SUPPLEMENTARY METHODS

Phylogeny construction.

Our in-house custom pipeline was used for phylogeny construction (diCenzo et al. 2017). This pipeline involved the use of GNU Bash v4.3.48(1), Perl v5.22.1, and HMMER v3.1b2 (Eddy 2009). The protein fasta files for all α-proteobacteria (822 proteomes) and β-proteobacteria (1000 proteomes) available through the National Center for Biotechnology Information (NCBI) annotated as 'Complete' or 'Chromosome' were downloaded (as of November 7th 2017). One randomly chosen, representative proteome for each species was selected and used for the phylogenetic analysis. The MarkerScanner script of the AMPHORA2 pipeline (Wu and Scott 2012) was run to extract orthologs of 31 highly conserved proteins from each proteome; only the 23 proteins (Frr, NusA, RplB, RplC, RplD, RplK, RplL, RplM, RplN, RplP, RplS, RplT, RpmA, RpoB, RpsB, RpsC, RpsE, RpsI, RpsJ, RpsK, RpsM, RpsS, Tsf) found in single-copy in each proteome and present in at least 95% of the analyzed proteomes were kept. Each set of orthologs were aligned using Mafft v7.310 (Katoh and Standley 2013) with the 'localpair' option and with 10 threads. Alignments were then trimmed using TrimAl v1.2rev59 (Capella-Gutiérrez et al. 2009) and the 'automated1' option, following which the alignments were concatenated. Finally, a maximum likelihood phylogeny was built with the RAxML v8.2.10 algorithm (Stamatakis 2014) on the Cipres webserver (Miller et al. 2010) using the LG amino acid substitution model and the CAT rate heterogeneity. The final tree represents the bootstrap best tree following 108 bootstrap replicates. The final tree was visualized using FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Identification of the NodABC and NifHDK symbiotic proteins.

A custom pipeline based on the use of hidden Markov models (HMM) was built to identify orthologs of interest. This pipeline involved the use of GNU Bash v4.3.48(1), Perl v5.22.1, and HMMER v3.1b2 (Eddy 2009). All 1,822 proteomes for the α-proteobacteria and β-proteobacteria were combined as a single fasta file. The complete Pfam-A v31.0 (16,712 HMMs) and TIGERFAM v15.0 (4,488 HMMs) HMM databases (Haft et al. 2013, Finn et al. 2016) were downloaded, hmmconvert used to ensure consistent formatting, the two databases combined into a single HMM database, and hmmpress used to make a searchable database. Additionally, the HMM seed alignments were downloaded from the TIGRFAM database (Haft et al. 2013) for NodA (TIGR04245), NodB (TIGR04243), NodC (TIGR04242), NifH (TIGR01287), NifD (TIGR01282), and NifK (TIGR01286).

For each HMM seed alignment, a HMM was built using hmmpress, and the output was then searched against the complete set of proteins using hmmsearch. The output was parsed, and the amino acid sequences for each of the hits (regardless of e-value) were collected. Each set of sequences were then searched against the combined HMM database using hmmscan, and the output parsed to identify the top scoring HMM hit for each query protein. Proteins were annotated as follows: NodA if the top hit was TIGR04245 (TIGRFAM) or NodA (Pfam); NodB if the top hit was TIGR04243 (TIGRFAM); NodC if the top hit was TIGR04242 (TIGRFAM); NifH if the top hit was TIGR01287 (TIGRFAM) or Fer4_NifH (Pfam); NifD if the top hit was TIGR01282 (TIGRFAM), TIGR01860 (TIGRFAM), or TIGR01861 (TIGRFAM); NifK if the top hit was TIGR02932 (TIGRFAM), TIGR02931 (TIGRFAM), or TIGR01286 (TIGRFAM). Each of the 1,822 proteomes included in the analysis were then annotated as having NodABC if each of the NodA, NodB, and NodC proteins were detected in the proteome, and as having

NifHDK if each of the NifH, NifD, and NifK proteins were detected in the proteome. A species was said to have NodABC if at least one strain had NodABC; a species was said to have NifHDK if at least one strain had NifHDK; and a species was said to have both NodABC and NifHDK if at least one strain had NodABC and NifHDK.

