Supplementary information for manuscript:

Horizontal Transfers and Gene Losses in the Phospholipid Pathway of *Bartonella* Reveal Clues about Early Ecological Niches

Authors: Qiyun Zhu, Michael Kosoy, Kevin J. Olival, Katharina Dittmar

Contents

Figure S1 Core genome phylogeny of *Bartonella*

Figure S2 Standard phylogenetic analysis of *Bartonella*

Figure S3 Genomic contexts of *gpsA* (Rh), (Ar) and (Se) in *Bartonella* and other bacterial groups

Figure S4 Phylogeny of the Ugp system in *Bartonella*

Figure S5 Phylogeny of *Bartonella plsX* gene

Table S1 Additional genomic data assessed in this study

Table S2 Summary of gene sequence alignments analyzed in this study

Table S3 Selection test results

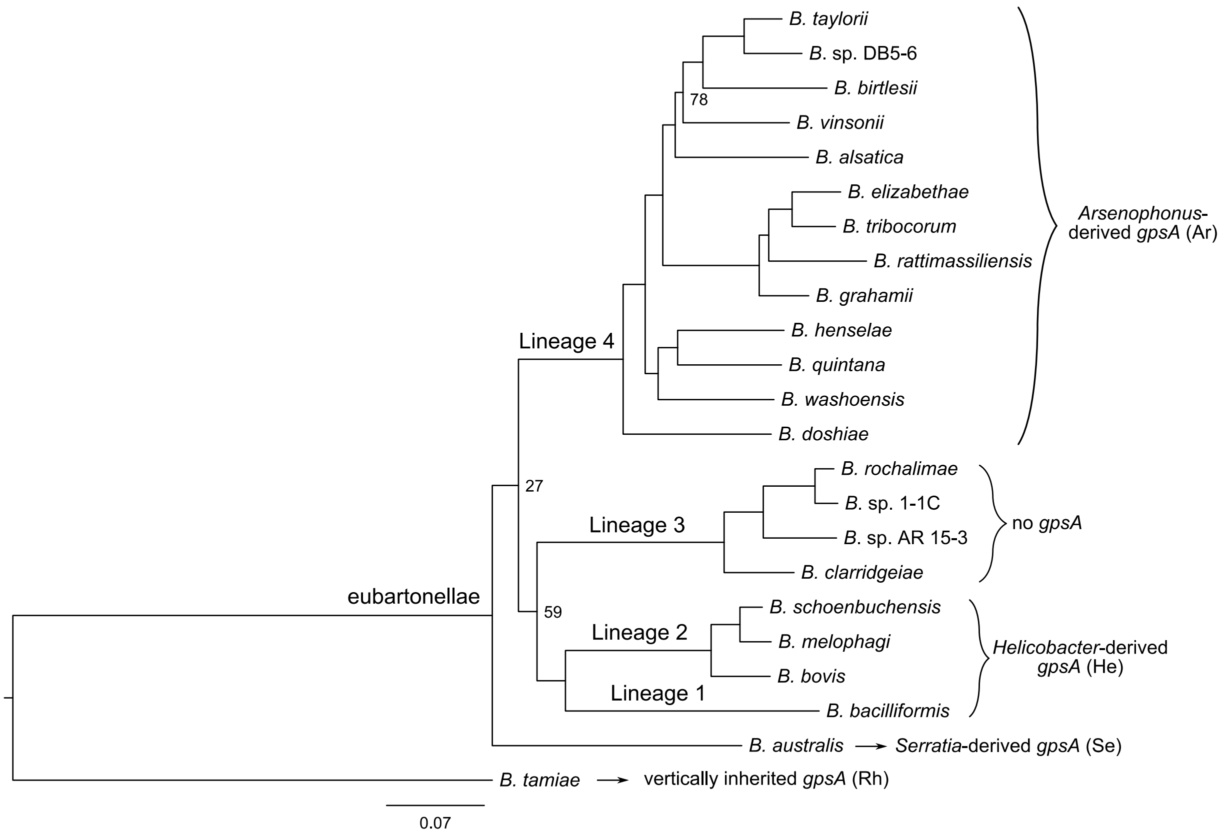

Figure S1 Core genome phylogeny of *Bartonella*

The tree was reconstructed based on 516 single-copy genes shared by 23 *Bartonella* species using the maximum likelihood (ML) method as implemented in RAxML. Major *Bartonella* lineages according to Engel et al. (2011) are indicated. Node labels are bootstrap support values (out of 100 replicates). Nodes without labels are fully supported (bootstrap = 100).

