S2 Table – Predictions and results for the genome evolution of *Bradyrhizobium* epidemics

Predictions for population genetic parameters of the CHR versus SI genome regions.

Model	Scenario	Predictions for parameters in CHR versus SI							
		GC content	**Π_1**	**Haplotype number[2]**	**Hd_3**	**Strain Richness[4]**	**Linkage within region[5]**	**Abundance[6]**	**Tajima's D_7**
SI Sweep	fixation of alleles localized to SI (see Sullivan & Ronson 1995)	differs in recently transferred genome regions (SI)	lower in SI	lower in SI	lower in SI	lower in SI	higher in SI	skewed in SI	negative in SI
CHR Sweep	fixation of alleles localized to CHR	differs in recently transferred genome regions (SI)	lower in CHR	lower in CHR	lower in CHR	lower in CHR	higher in CHR	skewed in CHR	negative in CHR
Whole genome sweep	novel alleles sweeps through CHR and SI (see Diep et al. 2006)	Similar across genome	NA[8]	NA[8]	NA[8]	NA[8]	NA[8]	skewed genome wide	negative in CHR & SI
Stasis	SI genome faithfully co-transmitted (see Juhas et al. 2007)	Similar across genome	Similar across genome	Similar across genome	Similar across genome	Similar across genome	Similar across genome	no skew	0 or greater for both

1. Π (Nucleotide diversity) is predicted to be reduced following a selective sweep.
2. A reduction in the number of haplotypes is predicted following a selective sweep.
3. Hd (haplotype diversity) is predicted to be reduced following a selective sweep.
4. Lower strain richness is predicted following a selective sweep.
5. Linkage is expected to be high in regions that have experienced a selective sweep.
6. A skewed distribution with a few highly abundant haplotypes and many rare ones is predicted following a selective sweep.
7. Tajima's D is a test to distinguish DNA evolving under neutral, directional or balancing selection. Selective sweeps are directional and should exhibit $D < 0$.
8. No prediction.

Results for tests of evolutionary-genomic scenarios.

Parameters are reported in comparisons between CHR and SI genome regions (Data summarized from S1, S4). Data are binned from the four well sampled clades (>20 isolates).

Clade	GC content	π	# haplotypes	Hd	Strain Richness	Linkage within region	Haplotype Abundance	Tajima's D
***B. canariense* n=244**	differs	similar across genome	depressed in CHR	depressed in CHR	depressed in CHR	higher in CHR	skewed in CHR	negative in SI / negative in CHR
***B. japonicum* n=27**	differs	similar across genome	depressed in CHR	similar	depressed in CHR	higher in CHR	skewed in CHR / skewed in SI	negative in SI / negative in CHR
***B. sp.* novel I n=54**	differs	similar across genome	depressed in CHR	depressed in CHR	depressed in CHR	higher in CHR	skewed in CHR	not significant
***B. yuanmingense* n=24**	differs	depressed in SI	similar	similar	similar	higher in CHR	skewed in CHR / skewed in SI	not significant

Legend
Data supports SI Sweep (fixation of alleles localized to SI loci)
Data supports CHR Sweep (fixation of alleles localized to CHR loci)
Data supports Whole genome sweep (novel alleles sweeps through CHR and SI)
Data supports Stasis (CHR and SI faithfully cotransmitted)
No single hypothesis supported

S3 Table – Population genetic parameters for loci and genomic regions.

