Description of Metagenome-assembled Genome Binning and Justification for Phylobins

Initially, MAGs were generated using automated binning algorithms CONCOCT [40] and MetaBAT [37] and a custom method based on hdbscan clustering of a t-SNE ordination (theta=0.5, perplexity = 30, $dims_{initial}$ = 50, $dims_{final}$ = 2) based on pentanucleotide frequencies in a process similar to VizBin [41]. The custom method is available in Jupyter notebook format in the Supplementary Data package. Consensus bins were then produced using DASTool [42]. DASTool produced 12 and 10 MAGs from enriched and unenriched contigs sets, respectively, which was half the number of complete phylobins (22 and 20). The DASTool MAG set was missing several of the most abundant ^{13}C-enriched taxa in the metagenome, including members of *Sphingomonadales*, *Micrococcales*, *Chloroflexales*, *Xanthomonadales*, *Vampirovibrionales*, *Streptomycetales*, *Rhodobacterales* and *Pseudomonadales*. Representative taxa from these MAGs were known to contain ^{13}C-labeled DNA from a prior DNA-SIP study performed on this DNA sample (Pepe-Ranney *et al*., 2016). Most missing MAGs were generated by at least one of the other binning algorithms, making it possible to assemble a more complete set of MAGs. However, we observed major issues with the cross-contamination of all MAGs produced by automated software evidenced by the LCA classifications of contigs (Table S9). On average, 10% of contigs in any MAG produced by DASTool bins were classified to taxonomic families outside of their phylogenetic lineage. For example, in a bin primarily composed of contigs classified to groups within *Verrucomicrobia* ("DASTool_tSNE.051": "*Rubritaleaceae*", "*Verrucomicrobiaceae*", "*unclassified Verrucomicrobia*", "*unclassified Verrucomicrobiales*") there were a total of 88 contigs (n_{total} = 514) belonging to other taxonomic groups abundant in our dataset, including *Planctomycetacea* and several of the proteobacterial families. In this bin, a total of 16% of all bases belonged to outgroups (n_{size} = 3.8 Mb). This bin would be assigned as relatively high-quality based on a CheckM assessment: 85% complete and 13% contamination [43]. Yet, the observed cross-

contamination posed a major concern for the accuracy of functional gene-based characterizations, which was problematic in all binning methods, including our custom method (28%, or 8.2% of contigs when subsetted to the most complete MAGs), MetaBAT (12% and 6.5%, respectively) and CONCOCT (18% and 6.3%, respectively). To resolve the cross-contamination issue and ensure a complete representation of taxonomic groups in our dataset, we opted to use a phylobin strategy, where contigs were grouped based on LCA classification at the rank Order. We interpreted results from phylobins with the following limitations in mind: (i) phylobins are more prone to false-negative error (missing genes due to the lack of sensitivity of LCA, particularly for underrepresented clades) and (ii) the aggregation of multiple genomes could confound differences among closely related taxa (which is of interest to our research questions). The loss of resolution of individual genomes phylobins was compensated for by performing all analyses in parallel on reference genomes chosen based on the similarity of full-length 16S rRNA genes recovered in our study.

Explaining Phylobin Sizes

The large size of phylobins can be attributed to natural pangenomic diversity (intra-species) and to the grouping of mixed populations by Order (inter-species). The exact number of species variants contributing to the pangenomic diversity of each phylobin varied according to estimates of single nucleotide polymorphisms (SNPs) in putatively single-copy genes (Table S10). In phylobins expected to contain a single genus, the median number of SNPs per single-copy gene ranged from 0.38 per 100 bp (in strongly ^{13}C-enriched *Cellvibrionales*) to 1.62 (in strongly ^{13}C-enriched *Micrococcales*). The background rate of SNPs in our assembly was 0.18 per 100 bp, indicating the number of species/strain variants ranged from 2 to 9 in single-genus phylobins.