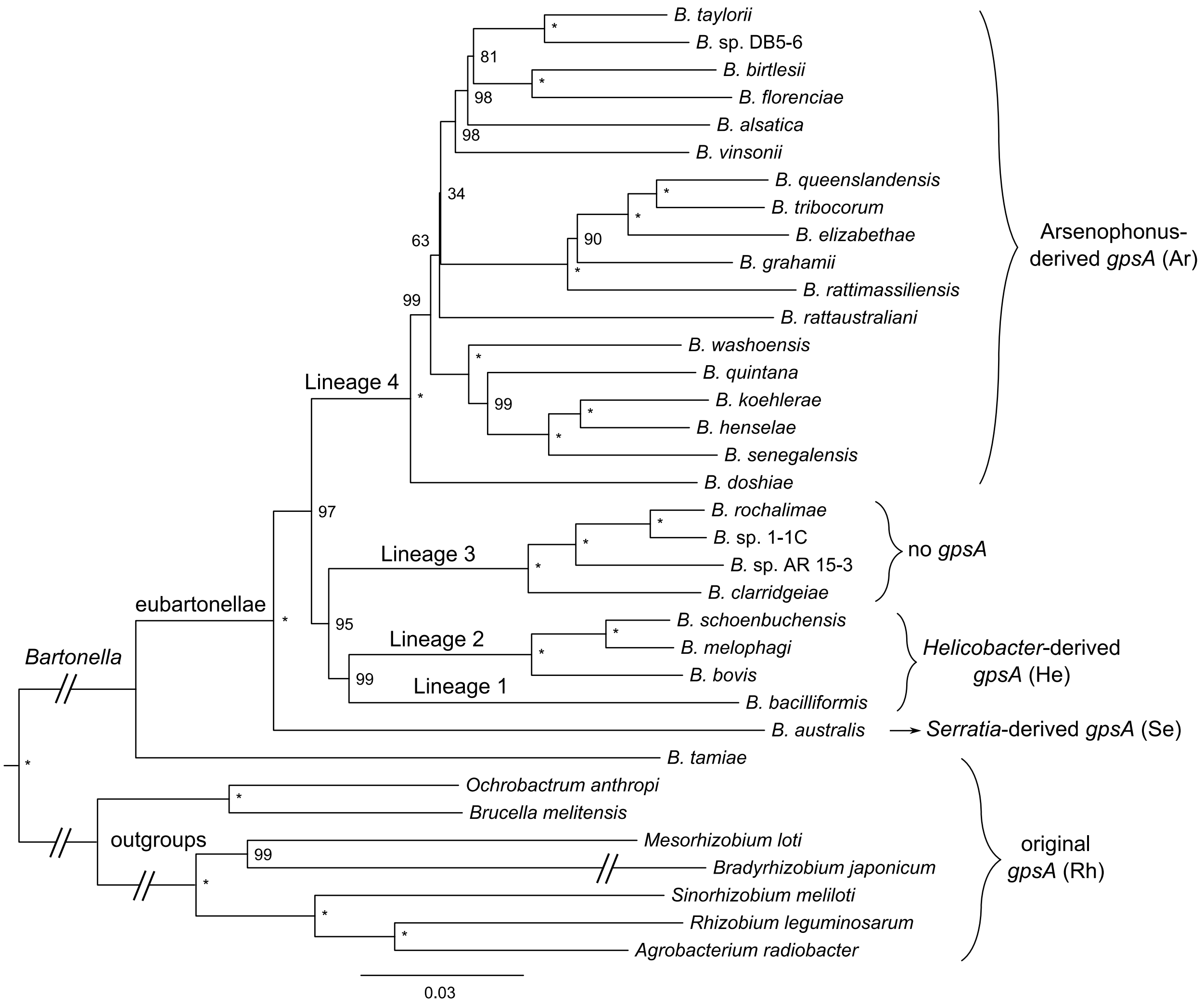

Figure S2 Standard phylogenetic analysis of *Bartonella*

Tree topology based on commonly used genetic markers of 28 *Bartonella* species, rooted to Rhizobiales. *Bartonella* species are the ingroup. Major *Bartonella* lineages according to Engel et al. (2011) are indicated. Node labels are bootstrap support values (out of 100 replicates). Asterisks (*) indicate 100% support.

