HHS Public Access

Author manuscript

*Nat Microbiol*. Author manuscript; available in PMC 2017 March 12.

Published in final edited form as:

*Nat Microbiol*. ; 1: 16160. doi:10.1038/nmicrobiol.2016.160.

# Exploiting rRNA Operon Copy Number to Investigate Bacterial Reproductive Strategies

**Benjamin R.K. Roller**[1,2,+], **Steven F. Stoddard**[1], and **Thomas M. Schmidt**[1,3,4,*]

[1] Department of Internal Medicine, University of Michigan, Ann Arbor, MI, 48109, USA

[2] Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI, 48824, USA

[3] Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, 48109, USA

[4] Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI, 48109, USA

## Summary

The potential for rapid reproduction is a hallmark of microbial life, but microbes in nature must also survive and compete when growth is constrained by resource availability. Successful reproduction requires different strategies when resources are scarce compared to when they are abundant[1,2], but a systematic framework for predicting these reproductive strategies in bacteria has not been available. Here we show that the number of ribosomal RNA operons (*rrn*) in bacterial genomes predicts two important components of reproduction – growth rate and growth efficiency – which are favored under contrasting regimes of resource availability[3,4]. We find that the maximum reproductive rate of bacteria doubles with a doubling of *rrn* copy number, while the efficiency of carbon use is inversely related to maximal growth rate and *rrn* copy number. We also identify a feasible explanation for these patterns: the rate and yield of protein synthesis mirror the overall pattern in maximum growth rate and growth efficiency. Furthermore, comparative analysis of genomes from 1,167 bacterial species reveals that *rrn* copy number predicts traits associated with resource availability, including chemotaxis and genome streamlining. Genome-wide patterns of orthologous gene content covary with *rrn* copy number, suggesting convergent evolution in response to resource availability. Our findings indicate that basic cellular processes adapt in



* To whom correspondence should be addressed. Tel: 1-(734)-763-8206; Fax: 1-(734)-615-5534; schmidti@umich.edu.
+Current address: Center for Adaptation to a Changing Environment, ETH Zürich, 8092 Zürich, Switzerland

**Author contributions** B.R.K.R. and T.M.S. conceived the study, interpreted results and wrote the paper. B.R.K.R. performed experiments, phylogenetic inferences, and statistical analyses. S.F.S. provided custom relational databases, queries, and Perl scripting to integrate the *rrn*DB, KEGG, and NCBI taxonomy for genomic analyses.

Accession numbers

Genbank assembly accession numbers for genomes used in this analysis were downloaded from KEGG and are listed in Supplementary Table 4.

Code availability

Computer code for statistical analyses are available on request.

Competing financial interests

The authors declare no competing financial interests.

contrasting ways to long-term differences in resource availability. They also establish a basis for predicting changes in bacterial community composition in response to resource perturbations using *rrn* copy number measurements[5] or inferences[6,7].

---

Resource availability selects for different aspects of bacterial reproduction. High resource concentrations impose strong selective pressures for rapid reproduction[3,4], so bacteria which are favored during resource abundance – copiotrophs – typically grow quickly[1,8]. Rapid bacterial growth requires a substantial increase in cellular ribosome content relative to slow growth. Copiotrophs achieve a high ribosome content, in part, by maintaining multiple *rrn* copies in their genomes[8-10], particularly near the origin of replication which effectively amplifies rRNA gene dosage during rapid growth[11].

Resource scarcity, especially in spatially structured environments, selects for efficient resource utilization in theory and in the laboratory[3,4]. However, it is unclear if bacteria which are favored by low-resource conditions in nature – oligotrophs – share adaptations that confer efficient growth. It has been suggested, but not yet demonstrated, that oligotrophic bacteria are generally more efficient[12], i.e. produce more offspring per unit resource consumed than copiotrophic bacteria. Oligotrophic bacteria do tend to encode fewer *rrn* copies in their genomes than copiotrophs[1,13], leading us to hypothesize that *rrn* copy number predicts two components of fitness related to resource availability: the rate and efficiency of growth. However, closely related bacteria tend to have similar traits[14], including similar numbers of *rrn* copies[6]. Therefore, we also address the alternative hypothesis that evolutionary history explains relationships between growth characteristics and the number of *rrn* copies, rather than adaptation to resource availability.

First, we re-examined the previously observed rank-order correlation between maximum recorded growth rate and *rrn* copy number[11]. We used phylogenetic regression and model selection criteria to quantitatively assess this relationship, while also including new species. The best-fit model to emerge was a log-log correlation; the doubling of maximal recorded growth rate corresponded to a doubling of *rrn* copy number (Fig. 1a) and this relationship could not be attributed to evolutionary history (Table 1, Supplementary Table 1). Therefore, a $\log_2$ transformation of *rrn* copy number, hereafter referred to as $\log_2$-*rrn*, was examined in subsequent analyses. The deletion of *rrn* copies has been causally linked to a loss of rapid growth in previous experiments. For instance, in *Escherichia coli*, the deletion of one or two *rrn* copies diminishes growth rates of cultures in rich medium although cells are able to compensate for those deletions in less optimal media[9]. Deletion of *rrn* copies also extends the time required for cultures of *E. coli* to adapt to nutrient and temperature change[15]. Taken together, the log-log correlation and previous experiments suggest that selection for rapid growth across the bacterial domain was accompanied by an expansion of *rrn* copy number, which directly contributes to rapid growth phenotypes.

Next, we measured carbon use efficiency (CUE) for eight phylogenetically diverse bacteria grown in a mineral medium containing succinate as the sole carbon and energy source. CUE was negatively correlated with the maximum growth rate reported for these eight species, and this relationship was also not driven by shared evolutionary history (Fig. 1b). The highest growth efficiencies were from a model ocean oligotrophic bacterium, *Sphingopyxis*

*alaskensis* RB2256[1], and the soil oligotrophs *Rhodospirillaceae* sp. PX3.14 and *Acidobacteriaceae* sp. TAA166, which were isolated using nutrient-poor media and long incubation times. The lowest growth efficiency was observed for a strain of the copiotrophic bacterium *Escherichia coli.* Overflow metabolism is not a likely cause of the observed CUE variation because unlike the bacteria investigated, known succinate fermenters are strict anaerobes. These findings are consistent with theory that suggests a rate-efficiency tradeoff is inevitable for heterotrophic metabolism[3].

Protein synthesis is an attractive target for explaining variation in the growth efficiency of bacteria because protein is an abundant macromolecule in bacterial cells, it is expensive to make (50-60% of ATP is used for polymerizing amino acids into protein during balanced growth[16]), and a high ribosome content is necessary for shifting to rapid growth[9,17]. We measured protein yield (protein produced per $O_2$ consumed) for the same eight species as above and found that it is negatively correlated with $\log_2$-*rrn* copy number even after controlling for evolutionary history (Fig. 2a, Table 1, Supplementary Table 1). This is in contrast to the positive relationship between translational power – the translation rate per ribosome – and $\log_2$-*rrn* copy number that was measured in many of the same bacterial species[10] (Fig. 2b, Table 1, Supplementary Table 1). These observations are consistent with a model in which the speed and processivity of ribosomes is intimately linked with variation in the rate and efficiency of bacterial growth[10,18]. Protein yield measurements can also be converted into CUE units (details in methods) and these $^3$H-CUE estimates validate the $^{14}$C-CUE findings; CUE is negatively correlated with maximum growth rate and $\log_2$-*rrn* (Table 1, Supplementary Table 1). To our knowledge, this is the first phylogenetically robust evidence that efficient growth is an adaptation of bacteria with few *rrn* operons.

