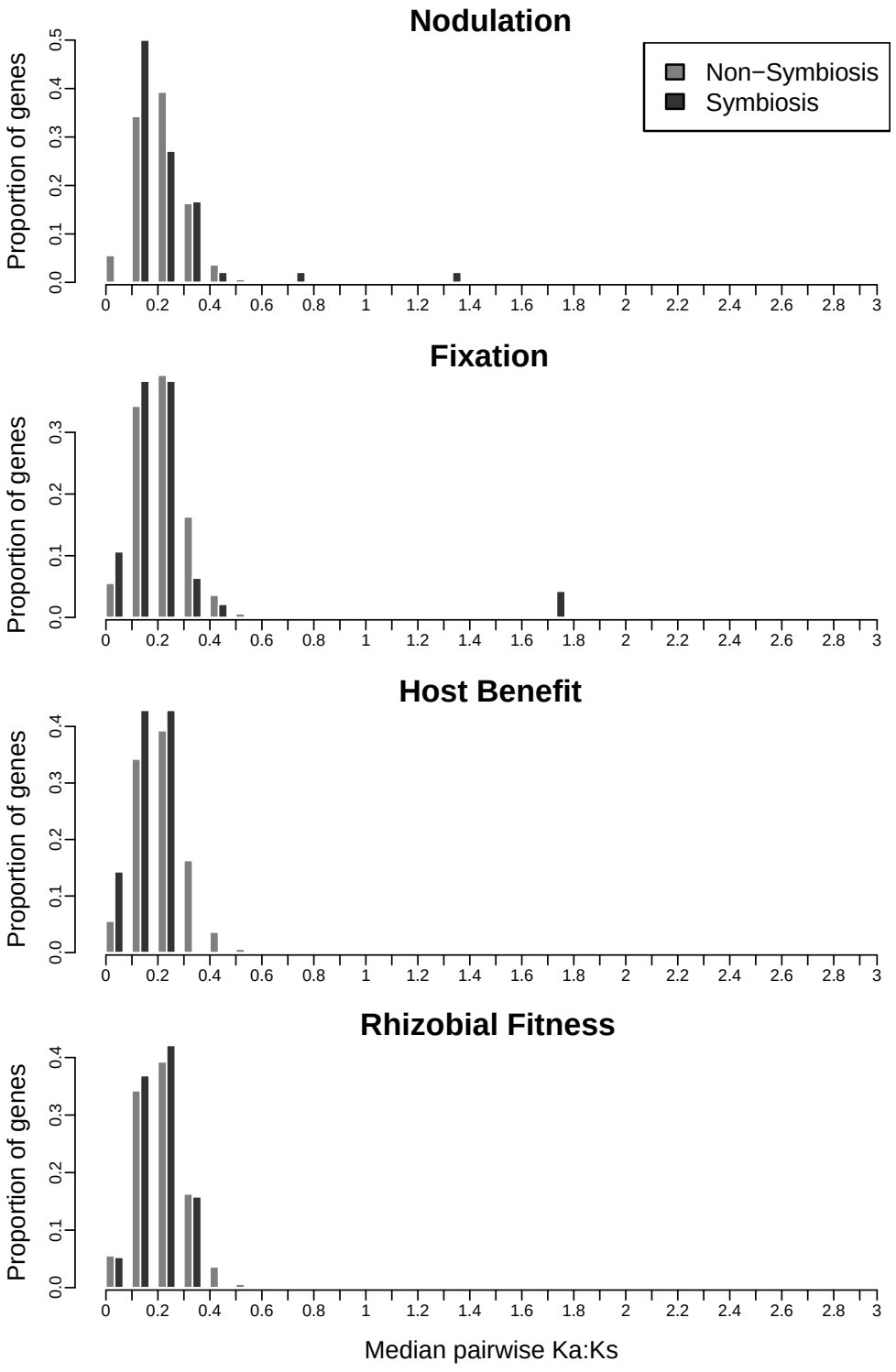
Supplemental Material from:

Comparative genomics reveals high rates of horizontal transfer and strong purifying selection on rhizobial symbiosis genes

Brendan Epstein, Peter Tiffin



Supplementary Figure S1: Distributions of median pairwise Ka:Ks for symbiosis genes (dark gray) and non-symbiosis genes (light gray). Only genes with ≧ 3 valid comparisons were plotted.

