A sparse mapping of structure to function in microbial communities

Summary: Simple models quantitatively predict metabolite dynamics in denitrifying bacterial communities from gene content alone.

Karna Gowda,[1,2] Derek Ping,[3] Madhav Mani,[4,5,6*] Seppe Kuehn[1,2*]

[1]Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637, USA

[2]Center for the Physics of Evolving Systems, University of Chicago, Chicago, IL 60637, USA

[3]Department of Physics, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[4]Department of Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL 60208, USA

[5]Department of Molecular Biosciences, Northwestern University, Evanston, IL 60208, USA

[6]NSF-Simons Center for Quantitative Biology, Northwestern University, Northwestern University, Evanston, IL 60208, USA

*To whom correspondence should be addressed;
E-mail: madhav.mani@gmail.com, seppe.kuehn@gmail.com.

The metabolic function of microbial communities emerges through a complex hierarchy of genome-encoded processes, from gene expression to interactions between diverse taxa. Therefore, a central challenge for microbial ecology is deciphering how genomic structure determines metabolic function in communities. Here we show, for the process of denitrification, that community metabolism is quantitatively predicted from the genes each member of the community possesses. For each strain in a set of bacterial isolates, the dynamics of nitrate and nitrite reduction are quantitatively encoded in the presence or absence of denitrification genes. We correctly predict metabolite dynamics in communities using a consumer-resource model that sums the contribution of each strain. Our results enable predicting metabolite dynamics from metagenomes, designing denitrifying communities and discovering how genome evolution impacts metabolism.

Introduction

Microbial metabolism plays an essential role in sustaining life on Earth. Working collectively in complex communities, microbes are key players in global nutrient cycles (*1*), wastewater treatment (*2*) and human health (*3*). As such, a key challenge in microbial ecology is understanding how emergent community metabolism is determined by the taxonomic and genomic structure of a community. Addressing this structure-function problem is critical for functionally interpreting community gene content (*4*), elucidating the evolutionary principles of community metabolism (*5, 6*) and designing synthetic communities (*7*).

Understanding how community metabolism is genomically encoded requires mapping the genotypes of each community member to metabolic phenotypes, and then deciphering how interactions between distinct populations contribute to collective metabolism. Quantitatively mapping genotypes to metabolic phenotypes for naturally-occurring bacteria is challenging due to substantial genetic and phenotypic variation in the wild (*8*). Moreover, interactions within communities depend on extracellular metabolites (*9*), abiotic factors (*10*), cooperation (*11*) and higher-order effects (*12*). While constraint-based models have found some success in predicting collective metabolism from genomes (*13–15*), these methods require significant manual refinement (*16*), complicating the prospect of making predictions from the genomes of non-model organisms or metagenomes of communities.

Despite the complexities of relating structure to function in microbial communities, sequencing surveys have shown that the functional gene content of a community correlates with local metabolite concentrations and environmental variables (*17–19*), hinting that genomic structure might be predictive of metabolic properties at the community-level. In the laboratory, enrichment experiments have revealed that taxa with conserved metabolic phenotypes are assembled to degrade exogenously supplied organic carbon (*20, 21*). So while it appears that structure at the genomic level may endow communities with specific functional capabilities, we

cannot quantitatively predict community metabolic function from genomic structure.

To address this challenge, we posed a prediction problem: can the dynamic flux of metabolites through a microbial community be quantitatively predicted from knowledge of the genomes of each strain present? Exploiting a library of diverse naturally-isolated and sequenced bacteria, quantitative measurements, modeling of metabolite dynamics, and an interpretable statistical modeling framework, we showed that the flux of metabolites through a community can be quantitatively predicted from the presence and absence of genes in the relevant metabolic pathway.

Results

We used denitrification as a model metabolic process for mapping genomic structure to community metabolic function (Fig. 1A). Denitrification is a form of anaerobic respiration whereby microbes use oxidized nitrogen compounds as electron acceptors, driving a cascade of four successive reduction reactions, $NO_3^- \rightarrow NO_2^- \rightarrow NO \rightarrow N_2O \rightarrow N_2$ (*22*). As a biogeochemical process, denitrification is essential to nitrogen cycling at a global scale through activity in soils, freshwater systems, and marine environments (*23*), and impacts human health through activity in wastewater treatment plants (*2*) and in the human gut (*24*). The process is performed by taxonomically-diverse bacteria (*25*) that are typically facultative anaerobes. The denitrification pathway is known to be modular, with some strains performing all four steps in the cascade and others performing one or a nearly arbitrary subset of reduction reactions (*26*). Denitrification in nature is therefore a collective process, where a given strain can produce electron acceptors that can be utilized by other strains (*9*).

We focused experimentally on the first two steps of denitrification: the conversion of nitrate (NO_3^-) to nitrite (NO_2^-) and subsequently nitric oxide (NO) (Fig. 1A). Nitrate and nitrite are soluble, enabling high throughput measurements of metabolite dynamics. Rather than working only with well-studied model organisms, we isolated 61 diverse bacterial strains spanning α-, β-, and γ-proteobacteria from local soils using established techniques (Materials and Methods). Each strain was obtained in axenic culture and was characterized as performing one or both of the first two steps of denitrification in a chemically-defined, electron acceptor-limited medium containing a single non-fermentable carbon source (succinate). Each of these strains was therefore classified into one of three possible phenotypes (Fig. 1A): (1) Nar/Nir strains that perform both nitrate and nitrite reduction ($NO_3^- \rightarrow NO_2^- \rightarrow NO$), (2) Nar strains that perform only nitrate reduction ($NO_3^- \rightarrow NO_2^-$), and (3) Nir strains that perform only nitrite reduction ($NO_2^- \rightarrow NO$). In addition to these 61 isolates, our strain library also included the full denitrifier *Paracoccus denitrificans* ATCC 19367.