If *rrn* copy number truly predicts adaptation to resource availability, then traits associated with copiotrophy and oligotrophy should also be related to *rrn* copy number. For example, streamlining theory predicts that selection for efficient resource use influences the genomes of oligotrophic bacteria[12], so it should drive gene loss and smaller genome size in bacteria with few *rrn* copies. We measured and inferred several resource-associated traits for genomes representing 1,167 unique bacterial species. Consistent with our underlying hypothesis, genome size is positively correlated with $\log_2$-*rrn* (Table 1, Supplementary Table 2, & Supplementary Fig. 1). This pattern is unlikely to be caused by drift-induced genome degradation in low *rrn* bacteria because we excluded genomes of obligately symbiotic bacteria where this process is well documented[12].

The composition of the thiamine biosynthesis pathway provides independent evidence in support of genome streamlining in bacteria with few *rrn* copies. Thiamine and its molecular precursors are secreted by microbes in many environments[19,20], allowing auxotrophs, which save on thiamine biosynthetic costs, to become dominant community members[12,19]. A higher number of *de novo* thiamine biosynthesis steps were found with increasing *rrn* copy number, and the steps missing from low *rrn* genomes are biased towards the genes that do not participate in recycling exogenous thiamine or its metabolic precursors (Table 1, Supplementary Fig. 2 & Supplementary Table 2). Presumably, this allows bacteria to save energy in low nutrient environments by exploiting exogenous reserves of this important metabolic cofactor, as has been observed for representative cultivars of the ubiquitous

SAR11 clade of marine oligotrophs[19]. Selection for trimming of the thiamine biosynthesis pathway is consistent with the major postulate from the Black Queen Hypothesis, namely that "...the loss of a costly, leaky function is selectively favored at the individual level and will proceed until the production of public goods is just sufficient to support the equilibrium community."[21].

Previous studies suggest that photoautotrophy is an adaptation to oligotrophic environments[22]. Extending this idea, we found that the probability of a bacterium being autotrophic, *i.e.*, encoding any of four carbon-fixing pathways inferred from genomes, was negatively correlated with $\log_2$-*rrn*. However, statistical support for the relationship is diminished when including phylogeny in the regression model (Table 1, Supplementary Table 2, & Supplementary Fig. 3), likely because there are too few autotrophs outside of the cyanobacteria to maintain the pattern after controlling for phylogeny.

A postulated adaptation for access to high resource concentrations is chemotactic motility, an energetically expensive foraging strategy that bacteria use to track nutrient gradients[23]. As predicted, the probability of encoding chemotaxis is positively correlated with $\log_2$-*rrn* copy number and this is not attributable to evolutionary history (Table 1, Supplementary Table 2, & Supplementary Fig. 4). Copiotrophs also tend to use specialized transporters, such as the phosphoenolpyruvate:carbohydrate phosphotransferase systems (PTS)[1], and we found that PTS transporter richness is positively related to $\log_2$-*rrn* copy number. However, possessing many distinct PTS transporters is phylogenetically restricted to a subset of Firmicutes and Gammaproteobacteria. Incorporating phylogeny into the PTS regression diminishes the magnitude of the correlation and we cannot exclude the alternative hypothesis of shared evolutionary history driving this previously reported pattern (Table 1, Supplementary Table 2, & Supplementary Fig. 5).

In addition to addressing specific hypotheses, we also explored the same bacterial genomes to determine if bacteria with similar *rrn* copy number shared an increased number of orthologs and functional pathways. Using phylogenetic principal components analysis (pPCA), we evaluated genome content at two levels of annotation in the KEGG database[24]: orthologs (n = 7,119) and modules (n = 418) – combinations of orthologous genes encoding for functionally interactive proteins in either a pathway or enzyme complex. The analyses revealed a robust relationship between *rrn* copy number and genome content, regardless of annotation level assessed (Fig. 3a and 3b). The first six pPCA axes in both the module and ortholog analyses represent approximately 35% and 28% of total variation in genome content similarity, respectively. All six pPCA axes from both analyses were correlated with $\log_2$-*rrn* using phylogenetic regression (Fig. 3c and 3d).

Bacteria with similar *rrn* copy number tend to share overall similarity in their profile of genes and pathways, suggesting a signal of convergent evolution in genome content among disparately related bacteria. Interestingly, the results do not identify a prototypical genome content for either copiotrophic or oligotrophic bacteria. Perhaps this is not surprising given that: 1) the analysis included metabolically diverse bacteria from environments exerting multiple selective pressures beyond resource availability, 2) individual pPCA axes explain small amounts of genome content variation (<10%), and 3) the influence of genome content

that is unique to a phylogenetic group was minimized by controlling for evolutionary history. Some biological functions had relatively strong loadings on axes correlated with $\log_2$-*rrn* which were consistent for analyses at both annotation levels, such as the type III secretion system (Supplementary Table 3). However, caution is warranted in interpreting these trends and future work is required to establish a link between these traits and reproductive strategies.

The emerging view from this and other research suggests that copiotrophic bacteria are adept at sensing and responding to the ephemeral patchiness of the external environment, while oligotrophic bacteria are less likely to respond to changes in the external environment and instead rely on efficient resource utilization to compete in the dilute matrix that characterizes many environments on Earth[1,12]. These interpretations are consistent with studies of microbial community dynamics tracking resource perturbations and *rrn* copy number in nature. For instance, in response to resource inputs from the Deepwater Horizon oil spill, there was a bloom of hydrocarbon-degrading bacteria classified as members of the genus *Colwellia*[25] – cultivated representatives of this genus have 9 *rrn* copies[5]. In terrestrial environments, early successional bacteria encode more *rrn* copies than late successional bacteria[26,27], and bacteria that responded most quickly to the addition of 2,4-D (an herbicide that is metabolized by bacteria) had more *rrn* copies than those that responded slowly[8]. In host-associated microbial communities, a bloom of high *rrn* copy number *Enterobacteriaceae* during antibiotic-associated diarrhea[28] was coincident with the temporary increase in carbohydrates entering the colon.

Determining the number of *rrn* copies in bacterial genomes typically requires the characterization of genomic DNA from pure cultures or gapless assemblies, however computational techniques are now available for inferring *rrn* copy number from 16S rRNA gene sequences[6,7]. Inference of *rrn* copy number offers the potential to generate testable predictions for complex microbial communities. For example, variation in carbon use efficiency (CUE) has strong implications for organic matter decomposition, but is assigned a fixed value for all microbial communities in prominent carbon cycling models[29]. Global soil carbon model projections will likely be improved when CUE can vary, rather than being fixed as a constant[29,30], because all other things being equal, communities dominated by low-*rrn* bacteria should have a higher CUE than high-*rrn* dominated communities.

Guided by ecological and evolutionary theory, we examined variation in bacterial reproduction and found that *rrn* copy number is a reliable and generalizable proxy for bacterial adaptation to resource availability. Strategies for rapid or efficient reproduction have genomic signatures, but they are also emergent properties of cellular metabolism that are associated with protein synthesis phenotypes. These insights can help generate a more holistic understanding of bacterial fitness in natural habitats, which will be essential if we are to ever manage microbiomes for human and environmental health.

