Impact of recombination on the base composition of Bacteria and Archaea

Supplementary Material

Louis-Marie Bobay and Howard Ochman

Table of content:

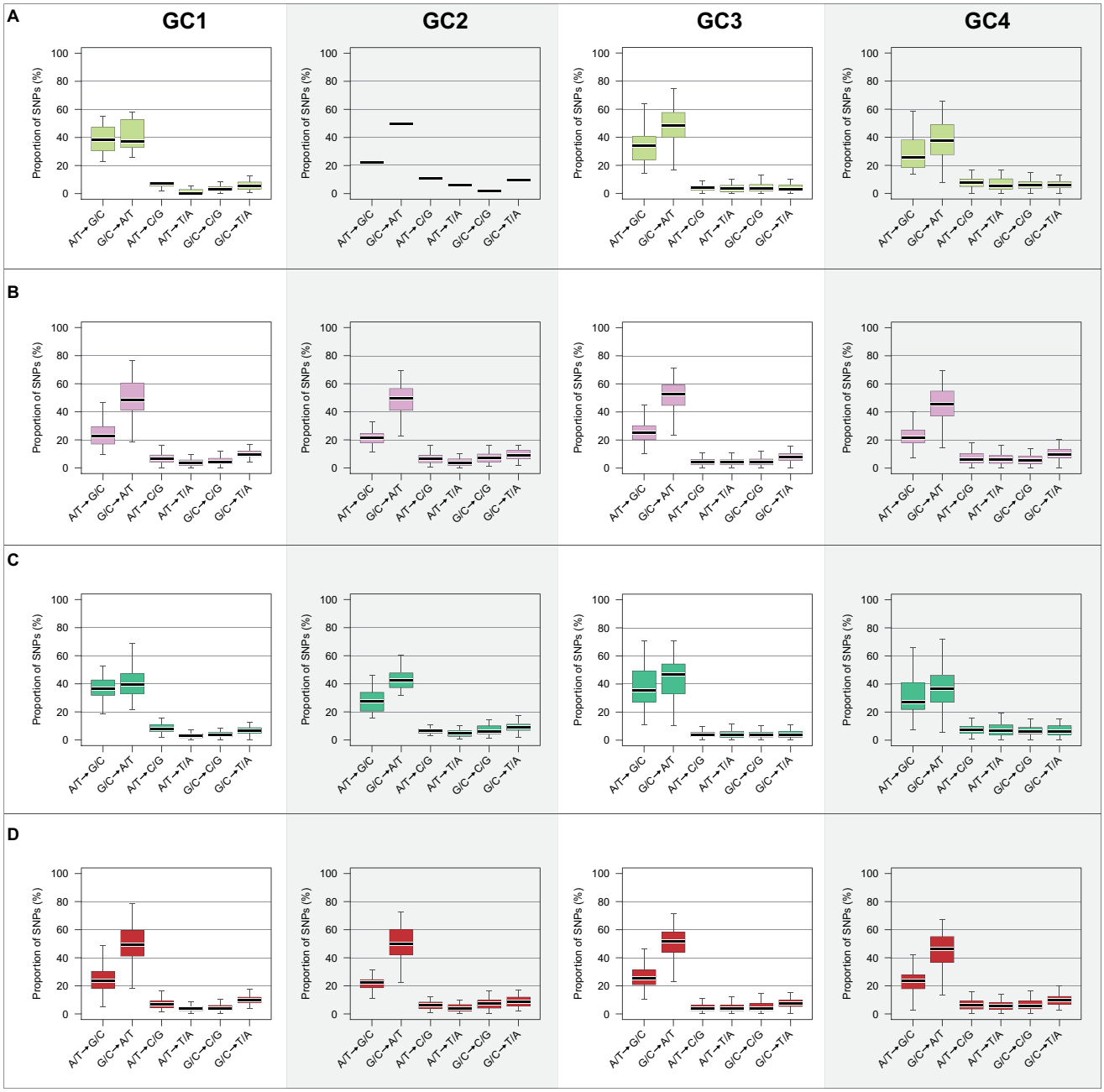

Figure S1. Comparison of nucleotide changes introduced by recombination and mutation at each codon position. Cumulative proportions of SNPs at first codon positions (GC1), second codon positions (GC2), third codon positions (GC3), and four-fold degenerate sites ($GC4_{fold}$) for each of the six types of nucleotide changes, as calculated for: (**A**) all alleles introduced by recombination (dark green); (**B**) new alleles introduced recombination (pale green); (**C**) all alleles introduced by mutations (dark red); (**D**) new alleles introduced by mutations (pale red). Values were normalized by the GC-contents at each codon position for each species prior to calculating overall proportions, and species with fewer than 50 polymorphic sites for a given category of alleles were excluded.