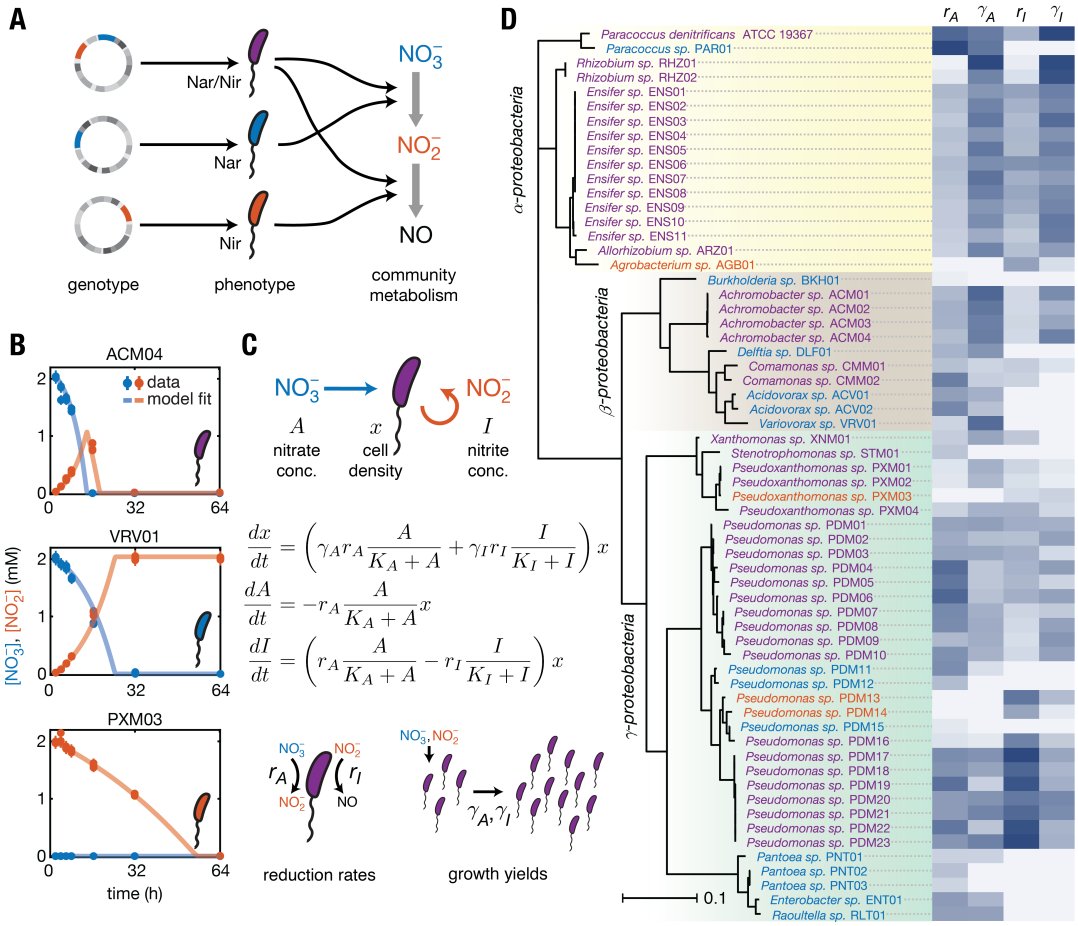

Figure 1: **Characterizing phenotypic variation of bacterial isolates.** (**A**) Genomic structure was statistically mapped to community metabolic function via a quantitative parameterization of individual strain phenotypes across a library of natural isolates. (**B**) Example batch culture metabolite dynamics for Nar/Nir, Nar, and Nir isolates, along with fits to a consumer-resource model (**C**). The model is parameterized by reduction rates r_A and r_I and yields γ_A and γ_I, for growth on nitrate and nitrite respectively (see Materials and Methods). (**D**) Phylogenetic tree and normalized consumer-resource parameters for 62 denitrifying strains. Scale bar indicates estimated number of substitutions per site of the 16S rRNA gene. Darker colors indicate larger values of the normalized parameters.

Parameterizing metabolite dynamics

We quantified the denitrification dynamics of each strain in our library in monoculture. To accomplish this, strains were inoculated at low starting densities into 96-well plates containing chemically-defined medium with either nitrate or nitrite provided as the sole electron acceptor, and then incubated under anaerobic conditions. Small samples ($10\,\mu\mathrm{L}$) were then taken at logarithmically-spaced time intervals over a period of $64\,\mathrm{h}$ and assayed for nitrate and nitrite concentrations (Materials and Methods, Fig. S1 and S2). At the end of the time course, optical density was assayed. The measurement resulted in a time series of nitrate and nitrite production/consumption dynamics in batch culture (points, Fig. 1B). For each strain in the library, these experiments were performed across a range of initial cell densities and nitrate/nitrite concentrations (Materials and Methods). We used these data to parameterize a simple consumer-resource model (Fig. 1C) describing denitrification dynamics for each strain in monoculture (Supplementary Text, Fig. S3 and S4). The model allowed us to quantitatively describe the phenotype of each strain in the library in terms of at most four parameters: r_A and r_I, which capture rates of nitrate and nitrite reduction, and γ_A and γ_I, which describe yields for nitrate and nitrite, respectively. Substrate affinities (K_*) were fixed to a small value since these parameters were not well constrained by the data (Supplementary Text, Fig. S5). The models for Nar and Nir strains correspond to setting $r_I = 0$ or $r_A = 0$, respectively. Yields (γ_*) were inferred using endpoint optical density measurements, and rates (r_*) were inferred by fitting the observed nitrate and nitrite dynamics to the consumer-resource model (Fig. 1C). Remarkably, with the exception of a small number of strains that were excluded from the library (Supplementary Text, Fig. S6), a single set of parameters quantitatively described metabolite dynamics for each strain across a range of initial cell densities and nitrate/nitrite concentrations (Supplementary Text, Fig. S4).