# Methods

## Strains and growth conditions

The following bacterial strains were used in growth efficiency experiments: *Vibrio natriegens* ATCC 14048[31], *Bacillus subtilis* Marburg ATCC 6051[32], *Escherichia coli* K12 MG1655[33], *Pseudomonas* sp. HF3[10,34], *Arthrobacter* sp. EC5[8,10], *Rhodospirillaceae* sp. PX3.14[8,10], *Sphingopyxis alaskensis* RB2256[10,35,36], and *Acidobacteriaceae* sp. TAA166[37].

The medium used for all experiments contained basal salts supplemented with trace elements and vitamins[37] but modified as follows (final concentrations reported): 2mM $Na_2SO_4$, 2μg/L $CuCl_2 \cdot 2H_2O$, 150μg/L pyridoxine hydrochloride, 10mM 3-(N-morpholino)propanesulfonic acid (MOPS) buffer, and 10mM sodium succinate as the sole carbon and energy source. *Vibrio natriegens* ATCC 14048 cultures were supplemented with 20g/L NaCl.

All growth experiments were performed with 30ml cultures in 300ml or 500ml sidearm flasks incubated at 25°C and shaken at either 100 rpm (*Acidobacteriaceae* sp. TAA166) or 200 rpm (all other strains). Optical density (OD) was measured over time on a Spec20D+ spectrophotometer at wavelengths of either 600nm (*Acidobacteriaceae* sp. TAA166) or 420nm (all other strains) to maximize measurement sensitivity at higher and lower OD values, respectively. Cells were recovered from freezer stock in batch culture and allowed to reach unconstrained and balanced growth by multiple transfers to fresh medium during exponential growth. Biological replicates were independently inoculated into separate recovery flasks from the same freezer stock and remained independent throughout all measurements. Biological replicates were often performed across different days and medium batches for protein yield, but not $^{14}C$ carbon use efficiency due to impracticalities in that experimental design. Unconstrained and balanced growth was empirically determined for each strain when the growth rate no longer improved upon transfer to fresh medium, typically after a dilution of at least 1,000 fold from exponentially growing recovery culture. Bacteria included in all analyses in this publication are listed in Supplementary Table 4.

## Maximum recorded growth rate determination

The maximum recorded growth rates for a diverse collection of 176 bacteria with known *rrn* copy number[11], and for the 8 strains measured for efficiency in this study, were gathered from the literature[10,11,31]. The growth of strain TAA166 in this study exceeded the maximum recorded growth rate from the literature[38], so the value from this study was utilized.

## Protein yield and carbon use efficiency measurements

All protein yield and carbon use efficiency measurements were obtained at least two doublings prior to departing from unconstrained and balanced growth. Three biological replicate cultures were used for protein yield measurements of each strain. Two biological replicate cultures were used for carbon use efficiency measurements of each strain.

Protein yield was measured using $^{3}$H-leucine incorporation and oxygen ($O_2$) consumption. Leucine was used because its is one of the least variable amino acids in protein on a mol% basis[39]. The amount of radiolabeled leucine added was optimized on the most rapidly growing strains, resulting the addition of 250nCi $^{3}$H-leucine (specific activity 0.5Ci/mol) to 30ml cultures of all strains. Incorporation of $^{3}$H-leucine was measured by removing three, 1ml aliquots (technical replicate level 1) from cultures at multiple time points to ensure that leucine was not depleted over the course of the experimental measurement. $^{3}$H-leucine incorporation was converted into protein production and biomass carbon (C) units with widely used conversion factors from a marine microbial community[39] – mol% leucine in protein (7.3), intracellular isotope dilution (1.71), and protein:C dry weight ratio (0.86). Technical replicate measurements were averaged for each biological replicate culture.

$O_2$ consumption was determined by measuring the specific $O_2$ consumption rate (nmol $O_2$ $OD^{-1}$ $min^{-1}$) in sub-samples of cultures during unconstrained and balanced growth with the Unisense microrespiration system. The Clark-type oxygen microsensor was calibrated with two sterile medium solutions: anoxic media and media equilibrated to ambient oxygen concentrations. The anoxic medium was generated by incubating sterile media in a calibration chamber linked via a diffusion membrane to an antioxidant solution of 0.1M sodium ascorbate and 0.1M NaOH. Oxic media was equilibrated to ambient oxygen concentrations by stirring sterile medium with a stir bar while exposed to the laboratory atmosphere. Two 500µl aliquots (technical replicate level 1) of each growing culture were removed during each sub-sampling event (technical replicate level 2) from the parent culture (biological replicate), placed in 500µl microrespiration chambers with miniature magnetic stir bars on top of magnetic stir rack (100 RPM), and submerged in a 25°C circulating water bath. $O_2$ concentrations were recorded once per second for six to ten minutes per sample using SensorTrace Basic software (Unisense). The specific $O_2$ consumption rate was estimated by dividing the measured $O_2$ consumption rate by the mean OD over a 6-10 minute measurement window. This mean OD was calculated from the growth equation of the parent culture. Mean specific $O_2$ consumption rate was calculated across at least three separate sub-sampling events (technical replicate level 2). The mean specific $O_2$ consumption rate was then multiplied by the integral of the growth equation for the culture (biological replicate) during $^{3}$H-leucine incorporation to determine the total amount of $O_2$ consumed in the parent culture. The total $O_2$ consumed and protein produced in the parent culture, calculated from $^{3}$H-leucine experiments, was used to quantify protein yield. $O_2$ consumption values were also converted into carbon dioxide ($CO_2$) production by using a respiratory quotient of 8/7. This calculation assumes complete oxidation of the carbon source and that all oxygen consumption is due to respiration. Converting $O_2$ consumption to $CO_2$ production allows for an indirect measurement of carbon use efficiency when coupled with C-transformed $^{3}$H-leucine data.

Carbon use efficiency (CUE) was measured directly by tracking the fate of $^{14}$C-succinate into $CO_2$ and biomass. $^{14}$C-succinate (0.25µCi) was added to two 30ml cultures (biological replicates) growing in 500ml flasks where unlabeled succinate was present in excess of biosynthetic demand. Cultures were sealed after radiochemical addition to trap $^{14}CO_2$ in the culture flask headspace. Sealed cultures were incubated for 1 generation or less after $^{14}$C addition, then terminated by addition of ice-cold trichloroacetic acid (TCA, 5% final

concentration) while the culture vessel was surrounded by ice. TCA also acidified the medium, releasing dissolved $CO_2$ as a gas into the headspace. Cultures were then flushed with $N_2$ for 2 hours while kept at 4°C, with the outflow gas entering 3 stoppered gas traps (1:1, phenethylamine:methanol) connected in series to the culture vessel with tygon tubing and needles. $^{14}CO_2$ was retained in the liquid phase of gas traps and was quantified by transferring three, 1 ml aliquots of each gas traps contents (technical replicates) to scintillation cocktail (Biosafe-II) and scintillation counting (Beckman Coulter).

