Why Environmental Biomarkers Work: Transcriptome-Proteome Correlations and Modeling of Multi-Stressor Experiments in the Marine Bacterium *Trichodesmium*

Nathan G. Walworth[1+], Mak A. Saito[2+*], Michael D. Lee[3,4], Matthew R. McIlvin[2], Dawn M. Moran[2], Riss M. Kellogg[2], Fei-Xue Fu[1], David A. Hutchins[1], and Eric A. Webb[1*]

[1] Marine and Environmental Biology, Department of Biological Sciences, University of Southern California, Los Angeles, CA, 90089, USA.

[2]Marine Chemistry and Geochemistry Department, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA

[3]Blue Marble Space Institute of Science, Seattle, WA, 98104, USA

[4]Exobiology Branch, NASA Ames Research Center, Moffett Field, CA, 94035, USA

[+]Co-First Authors

[*]**Corresponding authors:** Mak Saito msaito@whoi.edu and Eric A. Webb eawebb@usc.edu;

June 22, 2021

Keywords:

Transcriptome-proteome, environmental biomarkers, marine microbes, Trichodesmium, metaproteomics

Abstract

Ocean microbial communities are important contributors to the global biogeochemical reactions that sustain life on Earth. The factors controlling these communities are being increasingly explored through the use of metatranscriptomic and metaproteomic environmental biomarkers, despite ongoing uncertainty about the coherence between RNA and protein signals. Using published proteomes and transcriptomes from the abundant colony-forming cyanobacterium *Trichodesmium* (strain *T. erythraeum* IMS101) grown under varying Fe and/or P limitation and/or co-limitation in low and high CO_2, we observed robust correlations of stress induced proteins and RNAs (i.e., those involved in transport and homeostasis) that can yield useful information on nutrient status under low and/or high CO_2. Conversely, transcriptional and translational correlations of many other central metabolism pathways exhibit broad discordance. A cellular RNA and protein production/degradation model demonstrates how biomolecules with small initial inventories, such as environmentally responsive proteins, can achieve large increases in fold-change units, as opposed to those with higher basal expression and inventory such as metabolic systems. Microbial cells, due to their close proximity to the environment, tend to show large adaptive responses to environmental stimuli in both RNA and protein that result in transcript-protein correlations. These observations and model results demonstrate a multi-omic coherence for environmental biomarkers and provide the underlying mechanism for those observations, supporting the promise for global application in detecting responses to environmental stimuli in a changing ocean.

Introduction

A recurring question in the interpretation of transcriptome and proteome datasets is the extent to which they co-vary. Messenger RNA (mRNA) and protein levels have been reported to generally be uncorrelated within a single cell, and only modestly correlated in populations of cells, due to differences in half lives and degradation rates or phase variation within a population, respectively [1]. It has also been shown that genes and their corresponding proteins associated with different cellular processes (e.g., core central vs stress metabolism) may retain varying degrees of correlation [2]. While observation of correlations in larger organisms may be challenging due to internal tissues being more remote environmental signals, microbes due to their small size and immersion within the environment often maintain multiple adaptive response capabilities to common environmental stimuli. Microbial datasets with transcriptomic and proteomic methods applied to the same experiment(s) are becoming more common and hence could aid in detection and interpretation of key environmental processes. In recent years, concurrent measurements of transcripts and proteins have been conducted in marine microbes such as *Pelagibacter* [3], *Prochlorococcus* [4], the marine diatom *Thalassiosira pseudonana* [5], the brown alga *Aureococcus anophagefferens* [6], the polar alga *Phaeocystis antarctica* [7], and the diazotrophic cyanobacterium *Trichodesmium* [8]. These studies also observed varying extents of transcriptome-proteome coupling, with correlations observed particularly for genes involved in responding to environmental stresses, such as P, Fe, or vitamin B_{12} limitation. Despite these efforts, there is arguably a lack of consensus in the marine ecology community regarding the extent that RNA transcripts and proteins should correlate, with many having the opinion that correlations do not occur.

In this study we synthesize the results from a number of *Trichodesmium* transcriptome-proteome datasets across a spectrum of conditions to understand their transcriptional and translational responses on a mechanistic level. *Trichodesmium spp*. are filamentous, buoyant microorganisms that can commonly grow in macroscopic colonies in close association with other microbes and have the capability to form massive blooms [9]. Given their ability to fix both carbon and nitrogen, *Trichodesmium* spp. have relevance to both global productivity and biogeochemistry [10,11]. Together with other marine microbes, they can impact both ecosystem stability and climate feedbacks [12]. *Trichodesmium* is among one of several oceanic cyanobacteria that are globally significant sources of N [10,13], as well as unicellular forms (*Crocosphaera* spp.,

Candidatus Atelocyanobacterium thalassa and *Cyanothece*) and other heterocystous, symbiotic forms [11], that can have relatively high cell numbers [14] and biogeochemical importance [15,16]. *Trichodesmium* is considered one of the most important diazotrophs in many tropical and subtropical regimes [10,17]. As a result, research emphasis has been placed on developing field-ready, nutrient-limitation/stress markers to define the factors that control growth and N_2 fixation in this important genus. These combined efforts have shown that iron (Fe) and phosphorus (P) primarily limit *Trichodesmium* across much of the global oceans [10,11,18,19]. However, comparatively fewer studies have been conducted examining the interactions among future high CO_2 concentrations, Fe, and P in the context of holistic, molecular physiology.

A motivation in understanding the dynamics of RNA and protein responses is the interpretation of natural microbial populations, and in particular the expression of genes that can provide clues to ecological or biogeochemical processes at play. Advances in sequencing (e.g., metagenomics/metatranscriptomics) and mass spectrometry (e.g., proteomics) technologies have enabled worldwide microbial characterizations from different biogeochemical regimes (e.g., Global Ocean Sampling Dataset [20]). While these data have revolutionized how we think about microbially mediated processes, extrapolating biogeochemistry from relative abundances in metagenome data may result in biased interpretation. For example, it is now recognized that globally abundant microbial species initially identified via metagenomics may not necessarily be dominant members of the transcriptionally active microbial community that is primarily responsible for biogeochemical turnover [21]. While metagenomics studies serve as powerful hypothesis-generating datasets and have unmasked previously underappreciated ecosystem processes (e.g., proteorhodopsin diversity [20,22]), they have also underscored the need to understand the dynamics of intracellular molecular processes in the context of organismal physiology both in the laboratory and field. Thus, it is important to investigate molecular dynamics of biogeochemically-important microbes under differing nutrient and temporal regimes to resolve mRNA-protein dynamics that ultimately drive important biogeochemical processes. Here we focus primarily on N_2 fixation. In past studies we characterized the physiological and evolutionary responses of the globally important, photoautotrophic, N_2-fixing cyanobacterium, *Trichodesmium erythraeum* IMS101 (hereafter IMS101) to high CO_2. Using these cell lines as starting genetic backgrounds (i.e., replicates adapted to both 380 µatm and 750 µatm CO_2), we then conditioned them to iron-limited, phosphorus-limited, and iron/phosphorus (Fe/P) co-

limitation to define their predicted climate change-impacted Fe- and P-stress molecular responses [23]. Consistent with other studies [24], this demonstration of the fitness advantage conferred by Fe/P 'balanced limitation' compared to single limitation alone appears to contradict the long-standing Liebig limitation model and has implications for global biogeochemical cycles in both the current and future ocean such as increased exogenous nitrogen scavenging coupled to decreases in N_2 fixation [23,25,26].

Here we build on and synthesize our prior results to examine steady-state transcriptional/translational/physiological relationships in the context of environmentally relevant (e.g., Fe and/or P affected) nutrient regimes using our high and low CO_2-adapted, IMS101 cells lines [23,25,27]. As mentioned, numerous studies utilizing diverse techniques ranging from chemical quotas, enzyme activities to gene or protein-based molecular stress markers have pointed to the importance of Fe and P to *Trichodesmium* N_2 fixation *in situ* [18,28,29]. Despite the value of molecular markers for indicating species specific bioavailability of nutrients, concerns persist that gene/protein expression can be decoupled with biogeochemically relevant rates like N_2 fixation (e.g. [30]) and thus give an inconsistent view of *in situ* limitation. We performed meta-analyses of our published, long-term Fe- and P-single limited and co-limited proteomic and transcriptomic datasets [23,25] to determine which genes and proteins give congruent expression patterns during mid-day across 8 experimental conditions vs those that do not. This analysis shows that major components of core metabolic pathway genes and proteins show incongruent responses to environmental perturbations whereas many well-classified nutrient-responsive genes/proteins show consistent correlations under both low and high CO_2 regimes. A model of the cellular inventories for RNA and protein biomolecules that compares the timing of biosynthesis and degradation was created that provides a mechanistic explanation for the observed coherence of responses to environmental perturbations. Finally, the environmental biomarkers that have been characterized thus far in marine microbes are briefly reviewed.

Materials and Methods

Culturing and Molecular Extractions

Detailed culturing and protein [23] and RNA [31] extraction methods can be found in our previously published papers.

Differential expression analysis

Differential expression data was from previously described datasets [25].

Multivariate and Pairwise Analyses

Redundancy analysis was conducted on TMM-normalized transcript levels and nutrient concentrations using the vegan package [32]. RDA ordinations were compared via the Procrustes test with the "protest" function in vegan using default settings.

Nonmetric multidimensional scaling (NMDS) was conducted using the "metaMDS" function from the vegan package with default settings (Bray-Curtis dissimilarities computed using Wisconsin double standardization) except for "autotransform=FALSE". Hierarchical clustering with multiscale bootstrap resampling was conducted on Bray-Curtis dissimilarities calculated from TMM-normalized transcript levels using the pvclust package [33] with the following settings: method.dist="euclidean", method.hclust="average", nboot=1000.

The "cor.test" function in "R" was used to calculate Pearson correlation coefficients on log2 fold changes (relative to the r380 condition; see below) of average TMM-normalized transcript levels vs log2 fold changes of average normalized spectral counts of the corresponding protein product detected in Walworth et al. 2016 [23]. Significance tests between mean correlation coefficients were carried out via the "perm" package[34] using a two-sample permutation test with Monte Carlo simulation and 2000 permutations.

Gene Ontology (GO) enrichment analysis

As previously described in Walworth et al. 2016 [31], Gene Ontology (GO) mappings for *Trichodesmium* were downloaded from the Genome2D web server (http://genome2d.molgenrug.nl). The "phyper" function in "R" (R Core Team 2014) was used to test for significant enrichment of GO pathways and p-values were corrected with the Benjamin and Hochberg method [35] using the "p.adjust" function ($p \leq 0.05$). Finally, genes in enriched GO categories were manually checked.