We also enumerated the number of single-copy genes present in each phylobins to estimate the number of composited genomes. The median number of single-copy genes correlated with total phylobin size ($r = 0.86$; $p < 0.001$), demonstrating that most phylobins were comprised of several genomes, the largest containing a median 19 single-copy genes in the strongly ^{13}C-enriched *Rhizobiales* phylobin (Table S10). However, the use of single-copy genes to accurately estimate the number of genome duplications was found to be flawed. We observed that 15% of predicted single-copy genes occurred within the first or last 250 bp of contigs (> 7.5 Kb), which should occur by chance 3.4% of the time based on single-copy gene and contig lengths. This demonstrated that the highly conserved single-copy genes were causing breaks in the assembly, a well-studied problem for assembling the highly conserved 16S rRNA gene (Pericard *et al.*, 2018). These breaks create the potential to double count single-copy genes: once on each side of the break. To test whether this was occurring, we ran CheckM on 50 high-quality ($n_{contigs}$ =5), mid-quality ($n_{contigs}$ =50) and low-quality ($n_{contigs}$ = 405) genomes from the NCBI database 'refseq,' belonging to the order *Rhizobiales*. We identified 84 instances where the same single-copy gene occurring at the end of contigs were counted as 'contaminants' by CheckM (i.e. single-copy genes occurring more than once). The number of double-counted single-copy genes scaled with assembly quality with 5 out of 50 (10%) genomes from both high and mid-quality assembly having at least one instance of double-counting, while 23 / 50 (46%) of poorly assembled genomes had instances of double-counting. We can assume that the increased genomic diversity in a metagenome would further exacerbate problems with assembling highly conserved single-copy genes. Therefore, we cannot conclude with certainty the number of duplicated genomes per phylobins by enumerating the duplication of single-copy genes. Given these limitations, we believe our approach to complement the analysis of phylobins with representative genomes. Phylobins represent coherent ecological

units based on ^{13}C-assimilation and the phylogenetic conservation of traits within closely related organisms, while representative genomes are not subject to cross-contamination.

Single-copy genes were annotated using HMMs provided by BUSCO using HMMer. SNPs were enumerated via read mapping to all annotated single-copy genes with BBMap and SNP calling with samtools (mpileup; bcftools and vcfutils). Annotated sequences shorter than 90% the length of the reference single-copy genes were excluded from SNP calling. Ns were not considered as SNPs. The total number of SNPs for each gene family was normalized to the total number of copies in each family (i.e. direct correspondence: # SNPs in each single-copy gene family / # copies of that gene family in each MAG). In our validation of CheckM results, we used the output 'hmmer.analyze.txt' to identify single-copy genes occurring at the beginning or end of contigs and 'bin_stats_ext.tsv' to identify which were gene families were ruled as 'contaminants' (i.e. double counted).

Anomaly in Recovering Full-length 16S rRNA gene for Vampirovibrio

16S rRNA genes classified to *Vampirovibrio* were abundant in shotgun metagenome fragments and in 16S rRNA gene amplicon data, but a representative full-length sequence was not assembled by MATAM. We manually recovered a full-length 16S rRNA gene by mapping reads to the 16S rRNA gene from *V. chorellavorus* (LAPX00000000) with bowtie2 [31] (resulting in a 218x read depth) and used samtools 'mpileup' to create a consensus sequence (97% similarity to *V. chorellavorus*) [32].

Overview of Multiple De Novo Assembly and Merging Method

A total of 1.1 trillion reads remained after quality filtering and were distributed across eight density gradient fractions with a maximum read total of 181 million in the second heaviest fraction 7 (F_7) and a minimum of 9.8 million in the lightest fraction (F_{13}). The sequencing depth achieved

covered 80% of the estimated genomic diversity in the community according to nonpareil (Rodriguez and Konstantinidis, 2014). Metagenome assembly produced a total of 356,131 contigs greater than 2.5 Kb, amounting to a total length of 1.8 Gb (ex. ~ 230 genomes of 8 Mb). Merging *de novo* and secondary assemblies did not significantly increase the total length of the assembly, but increased the N_{50} by ~ 23%, doubling the number of contigs longer than 25 Kb (Table S2). Merging assemblies introduced single nucleotide ambiguities in the assembly at an average of 180 per 100 Kb (0.18%). The lack of consensus at each ambiguous base position was confirmed using SAMtools 'mpileup', revealing the prevalence of single-nucleotide polymorphisms in the assembly.

Overview of Designating Contig Enrichment Using Random Forest Classification

The Random Forest model used to predict enrichment status had an overall accuracy of 89.1% with high sensitivity and specificity for both strongly enriched (98.3% and 86.4%, respectively) and unenriched contigs (100% and 93%) (Table S4). The greatest source of inaccuracy was in classifying weakly enriched and bimodal gradient profiles. This had minimal impact on results, since weakly enriched were most frequently misclassified as strongly enriched (64%), and, in most analyses, these designations were treated both as enriched. Few contigs were identified as having bimodal distributions with respect to contig ^{13}C-enrichment (< 0.6% of total contigs) and these were excluded Contigs identified as bimodal represented a small fraction of the total and were excluded from analyses as they did not form coherent contig clusters.