All radiolabeled biomass from protein yield and carbon use efficiency experiments was precipitated using ice-cold TCA (5% final concentration) and replicate 1 ml aliquots (technical replicates) centrifuged at 11,000g for 10 minutes. Cell pellets were sequentially washed/centrifuged with 1ml 5% TCA, 1ml 80% ethanol, and finally resolubilized using 0.2ml NaOH (1M) at 90°C for one hour. Resuspended pellets were transferred into scintillation cocktail (Biosafe-II) and the radiolabel was quantified in a scintillation counter (Beckman Coulter) after 1 week to allow chemiluminescence from NaOH to diminish. Technical replicates were averaged (mean) to quantify radiolabel in biomass of biological replicate cultures. Model II regression (major axis estimation) of the two CUE methods among all strains indicated that measured CUE values from the direct ($^{14}C$) and indirect ($^{3}H$-leucine and $O_2$ converted to C units) methods were highly correlated. The slope of this regression was indistinguishable from 1 ($R^2 = 0.73$; slope = 1.44, 95% CI = 0.8-3.1), indicating the methods are approximately equivalent.

## Comparative genomics and phylogeny construction

The April 28, 2014 version (Release 70.0) of the Kyoto encyclopedia of genes and genomes (KEGG) database[24,40] was downloaded to construct the dataset of sequenced bacterial genomes. Obligate symbiotic, commensal, and parasitic bacteria were excluded from the dataset if these genomes were described in the literature to possess signatures of genome degradation via genetic drift[12], *e.g.*, high pseudogene counts, elevated rates of non-synonymous substitutions, or expansion of noncoding genetic elements. Bacteria were also excluded if their rRNA (*rrn)* operon copy number could not be accurately estimated from KEGG data (*rrn*DB version 4.3.3)[5]. All genomes were then subjected to manual curation and a single representative genome was chosen for each unique bacterial species. Representative genomes were selected using the following hierarchical criteria: 1) genome of Type strain of species available, 2) the central tendency of *rrn* operon copy number distribution for the species is accurately reflected by the genome, and 3) greatest number of annotated orthologous genes in KEGG orthology system are present in the genome. This resulted in 1,167 genomes that passed all selection criteria. The final dataset consisted of the *rrn* operon copy number, as well as the presence/absence of every asserted ortholog (K0) and module (M0), for 1,167 genomes extracted from the KEGG database.

A bacterial genome was scored as possessing chemotactic motility if the ortholog for the genes cheA (K03407), cheY (K03413), fliM (K02416), and fliN (K02417) were all present. These are orthologs for the following proteins: chemotaxis histidine kinase (CheA), chemotaxis response regulator which binds the flagellar motor (CheY), and flagellar motor switch proteins that are bound by CheY (FliM & FliN). Chemotactic systems are not

genetically uniform among model bacteria, so this definition was used to ensure that the genomes of four phylogenetically diverse model chemotactic bacteria, *Rhodobacter sphaeroides*, *Escherichia coli*, *Bacillus subtilis*, and *Rhizobium leguminosarum* bv. *viciae*, were scored as possessing chemotaxis[41,42].

A bacterial genome was scored as being autotrophic by a combination of manual curation and genome annotation. The first step used KEGG annotation to identify genomes that possessed at least one complete KEGG module for one of four autotrophic pathways: the Calvin cycle (M00165), the reductive TCA cycle (M00173), the 3-Hydroxypropionate cycle (M00376), or the Wood-Ljungdahl pathway (M00377). Microbes possessing the Wood-Ljungdahl pathway were then hand-curated and genomes that were not explicitly described as capable of fixing carbon dioxide as their sole source of biosynthetic carbon were excluded. Additionally, organisms possessing a sub-module of the Calvin cycle (M00166 or M00167) or that were from genera that were likely autotrophs were manually curated and those that were explicitly described as capable of fixing carbon dioxide as their sole source of biosynthetic carbon were added to the list of autotrophic genomes. Oxygenic photoautotrophs were subject to the same requirements as chemoautotrophs and in addition must possess photosystem I and II (M00163 and M00161, respectively).

KEGG has annotated a total of 25 distinct PTS transporter types as separate modules. For each genome these modules were summed to estimate PTS transporter richness.

The number of orthologs present in the *de novo* thiamine biosynthesis pathway was assessed for all 1,167 genomes. A total of 10 biosynthetic steps were considered from the canonical bacterial *de novo* biosynthesis pathway[43], with 12 orthologs catalyzing these steps annotated in the dataset. The ortholog of tenI (K10810) was annotated in only 41 of 1,167 genomes in the dataset. Therefore, the maximum possible number of *de novo* thiamine synthesis orthologs encoded in a genome was considered to be 11. These 11 orthologs were split into two categories, those involved in recycling thiamine or those uninvolved in recycling. Thiamine pyrophosphate is the bioactive co-factor molecule, and salvage of exogenous thiamine pyrophosphate metabolic precursors is possible. Orthologs which catalyze reactions downstream of salvage input steps were considered to be involved in recycling. Orthologs which catalyze reactions upstream of salvage input steps were considered to be uninvolved in recycling[43]. The 8 orthologs uninvolved in recycling include: dxs (K01662); thiO (K03153) or thiH (K03150); thiF (K03148); thiS (K03154); thiI (K03151); iscS (K04487); thiG (K03149); thiC (K03147). The 3 orthologs involved in recycling include: thiD (K00941) or thiDE (K14153); thiE (K00788) or thiDE (K14153); thiL (K00946).

Aligned 16S rRNA-encoding gene sequences for all bacteria in this study were downloaded from the SILVA ribosomal RNA database project[44]. If an aligned sequence from the utilized genome was not available in SILVA for any of the 1,167 bacterial genomes in the phylogenetic regression analysis, an aligned sequence from a separate sequencing effort on the same strain or from the type strain of that species was utilized. Three base phylogenetic tree sets were built using the ARB software package[45] using maximum likelihood estimation (RAxML) to generate the ten most likely trees for each set. One tree set was built for the 1,167 genome dataset, a second tree set was built for the 184 bacteria on which growth rate

and efficiency was recorded, and a third tree set was built for the analysis of previously published translational power data[10]. Of the ten most likely trees in each tree set, likelihood values and a topology consistent with previous studies were used to choose a single tree for all downstream comparative phylogenetic analyses. For all tree sets, only positions conserved in 50% of species in SILVA's Living Tree Project[46] were utilized to infer the phylogenetic tree. In the 1,167 genome tree set, this approach led to 110 distinct instances where pairs of species have zero unique branch length because all sequence positions utilized for tree-building were identical. Downstream phylogenetic comparative methods require each tip in the tree to have a branch length, so an arbitrary and small branch length was used to discriminate between species with identical sequences. This branch length was an order of magnitude smaller than the smallest branch length observed in the tree. Five archael sequences were used to root all trees, and were subsequently pruned from the tree along with any sequence that was not part of a given analysis.

## Statistical analyses

The R statistical programming language was used for all statistical analyses[47]. The following R packages were used for the indicated analysis technique: base R was used for linear regression analyses; lmodel2 was used for model II (major axis) regression; ape was used to import NEXUS formatted tree files and store tree objects; phylolm[48] and MCMCglmm[49] were used for phylogenetic regression; phytools[50] was used for phylogenetic principal coordinates analysis (pPCA); ggplot2[51] and scatterplot3d[52] were used for plotting figures. Phylogenetic trees were exported from ARB[53] in Newick format, imported into the Mesquite software environment[54] for converting to NEXUS format for use in R. In general, residual distributions were visually assessed for all regression analyses and did not appear to violate assumptions of the implemented tests. Specialized methods were implemented when statistical assumptions were violated (*e.g.* phylogenetic non-independence of comparative data) and they are discussed in the following sections. A phylogenetic Model II (ranged major axis) regression approach was also used in the $^{14}$CUE versus growth rate analysis to account for residual error in both dependent and independent variables. The model II regression output did not differ from standard phylogenetic regression, so was not reported for simplicity. Corrected Akaike information criterion (AICc) was used to determine the relative probability of regression models that differed only in the transformation of the predictor variable *rrn* copy number.

