Document S1. Supporting evidence for the REIS assembly. This supplement marshals evidence that our assembly is accurate and comes from a single endosymbiont with a uniform phylogenetic placement within *Rickettsia*. (A) Brief background information regarding the Wikel tick colony and the nature of transovarial transmission of spotted fever group rickettsiae. (B-E) Analyses of 16S rDNA sequences and of 18 combined proteins encoding the *rvh* T4SS, demonstrating that diverse gene sets consistently support REIS as a basal member of the SFG group, yet with high similarity to previously identified endosymbionts whose genome sequences are not yet available. (F-H) Analysis of 1499 REIS proteins with top BLASTP hits to proteins encoded in other *Rickettsia* genomes and 742 REIS proteins with top BLASTP hits to proteins encoded in non-*Rickettsia* genomes, showing that nearly all proteins have congruent phylogenies such that the assembly has not likely included accidental fragments from other organisms. Note: more descriptive legends follow the figures.

A

The Wikel Ixodes scapularis *colony*. Established by Dr. Stephen Wikel (Quinnipiac University School of Medicine) in 1996, the *I. scapularis* Wikel colony was seeded with approximately 30 pairs of field collected adult male and female ticks from populations in New York, Oklahoma and a Lyme disease endemic area of Connecticut. By 2004 (the time of genome sequencing), the colony had been continuously in-bred for approximately 12 generations, limiting the degree of polymorphism typical of *I. scapularis* natural populations with normal levels of gene flow (8). Importantly, since its establishment, the Wikel colony has not been supplemented with field-collected material, and is very carefully maintained. Exposure of the colony to known denizens of *I. scapularis* (several of which are pathogens), including *Babesia*, *Borrelia*, *Bartonella*, *Anaplasma*, and flavivirus, has not been detected, with ticks being maintained on laboratory animals (pathogen-free blood sources). Furthermore, prior to genomic DNA extraction, the ticks were surface sterilized to eliminate microbial contamination. Thus, because of its age, inbred nature and careful maintenance, the only likely microorganisms residing within the Wikel colony are those associated with transovarial transmission.

Transovarial transmission of rickettsiae. Transovarial transmission in some arthropod vectors entails the transfer of bacteria from parent arthropod directly to offspring arthropod via the ovaries. This mode of transmission is characteristic of many rickettsial species, especially spotted fever group (SFG) rickettsiae, wherein single bacterial species are maintained in their arthropod hosts through generations, with the competitive exclusion of congeners (a mechanism that remains poorly understood) (1-7). Thus, the detection of a SFG rickettsial symbiont in the *I. scapularis* genome assembly is not surprising. From the initial analysis of the preliminary *I. scapularis* assembly, it was determined that a single bacterial species was likely present (based on 16S rDNA analysis). Any non-rickettsial bacteria present in the field-collected *I. scapularis* individuals that seeded the Wikel colony would not have persisted in subsequent generations because they are disseminated only by transstadial transmission or acquired via horizontal transmission from infected blood sources. Furthermore, due to the exclusion principal of transovarial transmission, multiple rickettsial species (or subspecies) would have culminated in one endosymbiont that continually propagated through the ensuing *I. scapularis* generations. Thus, such possibilities are null considering 12 generations of laboratory maintenance.

In summary, the dynamics of tick-borne SFG rickettsiae transmission, coupled with the history and maintenance of the Wikel colony, lead us to believe that a single rickettsial species is present in the *I. scapularis* genome assembly.