Fitting our consumer-resource model to data for each strain yielded a quantitative description of the dynamic metabolic phenotype of each strain in the library (Fig. 1B and D). We observed large variability between taxa, with coefficients of variation for both rates (r_A, r_I) and yields (γ_A, γ_I) around 60%. We also observed some patterns of phylogenetic conservation; for example α-proteobacteria produced generally higher yields than β- or γ-proteobacteria, and a clade of *Pseudomonas* sp. isolates showed consistently higher rates of nitrite reduction than most other strains (Fig. 1D). Despite these patterns, the prevalence of each of the three phenotypes is not strongly dependent on phylogeny, with each phenotype present across the tree (Fig. 1D). The latter observation is consistent with pervasive horizontal gene transfer of denitrifying enzymes (*27, 28*). Finally, we did not observe a correlation or trade-off between rates and yields (Fig. S7).

Predicting metabolite dynamics from genomes

Armed with a quantitative parameterization of the metabolite dynamics for all strains in our library, we next set out to characterize the measured phenotypic variation across strains using genomic differences between isolates. We first performed whole genome sequencing on all 62

strains in the library. Then we assembled and annotated each genome (Materials and Methods) and determined the complement of denitrification genes possessed by each strain, exploiting the fact that the molecular and genetic basis of denitrification is well-understood (*22*). We identified not only the reductases that perform the reduction of the oxidized nitrogen compounds, but also the sensors/regulators (*29*) and transporters (*30*) known to be involved in denitrification (Materials and Methods). The presence and absence of each gene (or set of genes encoding proteins that form a complex) in each genome is presented in Fig. 2A. Patterns of gene presence/absence agree well with known features of the denitrification pathway, including the mutual exclusion of the two reductases performing nitrite reduction (*NirS* and *NirK*) (*17, 28*). Further, in almost all cases strains possessing nitrate and/or nitrite reductase performed the associated reactions in culture (with the only exception being the Nar strain *Acidovorax* sp. ACV01, which possesses both nitrate and nitrite reductase). This is in agreement with previous work demonstrating that bacterial genomes lose nonfunctional content due to streamlining (*31*).

Next we showed that the presence and absence of denitrification genes in each strain was sufficient to quantitatively predict metabolite dynamics in monoculture. Specifically, we constructed a linear regression where the measured phenotypic parameters of our consumer-resource model, which faithfully capture the denitrification dynamics for each strain, were predicted on the basis of gene presence and absence (e.g., Fig. 2B). The regression coefficients for each gene in the pathway quantify the impact of the presence of the gene on a given phenotypic parameter. We used L_1-regularized regression (LASSO) to avoid overfitting (Supplementary Text, Fig. S8 to S10), performing independent regressions for each of the phenotypic parameters in our consumer-resource model. LASSO yielded sparse regression models, revealing that presence/absence of a small set of genes is highly predictive of the phenotypic parameters for all strains in our diverse library (Fig. 2C to J). The in-sample coefficients of determination (R^2_{fit}) of our regressions were between 0.50 and 0.76 depending on the phenotypic parameter. Crucially, our regression approach generalized out-of-sample, as determined by iterated cross-validation (Supplemental Text, Fig. S9), albeit with slightly lower predictive power ($\bar{R}^2_{CV}$ between 0.26 to 0.54). Since, in general, traits may exhibit phylogenetic correlation (*8*), and our library contains a few clades comprising very closely related strains (e.g., ENS01–08, PDM20–23, Fig. 1D), we considered whether our regression utilized phylogenetic correlations in gene presence/absence and denitrification phenotypic parameters to achieve predictive power. We investigated this by collapsing clades containing strains with identical 16S rRNA sequences down to a single randomly-selected representative, and performing regressions again on this reduced set of strains. For these regressions we found that the predictive power and coefficients were similar to those for the full dataset (Supplementary Text, Fig. S11), supporting the claim that our regression is not simply detecting phylogenetic correlations between traits and genotypes. Altogether these results demonstrate that, across a diverse set of natural isolates, knowledge of the genes a denitrifying strain possesses is sufficient to accurately predict the rates and biomass yields of that strain on nitrate and nitrite.

Our regression approach leveraged biological knowledge of the denitrification pathway to predict metabolite dynamics, in effect presuming that denitrification gene content is the best