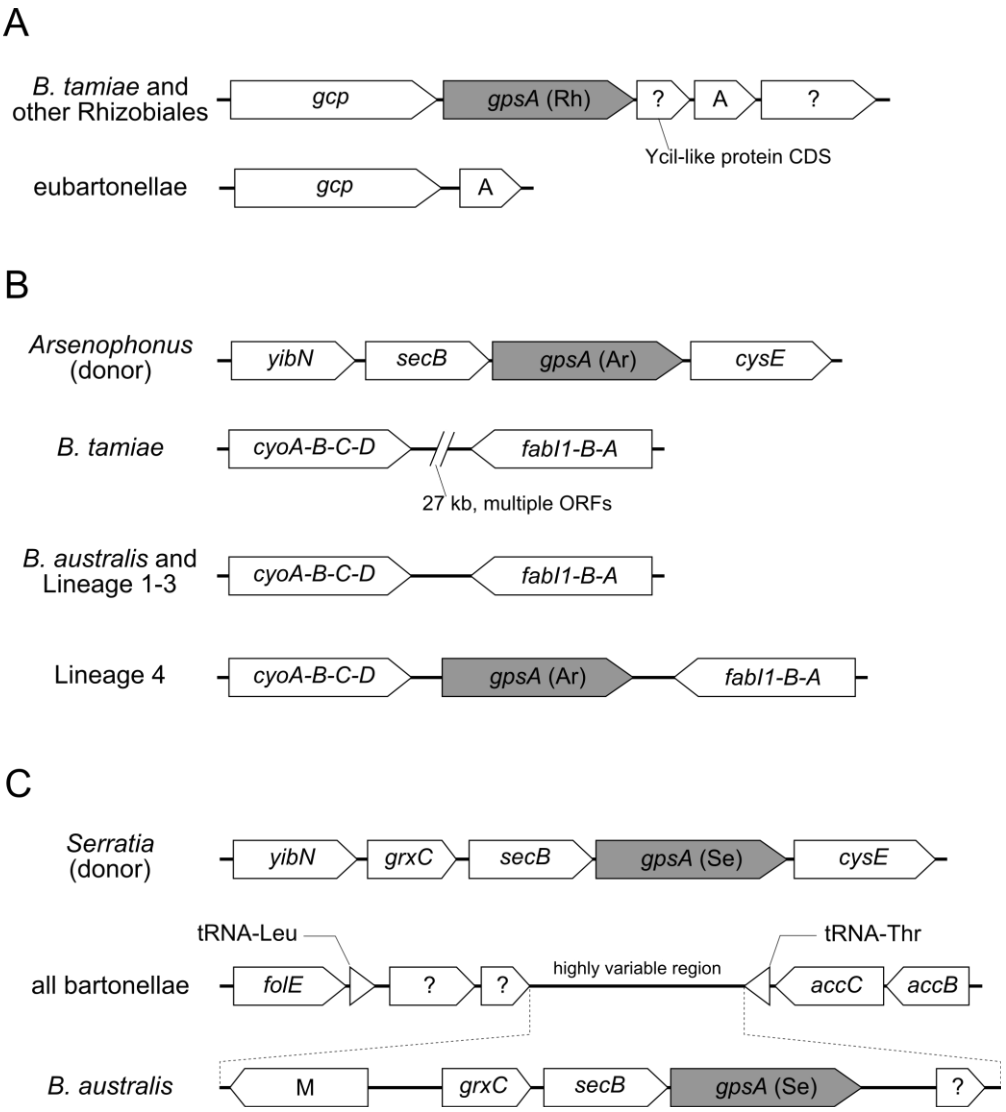

Figure S3 Genomic contexts of *gpsA* (Rh), (Ar) and (Se) loci in *Bartonella* and other bacterial groups

Genes are represented by boxes. Lengths of genes and intergenic regions are not drawn to scale. Tandemly arranged genes that share the same symbol (a putative operon) are merged. "M" in (C) is a multi-copy ORF (also see Figure 4). Other ORFs annotated as hypothetical genes are indicated by "?". Scenarios illustrated in the panels are: (A) Loss of *gpsA* (Rh) in stem eubartonella. (B) HGT of *gpsA* (Ar) from *Arsenophonus* to stem L4 *Bartonella*. The *gpsA* (Ar) gene is present in all Lineage 4 bartonellae but not in other lineages. (C) HGT of *gpsA* (Se) from *Serratia* to *B. australis*. All bartonellae share the structure shown in the middle row, in which only *B. australis* has *gpsA* (Se) within this structure as shown in the bottom row.

