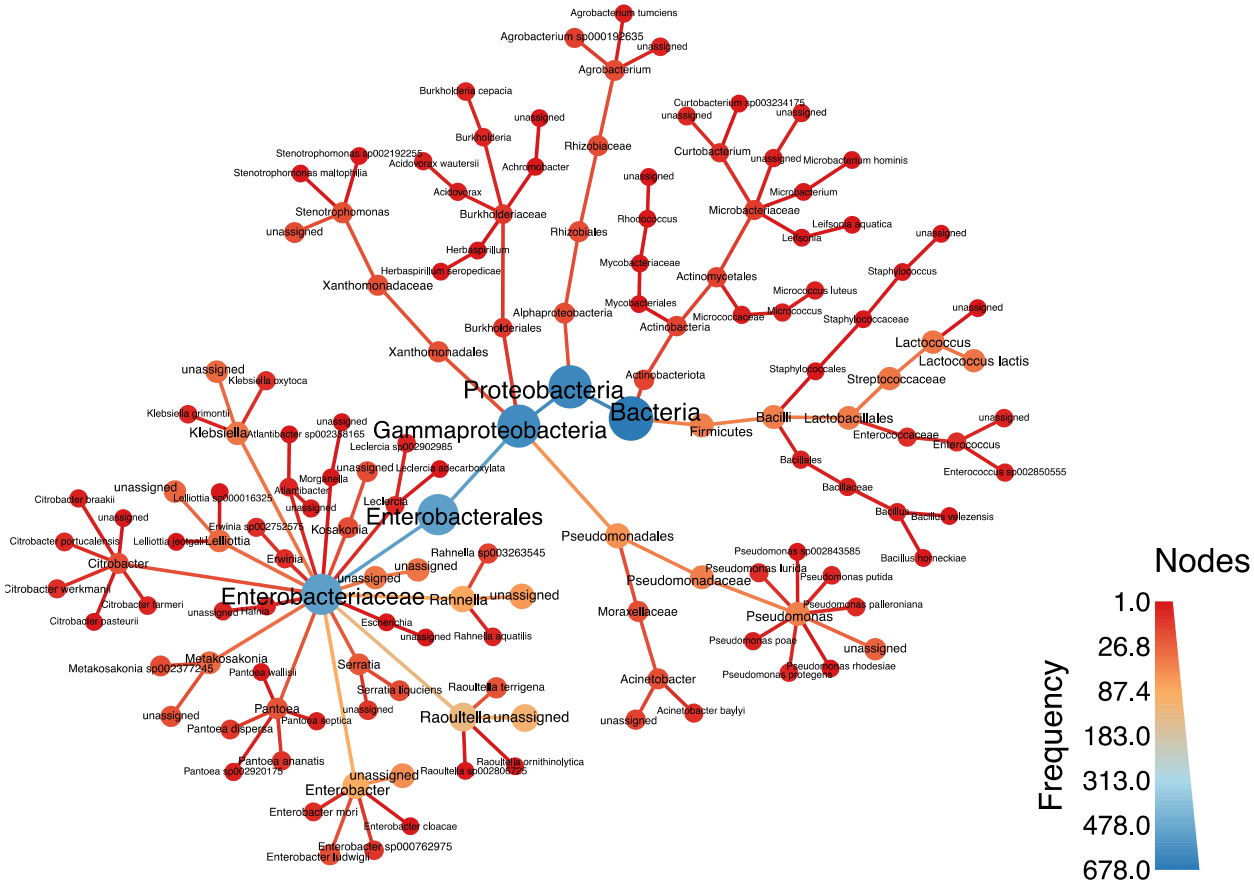