Evidence for the Absence of Effect of GC Content on Designation of 'Weak' Enrichment

Contrary to what would be expected if GC content were driving differences in 'weak' enrichment, the weakly enriched phylobins had lower GC content than corresponding unenriched phylobins for Sphingomonadales (μ_{weak} = 64.1% GC vs. $\mu_{unenriched}$ = 67.0%; $p < 0.001$) and

Planctomycetales (μ_{weak} = 63.9% vs. $\mu_{unenriched}$ = 66.5%; $p < 0.001$). No difference was observed in the GC content between weakly and strongly enriched Planctomycetales phylobins (μ_{weak} = 63.9% vs. μ_{strong} = 63.7%; $p = 0.08$), and the higher GC content in the strongly-enriched Sphingomonadales phylobin (μ_{strong} = 65.5% vs. μ_{weak} = 64.1%; $p < 0.001$) was too slight to account for the size of shift in density observed ($\Delta BD_{theoretical.}$ = 0.0012 g/mL vs. $\Delta BD_{obs.}$ = 0.0104 g/mL). The distinct clustering of contig sets by pentanucleotide signatures provided additional evidence that weakly enriched phylobins represented unique populations exhibiting a lower degree of ^{13}C-enrichment (Figure S4b).

Assigning Cellulolytic Potential

All open-reading frames annotated to the following endoglucanases-containing glycosyl hydrolase families (>= 60% similarity across 90% of the gene) were deemed to confer cellulolytic ability: GH5(1), GH5(2), GH5(4), GH5(5), GH5(7), GH5(8), GH5(9), GH5(11), GH5(12), GH5(13), GH5(15), GH5(16), GH5(18), GH5(19), GH5(22), GH5(23), GH5(24), GH5(25), GH5(26), GH5(27), GH5(28), GH5(29), GH5(30), GH5(31), GH5(36), GH5(37), GH5(38), GH5(39), GH5(40), GH5(41), GH5(43), GH5(44), GH5(45), GH5(46), GH5(47), GH5(48), GH5(49), GH5(50), GH5(51), GH5(53), GH6, GH7, GH8, GH9, GH12, GH16, GH26, GH44, GH45, GH48, GH51, GH61, GH74, and GH131. Chitinase-containing glycosyl hydrolase families were GH18 and GH19.

Preparation of Metaproteome

Total protein was extracted from soil using the NoviPure Soil Protein Kit (QIAGEN), as per manufacturer's instructions, up until protein precipitation and acetone washes. The protein pellets were dissolved in the lowest possible volume of extraction buffer (4% SDS, 100mM DTT in 100mM Tris-HCL) in a 1.5mL tube and incubated at 95 °C for 5 minutes, cooled and sonicated in a water bath. Any resulting debris was removed by centrifugation at 15,000g for 10 minutes. Supernatant was transferred to a new 1.5mL tube for Filter Aided Sample Preparation according

to methods described in https://www.biochem.mpg.de/4855967/FASP-Protocol.pdf. Samples were desalted using MicroSpin Columns, C18 Silica, The Nest Group, Inc. (Part # SEM SS18V) and MS grade solvents using manufacturer's instructions. Th bicinchoninic acid assay (BCA) was used for quantifying peptide concentration.

Chromatographic separation of peptides was performed with a Waters nano-Acquity M-Class dual pumping UPLC system (Milford, MA) configured for on-line trapping of a 5 µL injection with a 3 µL/min reverse-flow elution onto the analytical column flowing at 300 nL/min. Columns were packed in-house using 360 µm o.d. fused silica (Polymicro Technologies Inc., Phoenix, AZ) with 5-mm sol-gel frits for media retention (Maiolica *et al.*, 2005) and contained Jupiter C18 media (Phenomenex, Torrence, CA) in 5µm particle size for the trapping column (150 µm i.d. x 4 cm long) and 3µm particle size for the analytical column (75 µm i.d. x 70 cm long). Mobile phases consisted of (A) 0.1% formic acid in water and (B) 0.1% formic acid in acetonitrile with the following gradient profile (min, %B): 0, 1; 2, 8; 20, 12; 75, 30; 97, 45; 100, 95; 110, 95; 115, 1; 150, 1.

