ARTICLE

Microbial carbon use efficiency predicted from genome-scale metabolic models

Mustafa Saifuddin[1], Jennifer M. Bhatnagar [1], Daniel Segrè [1] & Adrien C. Finzi[1]

Respiration by soil bacteria and fungi is one of the largest fluxes of carbon (C) from the land surface. Although this flux is a direct product of microbial metabolism, controls over metabolism and their responses to global change are a major uncertainty in the global C cycle. Here, we explore an in silico approach to predict bacterial C-use efficiency (CUE) for over 200 species using genome-specific constraint-based metabolic modeling. We find that potential CUE averages 0.62 ± 0.17 with a range of 0.22 to 0.98 across taxa and phylogenetic structuring at the subphylum levels. Potential CUE is negatively correlated with genome size, while taxa with larger genomes are able to access a wider variety of C substrates. Incorporating the range of CUE values reported here into a next-generation model of soil biogeochemistry suggests that these differences in physiology across microbial taxa can feed back on soil-C cycling.

[1] Boston University, Boston, MA 02215, USA. Correspondence and requests for materials should be addressed to A.C.F. (email: afinzi@bu.edu)

Soil respiration is one of the largest exchanges of carbon (C) from the land surface to the atmosphere, releasing an estimated amount of 98 ± 12 Pg C/year from soil as CO_2[1,2]. Heterotrophic respiration (R_H) by soil bacteria and fungi can account for a large proportion of total global soil respiration (35–69 Pg C/year[3,4]). Although this flux is a direct product of microbial metabolism, controls over physiology and their responses to global change are a major uncertainty in the global C cycle[5,6].

Soil respiration rates are influenced by microbial physiology because cellular metabolism dictates what fraction of the C consumed by soil bacteria and fungi is allocated to respiration, biomass, extracellular enzyme production, and other metabolic functions. Thus, the partitioning of microbial C can have a direct impact on the global C cycle[5]. Variations in C partitioning among microbial taxa is poorly characterized, yet critical to understanding relationships between physiology, community composition, and soil C cycling.

Carbon use efficiency (CUE) measures the partitioning of C between microbial biomass and respiration[7]. Empirical estimates of microbial CUE range from near zero to over 0.8[8,9]. Most biogeochemical models use a fixed value selected between 0.15 and 0.6, typically without careful parameterization[8]. Some of the variation observed in CUE may be attributed to the sensitivity of CUE to abiotic factors such as temperature and pH[8,10]. However, an additional and often neglected source of variation in CUE may be due to physiological differences between soil microbial groups and their differential capacities for accessing particular substrate types[7,11,12]. This single parameter in microbial biogeochemistry models has direct impacts on estimates of greenhouse gas emissions and terrestrial C storage[5,6], and the sustainability of bioenergy cropping systems[13], making it necessary to survey how CUE varies both among taxa and across substrate types.

Characterizing microbial C metabolism is particularly important in the context of global change, which may alter the structure and activity of microbial communities and their access to substrates[14]. Previous work on functional and physiological variability suggests that defining bacterial taxa along a spectrum of copiotrophy (fast-growing, adapted to high substrate availability) to oligotrophy (slow-growing, adapted to limiting resource concentrations) may be one useful approach for understanding how groups respond to changes in temperature and resource availability[15,16]. Classification schemes based on trophic strategy may be useful from a biogeochemical perspective if differences in growth strategies correspond to variation in CUE. For example, copiotrophs are hypothesized to show lower CUE than oligotrophs[12,17], and this could potentially alter the CUE of microbial communities observed to shift toward a greater proportion of copiotrophic bacteria in response to global change manipulations that increase substrate availability, such as soil warming[18,19].

Observations from global change experiments and genome-based estimates of minimal generation times have shown some support for the classification of particular phyla, as oligotrophic or copiotrophic[14]. However, these phylum-level classifications are not consistent across studies[20,21]. An improved understanding of the phylogenetic structure of biogeochemically relevant traits is needed to identify how microbial community structure impacts C cycling[22,23]. Many functional genes show strong conservation within prokaryotes, leading to the possibility for strong phylogenetic structure in functional traits, particularly those that emerge from the coordinated activity of multiple genes[24]. For example, bacterial traits such as growth rates in the presence of labile C show strong phylogenetic signals, whereas other traits such as responses to priming show shallow phylogenetic signals[22,23]. Thus, the level of phylogenetic resolution required to characterize variation in CUE across bacterial taxa remains unclear.

In addition to understanding the phylogenetic structure of variation in CUE, it may be useful to explore whether particular genomic traits predict CUE. For example, copy numbers of ribosomal RNA operons are inversely related to growth efficiency in bacteria, providing a method for predicting growth efficiencies from genomes[12]. Similarly, genomic traits have been useful for predicting microbial trophic strategies and biogeography, with bacterial taxa with larger genomes occupying a wider range of habitat types[25] and dominating communities where resources are available in diverse forms but limiting concentrations[26]. Comparable efforts for predicting CUE from genomic traits are necessary to help overcome challenges with measuring taxa-specific CUE for highly diverse soil bacterial communities.

In the environment, microbial taxa are exposed to variations in substrate chemistry and supply rates that impact rates of C uptake and growth. These abiotic factors are likely to interact with intrinsic differences in physiology among taxa to ultimately determine CUE. For example, observations from bacterial cultures show that CUE increases with limiting resource concentration and with the free energy content of available resources[17]. These patterns are overlaid with differences between taxa, with potentially oligotrophic groups showing less responsivity to limiting resource availability than copiotrophic taxa[17]. Thus, estimates of CUE must consider both biotic and abiotic sources of variability, including bacterial physiology, substrate availability, and substrate chemistry.

Prior work on estimating CUE is limited to a small set of individual microbial taxa, or involves mixed, whole communities[10,11,17]. Direct measurements of CUE have been made using a wide range of methods including calorespirometry[11] and stable isotope approaches[23,27]. CUE has also been estimated indirectly for whole communities based on environmental variables such as resource stoichiometry[10,28]. These methods can lead to CUE estimates that vary by a factor of two or more, making direct inter-comparisons challenging[10].

Using a consistent methodology to measure CUE across a broad range of microbial taxa is necessary to determine how physiological variation in resource use between taxa impacts CUE. Metabolic models of bacterial physiology can be generated from annotated genomes[29] and can be used to estimate taxa-specific biological fluxes, including biomass growth and C uptake[30]. Here, we explore an in silico approach to generate theoretical predictions of CUE for over 200 taxa using genome-scale constraint-based metabolic modeling. We find that intrinsic physiological differences between taxa can lead to >300% variation in CUE, which is far greater than that assumed in global models where CUE is either fixed or varies solely in relation to abiotic factors. We find that CUE is primarily structured at subphylum phylogenetic levels and is correlated negatively with genome size and GC content. These findings provide a framework for predicting CUE from genomic traits and for inferring potential impacts of shifts in bacterial community composition on C cycling. Using a recent ecosystem model of heterotrophic soil respiration (DAMM-MCNiP), we demonstrate that accounting for the observed variation in microbial physiology across taxa alone can have persistent implications for estimates of soil C emissions and soil C pool sizes.