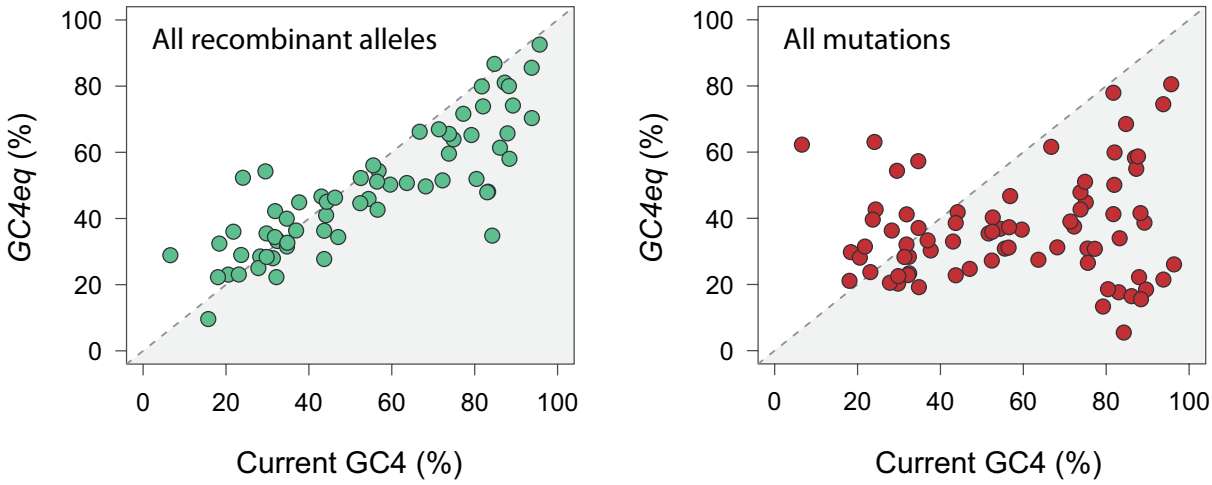

Figure S2. Equilibrium GC content relative to actual GC content. *GC4eq* is the expected GC-content at four-fold degenerate sites for a given species when based on all alleles introduced by recombination (dark green, left panel) and all alleles introduced by mutations (dark red, right panel). *GC4eq* values were normalized by the GC-contents at four-fold degenerate sites for each species, and species with fewer than 50 polymorphic sites for any given category of allele were excluded. Points in shaded area below the diagonal denote species that are GC-rich relative to the input of polymorphisms by recombination or mutation.