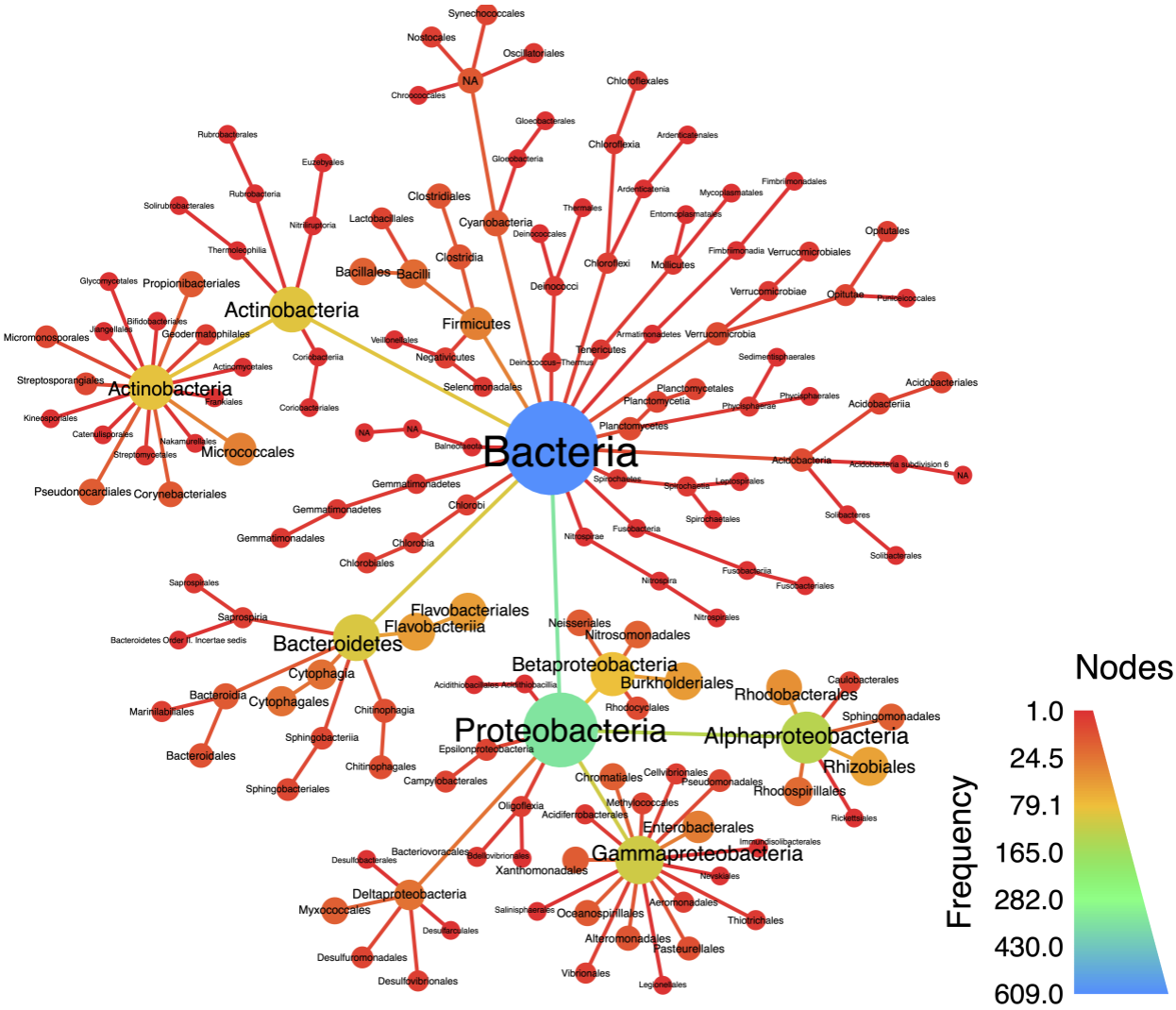
Supporting Information