Results

CUE from manually curated models varies by taxa and substrate. We first calculated CUE using the set of 13 manually curated, published metabolic models from diverse environments found in the BiGG database[31]. Flux balance analysis was
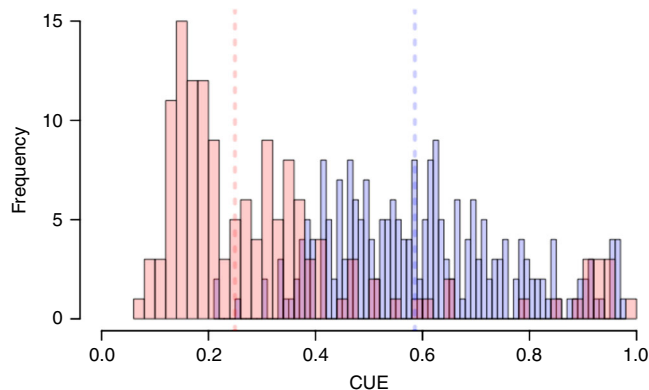
Fig. 1 Potential and substrate-limited CUE. Histogram of predicted potential CUE (purple) and predicted CUE under lysine-limitation (pink) across taxa

performed for each metabolic model with C supplied exclusively through one of 14 individual C-containing metabolites, and CUE was calculated as the proportion of C assimilated into biomass relative to C uptake. We observe a mean CUE of 0.53 ± 0.25 (S.E.) across taxa and substrate types, suggesting that nearly half of consumed C is lost via respiration on average (Supplementary Table 1). However, these models also indicate wide variation in mean CUE between individual taxa (0.14 ± 0.07 to 0.84 ± 0.17, Supplementary Table 1) and equally large variation in mean CUE across substrate types (0.26 ± 0.24 to 0.66 ± 0.20, Supplementary Table 1).

Potential CUE from draft models varies by taxa and substrate. Potential CUE values represent intrinsic variation in CUE based on genomic differences between taxa, and these values were most useful for comparisons between taxa and for identifying relationships between genome traits and CUE. Potential CUE ranges from 0.22 to 0.98 across all taxa, with a mean of 0.62 ± 0.17 (Fig. 1). The range of potential CUE values from this analysis corresponds to the high end of parameter settings currently used in microbial models of the C cycle (0.15–0.68).

To assess the impact of substrate chemistry on CUE, we calculated the dependence of biomass production on all transport reactions associated with C uptake and secretion. We then identified the set of C-containing metabolites that most commonly limited biomass production across the full set of taxa in our analysis, and calculated CUE after reducing the availability of each of these constraining metabolites individually. The most common constraining reactions were related to amino acid and dipeptide uptake (Supplementary Table 2, Fig. 2). When uptake of individual constraining metabolites was set to reduce biomass production by 75%, mean CUE across all 18 constraining metabolites was 0.29 ± 0.19. This corresponds to an average decline in CUE of 0.33, or a 53% reduction in CUE, compared to the potential CUE scenario.

Potential CUE is associated with genomic traits. Potential CUE shows a significant phylogenetic signal ($K = 0.99$, $p < 0.01$, Fig. 3), indicating a Brownian pattern of trait evolution, with closely related taxa showing similarity in potential CUE values. The class ($CI = 0.02 \pm 0.019$, Supplementary Table 3) and order ($CI = 0.016 \pm 0.020$, Supplementary Table 3) levels explained the most variation in CUE. Therefore, these phylogenetic levels may be more appropriate than the phylum level for considering relationships between C cycling and bacterial community composition.

Consistent with our observations based on the BIGG models (Fig. 4), we found a negative correlation between potential CUE
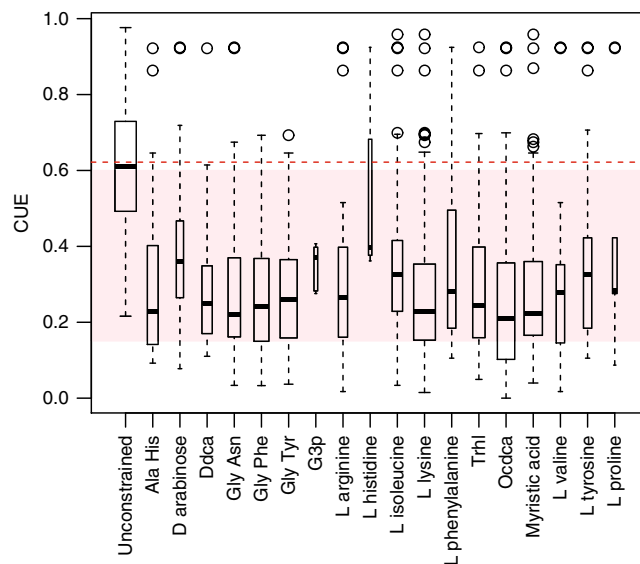

Fig. 2 Substrate-limited CUE. Boxplot of average CUE values across all taxa under potential and constrained scenarios. Boxplot width is proportional to number of models with a given constraining reaction. Dashed red line shows average for potential CUE. Shaded region shows range of values typically used in biogeochemical models. Solid lines within boxplots show median. Bottom and top edges of boxes represent 25th and 75th percentiles, respectively. Whiskers demarcate minimum and maximum datapoints within 1.5× of the interquartile range

and GC content in the larger set of metabolic models from kBase (Pseudo $R^2 = 0.20$, Supplementary Table 4). In addition, potential CUE is significantly negatively correlated with genome size (Pseudo $R^2 = 0.36$, Supplementary Table 4, Fig. 5), the number of genes coded for within a genome (Pseudo $R^2 = 0.34$, Supplementary Table 4) and the number of transport reactions associated with C uptake and secretion (Pseudo $R^2 = 0.50$, Supplementary Table 4).

Variation in potential CUE impacts ecosystem-level C cycling. Under a scenario in which the microbial community exhibited high efficiency (CUE = 0.9), soil organic C pool sizes were nearly twice as large following 100 years of simulation compared to the low efficiency scenario (CUE = 0.2, Fig. 6). This was driven, in part, by large sustained increases in microbial biomass, with the highly efficient microbial communities producing nearly four times greater microbial biomass than low efficiency communities over the same time span. Despite this large increase in microbial biomass, rates of respiration were reduced by 25% compared to the low efficiency communities (Fig. 6).