RNA-Protein Model

A model simulating the biosynthesis of RNA transcripts and proteins was created in MATLAB (MathWorks) using parameters listed in Table 1. One-minute time steps over a duration of 3 days (4320 min) within a population of 100 cells was modeled. Expression was modeled with a signal parameter that varied between 0 and 1 representing the gene being turned on or off with a simple linear ramping function. Basal signal level (s_0) was set to 0.01, representing 1% of the population or 1 cell within 100 cells with gene expression to avoid zero

denominator in fold-change estimates. In an alternate scenario representing metabolic functions in continuous use, a basal expression of 20% (0.2) was used. Rate constants for production of RNA and protein (k_{RNA}, $k_{protein}$), as well as their decay (k_{-RNA}, $k_{-protein}$) were obtained from experimental measurements in *Escherichia coli* and *Saccharomyces cerevisiae* where the degradation of each type of biomolecule was inhibited and measured [36,37]. Change in the inventories of RNA and protein biomolecules was estimated using simple production and loss terms (Equations 1 and 2):

$$\frac{dRNA}{dt} = (k_{RNA} * s) - (k_{-RNA} * [RNA]) \quad (1)$$

$$\frac{dProtein}{dt} = (k_{Protein} * [RNA]) - (k_{-Protein} * [Protein]) \quad (2)$$

Turnover time estimates of these biomolecules are, to our knowledge, unavailable for marine organisms. While they may be different, presumably slower based on observed growth rates, it seems a reasonable assumption that the large difference between RNA and protein decay would be maintained. Hence the trends observed here are likely to be consistent or further enhanced, particularly if the RNA degradation activity in marine microbes is slower than that of *E. coli* as might be expected.

This model has similarities to that of Moran *et al.* 2013 [38], but is applied here to focus on any coordination between RNA and protein and how results affect fold change estimates. Moran *et al*. instead explored random triggers for transcript production using 10 random events per day of 1 transcript per event, and then simulated environmental stimuli by increasing to 3 transcripts per event. In contrast, the model in the present study estimates production of RNA and protein molecules based on their upstream signal (signal, transcripts) as described in equations 1-2, rather than assigning a constant number of seven protein copies per transcript. Moran *et al.* observed the total number of mRNA molecules in a typical marine bacterium was only ~200 per cell[38]. These observations were incorporated into the model by allowing k_{RNA} to be 100 to enable signal input to be transcribed into 100 transcripts across a population of 100 cells, and then tuning the protein production $k_{Protein}$ parameter to produce 6-7 proteins per transcript also based

on Moran et al., 2013. Similarly Moran et al. used a RNA half-life of 1.5 minutes, whereas this model uses a more conservative RNA half-life of 6.8 minutes based on a tiling microarray study in *E. coli*[36].

Results and Discussion

Global comparison of CO_2-impacted, nutrient-limited molecular dynamics

A comprehensive depiction of the prior experimental design can be found in Fig. S1a and in Walworth et al. 2016 [23] in addition to physiological and proteomic results. To investigate shifts in molecular metabolism underlying Fe-limited, P-limited, and Fe/P co-limited metabolisms as they interact with predicted future CO_2 conditions, we examined the global transcriptional output in each scenario from Walworth et al. 2018 [8] and their corresponding translational correlations with proteins identified in Walworth et al. 2016 [23]. Across all treatments, we detected transcription of ≥96% and translation of 37% (see below) of the IMG annotated genes. Instances where comparisons between transcripts and proteins were not possible due to one of those biomolecules not being detected were excluded. Initial examination of the dataset in log2 fold change space revealed a cluster of points around the origin with a subset of data extending into the northeast and southwest quadrants implying a co-varying relationship between some transcript-protein pairs (Fig. 1). It is important to note that non-translated mRNAs can also potentially represent 'irrelevant expression', if these transcripts are not regulatory. Thus, there is much to be learned from future studies reconciling the constantly developing techniques of RNA and protein based 'omics. Yet, there have been few studies to date comparing multi-condition proteomes with concomitantly taken transcriptomes.

To this end, we first conducted exploratory analysis on the global transcriptional output without constraining for environmental variables (see below) in order to observe how replicates clustered without imparting treatment information (i.e. no assumptions). We used both nonmetric multidimensional scaling (NMDS) and hierarchical clustering with multiscale bootstrap resampling (replicates = 1,000) [40]. NMDS revealed consistent nutrient-limited patterns across replicates (Fig. S1b), which was further supported by hierarchical clustering (Fig. S1c) that generated two high-confidence clusters with AU-bootstrap p-values > 0.95 (i.e. rejection of the null hypothesis that the clusters do not exist at the 0.05 significance level). In these unconstrained analyses, the 380-Fe condition (low-CO_2, low iron) demonstrated greater inter-

replicate transcriptional variation, although both replicates were indeed Fe-limited as evidenced by both reduced growth [23] and several upregulated Fe-stress genes (Fig. S1b; see below and Walworth et al. 2018 [8]). Interestingly, Fe-limited treatments (380- and 750-Fe) clustered with the nutrient replete ones (r380 and r750) supporting the notion that P-limitation evokes a more varied molecular response than Fe-limitation relative to replete nutrient regimes (Fig. S1c). This trend was further corroborated through hierarchical clustering of transcript levels of all differentially expressed (DE) genes (i.e., those DE in any treatment relative to r380; n = 1943; Fig. 2a). To better resolve sources of variation producing these groupings, we further clustered upregulated (n = 951; Fig. 2b) and downregulated (n = 1074; Fig. 1c) transcripts, which yielded strikingly different trends among Fe-limited and replete conditions. Upregulated genes exhibited significantly different (bootstrap confidence level > 0.95) clusters among Fe-limited and CO_2 regimes relative to those produced by clustering all DE genes while those of downregulated genes maintained similar groupings to those of DE genes. These patterns suggest that Fe-limited and replete clusters produced from all DE genes (Fig. 2a) may have been driven more so by downregulated transcriptional variation of the DE gene pool. Hence, mechanistic adaptations to high CO_2 and responses to Fe-limitation both seem to be more similar to each other than to molecular responses under P-limitation. Interestingly, both up- and downregulated P-limited clusters were statistically analogous to those produced by clustering all DE genes (Fig. 2a) and total genes (Fig. S1c), which further corroborates the notion that P-limitation evokes a highly conserved yet more varied transcriptional rearrangement than either high CO_2 or Fe-limitation.

Next, we employed constrained multivariate methods that attempts to explain variation based on the given environmental variables (e.g. RDA analysis) for both transcripts and proteins. When relating nutrient-limited proteome variation identified via redundancy analysis (RDA) in Walworth et al. 2016 [23] (n = 1248; 37% of genome coding potential) to that of their corresponding gene transcripts (corresponding-RNA-RDA; Fig. 2d), similar nutrient-limited patterns across replicates were observed with most of the proportional variance explained by the environmental variables for both transcripts (adjusted $R^2 = 0.56$) and proteins (adjusted $R^2 = 0.59$), respectively. Moreover, a statistically significant correlation was observed ($r = 0.75$; $P < 0.001$) when comparing the corresponding-RNA-RDA and prot-RDA ordinations via permutational, least-squares orthogonal analysis (i.e., Procrustes analysis; Methods) indicating that, overall relative transcriptional variation among treatments was captured in the

corresponding proteome variation (Fig. 2e).

Performing the same RDA analysis on the global transcriptional output (global-RNA-RDA; Fig. 2f; adjusted $R^2 = 0.50$; n = 4862) displayed a strikingly consistent ordination relative to that of both the prot-RDA (represented in Figure 1c in Walworth et al. 2016[23] and corresponding-RNA-RDA (Fig. 2d). In fact, when analyzing congruence of the global-RNA-RDA to both the prot-RDA [23] and the corresponding-RNA-RDA, respectively, a robust, statistically significant correlation was observed relative to both the prot-RDA ($r = 0.75$, $P < 0.001$) and the correlated-RNA-RDA ($r = 0.98$; $P < 0.001$). Taken together, these analyses indicate that both transcriptional and translational variation inherent in the overlapping 37% of the detected proteome and 96% of the detected transcriptome are not only representative of one another across nutrient-limited conditions, but also of global transcriptional variation. In other words, this congruence implicates that the detected proteome variation is indeed representative of overall treatment-specific, transcriptional variation in IMS101 for Fe, P, Fe/P, and CO_2.

Another striking observation is that 380-Fe/P replicates were more closely associated to P-limited treatments (380- and 750-P) while 750-Fe/P replicates were more correlated with Fe-limited ones (380- and 750-Fe) for both ordination-based (RDA, Fig 2d,f; NMDS, Fig. S1b) and clustering analyses (Fig. 2a,b,c and Fig. S1c). Hence, these trends imply that regardless of the amount of variation explained by the given environmental variables, the global transcriptional profile of Fe/P co-limited metabolism is more similar to that of P-limitation when adapted to current CO_2 levels (380-Fe/P) but is more similar to Fe-limitation when adapted to high CO_2 (750-Fe/P).

From global to pairwise mRNA-protein correlations

Bacterial messenger RNA (mRNA) typically have much shorter half-lives than those of proteins and can be rapidly transcribed in response to quickly changing environmental conditions [41]. For example, in an *E. coli* cell, many mRNAs are typically degraded within minutes, whereas their corresponding proteins can have longer lifetimes than the cell cycle [42]. These data indicate that intracellular mRNA copy number at any instant may typically reflect recent transcriptional activity, while protein levels in the same instant can represent a relatively longer history of accumulated expression. Indeed, simultaneous measurements of mRNA and their corresponding

protein concentrations have been shown to be uncorrelated within a single cell at a single time point, while mRNA levels integrated over many cells in a population generally correlate with protein concentrations [42]. Different environmental variables invoking specific metabolic pathways can also impact overall mRNA-protein correlations [1,2].

Unlike prior other microbial studies investigating mRNA-protein correlations across one or two treatments [2], we examined correlations across eight different growth conditions. As in most organisms to date, IMS101 transcript levels modestly correlated to their corresponding protein abundances across all treatments (n = 1,246; squared Pearson correlation coefficient, r^2 = 0.36), suggesting that other forms of regulation might need to be invoked to explain the majority of variation [1]. Upon calculating Pearson correlation coefficients across all treatments for genes that were differentially expressed (DE) in at least one experimental condition relative to r380 versus those that were not called as differentially expressed (NDE) in any condition, we observed a significantly greater mean correlation coefficient for DE (r^2 = 0.35) vs NDE (r = 0.03) genes ($p < 0.0005$; two-sample permutation test with Monte Carlo simulation). This same robust trend was also observed when comparing average correlation coefficients for DE vs. NDE genes for each pairwise treatment comparison (Fig. S2). Hence, this greater correlation suggests that transcription and translation of environmentally responsive genes and their protein products may be more closely coupled on average than those that do not respond to changing environmental conditions.