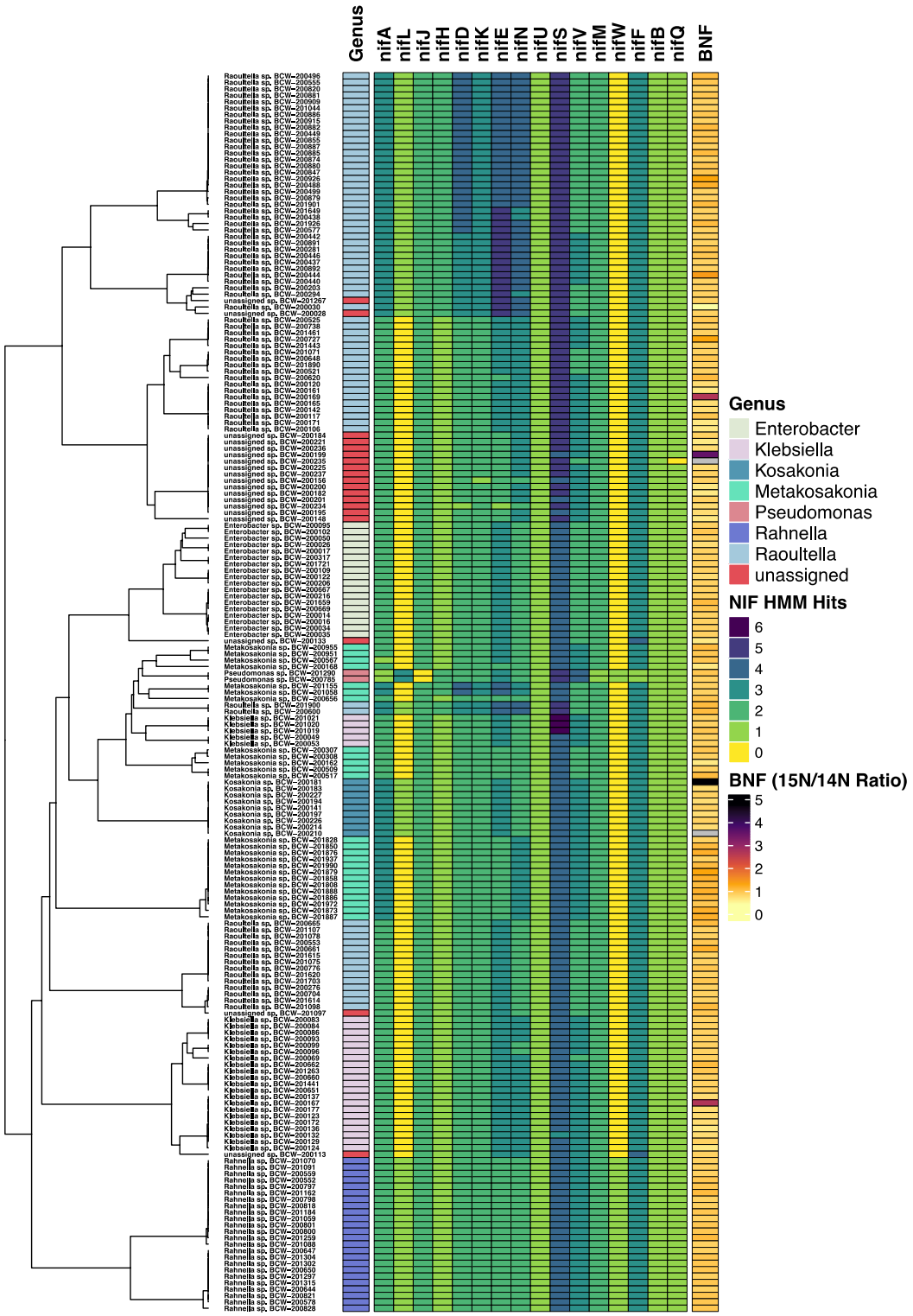
Supplementary Figures



S1 Fig. Taxonomic classification of genome bins from mucilage isolates. Heat Tree showcasing the phylogenetic diversity of all Metabat [1] genome bins derived from isolate draft genome assemblies. Node size corresponds to the frequency of occurrence at each level of taxonomic classification. Dark blue corresponds to the highest observation count of a taxonomic assignment and dark red indicates low observation frequency.