Discussion
Genome-specific metabolic models have typically been used to explore variation in growth and microbial community interactions for small sets of microbial taxa[32,33]. To date, this approach has not been applied to better understand microbial CUE, a key parameter in emerging microbial models of the C cycle. Here, we show large, phylogenetically structured variation in potential CUE attributed to differences in physiology among >200 individual bacterial taxa. We observed that CUE was sensitive to substrate chemistry, substrate supply, and variation in microbial physiology between taxa. The intrinsic variation in CUE we observed among taxa is as large as that previously attributed to abiotic factors such as temperature and substrate chemistry[8,10]. For example, the temperature sensitivity of CUE for whole
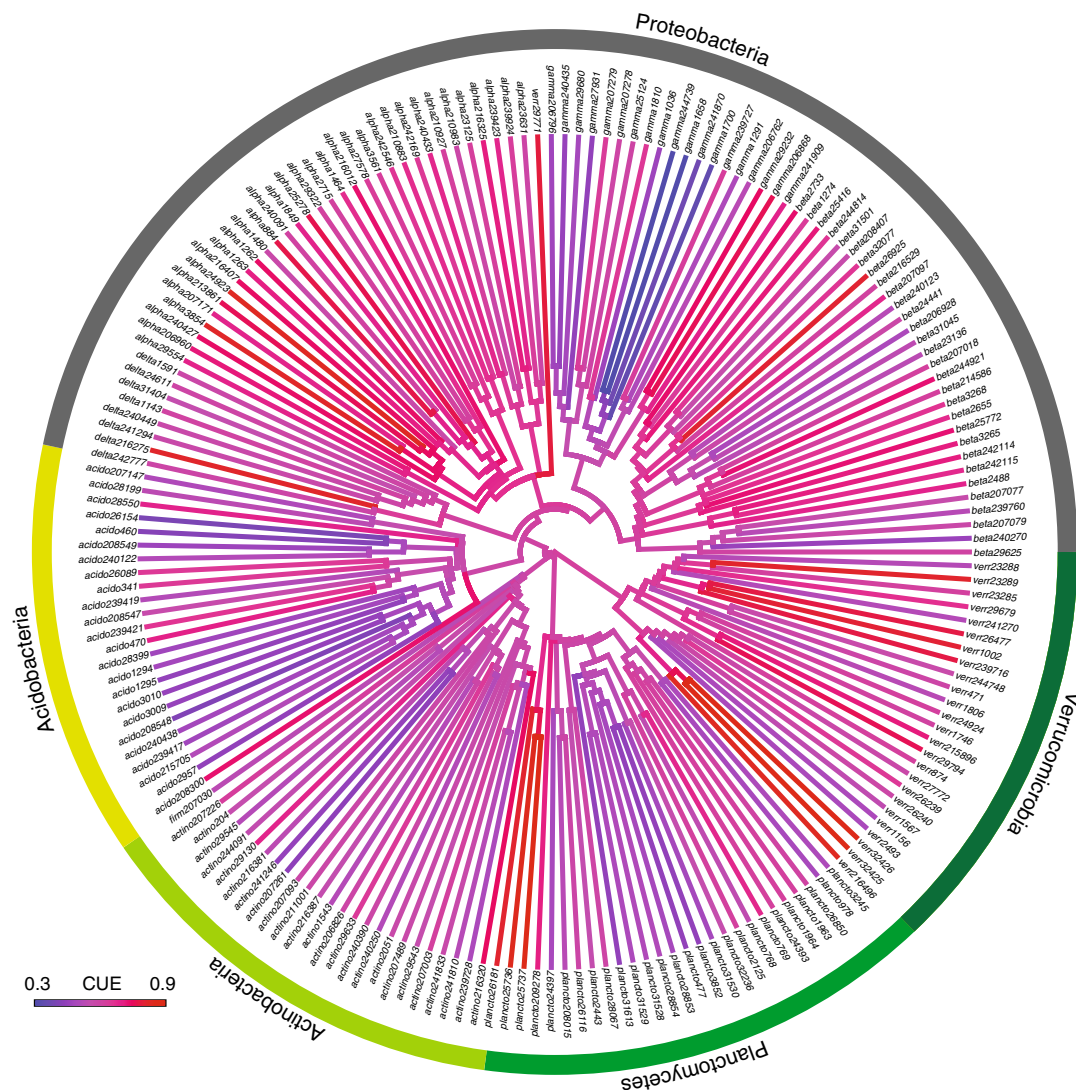
Fig. 3 Phylogenetic heatmap of potential CUE. Phylogenetic heatmap of potential CUE values from draft metabolic models. Labels on tips correspond to kBase accession numbers

communities has been modeled as declining 0.4 units over a range of 25 °C[5], while we observed over 0.6 units of variation in potential CUE between individual bacterial taxa. We detected a significant phylogenetic signal in potential CUE corresponding to clustering at sub-phylum levels, and we found that potential CUE was negatively correlated with particular genome traits, including genome size and GC content. In addition, we identified a particular set of amino acids, dipeptides, fatty acids, and carbohydrates that resulted in large reductions in CUE when their availability was constrained. Finally, we found that the range of variation we observed in CUE across taxa could have major implications for estimates of respiration and C storage at the ecosystem level.

Overall, we observed a mean potential CUE of 0.62 ± 0.17 (Fig. 1), which may represent a mean maximum CUE for bacteria in the absence of resource limitation. Potential CUE values represent intrinsic variation in CUE based on genomic differences between taxa, and these values were most useful for comparisons between taxa and for identifying relationships between genome traits and CUE. The range of observed potential CUE values from this analysis corresponds to the high end of parameter settings currently used in microbial models of the C cycle (0.15–0.68). However, empirical measurements of CUE extend well above

this mean in the absence of resource limitation[9]. When the availability of metabolites was constrained to reflect more reasonable expectations of resource limitation in the environment, we observed consistent declines of approximately 53% with CUE averaging 0.29 ± 0.19 (Supplementary Table 2).

Potential CUE varied across bacterial lineages, although not at the phylum-level. The significant phylogenetic signal in potential CUE indicates a Brownian pattern of trait evolution, with closely related taxa showing similarity in potential CUE values. However, we did not observe significant differences in potential CUE between bacterial phyla, and the greatest level of variation was structured at finer phylogenetic resolutions, including the class and order levels (Supplementary Table 3). Similar conclusions warning against broad phylum-level generalizations regarding carbon use traits have emerged from recent work using stable-isotope approaches[22].