1. **Burgdorfer, W.** 1989. Ecological and Epidemiological considerations of Rocky Mountain spotted fever and scrub typhus, p. 34-47. *In* D. H. Walker (ed.), Biology of Rickettsial Diseases, vol. 1. , vol. 1. CRC Press, Boca Raton, FL.
2. **Burgdorfer, W.** 1975. A review of Rocky Mountain spotted fever (tick-borne typhus), its agent, and its tick vectors in the United States. J Med Entomol **12:**269-278.
3. **Burgdorfer, W., S. F. Hayes, and M. A. J.** 1981. Nonpathogenic rickettsiae in *Dermacentor andersoni*: a limiting factor for *Rickettsia rickettsii*, p. 585-594. *In* W. Burgdorfer and R. L. Anacker (ed.), Rickettsiae and Rickettsial Diseases. Academic Press, New York, NY.
4. **Hayes, S. F., and W. Burgdorfer.** 1990. Interactions between rickettsial endocytobionts and their tick hosts. . *In* S. A. Gossner (ed.), Insect Endocytobiosis: Morphology, Physiology, Genetics, Evolution. CRC PRess, Boca Raton, FL.
5. **Macaluso, K. R., A. Mulenga, J. A. Simser, and A. F. Azad.** 2003. Differential expression of genes in uninfected and rickettsia-infected *Dermacentor variabilis* ticks as assessed by differential-display PCR. Infect Immun **71:**6165-6170.
6. **Macaluso, K. R., D. E. Sonenshine, S. M. Ceraul, and A. F. Azad.** 2001. Infection and transovarial transmission of rickettsiae in *Dermacentor variabilis* ticks acquired by artificial feeding. Vector Borne Zoonotic Dis **1:**45-53.
7. **Macaluso, K. R., D. E. Sonenshine, S. M. Ceraul, and A. F. Azad.** 2002. Rickettsial infection in *Dermacentor variabilis* (Acari: Ixodidae) inhibits transovarial transmission of a second *Rickettsia*. J Med Entomol **39:**809-813.
8. **Pagel Van Zee, J., N. S. Geraci, F. D. Guerrero, S. K. Wikel, J. J. Stuart, V. M. Nene, and C. A. Hill.** 2007. Tick genomics: the *Ixodes* genome project and beyond. Int J Parasitol **37:**1297-1305.

B

Accession	Description	Max score	Tot. score	Query coverage	E value	Max ident
D84558.1	Rickettsia sp. (Ixodes scapularis symbiont)	2617	2617	100%	0	99%
DQ100164.1	Rickettsia monacensis strain IrR/Munich	2591	2591	100%	0	99%
AF141908.1	Rickettsia sp. IRS4	2586	2586	100%	0	99%
AF141907.1	Rickettsia sp. IRS3	2580	2580	100%	0	99%