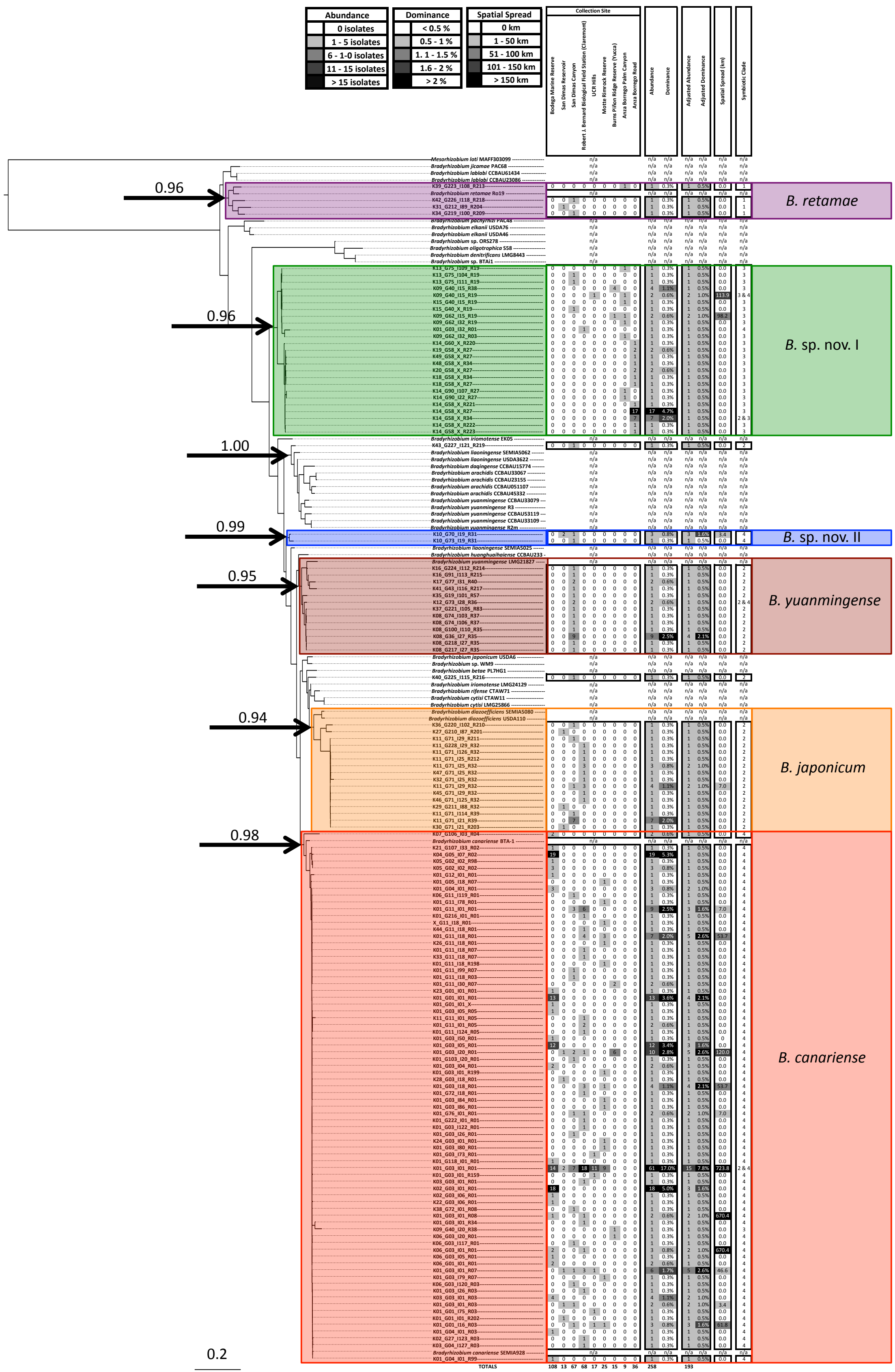
GC content is listed as a percentage. Nucleotide diversity (π) is the average number of nucleotide differences per site between any two randomly chosen sequences from a population. Haplotype diversity (Hd) is the probability that two isolates drawn at random are the same haplotype. Strain richness is calculated by dividing the number of haplotypes by the number of isolates collected. Mean F_{ST} was calculated between collection sites for each genome region using the Weir-Cockerham method.

Locus	# Isolates	# haplotypes	# bp used	# variable sites	GC content	π (nucleotide diversity)	ω (ka/ks)	Hd (haplotype diversity)
dnak	357	48	209	76	65.8	0.04	0.33	0.749
***gln*II**	358	46	560	172	63.2	0.04	0.16	0.803
ITS	321	48	1026	235	52.2	0.03	n/a	0.764
***rec*A**	357	44	414	126	68.4	0.04	0.18	0.746
***nif*D**	349	87	758	186	56.4	0.02	0.12	0.966
nodD-A	351	98	856	210	55.1	0.02	0.18	0.950
***nod*Z**	351	54	429	80	56.1	0.01	0.09	0.835
***nol*L**	354	68	650	142	53.1	0.03	0.30	0.900
CHR	358	138	2209	606	59.3	0.03	0.19	0.947
SI	357	226	2693	535	55.1	0.02	0.17	0.993
between genome regions	n/a	n/a	4942	1030	56.9	0.02	n/a	0.994
across genome	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