Certain genomic traits, such as GC content and genome size, can be useful predictors of bacterial niche preferences and the response of bacterial communities to environmental changes[12,25,34]. Bacteria with larger genomes must allocate greater resources towards maintenance, while smaller genomes can exhibit greater efficiency[35]. In soil, taxa with larger genomes tend to dominate communities where substrates are available in
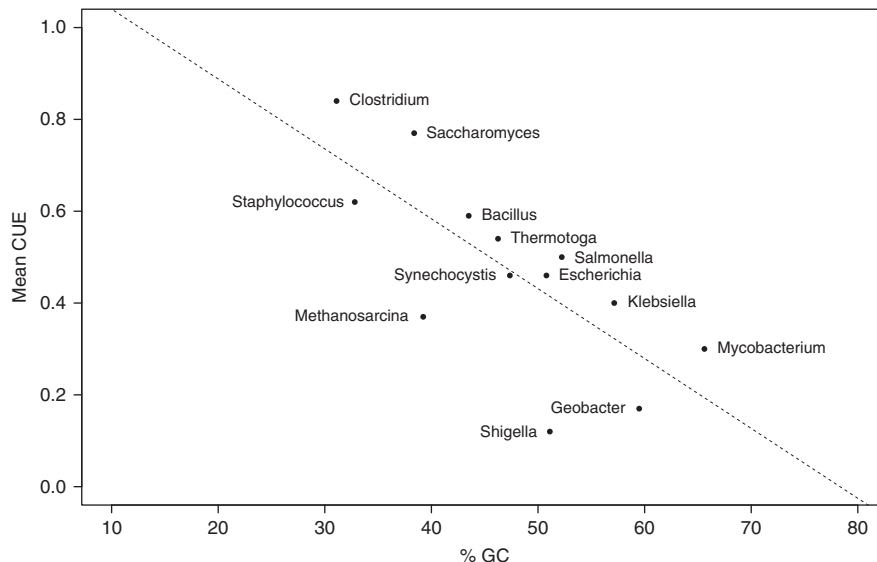
Fig. 4 Mean CUE versus GC content. Mean CUE versus GC content for manually curated metabolic models. Species in order of increasing GC content are *Clostridium ljungdahlii DSM 13528, Staphylococcus aureus subsp. aureus N315, Saccharomyces cerevisiae S288c, Methanosarcina barkeri str. Fusaro, Bacillus subtilis subsp. subtilis str. 168, Thermotoga maritima MSB8, Synechocystis sp. PCC 6803, Escherichia coli str. K-12 substr. MG1655, Shigella boydii Sb227, Salmonella enterica subsp. enterica serovar Typhimurium str. LT2, Klebsiella pneumoniae subsp. pneumoniae MGH 78578, Geobacter metallireducens GS-15, Mycobacterium tuberculosis H37Rv*. Mean CUE was calculated from CUE on growth on each of the following C-sources individually: D-Glucose, Fumarate, Acetate, Acetaldehyde, 2-Oxoglutarate, Ethanol, Formate, D-Fructose, L-Glutamine, L-Glutamate, D-lactate, L-Malate, Pyruvate, Succinate
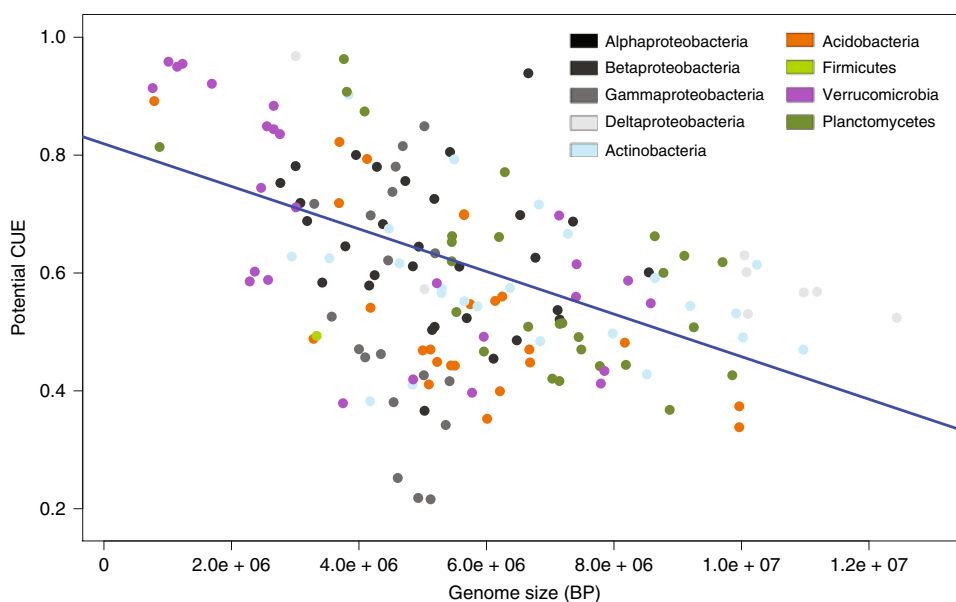

Fig. 5 Potential CUE versus genome size. Potential CUE regressed against genome size (bp). Blue lines show GLS fit. Points are colored by phylum

diverse forms but with limiting concentrations[26], and bacterial taxa with larger genomes tend to occupy a wider range of habitat types[25]. In our analysis, potential CUE declined by 0.04 units per additional Mbp in a genome (Fig. 5). However, there was a tradeoff between efficiency and access to substrates, as taxa with larger genomes were able to access a larger breadth of C sources at the cost of reductions in potential CUE. Thus, taxa with the highest CUEs may be less adaptable to changes in substrate chemistry, representing a more specialized trophic strategy. Prior studies also observe copiotrophic taxa having large numbers of genes associated with transport proteins, which would correspond to large numbers of transport reactions associated with C uptake and secretion[34].

Nutrient limitation can lead to shifts in community composition that favor GC-poor genomes, potentially due to the greater energetic cost of producing GTP and CTP bases[36]. Consistent with these findings, we observed a strong negative correlation between CUE and GC content ($R^2 = 0.522$; Fig. 4). Thus, environmental changes that favor GC-poor genomes may also have ramifications for C cycling through correlated increases in CUE and corresponding reductions in CO_2 emissions.

In the environment, microbial taxa are exposed to variation in substrate chemistry and availability, which can impact rates of C uptake and growth[37]. In prior studies, CUE shows sensitivity to substrate availability and stoichiometry at both the organismal[9,17] and community levels[8,28]. In our analysis, we identified several
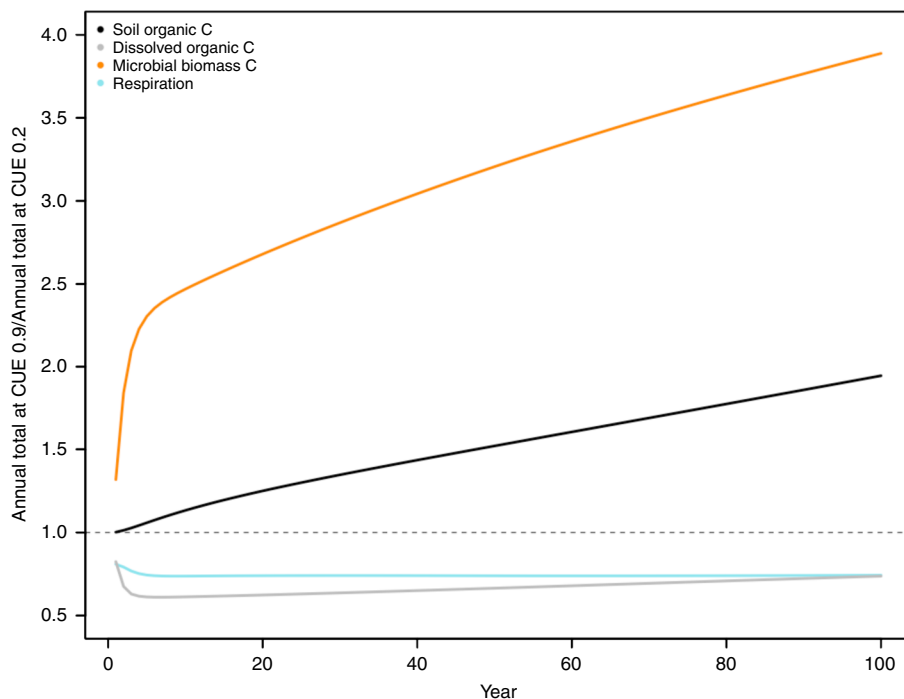
Fig. 6 Ecosystem C stocks and fluxes with variable microbial communities. Annual totals for C cycle pools and respiration rates for models for high efficiency taxa (CUE = 0.9) relative to low efficiency taxa (CUE = 0.2) across 100 years. Dashed line represents no difference in model estimates at the two CUE values

specific amino acids and dipeptides whose availabilities limited CUE. These findings comport with patterns of amino acid uptake by bacteria in the environment[38,39] and the incorporation of amino acids, such as alanine, directly into cell wall components. Amino acids represent a key input of N in soil[40], and rapid uptake by microbes results in short residence time of these compounds in soil[41,42]. It is hypothesized that microbes consume amino acids primarily as a C source, which may support the large impact of constraining amino acid availability on CUE we observed. Similarly, dipeptides contain higher C:N ratios than their component amino acid monomers, and their uptake is greater than that of amino acids[43].

Structured variability in soil organic matter, chemistry in soil could favor particular bacterial taxa over others based on their capacity to consume available C sources. In our analysis, taxa with the ability to consume a wide range of metabolites showed the lowest potential CUE values because of increased uptake of non-essential C-containing metabolites (Supplementary Table 4). In contrast, taxa with fewer exchange reactions were able to maintain higher CUE in the potential environment through reduced C uptake. These differences may be related to differences between copiotrophs and oligotrophs in terms of resource specialization, with less-specialized copiotrophic taxa showing lower CUE. Prior studies also observe copiotrophs having large numbers of genes associated with transport proteins, which would correspond to large numbers of C-containing exchange reactions in this analysis[34].