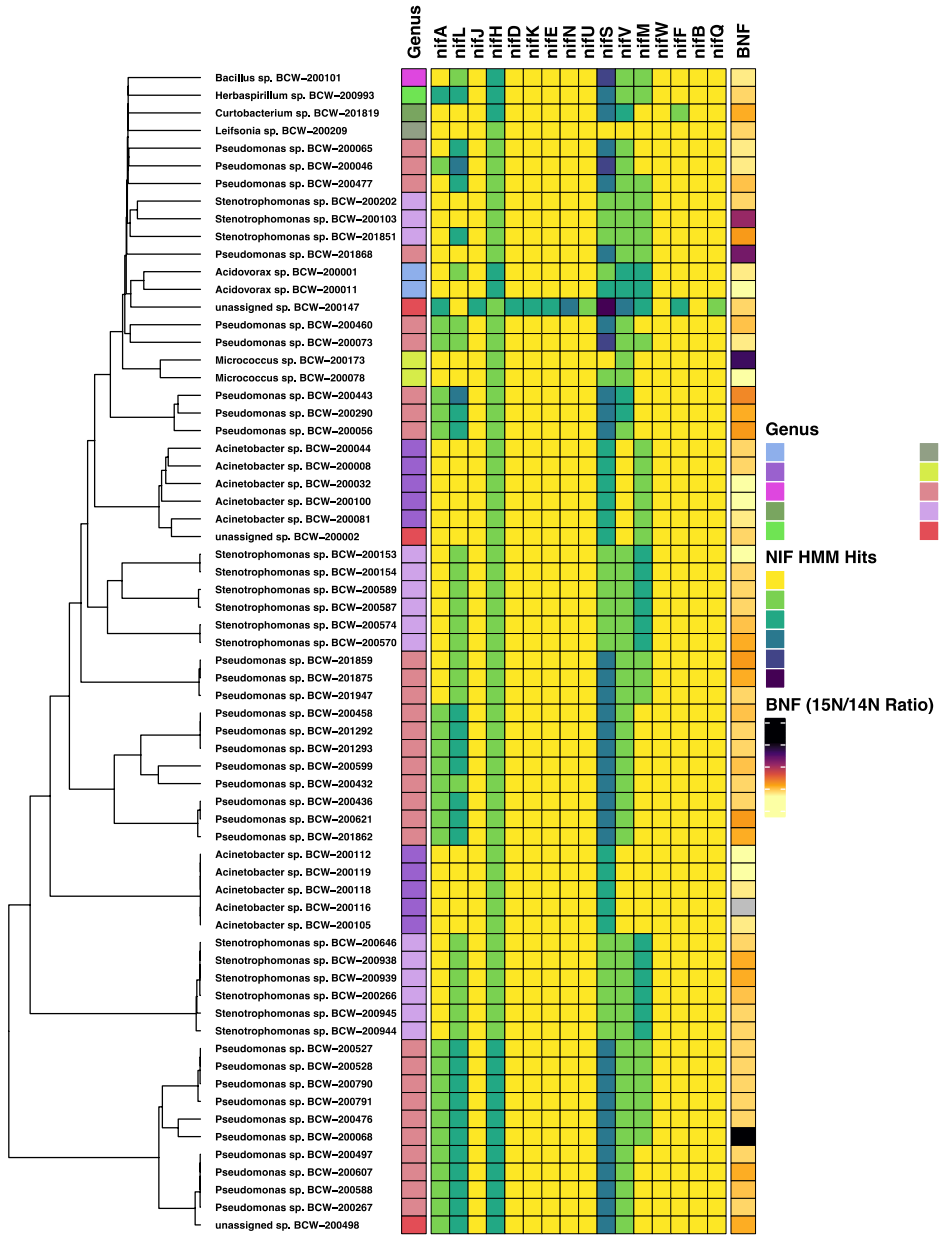


S2 Fig. Observation counts of classified taxa in the mucilage metagenome. Quality trimmed short reads of the OLMM00 mucilage metagenome were input to Kraken2 [2] and classified using the RefSeq complete database for microbial genomes installed with built-in commands of Kraken2. Bracken2 [3] was used to re-estimate classified read counts and the data was imported to R in biom-format. Taxa with readcounts less than 500 were filtered using Phyloseq [4] and the data were visualized using MetacodeR [5]. Terminal nodes represent taxa classified at the order level and node size corresponds to frequency of observation for each taxonomic level.

