Fig. S2. Additional information supporting the use case "Application to Comparative Genomics: Erythritol Utilization in Brucella". (A) Genes encoding enzymes involved in erythritol transport and catabolism in *Brucella* spp., with emphasis on candidate pseudogenes across 16 genomes [1, 2]. (B) Whole genome-based phylogeny estimation for 107 Rhizobiales taxa. One outgroup (Sphingomonadales) is used to root the tree. Arrow points to inset showing cladogram of the 41 *Brucella* genomes. Red bar marks the origin of the *abortus* clade, and red asterisk denotes the monophyly of *B. abortus* strains S19 and NCTC 8038 (see text for details). The following pipeline was implemented to estimate Rhizobiales phylogeny: BLAT (refined BLAST algorithm) [3] searches were performed to identify similar protein sequences between all genomes, including the outgroup taxon. To predict initial homologous protein sets, mcl [4] was used to cluster BLAT results, with subsequent refinement of these sets using in-house hidden Markov models [5]. These protein families were then filtered to include only those with membership in >80% of the analyzed genomes (85 or more taxa included per protein family). Multiple sequence alignment of each protein family was performed using MUSCLE (default parameters) [6, 7], with masking of regions of poor alignment (length heterogeneous regions) done using Gblocks (default parameters) [8, 9]. All modified alignments were then concatenated into one dataset. Tree-building was performed using FastTree [10]. Support for generated lineages was estimated using a modified bootstrapping procedure, with 100 pseudoreplications sampling only half of the aligned protein sets per replication (NOTE: standard bootstrapping tends to produce inflated support values for very large alignments). Local refinements to tree topology were attempted in instances where highly supported nodes have subnodes with low support. This refinement is executed by running the entire pipeline on only those genomes represented by the node being refined (with additional sister taxa for rooting purposes). The refined subtree was then spliced back into the full tree. More information pertaining to this phylogeny pipeline is available at PATRIC (see "Phylogeny FAQs" at http://enews.patricbrc.org/faqs/). (C) Application of the Multiple Sequence Alignment Viewer tool to evaluate the origin and diversification of the ATP-binding (*eryE*), permease (*eryF*), and substrate-binding (*eryG*) components of erythritol ABC transporters 1 and 2. Using the 'Protein Family Sorter' tool, orthologous proteins for each *ery* protein (three FIGfams for EryE, five FIGfams for EryF and four FIGfams for EryG) were extracted from the interactive heatmap. Specifically, once a set of FIGfams was captured, the "show proteins" option was selected. From this table, all proteins were selected (65 for EryE, 72 for EryF and 56 for EryG). Next, the "Integrated Protein Tree and Alignment" option was selected, resulting in the display of the full length multiple sequence alignment coupled with an estimated phylogeny (the Multiple Sequence Alignment Viewer tool). For more information regarding the Multiple Sequence Alignment Viewer tool, see the "Alignment FAQs".

1. Crasta OR, Folkerts O, Fei Z, Mane SP, Evans C, Martino-Catt S, Bricker B, Yu G, Du L, Sobral BW: **Genome sequence of *Brucella abortus* vaccine strain S19 compared to virulent strains yields candidate virulence genes**. *PLoS One* 2008, **3**(5):e2193.
2. Tsolis RM, Seshadri R, Santos RL, Sangari FJ, Lobo JM, de Jong MF, Ren Q, Myers G, Brinkac LM, Nelson WC *et al*: **Genome degradation in *Brucella ovis* corresponds with narrowing of its host range and tissue tropism**. *PLoS One* 2009, **4**(5):e5519.
3. Kent WJ: **BLAT--the BLAST-like alignment tool**. *Genome Res* 2002, **12**(4):656-664.
4. Van Dongen S: **Graph Clustering Via a Discrete Uncoupling Process**. *SIAM J Matrix Anal & Appl* 2008, **30**:121-141.
5. Durbin R, Eddy S, Krogh A, Mitchison G: **Biological sequence analysis: probabilistic models of proteins and nucleic acids**: Cambridge University Press; 1998.
6. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity**. *BMC Bioinformatics* 2004, **5**:113.

7. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**(5):1792-1797.
8. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis**. *Mol Biol Evol* 2000, **17**(4):540-552.
9. Talavera G, Castresana J: **Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments**. *Syst Biol* 2007, **56**(4):564-577.
10. Price MN, Dehal PS, Arkin AP: **FastTree 2--approximately maximum-likelihood trees for large alignments**. *PLoS One* 2010, **5**(3):e9490.