Phylogenetic regression was performed to account for the effect of shared evolutionary history among life history traits and *rrn* copy number. These statistical methods entail making decisions about how to model trait evolution. The linear phylogenetic regressions for growth rate, carbon use efficiency, translational power, and protein yield were performed using Pagel's lambda as an evolutionary model. Phylogenetic linear regression for genome size was performed using an Ornstein-Uhlenbeck random root evolutionary model. Alternative evolutionary models were explored for these analyses, but in general these models were of comparable fit, or less likely, based on AICc values. Analyses with alternative evolutionary models did not result in outcomes that challenge the reported findings or their interpretations. Phylogenetic Poisson regression of thiamine biosynthesis and PTS transporters were performed using generalized estimating equations[48], but

convergence issues were noted for some analyses so alternative methods were also utilized. This included using Bayesian approaches for phylogenetic regression (MCMCglmm) and converting the variables from counts to binary variables (presence or absence of the two thiamine sub-pathways) for a phylogenetic logistic regression analysis. Phylogenetic logistic regression of thiamine biosynthesis, chemotactic motility, autotrophy, and oxygenic photoautotrophy uses Ives & Garland's binary model of trait evolution[55].

Phylogenetic Poisson regression for thiamine biosynthesis and PTS transporter richness, was also performed in MCMCglmm. This method is comparable to using Pagel's lambda as evolutionary model[48]. A prior must be specified for the residual variance structure of the covariates and phylogenetic effects in MCMCglmm analyses. The inverse Wishart distribution was used in both cases (variance limit V=1 and belief parameter nu = 0.002). Package guidelines (MCMCglmm package course notes; http://cran.r-project.org/web/packages/MCMCglmm) recommended these parameters for a flat prior, and it has been used in similar phylogenetic comparative analyses of bacteria[56]. A total of 2,810,001 iterations were run for each thiamine model and the PTS model, with a burn-in of 10,000 and thinning of 2,800, which saves 1,000 total iterations for each analysis. This minimized auto-correlation in posterior parameter sampling.

Covariance-based phylogenetic principal components analysis (pPCA) was performed using a Brownian motion evolutionary model on two datasets: 1) 7,119 orthologs and 2) 418 modules. pPCA combines variables into new axes which maximally summarize variation in low-dimensional space while accounting for non-independence of data points due to shared evolutionary history. The coordinates for each genome on the new pPCA axes were further analyzed using phylogenetic regression against the explanatory variable *rrn* copy number. Therefore, orthologs of 5S, 16S, 23S, and tRNA genes were removed from the ortholog dataset prior to performing pPCA. These genes are part of the *rrn* operon and if they were not removed they confound downstream regression of pPCA axes against *rrn* copy number.

A random root Ornstein-Uhlenbeck evolutionary model was used in all regressions of pPCA axes against *rrn* copy number. Of the first ten pPCA axes from each analysis, nine ortholog axes and eight module axes were significantly correlated with $\log_2$-*rrn* when it was the sole explanatory variable (Supplementary Table 5). We have demonstrated that genome size is correlated with $\log_2$-*rrn* and needed to confirm that this confounding variable wasn't responsible for driving the genome similarity, e.g. if smaller genomes more similar due to more shared housekeeping genes. Controlling for $\log_2$-transformed genome size by including it as a covariate in these regressions did not diminish the significance of correlations with $\log_2$-*rrn*, however the direction of the correlation did change sign for one pPCA axis in the modules analysis (Supplementary Table 5). This indicates genome size doesn't drive the genome content similarity patterns observed. The 100 largest loadings on the top 6 pPCA axes for both module and ortholog analyses are included in Supplementary Table 3.

No statistical methods were used to predetermine sample size and no randomization method was used for sample processing.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

Support for this work was provided by the Department of Energy Office of Science Graduate Fellowship Program (DOE SCGF), made possible in part by the American Recovery and Reinvestment Act of 2009, administered by ORISE-ORAU under contract no. DE-AC05-06OR23100; the National Science Foundation's Long-Term Ecological Research Program through grant no. DEB 1027253 and the National Institutes of Health (GM0099549). We acknowledge Alex Schmidt, Clive Waldron, Arvind Venkataraman, Byron Smith, and Matt Hoostal for their helpful comments on this manuscript.

## References

1. Lauro F, et al. The genomic basis of trophic strategy in marine bacteria. Proc Natl Acad Sci USA. 2009; 106:15527–15533. [PubMed: 19805210]
2. Roller BRK, Schmidt TM. The physiology and ecological implications of efficient growth. ISME J. 2015; 9:1481–1487. [PubMed: 25575305]
3. Pfeiffer T, Schuster S, Bonhoeffer S. Cooperation and competition in the evolution of ATP-producing pathways. Science. 2001; 292:504–507. [PubMed: 11283355]
4. Bachmann H, et al. Availability of public goods shapes the evolution of competing metabolic strategies. Proc Natl Acad Sci USA. 2013; 110:14302–14307. [PubMed: 23940318]
5. Stoddard SF, Smith BJ, Hein R, Roller BRK, Schmidt TM. rrnDB: improved tools for interpreting rRNA gene abundance in Bacteria and Archaea and a new foundation for future development. Nucleic Acids Research. 2015; 43:D593–D598. [PubMed: 25414355]
6. Kembel SW, Wu M, Eisen JA, Green JL. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. PLoS Comp Biol. 2012; 8:e1002743.
7. Angly FE, et al. CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. Microbiome. 2014; 2:1–13. [PubMed: 24468033]
8. Klappenbach JA, Dunbar JM, Schmidt TM. rRNA operon copy number reflects ecological strategies of bacteria. Appl Environ Microbiol. 2000; 66:1328–1333. [PubMed: 10742207]
9. Stevenson BS, Schmidt TM. Life history implications of rRNA gene copy number in Escherichia coli. Appl Environ Microbiol. 2004; 70:6670. [PubMed: 15528533]
10. Dethlefsen L, Schmidt TM. Performance of the translational apparatus varies with the ecological strategies of bacteria. Journal of Bacteriology. 2007; 189:3237–3245. [PubMed: 17277058]
11. Vieira-Silva S, Rocha EPC. The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics. PLoS Genet. 2010; 6:e1000808. [PubMed: 20090831]
12. Giovannoni SJ, Thrash JC, Temperton B. Implications of streamlining theory for microbial ecology. ISME J. 2014:1–13. doi:10.1038/ismej.2014.60.
13. Eichorst SA, Kuske CR, Schmidt TM. Influence of plant polymers on the distribution and cultivation of Bacteria in the phylum Acidobacteria. Appl Environ Microbiol. 2011; 77:586–596. [PubMed: 21097594]
14. Martiny AC, Treseder K, Pusch G. Phylogenetic conservatism of functional traits in microorganisms. ISME J. 2013; 7:830–838. [PubMed: 23235290]
15. Condon C, Liveris D, Squires C, Schwartz I, Squires CL. rRNA operon multiplicity in Escherichia coli and the physiological implications of rrn inactivation. Journal of Bacteriology. 1995; 177:4152–4156. [PubMed: 7608093]
16. Stouthamer AH. A theoretical study on the amount of ATP required for synthesis of microbial cell material. Antonie Van Leeuwenhoek. 1973; 39:545–565. [PubMed: 4148026]
17. Fegatella F, Lim J, Kjelleberg S, Cavicchioli R. Implications of rRNA operon copy number and ribosome content in the marine oligotrophic ultramicrobacterium Sphingomonas sp. strain RB2256. Appl Environ Microbiol. 1998; 64:4433–4438. [PubMed: 9797303]