Table S2: References for gene annotation searches. We chose gene names and terms that are associated with the “classic” fixation and nodulation genes—those that have long been recognized as having a direct role in symbiosis. References are intended to provide a general review of nodulation and nitrogen-fixation genes in rhizobia, including lists of important genes, which informed our strategy for identifying nodulation and fixation genes that are directly involved in legume-rhizobia symbiosis.

Triplett & Sadowsky 1992	Review of nodulation genes	Triplett E, Sadowsky M. 1992 Genetics of competition for nodulation of legumes. Annual review of microbiology 46, 399–428. (doi:10.1146/annurev.mi.46.100192.002151)
Cooper 2007	Review of nodulation genes	Cooper JE. 2007 Early interactions between legumes and rhizobia: disclosing complexity in a molecular dialogue. Journal of Applied Microbiology 103, 1355–1365. (doi:10.1111/j.1365-2672.2007.03366.x)
Niner & Hirsch 1998	Review of genes indirectly involved in symbiosis	Niner BM, Hirsch AM. 1998 How Many Rhizobium Genes, in Addition to nod, nif/fix, and exo, are Needed for Nodule Development and Function? Symbiosis 24, 51-102.
Masson-Boivin et al. 2009	Review of nodulation and fixation genes	Masson-Boivin C, Giraud E, Perret X, Batut J. 2009 Establishing nitrogen-fixing symbiosis with legumes: how many rhizobium recipes? Trends in Microbiology 17, 458–466. (doi:10.1016/j.tim.2009.07.004)
Geddes et al. 2020	Review of nodulation and fixation genes and genomic arrangement	Geddes BA, Kearsley J, Morton R, diCenzo GC, Finan TM. 2020 Chapter Eight - The genomes of rhizobia. In Advances in Botanical Research (eds P Frendo, F Frugier, C Masson-Boivin), pp. 213–249. Academic Press. (doi:10.1016/bs.abr.2019.09.014)
Staehelin & Krishnan 2015	Review of *nop* genes (nodulation)	Staehelin C, Krishnan HB. 2015 Nodulation outer proteins: double-edged swords of symbiotic rhizobia. Biochemical Journal 470, 263–274. (doi:10.1042/BJ20150518)
De Meyer et al. 2016	Nodulation & fixation genes in Betaproteobacteria	De Meyer SE et al. 2016 Symbiotic Burkholderia Species Show Diverse Arrangements of nif/fix and nod Genes and Lack Typical High-Affinity Cytochrome cbb3 Oxidase Genes. MPMI 29, 609–619. (doi:10.1094/MPMI-05-16-0091-R)

Table S4: Five part table (A-E) with all statistics and tests. A & B) Means, medians of each statistic for each gene category. C & D) The proportion of randomly sampled datasets that had a median (C) or mean (D) greater than the median or mean, respectively, of the symbiosis genes. Larger numbers in C & D indicate more evidence that the symbiosis genes have a value less than that of non-symbiosis genes. E) The p-values of t-tests comparing symbiosis gene categories to each other. The pairwise median Ka:Ks calculated for all genes and for only genes with more than 2 valid comparisons, because the values for genes with small number of comparisons can be driven by a single sequence. The delta statistic was also calculated for all genes, with different numbers of copies treated as different categories, or with only genes that had more than two sequences (because delta is based on phylogenetic signal, which makes more sense with a larger number of sequences), treating one or more copies as "present". The p-values are from t-tests comparing symbiosis genes to all other genes or different categories of symbiosis genes to each other. Random gene samples were chosen either from all other genes ("All genes") or from genes that were present in the same number of strains and had a similar copy number to the symbiosis genes ("Matched").