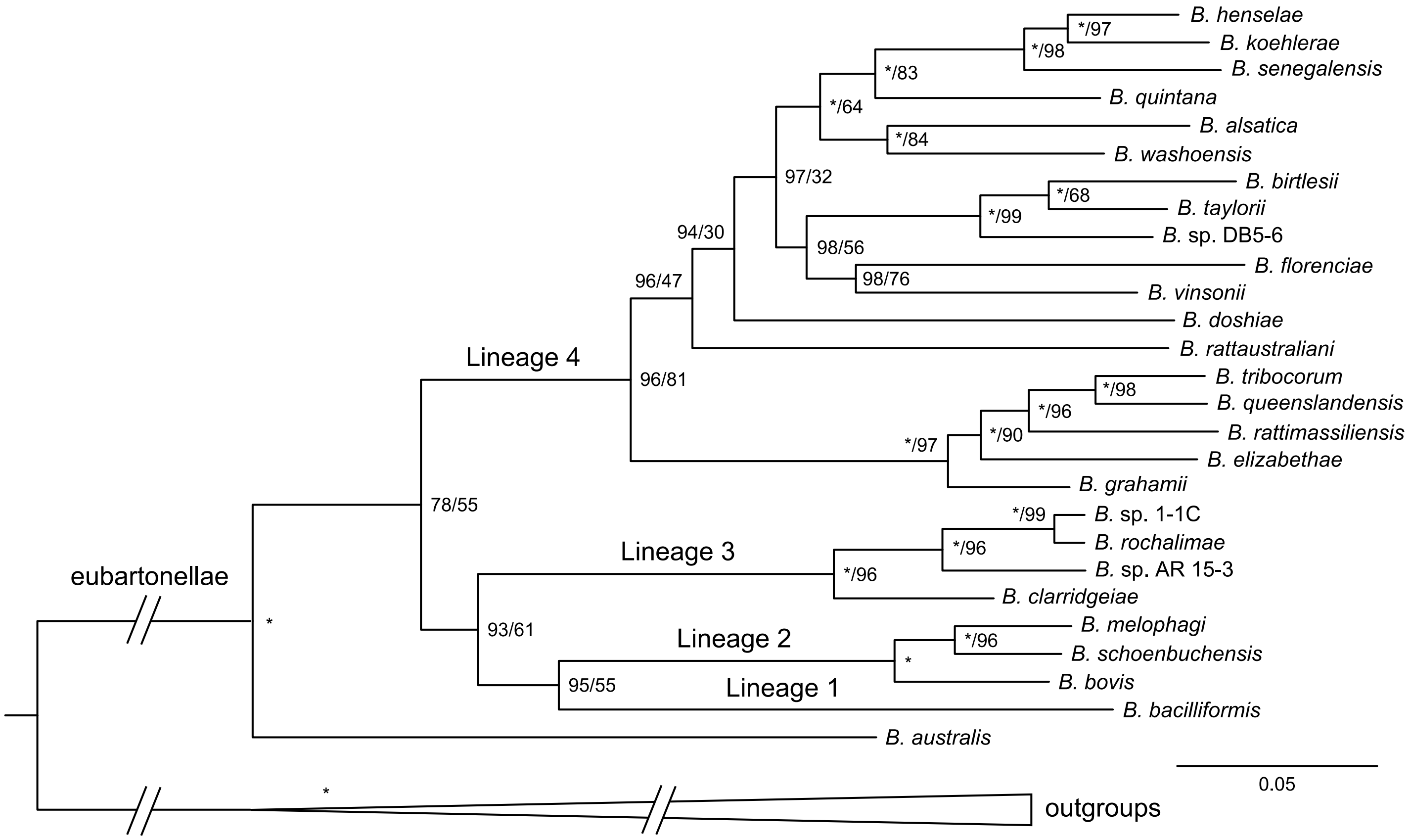

Figure S4 Phylogeny of the Ugp system in *Bartonella*

Four genes that constitute the *ugp* operon: *ugpB*, *A*, *E* and *C*, were concatenated, a phylogenetic tree was inferred in MrBayes. Node labels are Bayesian posterior probabilities / maximum-likelihood bootstrap support values. Asterisks (*) indicate 100% support. The tree is rooted to Rhizobiales outgroups (Table S1), which are collapsed into a triangle.