Pathway-specific molecular relationships

To search for metabolic pathways harboring highly correlated transcript and protein abundances, we calculated Pearson's correlation coefficients (PC) and kept genes retaining PC values ≥ 0.7 and p-value < 0.05. We then mapped genes onto their Gene Ontology (GO) pathways and searched for significant GO enrichment using the hypergeometric test with Benjamini-Hochberg correction (FDR ≤ 0.05; Dataset S1). Enriched GO pathways included translation, nitrogen fixation, oxidation-reduction, calcium ion binding, proteolysis, iron-sulfur cluster binding, and photosynthetic electron transport/light reactions. Hence, these data suggest that biosynthesis of these protein products increases with increasing transcript accumulation. Therefore, their abundances may be more representative of instantaneous cellular activity rather

than those with low PC values, in which the latter may be a product of constitutive activity and/or reduced protein degradation. For example, well-characterized Fe- and P-stress genes exhibited high PC values in conjunction with genes involved in photosynthetic light reactions (Fig. 3). In all cases, the interaction of high CO_2 with limiting Fe irrespective of fluctuating P (750-Fe and 750-Fe/P) yields increases in both transcript and protein levels for all Fe-stress genes relative to their low CO_2/low Fe conditions (i.e. 380-Fe and 380-Fe/P). Hence, the expression of these Fe-stress biomarkers increases under limiting Fe in future ocean conditions suggesting their profiles to be robust for *in situ* profiling. The genes involved in photosynthesis light reactions (Fig. 3c) exhibit parallel trends possibly suggesting co-regulation by iron limitation or growth rate. Interestingly, P-stress transcript and protein levels robustly respond to limiting P irrespective of Fe and CO_2, but increased CO_2 has little to no effect on P-stress molecular machinery (Fig. 3b). Additionally, these data suggest that different subunits of the same phosphate transport system (i.e. Pst) could respond differently to fluctuating P as transcript and protein levels of, for example, the PstB subunit increased more considerably than those of PstS under limiting P (Fig. 3b).

Conversely, GO-pathways specific to the NDE pool with low PC values involve general biosynthetic processes, magnesium ion binding, proteolysis, respiration (e.g., TCA cycle), nucleotide metabolism, inosine monophosphate (IMP) biosynthesis, and amino acid metabolism (Fig. 4). Of note, decoupling of TCA gene expression and enzyme abundance in yeast has also been previously observed [2]. Interestingly, the GO pathway assigned to photosystem I reaction center (GO: 0003989) experience significant enrichment of genes with low PC values (Dataset S1) while conversely, the overall photosynthesis light reaction GO pathway (GO:0019684) was enriched with genes with high PC values (Fig. 3c). Hence, it is important to distinguish between potentially constitutive vs. responsive components of large, multifaceted pathways with many components that can have widely differing roles (e.g., energy flow vs. structure). Overall, it seems the persistent activity of certain core pathways to maintain basic cellular homeostasis, irrespective of changing environmental conditions, may result in more decoupling of transcription and translation on average (Fig. 4), relative to pathways that synthesize new transcripts to generate protein products specific to particular environmental triggers (e.g., Fig. 3).

RNA-Protein Modeling: Influence of biomolecule inventories on fold-change dynamics

In order to explore the potential causality of RNA-protein coupling and decoupling we generated a model for production and decay of both biomolecules (Fig. 5). Model parameters include previously published values for RNA and protein degradation based on studies of whole transcriptome and proteome degradation measurements from model organisms (*E. coli* [36] and yeast [37], respectively) as well as delays between RNA and protein expression observed in *Prochlorococcus* [4] (Table 1). Notably, RNA decay rates are approximately two orders of magnitude higher (~60-275 fold) than that of protein degradation. While differences in degradation rates occur between different specific transcripts and protein molecules, there is clearly a large difference between transcript and protein degradation rates, as expected based on the transient information transmission role of RNA versus the long-term functional roles of proteins. The choice of *E. coli* and yeast as organisms for obtaining decay parameters will likely vary from that of various marine microbes, with both degradation processes likely being slower within slower growing marine microbes such as *Trichodesmium*. However, the large difference in cellular decay rates of RNA and protein rates is almost certainly a universal phenomenon, and hence we expect the general trends of temporal overlap between transcripts and protein inventories modeled here to be relevant even if future taxon-specific activities vary somewhat from those used here. In particular, if the RNA degradation constant is lower in slower growing organisms due to lower abundance and/or efficiency of RNases, an increased temporal overlap between transcripts and proteins would result in greater opportunity for correlation between the two types of biomolecules. Additional parameters in the model include production of RNA and protein, as well as constants for temporal offsets between signal transduction and transcription (d_{RNA}), and transcription and translation processes ($d_{protein}$). For the latter, 3h was used, similar to the range observed in *Prochlorococcus* [4]. Finally, the model includes initial condition parameters whose influence can be explored including basal expression levels (1% and 20%) and time settings for onset, duration, and decay of signals to initiate transcription. The notion of the transcriptional signal is generalized here, whether they occur from riboswitches, two component regulatory systems, sigma factors or other mechanisms. Moreover, the onset and decay of the regulatory signaling is also idealized here as a linear ramp (using a “long slow ramp” that repeats daily), although it could be further qualified to have positive feedbacks to increase sensory capabilities such as observed in increased expression of two component regulatory systems, e.g., phosphate regulatory systems [43]. The influences of post-translational modifications (PTMs) were

purposely not included in this model in order to allow an examination of potential range of differences in RNA and protein fold-change expression levels prior to often invoked additional PTM regulatory controls.

The RNA-Protein expression model results imply that changes in basal expression level can have a major influence on the extent of RNA-proteome coupling and decoupling observed experimentally. For example, when there is no significant basal expression (1% or s_0=0.01; where a non-zero value was chosen to allow fold-change ratio estimate) as would be expected prior to an environmental stimulus such as phosphate or iron stress, both RNA and protein signals respond with a large increase in inventory and fold-change (Fig. 6 and 7 left panels). In contrast, when there was a 20% basal level of expression (s_0=0.2), as expected for routinely used metabolic machinery, a significant inventory of protein product accumulates (Fig. 6 right panels), causing the fold change signal to become muted (Fig. 6). These results are also displayed as a movie format (Supplemental Movie M1) to display the temporal unfolding of these transcript-protein inventories and their influence on fold changes results. The fold change decrease is accentuated if calculated each day as the protein inventory accumulates (e.g., the choice of what to normalize expression to in the fold change estimate). Notably, proteins that respond to environmental stress are consistent with the lower basal expression model that yields correlations between RNA and protein. In *Trichodesmium*, photosynthetic transcripts and proteins were also observed to be correlated (Fig 4c). While generally considered metabolic proteins with high inventories, this observed correlation can be particularly driven by iron limitation scenarios, which can remodel overall machinery by reducing the iron-rich photosystem I (PSI) in favor of PSII [23]. Iron stress transcripts and proteins like *isiA* can also concurrently increase in abundance solely in times of enhanced iron stress to form PSI-IsiA-chlorophyll–protein–antenna super-complexes [44] while in other conditions like phosphorus stress, they show little correlation due to the lack of an environmental trigger. Additionally, diazotrophs have been shown to have strong diel cycles that affect both transcripts and proteins of photosynthetic and nitrogen fixation enzymes in order to promote intracellular oxygen protection and iron conservation efforts [45,46]. Hence, the choice of when to sample transcriptomic and proteomic experiments can also be consequential. The short-reported half-lives of RNA molecules cause a rapid decrease in transcripts if they are not continually being produced [38]. The expected large two order of magnitude difference in half lives of RNA and protein molecules results in a

counterclockwise progression on RNA and protein space plots commonly used for assessing coherence between these biomolecules (Fig. 8). Choice of sampling time is typically estimated to allow sufficient time for RNA and protein biosynthesis, where if too early there is insufficient time for protein synthesis, and if too late RNA transcripts may have begun degradation [7]. Notably, the magnitude of these counterclockwise traces is affected by the basal expression level, where low basal expression creates an easily observable pattern due to the high fold change values for transcripts and proteins, while high basal expression results in the circular trace to be compressed near the origin. This is indeed what many experimental datasets look like with metabolic proteins found near the origin in log2 transcript-protein space (Fig. 1), while genes that respond strongly to environmental cues have stronger signals that project into the northeast quadrant of these plots. The model results here may be overly conservative, as it seems likely that marine microbes with much slower reproduction rates than *E. coli* or yeast would also have lower concentrations of RNases and concurrent longer residence times of RNA, as mentioned above. If lower RNase activity were common among marine microbes like *Trichodesmium*, the overlapping time window for co-existence of RNA and proteins would be enhanced.

Based on this simple modeling approach, the (often negative) conclusions about the coherence of RNA and protein expression in microbes appear to have been clouded by use of relative units of fold change that discriminate against metabolic proteins with larger inventories. In these cases, there may be highly abundant proteins that change in inventory by 20% for example, which would be a large change in copy number, but a small change in fold change ratios. Use of absolute units for both transcripts and proteins and efforts to calibrate with standards to achieve copy number per cell estimates would resolve this problem. Importantly however for environmental applications, the successful use of environmentally responsive genes in both RNA and protein space can be explained by their large signals relative to basal expression that allows easy detection of large fold changes. In short, the lack of observed correlations of metabolic systems represented in transcripts and proteins can be explained, at least partially, by an over-reliance on ratios of data, and those observations should not detract from the statistically significant correlations of environmentally response genes in RNA-protein space and their application as research and monitoring tools in environmental microbiology and biogeochemistry.

Together these empirical observations and model results validate the utility and provide a theoretical basis behind proteins that respond to environmental stimuli in marine microbes. Because organisms that comprise natural ecosystems are frequently limited in their productivity by nutritional elements, they often have evolved specific adaptive responses to nutritional scarcity. As such, these proteins and their mRNA transcripts represent ideal environmental biomarkers to detect controls on ecosystem productivity. Biomarkers for nutritional stress in marine phytoplankton and bacteria that have been identified are shown in Table 2, including those for nitrogen [43,48-50], phosphorus [5-6,51-53], iron [18-19, 54-61], vitamin B_{12}[62,63], and zinc[64]. Notably, proteins involved in adaptive responses in marine microbes are often distinct from those model organisms because of their very different environmental conditions. As a result, several of these marine biomarkers have only recently been discovered and it is almost certain that more remain to be discovered. In some cases, biomarkers for required elements such as cobalt do not appear to have adaptive responses (or their chemical species they respond to has not yet been identified)[65] and corresponding biomarkers perhaps due to the nutritional burden of such systems in highly streamlined genomes. With the natural environmental undergoing rapid and unprecedented changes, the ability to observe and understand how the oceans are changing on a global scale is possible through the application of environmental biomarkers. Demonstration that RNA and protein do correlate for environmentally responsive genes and the theory behind when and why these correlations occur is valuable in increasing confidence for study of metaproteomic and other meta-‘omic studies to the global oceans [66-70].

