sp. MR-7
3	94122	94122.58347	Shewanella sp. ANA-3
4	675810	675810.4111	Vibrio sp. RC341
4	675815	675815.4162	Vibrio sp. RC586
5	314267	314267.5426	Sulfitobacter sp. NAS-14.1
5	52598	52598.54191	Sulfitobacter sp. EE-36
6	439495	439495.5471	Pseudovibrio sp. JE062
6	911045	911045.8237	Pseudovibrio sp. FO-BEG1
7	164756	164756.5847	Mycobacterium sp. MCS
7	164757	164757.5849	Mycobacterium sp. JLS
7	189918	189918.5849	Mycobacterium sp. KMS
8	355249	355249.6692	Streptomyces sp. Tu6071
8	465543	465543.4842	Streptomyces sp. SPB74
8	591157	591157.5582	Streptomyces sp. SPB78
8	683219	683219.619	Streptomyces sp. SA3_actG
9	1000568	1000568.678	Megasphaera sp. UPII 199-6
9	699218	699218.4657	Megasphaera genomosp. type_1
10	481743	481743.4113	Paenibacillus sp. Y412MC10
10	908341	908341.6615	Paenibacillus sp. HGF5
11	457403	457403.682	Fusobacterium sp. 11_3_2
11	457405	457405.5561	Fusobacterium sp. 7_1
11	469601	469601.6819	Fusobacterium sp. 21_1A
11	469603	469603.4085	Fusobacterium sp. 3_1_33
11	556264	556264.5563	Fusobacterium sp. D11
12	469602	469602.4779	Fusobacterium sp. 3_1_27
12	469604	469604.56	Fusobacterium sp. 3_1_36A2
12	469606	469606.5561	Fusobacterium sp. 4_1_13
13	1029718	1029718.714	Candidatus Arthromitus sp. SFB-mouse
13	1041809	1041809.678	Candidatus Arthromitus sp. SFB-mouse
14	245018	245018.4596	butyrate-producing bacterium SSC/2
14	411484	411484.5455	Clostridium sp. SS2/1
14	658089	658089.6188	Lachnospiraceae bacterium 5_1_63FAA
15	552395	552395.6358	Lachnospiraceae bacterium 4_1_37FAA
15	658088	658088.6642	Lachnospiraceae bacterium 9_1_43BFAA

Supplementary Table 20

The 130 genomes used for the benchmarking shown in Figure 1b.