Locus	Average # nucleotide differences/ site (k)	Strain Richness	average linkage among loci|D'|	Estimate of Recombination (R)/gene	Estimate of R between adjacent sites	Minimum # of Recombination Events	Mean F_{ST}
dnak	0.0431	0.13	0.961	0.001	0.000	10	0.09
***gln*II**	0.0393	0.13	0.953	0.080	0.000	38	0.14
ITS	0.0254	0.15	0.932	0.001	0.000	30	0.12
***rec*A**	0.0386	0.12	0.981	0.300	0.001	25	0.12
***nif*D**	0.0185	0.25	0.978	0.001	0.000	24	0.07
nodD-A	0.0012	0.28	0.997	1.200	0.001	13	0.08
***nod*Z**	0.0140	0.15	0.994	0.700	0.002	6	0.06
***nol*L**	0.0262	0.19	0.987	0.001	0.000	11	0.10
CHR	0.0300	0.39	0.925	0.001	0.000	108	0.12
SI	0.0193	0.63	0.968	1.300	0.001	46	0.08
between genome regions	n/a	n/a	0.937	n/a	n/a	1	n/a
across genome	0.0227	n/a	0.946	0.0010	0.0000	124	n/a

S4 Figure – Phylogram of concatenated chromosomal loci (CHR)

The CHR tree was reconstructed in PhyML 3.0 using four loci (*dnak*, *glnII*, ITS*,* and *recA*) and is shown with associated heatmaps for abundance, dominance, and spatial range. Species-like clades are designated as highly supported (Shimodaira-Hasegawa support >0.90), non-nested monophyletic clades including no more than one reference species. Collection site heat map shows number of times the haplotype was isolated at that site. Abundance heat map indicates the number of times the haplotype was isolated out of a total of 358 isolates. Dominance heat map indicates the percentage a haplotype represents out of the total 358 isolates. Adjusted abundance and dominance refer to the same calculations, but only counting identical haplotypes from unique GPS locations. Spatial range is the maximum distance between collection sites as calculated from the geographic midpoint of each collection site.



Abundance
0 isolates
1 - 5 isolates
6 - 1-0 isolates
11 - 15 isolates
> 15 isolates
Dominance
< 0.5 %
0.5 - 1 %
1. 1 - 1.5 %
1.6 - 2 %
> 2 %
Spatial Spread
0 km
1 - 50 km
51 - 100 km
101 - 150 km
> 150 km
Collection Site
Bodega Marine Reserve
San Dimas Reservoir
San Dimas Canyon
Robert J. Bernard Biological Field Station (Claremont)
UCR Hills
Motte Rimrock Reserve
Burns Piñon Ridge Reserve (Yucca)
Anza Borrego Palm Canyon
Anza Borrego Road
Abundance
Dominance
Adjusted Abundance
Adjusted Dominance
Spatial Spread (km)
Symbiotic Clade
0.96
B. retamae
0.96
B. sp. nov. I
1.00
0.99
B. sp. nov. II
0.95
B. yuanmingense
0.94
B. japonicum
0.98
B. canariense
0.2
TOTALS
108 13 67 68 17 25 15 9 36
258
193

S5 Table – Population genetic parameters of well-sampled species-like clades.

Population genetic parameters are calculated separately for *Bradyrhizobium* clades with the greatest number of haplotypes (*B. canariense*, *B. japonicum*, *B.* sp. Novel I, and *B. yuanmingense*). Dataset included 4,942 sites; 1,030 polymorphic sites; and 374,545 pairwise comparisons. GC content listed as a percentage. Strain richness is calculated by dividing the number of haplotypes by the number of isolates collected. Haplotype diversity (Hd) is the probability that two isolates drawn at random are the same haplotype. Nucleotide diversity (π) is the average number of nucleotide differences per site between any two randomly chosen sequences from a population. Recombination rates (R) are measured between genes and sites.

Clade	Locus	# Isolates	# haplotypes	# nt sequenced	# variable sites	GC content	Strain Richness	Hd	π
B. canariense	dnak	243	18	209	44	65%	0.07	0.49	0.011
	glnII	244	17	560	100	63%	0.07	0.61	0.008
	ITS	244	26	1185	105	52%	0.11	0.62	0.003
	recA	243	15	414	66	69%	0.06	0.48	0.007
	nifD	242	52	758	95	56%	0.21	0.94	0.008
	nodDA	243	62	865	138	55%	0.26	0.92	0.006
	nodZ	244	29	429	49	56%	0.12	0.70	0.003
	nolL	242	37	650	91	53%	0.15	0.82	0.005
	CHR	242	71	2368	315	59%	0.29	0.91	0.006
	SI	239	145	2702	368	55%	0.61	0.99	0.006
	genome	237	166	5070	660	57%	0.7	0.99	0.006