Conclusions

In summary, these data elucidate broad, nutrient-limited mRNA-protein dynamics following long-term high CO_2 adaptation under multiple nutrient limitation regimes in a biogeochemically-important organism. Overall, nutrient-limited proteome variation representing 37% of the IMS101 coding potential was congruent with that of its corresponding transcripts, as well as with global transcriptional variation. Pairwise mRNA-protein transcript levels were modestly correlated to protein abundances, yet environmentally responsive transcripts were significantly more correlated than those that did not respond to any of the 8 conditions. Well-characterized stress mRNA and protein biomarkers responded robustly to their respective limiting nutrients following long-term CO_2 adaptation, thereby supporting their future use under

ocean acidification conditions. Other responsive pathways harboring genes with high PC values suggest that they could be good indicators of cellular physiology although certain major pathways (e.g., photosynthesis) contain both environmentally responsive (e.g., light reactions) and unresponsive (e.g., photosystem I) genes. Hence, it is important to consider the functions of different components that make up larger biochemical pathways when investigating energy and/or nutrient flux through the cell. It is as equally important to consider the nature of the environmental conditions being tested. As discussed above, environmentally responsive iron stress genes that form complexes with high-inventory photosystems (e.g. IsiA) in iron limitation can yield high RNA-protein correlations across treatments, which may not be the case if another set of conditions were tested that lacked iron stress. These data not only fill knowledge gaps relating to long-term, nutrient-limited mRNA-protein dynamics, but also to differences and commonalities among molecular metabolisms of interrelated single- and co-limited physiologies under increasing CO_2. A model of transcript and protein production and decay was able to reproduce correlations (in fold change units) observed for environmentally responsive genes with small initial inventories, consistent with the observations, and demonstrating the mechanics of how both transcripts and proteins can be useful in assessing ecosystem and biogeochemical function. Although these data highlight molecular dynamics fueling important biogeochemical processes, future studies examining mRNA-protein correlations over longer temporal periods (e.g., diel to several cell cycles) will help to elucidate transcriptional and corresponding protein synthesis times in the face of environmental change.

Acknowledgements: This work was supported by US National Science Foundation Grants OCE 1851222 and OCE 1657757 and (to D.A.H., E.A.W., and F.-X.F.), OCE1924554 and OCE1850719, G.B Moore Foundation, and NIH R01 GM135709 grants (to M.A.S.).

Conflict of Interest

The authors declare no conflict of interest.

Supplementary Information

Supplementary Figures S1 and S2 show the experimental design and Pearson correlation coefficients, respectively. Supplemental Movie M1 provides a time course of model results transcript and protein inventories and the influence on log2 fold change results. Supplemental Dataset 1 provides the combined transcriptome-proteome dataset as an Excel spreadsheet. Tabs

within the sheet include normalized biological duplicate transcript counts, biological triplicate protein spectral counts, the resulting log2 fold change dataset for transcripts and proteins, and a key for gene annotations.

Availability of Data and Materials

All RNA-Seq data used in this study have been deposited as raw fastq files in NCBI's Gene Expression Omnibus [47] and are accessible through GEO Series accession number GSE94951. All protein spectral count data used in the above analyses can be found in Walworth et al. 2016 [23] in Supplementary Data 4 (DOI: 10.1038/ncomms12081). Combined transcriptome-proteome dataset and gene annotations are provided as a Supplemental Dataset 1. Raw mass spectra datasets used in the above analyses are available at ProteomeXchange under the identifier PXD010515 (DOI: 10.6019/PXD010515). Code for the RNA-Protein model is available on GitHub under M.A.S's repository: https://github.com/maksaito/RNA-Protein_MATLAB_model.

Author contributions

F.-X.F., D.A.H., M.A.S. and E.A.W. designed research; N.G.W., M.D.L., M.A.S., D.M.M., M.R.M. and F.-X.F. performed research; N.G.W., M.D.L., F.-X.F., D.A.H., and E.A.W. analyzed data; M.A.S. developed the RNA-Protein model, and N.G.W., M.A.S., M.D.L., R.M. K., F.-X.F., D.A.H., and E.A.W. wrote the paper.

References

(1) Vogel, C.; Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* **2012**, *13* (4), 227–232.

(2) Newman, J. R. S.; Ghaemmaghami, S.; Ihmels, J.; Breslow, D. K.; Noble, M.; DeRisi, J. L.; Weissman, J. S. Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* **2006**, *441* (7095), 840–846.

(3) Sowell, S. M.; Norbeck, A. D.; Lipton, M. S.; Nicora, C. D.; Callister, S. J.; Smith, R. D.; Barofsky, D. F.; Giovannoni, S. J. Proteomic Analysis of Stationary Phase in the Marine Bacterium "*Candidatus* Pelagibacter ubique." *Appl. Environ. Microbiol.* **2008**, *74* (13), 4091–4100.

(4) Waldbauer, J. R.; Rodrigue, S.; Coleman, M. L.; Chisholm, S. W. Transcriptome and Proteome Dynamics of a Light-Dark Synchronized Bacterial Cell Cycle. *PLoS ONE* **2012**, *7* (8), e43432.

(5) Dyhrman, S. T.; Jenkins, B. D.; Rynearson, T. A.; Saito, M. A.; Mercier, M. L.; Alexander, H.; Whitney, L. P.; Drzewianowski, A.; Bulygin, V. V.; Bertrand, E. M.; et al. The Transcriptome and Proteome of the Diatom *Thalassiosira pseudonana* Reveal a Diverse Phosphorus Stress Response. *PLoS ONE* **2012**, *7* (3), e33768–10.

(6) Wurch, L. L.; Bertrand, E. M.; Saito, M. A.; Van Mooy, B. A. S.; Dyhrman, S. T. Proteome Changes Driven by Phosphorus Deficiency and Recovery in the Brown Tide-Forming Alga *Aureococcus anophagefferens*. *PLoS ONE* **2011**, *6* (12), e28949.

(7) Bender, S. J.; Moran, D. M.; McIlvin, M. R.; Zheng, H.; McCrow, J. P.; Badger, J.; DiTullio, G. R.; Allen, A. E.; Saito, M. A. Colony formation in *Phaeocystis antarctica* : connecting molecular mechanisms with iron biogeochemistry. *Biogeosciences* **2018**, *15* (16), 4923–4942.

(8) Frischkorn, K.R., Haley, S.T. and Dyhrman, S.T. Transcriptional and proteomic choreography under phosphorus deficiency and re-supply in the N_2 fixing cyanobacterium *Trichodesmium erythraeum*. *Frontiers in microbiology* **2019**. *10*, 330.

(9) Capone, D. G.; Zehr, J. P.; Paerl, H. W.; Bergman, B.; Carpenter, E. J. *Trichodesmium*, a Globally Significant Marine Cyanobacterium. *Science* **1997**, *276* (5316), 1221–1229.

(10) Hutchins, D. A.; Fu, F. Microorganisms and ocean global change. *Nature Microbiology* **2017**, *2*, 1–11.

(11) Sohm, J. A.; Webb, E. A.; Capone, D. G. Emerging patterns of marine nitrogen fixation. *Nature Reviews Microbiology* **2011**, 1–10.

(12) Hutchins, D. A.; Mulholland, M. R.; Fu, F.-X. Nutrient cycles and Marine Microbes in a CO_2-Enriched Ocean. *Oceanography* **2009**, *22* (4), 128–145.

(13) Zehr, J. P. Nitrogen fixation by marine cyanobacteria. *Trends in Microbiology* **2011**, *19* (4), 162–173.

(14) Luo, Y. W.; Doney, S. C.; Anderson, L. A.; Benavides, M.; Berman-frank, I.; Bode, A.; Bonnet, S.; Boström, K. H.; Böttjer, D.; Capone, D. G.; et al. Database of diazotrophs in global ocean: abundance, biomass and nitrogen fixation rates. *Earth System Science Data* **2012**, *4* (1), 47–73.

(15) Montoya, J. P.; Holl, C. M.; Zehr, J. P.; Hansen, A. High rates of N_2 fixation by unicellular diazotrophs in the oligotrophic Pacific Ocean. *Nature* **2004**, *430* (7003), 1027–1032.

(16) Karl, D. M.; Church, M. J.; Dore, J. E.; Letelier, R. M.; Mahaffey, C. Predictable and efficient carbon sequestration in the North Pacific Ocean supported by symbiotic nitrogen fixation. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109* (6), 1842–1849.

(17) Bergman, B.; Sandh, G.; Lin, S.; Larsson, J.; Carpenter, E. J. *Trichodesmium*– a widespread marine cyanobacterium with unusual nitrogen fixation properties. *FEMS Microbiol Rev* **2013**, *37* (3), 286–302.

(18) Webb, E. A.; Jakuba, R. W.; Moffett, J. W. Molecular assessment of phosphorus and iron physiology in *Trichodesmium* populations from the western Central and western South Atlantic. *Limnol. Oceanogr.* **2007**, *52* (5), 2221–2232.

(19) Chappell, P. D.; Moffett, J. W.; Hynes, A. M.; Webb, E. A. Molecular evidence of iron limitation and availability in the global diazotroph *Trichodesmium*. *ISME J.* **2012**, *6* (9), 1728–1739.

(20) Venter, J. C.; Remington, K.; Heidelberg, J. F.; Halpern, A. L.; Rusch, D.; Eisen, J. A.; Wu, D.; Paulsen, I.; Nelson, K. E.; Nelson, W.; et al. Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science* **2004**, *304* (5667), 66–74.

(21) Dupont, C. L.; McCrow, J. P.; Valas, R.; Moustafa, A.; Walworth, N.; Goodenough, U.; Roth, R.; Hogle, S. L.; Bai, J.; Johnson, Z. I.; et al. Genomes and gene expression across light and productivity gradients in eastern subtropical Pacific microbial communities. *ISME J* **2014**, *9* (5), 1076–1092.

(22) Moran, M. A. The global ocean microbiome. *Science* **2015**, *350* (6266), aac8455–aac8455.

(23) Walworth, N. G.; Fu, F.-X.; Webb, E. A.; Saito, M. A.; Moran, D.; Mcllvin, M. R.; Lee, M. D.; Hutchins, D. A. Mechanisms of increased *Trichodesmium* fitness under iron and phosphorus co-limitation in the present and future ocean. *Nat. Commun.* **2016**, *7*, 12081.

(24) Garcia, N. S.; Fu, F.; Sedwick, P. N.; Hutchins, D. A. Iron deficiency increases growth and nitrogen-fixation rates of phosphorus-deficient marine cyanobacteria. *ISME J.* **2015**, *9* (1), 238–245.

(25) Walworth, N. G.; Fu, F.-X.; Lee, M. D.; Cai, X.; Saito, M. A.; Webb, E. A.; Hutchins, D. A. Nutrient-Colimited *Trichodesmium* as a Nitrogen Source or Sink in a Future Ocean. *Appl. Environ. Microbiol.* **2018**, *84* (3), e02137–17–14.