A

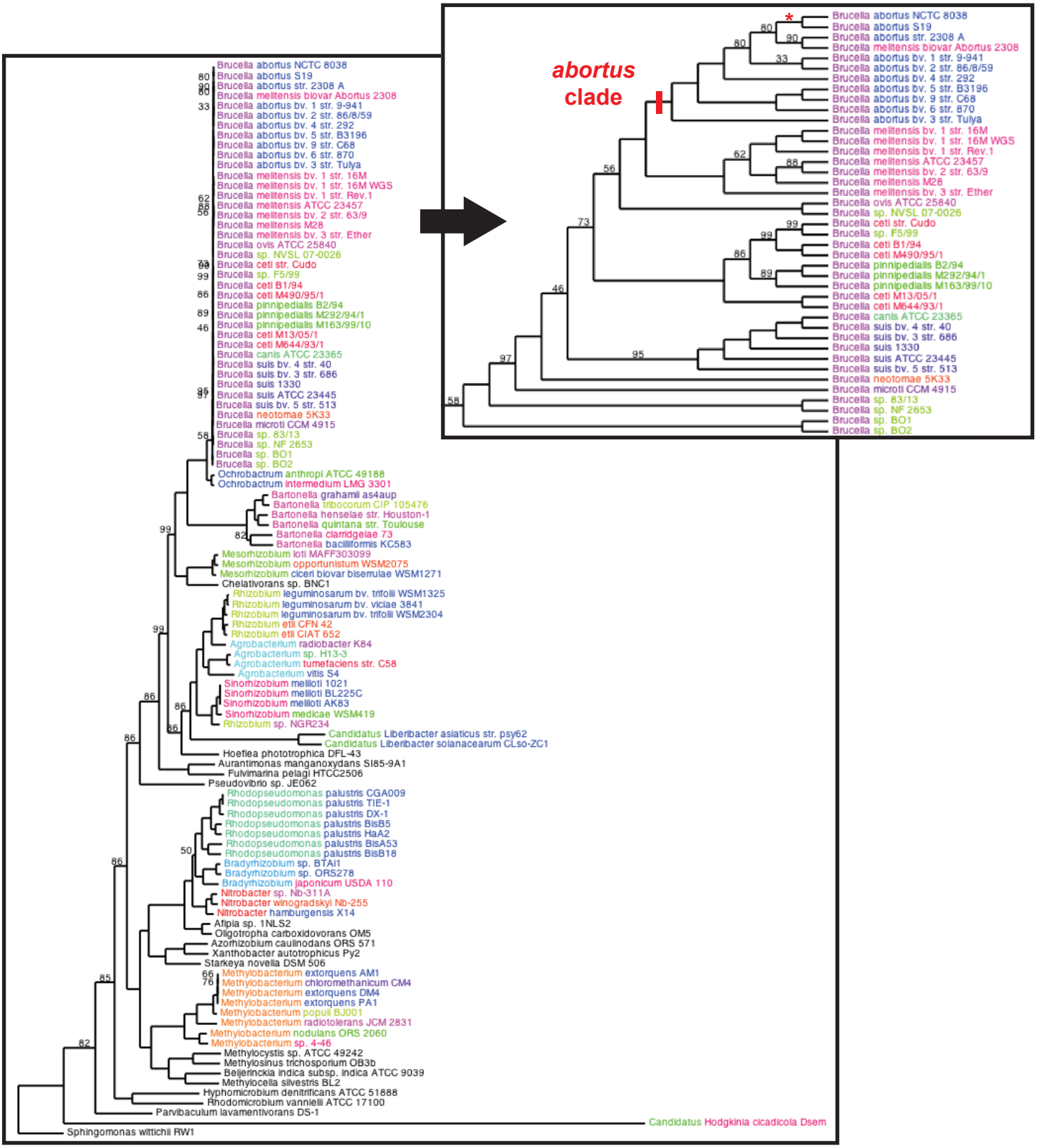
Gene	PATRIC Annotation	Putative Pseudogenes			Ref. [3]
		Genome[1]	Mutation[2]	Effect	
eryA	Erythritol kinase EryA	C	D (1)	frameshift	--
		E,G	D (1)	nonsense/premature stop	--
		I	PM	altered start site; short	T
		N	I (1)	nonsense/premature stop	--
		P	amb.s	disrupted reading frame	--
eryB	Erythritol phosphate dehydrogenase EryB	--	--	--	--
eryC	Possible D-erythrulose 4-phosphate dehydrogenase EryC	B	D (703)	not annotated	C
eryD	Erythritol transcriptional regulator EryD	B	D (703)	altered start site; short	C
		I	D (7)	nonsense/premature stop	T
--	Predicted erythritol ABC transporter 2, hypothetical lipoprotein	E,G	D (1)	nonsense/premature stop	--
eryE	Predicted erythritol ABC transporter 2, ATP-binding component	O	D (1)	frameshift	--
eryF	Predicted erythritol ABC transporter 2, permease component	A,B	D (67)	altered start site; short	--
		D-H	D (1)	nonsense/premature stop	--
		I	D (2)	altered start site; short	T
		J-M	D (1)	nonsense/premature stop	--
eryG	Predicted erythritol ABC transporter 2, substrate-binding component	I,O	D (1)	frameshift	T