18. Kurland CG. Translational accuracy and the fitness of bacteria. Annu Rev Genet. 1992; 26:29–50. [PubMed: 1482115]
19. Carini P, et al. Discovery of a SAR11 growth requirement for thiamin's pyrimidine precursor and its distribution in the Sargasso Sea. ISME J. 2014; 8:1727–1738. [PubMed: 24781899]
20. Strzelczyk E, Leniarska U. Production of B-group vitamins by mycorrhizal fungi and actinomycetes isolated from the root zone of pine (Pinus sylvestris L.). Plant and soil. 1985
21. Morris JJ, Lenski RE, Zinser ER. The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss. mBio. 2012; 3:e00036–12–e00036–12. [PubMed: 22448042]
22. Raven JR, Andrews M, Quigg A. The evolution of oligotrophy: implications for the breeding of crop plants for low input agricultural systems. Annals of Applied Biology. 2005; 146:261–280.
23. Taylor JR, Stocker R. Trade-offs of chemotactic foraging in turbulent water. Science. 2012; 338:675–679. [PubMed: 23118190]
24. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research. 2016; 44:D457–D462. [PubMed: 26476454]
25. Redmond MC, Valentine DL. Natural gas and temperature structured a microbial community response to the Deepwater Horizon oil spill. Proc Natl Acad Sci USA. 2012; 109:20292–20297. [PubMed: 21969552]
26. Shrestha PM, Noll M, Liesack W. Phylogenetic identity, growth-response time and rRNA operon copy number of soil bacteria indicate different stages of community succession. Environmental Microbiology. 2007; 9:2464–2474. [PubMed: 17803772]
27. Nemergut DR, et al. Decreases in average bacterial community rRNA operon copy number during succession. 2015; 10:1147–1156.
28. Young VB, Schmidt TM. Antibiotic-associated diarrhea accompanied by large- scale alterations in the composition of the fecal microbiota. Journal of Clinical Microbiology. 2004; 42:1203–1206. [PubMed: 15004076]
29. Wieder WR, Bonan GB, Allison SD. Global soil carbon projections are improved by modelling microbial processes. Nature Climate Change. 2013; 3:1–4.
30. Lee ZM, Schmidt TM. Bacterial growth efficiency varies in soils under different land management practices. Soil Biology and Biochemistry. 2014; 69:282–290.
31. Eagon R. Pseudomonas natriegens, a marine bacterium with a generation time of less than 10 minutes. Journal of Bacteriology. 1962; 83:736. [PubMed: 13888946]
32. Conn HJ. The identity of Bacillus subtilis. The Journal of Infectious Diseases. 1930
33. Datta S, Costantino N, Court DL. A set of recombineering plasmids for gram- negative bacteria. Gene. 2006; 379:109–115. [PubMed: 16750601]
34. Gorlach K, Shingaki R, Morisaki H, Hattori T. Construction of eco-collection of paddy field soil bacteria for population analysis. Journal of General and Applied Microbiology. 1994; 40:509–517.
35. Schut F, et al. Isolation of Typical Marine Bacteria by Dilution Culture: Growth, Maintenance, and Characteristics of Isolates under Laboratory Conditions. Appl Environ Microbiol. 1993; 59:2150–2160. [PubMed: 16348992]
36. Schut F, Gottschal JC, Prins RA. Isolation and characterisation of the marine ultramicrobacterium Sphingomonas sp. strain RB2256. FEMS Microbiol Rev. 1997; 20:363–369.
37. Stevenson BS, Eichorst SA, Wertz JT, Schmidt TM, Breznak JA. New Strategies for Cultivation and Detection of Previously Uncultured Microbes. Appl Environ Microbiol. 2004; 70:4748–4755. [PubMed: 15294811]
38. Eichorst SA, Breznak JA, Schmidt TM. Isolation and Characterization of Soil Bacteria That Define Terriglobus gen. nov., in the Phylum Acidobacteria. Appl Environ Microbiol. 2007; 73:2708–2717. [PubMed: 17293520]
39. Simon M, Azam F. Protein content and protein synthesis rates of planktonic marine bacteria. Marine ecology progress series. 1989; 51:201–213.
40. Kanehisa M, et al. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Research. 2013; 42:D199–D205. [PubMed: 24214961]
41. Miller, LD.; Russell, MH.; Alexandre, G. Advances in Applied Microbiology. Vol. 66. Elsevier Inc.; 2009. p. 53-75.

42. Porter SL, Wadhams GH, Armitage JP. Signal processing in complex chemotaxis pathways. Nat Rev Micro. 2011; 9:153–165.
43. Jurgenson CT, Begley TP, Ealick SE. The Structural and Biochemical Foundations of Thiamin Biosynthesis. Annu. Rev. Biochem. 2009; 78:569–603. [PubMed: 19348578]
44. Quast C, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research. 2013; 41:D590–D596. [PubMed: 23193283]
45. Westram, R., et al. Handbook of Molecular Microbial Ecology: Metagenomics and Complementary Approaches. de Bruijn, FJ., editor. Vol. 1. John Wiley & Sons; 2011. p. 399-406.
46. Munoz R, et al. Release LTPs104 of the All-Species Living Tree. Syst Appl Microbiol. 2011; 34:169–170. [PubMed: 21497273]
47. R Core Team. [23rd July 2015] R: A Language and Environment for Statistical Computing. 2014. www.R- project.org Available at:
48. Tung Ho LS, Ane C. A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait Evolution Models. Systematic biology. 2014; 63:397–408. [PubMed: 24500037]
49. Hadfield J. MCMC Methods for Multi-response Generalized Linear Mixed Models: The MCMCglmm R Package. Journal of Statistical Software. 2010; 33
50. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution. 2012; 3:217–223.
51. Wickham, H. ggplot2: elegant graphics for data analysis. Springer; 2009.
52. Ligges U, Mächler M. Scatterplot3d - an R Package for Visualizing Multivariate Data. Journal of Statistical Software. 2003; 8:1–20.
53. Ludwig W, et al. ARB: a software environment for sequence data. Nucleic Acids Research. 2004; 32:1363–1371. [PubMed: 14985472]
54. Maddison, WP.; Maddison, DR. [25 March 2016] Mesquite: a modular system for evolutionary analysis. 2015. www.mesquiteproject.org Available at: http://mesquiteproject.org.
55. Ives AR, Garland T. Phylogenetic logistic regression for binary dependent variables. Systematic biology. 2010; 59:9–26. [PubMed: 20525617]
56. Kümmerli R, Schiessl KT, Waldvogel T, McNeill K, Ackermann M. Habitat structure and the evolution of diffusible siderophores in bacteria. Ecol Lett. 2014; 17:1536–1544. [PubMed: 25250530]