NR_042727.1	Rickettsia tamurae strain AT-1	2564	2564	100%	0	99%
EU072493.1	Uncultured Rickettsia sp. clone 11122	2564	2564	100%	0	99%
DQ365809.1	Rickettsia raoultii strain Marne	2564	2564	100%	0	99%
L36102.1	Rickettsia sp. (16S rRNA) gene	2564	2564	100%	0	99%
AF120026.1	Rickettsia sp. RpA4 strain RpA4	2556	2556	99%	0	99%
CP000683.1	Rickettsia massiliae MTU5,	2558	2558	100%	0	99%
NR_025919.1	Rickettsia massiliae strain Mtu1	2558	2558	100%	0	99%
L36106.1	Rickettsia massiliae (16S rRNA) gene	2558	2558	100%	0	99%
EU036982.1	Rickettsia raoultii strain Elanda-23/95	2551	2551	99%	0	99%
AF120024.1	Rickettsia sp. DnS28 strain DnS28	2551	2551	99%	0	99%
EF589609.1	Rickettsia sp. TwKM01	2553	2553	100%	0	99%
NR_043755.1	Rickettsia raoultii strain Khabarovsk	2553	2553	100%	0	99%
L36224.1	Rickettsia slovaca (strain 13-B)	2553	2553	100%	0	99%
NR_025921.1	Rickettsia rhipicephali strain 3-7-6	2553	2553	100%	0	99%
AF120025.1	Rickettsia sp. DnS14 strain DnS14	2545	2545	99%	0	99%
AB604667.1	Rickettsia endosymbiont (Curculio hilgendorfi)	2534	2534	99%	0	99%
AB604665.1	Rickettsia endosymbiont (Curculio camelliae)	2534	2534	99%	0	99%
L28944.1	Rickettsia sp. 16S ribosomal DNA sequence	2529	2529	99%	0	99%
AP011533.1	Rickettsia japonica YH DNA,	2547	2547	100%	0	99%
CP000766.2	Rickettsia rickettsii str. Iowa,	2547	2547	100%	0	99%
CP001227.1	Rickettsia peacockii str. Rustic	2547	2547	100%	0	99%
AE006914.1	Rickettsia conorii str. Malish 7	2547	2547	100%	0	99%
CP000848.1	Rickettsia rickettsii str. 'Sheila Smith'	2547	2547	100%	0	99%
EF589608.1	Rickettsia sp. IG-1	2547	2547	100%	0	99%
CP000053.1	Rickettsia felis URRWXCal2	2547	2547	100%	0	99%
AF394903.1	Rickettsia sp. FLA1	2547	2547	100%	0	99%
DQ062433.1	Rickettsia peacockii	2547	2547	100%	0	99%
L36220.1	Rickettsia sp. (strain TT-118)	2547	2547	100%	0	99%
L36213.1	Rickettsia japonica (strain YM)	2547	2547	100%	0	99%
L36107.1	Rickettsia conorii	2547	2547	100%	0	99%
U12460.1	Rickettsia conorii ITT-597	2547	2547	100%	0	99%
U11021.1	Rickettsia rickettsii	2547	2547	100%	0	99%
NR_041934.1	Rickettsia conorii strain Malish 7	2542	2542	99%	0	99%
L36105.1	Rickettsia conorii	2542	2542	99%	0	99%