Accounting for the variation in CUE, we observed across taxa can have significant consequences to ecosystem-level estimates of C pool sizes and respiration rates (Fig. 6). Soil organic C pool sizes were reduced by almost half when a community shift towards low efficiency bacteria (CUE = 0.2, Fig. 6) was modeled compared to a community comprised of high efficiency bacteria (CUE = 0.9). These values represent the extremes of our potential CUE observations and therefore represent the widest range of expected outcomes. The change in soil organic C pool sizes we

observed was driven, in part, by large sustained increases in microbial biomass, with the highly efficient microbial communities producing nearly four times greater microbial biomass than low efficiency communities over the same time span. Despite this large increase in microbial biomass, rates of respiration were reduced by 25% compared to the low efficiency communities (Fig. 6). Thus, accounting for variation in CUE among taxa alone can have significant consequences to ecosystem-level estimates of C storage and respiration rates, and these differences can persist even at decadal timescales. While DAMM-MCNiP was parameterized and validated for a specific temperate forest ecosystem, the general model structure and its dependence on CUE are representative of soil C models used across several ecosystem types[44], suggesting that variation in CUE across taxa is likely to have important implications for soil C cycling more broadly. In this analysis, only CUE was varied across model runs to isolate the impact of variation in microbial physiology while holding other parameter settings constant. However, shifts in microbial composition and physiology may also be concomitant with shifts in other parameters, such as those describing enzyme activities, making it critical to characterize these potential interacting effects to fully describe parameter uncertainty in the future.

Direct comparisons between the values observed here and those in other studies are challenging as potential CUE may not have exact parallels to empirical observations in which CUE has been measured for a small number of individual taxa or complex soil communities. We interpret potential CUE values as intrinsic variation based on genetic differences between taxa that may be most useful in terms of exploring comparisons between taxa and for identifying relationships between genome traits and CUE. Despite the challenges with measuring CUE, developing empirical approaches[11,23] to directly estimate taxa-specific CUE will be necessary and useful for validating these observations.

It is important to note that the limited capacity of ascribing functions to genes through annotation pipelines, the poor knowledge of taxon-specific microbial biomass composition, and

the need to implement a gap-filling algorithm to compensate for missing reactions in genome scale reconstructions can each impact flux estimates generated through FBA models. Despite these limitations, models generated through the pipeline used for our draft predictions have been demonstrated to closely match empirical phenotype data[32], have been used to explore complex community interactions[45], and have been shown to successfully predict community structure and environmental metabolomics[46]. We expect that future advancements in genome annotation and metabolic model construction paired with taxa-specific empirical observations of CUE could use the same conceptual framework proposed here to provide predictions with improved precision and fidelity. Collecting empirical measurements of taxa-specific bacterial CUE across a range of substrate types using a consistent methodology is recommended for further validation of these hypotheses.

The range of potential CUE values we observed between taxa is comparable to that observed in other studies in which wide ranges of CUE are attributed to differences between ecosystems or due to abiotic factors[10]. Soil microbial communities undergo shifts in composition under global change[47], and these changes may alter the overall soil microbial CUE if particular taxa with uniquely high or low CUE values are favored based on growth strategy or substrate preference. Failing to account for relationships between CUE and microbial community composition may cause ecosystem models to miss important biotic feedbacks that can impact respiration fluxes and soil-C balance[6]. This analysis explores a method for generating draft predictions of taxa-specific CUE from metabolic models and identifies genome size and GC content as traits that may link genomic variation with C utilization strategies. We show that large phylogenetic variation in CUE between individual taxa make microbial physiology and community composition important factors to consider when estimating microbial contributions to C cycling.

Methods

Metabolic modeling. Genome-scale metabolic modeling (also known as stoichiometric modeling, or constraint-based modeling) can be used to quantitatively analyze the complete set of metabolic reactions in an organism. This approach has been successfully used to represent bacterial metabolism and growth patterns[32,48], uptake and secretion[49,50], and complex community interactions[33] in silico. The metabolic model for a given organism can be generated by extracting the list of all biochemical reactions available to an organism from its annotated genome. In addition to intracellular reactions, the model includes exchange reactions, which involve uptake or secretion of metabolites, either through genome-encoded transporters, or expected free diffusion through the membrane. For convenience of subsequent mathematical analysis, this list is converted into a stoichiometric matrix, $\mathbf{S}$, whose element S_{ij} corresponds to the stoichiometric coefficients of metabolite i in reaction j. Due to incomplete genome annotations, gapfilling is often required to supplement models with additional reactions before models are capable of producing a nonzero biomass flux.

Genome-scale metabolic models were selected for analysis from two separate databases. Thirteen microbial models were selected from the Biochemically, Genetically, and Genomically structured knowledgebase of metabolic reconstructions (BiGG), which contains a small set of manually curated metabolic models from diverse environments[31]. We utilized the following thirteen taxa in this analysis: *Clostridium ljungdahlii DSM 13528, Staphylococcus aureus subsp. aureus N315, Saccharomyces cerevisiae S288c, Methanosarcina barkeri str. Fusaro, Bacillus subtilis subsp. subtilis str. 168, Thermotoga maritima MSB8, Synechocystis sp. PCC 6803, Escherichia coli str. K-12 substr. MG1655, Shigella boydii Sb227, Salmonella enterica subsp. enterica serovar Typhimurium str. LT2, Klebsiella pneumoniae subsp. pneumoniae MGH 78578, Geobacter metallireducens GS-15,* and *Mycobacterium tuberculosis H37Rv*. Mean CUE was calculated from CUE on growth on each of the following C-sources individually: D-Glucose, Fumarate, Acetate, Acetaldehyde, 2-Oxoglutarate, Ethanol, Formate, D-Fructose, L-Glutamine, L-Glutamate, D-lactate, L-Malate, Pyruvate, Succinate. Due to the limited number of manually curated microbial metabolic models, we expanded our approach to include models generated using automated pipelines (described below) for over 200 bacterial taxa from phyla commonly observed in soil environments.

Genome selection. The BiGG database primarily includes microbial models associated with the human microbiome, limiting our capacity to extrapolate our

findings from these well-curated metabolic models to environmental microbial communities. We expanded our approach to target bacterial genomes belonging to phyla commonly observed in soil environments[19,47], which are of particular interest due to their major contributions to global respiration. We queried the Department of Energy's kBase for over 200 taxa and used automated pipelines to construct a large set of draft metabolic models.

The Department of Energy systems biology knowledgebase[51] (kBase) was searched in March 2016 for bacterial genomes belonging to phyla that have been observed to dominate forest soil bacterial community composition based on 16S ribosomal RNA and DNA sequencing[15,47]. A total of 23,530 genomes belonging to the six selected phyla were identified in kBase, corresponding to 1064 unique genera. For each phylum, at least 25 genomes were selected for analysis. For phyla with more than 50 available genomes, the full list of unique genera was scanned to target genera that have been observed in soil environments when possible (Supplementary Fig. 2).

The Build Metabolic Model tool was used in kBase to generate metabolic models from 231 selected genomes[32]. Model construction in kBase involves functional annotation of the genome to identify metabolic genes and their associated biochemical reactions using the *Rapid Annotation* of microbial genomes using *Subsystems Technology* (RAST) genome annotation pipeline and the model SEED framework[32,52]. Draft metabolic models were gapfilled using the Gapfill Metabolic Model tool in kBase to add the minimal set of reactions required to produce biomass on complete media, which contains all possible metabolites available for uptake[53]. Gapfilling on complete media results in conservative gapfilling by assuming that metabolites necessary for growth but not produced intracellularly based on genome annotation are available in the environment.