[1] A, *Brucella abortus* NCTC 8038; B, *Brucella abortus* S19; C, *Brucella canis* ATCC 23365; D, *Brucella ceti* B1/94; E, *Brucella ceti* M13/05/1; F, *Brucella ceti* M490/95/1; G, *Brucella ceti* M644/93/1; H, *Brucella ceti* str. Cudo; I, *Brucella ovis* ATCC 25840; J, *Brucella pinnipedialis* B2/94; K, *Brucella pinnipedialis* M163/99/10; L, *Brucella pinnipedialis* M292/94/1; M, *Brucella* sp. F5/99; N, *Brucella* sp. NF 2653; O, *Brucella* sp. NVSL 07-0026; P, *Brucella suis* bv. 3 str. 686.

[2] D, deletion; PM, point mutation; I, insertion. Numbers in parentheses denote nucleotides.

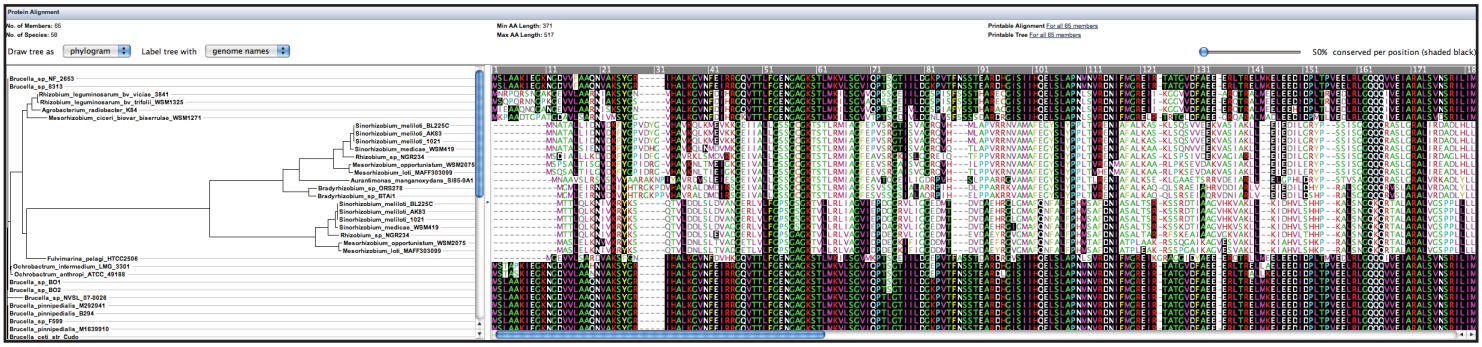
[3] T, Tsolis et al. [1]; C, Crasta et al. [2].