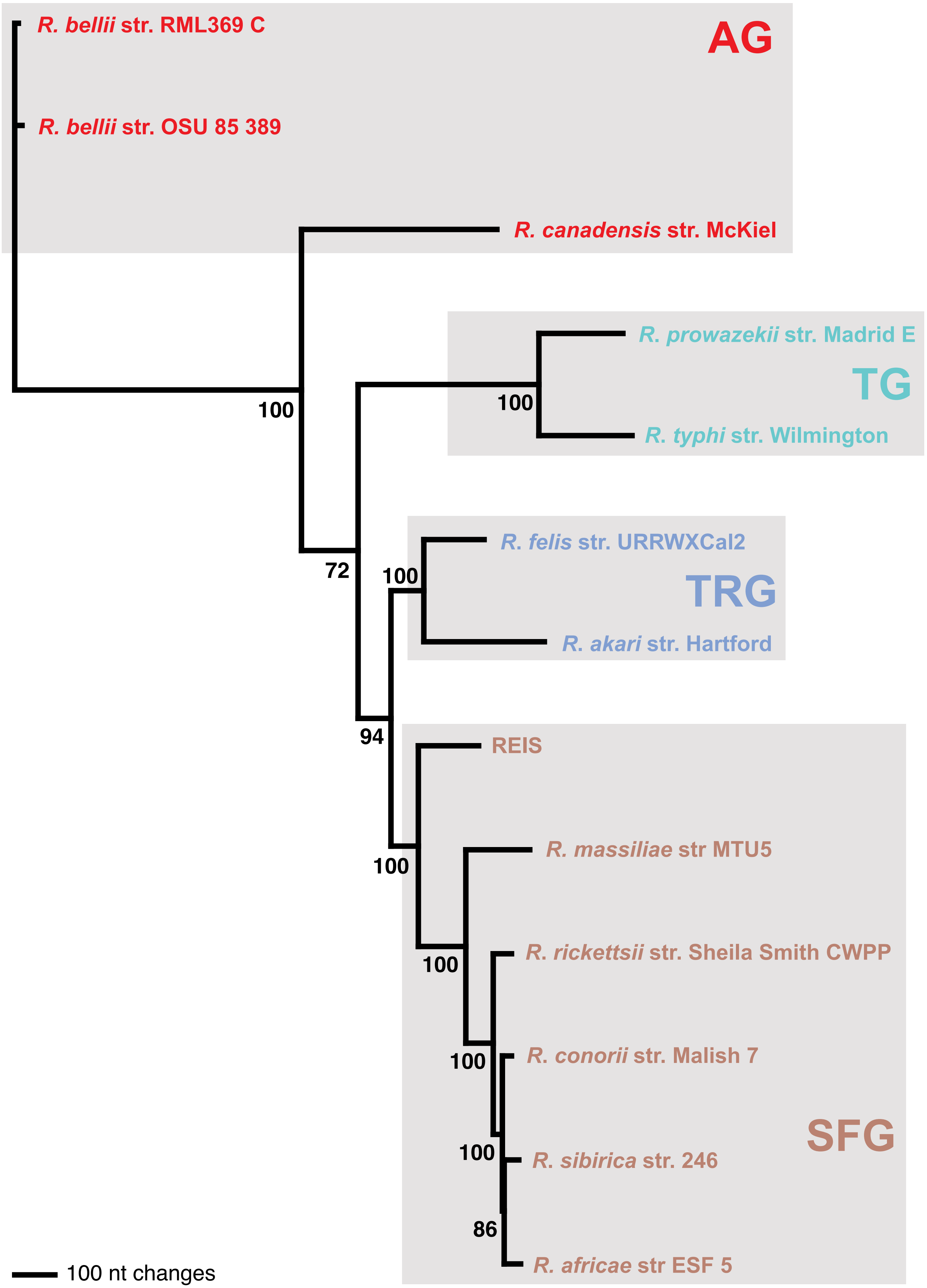
C

```
>dbj|D84558.1|IXORDNAA  Rickettsia sp. (Ixodes scapularis symbiont) gene for 16S rRNA,
partial sequence
Length=1420

 Score = 2617 bits (1417),  Expect = 0.0
 Identities = 1420/1421 (99%), Gaps = 1/1421 (0%)
 Strand=Plus/Plus


REIS    181   GGATGAGCCCGCGTCAGATTAGGTCGTTGGTGAGGTAACGGCTCACCAAGCCGACGATCT  240
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
D84558  181   GGATGAGCCCGCGTCAGATTAGGTCGTTGGTGAGGTAACGGCTCACCAAGCCGACGATCT  240


REIS    241   GTAGCTGGTCTGAGAGGATGATCAGCCACACTGGGACTGAGACACGGCCCAGACTCCTAC  300
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||| ||||
D84558  241   GTAGCTGGTCTGAGAGGATGATCAGCCACACTGGGACTGAGACACGGCCCAGACT-CTAC  299
                                                                     *

REIS    301   GGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGAAAGCCTGATCCAGCAATACCGAGT  360
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
D84558  300   GGGAGGCAGCAGTGGGGAATATTGGACAATGGGCGAAAGCCTGATCCAGCAATACCGAGT  359
```