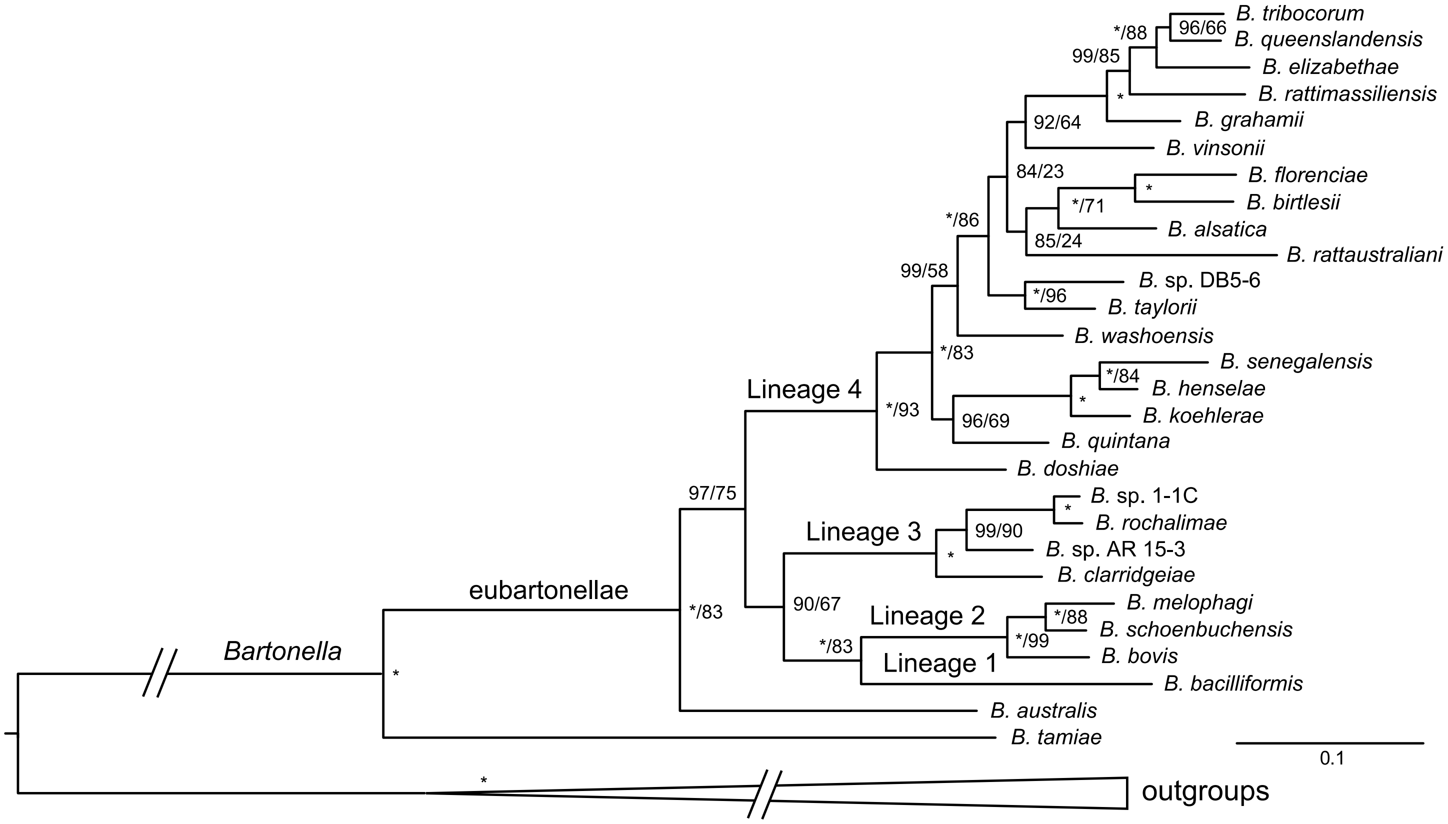

Figure S5 Phylogeny of *Bartonella plsX* gene

Node labels are Bayesian posterior probabilities / maximum-likelihood bootstrap support values. Asterisks (*) indicate 100% support. The tree is rooted to Rhizobiales outgroups (Table S1), which are collapsed into a triangle.