(26) Hutchins, D. A.; Boyd, P. W. Marine phytoplankton and the changing ocean iron cycle. *Nat. Clim. Change* **2016**, *6* (12), 1072–1079.

(27) Hutchins, D. A.; Walworth, N. G.; Webb, E. A.; Saito, M. A.; Moran, D.; McIlvin, M. R.; Gale, J.; Fu, F.-X. Irreversibly increased nitrogen fixation in *Trichodesmium* experimentally adapted to elevated carbon dioxide. *Nat. Commun.* **2015**, *6*.

(28) Sañudo-Wilhelmy, S. A.; Kustka, A. B.; Gobler, C. J.; Hutchins, D. A.; Yang, M.; Lwiza, K.; Burns, J.; Capone, D. G.; Raven, J. A.; Carpenter, E. J. Phosphorus limitation of nitrogen fixation by *Trichodesmium* in the central Atlantic Ocean. *Nature* **2001**, *411*, 66–69.

(29) Mills, M. M.; Ridame, C.; Davey, M.; LaRoche, J.; Gelder, R. J. Iron and phosphorus co-limit nitrogen fixation in the eastern tropical North Atlantic. *Nature* **2004**, *429*, 292–294.

(30) Levitan, O.; Sudhaus, S.; LaRoche, J.; Berman-Frank, I. The Influence of pCO2 and Temperature on Gene Expression of Carbon and Nitrogen Pathways in *Trichodesmium* IMS101. *PLoS ONE* **2010**, *5* (12), e15104–e15111.

(31) Walworth, N. G.; Lee, M. D.; Fu, F.-X.; Hutchins, D. A.; Webb, E. A. Molecular and physiological evidence of genetic assimilation to high CO_2 in the marine nitrogen fixer *Trichodesmium*. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, 201605202–201605208.

(32) Oksanen, J.; Blanchet, F. G.; Kindt, R. vegan: Community Ecology Package. *Community ecology …* **2013**.

(33) Suzuki, R.; Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **2006**, *22* (12), 1540–1542.

(34) Fay, M. P.; Shaw, P. A. Exact and Asymptotic Weighted Logrank Tests for Interval Censored Data: The interval R package. *Journal of statistical software* **2010**, *36* (2).

(35) Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **1995**, *57* (1), 289–300.

(36) Selinger, D. W.; Saxena, R. M.; Cheung, K. J.; Church, G. M.; Rosenow, C. Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation. *Genome Research* **2003**, *13* (2), 216–223.

(37) Pratt, J. M.; Petty, J.; Riba-Garcia, I.; Robertson, D. H. L.; Gaskell, S. J.; Oliver, S. G.; Beynon, R. J. Dynamics of Protein Turnover, a Missing Dimension in Proteomics. *Mol Cell Proteomics* **2002**, *1* (8), 579–591.

(38) Moran, M. A.; Satinsky, B.; Gifford, S. M.; Luo, H.; Rivers, A.; Chan, L.-K.; Meng, J.; Durham, B. P.; Shen, C.; Varaljay, V. A.; et al. Sizing up metatranscriptomics. *ISME J.* **2013**, *7* (2), 237–243.

(39) Santoro, A. E.; Dupont, C. L.; Richter, R. A.; Craig, M. T.; Carini, P.; McIlvin, M. R.; Yang, Y.; Orsi, W. D.; Moran, D. M.; Saito, M. A. Genomic and proteomic characterization of "*Candidatus Nitrosopelagicus brevis*": An ammonia-oxidizing archaeon from the open ocean. *Proc Natl Acad Sci USA* **2015**, *112* (4), 1173–1178.

(40) Ramette, A. Multivariate analyses in microbial ecology. *FEMS Microbiology Ecology* **2007**, *62* (2), 142–160.

(41) Rauhut, R.; Klug, G. mRNA degradation in bacteria. *FEMS Microbiol Rev* **1999**, *23* (3), 353–370.

(42) Taniguchi, Y.; Choi, P. J.; Li, G.-W.; Chen, H.; Babu, M.; Hearn, J.; Emili, A.; Xie, X. S. Quantifying *E. coli* Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. *Science* **2010**, *329* (5991), 533–538.

(43) Saito, M. A.; McIlvin, M. R.; Moran, D. M.; Goepfert, T. J.; DiTullio, G. R.; Post, A. F.; Lamborg, C. H. Multiple nutrient stresses at intersecting Pacific Ocean biomes detected by protein biomarkers. *Science* **2014**, *345* (6201), 1173–1177.

(44) Morrissey, J. M.; Bowler, C. Iron utilization in marine cyanobacteria and eukaryotic algae. *Front. Microbiol.* **2012**, *3*, 1–13.

(45) Mohr, W.; Intermaggio, M. P.; LaRoche, J. Diel rhythm of nitrogen and carbon metabolism in the unicellular, diazotrophic cyanobacterium *Crocosphaera watsonii* WH8501. *Environ. Microbiol.* **2010**, *12* (2), 412–421.

(46) Saito, M. A.; Bertrand, E. M.; Dutkiewicz, S.; Bulygin, V. V.; Moran, D. M.; Monteiro, F. M.; Follows, M. J.; Valois, F. W.; Waterbury, J. B. Iron conservation

by reduction of metalloenzyme inventories in the marine diazotroph *Crocosphaera watsonii*. *Proc Natl Acad Sci USA* **2011**, *108* (6), 2184–2189.

(47) Edgar, R.; Domrachev, M.; Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **2002**, *30* (1), 207–210.

(48) Lindell, D, Padan, E and Post, AF. Regulation of ntcA expression and nitrite uptake in the marine *Synechococcus sp*. strain WH 7803. *J. Bacteriol*, **1998**. *180*(7), 1878-1886.

(49) Smith, D.P., Thrash, J.C., Nicora, C.D., Lipton, M.S., Burnum-Johnson, K.E., Carini, P., Smith, R.D. and Giovannoni, S.J. Proteomic and transcriptomic analyses of "*Candidatus Pelagibacter ubique*" describe the first PII-independent response to nitrogen limitation in a free-living Alphaproteobacterium. *MBio* **2013,** *4*(6), e00133-12.

(50) Bender, S.J., Durkin, C.A., Berthiaume, C.T., Morales, R.L. and Armbrust, E.. Transcriptional responses of three model diatoms to nitrate limitation of growth. *Frontiers in Marine Science* **2014**, *1*, 3.

(51) Orchard, E.D., Webb, E.A. and Dyhrman, S.T. Molecular analysis of the phosphorus starvation response in *Trichodesmium spp*. *Environ Microbiol* **2009**, *11*(9), 2400-2411.

(52) Scanlan, D.J., Silman, N.J., Donald, K.M., Wilson, W.H., Carr, N.G., Joint, I. and Mann, N.H. An immunological approach to detect phosphate stress in populations and single cells of photosynthetic picoplankton. *Applied Environ Microbiol* **1997**, *63*(6), 2411-2420.

(53) Pereira, N., Shilova, I.N. and Zehr, J.P. Use of the high-affinity phosphate transporter gene, pstS, as an indicator for phosphorus stress in the marine diazotroph *Crocosphaera watsonii* (Chroococcales, Cyanobacteria). *J Phycol* **2019**, *55*(4), 752-761.

(54) Morrissey, J., Sutak, R., Paz-Yepes, J., Tanaka, A., Moustafa, A., Veluchamy, A., Thomas, Y., Botebol, H., Bouget, F.Y., McQuaid, J.B. and Tirichine, L. A novel protein, ubiquitous in marine phytoplankton, concentrates iron at the cell surface and facilitates uptake. *Current Biology* **2015**, *25*(3), 364-371.

(55) McQuaid, J.B., Kustka, A.B., Oborník, M., Horák, A., McCrow, J.P., Karas, B.J., Zheng, H., Kindeberg, T., Andersson, A.J., Barbeau, K.A. and Allen, A.E. Carbonate-sensitive phytotransferrin controls high-affinity iron uptake in diatoms. *Nature* **2018**, *555*(7697), 534-537.

(56) Behnke, J. and LaRoche, J. Iron uptake proteins in algae and the role of Iron Starvation-Induced Proteins (ISIPs). *Eur J Phycol* **2020**, *55*(3), 339-360.

(57) Roche, J.L., Murray, H., Orellana, M. and Newton, J. Flavodoxin Expression as an Indicator of Iron Limitation in Marine Diatoms1. *J Phycology* **1995**, *31*(4), 520-530.

(58) Chappell, P.D. and Webb, E.A. A molecular assessment of the iron stress response in the two phylogenetic clades of *Trichodesmium*. *Environ Microbiol* **2010**, *12*(1), 13-27.

(59) Held, N.A., Webb, E.A., McIlvin, M.M., Hutchins, D.A., Cohen, N.R., Moran, D.M., Kunde, K., Lohan, M.C., Mahaffey, C., Woodward, E.M.S. and Saito, M.A. Co-occurrence of Fe and P stress in natural populations of the marine diazotroph *Trichodesmium*. *Biogeosciences* **2020**, *17*(9), 2537-2551.

(60) Chappell, P.D., Moffett, J.W., Hynes, A.M. and Webb, E.A. Molecular evidence of iron limitation and availability in the global diazotroph *Trichodesmium*. *ISME J* **2012**, *6*(9), 1728-1739.

(61) Smith, D.P., Kitner, J.B., Norbeck, A.D., Clauss, T.R., Lipton, M.S., Schwalbach, M.S., Steindler, L., Nicora, C.D., Smith, R.D. and Giovannoni, S.J. Transcriptional and translational regulatory responses to iron limitation in the globally distributed marine bacterium *Candidatus Pelagibacter ubique*. *PLoS One* **2010**, *5*(5), 10487.

(62) Bertrand, E.M., Allen, A.E., Dupont, C.L., Norden-Krichmar, T.M., Bai, J., Valas, R.E. and Saito, M.A. Influence of cobalamin scarcity on diatom molecular physiology and identification of a cobalamin acquisition protein. *Proc Natl Acad Sci USA* **2012**, *109*(26), E1762-E1771.

(63) Bertrand, E.M., Moran, D.M., McIlvin, M.R., Hoffman, J.M., Allen, A.E. and Saito, M.A. Methionine synthase interreplacement in diatom cultures and communities: Implications for the persistence of B_{12} use by eukaryotic phytoplankton. *Limnol Oceanogr* **2013**, *58*(4), 1431-1450.

(64) Kellogg R.M., M.A. Moosburner, N.R. Cohen, N.J. Hawco, M.R. McIlvin, D.M. Moran, A.E. Allen, M.A. Saito. Two novel zinc and cobalt - responsive proteins in marine eukaryotic phytoplankton and implications for Zn scarcity in the surface oceans. Submitted.