Clade	Locus	k	average linkage among loci|D'|	R/gene	R between adjacent sites	Minimum # of Recombination Events	Tajima's D	Tajima's D Statistical Significance
B. canariense	dnak	0.0115	0.986	0.0010	0.0000	4	-2.19	$p < 0.01$
	glnII	0.0077	0.991	0.0010	0.0000	7	-2.28	$p < 0.01$
	ITS	0.0030	0.999	0.0010	0.0000	5	-2.45	$p < 0.01$
	recA	0.0068	0.987	0.0010	0.0000	7	-2.25	$p < 0.01$
	nifD	0.0084	0.994	2.3000	0.0030	5	-1.88	$p < 0.05$
	nodDA	0.0064	0.994	0.0010	0.0000	11	-2.40	$p < 0.01$
	nodZ	0.0030	0.996	0.0010	0.0000	2	-2.44	$p < 0.01$
	nolL	0.0046	0.991	0.0010	0.0000	4	-2.47	$p < 0.001$
	CHR	0.0055	0.990	0.001	0.0000	24	-2.4	$p < 0.01$
	SI	0.0059	0.978	0.300	0.0001	25	-2.37	$p < 0.01$
	genome	0.0056	0.972	0.001	0.0000	49	-2.40	$p < 0.01$

Clade	Locus	# Isolates	# haplotypes	# nt sequenced	# variable sites	GC content	Strain Richness	Hd	π
B. japonicum	dnak	27	9	209	42	66%	0.33	0.51	0.024
	glnII	27	5	560	47	63%	0.19	0.28	0.009
	ITS	27	8	1228	96	53%	0.30	0.66	0.010
	recA	27	7	414	36	69%	0.26	0.66	0.011
	nifD	27	14	758	57	57%	0.52	0.78	0.011
	nodDA	27	8	868	38	55%	0.30	0.60	0.005
	nodZ	27	13	429	32	56%	0.48	0.89	0.015
	nolL	27	4	650	23	52%	0.15	0.21	0.003
	CHR	27	16	2411	221	59%	0.59	0.92	0.011
	SI	27	19	2705	150	55%	0.7	0.93	0.008
	genome	27	21	5116	371	57%	0.78	0.95	0.009

Clade	Locus	k	average linkage among loci|D'|	R/gene	R between adjacent sites	Minimum # of Recombination Events	Tajima's D	Tajima's D Statistical Significance
B. japonicum	dnak	0.0239	0.980	0.0010	0.0000	6	-2.18	$p < 0.01$
	glnII	0.0089	1.000	0.0010	0.0000	0	-2.27	$p < 0.01$
	ITS	0.0096	1.000	0.0010	0.0000	0	-2.10	$p < 0.05$
	recA	0.0111	0.982	0.0010	0.0000	2	-2.10	$p < 0.05$
	nifD	0.0109	0.987	0.5000	0.0007	6	-1.75	$0.10 > p > 0.05$
	nodDA	0.0054	1.000	0.0010	0.0000	0	-2.01	$p < 0.05$
	nodZ	0.0149	1.000	10.4000	0.0243	0	-1.03	$p > 0.10$
	nolL	0.0031	1.000	0.0010	0.0000	0	-2.41	$p < 0.01$
	CHR	0.0108	0.990	0.001	0.0000	9	-2.22	$p < 0.01$
	SI	0.0080	0.983	2.000	0.0007	9	-1.84	$p < 0.05$
	genome	0.0093	0.988	0.001	0.0000	18	-2.08	$p < 0.05$

Clade	Locus	# Isolates	# haplotypes	# nt sequenced	# variable sites	GC content	Strain Richness	Hd	π
	dnak	54	10	209	22	68%	0.19	0.66	0.011
	glnII	54	7	560	49	63%	0.13	0.56	0.008
	ITS	17	6	1307	105	55%	0.35	0.71	0.018
	recA	54	10	414	38	67%	0.19	0.73	0.011
	nifD	50	12	758	28	56%	0.24	0.88	0.004
B. nov. I	nodDA	50	19	868	68	56%	0.38	0.89	0.018
	nodZ	52	8	429	21	56%	0.15	0.75	0.011
	nolL	52	16	650	65	53%	0.31	0.83	0.011
	CHR	17	11	2490	202	59%	0.65	0.88	0.017
	SI	46	30	2705	168	55%	0.65	0.97	0.011
	genome	17	14	5195	350	57%	0.82	0.97	0.014