Mass spectroscopy (MS) was performed using a Q-Exactive HF mass spectrometer (Thermo Scientific, San Jose, CA) outfitted with a home-made nano-electrospray ionization interface. Electrospray emitters were prepared using 150 µm o.d. x 20 um i.d. chemically etched fused silica (Kell *et al.*, 2006). The ion transfer tube temperature and spray voltage were 325ºC and 2.4 kV, respectively. Data were collected for 100 min following a 15 min delay from sample injection. FT-MS spectra were acquired from 400-2000 m/z at a resolution of 60k (AGC target 3e6) and while the top 12 FT-HCD-MS/MS spectra were acquired in data dependent mode with an isolation window of 2.0 m/z and at a resolution of 15k (AGC target 1e5) using a normalized collision energy of 30 and a 45 sec exclusion time.

Seven fractions per sample, from the preceding UPLC separation, were subsequently submitted for LC-MS/MS using an LTQ Orbitrap Velos mass spectrometer (ThermoFisher, Waltham MA) coupled to a high-performance reverse phase liquid chromatography system. Data were collected for 80 minutes using a Top-10 shotgun approach involving a single MS level survey scan (350m/z – 1800m/z) followed by 10 data-dependent MS/MS (a.k.a. MS2) scans using the ten most abundant signals from the survey scan as precursors. MS/MS spectra were developed using a normalized collision energy of 35, isolation width of 2m/z, and a +/-10ppm rolling dynamic exclusion list wherein previously selected survey scan signals were not fragmented again for 15 seconds.

Peptides were identified from LC-MS/MS data using predicted protein sequences from the metagenome. MS/MS spectra from each LC-MS/MS dataset were extracted using MSConvert (https://github.com/ProteoWizard/pwiz) to ASCII text format (.dta) and MSGFPlus (Kim and Pevzner, 2014) used to compare spectra against candidate sequences from the FASTA files. Due to size limitations, each FASTA file was split into 25 equal sections and MS/MS spectra searched against each section, with results recombined using MSGF Spectral Probability (Kim *et al.*, 2008) used to determine the best Peptide-to-Spectrum match (PSM). Additionally, only tryptic peptides (i.e. enzymatic cleavage at Lysing or Arginine) were considered as PSM candidates, with a +/-15ppm precursor signal mass tolerance applied. Oxidized Methionine, a commonly observed port-translational modification in proteomics experiments, was included in the search approaches, as well as using a Target/Decoy approach to assess false discovery rate (FDR) (Elias and Gygi, 2007).

MSGFPlus search results were collated to tab-separated text format using in-house developed MAGE Extractor, wherein only one protein is reported for each PSM, specifically the first in the randomly sorted peptide FASTA. Subsequently the results were imported into a

Microsoft SQL Server relational database and filtered by adjusting the MSGFPlus supplied Q-Value to as close to 1% FDR as was practical (FDR = $[n_{decoy}/n_{passed\ filter}]$ x 100).

An analysis of auxotrophy in the metaproteome was not attempted given the low specificity of short peptide sequences to highly conserved biosynthetic genes.

References

Elias JE and Gygi SP. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*, **4**(3): 207-14.

Kelly RT, JS Page, Q Luo, RJ Moore, DJ Orton, K Tang, and RD Smith. 2006. "Chemically Etched Open Tubular and Monolithic Emitters for Nanoelectrospray Ionization Mass Spectrometry." *Analytical Chemistry* **78**(22):7796-7801

Kim S, Gupta N, Pevzner PA. 2008. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J Proteome Res*, **7**(8): 3354-63.

Kim S and Pevzner PA. 2014. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun*, **31**: 5-5277.

Maiolica A, Borsotti D, Rappsilber J. 2005. "Self-made frits for nanoscale columns in proteomics." Proteomics **5**: 3847-3850

Pepe-Ranney C, Campbell AN, Koechli CN, Berthrong S, Buckley DH. 2016. Unearthing the ecology of soil microorganisms using a high-resolution DNA-SIP approach to explore cellulose and xylose metabolism in soil. Front Microbiol, **7**: 1–17.

Pericard P, Dufresne Y, Couderc L, Blanquart S, Touzet H. 2018. MATAM: Reconstruction of phylogenetic marker genes from short sequencing reads in metagenomes. Bioinformatics, **34**: 585–591.