(65) Hawco, N.J., McIlvin, M.M., Bundy, R.M., Tagliabue, A., Goepfert, T.J., Moran, D.M., Valentin-Alvarado, L., DiTullio, G.R. and Saito, M.A. Minimal cobalt metabolism in the marine cyanobacterium *Prochlorococcus*. *Proc Natl Acad Sci USA* **2020**, *117*(27), 15740-15747.

(66) Saito, M.A., Dorsk, A., Post, A.F., McIlvin, M.R., Rappé, M.S., DiTullio, G.R. and Moran, D.M. Needles in the blue sea: Sub-species specificity in targeted protein biomarker analyses within the vast oceanic microbial metaproteome. *Proteomics* **2015**, *15*(20), 3521-3531.

(67) Saito, M.A., Bertrand, E.M., Duffy, M.E., Gaylord, D.A., Held, N.A., Hervey IV, W.J., Hettich, R.L., Jagtap, P.D., Janech, M.G., Kinkade, D.B. and Leary, D.H. Progress and challenges in ocean metaproteomics and proposed best practices for data sharing. *J Proteome Res* **2019**, *18*(4), 1461-1476.

(68) Saito, M.A., Saunders, J.K., Chagnon, M., Gaylord, D.A., Shepherd, A., Held, N.A., Dupont, C., Symmonds, N., York, A., Charron, M. and Kinkade, D.B. Development of an Ocean Protein Portal for Interactive Discovery and Education. *J Proteome Res* **2021**. 20, 1, 326-336.

(69) Villar, E., Vannier, T., Vernette, C., Lescot, M., Cuenca, M., Alexandre, A., Bachelerie, P., Rosnet, T., Pelletier, E., Sunagawa, S. and Hingamp, P. The Ocean Gene Atlas: exploring the biogeography of plankton genes online. *Nucleic Acids Research* **2018**, *46*(W1), W289-W295.

(70) Ustick, L.J., Larkin, A.A., Garcia, C.A., Garcia, N.S., Brock, M.L., Lee, J.A., Wiseman, N.A., Moore, J.K. and Martiny, A.C. Metagenomic analysis reveals global-scale patterns of ocean nutrient limitation. *Science* **2021**, *372*(6539), 287-291.

Table 1. Parameters in transcriptome and proteome expression model

Constant	Description	Value	Reference
k_{RNA}	Production of RNA	100	
k_{-RNA}	Degradation of RNA	0.102 min^{-1} (6.8 min half-life)	1 (*E. coli*)
d_{RNA}	Delay between signal and RNA	1 min	
$k_{protein}$	Production of protein	3.5	3
$k_{-protein}$	Degradation of protein	0.0017 min^{-1} (6.9 h half-life)	2 (Yeast)
$d_{protein}$	Delay between RNA and protein	180 min	
s_0	Basal signal level 1%	0.01	
s_1	Increased signal due to environmental cue	0.2	
s_{0alt}	Basal signal 20%	0.2	
s_{1alt}	Increased signal due to environmental cue (+0.2)	0.4	
t_{start}, $t_{plateau1}$, $t_{plateau2}$, t_{end}	Time of onset of signal (t_{start}), begin plateau ($t_{plateau1}$), end plateau ($t_{plateau2}$), end of decrease in signal $t_{end;}$ ("long slow ramp")	600, 650, 700, 1200 min	
t_0 t_0 by day	Time point denominator for fold change	9 h (540 min) 540, 1980, 3420 min	

[1] Selinger et al., 2003
[2] Pratt et al., 2002; 6.9 – 34 h protein half lives reported
[3] Tuned using 1 minute signal to be 6-7 proteins per transcript based on Moran et al., 2013

Table 2. Example Biomarkers for Nutrient Stress in Marine Microbes

Protein Name	Nutrient	Relevant Taxa	Example Refs
Urea Transporter	Nitrogen	Cyanobacteria	43, 48-50
NtcA and P-II N Regulatory Proteins	Nitrogen	Cyanobacteria	43, 48, 66
NtrX Regulatory Protein	Nitrogen	Pelagibacter	49
Ammonia transporter	Nitrogen	Cyanobacteria, Diatoms	43, 50
Alkaline Phosphatase, PhoA	Phosphorus	Cyanobacteria, Diatoms	5, 6, 51
Phosphate Transporter, PstS	Phosphorus	Cyanobacteria	51,52,53
Iron Transporter IdiA/FutA	Iron	Cyanobacteria	18, 19, 58, 59, 60
Flavodoxin IsiB and Ferredoxin	Iron	Cyanobacteria, Eukaryotic Algae	57, 43, 7
Iron Transporter SfuC	Iron	*Pelagibacter*	61
ISIP2	Iron	Eukaryotic Algae	7, 54, 55
ISIP3	Iron	Eukaryotic Algae	7, 54
B_{12} transporter CBA1	Vitamin B_{12}	Eukaryotic Algae	62-63
Methioine Synthase MetE and MetH	Vitamin B_{12}	Eukaryotic Algae	62-63
Zn chaperone ZCRP-A	Zinc, Cobalt	Eukaryotic Algae	64
Zn membrane protein ZCRP-B	Zinc, Cobalt	Eukaryotic Algae	64
Unknown	Co, Ni, Mn, vitamins, etc.		65

Figure Legends

Figure 1. Overall comparison of proteome-transcriptome data in log2 fold space for eight experimental treatments, where treatments varied CO_2, iron and phosphorus (see Fig. S1 for full experimental design) in cultures of the marine cyanobacterium *Trichodesmium erythraeum* IMS101. Legend shows symbols of seven treatments relative to the 380ppm CO_2 replete treatment. Transcript data is the average of biological duplicates, protein data is the average of biological triplicates.

Figure 2. Global molecular analysis. (a) The top panel shows a redundancy (RDA) analysis of normalized transcript abundances of genes whose protein products were also detected in Walworth et al. (2016a). The middle panel shows a Procrustes analysis between corresponding-RNA-RDA and prot-RDA ordinations indicating close agreement evidenced by small vector residuals ($r = 0.75$; $P < 0.001$). The bottom panel displays the same redundancy analysis as in the top panel except now including all detected gene transcripts. (b) Hierarchical clustering of Bray-Curtis dissimilarities with multiscale bootstrap resampling calculated from normalized transcripts of differentially expressed (top panel), upregulated (bottom left), and downregulated (bottom right) genes. Numbers at dendrogram nodes are approximately unbiased p-values calculated from multiscale bootstrap resampling.

Figure 3. Scatterplots of log2 fold changes of transcript and protein abundances for select, highly correlated genes. (a) Shown are log2 fold change of normalized protein (y-axis) and transcript (x-axis) abundances for Fe-stress genes (b) Shown are log2 fold change of normalized protein (y-axis) and transcript (x-axis) abundances for P-stress gene. Ellipses denote different subunits of the Pst transport system. (c) Shown are log2 fold change of normalized protein (y-axis) and transcript (x-axis) abundances for photosynthesis light reaction genes. Different symbols denote different treatments.

Figure 4. Scatterplots of log2 fold changes of transcript and protein abundances for select core metabolism genes. Shown are log2 fold change of normalized protein (y-axis) and transcript (x-axis) abundances for TCA cycle, amino acid binding, and IMP biosynthesis genes. Different symbols denote different treatments.

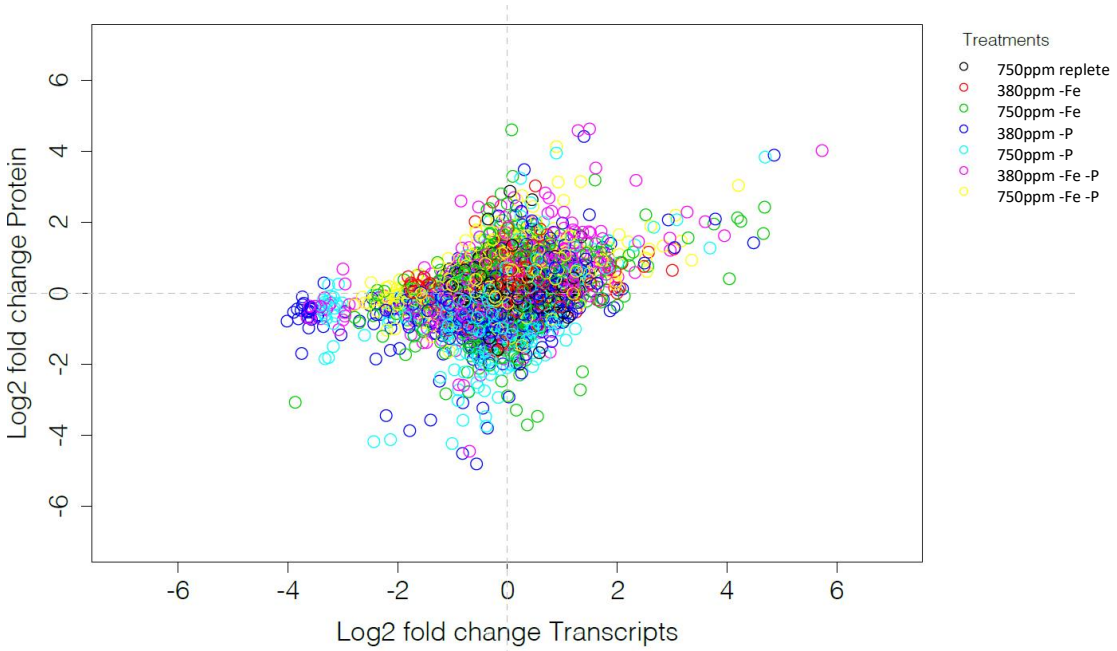
Figure 5. Representation of transcriptome-proteome model. Parameters described in Table 1.

Figure 6. RNA-Protein expression model results. Low basal expression of 1% (left panels) and moderate basal expression (right panels), with a simulated daily expression pattern. Short and long half-lives of RNA and protein, respectively, result different inventories over time. When basal expression is increased to 20% protein inventories begin to accumulate with between successive daily cycles.