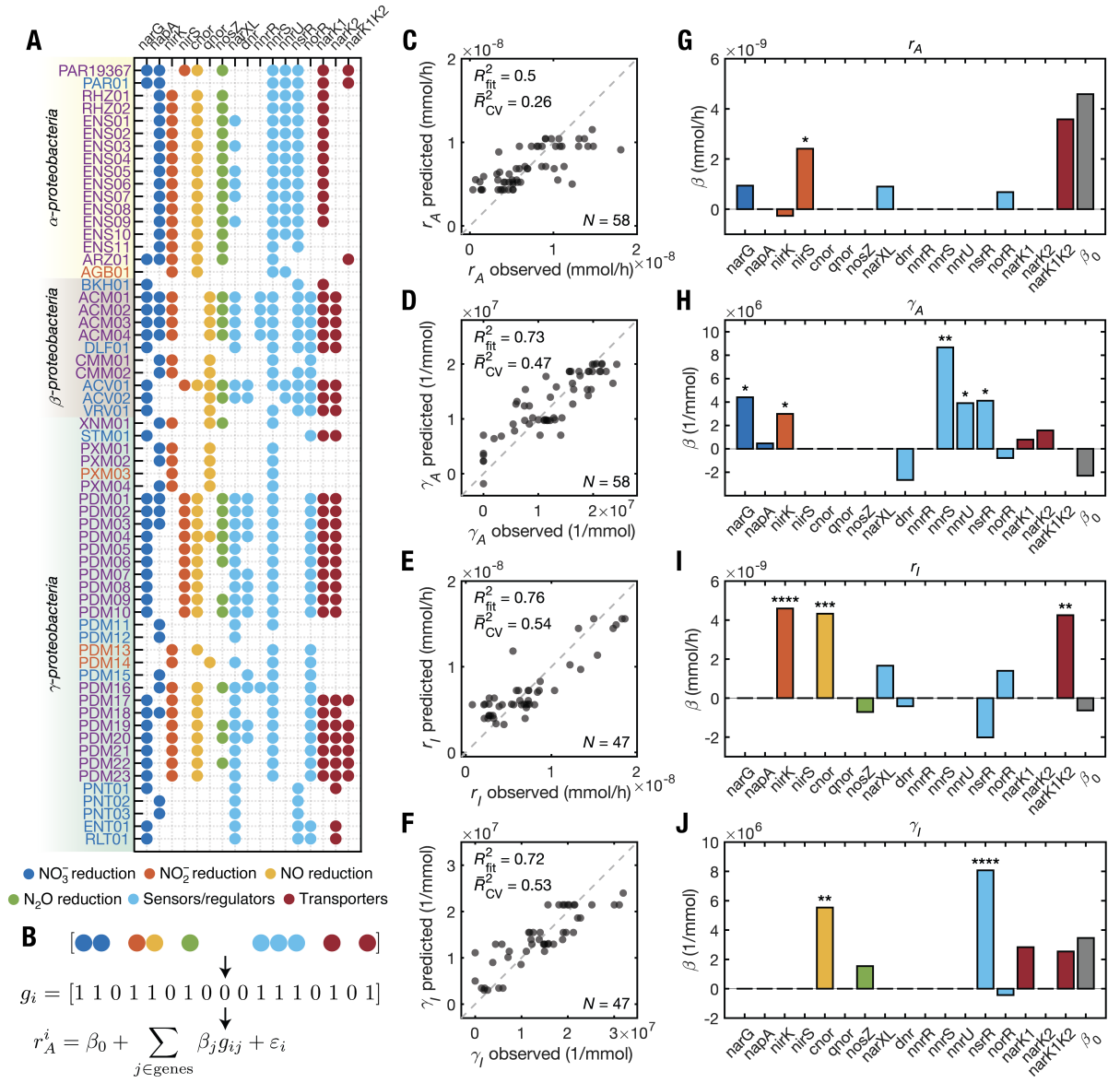

Figure 2: **Gene presence/absence predicts metabolite dynamics of individual strains.** (**A**) Denitrification gene presence/absence for 62 denitrifying strains. (**B**) Observed consumer-resource parameters for each strain (e.g., nitrate reduction rate r_A) were linearly regressed against gene presence/absence via L_1-regularized regression, resulting in regression coefficients β_j for each gene j, an intercept β_0, and a noise term ϵ_i for each observation i. (**C** to **F**) Regression applied to each consumer-resource parameter resulted in predicted values of the parameters for each strain, which are plotted against observed values. The in-sample coefficients of determination for these data (R^2_{fit}) and the out-of-sample coefficients of determination estimated via iterated cross-validation ($\bar{R}^2_{\text{CV}}$) are shown. (**G** to **J**) Regressions yielded estimated coefficients $\vec{\beta}$, along with intercepts β_0, which are shown in the bar charts. Asterisks indicate significance level for each regression coefficient (Supplementary Text, Fig. S10).

genomic feature for prediction. To investigate whether this assumption is correct, we asked whether other genomic properties could better predict metabolite dynamics. First, we tested the predictive capability of sets of randomly selected genes. We chose sets of 17 random genes that were not strongly correlated with any denitrification genes, but retained the same marginal frequency distribution as denitrification genes in the population. We found that regressions using these randomly-selected genes have, on average, much less predictive power than regressions using the denitrification genes (Supplementary Text, Fig. S12). We note that this result provided further evidence that regressions on denitrification gene presence/absence are not simply detecting phylogenetic correlations, since random genes would be expected to perform equally well on average if phylogenetic structure dominated the phenotypic parameters. Second, we tested whether 16S rRNA copy number or genome size improves the predictive ability of denitrification gene presence/absence regressions. 16S rRNA copy number has been observed to correlate positively with maximal growth rate in nutrient rich conditions (*32, 33*), and smaller genomes are associated with faster growth (*31, 33*). We found that these predictors do not meaningfully alter the regressions or improve their predictive ability (Supplementary Text, Fig. S13 and S14). In summary, our statistical analyses provided evidence that denitrification gene presence/absence outperforms arbitrary sets of genes and coarse genomic features.

Why did our sparse regression models select specific genes in the denitrification pathway to quantitatively predict metabolite dynamics? To address this question we examined the regression coefficients in the context of what is known about the denitrification pathway. We found that in many cases the sign and magnitude of the regression coefficients agree qualitatively with known mechanistic properties of the associated enzymes. Previous comparisons between membrane-bound and periplasmic nitrate reductases (encoded by *narG* and *napA*, respectively) in multiple bacterial species showed that the membrane-bound enzyme exhibits higher nitrate reduction activity *in vitro* than the periplasmic enzyme (Table S1). This accords with the large positive coefficient for *narG* we observed in the nitrate reduction rate regression (Fig. 2G). Similarly, in the nitrite reduction rate regression we observed a large positive coefficient for the gene encoding the copper-based nitrite reductase (*nirK*) (Fig. 2I), which in previous studies showed markedly higher activity *in vitro* (Table S2) and *in vivo* (*34*) compared to the alternate nitrite reductase enzyme encoded by *nirS*. Further, our regression coefficients showed larger contributions of *narG* versus *napA* to yield on nitrate (Fig. 2H), and similarly *cnor* versus *qnor* to yield on nitrite (Fig. 2J). Both these observations are consistent with the fact that the genes encoded by *narG* and *cnor* contribute more to the proton motive force (and therefore to ATP generation) than their alternatives, *napA* and *qnor*, respectively (*35*). Finally, the transporter encoded by the gene *narK1K2* is a fusion of the nitrate/H^+ symporter *NarK1* and the nitrate/nitrite antiporter *NarK2*, the latter of which is crucial for exchanging nitrate and nitrite between the cytoplasm and periplasm during denitrification when the membrane-bound nitrate reductase is utilized. In *Paracoccus denitrificans*, this fusion has been shown to have substantially higher affinity for nitrate than *NarK2* alone, resulting in higher growth rates under denitrifying conditions (*36*). Remarkably this agrees with what we found in the nitrate and nitrite reduction rate regressions, where we observed large positive contributions of *narK1K2* (Fig. 2G and I). Taken together,