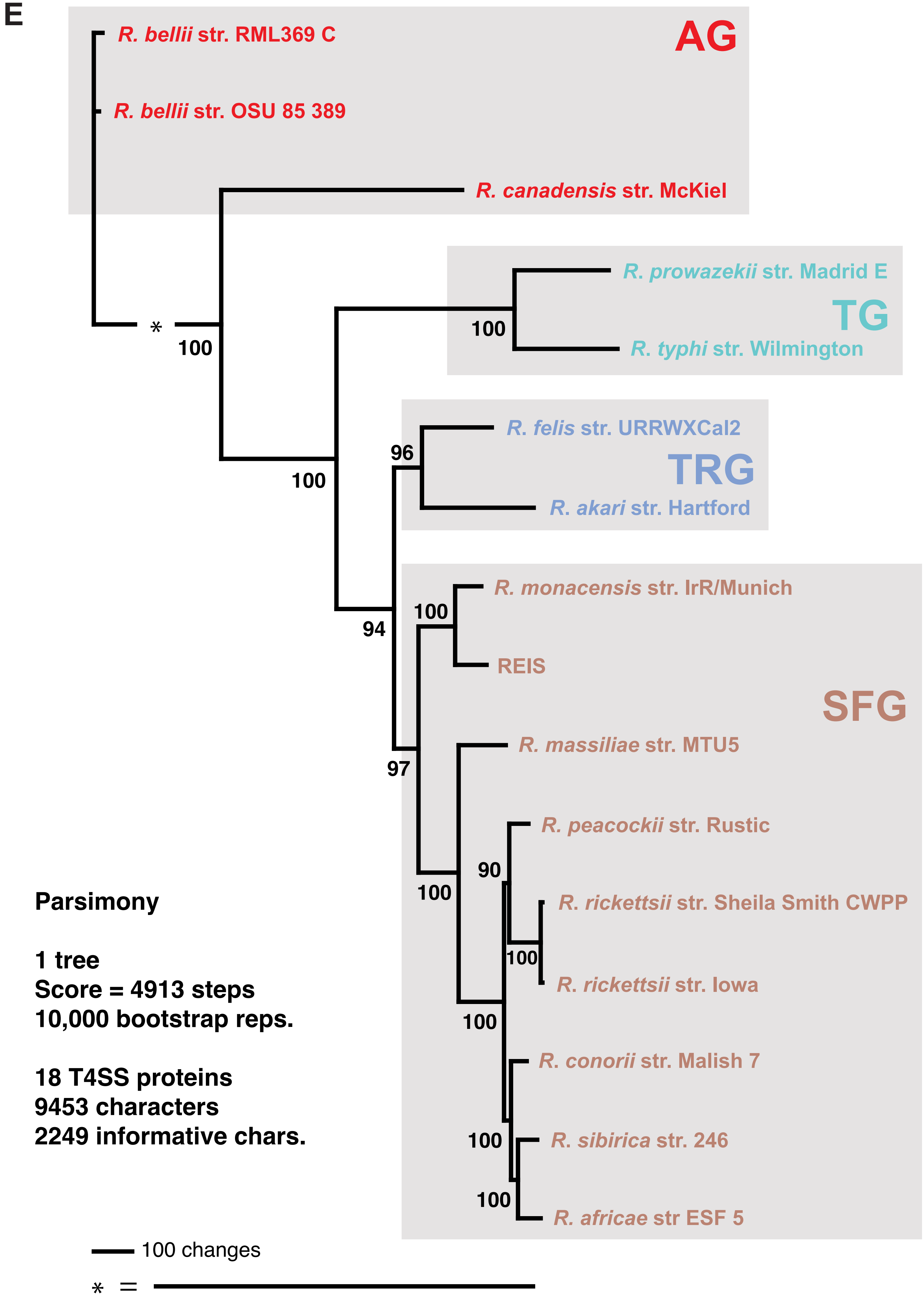
(B) Blastn hits using REIS 16S rDNA sequence as query (retrieved from PATRIC, VBIRicEnd40569_r031). Only subjects scoring a maximum identity of 99% are listed. NOTE: 36 nts from the 5'-end and 51 nts from the 3'-end of the sequence were trimmed to target the region predominantly sequenced by rickettsial researchers (using the entire REIS gene inflates similarity statistics shared with full length 16S rDNA sequences from complete genome sequencing projects). The top hit to the 16S rDNA sequence from a rickettsial endosymbiont of *I. scapularis* (GenBank accession no. D84558) is highlighted.