A: Means

	Family size	Genomes	Copies per genome	duplication	transfer	Pairwise median Ka/Ks all genes	Pairwise median Ka/Ks genes with > 2 comparisons	R^2	delta, all genes, copy-number as categorical trait	delta, genes with > 2 sequences, presence / absence only
Non-Symbiosis	5.9	5.1	1.07	0.02	0.12	0.24	0.23	0.62	34.0	25.6
Nodulation	14.9	11.0	1.22	0.02	0.23	0.26	0.26	0.32	14.2	25.8
Fixation	23.0	17.4	1.26	0.04	0.14	0.26	0.26	0.49	4.1	17.1
Benefit	25.2	15.7	1.37	0.03	0.20	0.17	0.17	0.33	2.0	1.4
Fitness	27.1	16.3	1.45	0.01	0.21	0.21	0.22	0.28	7.8	12.3

B: Medians

Non-Symbiosis	2	2	1	10^{-7}	0.08	0.22	0.22	0.76	2.9	3.3
Nodulation	8	6	1	10^{-7}	0.24	0.20	0.20	0.23	2.1	1.6
Fixation	27	22	1	10^{-7}	0.14	0.21	0.20	0.54	1.9	2.2
Benefit	19.5	19	1.08	0.002	0.20	0.15	0.15	0.29	1.8	1.1
Fitness	18	17	1.13	10^{-7}	0.18	0.22	0.23	0.19	1.8	1.9

C: Proportion of randomly sampled datasets with a median value > median value of empirical estimate (mean for copies per genome)

random samples vs.:	Family size	Genomes	Copies per genome	duplication	transfer	Pairwise median Ka/Ks all genes	Pairwise median Ka/Ks genes with > 2 comparisons	R^2	delta, all genes, copy-number as categorical trait	delta, genes with > 2 sequences, presence / absence only
All genes:										
Nodulation	0	0	0.009	0.011	0.001	0.843	0.86	0.999	1	0.999
Fixation	0	0	0.003	0.017	0.141	0.752	0.826	0.873	1	0.913
Benefit	0	0	0.015	0.203	0.138	0.949	0.934	0.881	0.965	0.971
Fitness	0	0	0.011	0.119	0.126	0.528	0.381	0.985	0.999	0.897
Matched:										
Nodulation				0.002	0	0.99	0.997	0.997	0.991	0.999
Fixation				0	0	0.979	0.992	0.886	0.999	0.992
Benefit				0.067	0.006	0.992	0.996	0.965	0.891	0.959
Fitness				0.008	0.041	0.904	0.734	0.855	0.954	0.945

D: Proportion of randomly sampled datasets with a mean value > mean value of empirical estimate

random samples vs.:	Family size	Genomes	Copies per genome	duplication	transfer	Pairwise median Ka/Ks all genes	Pairwise median Ka/Ks genes with > 2 comparisons	R^2	delta, all genes, copy-number as categorical trait	delta, genes with > 2 sequences, presence / absence only
All genes:										
Nodulation	0	0	0.009	0.299	0.003	0.171	0.128	1	0.983	0.444
Fixation	0	0	0.003	0.123	0.286	0.145	0.104	0.947	1	0.693
Benefit	0.001	0	0.015	0.191	0.157	0.924	0.918	0.935	0.992	0.964
Fitness	0.001	0	0.011	0.347	0.078	0.725	0.606	0.997	0.927	0.572
Matched:										
Nodulation				0.631	0	0.207	0.209	1	0.791	0.564
Fixation				0	0	0.041	0.034	0.982	0.998	0.876
Benefit				0.066	0.005	0.996	0.998	0.995	0.949	0.947
Fitness				0.161	0.001	0.993	0.982	0.995	0.71	0.594

E: P-values from t-tests comparing symbiosis gene categories

	Family size	Genomes	Copies per genome	duplication	transfer	Pairwise median Ka/Ks all genes	Pairwise median Ka/Ks genes with > 2 comparisons	R^2	delta, all genes, copy-number as categorical trait	delta, genes with > 2 sequences, presence / absence only
nodulation vs. fixation	0.01	<0.001	0.55	0.17	< 0.001	0.94	0.9	0.007	0.12	0.56
benefit vs. fitness	0.86	0.89	0.8	0.3	0.75	0.16	0.07	0.7	0.32	0.21
nodulation vs. benefit	0.29	0.27	0.54	0.53	0.36	0.01	0.02	0.95	0.06	0.04
nodulation vs fitness	0.05	0.03	0.17	0.38	0.68	0.14	0.29	0.56	0.45	0.34
fixation vs. benefit	0.82	0.68	0.67	0.8	0.08	0.06	0.07	0.22	0.06	0.11
fixation vs. fitness	0.51	0.64	0.29	0.06	0.07	0.28	0.42	0.005	0.52	0.7