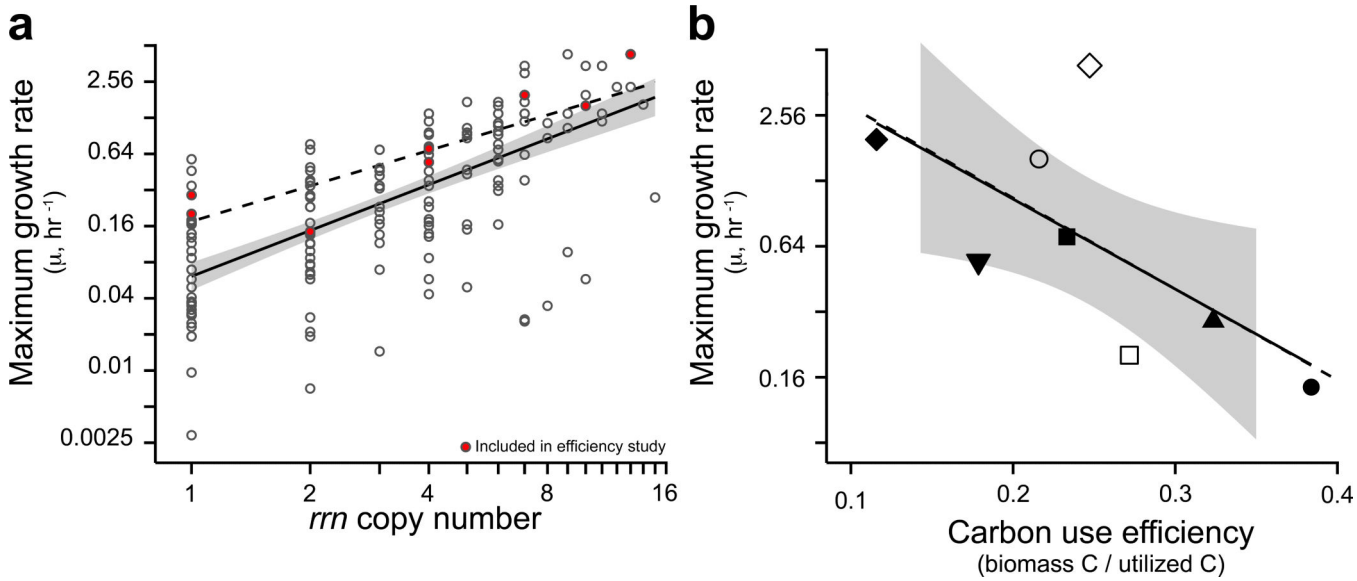


**Figure 1. Maximum growth rate is related to a bacterium's *rrn* copy number (a, n=184) and carbon use efficiency (b, n=8)**

Non-phylogenetic OLS regression (solid lines with 95% CI) reveals that these traits are correlated (**a**, $p < 0.001$; **b**, $p = 0.037$) and phylogenetic regression (dashed lines) demonstrates that the relationships can't be explained by shared ancestry (**a**, $p < 0.001$; **b**, $p = 0.036$). Mean carbon use efficiency is plotted in panel b from two independent flasks, *i.e.* biological replicates. Species represented in panel b are: *Sphingopyxis alaskensis* RB2256 (▲), *Acidobacteriaceae* sp. TAA166 (□), *Rhodospirillaceae* sp. PX3.14 (●), *Pseudomonas* sp. HF3 (■), *Arthrobacter* sp. EC5 (▼), *Escherichia coli* K12 MG1655 (◆), *Bacillus subtilis* Marburg ATCC 6051 (○), *Vibrio natriegens* ATCC 14048 (◇).

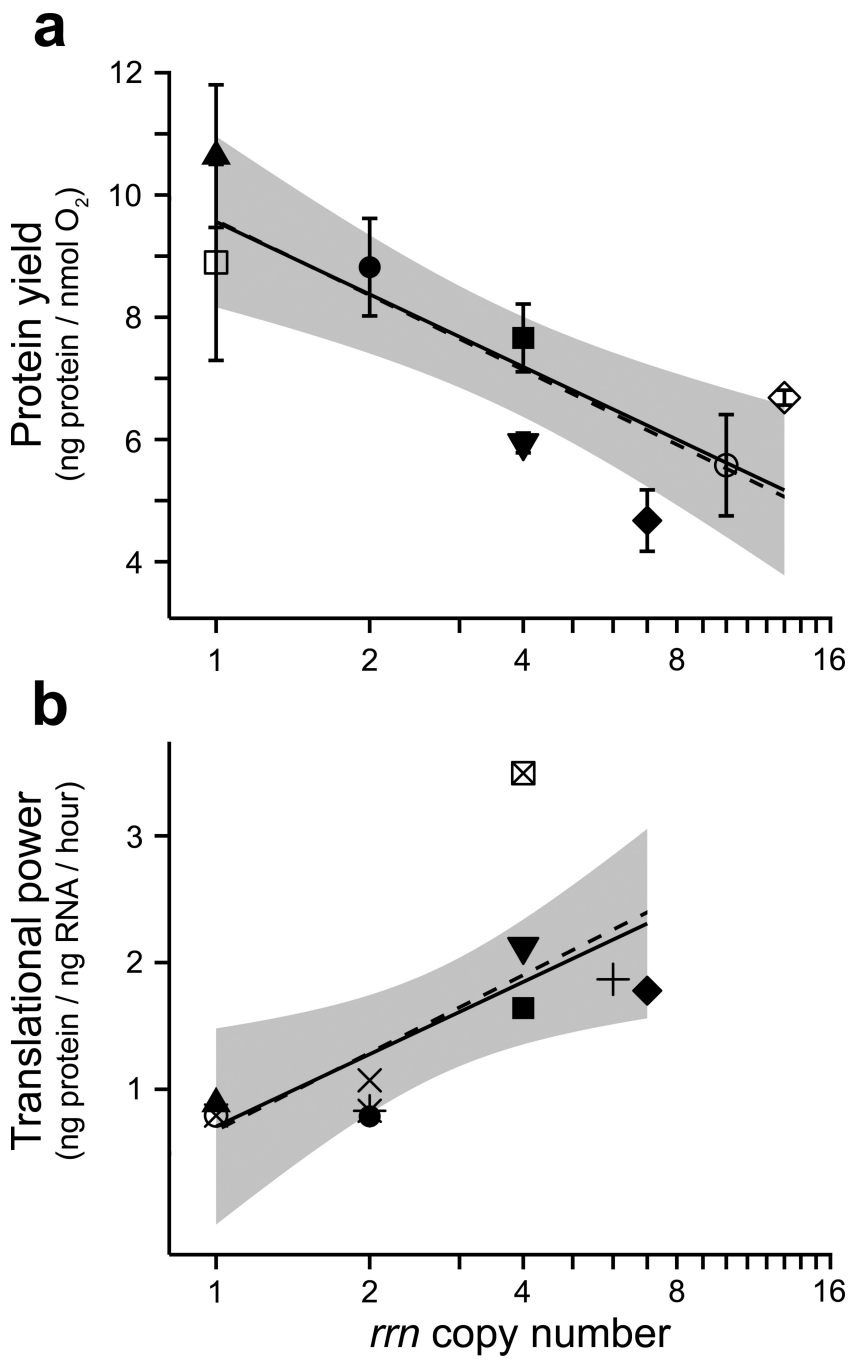


**Figure 2. Protein synthesis yield (a, n=8) and rate[10] (b, n=10) are correlated with $\log_2$-*rrn* copy number, but in opposite directions**

Non-phylogenetic OLS regression (solid lines with 95% CI) reveals these traits are correlated and phylogenetic regression (dashed line) demonstrates evolutionary history is not responsible for the relationship (Table 1, Supplementary Table 1). Mean protein yield and mean translational power are plotted, with error bars representing standard error of three independent flasks, *i.e.* biological replicates. Species represented in panel a are: *Sphingopyxis alaskensis* RB2256 (▲), *Acidobacteriaceae* sp. TAA166 (□), *Rhodospirillaceae* sp. PX3.14 (●), *Pseudomonas* sp. HF3 (■), *Arthrobacter* sp. EC5 (▼),

*Escherichia coli* K12 MG1655 (◆), *Bacillus subtilis* Marburg ATCC 6051 (○), *Vibrio natriegens* ATCC 14048 (◇). Species in panel b are the same as panel a, with the following changes and additions: *Escherichia coli* B REL607 (◆), *Comamonadaceae* sp. HS5 (⊗), *Mycobacterium* sp. PX3.15 (×), *Sphingobacteriaceae* sp. LC9 (✳), *Oxalobacteriaceae* sp. EC4 (☒), *Sphingobacteriaceae* sp. EC2 (+).