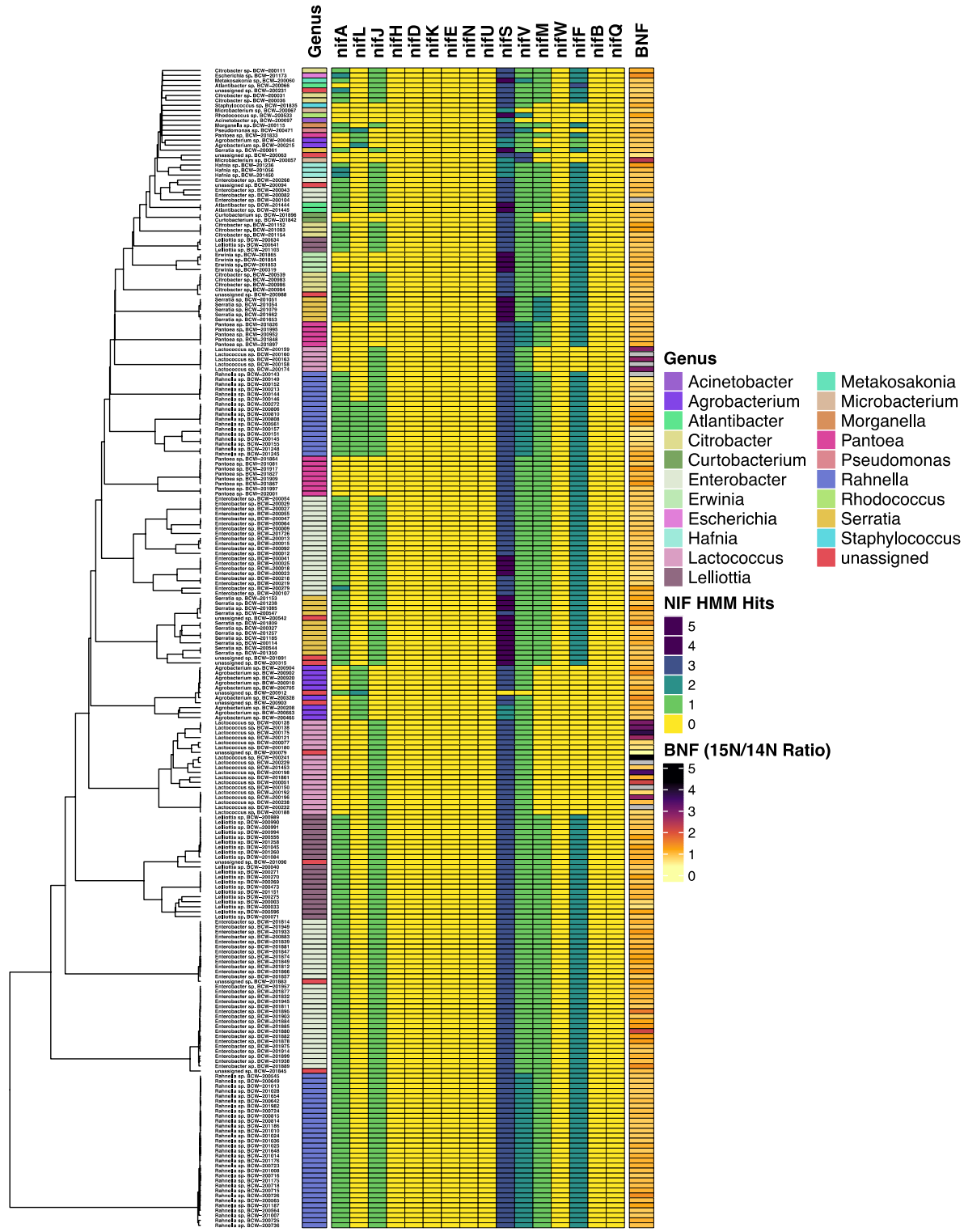


S3 Fig. *Nif* gene profile and BNF assay performance of Dos Santos Positive isolates. *Nif* gene profiles for isolate genomes of the DSP group were extracted and visualized independently