Table S1 Additional genomic data assessed in this study

Genus	Species	Strain	NCBI Assembly
Outgroups of *Bartonella* (Figure 2, S1-4)			
Ochrobactrum	*O. anthropi*	ATCC 49188	GCF_000017405
Brucella	*B. melitensis*	bv. 1 str. 16M	GCF_000007125
Mesorhizobium	*M. loti*	MAFF303099	GCF_000009625
Sinorhizobium	*S. meliloti*	1021	GCF_000006965
Rhizobium	*R. leguminosarum*	bv. viciae 3841	GCF_000009265
Agrobacterium	*A. radiobacter*	K84	GCF_000016265
Bradyrhizobium	*B. japonicum*	USDA 110	GCF_000011365
Other Rhizobiales (Figure 3A)			
Bartonella	*B. tamiae*	Th307	GCF_000279995
Brucella	*B. abortus*	S19	GCF_000018725
	B. microti	CCM 4915	GCF_000022745
	B. pinnipedialis	B2/94	GCF_000221005
	B. ovis	ATCC 25840	GCF_000016845
	B. canis	ATCC 23365	GCF_000018525
	B. suis	1330	GCF_000007505
Ochrobactrum	*O.* sp. CDB2	-	GCF_000344725
	O. intermedium	M86	GCF_000332835
Mesorhizobium	*M. ciceri*	bv. biserrulae WSM1271	GCF_000185905
	M. opportunistum	WSM2075	GCF_000176035
	M. australicum	WSM2073	GCF_000230995
Nitratireductor	*N. aquibiodomus*	RA22	GCF_000265055
	N. indicus	C115	GCF_000300515
	N. pacificus	pht-3B	GCF_000300335
Rhizobium	*R. etli*	CFN 42	GCF_000092045
	R. tropici	CIAT 899	GCF_000330885
Agrobacterium	*A. fabrum*	str. C58	GCF_000092025
	A. sp. H13-3	-	GCF_000192635
	A. vitis	S4	GCF_000016285
Sinorhizobium	*S. medicae*	WSM419	GCF_000017145
	S. fredii	NGR234	GCF_000018545
Bradyrhizobium	*B.* sp. ORS 278	-	GCF_000026145
Rhodopseudomonas	*R. palustris*	CGA009	GCF_000195775
Xanthobacter	*X. autotrophicus*	Py2	GCF_000017645
Methylobacterium	*M. extorquens*	PA1	GCF_000018845
Helicobacteraceae and other Campylobacterales (Figure 3B)			
Helicobacter	*H. pylori*	G27	GCF_000021165
	H. pylori	26695	GCF_000307795
	H. pylori	B38	GCF_000091345
	H. pylori	J99	GCF_000008785
	H. acinonychis	str. Sheeba	GCF_000009305
	H. bizzozeronii	CIII-1	GCF_000237285
	H. felis	ATCC 49179	GCF_000200595
	H. mustelae	12198	GCF_000091985
	H. cinaedi	CCUG 18818	GCF_000155475
	H. hepaticus	ATCC 51449	GCF_000007905
	H. canadensis	MIT 98-5491	GCF_000162575
	H. winghamensis	ATCC BAA-430	GCF_000158455
Wolinella	*W. succinogenes*	DSM 1740	GCF_000196135
Campylobacter	*C. concisus*	13826	GCF_000017725
	C. fetus	subsp. fetus 82-40	GCF_000015085
Arcobacter	*A. butzleri*	JV22	GCF_000185325