(C) Blastn information and partial region showing the single difference between the REIS and D84558 16S rDNA sequences. Note: the next closest sequence to the REIS 16S rDNA is from *R. monacensis* str. IrR/Munich (DQ100164), which is known from European ixodid ticks (particularly *I. ricinus*).

D

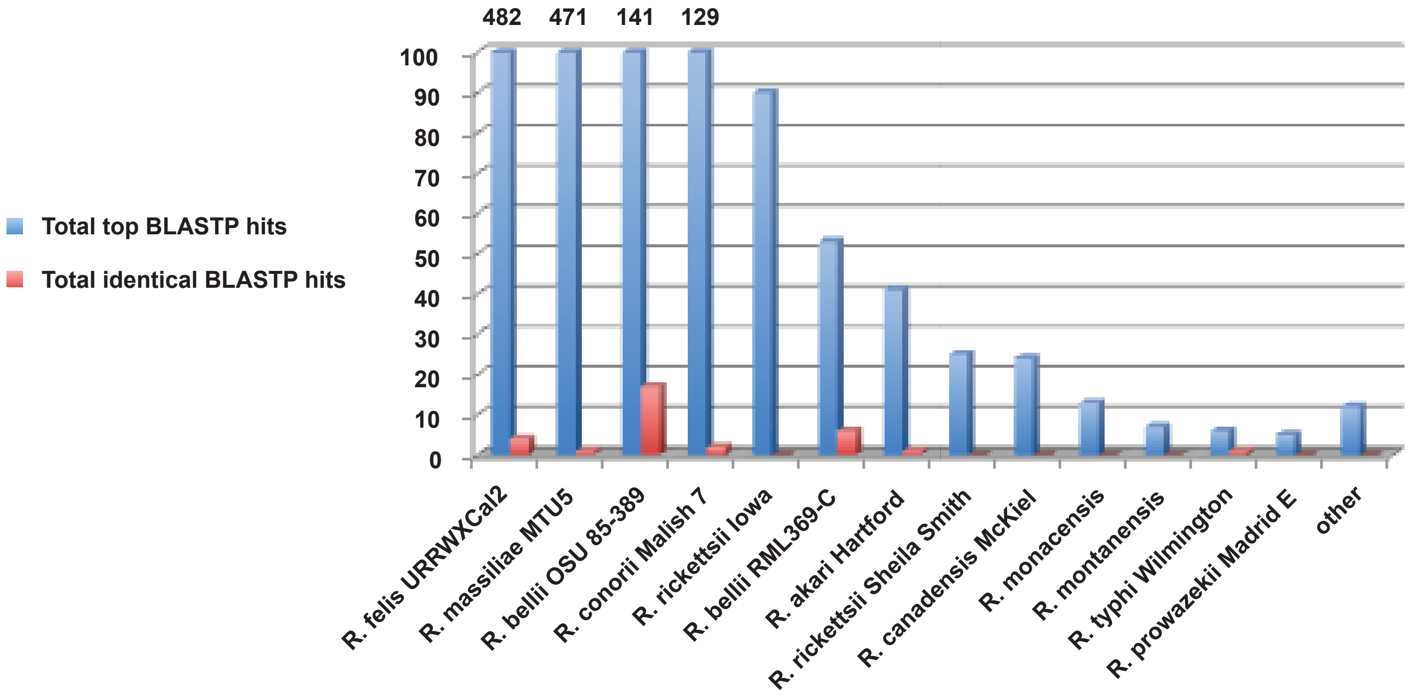

R. bellii str. RML369 C
AG
R. bellii str. OSU 85 389
R. canadensis str. McKiel
100
R. prowazekii str. Madrid E
TG
100
R. typhi str. Wilmington
R. felis str. URRWXCal2
72
100
TRG
R. akari str. Hartford
94
REIS
R. massiliae str MTU5
100
100
R. rickettsii str. Sheila Smith CWPP
100
R. conorii str. Malish 7
100
R. sibirica str. 246
SFG
86
R. africae str ESF 5
100 nt changes

(D) Phylogeny estimation of rickettsial 16S rDNA sequences, illustrating that REIS is indeed a well-diverged lineage in relation to the other *Rickettsia* spp. with an available complete genome sequence. Note: this estimated phylogeny is consistent with the species tree shown in Fig. 3 of the manuscript.


E
AG
R. bellii str. RML369 C
R. bellii str. OSU 85 389
R. canadensis str. McKiel
TG
R. prowazekii str. Madrid E
R. typhi str. Wilmington
TRG
R. felis str. URRWXCal2
R. akari str. Hartford
SFG
R. monacensis str. IrR/Munich
REIS
R. massiliae str. MTU5
R. peacockii str. Rustic
R. rickettsii str. Sheila Smith CWPP
R. rickettsii str. Iowa
R. conorii str. Malish 7
R. sibirica str. 246
R. africae str ESF 5
*
100
100
100
96
94
97
100
100
90
100
100
100
100
Parsimony
1 tree
Score = 4913 steps
10,000 bootstrap reps.
18 T4SS proteins
9453 characters
2249 informative chars.
100 changes
* =

(E) Phylogeny estimation of 18 proteins encoding the *rvh* T4SS with the inclusion of unpublished sequences from *R*. *monacensis* str. IrR/Munich (courtesy of Ulrike Munderloh and Tim Kurtti, U. of Minnesota). REIS and *R*. *monacensis* str. IrR/Munich form a monophyletic group, but with each lineage well diverged from their common ancestor (>50 aa differences across 18 proteins).