using the ComplexHeatmap [6] package in R. Isolate genomes were clustered using the dendrogram output from Sourmash [7] and each genome row is presented along with annotations to indicate lowest common ancestor (LCA) classification data at the genus level.



S4 Fig. *Nif* gene profile and BNF assay performance of Semi-Dos Santos isolates. *Nif* gene profiles for isolate genomes of the SDS group were extracted and visualized independently using the ComplexHeatmaps [6] package in R. Isolate genomes were clustered using the dendrogram output from Sourmash [7] and each genome row is presented along with annotations to indicate LCA classification data at the genus level.



S5 Fig. *Nif* gene profile and BNF assay performance of Dos Santos Negative isolates. *Nif* gene profiles for isolate genomes of the DSN group were extracted and visualized independently

using the ComplexHeatmaps [6] package in R. Isolate genomes were clustered using the dendrogram output from Sourmash [7] and each genome row is presented along with annotations to indicate LCA classification data at the genus level.

Supplementary Tables

Supplementary tables are provided as sheets within a single Microsoft excel workbook file entitled, *Supporting_Information_Tables.xlsx*. Sheet names correspond to the following table legends that provide additional information related to the values presented in each table.

Supplementary Table Legends

S1 Table. Culturing conditions for Sierra Mixe maize bacterial isolates.
Isolation sources included samples collected from landrace maize varieties grown in the Sierra Mixe region of Oaxaca, Mexico. Medium types included Blood-Heart-Infusion Agar (BHI), M9 minimal medium, and a custom nitrogen-free minimal medium (NFM) – see Methods.

S2 Table. List of diazotrophic isolates and their BNF ratios. $^{15}N/^{14}N$ ratios (BNF ratios) for each isolate are presented alongside their unique isolate ID number. BNF ratios represent the quotient of summations for peak intensities of all statistically significant N-containing biomarkers under both enriched and non-enriched atmospheric conditions. N-containing biomarkers were determined by analysis of LC-MS data using Metaboanalyst [8]. See methods for further details on determination of N-containing biomarkers.

S3 Table. Bacterial whole genome sequencing library and assembly metrics. Genome assemblies generated with MEGAHIT [9] were assessed using QUAST [10] to generate the summary data presented that include number of contigs, total length of the assembly in megabase pairs (Mb), length of the largest contig in the genome assembly in kilobase pairs (Kb),

percentage of the genome comprised by guanine and cytosine (GC %). Mean fold coverage was computed by mapping input reads back to the genome assembly and computing an average value for the coverage of all contigs in the assembly. Genome bins were generated using Metabat [1].

S4 Table. Taxonomic classification of all sequenced bacterial isolate genomes. All sequenced isolate genomes were classified using the 'lca classify' function of Sourmash [7]. MinHash sketches were generated using a k-size of 31 and a scaled value set at 2000. Each genome sketch was queried against a database of MinHash sketches for Genome Taxonomy Database (GTDB) [11] version 89 (available at: https://osf.io/gs29b/).

S5 Table. Taxonomic classification of WGS bacterial isolate bins. Genome bins from all sequenced isolate genomes were generated using Metabat. These bins were subsequently classified using the 'lca classify' function of Sourmash [7]. MinHash sketches were generated using a k-size of 31 and a scaled value set at 2000. Each genome sketch was queried against a database of MinHash sketches for Genome Taxonomy Database (GTDB) [11] version 89 (available at: https://osf.io/gs29b/).