B

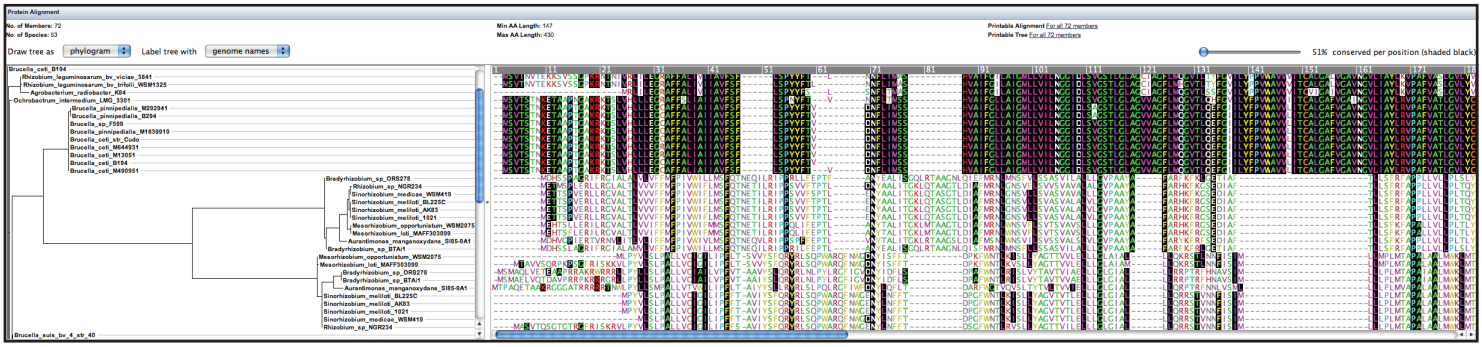

abortus clade
Brucella abortus NCTC 8038
Brucella abortus S19
Brucella abortus str. 2308 A
Brucella melitensis biovar Abortus 2308
Brucella abortus bv. 1 str. 9-941
Brucella abortus bv. 2 str. 86/8/59
Brucella abortus bv. 4 str. 292
Brucella abortus bv. 5 str. B3196
Brucella abortus bv. 9 str. C68
Brucella abortus bv. 6 str. 870
Brucella abortus bv. 3 str. Tulya
Brucella melitensis bv. 1 str. 16M
Brucella melitensis bv. 1 str. 16M WGS
Brucella melitensis bv. 1 str. Rev.1
Brucella melitensis ATCC 23457
Brucella melitensis bv. 2 str. 63/9
Brucella melitensis M28
Brucella melitensis bv. 3 str. Ether
Brucella ovis ATCC 25840
Brucella sp. NVSL 07-0026
Brucella ceti str. Cudo
Brucella sp. F5/99
Brucella ceti B1/94
Brucella ceti M490/95/1
Brucella pinnipedialis B2/94
Brucella pinnipedialis M292/94/1
Brucella pinnipedialis M163/99/10
Brucella ceti M13/05/1
Brucella ceti M644/93/1
Brucella canis ATCC 23365
Brucella suis bv. 4 str. 40
Brucella suis bv. 3 str. 686
Brucella suis 1330
Brucella suis ATCC 23445
Brucella suis bv. 5 str. 513
Brucella neotomae 5K33
Brucella microti CCM 4915
Brucella sp. 83/13
Brucella sp. NF 2653
Brucella sp. BO1
Brucella sp. BO2
Ochrobactrum anthropi ATCC 49188
Ochrobactrum intermedium LMG 3301
Bartonella grahamii as4aup
Bartonella tribocorum CIP 105476
Bartonella henselae str. Houston-1
Bartonella quintana str. Toulouse
Bartonella clarridgeiae 73
Bartonella bacilliformis KC583
Mesorhizobium loti MAFF303099
Mesorhizobium opportunistum WSM2075
Mesorhizobium ciceri biovar biserrulae WSM1271
Chelativorans sp. BNC1
Rhizobium leguminosarum bv. trifolii WSM1325
Rhizobium leguminosarum bv. viciae 3841
Rhizobium leguminosarum bv. trifolii WSM2304
Rhizobium etli CFN 42
Rhizobium etli CIAT 652
Agrobacterium radiobacter K84
Agrobacterium sp. H13-3
Agrobacterium tumefaciens str. C58
Agrobacterium vitis S4
Sinorhizobium meliloti 1021
Sinorhizobium meliloti BL225C
Sinorhizobium meliloti AK83
Sinorhizobium medicae WSM419
Rhizobium sp. NGR234
Candidatus Liberibacter asiaticus str. psy62
Candidatus Liberibacter solanacearum CLso-ZC1
Hoeflea phototrophica DFL-43
Aurantimonas manganoxydans SI85-9A1
Fulvimarina pelagi HTCC2506
Pseudovibrio sp. JE062
Rhodopseudomonas palustris CGA009
Rhodopseudomonas palustris TIE-1
Rhodopseudomonas palustris DX-1
Rhodopseudomonas palustris BisB5
Rhodopseudomonas palustris HaA2
Rhodopseudomonas palustris BisA53
Rhodopseudomonas palustris BisB18
Bradyrhizobium sp. BTAi1
Bradyrhizobium sp. ORS278
Bradyrhizobium japonicum USDA 110
Nitrobacter sp. Nb-311A
Nitrobacter winogradskyi Nb-255
Nitrobacter hamburgensis X14
Afipia sp. 1NLS2
Oligotropha carboxidovorans OM5
Azorhizobium caulinodans ORS 571
Xanthobacter autotrophicus Py2
Starkeya novella DSM 506
Methylobacterium extorquens AM1
Methylobacterium chloromethanicum CM4
Methylobacterium extorquens DM4
Methylobacterium extorquens PA1
Methylobacterium populi BJ001
Methylobacterium radiotolerans JCM 2831
Methylobacterium nodulans ORS 2060
Methylobacterium sp. 4-46
Methylocystis sp. ATCC 49242
Methylosinus trichosporium OB3b
Beijerinckia indica subsp. indica ATCC 9039
Methylocella silvestris BL2
Hyphomicrobium denitrificans ATCC 51888
Rhodomicrobium vannielii ATCC 17100
Parvibaculum lavamentivorans DS-1
Candidatus Hodgkinia cicadicola Dsem
Sphingomonas wittichii RW1

C

ATP-binding protein (*eryE*)



Permease (*eryF*)



Substrate binding protein (*eryG*)