Flux balance analysis. Flux balance analysis (FBA) allows for an estimation of metabolic fluxes, such as rates of C uptake and utilization, through a metabolic model based on linear optimization of a specified objective function, such as biomass production. FBA makes a steady-state assumption, circumventing the need for knowledge of kinetic parameters, and uses the stoichiometry of metabolic reactions to determine the feasible space of all possible combinations of reaction rates. By prescribing an optimization scheme, it is possible to identify specific points in this feasible space, resulting in putative predictions of all metabolic reaction rates in the organism, including uptake and secretion fluxes and growth. This approach requires specification of (1) a flux or set of fluxes to maximize (or minimize) and (2) upper and lower bounds for all reactions within the metabolic model. Upon specification of these inputs, FBA is able to estimate the particular combination of fluxes through all reactions in the model that satisfy the given conditions. FBA was performed in MATLAB R2014a using the *optimizeCbModel* command in the *COnstraint-Based Reconstruction and Analysis* (COBRA) Toolbox[54]. All FBA analyses were set to maximize bacterial biomass production in this analysis, in accordance with standard FBA assumptions[33].

C use efficiency. C use efficiency (CUE) is calculated as the proportion of C retained in biomass relative to total C uptake (Eq. 1). For a metabolic model with n exchange reactions, and where C is equal to the number of C atoms taken up or secreted in a given reaction:

$$\text{CUE} = \frac{\sum_1^n \text{Uptake flux}_i \times C_i - \sum_1^n \text{Secretion flux}_i \times C_i}{\sum_1^n \text{Uptake flux}_i \times C_i} \qquad (1)$$

For the set of manually curated models, the availability of one of 14 individual C sources was manipulated, and CUE was calculated under exclusive uptake of each metabolite separately. For the larger set of models from kBase, CUE was explored under two scenarios. (1) potential CUE was calculated by allowing a model to utilize all exchange reactions present, and (2) constrained CUE was calculated by limiting the availability of a single C-containing metabolite relative to the availability of all other metabolites. Potential CUE was calculated to explore intrinsic metabolic variation in CUE, and these values were most useful for comparisons between taxa and for identifying relationships between genome traits and CUE. All exchange reactions present in a model were made available for uptake by allowing for a default maximum flux of 1000 mmol grDW^{-1} h^{-1}, where grDW indicates the cellular biomass dry weight in grams. As CUE was calculated as a ratio of fluxes, values were not sensitive to the order of magnitude of maximum flux bounds as long as these were consistent across reactions. Certain models produced a respiration flux of 0 mmol grDW^{-1} h^{-1} and were excluded from subsequent analyses of CUE.

To calculate CUE under conditions of limited substrate availability, reactions in each metabolic model were first classified according to the following hierarchy: (1) exchange, (2) C-containing, (3) utilized when available, (4) essential to biomass production, and (5) constraining to biomass production (Supplementary Fig. 1). For a given model, all C-containing exchange reactions with a nonzero flux under potential conditions were classified as utilized. The maximum uptake flux for each individual utilized reaction was then set to 0 mmol grDW^{-1} h^{-1} and FBA was performed again to identify reactions that were essential for biomass production. Finally, maximum uptake for all essential reactions was individually set to 5% of the maximum uptake flux for all other metabolites (50 mmol grDW^{-1} h^{-1}), and FBA was performed again to detect the impact of constraining

particular essential reactions. Reactions that resulted in a reduction of the biomass flux by at least 5% were classified as constraining, meaning that the biomass production flux showed a direct response to the availability of metabolites dictated by these reactions.

Uptake fluxes for the most commonly occurring constraining reactions across all models were analyzed to determine the response of biomass production relative to availability for each metabolite (biomass/uptake). For 18 of the most commonly constraining reactions, the uptake flux corresponding to a 75% reduction in the biomass flux was identified for each model. For all models containing a given constraining reaction, FBA was performed after setting the maximum uptake flux for the constraining reaction to this reduced value while leaving all other exchange reaction fluxes potential. Constrained CUE was then calculated according to equation (1). Constrained CUE was compared to potential CUE for all models with a given constraining reaction using paired T-tests and Cohen's D calculated from the *lsr* package[55] in R Studio[56].

Model evaluation and empirical comparisons. To test the sensitivity of our results to the method of gapfilling, two parallel sets of models were constructed for each taxon. One set of models were gapfilled to achieve a minimum biomass flux of 0.1 mmol grDW^{-1} h^{-1} while a second set was more heavily gapfilled to achieve a (default) minimum biomass flux of 1000 mmol grDW^{-1} h^{-1}. A total of 246 exchange reactions, including 211 C-containing exchange reactions, were observed across 231 models gapfilled to the lower biomass threshold. A total of 318 exchange reactions, including 279 C-containing exchange reactions, were observed across 231 models gapfilled to the higher biomass threshold. On average, models gapfilled to the higher biomass threshold had only 8 additional C-containing exchange reactions. Gapfilling intensity had a significant impact on subsequent calculations including CUE, but inter-model comparisons and rank order were not strongly affected by gapfilling intensity. Potential CUE values calculated from models gapfilled at the two intensities were strongly correlated (Pearson's rank correlation coefficient = 0.7).

Phylogenetic analyses. The Build Phylogenetic Tree tool was used in the DOE kBase to generate a phylogenetic tree for 220 of the 231 genomes analyzed based on 49 highly conserved clusters of orthologous group (COG) families and the FastTree maximum likelihood method[57]. Branch lengths were computed according to the Grafen method[58] using the *compute.brlen* command in the *ape* package[59] in R Studio. To test for phylogenetic signals, Blomberg's K statistic[60] was calculated using the *multiPhylosignal* function in the *Picante* package[61] in R Studio. Mean differences in potential CUE between phyla were compared using phylogenetic ANOVA with the *phylANOVA* function in the *phytools* package[62] in R Studio.

In order to determine the taxonomic level which best describes variation in potential CUE, we used Phylocom to calculate the contribution index (CI) for each of the 191 nodes in the bacterial phylogeny[63]. The CI indicates how much a particular node on the phylogeny accounts for the total variation in potential CUE[64]. After calculating the CI for all 191 nodes in our analysis, we classified the subset of nodes where collective contributions accounted 90% of the variation in potential CUE based on the taxonomic level at which descendent species diverged.

The relationship between potential CUE and (1) the number of exchange reactions, (2) the number of C-containing exchange reactions, (3) genome size, (4) guanine-cytosine (GC) content, and (5) number of genes was assessed using phylogenetic generalized least-squares regression with the *nlme* package[65] in R Studio. The proportion of variance explained by each predictor was estimated using a pseudo R^2 value designed for nonlinear regression[66] using the *r.squaredLR* function in the *MuMIn* package[67], which estimates the improvement of the fit model relative to a null model based on a likelihood ratio test. The *stepAIC* function in the *MASS* package[68] was additionally used to determine the simplest regression model with multiple predictors.