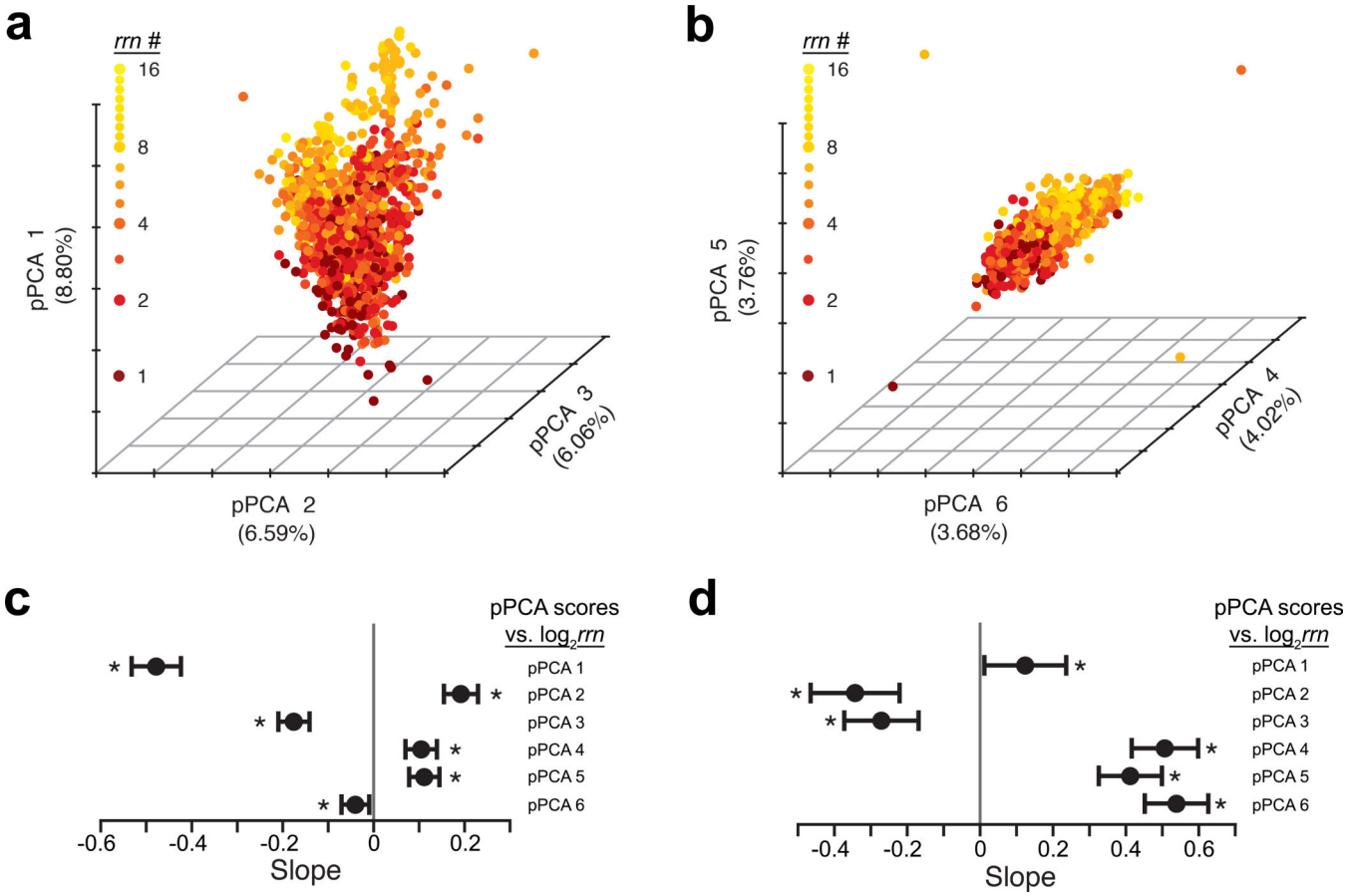


**Figure 3. Phylogenetic principal component analysis (pPCA) of genome content of 1,167 unique bacterial species**

KEGG modules (a) or orthologs (b) datasets were analyzed separately. pPCA genome scores were regressed against $\log_2$-*rrn* and slopes of phylogenetic linear regressions are reported below their corresponding analysis (c and d; slope estimate plotted with its 95% confidence interval and *indicates slope $p < 0.05$). pPCA axes displayed in a and b have the highest magnitude slopes from the first six pPCA axes.

## Table 1

Summary of re source-associated traits regressed against $\log_2$ *rrn* copy number

| Trait effect | Trait | Life history adaptation? | Effect size* | Significance |
|---|---|---|---|---|
| Resource metabolism | $\mu_{MAX}$ | Yes | $\mu_{MAX}$ doubles with *rrn* doubling | **p < 0.001**# |
| | Efficiency<br>$^3$H-CUE<br>$^{14}$C-CUE | Yes | −3.6% CUE with *rrn* doubling<br>−3.2% CUE with *rrn* doubling | **p = 0.004**#<br>p = 0.057# |
| | Translational power | Yes | 0.61 units## with *rrn* doubling | **p = 0.014** |
| | Protein yield | Yes | −1.14 units## with *rrn* doubling | **p = 0.016** |
| | Genome streamlining<br>Genome size<br>Thiamine Biosynthesis | Yes | + 6.6 kbp from 1-15 *rrn*<br>+ 3 biosynthetic steps from 1-15 *rrn* | **p < 0.001**#<br>**pMCMC<0.001**^ |
| | Autotrophy | Uncertain | −5.1% probability from 1-15 *rrn* | p = 0.131 |
| Resource uptake | PTS transporters | Uncertain | < +1 PTS transporter from 1-15 *rrn* | **pMCMC = 0.007**^ |
| Resource sensing | Chemotactic motility | Yes | +11% probability from 1-15 *rrn* | **p = 0.035** |

# Linear, logistic

^ Poisson phylogenetic regression models were used for these analyses.

* Effect size for Poisson regression models is expressed as the difference in predicted trait count for 15 *rrn* compared to 1 *rrn*. Effect size for logistic regression models is expressed as the difference in predicted probability of encoding the trait for 15 *rrn* compared to 1 *rrn*. These values were used to encompass the maximum and minimum observed *rrn* copy numbers for bacteria.

## Translational power (gProtein $gRNA^{-1}$ $hr^{-1}$) derived from[10], while protein yield (ng protein nmol $O_2^{-1}$) was measured in this study.