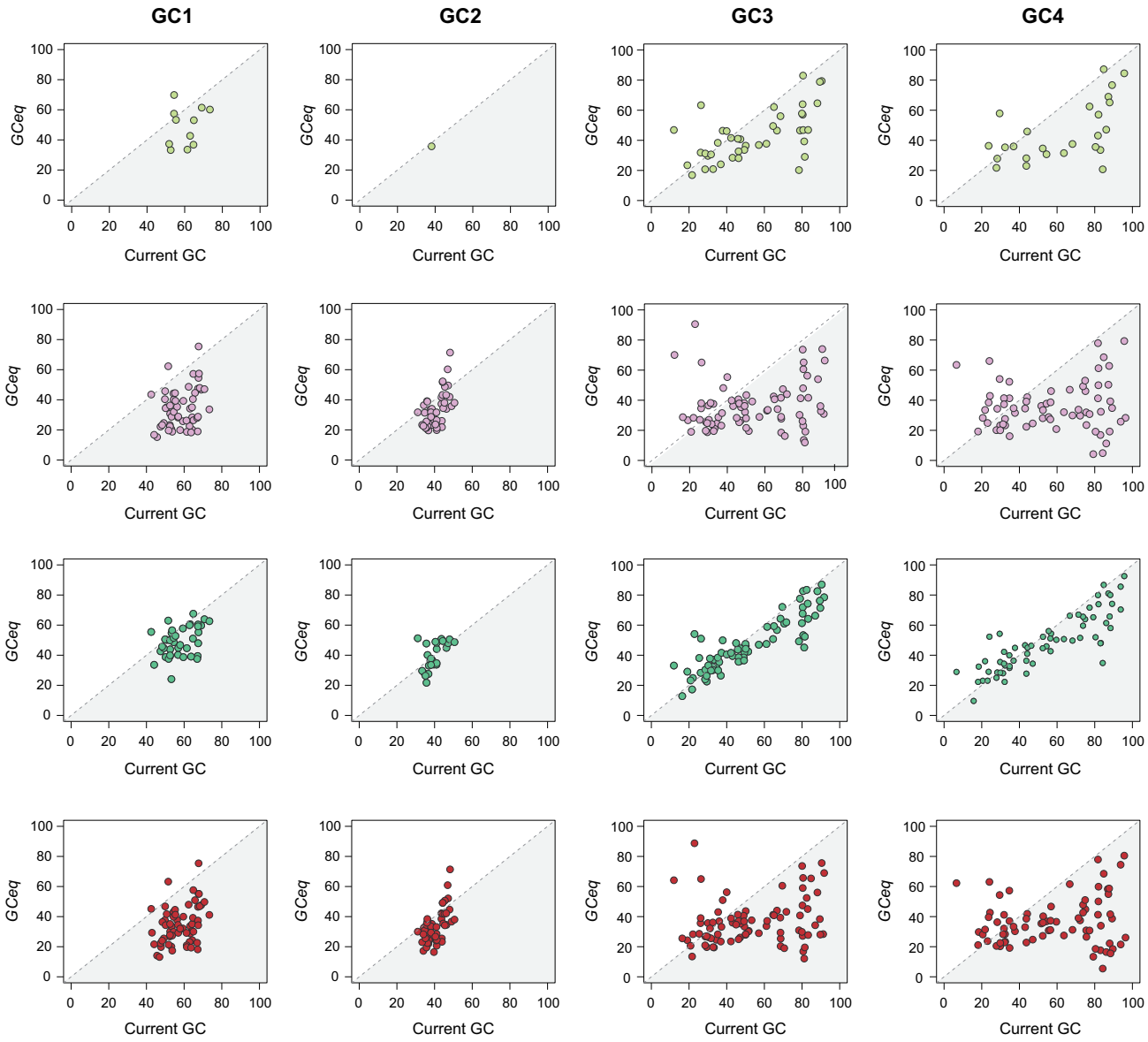

Figure S3. Equilibrium GC content inferred from new polymorphisms relative to actual GC content at each codon position. *GCeq* is the expected GC-content at first codon positions (GC1), second codon positions (GC2), third codon positions (GC3), and four-fold degenerate sites ($GC4_{fold}$) for a given species when based on new alleles introduced by recombination (pale green), new alleles introduced by mutations (pale red), all alleles introduced by recombination (dark green) and all alleles introduced by mutations (dark red). *GCeq* values were normalized by the GC-contents at each codon position for each species, and species with fewer than 50 polymorphic sites for any given category of allele were excluded. Points in shaded area below the diagonal denote species that are GC-rich relative to the input of polymorphisms by recombination or mutation.