S6 Table. Summary of pure isolate WGS. Pure isolate genome assemblies were generated using MEGAHIT [9], assessed using QUAST [10], and taxonomically classified using the 'lca classify' function of Sourmash [7] (see Methods). MinHash sketches were generated using a k-size of 31 and a scaled value set at 2000. NIF Group assignments indicate that isolate genomes were either positive for the presence of all six essential *nif* genes of the Dos Santos Model (DSP), had a semi-complete set of the six *nif* genes (SDS), or were negative for all six (DSN) [12].

S7 Table. Mucilage metagenome taxonomic classification and normalized readcounts. The OLMM00 mucilage metagenome previously reported by Van Deynze et al. in 2018 [13] was re-analyzed using Kraken2 [14] and the RefSeq complete database of microbial genomes [15]. Classified read counts were re-estimated using Bracken2 and exported in biom-format using Kraken-biom. The read counts were imported to R and normalized to counts per million using Phyloseq. Taxa with fewer than 500 classified reads were removed. Relative abundances were generated by dividing re-estimated read counts for each taxon by the sum total of all classified reads post-filter.

S8 Table. Summary of mucilage metagenome taxonomic observation frequencies. This table provides the corresponding values for taxonomic observation frequencies in the OLMM00 mucilage metagenome after re-estimation of Kraken2 [14] classified read counts by Bracken2 [3] that were used to generate S1 Fig.

S9 Table. Carbohydrate Active Enzyme Families used for CAZyme genome screening. The CAZy database [16] was referenced to generate a manually curated list of CAZyme families that were determined to by relevant for utilization of mucilage polysaccharide. Reported substrate specificities were considered for each family to form eleven custom groupings based on sugar residue activities.

S10 Table. Summary of GH family gene presence in pure isolate genomes. Values indicate the number of isolates with one or more unique genes identified within the genome that matched a HMM for GH families within a designated grouping: Ara – GH(62,127,137,142,146); Ara/Gal/GlcA/Man/Xyl - GH2; Ara/Xyl - GH(3,43,51,54); Fuc - GH(29,95,139,141,151); Fuc/GlcA/Xyl - GH30; Gal - GH(16,27,35,36,42,57,59,95,97,98,110,147,160,165); Gal/GlcA - GH4; Gal/GlcA/Fuc/Man/Xyl - GH1; Gal/Man/Xyl - GH 31; GlcA - GH(67,79,115,154); Man - GH(5,38,47,63,92,99,125,130,164). 'N isolates' indicates the total number of pure isolate genomes that were classified to the indicated genus and included in the analysis.

S11 Table. TCDB Accessions used for sugar transport gene detection. Transporter accession IDs, transporter type and descriptions were reported based on information available in the Transporter Classification Database (www.tcdb.org) [17]. Sugar transporters were selected by manually searching through the TCDB for bacterial sugar transport genes that corresponded to monosaccharide components of the mucilage polysaccharide from Sierra Mixe maize.

S12 Table. Summary of sugar transporter gene presence in pure isolate genomes. Values represent the total number of isolates with detected membrane transporters for each sugar component of mucilage polysaccharide. 'N Isolates' corresponds to the number of isolate genomes sequenced under the given taxonomic assignment at the genus level. Transporter mechanisms for each sugar were grouped based on the type of sugar transport to calculate sum totals for each isolate class.

S13 Table. Summary for the Pan-genome of isolates possessing alternative *nif* genes. The pan-genome for isolates containing alternative *nif* genes was generated using Roary 3.12.0 [18]. The summary of features was generated as standard output from running the bioinformatic pipeline with GFF files for each isolate generated using Prokka 1.12 [19].

References

1. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ. 2015;3:e1165. Epub 2015/09/04. doi: 10.7717/peerj.1165. PubMed PMID: 26336640; PubMed Central PMCID: PMCPMC4556158.

2. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20(1):257. Epub 2019/11/30. doi: 10.1186/s13059-019-1891-0. PubMed PMID: 31779668; PubMed Central PMCID: PMCPMC6883579.

3. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. Peerj Computer Science. 2017;3:e104. doi: 10.7717/peerj-cs.104. PubMed PMID: WOS:000425411300002.

4. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One. 2013;8(4):e61217. Epub 2013/05/01. doi: 10.1371/journal.pone.0061217. PubMed PMID: 23630581; PubMed Central PMCID: PMCPMC3632530.

5. Foster ZS, Sharpton TJ, Grunwald NJ. Metacoder: An R package for visualization and manipulation of community taxonomic diversity data. PLoS Computational Biology. 2017;13(2):e1005404. Epub 2017/02/22. doi: 10.1371/journal.pcbi.1005404. PubMed PMID: 28222096; PubMed Central PMCID: PMCPMC5340466.

6. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32(18):2847-9. Epub 2016/05/22. doi: 10.1093/bioinformatics/btw313. PubMed PMID: 27207943.

7. Brown CT, Irber L. sourmash: a library for MinHash sketching of DNA. J Open Source Software. 2016;1(5):27.

8. Xia JG, Sinelnikov IV, Han B, Wishart DS. MetaboAnalyst 3.0-making metabolomics more meaningful. Nucleic Acids Research. 2015;43(W1):W251-W7. doi: 10.1093/nar/gkv380. PubMed PMID: WOS:000359772700039.

9. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31(10):1674-6. Epub 2015/01/23. doi: 10.1093/bioinformatics/btv033. PubMed PMID: 25609793.

10. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072-5. Epub 2013/02/21. doi: 10.1093/bioinformatics/btt086. PubMed PMID: 23422339; PubMed Central PMCID: PMCPMC3624806.

11. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. 2019. Epub 2019/11/16. doi: 10.1093/bioinformatics/btz848. PubMed PMID: 31730192.

12. Dos Santos PC, Fang Z, Mason SW, Setubal JC, Dixon R. Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. BMC Genomics. 2012;13(1):162. Epub 2012/05/05. doi: 10.1186/1471-2164-13-162. PubMed PMID: 22554235; PubMed Central PMCID: PMCPMC3464626.

13. Van Deynze A, Zamora P, Delaux PM, Heitmann C, Jayaraman D, Rajasekar S, et al. Nitrogen fixation in a landrace of maize is supported by a mucilage-associated diazotrophic microbiota. PLoS Biol. 2018;16(8):e2006352. Epub 2018/08/08. doi: 10.1371/journal.pbio.2006352. PubMed PMID: 30086128; PubMed Central PMCID: PMCPMC6080747 sponsors. Author Cristobal Heitmann was unable to confirm authorship or contributions himself, and this was carried out collectively by the other co-authors.

14. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3):R46. Epub 2014/03/04. doi: 10.1186/gb-2014-15-3-r46. PubMed PMID: 24580807; PubMed Central PMCID: PMCPMC4053813.

15. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007;35(Database issue):D61-5. Epub 2006/11/30. doi: 10.1093/nar/gkl842. PubMed PMID: 17130148; PubMed Central PMCID: PMCPMC1716718.

16. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res. 2009;37(Database issue):D233-8. Epub 2008/10/08. doi: 10.1093/nar/gkn663. PubMed PMID: 18838391; PubMed Central PMCID: PMCPMC2686590.

17. Saier MH, Jr., Reddy VS, Tsu BV, Ahmed MS, Li C, Moreno-Hagelsieb G. The Transporter Classification Database (TCDB): recent advances. Nucleic Acids Res. 2016;44(D1):D372-9. Epub 2015/11/08. doi: 10.1093/nar/gkv1103. PubMed PMID: 26546518; PubMed Central PMCID: PMCPMC4702804.

18. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691-3. Epub 2015/07/23. doi: 10.1093/bioinformatics/btv421. PubMed PMID: 26198102; PubMed Central PMCID: PMCPMC4817141.

19. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068-9. Epub 2014/03/20. doi: 10.1093/bioinformatics/btu153. PubMed PMID: 24642063.