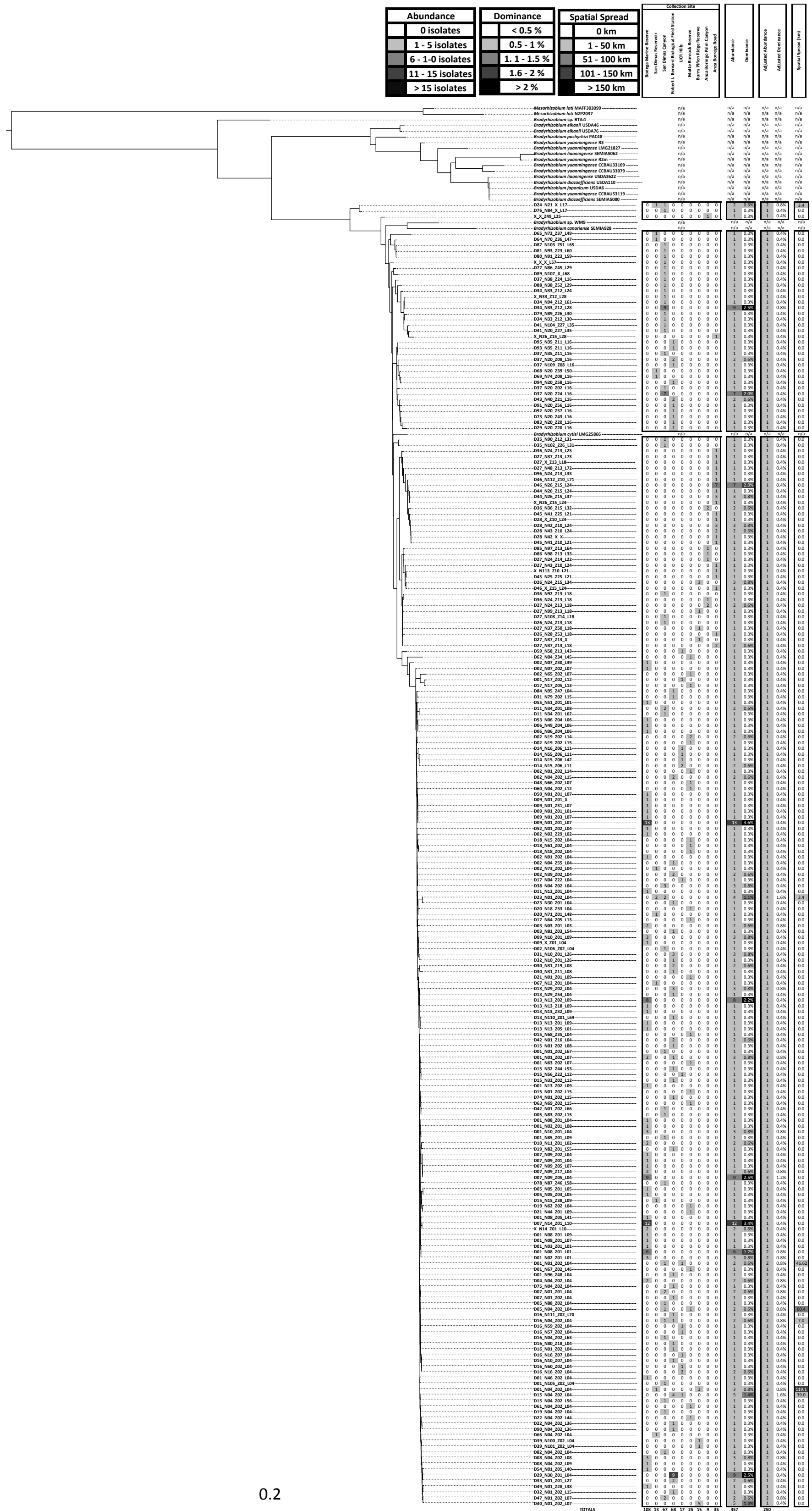
Clade	Locus	k	average linkage among loci|D'|	R/gene	R between adjacent sites	Minimum # of Recombination Events	Tajima's D	Tajima's D Statistical Significance
	dnak	0.0110	0.996	0.0010	0.0000	2	-1.64	0.10 > p > 0.05
	glnII	0.0080	0.999	0.0010	0.0000	1	-2.06	p < 0.05
	ITS	0.0183	1.000	0.0010	0.0000	0	-0.98	p > 0.10
	recA	0.0116	0.995	0.0010	0.0000	1	-1.44	p > 0.10
	nifD	0.0040	0.998	0.0010	0.0000	1	-1.73	0.10 > p > 0.05
B. nov. I	nodDA	0.0177	0.983	2.5000	0.0029	4	0.00	p > 0.10
	nodZ	0.0112	1.000	2.7000	0.0063	0	-0.02	p > 0.10
	nolL	0.0110	0.998	0.001	0.0000	1	-1.76	0.10 > p > 0.05
	CHR	0.0048	0.996	0.001	0.0000	2	-1.27	p > 0.10
	SI	0.0113	0.963	3.600	0.0013	8	-0.76	p > 0.10
	genome	0.0144	0.967	0.001	0.0000	9	-1.22	p > 0.10