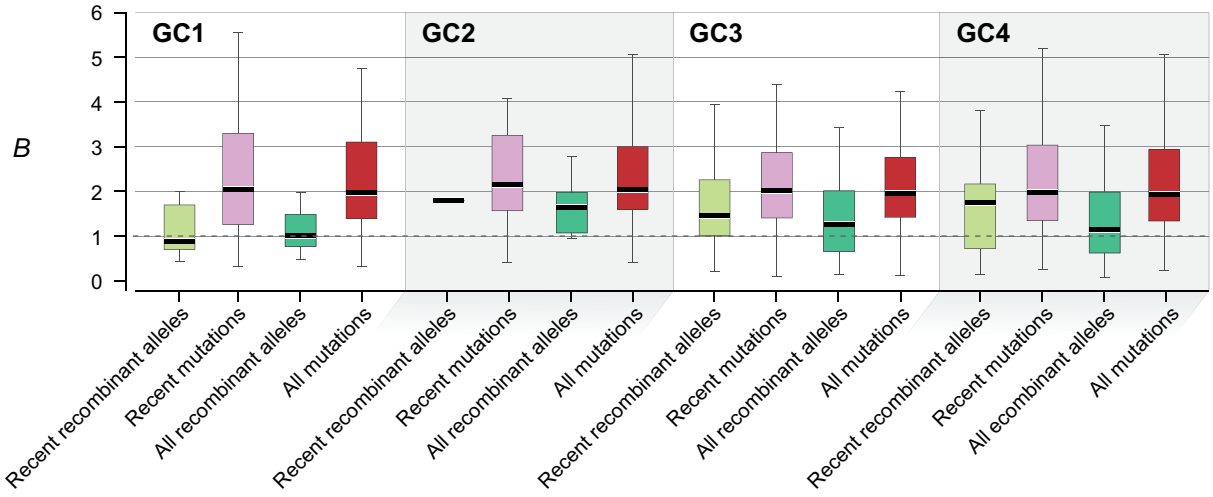

Figure S4. Impact of recombination and mutations on genomic nucleotide composition and codon usage. The metric *B* represents the number of changes from G or C to A or T relative to the number of changes from A or T to G or C at four-fold degenerate sites. $B > 1$ indicates an enrichment towards A and T, and $B < 1$ indicates an enrichment toward G and C. Values shown are for all alleles introduced by recombination (dark green); new alleles introduced recombination (pale green); all alleles introduced by mutations (dark red); new alleles introduced by mutations (pale red). Values were normalized by the GC-contents at the corresponding codon postion for each species prior to calculating overall proportions, and species with fewer than 50 polymorphic sites for a given category of alleles were excluded.

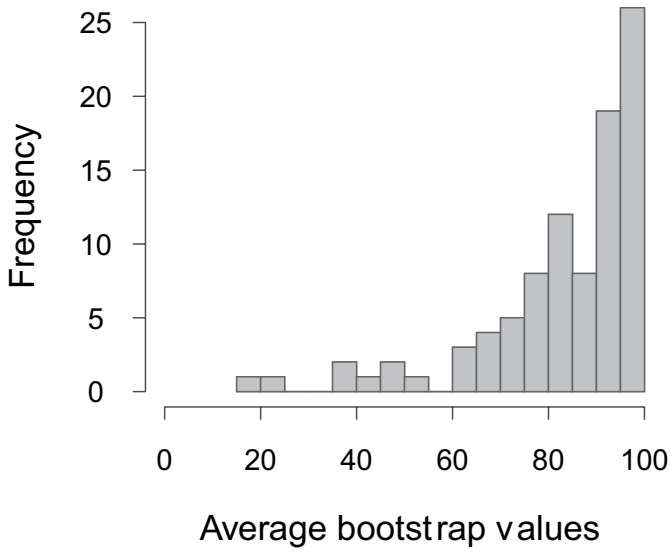

Figure S5. Distribution of average bootstrap values for the species phylogenies. For each species tree, 100 bootstrap replicates were computed, averaged across all nodes to produce an average bootstrap value for a species. The list of average bootstrap values of each species is given in Table S1.