these observations suggest that the regressions exploited mechanistic aspects of the denitrification process to predict metabolite dynamics. However, for many coefficients in our regression, notably regulators, there is no clear interpretation, and definitive proof that these coefficients are mechanistically informative will require genetic manipulation of diverse bacteria.

We conclude that sparse statistical models quantitatively map the genotype of each strain in our library to its associated metabolite dynamics via a consumer-resource model. Since the consumer-resource model quantified the metabolites and growth dynamically over a range of initial conditions (Fig. S4), this mapping makes no steady state assumptions and works across a range of environments. Moreover, since our approach utilized genomic variation across a diverse library of natural isolates rather than only model organisms, and since the regressions generalized well out-of-sample, we expect the models can predict metabolite dynamics for new isolates using only genome sequence data.

Predicting metabolite dynamics in communities

We next asked whether knowledge of the consumer-resource phenotypic parameters for individual strains permitted quantitative predictions of metabolite dynamics in communities. Since the phenotypic parameters are sparsely encoded by the genomes of each strain (Fig. 2), predicting community metabolite dynamics from the consumer-resource model would provide a direct mapping from gene content to community metabolism. To address this question, we extended to our modeling formalism to N-strain communities by adding the rate contributions of each strain to the dynamics of nitrate and nitrite (Fig. 3B, Supplementary Text). This model assumes that strains interact only via cross-feeding and resource-competition for electron acceptors. This "additive" model also assumes that the rates and yields on nitrate and nitrite for strains in pair culture are the same as in monoculture. As a result, the model provides predictions for N-strain community metabolite dynamics without any additional free parameters.

We first tested the ability of this approach to predict metabolite dynamics in all pair combinations of 12 diverse strains from our library (4 Nar/Nir, 4 Nar, 4 Nir). We assembled communities in 96-well plates containing chemically-defined medium and sampled over a $64\,\mathrm{h}$ period to measure concentrations of nitrate and nitrite (Materials and Methods). Remarkably, we found that the additive model accurately predicted the metabolic dynamics for most 2-strain communities (Fig. 3, Fig. S15 and S16). Specifically, the third column of Fig. 3A shows the zero-free-parameter predictions (lines) of denitrification dynamics in 2-strain communities, which agreed well with measurements (points). The 2-strain community predictions include non-trivial dynamics such as two Nar strains exhibiting faster nitrate reduction as a collective or an increase in the transient levels of nitrite in a Nar/Nir + Nar community.

We quantified the quality of the additive model predictions by computing a normalized root-mean-square error (NRMSE, see caption of Fig. 3). NRMSE in the range 0–2 indicates predictions in 2-strain communities that are similar in quality to the fits of the constituent monocultures. We found that most 2-strain communities have low NRMSE, indicating that our model successfully predicted metabolite dynamics in most cases, given only knowledge of the mono-