Genus	Species	Strain	NCBI Assembly
Enterobacteriales and other Gammaproteobacteria (Figure 3C)			
Arsenophonus	*A. nasoniae*	DSM 15247	GCA_000429565
Providencia	*P. alcalifaciens*	DSM 30120	GCF_000173415
	P. rettgeri	DSM 1131	GCF_000158055
	P. rustigianii	DSM 4541	GCF_000156395
	P. stuartii	ATCC 25827	GCF_000154865
Photorhabdus	*P. asymbiotica*	ATCC 43949	GCF_000196475
	P. luminescens	subsp. laumondii TTO1	GCF_000196155
Xenorhabdus	*X. bovienii*	SS-2004	GCF_000027225
	X. nematophila	ATCC 19061	GCF_000252955
Proteus	*P. mirabilis*	HI4320	GCF_000069965
	P. penneri	ATCC 35198	GCF_000155835
	P. vulgaris	CCUG 19017	GU254061
Morganella	*M. morganii*	subsp. morganii KT	GCF_000286435
Serratia	*S. proteamaculans*	568	GCF_000018085
	S. liquefaciens	ATCC 27592	GCF_000422085
	S. plymuthica	AS9	GCF_000214235
	S. symbiotica	str. 'Cinara cedri'	GCF_000238975
	S. marcescens	WW4	GCF_000336425
Yersinia	*Y. enterocolitica*	subsp. palearctica 105.5R(r)	GCF_000192105
	Y. kristensenii	ATCC 33638	GCF_000173715
	Y. pestis	A1122	GCF_000222975
	Y. rohdei	ATCC 43380	GCF_000173775
Candidatus Hamiltonella	*H. defensa*	5AT	GCF_000021705
Dickeya	*D. dadantii*	Ech703	GCF_000023545
	D. zeae	Ech1591	GCF_000023565
Pectobacterium	*P. atrosepticum*	SCRI1043	GCF_000011605
Candidatus Blochmannia	*B. floridanus*	-	GCF_000043285
	B. pennsylvanicus	str. BPEN	GCF_000011745
	B. vafer	str. BVAF	GCF_000185985
Wigglesworthia	*W. glossinidia*	endosymbiont of *Glossina morsitans morsitans* (Yale colony)	GCF_000247565
Baumannia	*B. cicadellinicola*	str. Hc (*Homalodisca coagulata*)	GCF_000013185
Sodalis	*S. glossinidius*	str. 'morsitans'	GCF_000010085
Escherichia	*E. albertii*	albertii TW07627	GCF_000155105
	E. coli	2534-86	GCF_000225065
	E. fergusonii	ATCC 35469	GCF_000026225
Shigella	*S. boydii*	Sb227	GCF_000012025
	S. dysenteriae	Sd197	GCF_000012005
	S. flexneri	K-315	GCF_000268165
	S. sonnei	Ss046	GCF_000092525
Citrobacter	*C. koseri*	ATCC BAA-895	GCF_000018045
	C. rodentium	ICC168	GCF_000027085
Salmonella	*S. bongori*	NCTC 12419	GCF_000252995
	S. enterica	subsp. enterica serovar Typhi str. P-stx-12	GCF_000245535
Enterobacter	*E. cloacae*	subsp. dissolvens SDM	GCF_000235765
	E. radicincitans	DSM 16656	GCF_000280495
	E. sp. 638	-	GCF_000016325
Yokenella	*Y. regensburgei*	ATCC 43003	GCF_000239335
Klebsiella	*K. oxytoca*	E718	GCF_000276705
	K. pneumonia	KCTC 2242	GCF_000220485
	K. variicola	At-22	GCF_000025465
Cronobacter	*C. sakazakii*	ES15	GCF_000263215
	C. turicensis	z3032	GCF_000027065
Erwinia	*E. amylovora*	ATCC 49946	GCF_000027205

Genus	Species	Strain	NCBI Assembly
	E. billingiae	Eb661	GCF_000196615
	E. pyrifoliae	DSM 12163	GCF_000026985
	E. tasmaniensis	ET1/99	GCF_000026185
Pantoea	*P. ananatis*	AJ13355	GCF_000270125
	P. sp. GM01	-	GCF_000282675
	P. sp. YR343	-	GCF_000282695
	P. vagans	C9-1	GCF_000148935
Actinobacillus	*A. pleuropneumoniae*	serovar 3 str. JL03	GCF_000016685
Haemophilus	*H. sputorum*	CCUG 13788	GCF_000238795
Pasteurella	*P. dagmatis*	ATCC 43325	GCF_000163475
	P. multocida	subsp. multocida str. Pm70	GCF_000006825
Vibrio	*V. cholera*	O395	GCF_000021625
	V. cyclitrophicus	ZF14 subcontig_64	GCF_000256265
	V. fischeri	SR5	GCF_000241785
	V. parahaemolyticus	K5030	GCF_000182465