F



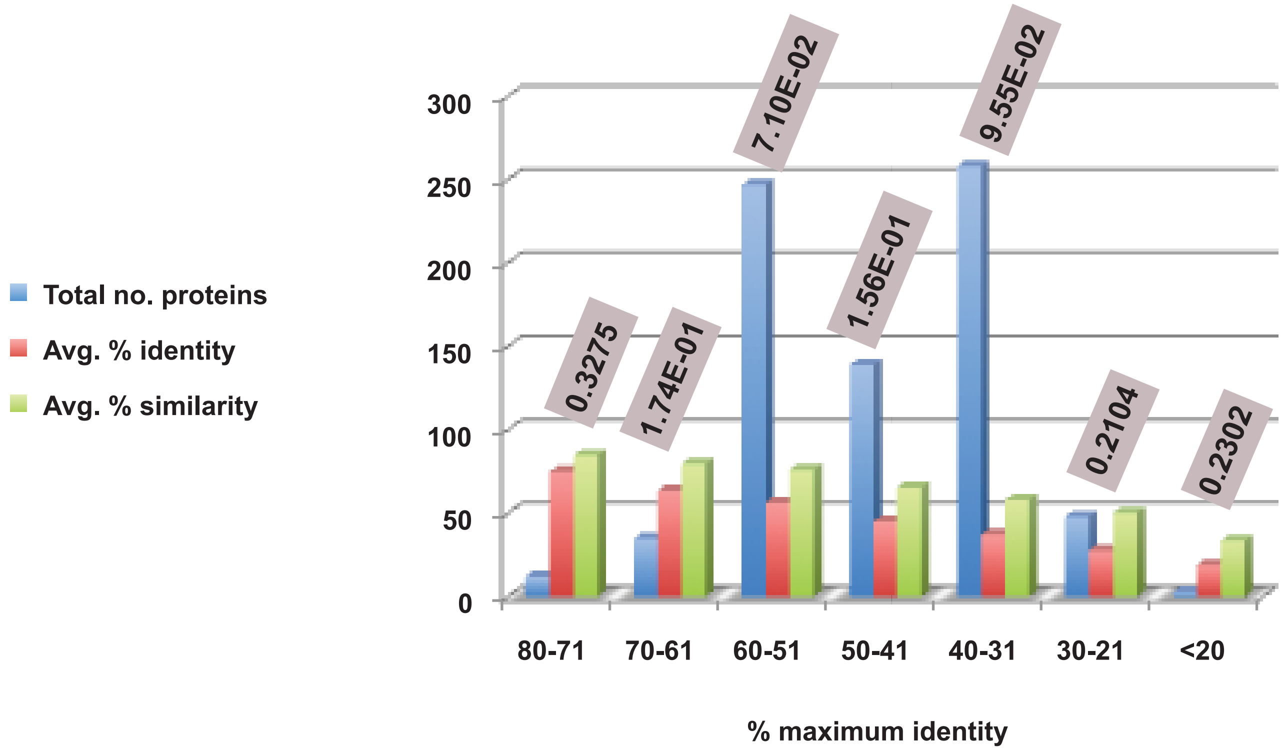
G

REIS	L	Genome													
no.	(aa)	Br	Bo	Ca	Pr	Ty	Fe	Ak	Ma	Pe	Ri	Rw	Co	Si	Af
116	86	62	58	99	99	99	ID	94	ID	98	99	99	ID	ID	ID
197	71	93	93	97	99	ID	99	97	97	97	99	99	99	97	99
591	97	87	87	94	86	82	99	98	ID	99	99	99	ID	ID	ID
841	388	ID	ID						90						
842	594	ID	99							94					
843	55	ID	ID												
844	102	ID	ID						93	92					
845	183	ID	ID						98	93					
846	148	ID	ID						91	92					
848	472	99	ID						90	89					
849	84	ID	ID						92						
852	68	ID	ID												
853	322	ID	ID						91						
854	144	ID	ID						70						
856	277	99	ID						79						
857	444	ID	ID						89						
858	905	ID	ID						91						
859	397	ID	ID						65						
860	562	ID	ID						92						
863	97	ID	ID	76			44		70						
864	215	ID	ID												
865	48	ID	ID												
867	248	ID	ID						70						
868	316	ID	ID						48						
871	569	ID	ID						76						74
918	68	ID	ID												
1114	87	80	79	91	91	87	99	90	ID	98	98	98	95	98	98
1328	475	94	94	98	95	95	ID	98	99	99	99	99	98	99	98
1832	127	92	92	98	98	98	ID	98	99	ID	99	99	99	99	99
1845	122	98	98	99	98	98	ID	98	98	98	98	98	98	98	98
1899	78	94	88	95	91	90	ID	99	ID	ID	99	99	ID	ID	ID