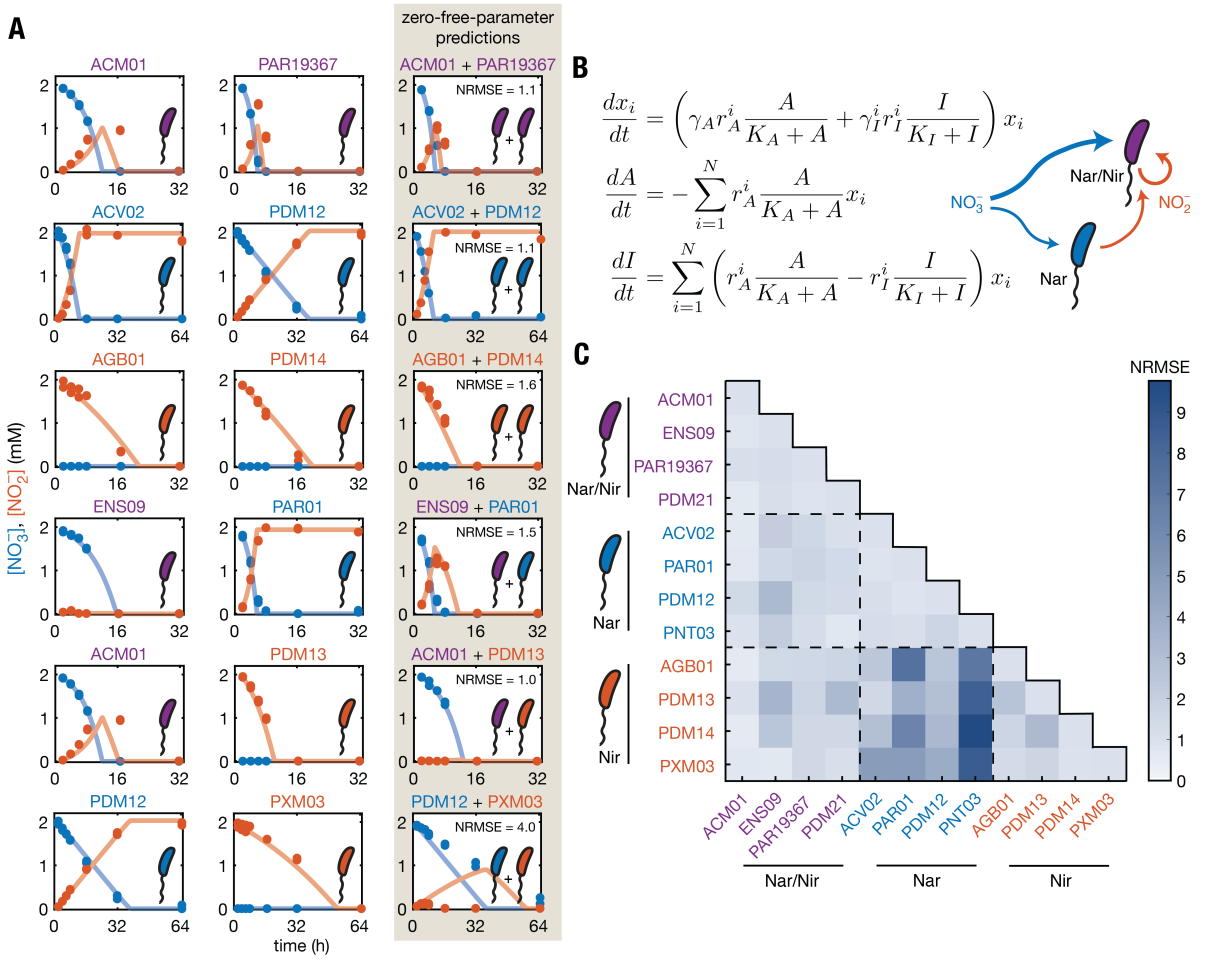

Figure 3: 2**-strain community metabolite dynamics are predictable from single-strain information.** (**A**) Examples of pair culture dynamics from each combination of the three denitrification phenotypes (Nar/Nir, purple; Nar, blue; Nir; red). The first two columns of panels show metabolite dynamics for each of two strains cultured individually, and the third column shows the metabolite dynamics of the two strains in co-culture. Curves show the predictions of an additive null model (**B**) that assumes interaction only via cross-feeding and resource competition. (**C**) Normalized root-mean-square error (NRMSE) values quantifying the quality of model predictions for all pairs of 12 strains. NRMSE was computed as $NRMSE_{ij} = RMSE_{ij}/\sqrt{(RMSE_i^2 + RMSE_j^2)/2}$, where RMSE_{ij} is the root-mean-square error between model predictions and observed metabolite concentrations of strains i and j in pair culture, and RMSE_i and RMSE_j are similarly the RMSEs of strains i and j in monoculture. NRMSE values between 0 and 2 indicate fits of similar quality to the corresponding monocultures.

culture rates and yields for each strain. The success or failure of the model depended on the phenotypes of the strains present. The model successfully predicted 2-strain metabolite dynamics for most types of communities (e.g., Nar/Nir + Nar or Nar + Nar) but failed only in the case where Nar strains were cultured with Nir strains (Fig. 3A and C, Fig. S17). We speculate that the failure of the model to predict metabolite dynamics in Nar + Nir communities was caused by excretion of nitric oxide by the Nir strain, which can be cytotoxic to strains that do not express nitric oxide reductase (*37*), and may consequently slow Nar strain growth. Although both Nar/Nir and Nir strains are capable of generating extracellular nitric oxide, Nir isolates have been observed in a previous study to transiently generate nitric oxide at higher concentrations (*26*), possibly explaining why the 2-strain additive model fails only to predict Nar + Nir communities.

We next asked whether information from monocultures also successfully predicted metabolite dynamics in 3-strain communities. We applied the additive model to predicting the nitrate and nitrite dynamics in 81 random combinations of 3 strains from the 12-strain subset. In communities that did not contain a Nar + Nir pair (e.g., Fig. 4A), we found that prediction accuracy was high (grey points, Fig. 4B, Fig. S18). This again indicated that in most combinations of phenotypes, community dynamics were predictable from consumer-resource parameters for each strain in the community. However, in communities that contained a Nar + Nir pair, predictions were relatively poor (yellow points, Fig. 4B, Fig. S18), suggesting that interactions between Nar and Nir phenotypes that were not captured in the additive model were again driving low prediction accuracy.

To address the impact of interactions between Nar and Nir strains not accounted for by our additive model in 3-strain communities, we took a coarse-graining approach. We asked whether the collective metabolism of Nar + Nir pairs could be treated as modules within 3-strain communities. To accomplish this we re-fitted nitrate and nitrite reduction rates (r_A, r_I) to pair-culture data for each Nar + Nir pair, leaving yields fixed (Fig. 4C, Supplementary Text, Fig. S19). This resulted in effective nitrate and nitrite reduction rates ($\tilde{r}_A$, $\tilde{r}_I$) for each Nar + Nir pair. We then used these rates to make predictions for 3-strain communities that included a Nar + Nir pair (e.g., Fig. 4D). For 3-strain communities that included multiple Nar + Nir pairs (e.g., Nar + Nar + Nir), we developed simple rules for determining the effective rates from the rates for each Nar + Nir pair present (Supplementary Text). We found that the metabolite dynamics in 3-strain communities containing Nar + Nir pairs were quantitatively well-predicted by this coarse-graining approach (yellow points, Fig. 4B). We conclude that treating Nar + Nir pairs as effective modules within larger communities recovers the predictive power of the additive consumer-resource model.