Clade	Locus	# Isolates	# haplotypes	# nt sequenced	# variable sites	GC content	Strain Richness	Hd	π
B. yuanmingense	dnak	24	8	209	20	67%	0.33	0.66	0.016
	glnII	24	13	560	69	63%	0.54	0.86	0.023
	ITS	24	9	1113	123	52%	0.38	0.66	0.022
	recA	24	10	414	64	66%	0.42	0.75	0.033
	nifD	22	9	758	43	57%	0.41	0.71	0.012
	nodDA	23	11	859	62	55%	0.48	0.74	0.015
	nodZ	22	6	429	22	55%	0.27	0.59	0.012
	nolL	24	12	650	61	52%	0.50	0.83	0.021
	CHR	23	12	2313	198	59%	0.52	0.85	0.019
	SI	21	13	2696	185	55%	0.62	0.83	0.015
	genome	21	18	5009	362	57%	0.86	0.97	0.017

Clade	Locus	k	average |D'|	R/gene	R between adjacent sites	Minimum # of Recombination Events	Tajima's D	Tajima's D Statistical Significance
B. yuanmingense	dnak	0.0163	0.977	0.5000	0.0024	3	-1.33	p > 0.10
	glnII	0.0227	0.989	0.0010	0.0000	5	-1.41	p > 0.10
	ITS	0.0222	0.961	0.0010	0.0000	7	-1.40	p > 0.10
	recA	0.0329	0.964	1.3000	0.0031	10	-0.98	p > 0.10
	nifD	0.0121	1.000	0.0010	0.0000	0	-0.94	p > 0.10
	nodDA	0.0151	0.998	1.3000	0.0015	2	-1.04	p > 0.10
	nodZ	0.0124	0.995	0.0010	0.0000	2	-0.74	p > 0.10
	nolL	0.0214	0.985	0.5000	0.0008	7	-0.58	p > 0.10
	CHR	0.0190	0.980	0.001	0.0000	17	-0.95	p > 0.10
	SI	0.0153	0.972	0.4000	0.0001	9	-0.92	p > 0.10
	genome	0.0170	0.950	6.2000	0.0012	22	-0.82	p > 0.10