(F) Distribution across other *Rickettsia* genomes of REIS proteins with top BLASTP hits to *Rickettsia* proteins. For each genome, the total number of top BLASTP hits (blue) and total number of identical BLASTP hits (red) is shown. 'Other' includes the following: *R. raoultii* (3 proteins), *R. helvetica* (2), *R. rhipicephali* (2), *R. africae* (1), *R. australis* (1), rickettsial endosymbiont of *Ixodes scapularis* (1), *Rickettsia* sp. IRS 4 (1), *R. parkeri* (1).

(G) REIS proteins identical to proteins in other *Rickettsia* genomes, and the distribution of these proteins across the other *Rickettsia* genomes. ID (highlighted yellow), identical. All other numbers depict maximum sequence identity determined from BLASTP analysis. Black cells depict absence of significant homolog or presence of distant paralog. The pink shading depicts the proteins encoded within the RAGE-Be, which is highly similar across REIS and *R. bellii* genomes (see manuscript text for details). Annotation for the 31 REIS proteins as follows: REIS_0116, acyl carrier protein; REIS_0197, translation initiation factor IF-1; REIS_0591, 50S ribosomal protein L28; REIS_0841, integrase; REIS_0842, leucine rich repeat domain protein; REIS_0843, conserved hypothetical protein; REIS_0844, conserved hypothetical protein; REIS_0845, conjugative transfer protein TraE; REIS_0846, conserved hypothetical protein; REIS_0848, F pilus assembly protein TraB; REIS_0849, conjugative transfer protein TraV; REIS_0852, prevent-host-death family protein; REIS_0853, conjugal DNA transfer protein TraU; REIS_0854, conjugative transfer protein TrbC; REIS_0856, F pilus assembly protein TraF; REIS_0857, F pilus assembly protein TraH; REIS_0858, conjugative transfer protein TraG; REIS_0859, tetratricopeptide repeat-containing protein; REIS_0860, putative conjugative transfer protein TraD; REIS_0863, conjugal transfer protein TraD; REIS_0864, guanosine polyphosphate pyrophosphohydrolase/synthetase; REIS_0865, conserved hypothetical protein; REIS_0867, guanosine polyphosphate pyrophosphohydrolase/synthetase; REIS_0868, signal transduction histidine kinase; REIS_0871, conserved hypothetical protein; REIS_0918, prevent-host-death family protein; REIS_1114, tetratricopeptide repeat-containing protein; REIS_1328, RNA pyrophosphohydrolase; REIS_1832, 30S ribosomal protein S11; REIS_1845, ribosomal protein L14; REIS_1899, conserved domain protein.

H


Total no. proteins
Avg. % identity
Avg. % similarity
300
250
200
150
100
50
0
0.3275
1.74E-01
7.10E-02
1.56E-01
9.55E-02
0.2104
0.2302
80-71
70-61
60-51
50-41
40-31
30-21
<20
% maximum identity

(H) Analysis of 742 REIS proteins with top BLASTP hits to proteins encoded in non-*Rickettsia* genomes. The BLASTP results were ranked according to maximum identity scores (%), and plotted in intervals of ten (x-axis). Note: no REIS proteins were identical to any proteins from non-*Rickettsia* organisms, with the highest identity score equal to 80%. For each interval of maximum identity the total number of proteins (blue), average percent identity (red) and average percent similarity (green) are shown. Values within gray boxes above each interval depict the average E-value across the total comparisons per interval.