Ecosystem modeling. The Dual-Arrhenius Michaelis-Menten Microbial C and Nitrogen Physiology (DAMM-MCNiP) model was used to estimate potential impacts of the observed variation in CUE on ecosystem-level C fluxes. DAMM-MCNiP models the effects of soil moisture and temperature on coupled C and nitrogen fluxes through soil pools and microbial biomass[69]. Specifically, the model uses Michaelis-Menten kinetics to describe the depolymerization of soil organic C and soil organic Nitrogen by microbial extracellular enzymes to produce dissolved organic C (DOC) and dissolved organic N (DON). Maximum reaction velocities are governed by temperature-sensitive Arrhenius functions. Uptake of DOC and DON by microbial biomass is governed by a second series of Arrhenius and Michaelis-Menten kinetic equations, which are sensitive to moisture-mediated O_2 availability. Following uptake, the parameter CUE is used to determine the partitioning of C between microbial biomass and soil respiration. DAMM-MCNiP has been parameterized to describe seasonal patterns in heterotrophic soil respiration at a temperate forest site, and is able to capture 56% of variation in empirical observations of seasonal heterotrophic respiration at an hourly scale[69] (RMSE = 0.25, R^2 adjusted = 56). The published model uses a default CUE value of 0.3 as used in several other ecosystem models. We modified the parameterization of CUE

in this model (while retaining all other parameter settings as described in detail in Abramoff et al. 2017, Supplementary Table 5) to reflect the range of variation observed in CUE across taxa. We then quantified the impact of this variation on model estimates over 100 repeated cycles of annual variation in temperature and moisture to assess long-term impacts.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
Data used in this analysis are available in the supplementary material and additionally available upon request.

References
1. Bond-Lamberty, B. & Thomson, A. Temperature-associated increases in the global soil respiration record. *Nature* **464**, 579–582 (2010).
2. Hashimoto, S. et al. Global spatiotemporal distribution of soil respiration modeled using a global database. *Biogeosciences* https://doi.org/10.5194/bg-12-4121-2015 (2015).
3. Bond-Lamberty, B., Wang, C. & Gower, S. T. A global relationship between the heterotrophic and autotrophic components of soil respiration?. *Glob. Chang. Biol.* https://doi.org/10.1111/j.1365-2486.2004.00816.x (2004).
4. Tian, H. et al. Global patterns and controls of soil organic carbon dynamics as simulated by multiple terrestrial biosphere models: Current status and future directions. *Glob. Biogeochem. Cycles* https://doi.org/10.1002/2014GB005021 (2015).
5. Allison, S. D., Wallenstein, M. D. & Bradford, M. A. Soil-carbon response to warming dependent on microbial physiology. *Nat. Geosci.* **3**, 336–340 (2010).
6. Li, J., Wang, G., Allison, S. D., Mayes, M. A. & Luo, Y. Soil carbon sensitivity to temperature and carbon use efficiency compared across microbial-ecosystem models of varying complexity. *Biogeochemistry*. https://doi.org/10.1007/s10533-013-9948-8 (2014).
7. Geyer, K. M., Kyker-Snowman, E., Grandy, A. S. & Frey, S. D. Microbial carbon use efficiency: accounting for population, community, and ecosystem-scale controls over the fate of metabolized organic matter. *Biogeochemistry*. https://doi.org/10.1007/s10533-016-0191-y (2016).
8. Manzoni, S., Taylor, P., Richter, A., Porporato, A. & Ågren, G. I. Environmental and stoichiometric controls on microbial carbon-use efficiency in soils. *New Phytologist* https://doi.org/10.1111/j.1469-8137.2012.04225.x (2012).
9. Keiblinger, K. M. et al. The effect of resource quantity and resource stoichiometry on microbial carbon-use-efficiency. *FEMS Microbiol. Ecol.* https://doi.org/10.1111/j.1574-6941.2010.00912.x (2010).
10. Sinsabaugh, R. L. et al. Stoichiometry of microbial carbon use efficiency in soils. *Ecol. Monogr.* https://doi.org/10.1890/15-2110.1 (2016).
11. Bölscher, T., Wadsö, L., Börjesson, G. & Herrmann, A. M. Differences in substrate use efficiency: impacts of microbial community composition, land use management, and substrate complexity. *Biol. Fertil. Soils*. https://doi.org/10.1007/s00374-016-1097-5 (2016).
12. Roller, B. R. K., Stoddard, S. F. & Schmidt, T. M. Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. *Nat. Microbiol.* https://doi.org/10.1038/nmicrobiol.2016.160 (2016).
13. Gelfand, I. et al. Sustainable bioenergy production from marginal lands in the US Midwest. *Nature* **493**, 7433 (2013).
14. Leff, J. W. et al. Consistent responses of soil microbial communities to elevated nutrient inputs in grasslands across the globe. *Proc. Natl Acad. Sci.* https://doi.org/10.1073/pnas.1508382112 (2015).
15. Fierer, N., Bradford, M. A. & Jackson, R. B. Toward an ecological classification of soil bacteria. *Ecology*. https://doi.org/10.1890/05-1839 (2007).
16. Sistla, S. A., Rastetter, E. B. & Schimel, J. P. Responses of a tundra system to warming using SCAMPS: A stoichiometrically coupled, acclimating microbeplantsoil model. *Ecol. Monogr.* https://doi.org/10.1890/12-2119.1 (2014).
17. Roller, B. R. K. & Schmidt, T. M. The physiology and ecological implications of efficient growth. *ISME J.* https://doi.org/10.1038/ismej.2014.235 (2015).
18. Frey, S. D., Drijber, R., Smith, H. & Melillo, J. Microbial biomass, functional capacity, and community structure after 12 years of soil warming. *Soil Biol. Biochem.* **40**, 2904–2907 (2008).
19. Fierer, N. et al. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc. Natl Acad. Sci.* https://doi.org/10.1073/pnas.1215210110 (2012).

20. Bernard, L. et al. Dynamics and identification of soil microbial populations actively assimilating carbon from13C-labelled wheat residue as estimated by DNA- and RNA-SIP techniques. *Environ. Microbiol.* https://doi.org/10.1111/j.1462-2920.2006.01197.x (2007).

21. Bastian, F., Bouziri, L., Nicolardot, B. & Ranjard, L. Impact of wheat straw decomposition on successional patterns of soil microbial community structure. *Soil Biol. Biochem.* https://doi.org/10.1016/j.soilbio.2008.10.024 (2009).

22. Morrissey, E. M. et al. Phylogenetic organization of bacterial activity. *ISME J.* https://doi.org/10.1038/ismej.2016.28 (2016).

23. Morrissey, E. M. et al. Bacterial carbon use plasticity, phylogenetic diversity and the priming of soil organic matter. *ISME J.* https://doi.org/10.1038/ismej.2017.43 (2017).

24. Martiny, A. C., Treseder, K. & Pusch, G. Phylogenetic conservatism of functional traits in microorganisms. *ISME J.* https://doi.org/10.1038/ismej.2012.160 (2013).

25. Barberán, A. et al. Why are some microbes more ubiquitous than others? Predicting the habitat breadth of soil bacteria. *Ecol. Lett.* https://doi.org/10.1111/ele.12282 (2014).

26. Konstantinidis, K. T. & Tiedje, J. M. Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc. Natl Acad. Sci.* https://doi.org/10.1073/pnas.0308653100 (2004).

27. Dijkstra, P. et al. High carbon use efficiency in soil microbial communities is related to balanced growth, not storage compound synthesis. *Soil Biol. Biochem.* https://doi.org/10.1016/j.soilbio.2015.06.021 (2015).