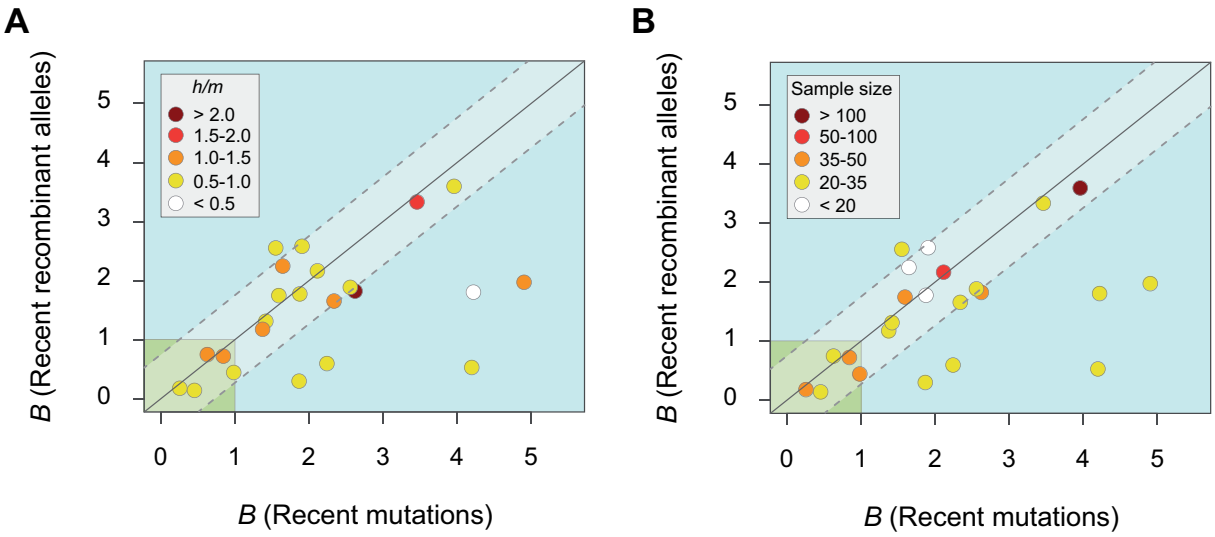

Figure S6. Nucleotide bias of recent recombinant alleles and recent mutations. The metric *B* represents the number of changes from G or C to A or T relative to the number of changes from A or T to G or C at four-fold degenerate sites. On each axis, a value of $B < 1$ indicates an enrichment toward G and C (green area), and value of $B > 1$ indicates an enrichment towards A and T (blue area). The solid diagonal line indicates identical nucleotide bias for recent recombinant alleles and recent mutations; the dotted lines represent half of the standard deviation. Values were normalized by the GC-contents at four-fold degenerate sites for each species prior to calculating overall proportions, and species with fewer than 50 polymorphic sites for a given category of alleles were excluded. (**A**) The color of dots denotes the recombination rate *h/m* for a species, as indicated in the key. (**B**) The color of dots denotes the number of strains for a given species, as indicated in the key.

Table S1. Description of species and core genome information.

Species	Genomes	Distinct strains	Core genes	*h/m* (average on core genes)	GC content (%) (average on core genes and standard deviation)	p-value	Spearman's Rho	Average Bootstrap support	Polymorphic sites (%)
Acinetobacter baumannii	1046	276	49	1,26	42.3 (2.98)	1,22E-01	-0,22	54,30	40,38
Acinetobacter pittii	20	15	2360	0,74	40.35 (3.11)	7,13E-08	0,11	97,60	13,75
Aggregatibacter actinomycetemcomitans	21	20	282	0,16	48.01 (3.67)	8,33E-01	0,01	97,10	8,63
Bacillus anthracis	95	22	2037	0,12	36.56 (3.02)	2,46E-01	-0,04	83,70	0,13
Bacillus cereus	168	130	970	1,18	37.26 (3.1)	2,33E-03	0,10	97,20	31,95
Bacillus subtilis	79	54	224	0,73	44.8 (3.05)	2,97E-03	0,20	90,60	25,18
Bacillus thuringiensis	56	36	1857	0,99	36.97 (3.07)	2,73E-01	-0,03	97,00	16,81
Bacteroides fragilis	93	36	597	0,25	45.21 (3.12)	3,38E-01	0,04	86,90	13,31
Bifidobacterium longum	44	28	734	1,65	61.72 (2.57)	2,19E-01	-0,05	71,00	7,91
Bordetella bronchiseptica	64	41	2340	1,30	68.58 (3.06)	2,93E-01	0,02	76,40	2,86
Borrelia burgdorferi	30	25	79	0,90	30.54 (4.57)	4,69E-01	0,09	83,10	1,94
Borrelia garinii	22	20	31	0,50	31.96 (3.53)	1,73E-01	0,26	84,20	2,97
Brucella militensis	64	34	950	0,33	58.49 (3.27)	2,71E-01	0,04	90,80	0,55
Brucella abortus	150	31	1498	0,38	58.63 (3.2)	7,45E-01	0,01	96,60	0,54
Brucella suis	51	16	2124	0,37	58.79 (2.93)	5,55E-02	-0,05	99,60	0,44
Burkholderia pseudomallei	189	120	172	1,26	67.61 (2.95)	6,95E-01	-0,03	37,10	2,50
Campylobacter jejuni	142	117	174	1,28	32.47 (3.51)	4,52E-02	0,15	88,50	11,27
Campylobacter coli	72	70	270	1,88	34.3 (2.91)	7,32E-01	0,02	84,10	13,50
Chlamydia trachomatis	92	56	752	2,76	42.03 (2.2)	1,19E-01	0,06	91,70	1,52
Chlamydia psittaci	47	19	127	0,30	40.1 (2.33)	7,01E-01	-0,03	92,90	35,32
Clostridium botulinum	53	37	46	0,92	31.32 (3.3)	6,24E-01	-0,08	78,00	8,73
Corynebacterium pseudotuberculosis	22	18	1260	0,73	53.3 (3.05)	7,37E-01	0,01	93,30	1,58
Coxiella burnetii	24	16	536	0,08	44.64 (4.1)	3,11E-01	-0,05	92,00	0,76
Enterobacter cloacae	199	99	426	1,19	55.45 (3.03)	9,64E-13	0,34	79,90	32,74
Enterobacter aerogenes	28	21	2081	0,83	57.15 (3.34)	9,20E-01	0,00	94,50	6,10
Enterococcus faecalis	362	158	158	1,16	39.56 (2.58)	4,52E-01	0,06	70,30	4,58
Enterococcus faecium	274	165	238	1,86	40.83 (2.18)	4,81E-01	0,05	61,80	7,30
Escherichia coli	2961	63	656	0,68	52.49 (2.82)	4,04E-01	0,03	96,20	31,44
Francisella tularensis	67	29	489	0,22	33.87 (3.3)	7,81E-01	0,01	96,80	9,80
Fusobacterium nucleatum	29	26	708	0,88	28.36 (3.3)	1,37E-01	-0,06	84,20	19,68
Gallibacterium anatis	22	21	1456	2,32	41.79 (3.08)	6,62E-03	-0,07	90,80	9,47
Gardnerella vaginalis	36	32	91	0,89	44.49 (3.81)	7,32E-04	0,35	89,60	35,42
Haemophilus influenzae	43	38	526	2,39	39.37 (3.02)	8,12E-01	-0,01	88,00	10,51
Helicobacter pylori	445	347	50	2,47	40.69 (2.55)	3,83E-01	0,13	46,30	36,81
Kingella kingae	42	42	914	1,53	48.48 (3.65)	4,22E-01	-0,03	96,50	3,29
Klebsiella pneumoniae	526	178	344	0,76	57.85 (3.68)	1,25E-01	0,08	66,60	12,65
Lactobacillus rhamnosus	31	21	286	1,13	47.95 (3.18)	2,94E-01	0,06	90,40	5,73
Lactobacillus casei	27	20	869	1,33	48.16 (2.71)	3,38E-01	-0,03	99,30	3,07
Lactobacillus paracasei	43	38	133	0,86	47.28 (2.81)	5,91E-01	-0,05	82,30	3,35