Discussion

Since gene content enabled the prediction of phenotypic parameters that in turn predicted community metabolism, our results indicate that functional gene content in simple communities

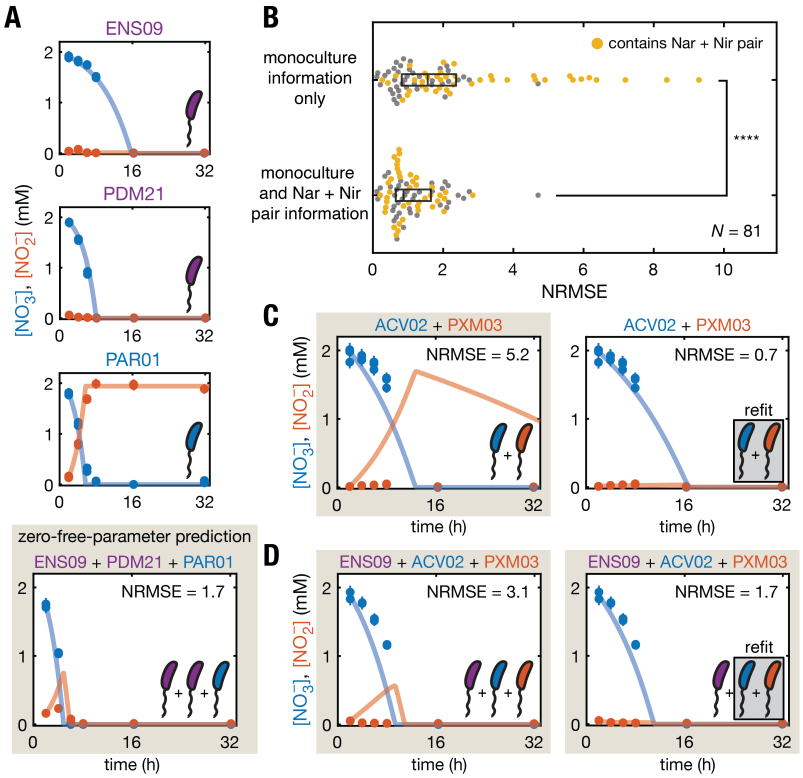

Figure 4: **Predicting metabolite dynamics in 3-strain communities.** (**A**) Metabolite dynamics for an example 3-strain (Nar/Nir + Nar/Nir + Nar) community. The first three panels show metabolite dynamics for each strain cultured individually, and the fourth panel shows the metabolite dynamics of the 3-strain community. Curves show the prediction of the additive null model (Fig. 3B). (**B**) NRMSE values quantifying quality of model predictions for 3-strain communities, comparing predictions using only parameters fit only to monocultures versus predictions that treated Nar + Nir pairs as effective modules. Mean NRMSE values were compared via t-test ($p = 7 \times 10^{-6}$). (**C**) Metabolite dynamics for an example Nar + Nir pair, where curves in the left panel show the prediction of the additive null model using only parameters fit to monocultures, and curves in the right panel show an effective refitting of the Nar and Nir strain reduction rate parameters to the pair culture data. (**D**) Metabolite dynamics for a 3-strain community containing a Nar/Nir strain and the Nar + Nir pair shown in panel **C**, where again curves in the left panel show the prediction of the additive null model for which all parameters are fit in monoculture, and the right panel shows the prediction where the Nar + Nir pair is treated as a module with rate parameters refit in pair culture.

can be interpreted in terms of metabolite dynamics at the community-level. This insight could eventually enable the prediction of metabolite dynamics in complex communities where functional gene content has been assigned to individual genomes (*38*). Soils and host-associated communities typically contain thousands of bacterial taxa, so testing the predictive power of the consumer-resource formalism in communities of many taxa in more complex environments will be essential. However, micron-scale spatial structure in soils suggests that denitrification may occur locally, in communities of just a few taxa (*39*), meaning that the rules of denitrification for simple communities could apply to natural contexts. Understanding the ecology of denitrification in complex contexts is essential for minimizing N_2O production from soils (*40*) and controlling bacterial nitric oxide production in mammalian hosts (*41*).

At the cellular level, the apparent mechanistic relevance of the regression coefficients in this study suggests that a statistical approach, coupled with large-scale culturing and phenotyping on libraries of isolates (*42, 43*), could be exploited to discover the salient features of genomes that determine other metabolic functions. Higher throughput measurements should enable a more detailed interrogation of genomic features, allowing us to extend our statistical approach to gene sequences, promoter architecture and synteny. These insights might then be used to design genomes and communities with predefined metabolic capabilities by the addition or deletion of specific genes (*44*).

The evolutionary and ecological basis for our mapping from genomic structure to community function remains to be discovered, but our results lend support to the idea that different genes in the denitrification pathway are adapted to different ecological niches (*17, 25*). Combining our understanding of how gene content determines phenotype with analyses of horizontal gene transfer and community assembly in specific niches could yield insights into how evolutionary and ecological processes combine to shape community structure and function.

References

1. M. J. Follows, S. Dutkiewicz, S. Grant, S. W. Chisholm, *Science* **315**, 1843 (2007).
2. H. Lu, K. Chandran, D. Stensel, *Water Research* **64**, 237 (2014).
3. S. Subramanian, *et al.*, *Nature* **510**, 417 (2014).
4. K. Anantharaman, *et al.*, *Nature Communications* **7**, 13219 (2016).
5. N. Molina, E. van Nimwegen, *Trends in Genetics* **25**, 243 (2009).
6. I. Sela, Y. I. Wolf, E. V. Koonin, *Physical Review X* **9**, 031018 (2019).
7. W. Shou, S. Ram, J. M. G. Vilar, *Proceedings of the National Academy of Sciences of the United States of America* **104**, 1877 (2007).
8. J. B. H. Martiny, S. E. Jones, J. T. Lennon, A. C. Martiny, *Science* **350** (2015).

9. E. E. Lilja, D. R. Johnson, *The ISME Journal* **10**, 1568 (2016).

10. D. M. Ward, *et al.*, *Philosophical Transactions of the Royal Society B: Biological Sciences* **361**, 1997 (2006).

11. O. X. Cordero, L.-A. Ventouras, E. F. DeLong, M. F. Polz, *Proceedings of the National Academy of Sciences* **109**, 20059 (2012).