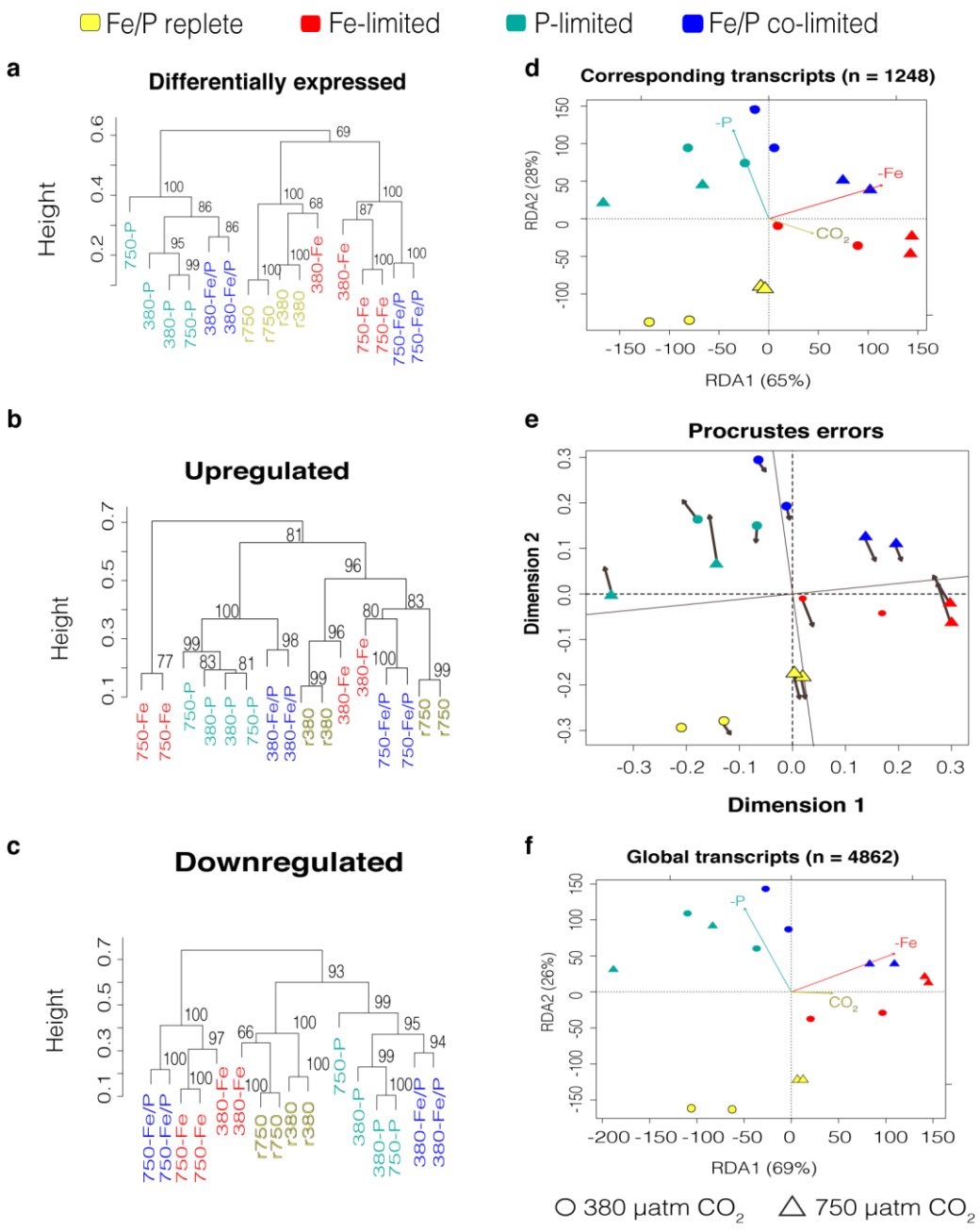
Figure 7. RNA and protein expression in fold change space. Model output from Figure 6 converted to fold-change space results in high values with low basal expression (1%; left panels) and muted values with some basal expression (20%; right panels). Fold-change was calculated relative to day 0 or each successive day in right panels.

Figure 8. Expression patterns in transcriptome-proteome space. The succession of RNA and protein biomolecules resulted in a counterclockwise path, with more pronounced fold change signals with low basal expression (left panels) compared to higher basal expression (center panels) where the signal circular progression falls close to the origin (top middle). Coherence between biomolecules was observed as in the northeast or southwest quadrants (dotted line). The resulting maximum fold change (right panels) observed was high with low basal expression (top right) and when relative to day 1, rather than successive days as the protein inventory accumulates (Figure 6). Environmental biomarkers tend to have lower basal expression and are expressed based on regulation by environmental cues.

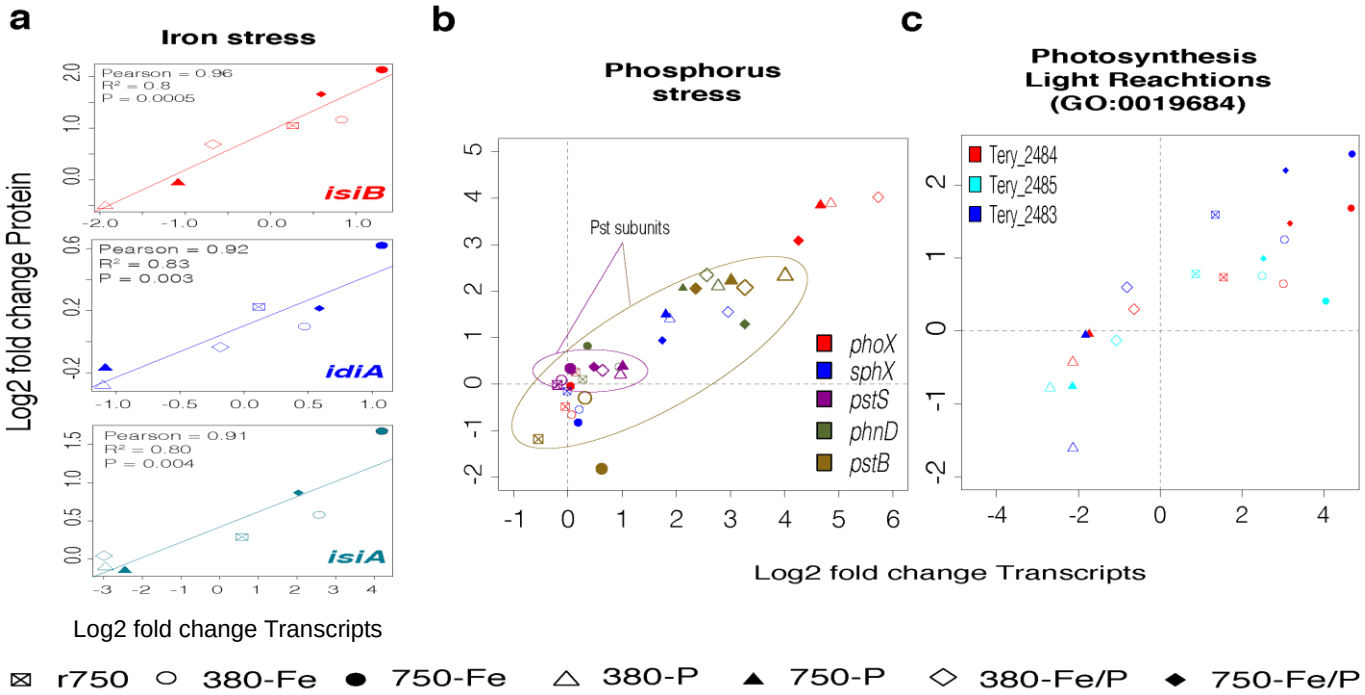
Figure 1


Treatments
750ppm replete
380ppm -Fe
750ppm -Fe
380ppm -P
750ppm -P
380ppm -Fe -P
750ppm -Fe -P
Log2 fold change Protein
Log2 fold change Transcripts
-6
-4
-2
0
2
4
6

824 Figure 2


Fe/P replete
Fe-limited
P-limited
Fe/P co-limited
a
Differentially expressed
Height
b
Upregulated
Height
c
Downregulated
Height
d
Corresponding transcripts (n = 1248)
RDA2 (28%)
RDA1 (65%)
-P
-Fe
CO2
e
Procrustes errors
Dimension 2
Dimension 1
f
Global transcripts (n = 4862)
RDA2 (26%)
RDA1 (69%)
-P
-Fe
CO2
380 µatm CO2
750 µatm CO2

825 Figure 3


a
Iron stress
Pearson = 0.96
R² = 0.8
P = 0.0005
isiB
Pearson = 0.92
R² = 0.83
P = 0.003
idiA
Pearson = 0.91
R² = 0.80
P = 0.004
isiA
Log2 fold change Protein
Log2 fold change Transcripts
b
Phosphorus stress
Pst subunits
phoX
sphX
pstS
phnD
pstB
c
Photosynthesis Light Reachtions (GO:0019684)
Tery_2484
Tery_2485
Tery_2483
Log2 fold change Transcripts
r750
380-Fe
750-Fe
380-P
750-P
380-Fe/P
750-Fe/P

826 Figure 4.

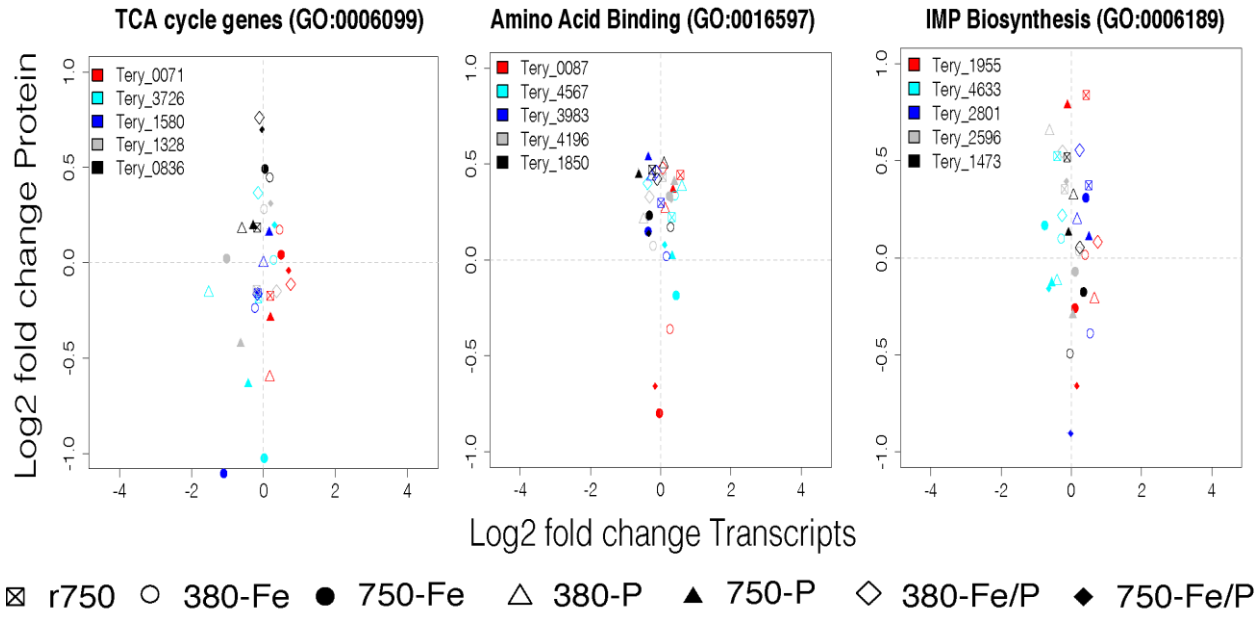

TCA cycle genes (GO:0006099)
Amino Acid Binding (GO:0016597)
IMP Biosynthesis (GO:0006189)
Log2 fold change Protein
Log2 fold change Transcripts
Tery_0071
Tery_3726
Tery_1580
Tery_1328
Tery_0836
Tery_0087
Tery_4567
Tery_3983
Tery_4196
Tery_1850
Tery_1955
Tery_4633
Tery_2801
Tery_2596
Tery_1473
r750
380-Fe
750-Fe
380-P
750-P
380-Fe/P
750-Fe/P

Figure 5.