Lactobacillus plantarum	28	25	702	0,29	46.71 (3.31)	1,38E-04	-0,14	67,00	25,30
Lactococcus lactis	39	35	382	0,68	37.93 (3.32)	4,87E-01	0,04	90,80	13,67
Legionella pneumophila	74	44	510	0,66	39.54 (2.91)	2,38E-01	0,05	95,70	13,36
Leptospira interrogans	198	55	217	0,81	38.28 (3.53)	7,85E-01	0,02	63,00	3,99
Leptospira kirschneri	23	17	1865	1,67	38.33 (3.24)	2,66E-01	-0,03	90,90	3,33
Leptospira santarosai	23	22	1899	1,17	43.6 (2.99)	1,93E-01	0,03	90,30	2,12
Listeria monocytogenes	331	63	427	0,75	39.31 (2.77)	9,07E-01	0,01	75,80	11,47
Methanosarcina mazei	63	30	1708	3,61	45.93 (3.44)	3,90E-01	0,02	94,40	2,42
Mycobacterium tuberculosis	1817	143	38	0,19	63.04 (1.74)	5,04E-02	-0,32	18,10	1,15
Mycobacterium abscessus	79	24	1898	2,24	65.01 (2.56)	4,54E-01	-0,02	88,80	5,13
Mycobacterium africanum	28	20	3336	0,18	65.85 (3.09)	8,45E-01	0,00	100,00	0,12
Mycobacterium avium	58	34	186	1,34	68.17 (2.58)	6,20E-01	0,04	93,00	1,93
Mycobacterium bovis	39	19	2085	0,33	65.31 (3.04)	2,69E-01	0,03	92,10	0,16
Neisseria meningitidis	200	153	318	4,28	54.56 (4.82)	3,88E-01	0,05	85,00	10,46
Neisseria gonorrhoeae	23	19	1163	2,08	54.94 (4.98)	1,92E-02	0,08	87,80	0,95
Oenococcus oeni	58	36	484	1,16	39.85 (3.62)	6,81E-01	-0,02	92,20	2,02
Pasteurella multocida	27	22	182	0,76	41.8 (2.86)	5,01E-01	0,05	80,40	3,33
Pectobacterium carotovorum	35	34	1852	1,08	53.84 (3.53)	1,62E-03	0,07	97,20	23,04
Peptoclostridium difficile	254	81	192	0,71	30.13 (2.9)	5,02E-01	-0,05	71,90	5,05
Prochlorococcus marinus	31	26	289	0,89	36.77 (2.77)	1,97E-02	0,14	89,10	47,02
Propionibacterium acnes	92	78	222	0,82	60.25 (2.88)	9,41E-01	0,01	65,20	4,66
Pseudomonas aeruginosa	725	310	104	0,83	63.83 (3.61)	3,07E-01	0,10	43,60	7,79
Pseudomonas fluorescens	57	54	1492	2,37	61.48 (2.98)	7,25E-03	0,07	94,80	42,69
Pseudomonas putida	35	33	526	1,21	62.15 (3.07)	9,03E-01	-0,01	96,90	38,61
Pseudomonas stuzeri	22	22	1620	1,32	64.36 (3.16)	7,56E-21	-0,23	98,10	42,03
Pseudomonas syringae	98	57	394	1,07	58.38 (3.14)	8,44E-07	0,25	92,40	33,70
Ralstonia solanacearum	29	24	1152	0,24	67.26 (3.16)	1,27E-02	-0,07	100,00	12,37
Rhizobium leguminosarum	30	30	2497	0,75	62.87 (2.56)	6,17E-01	0,01	97,90	33,52
Rhodococcus fascians	21	20	2136	0,39	65.04 (2.16)	3,75E-02	-0,05	99,30	32,25
Salinispora arenicola	47	41	1171	0,55	68.99 (2.62)	5,29E-02	-0,06	79,10	4,97
Salinispora pacifica	38	32	2667	0,51	70.42 (3.04)	7,10E-09	-0,11	99,30	18,47
Salmonella enterica	1139	64	279	0,69	53.75 (3.56)	6,20E-02	0,11	63,80	13,37
Serratia marcescens	36	23	1853	0,72	61.04 (3.99)	9,75E-01	0,00	98,10	12,39
Sinorhizobium meliloti	30	28	2294	1,38	63.2 (2.55)	3,45E-01	-0,02	70,70	2,83
Staphylococcus aureus	4221	370	30	0,34	34.23 (3.27)	5,09E-01	-0,13	24,90	6,29
Staphylococcus epidermidis	104	72	691	1,31	33.78 (3.01)	8,13E-01	-0,01	82,20	5,14
Stenotrophomonas maltophilia	29	27	119	0,85	66.28 (2.75)	4,61E-01	0,07	83,10	31,72
Streptococcus pneumoniae	327	237	118	1,48	42.62 (3.02)	2,14E-01	0,12	72,50	8,29
Streptococcus agalactiae	316	136	79	1,14	36.18 (3.15)	3,03E-01	-0,12	45,80	2,91
Streptococcus	41	39	362	1,28	43.18 (2.79)	2,43E-05	0,22	98,10	37,44