Pangenome analysis.

The pangenome of 20 strains of *Sinorhizobium meliloti* was calculated as follows. The nucleotide fasta files for 20 of the 22 *S. meliloti* genomes with a status of 'Complete' or 'Chromosome' were downloaded from NCBI (November 30th 2017). Of the two excluded sequences, one was a second sequence of the strain Rm41, and the other was Rm2011, which was excluded as it is derived from the same original nodule isolate of Rm1021 (included in the analysis). Each fasta file was annotated using Prokka v1.12-beta (Seemann 2014) to ensure consistent annotation. Prokka was called with the 'fast', 'norrna', and 'notrna' options. The Genbank flat files (gff) returned by Prokka were used for pangenome construction using Roary v3.7.0 (Page et al. 2015) called with the '-e' and '--mafft', and with eight threads. A phylogeny was produced from the core gene alignment return by Roary using FastTree v2.1.4 SSE3 (Price et al. 2010) using the 'nt' and 'gtr' options.

Metabolic modelling procedures.

The iGD726 metabolic model (diCenzo et al. 2018), encompassing the core metabolism of *S. meliloti*, was used throughout. For growth with glucose, the lower bound of the glucose exchange reaction (EX_cpd00027_e0) was set to -2.41 mmol/hr/gCDW. Prior to determining the flux distribution for growth with glucose, reactions rxn00159, rxn00161, and rxn00346 were turned off; rxn00346 was turned back on prior to testing reaction essentiality. For growth with succinate, the lower bound of the succinate exchange reaction (EX_cpd00036_e0) was set to -6.252 mmol/hr/gCDW. All other exchange reactions had a lower bound of -1000 mmol/hr/gCDW. Prior to determining the flux distribution for growth with succinate, reactions rxn00250 and rxn00346 were turned off; rxn00346 was turned back on prior to testing reaction essentiality.

All simulations were performed in Matlab 2017a (Mathworks) with scripts from the Cobra Toolbox (downloaded May 12, 2017 from the openCOBRA repository) (Schellenberger et al. 2011), and using the Gurobi 7.0.2 solver (www.gurobi.com), the SBMLToolbox 4.1.0 (Keating et al. 2006), and libSBML 5.15.0 (Bornstein et al. 2008). Flux Balance Analysis (FBA) was performed as implemented in the *optimizeCbModel* function. Optimal flux distributions were determined with FBA. Reaction essentiality was determined by iteratively deleting one reaction from the model and using FBA to predict the optimal growth rate of reaction-deleted model; if the model could not produce biomass, the reaction was said to be essential.