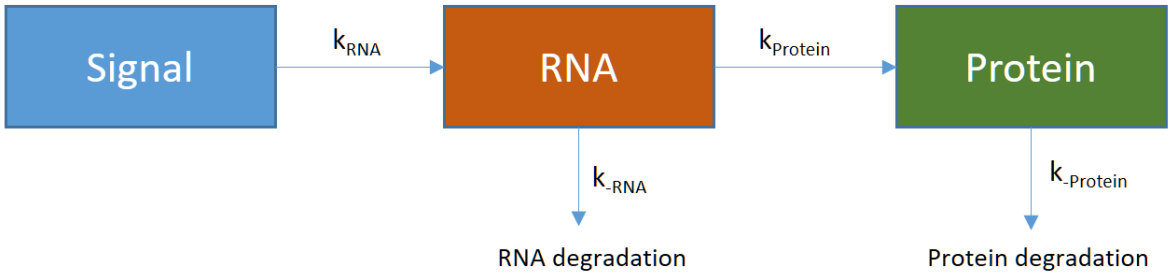

Signal
k_{RNA}
RNA
$k_{Protein}$
Protein
k_{-RNA}
RNA degradation
$k_{-Protein}$
Protein degradation

Figure 6.

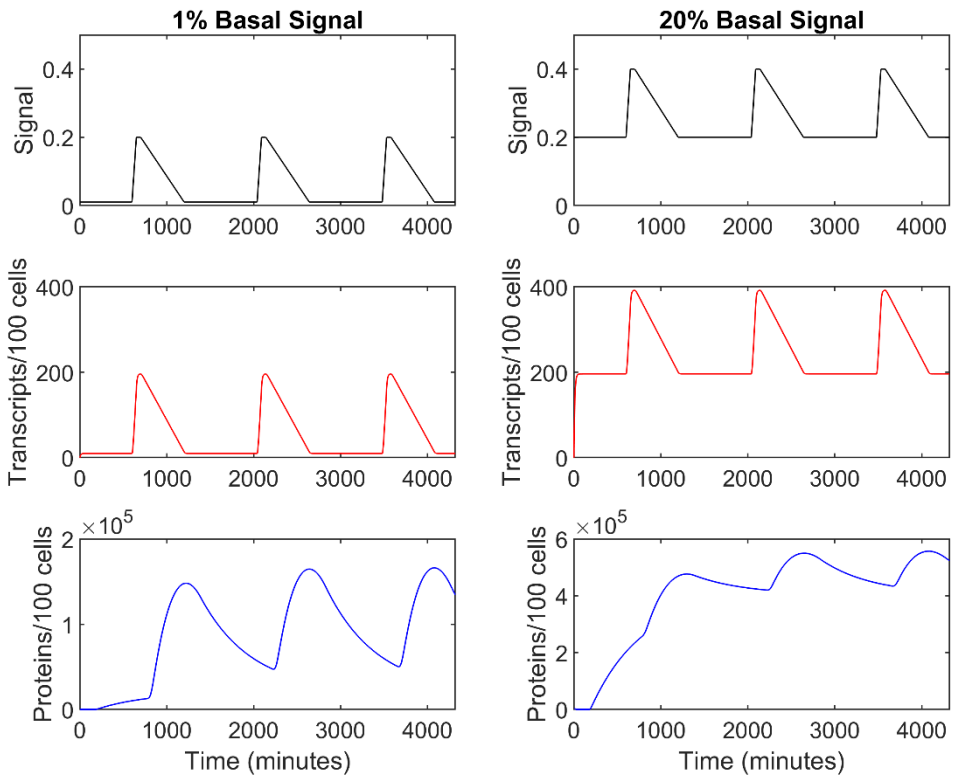

1% Basal Signal
20% Basal Signal
Signal
Transcripts/100 cells
Proteins/100 cells
×10^5
Time (minutes)
0
1000
2000
3000
4000
0.4
0.2
400
200
2
1
6
4

Figure 7.

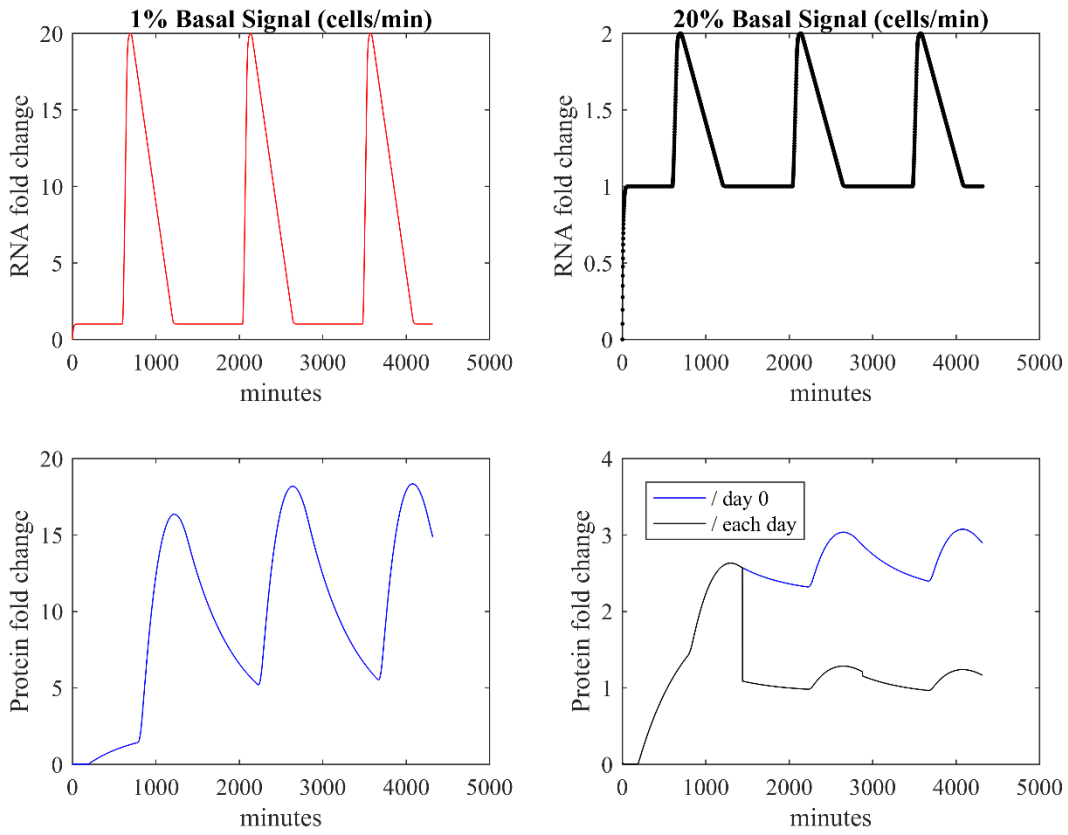
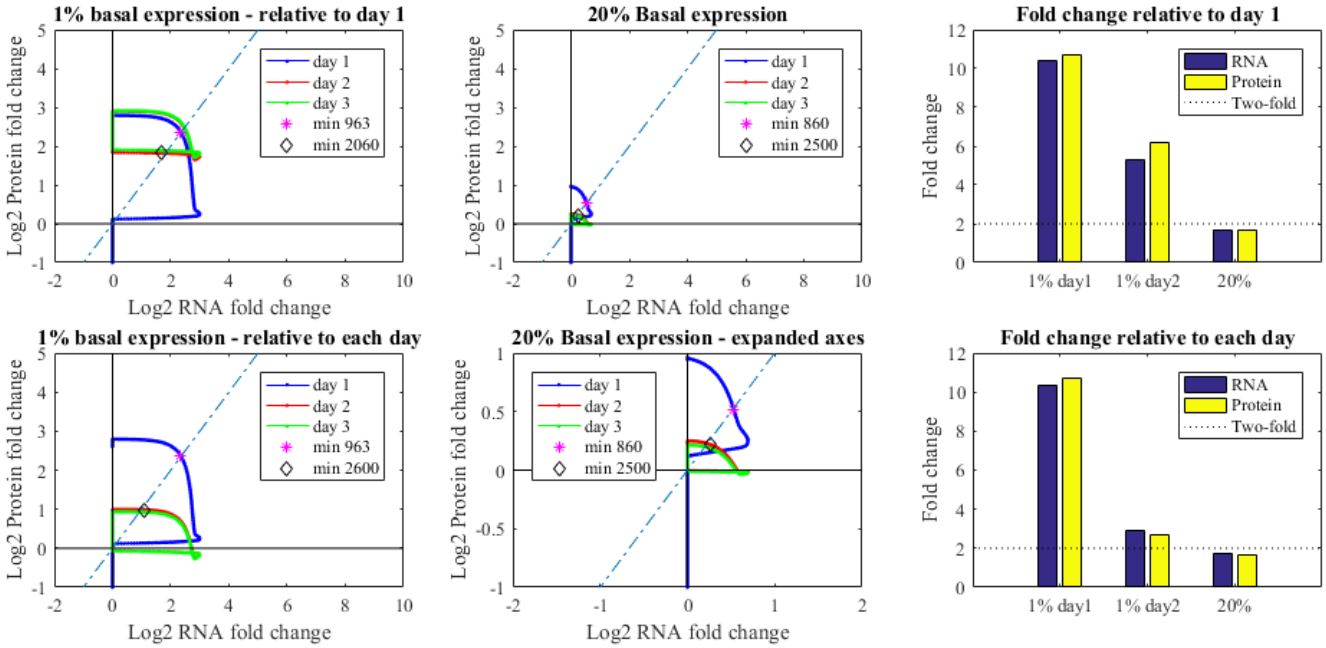

1% Basal Signal (cells/min)
20% Basal Signal (cells/min)
RNA fold change
Protein fold change
minutes
/ day 0
/ each day

Figure 8.


1% basal expression - relative to day 1
Log2 Protein fold change
Log2 RNA fold change
day 1
day 2
day 3
min 963
min 2060
20% Basal expression
min 860
min 2500
Fold change relative to day 1
Fold change
RNA
Protein
Two-fold
1% day1
1% day2
20%
1% basal expression - relative to each day
min 2600
20% Basal expression - expanded axes
Fold change relative to each day