mitis									
Streptococcus mutans	166	145	861	0,99	38.17 (3.17)	3,72E-01	-0,03	89,20	5,35
Streptococcus pyogenes	234	54	264	0,93	39.78 (3.19)	4,72E-02	0,12	81,80	4,19
Streptococcus sanguinis	24	21	791	0,93	46.24 (3.02)	4,26E-08	0,19	92,80	25,08
Streptococcus sobrinus	48	39	87	3,92	44.53 (3.8)	9,86E-02	0,21	82,00	3,27
Streptococcus suis	470	231	368	1,36	43.05 (3.91)	4,32E-12	0,35	78,30	34,95
Sulfolobus islandicus	20	18	1316	0,71	35.39 (3.06)	7,91E-01	-0,01	98,70	1,80
Vibrio cholerae	278	66	114	2,16	48.18 (3.09)	3,39E-01	-0,09	66,90	9,70
Vibrio cyclitrophicus	21	19	3354	2,12	45.03 (2.41)	9,55E-02	0,03	79,90	3,10
Vibrio parahaemolyticus	76	75	63	0,72	46.43 (2.32)	6,23E-01	0,07	35,80	4,16
Vibrio vulnificus	31	23	95	1,83	48.21 (3.0)	6,23E-02	-0,19	75,50	8,67
Xanthomonas axonopodis	90	29	1046	0,56	64.96 (2.9)	4,00E-01	0,03	96,80	11,03
Xanthomonas campestris	22	16	1355	0,34	65.88 (2.97)	1,23E-13	-0,20	100,00	23,43
Xanthomonas citri	51	21	1867	0,58	65.51 (3.07)	4,15E-01	0,02	99,30	1,47
Yersinia pseudotuberculosis	23	16	2888	0,89	49.04 (4.34)	1,02E-01	0,03	99,60	0,91