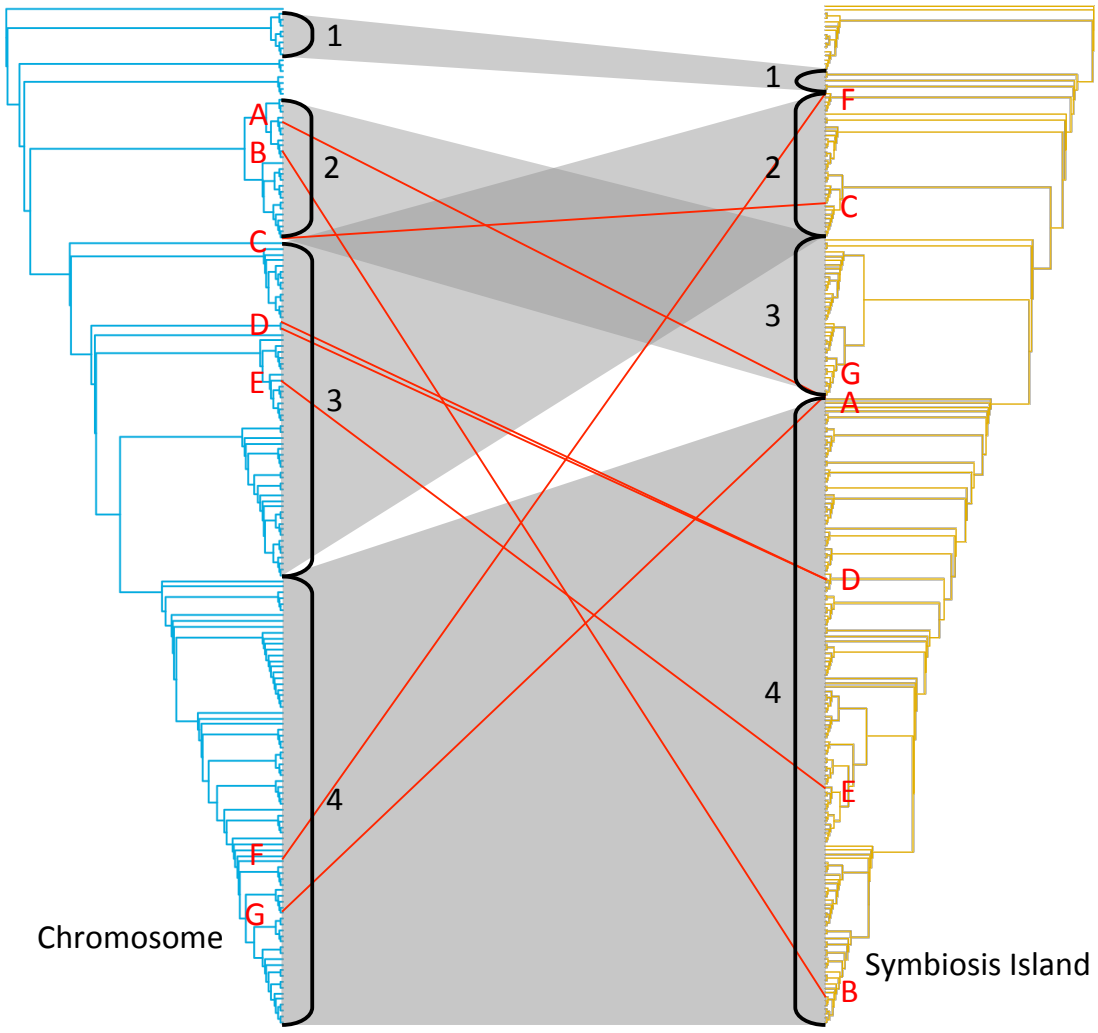
S6 Figure – Phylogram of concatenated symbiosis island loci (SI)

Phylogram of concatenated SI loci (*nifD*, *nodD-A*, *nodZ,* and *nolL*) reconstructed in PhyML 3.0 with associated heatmaps for abundance, dominance, and spatial range. Collection site heat map shows number of times the haplotype was isolated at that site. Abundance heat map indicates the number of times the haplotype was isolated out of a total of 357 isolates for which symbiosis island loci were amplified. Dominance heat map indicates the percentage a haplotype represents out of the total 357 isolates. Adjusted abundance and dominance refer to the same calculations, but only counting identical haplotypes from unique GPS locations. Spatial range is the maximum distance between collection sites as calculated from the geographic midpoint of each collection site.


Abundance
0 isolates
1 - 5 isolates
6 - 1-0 isolates
11 - 15 isolates
> 15 isolates
Dominance
< 0.5 %
0.5 - 1 %
1. 1 - 1.5 %
1.6 - 2 %
> 2 %
Spatial Spread
0 km
1 - 50 km
51 - 100 km
101 - 150 km
> 150 km
0.2

S7 Figure – TreeMap cophylogeny tanglegram of CHR and SI haplotypes.

Genome region cladograms reconstructed with PhyML using CHR loci (*dnaK*, *glnII*, ITS, and *recA*; blue tree) and SI loci (*nifD*, *nodD-A*, *nodZ*, and *nolL*; yellow tree). Patterns of associations between CHR and SI haplotypes are indicated by grey bars with probable SI transfer events indicated with red lines and letters. CHR clades and SI lineages that share patterns of genome region association are bracketed and numbered.



S8 – Dominant and epidemic haplotypes for CHR, SI, and the whole genome. Dominant haplotypes are assigned within sites, must be collected 5 times, and constitute at least 10% of the isolates. Among these, epidemic haplotypes must be dominant at a site and collected at another site that is >10km away as indicated by spatial range (marked with an asterisk). N is the number of total isolates per site. Abundance is the number of times that a haplotype was collected. Km is the total range of a haplotype across all collection locales. Anza Palm Canyon and San Dimas Reservoir had no dominant haplotypes.