12. A. Sanchez-Gorostiaga, D. Bajić, M. L. Osborne, J. F. Poyatos, A. Sanchez, *PLOS Biology* **17**, e3000550 (2019).

13. N. Klitgord, D. Segrè, *Current Opinion in Biotechnology* **22**, 541–546 (2011).

14. M. Mori, T. Hwa, O. C. Martin, A. D. Martino, E. Marinari, *PLOS Computational Biology* **12**, e1004913 (2016).

15. W. Harcombe, *et al.*, *Cell Reports* **7**, 1104 (2014).

16. C. J. Norsigian, X. Fang, Y. Seif, J. M. Monk, B. O. Palsson, *Nature Protocols* **15**, 1 (2020).

17. C. M. Jones, S. Hallin, *The ISME Journal* **4**, 633–641 (2010).

18. N. Fierer, *et al.*, *The ISME Journal* **6**, 1007 (2012).

19. S. Louca, L. W. Parfrey, M. Doebeli, *Science* **353**, 1272 (2016).

20. J. E. Goldford, *et al.*, *Science* **361**, 469 (2018).

21. M. S. Datta, E. Sliwerska, J. Gore, M. F. Polz, O. X. Cordero, *Nature Communications* **7** (2016).

22. W. G. Zumft, *Microbiology and Molecular Biology Reviews* **61**, 84 (1997).

23. S. Seitzinger, *et al.*, *Ecological Applications* **16**, 2064–2090 (2006).

24. T. Irrazabal, A. Belcheva, S. E. Girardin, A. Martin, D. J. Philpott, *Molecular Cell* **54**, 309 (2014).

25. D. R. H. Graf, C. M. Jones, S. Hallin, *PLoS ONE* **9**, e114118 (2014).

26. P. Lycus, *et al.*, *The ISME Journal* **11**, 2219 (2017).

27. K. Heylen, *et al.*, *Environmental Microbiology* **8**, 2012 (2006).

28. C. M. Jones, B. Stres, M. Rosenquist, S. Hallin, *Molecular Biology and Evolution* **25**, 1955 (2008).

29. D. A. Rodionov, I. L. Dubchak, A. P. Arkin, E. J. Alm, M. S. Gelfand, *PLoS Computational Biology* **1**, 17 (2005).

30. J. W. B. Moir, N. J. Wood, *Cellular and Molecular Life Sciences* **58**, 215–224 (2001).

31. M. Lynch, *Annual Review of Microbiology* **60**, 327 (2006).

32. B. R. K. Roller, S. F. Stoddard, T. M. Schmidt, *Nature Microbiology* **1**, 16160 (2016).

33. J. Li, *et al.*, *The ISME Journal* **13**, 2162 (2019).

34. A. B. Gloekner, A. Jiingst, W. G. Zumft, *Archives of Microbiology* **160**, 18–26 (1993).

35. S. J. Ferguson, D. J. Richardson, *The Enzymes and Bioenergetics of Bacterial Nitrate, Nitrite, Nitric Oxide and Nitrous Oxide Respiration* (Springer, 2004), vol. 2, p. 169–206.

36. A. D. Goddard, J. W. B. Moir, D. J. Richardson, S. J. Ferguson, *Molecular Microbiology* **70**, 667–681 (2008).

37. C. Braun, W. G. Zumft, *The Journal of Biological Chemistry* **266**, 22785–22788 (1991).

38. C. M. K. Sieber, *et al.*, *Nature Microbiology* **3**, 836 (2018).

39. R. Lensi, A. Clays-Josserand, L. Jocteur Monrozier, *Soil Biology and Biochemistry* **27**, 61 (1995).

40. M. A. Cavigelli, G. P. Robertson, *Ecology* **81**, 1402 (2000).

41. E. R. Hyde, *et al.*, *PLOS ONE* **9**, e88645 (2014).

42. S. A. Connon, S. J. Giovannoni, *Applied and Environmental Microbiology* **68**, 3878 (2002).

43. J. Kehe, *et al.*, *Proceedings of the National Academy of Sciences* **116**, 12804 (2019).

44. A. J. Shaw, *et al.*, *Proceedings of the National Academy of Sciences* **105**, 13769 (2008).

Acknowledgments

We thank Laura Troyer for assistance with isolating bacterial strains, and Elizabeth Ujhelyi and Annette Wells for assistance with sequencing. We acknowledge Cameron Pittelkow for access to corn and soybean fields in Savoy, Illinois, and the laboratory of Julie Zilles for providing the bacterial strain *Paracoccus denitrificans* ATCC 19367. We also thank James Sethna, William Metcalf, Jun Song and members of the Kuehn laboratory and Mani group for helpful discussions. **Funding:** This work was supported by the National Science Foundation Physics

Frontiers Center Program PHY 0822613 and PHY 1430124 (S.K.), James S. McDonnell Foundation Postdoctoral Fellowship Award #220020499 (K.G.), and the Simons Foundation Investigator Award #597491 (M.M.). **Authors contributions:** K.G.: Conceptualization, experimental design, data collection, formal analysis, coding, writing - original draft. D.P.: data collection. M.M: Conceptualization, formal analysis, supervision, writing - revision & editing. S.K.: Conceptualization, experimental design, formal analysis, supervision, writing - original draft. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Draft genome assemblies are deposited on NCBI (BioProject PRJNA660495). Annotation files and data used in the regressions are deposited on Open Science Framework (doi: 10.17605/OSF.IO/T3PRD). Isolates are available upon request.

Supplementary materials

Materials and Methods
Supplementary Text
Figs. S1 to S19
Tables S1 to S9
References *(45-85)*