We estimated the recombination rate *h/m* as the ratio of homoplasic (*h*) to non-homoplasic (*m*) alleles inferred along the core genome of each species. Homoplasies were defined with the distance-based method (see Materials and Methods). Spearman coefficients Rho and *p*-values were estimated by correlating the recombination rate *h/m* relative to the GC content of the core genes for each species. Species in bold correspond to the species displayed in Figures 6 and S6. Average bootstrap supports are defined as the average of bootstrap values estimated over all the nodes of each species tree, which were built with a maximum likelihood approach on the concatenate of core genes (see Materials and Methods).

Table S2. Identifying recombinant alleles in simulated data sets.

N=100, *rec*=10, *m*=1		
Generations	Inferred	Recombinant
G2000	10	10
G5000	**21**	**20**
G10000	14	14
G12000	2	2
G15000	22	22
G18000	14	14
G20000	21	21
G22000	8	8
G25000	68	68
G30000	**17**	**16**
N=100, *rec*=10, *m*=10		
Generations	Inferred	Recombinant
G2000	317	317
G5000	96	96
G10000	306	306
G12000	626	626
G15000	82	82
G18000	358	358
G20000	651	651
G22000	197	197
G25000	223	223
G30000	67	67
N=500, *rec*=10, *m*=1		
Generations	Inferred	Recombinant
G2000	54	54
G5000	26	26
G10000	163	163
G12000	53	53
G15000	67	67
G18000	207	207
G20000	146	146
G22000	105	105
G25000	58	58
G30000	424	424
N=500, *rec*=10, *m*=10		
Generations	Inferred	Recombinant
G2000	786	786
G5000	4151	4151
G10000	880	880
G12000	2517	2517
G15000	931	931
G18000	1224	1224
G20000	1313	1313
G22000	846	846
G25000	1791	1791
G30000	1214	1214

N=100, *rec*=50, *m*=1		
Generations	Inferred	Recombinant
G2000	18	18
G5000	36	36
G10000	89	89
G12000	81	81
G15000	139	139
G18000	12	12
G20000	56	56
G22000	40	40
G25000	10	10
G30000	49	49
N=100, *rec*=50, *m*=10		
Generations	Inferred	Recombinant
G2000	433	433
G5000	575	575
G10000	681	681
G12000	1616	1616
G15000	741	741
G18000	755	755
G20000	428	428
G22000	355	355
G25000	1502	1502
G30000	744	744
N=500, *rec*=50, *m*=1		
Generations	Inferred	Recombinant
G2000	847	847
G5000	538	538
G10000	784	784
G12000	320	320
G15000	235	235
G18000	821	821
G20000	316	316
G22000	499	499
G25000	208	208
G30000	218	218
N=500, *rec*=50, *m*=10		
Generations	Inferred	Recombinant
G2000	5214	5214
G5000	2585	2585
G10000	8977	8977
G12000	6104	6104
G15000	3100	3100
G18000	2572	2572
G20000	5577	5577
G22000	3462	3462
G25000	5289	5289
G30000	4319	4319

N=100, *rec*=100, *m*=1		
Generations	Inferred	Recombinant
G2000	18	18
G5000	48	48
G10000	57	57
G12000	57	57
G15000	90	90
G18000	37	37
G20000	126	126
G22000	129	129
G25000	16	16
G30000	106	106
N=100, *rec*=100, *m*=10		
Generations	Inferred	Recombinant
G2000	2358	2358
G5000	600	600
G10000	1544	1544
G12000	1453	1453
G15000	1540	1540
G18000	1918	1918
G20000	1641	1641
G22000	742	742
G25000	1848	1848
G30000	1687	1687
N=500, *rec*=100, *m*=1		
Generations	Inferred	Recombinant
G2000	775	775
G5000	674	674
G10000	394	394
G12000	527	527
G15000	146	146
G18000	283	283
G20000	458	458
G22000	931	931
G25000	448	448
G30000	380	380
N=500, *rec*=100, *m*=10		
Generations	Inferred	Recombinant
G2000	6814	6814
G5000	5753	5753
G10000	5662	5662
G12000	7214	7214
G15000	7151	7151
G18000	8866	8866
G20000	6326	6326
G22000	8495	8495
G25000	6002	6002
G30000	6852	6852