Table S2 Summary of gene sequence alignments analyzed in this study

Gene	No. of sequences	No. of ***Bartonella*** sequences	Length (bp)	Pairwise identity
gpsA (Rh)	34	2 (*B. tamiae*)	972	60.6%
gpsA (He)	21	5 (L1 &L2)	927	60.5%
gpsA (Ar & Se)	109	39 (L4 & *B. australis*)	984	63.6%
ugpB-A-E-C	35	30	4110	69.4%
plsX	35	30	1065	74.7%

Table S3 Selection test results

Gene	Tree	M0	Site models		Foreground branch	Branch model			Branch site models	
		ω	LRT (M2a vs. M1a)	LRT (M8a vs. M8)		ω_0	ω_1	LRT (one-ratio vs. two-ratio)	LRT (A1 vs. A)	Positive sites (BEB)
gpsA (Rh)	Figure 3A	0.0909	0.00	0.00	*B. tamiae* clade	0.0939	0.0035	16.90***	5.05*	26C, 80Y, 83L, 146E, 150K, 258N*, 271A, 284R, 302K, 314R*, 328S
					Agrobacterium + *Rhizobium* clade	0.0867	0.0991	1.45	-	
					Brucella clade	0.0913	0.0810	0.14	-	
					Ochrobactrum + *Brucella* clade	0.0919	0.0865	0.22	-	
					Mesorhizobium clade	0.0917	0.0856	0.23	-	
					Nitratireductor clade	0.0974	0.0525	13.81***	-	
					Sinorhizobium clade	0.0911	0.0891	0.02	-	
gpsA (He)	Figure 3B	0.0730	0.00	0.00	*Bartonella* (L1+L2) node	0.0754	0.0012	8.37***	3.19	62S, 102K*, 230Q
					Bartonella (L1+L2) clade	0.0298	0.1563	52.9***	0.00	102K, 118T, 124H, 230Q
gpsA (Ar) & (Se)	Figure 3C	0.0887	0.00	0.00	*Bartonella* (L4) node	0.0886	0.0889	0.00	0.88	102N*, 150I, 319G, 333T*
					Bartonella (L4) clade	0.0694	0.1742	240.00***	0.00	24A**, 37Y**, 38N**, 49H**, 54A**, 60Q*, 77A*, 79R**, 91H**, 95N*, 97Q**, 102N**, 104H**, 116H**, 121L*, 166S*, 176D*, 239L**, 267I**, 275I**, 282I*, 296R**, 297M**, 314H**, 319G**, 331S*, 333T**
					Arsenophonus node	0.0889	0.0369	0.83	-	
					Arsenophonus clade	0.0883	0.1194	1.42	-	
					Arsenophonus-Bartonella (L4) node	0.0893	0.0498	5.03*	-	
					B. australis tip	0.0877	0.2050	10.49**	0.00	40S, 107V, 108L, 123W*, 218V*, 284R, 301H, 303D
					Serratia clade (except for *B. australis*)	0.0909	0.0526	17.56***	-	
ugpB-A-E-C	Figure S4	0.0844	-	-	*B. australis* tip	0.0846	0.0825	0.04	-	
					Bartonella (L1) tip	0.0837	0.0986	1.55	-	
					Bartonella (L2) clade	0.0816	0.1160	12.64***	-	
					Bartonella (L3) clade	0.0856	0.0756	1.86	-	
					Bartonella (L4) clade	0.0921	0.0806	4.94*	-	
plsX	Figure S5	0.0826	0.00	2.50	*Bartonella* (L3) clade	0.0794	0.1100	3.83	-	

The table reports ω (dN/dS value), LRT (likelihood ratio test value, = $2(\ln L_1 - \ln L_0)$) and its significance level (indicated by *), as well as putative positively selected sites with significance level (indicated by *) estimated by the BEB method. Branches of interest were labeled as foreground branches in the analyses. “Node” refers to a single branch. “Clade” refers to the selected branch and all sub-branches after it. ω under the one-ratio model (M0) is the global ω for the whole tree. ω_1 and ω_0 under the branch model refer to the ω values for the foreground branch and the tree background, respectively. Coordinates and residues of putative positively selected sites refer to the protein sequence alignment in Figure 5.