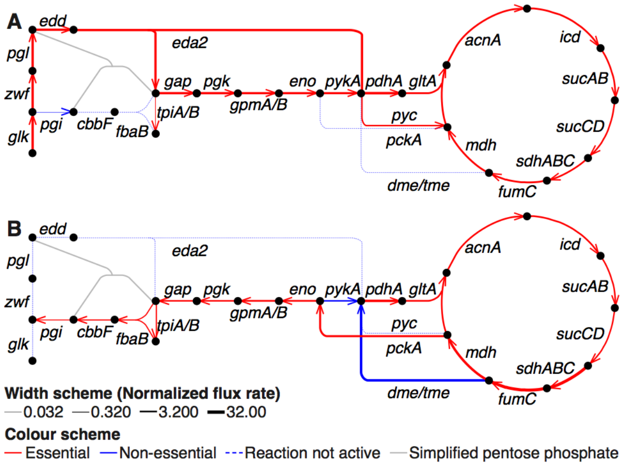

Figure S1. *In silico* predicted flux distributions and reaction essentialities. Example data obtained through metabolic modelling. Results from *in silico* metabolic modelling of *S. meliloti* metabolism using the iGD726 metabolic reconstruction are shown (diCenzo et al. 2018). Growth was simulated using (**A**) glucose or (**B**) succinate as the sole carbon source. The pathways represent central carbon metabolism, with the exception of the pentose phosphate pathway (for simplicity). The width of the line represents the amount of flux through the reaction; a doubling in the width corresponds to a ten-fold flux increase. Arrows indicate the direction of flux. Colours indicate if the reaction is essential (red) or non-essential (blue). Genes associated with reactions are based on the information present in iGD726, but they are not necessarily comprehensive. See the Supplementary Materials for details on the modelling procedures used. A version without gene names is provided as Figure 7 in manuscript.

SUPPLEMENTARY REFERENCES

Bornstein, B.J., Keating, S.M., Jouraku, A., and Hucka, M. 2008. LibSBML: an API library for SBML. Bioinformatics **24**(6): 880–881. doi:10.1093/bioinformatics/btn051.

Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics **25**(15): 1972–1973. doi:10.1093/bioinformatics/btp348.

diCenzo, G.C., Benedict, A.B., Fondi, M., Walker, G.C., Finan, T.M., Mengoni, A., and Griffitts, J.S. 2018. Robustness encoded across essential and accessory replicons in an ecologically versatile bacterium. PLOS Genet. **14**(4): e1007357. doi:10.1101/209916.

diCenzo, G.C., Zamani, M., Ludwig, H.N., and Finan, T.M. 2017. Heterologous complementation reveals a specialized activity for BacA in the *Medicago-Sinorhizobium meliloti* symbiosis. Mol. Plant Microbe Interact. **30**(4): 312–324. doi:10.1094/MPMI-02-17-0030-R.

Eddy, S.R. 2009. A new generation of homology search tools based on probabilistic inference. Genome Inform. **23**(1): 205–211. doi:10.1142/9781848165632_0019.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., and Bateman, A. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. **44**(D1): D279–D285. doi:10.1093/nar/gkv1344.

Haft, D.H., Selengut, J.D., Richter, R.A., Harkins, D., Basu, M.K., and Beck, E. 2013. TIGRFAMs and genome properties in 2013. Nucleic Acids Res. **41**(Database issue): D387–D395. doi:10.1093/nar/gks1234.

Katoh, K., and Standley, D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. **30**(4): 772–780. doi:10.1093/molbev/mst010.

Keating, S.M., Bornstein, B.J., Finney, A., and Hucka, M. 2006. SBMLToolbox: an SBML toolbox for MATLAB users. Bioinformatics **22**(10): 1275–1277. doi:10.1093/bioinformatics/btl111.

Miller, M.A., Pfeiffer, W., and Schwartz, T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. IEEE, New Orleans, LA. pp. 1–8. doi:10.1109/GCE.2010.5676129.

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., Fookes, M., Falush, D., Keane, J.A., and Parkhill, J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics **31**(22): 3691–3693. doi:10.1093/bioinformatics/btv421.

Price, M.N., Dehal, P.S., and Arkin, A.P. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLOS One **5**(3): e9490. doi:10.1371/journal.pone.0009490.

Schellenberger, J., Que, R., Fleming, R.M.T., Thiele, I., Orth, J.D., Feist, A.M., Zielinski, D.C., Bordbar, A., Lewis, N.E., Rahmanian, S., Kang, J., Hyduke, D.R., and Palsson, B.Ø. 2011. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat Protoc **6**(9): 1290–1307. doi:10.1038/nprot.2011.308.

Seemann, T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics **30**(14): 2068–2069. doi:10.1093/bioinformatics/btu153.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**(9): 1312–1313. doi:10.1093/bioinformatics/btu033.

Wu, M., and Scott, A.J. 2012. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics **28**(7): 1033–1034. doi:10.1093/bioinformatics/bts079.