NCBI Taxonomy ID	Genome Project ID	NCBI Taxonomy Name
1009464	51067	Gardnerella vaginalis HMP9231
1010838	66075	Mycobacterium avium subsp. paratuberculosis S397
1027843	66485	Chlamydophila psittaci 08DC60
1027844	66483	Chlamydophila psittaci C19/98
1027845	66481	Chlamydophila psittaci 02DC15
1027846	66479	Chlamydophila psittaci 01DC11
1029822	66655	Lactobacillus salivarius NIAS840
1031710	66839	Oligotropha carboxidovorans OM4
1035817	66401	Bifidobacterium longum subsp. longum KACC 91563
1036743	67413	Rhodospirillum rubrum F11
1037409	67463	Bradyrhizobium japonicum USDA 6
1041521	67679	Lactobacillus salivarius GJ-24
1042403	67865	Bifidobacterium animalis subsp. lactis CNCM I-2494
1045856	80739	Enterobacter cloacae EcWSU1
1048834	68273	Alicyclobacillus acidocaldarius subsp. acidocaldarius Tc-4-1
1049565	67293	Klebsiella pneumoniae KCTC 2242
1051072	68327	Streptococcus equi subsp. zooepidemicus ATCC 35246
1075106	71815	Bifidobacterium animalis subsp. lactis BLC1
1107880	75083	Leuconostoc mesenteroides subsp. mesenteroides J18
1112856	85495	Mycoplasma pneumoniae 309
1125630	78789	Klebsiella pneumoniae subsp. pneumoniae HS11286
349123	54165	Lactobacillus reuteri 100-23
351581	54341	Francisella tularensis subsp. holarctica FSC200
353496	16120	Lactobacillus delbrueckii subsp. bulgaricus 2038
406984	17947	Chlamydophila pneumoniae LPCoLN
436113	19245	Mycoplasma mycoides subsp. capri str. GM12
436113	39245	Mycoplasma mycoides subsp. capri str. GM12
445334	54845	Clostridium perfringens C str. JGS1495
445983	54817	Borrelia burgdorferi 156a
445984	54819	Borrelia burgdorferi Bol26
451754	54847	Clostridium perfringens B str. ATCC 3626
451755	54849	Clostridium perfringens E str. JGS1987
451757	54853	Clostridium perfringens NCTC 8239
476210	54935	Borrelia burgdorferi 118a
476211	54937	Borrelia burgdorferi 72a
476212	54933	Borrelia burgdorferi 94a
483214	54983	Methanobrevibacter smithii DSM 2375
486409	55085	Bifidobacterium animalis subsp. lactis HN019
488537	55051	Clostridium perfringens D str. JGS1721
492476	55013	Clostridium thermocellum JW20
497978	28333	Acinetobacter baumannii MDR-ZJ06
498736	55057	Borrelia burgdorferi 80a
498737	55061	Borrelia burgdorferi WI91-23
498738	55055	Borrelia burgdorferi 29805
498739	55063	Borrelia burgdorferi CA-11.2A
498740	55067	Borrelia burgdorferi 64b
510831	38289	Francisella tularensis subsp. tularensis NE061598
515609	54873	Ureaplasma parvum serovar 14 str. ATCC 33697
515611	54877	Ureaplasma urealyticum serovar 13 str. ATCC 33698
518603	54749	Ureaplasma urealyticum serovar 5 str. ATCC 27817
519851	54747	Ureaplasma parvum serovar 6 str. ATCC 27818
521002	55123	Methanobrevibacter smithii DSM 2374
521007	29357	Borrelia burgdorferi N40
521008	29359	Borrelia burgdorferi JD1
525262	40867	Corynebacterium jeikeium ATCC 43734
525280	55483	Erysipelothrix rhusiopathiae ATCC 19414
525283	49043	Fusobacterium nucleatum subsp. nucleatum ATCC 23726
525325	55497	Lactobacillus fermentum ATCC 14931
525330	55505	Lactobacillus johnsonii ATCC 33200
525338	55521	Lactobacillus plantarum subsp. plantarum ATCC 14917
525341	55541	Lactobacillus reuteri CF48-3A
525369	55525	Proteus mirabilis ATCC 29906
526347	54751	Ureaplasma urealyticum serovar 11 str. ATCC 33695
543302	55305	Alicyclobacillus acidocaldarius LAA1
548480	55465	Bifidobacterium longum subsp. longum ATCC 55813
548485	55517	Lactobacillus reuteri MM4-1A
550694	38067	Escherichia fergusonii B253
550747	54753	Ureaplasma urealyticum serovar 12 str. ATCC 33696
550748	54647	Ureaplasma urealyticum serovar 4 str. ATCC 27816
550773	54651	Ureaplasma urealyticum serovar 9 str. ATCC 33175
553201	55453	Rothia mucilaginosa ATCC 25296
553481	55387	Mycobacterium avium subsp. avium ATCC 25291
555217	30987	Zymomonas mobilis subsp. mobilis ATCC 10988
557601	55405	Acinetobacter baumannii AB900
572545	55637	Clostridium thermocellum DSM 2360
573059	32603	Desulfovibrio vulgaris RCH1
573236	32515	Bifidobacterium animalis subsp. lactis V9
585198	51715	Mobiluncus curtisii subsp. curtisii ATCC 35241
585517	55885	Lactobacillus reuteri MM2-3
585520	55883	Lactobacillus helveticus DSM 20075
586220	55887	Leuconostoc mesenteroides subsp. cremoris ATCC 19254
592014	47501	Acinetobacter baumannii 6013113
637389	55945	Acidithiobacillus caldus ATCC 51756
637887	41553	Clostridium thermocellum DSM 1313
644651	37877	Erwinia pyrifoliae DSM 12163
655813	47255	Streptococcus oralis ATCC 35037
657317	39161	Eubacterium rectale M104/1
657318	39159	Eubacterium rectale DSM 17629
663917	39745	Geobacter sulfurreducens KN400
667127	41361	Klebsiella pneumoniae subsp. rhinoscleromatis ATCC 13884
679196	42955	Lactobacillus gasseri 224-1
679936	40777	Sulfobacillus acidophilus DSM 10332
682147	49423	Gardnerella vaginalis AMD
682148	49673	Gardnerella vaginalis 5-1
686660	42953	Veillonella parvula ATCC 17745
696749	42153	Acinetobacter baumannii 1656-2
712961	43535	Lactobacillus salivarius CECT 5713
718254	45967	Enterobacter cloacae subsp. cloacae NCTC 9394
722911	45963	Bifidobacterium longum subsp. longum F8
762633	51619	Thermus thermophilus SG0.5JP17-16
767462	49153	Lactobacillus helveticus H10
768065	49335	Haloquadratum walsbyi C23
768728	50761	Lactobacillus salivarius ACS-116-V-Col5a
862962	84217	Bacteroides fragilis 638R
863638	50455	Clostridium acetobutylicum EA 2018
869211	50823	Spirochaeta thermophila DSM 6578
873517	61279	Capnocytophaga ochracea F0287
876138	46683	Streptococcus agalactiae FSL S3-026
879307	52049	Gardnerella vaginalis 315-A
886886	52391	Acinetobacter baumannii AB210
887326	61495	Mobiluncus curtisii ATCC 51333
887899	61497	Mobiluncus curtisii subsp. holmesii ATCC 35242
888027	53001	Lactobacillus delbrueckii subsp. lactis DSM 20072
888048	62541	Streptococcus parasanguinis ATCC 903
888745	68679	Streptococcus agalactiae ATCC 13813
888828	63561	Haemophilus parainfluenzae ATCC 33392
905067	60563	Streptococcus parasanguinis F0405
907287	53881	Mycoplasma hyopneumoniae 168
909954	56053	Lactobacillus johnsonii DPC 6026
910312	61039	Porphyromonas asaccharolytica PR426713P-I
930944	60483	Yersinia enterocolitica subsp. palearctica Y11
980514	62279	Acinetobacter baumannii TCDC-AB0715
983328	62521	Campylobacter fetus subsp. venerealis NCTC 10354
984894	66151	Chlamydophila psittaci Cal10
990315	63187	Xanthomonas campestris pv. raphani 756C
992402	63335	Acinetobacter baumannii ABNIH1
992403	63337	Acinetobacter baumannii ABNIH2
992404	63339	Acinetobacter baumannii ABNIH3
992405	63341	Acinetobacter baumannii ABNIH4
996285	63543	Pseudomonas stutzeri DSM 4166