Collection Sites	N	Loci	Haplotype	Abundance	Km
Anza - Roadside	36	CHR	K14_G58_X_R27	17	0
Anza - Roadside	36	CHR	K14_G58_X_R34	7	0
Anza - Roadside	36	SI	D46_N26_Z15_L24	7	0
Anza - Roadside	36	Genome	K14_G58_X_R34_D46_N26_Z15_L24	5	0
Bodega Marine Reserve	108	CHR	K04_G05_I07_R02	19	0
Bodega Marine Reserve	108	CHR	K02_G03_I01_R01	18	0
Bodega Marine Reserve	108	CHR	K01_G03_I01_R01*	14	724
Bodega Marine Reserve	108	CHR	K01_G01_I01_R01	13	0
Bodega Marine Reserve	108	CHR	K01_G03_I05_R01	12	0
Bodega Marine Reserve	108	SI	D09_N01_Z01_L07	13	0
Bodega Marine Reserve	108	SI	D07_N14_Z01_L10	12	0
Bodega Marine Reserve	108	Genome	K04_G05_I07_R02_D09_N01_Z01_L07	13	0
Bodega Marine Reserve	108	Genome	K02_G03_I01_R01_D07_N14_Z01_L10	12	0
Burns Piñon Ridge Reserve	15	CHR	K01_G03_I20_R01*	6	120
Burns Piñon Ridge Reserve	15	SI	D40_N01_Z02_L07	5	0
Motte Rimrock Reserve	25	CHR	K01_G03_I01_R01*	9	724
Bernard Field Station	68	CHR	K01_G03_I01_R01*	18	724
Bernard Field Station	68	SI	D29_N30_Z01_L04	9	0
Bernard Field Station	68	CHR	K01_G03_I01_R01_D29_N30_Z01_L04	7	0
San Dimas Canyon	67	CHR	K08_G36_I27_R35	9	0
San Dimas Canyon	67	CHR	K01_G03_I01_R01*	7	724
San Dimas Canyon	67	CHR	K11_G71_I21_R39	7	0
San Dimas Canyon	67	SI	D34_N33_Z12_L28	9	0
San Dimas Canyon	67	SI	D37_N20_Z24_L16	7	0
San Dimas Canyon	67	Genome	K11_G71_I21_R39_D37_N20_Z24_L16	7	0
UC Riverside	17	CHR	K01_G03_I01_R01*	11	724
UC Riverside	17	CHR	K01_G03_I20_R01	10	120

S9 Table – Intrapopulation and interpopulation parameters

Within population analyses are separated by genome region. Mean F_{ST} was calculated between collection sites for each genome region using the Weir-Cockerham method and excluded the following samples: 11LoS11_3 (chromosome haplotype: K31_G212_I89_R204), 11LoS20_2 (chromosome haplotype: K34_G219_I100_R209), 11LoS31_5 (chromosome haplotype: K39_G223_I108_R213), 12LoS4_2 (chromosome haplotype: K42_G226_I118_R218).

	Intrapopulation							Interpopulation
	Collection Site	**# Isolates**	**π**	***h***	**H_d**	**k**		**CHR Mean F_{ST}**
chromosome	**Anza Borrego (Palm Canyon)**	9	0.050	8	0.97	0.0548	**chromosome**	0.25
	Anza Borrego (Road)	36	0.002	12	0.75	0.0009		0.09
	Bodega Marine Reserve	108	0.009	19	1.00	0.0077		0.10
	Burns Piñon Ridge	15	0.051	6	0.79	0.0439		0.83
	Motte Rimrock Reserve	25	0.002	11	0.78	0.0018		0.10
	Bernard Field Station	68	0.029	34	0.92	0.0308		0.10
	San Dimas Canyon	67	0.059	37	0.95	0.0317		0.13
	San Dimas Reservoir	13	0.059	11	0.97	0.0602		0.10
	UC Riverside	17	0.012	7	0.60	0.0127		0.10
symbiosis island	**Anza Borrego (Palm Canyon)**	9	0.013	6	0.89	0.0052	**symbiosis island**	0.13
	Anza Borrego (Road)	35	0.008	26	0.97	0.0033		0.13
	Bodega Marine Reserve	108	0.002	12	0.64	0.0004		0.07
	Burns Piñon Ridge	15	0.016	8	0.87	0.0126		0.06
	Motte Rimrock Reserve	25	0.009	24	1.00	0.0089		0.07
	Bernard Biological Field Station	68	0.014	46	0.98	0.0149		0.06
	San Dimas Canyon	67	0.034	26	0.91	0.0082		0.08
	San Dimas Reservoir	13	0.032	12	0.99	0.0270		0.06
	UC Riverside	17	0.009	14	0.97	0.0093		0.07