28. Sinsabaugh, R. L., Manzoni, S., Moorhead, D. L. & Richter, A. Carbon use efficiency of microbial communities: Stoichiometry, methodology and modelling. *Ecol. Lett.* https://doi.org/10.1111/ele.12113 (2013).

29. Thiele, I. & Palsson, B. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* https://doi.org/10.1038/nprot.2009.203 (2010).

30. O'Brien, E. J., Monk, J. M. & Palsson, B. O. Using genome-scale models to predict biological capabilities. *Cell* https://doi.org/10.1016/j.cell.2015.05.019 (2015).

31. King, Z. A. et al. BiGG Models: a platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkv1049 (2016).

32. Henry, C. S. et al. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* https://doi.org/10.1038/nbt.1672 (2010).

33. Harcombe, W. R. et al. Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Rep.* https://doi.org/10.1016/j.celrep.2014.03.070 (2014).

34. Lauro, F. M. et al. The genomic basis of trophic strategy in marine bacteria. *Proc. Natl Acad. Sci.* https://doi.org/10.1073/pnas.0903507106 (2009).

35. Lynch, M. Streamlining and simplification of microbial genome architecture. *Annu. Rev. Microbiol.* https://doi.org/10.1146/annurev.micro.60.080805.142300 (2006).

36. Mann, S. & Chen, Y. P. P. Bacterial genomic G+C composition-eliciting environmental adaptation. *Genomics.* https://doi.org/10.1016/j.ygeno.2009.09.002 (2010).

37. Mooshammer, M. et al. Adjustment of microbial nitrogen use efficiency to carbon:Nitrogen imbalances regulates soil nitrogen cycling. *Nat. Commun.* https://doi.org/10.1038/ncomms4694 (2014).

38. Farrell, M. et al. Rapid peptide metabolism: a major component of soil nitrogen cycling? *Glob. Biogeochem. Cycles.* https://doi.org/10.1029/2010GB003999 (2011).

39. Broughton, R. C. I., Newsham, K. K., Hill, P. W., Stott, A. & Jones, D. L. Differential acquisition of amino acid and peptide enantiomers within the soil microbial community and its implications for carbon and nitrogen cycling in soil. *Soil Biol. Biochem.* https://doi.org/10.1016/j.soilbio.2015.05.003 (2015).

40. Jones, D. L. et al. Soil organic nitrogen mineralization across a global latitudinal gradient. *Glob. Biogeochem. Cycles.* https://doi.org/10.1029/2008GB003250 (2009).

41. Finzi, A. C. & Berthrong, S. T. The uptake of amino acids by microbes and trees in three cold-temperate forests. *Ecology.* https://doi.org/10.1890/04-1460 (2005).

42. Wilkinson, A., Hill, P. W., Farrar, J. F., Jones, D. L. & Bardgett, R. D. Rapid microbial uptake and mineralization of amino acids and peptides along a grassland productivity gradient. *Soil Biol. Biochem.* https://doi.org/10.1016/j.soilbio.2014.01.026 (2014).

43. Farrell, M., Prendergast-Miller, M., Jones, D. L., Hill, P. W. & Condron, L. M. Soil microbial organic nitrogen uptake is regulated by carbon availability. *Soil Biol. Biochem.* https://doi.org/10.1016/j.soilbio.2014.07.003 (2014).

44. Manzoni, S. & Porporato, A. Soil carbon and nitrogen mineralization: theory and models across scales. *Soil Biol. Biochem* **41**, 1355–1379 (2009).

45. Magnúsdóttir, S. et al. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat. Biotechnol.* https://doi.org/10.1038/nbt.3703 (2017).

46. Garza, D. R., Van Verk, M. C., Huynen, M. A. & Dutilh, B. E. Towards predicting the environmental metabolome from metagenomics with a mechanistic model. *Nat. Microbiol.* https://doi.org/10.1038/s41564-018-0124-8 (2018).

47. DeAngelis, K. M. et al. Long-term forest soil warming alters microbial communities in temperate forest soils. *Front. Microbiol.* https://doi.org/10.3389/fmicb.2015.00104 (2015).

48. Sridhara, V. et al. Predicting bacterial growth conditions from bacterial physiology. 0–26 https://doi.org/10.1101/002287 (2014).

49. Varma, A. & Palsson, B. O. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *Appl. Environ. Microbiol.* **60**, 3724–3731 (1994).

50. Edwards, J. S., Ibarra, R. U. & Palsson, B. O. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.* https://doi.org/10.1038/84379 (2001).

51. Arkin, A. P. et al. KBase: The United States department of energy systems biology knowledgebase. *Nat. Biotechnol.* https://doi.org/10.1038/nbt.4163 (2018).

52. Overbeek, R. et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkt1226 (2014).

53. Latendresse, M. Efficiently gap-filling reaction networks. *BMC Bioinform.* https://doi.org/10.1186/1471-2105-15-225 (2014).

54. Schellenberger, J. et al. Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox v2.0. *Nat. Protoc.* https://doi.org/10.1038/nprot.2011.308 (2011).

55. Navarro, D. J. *Learning Statistics with R: A Tutorial for Psychology Students and Other Beginners.* (2015).

56. R Core Team. *R: A Language and Environment for Statistical Computing.* (2017).

57. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2-Approximately maximum-likelihood trees for large alignments. *PLoS One.* https://doi.org/10.1371/journal.pone.0009490 (2010).

58. Grafen, A. The phylogenetic regression. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* https://doi.org/10.1098/rstb.1989.0106 (1989).

59. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics.* https://doi.org/10.1093/bioinformatics/btg412 (2004).

60. Blomberg, S. P., Garland, T. & Ives, A. R. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* https://doi.org/10.1111/j.0014-3820.2003.tb00285.x (2003).

61. Kembel, S. W. et al. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics.* https://doi.org/10.1093/bioinformatics/btq166 (2010).

62. Revell, L. J. phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* https://doi.org/10.1111/j.2041-210×.2011.00169.x (2012).

63. Webb, C. O., Ackerly, D. D. & Kembel, S. W. Phylocom: software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics.* https://doi.org/10.1093/bioinformatics/btn358 (2008).

64. Moles, A. T. et al. A brief history of seed size. *Science* https://doi.org/10.1126/science.1104863 (2005).

65. Pinheiro, J., Bates, D., DebRoy, S. & Sarkar, D. nlme: Linear and nonlinear mixed effects models. *R. Packag. version.* https://doi.org/10.1016/j.molliq.2016.04.126 (2007).

66. Burnham, K. P., Anderson, D. R., Burnham, K. P. & Anderson, D. R. Model selection and multimodel inference: a practical information-theoretic approach. 2nd edn (Springer-Verlag, New York, 2002).

67. Bartón, K. MuMIn: Multi-Model Inference. *R package version 1.40.4.* https://CRAN.R-project.org/package=MuMIn. Bates (2018).

68. Venables, W. N. & Ripley, B. D. Modern applied statistics with S. *Technometrics* https://doi.org/10.1198/tech.2003.s33 (2002).

69. Abramoff, R. Z., Davidson, E. A. & Finzi, A. C. A parsimonious modular approach to building a mechanistic belowground carbon and nitrogen model. *J. Geophys. Res. Biogeosci.* https://doi.org/10.1002/2017JG003796 (2017).

Author contributions
M.S. conducted metabolic modeling and data analysis, wrote and edited the paper; J.M.B. directed phylogenetic data analysis and edited the paper; D.S. directed metabolic modeling and edited the paper; A.C.F. conceived the idea, wrote and edited the paper.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-11488-z.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Peer review information: *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.