Supplementary Note: Speciation of *Escherichia/Shigella*

Escherichia/Shigella complex is known to be polyphyletic [19,38] with *Shigella* genomes branching out in different regions of the phylogenetic tree of the *Escherichia/Shigella* complex (also see Supplementary Fig. 7) that forms a large cluster in the specI species clustering excluding *Escherichia albertii*, but including *Escherichia fergusoni*. *Escherichia coli* is one of the best studied organisms and its speciation has been studied thoroughly [39]. It seems clear that *Escherichia coli* has a conserved, ancient backbone [39]. This backbone was also recovered as the core genome in a recent analysis of the *Escherichia coli* pangenome [40]. In addition, to the core genome Escherichia coli seems to have a rather large dispensable genome [40], supporting the notion that Escherichia coli seems to have an extensive mosaic structure [39]. More recently, there have also been claims that the *Escherichia*/*Shigella* complex is currently undergoing speciation into different environmental and host associated subgroups due to evolutionary events in the dispensable genome [41]. The same study detected a genetic continuum of the core genome within the *Escherichia* genus (more specifically between *E. coli* and *E. albertii*). In conclusion, our results coincide with the core genome hypothesis [42] as our clustering of this clade seems to coincide with the presence of a conserved core genome of the *Escherichia*/*Shigella* complex. The dispensable genome is not considered for species delineation by specI and if the dispensable genome has a strong impact on speciation, this might represent a limitation of the method. Analyses of the pan-genome of different species might help to explain the effect of the dispensable genome on speciation in the future.

38. Sims, G. & Kim, S.-H. Proceedings of the National Academy of Sciences of the United States of America 108, 8329–8334 (2011).

39. Welch, R. et al. Proceedings of the National Academy of Sciences of the United States of America 99, 17020–17024 (2002).

40. Rasko, D. et al. Journal of bacteriology 190, 6881–6893 (2008).

41. Luo, C. et al. Proceedings of the National Academy of Sciences of the United States of America 108, 7200–7205 (2011).

42. Riley, M. & Lizotte-Waniewski, M. Methods in molecular biology (Clifton, N.J.) 532, 367–377 (2009).