Full Title: Transcriptome-guided parsimonious flux analysis improves predictions with metabolic networks in complex environments

Short Title: Transcriptome-guided parsimonious flux analysis

Authors:

Matthew L. Jenior[1]

Thomas J. Moutinho Jr.[1]

Jason A. Papin[1,2,3]

Affiliations:

1. Department of Biomedical Engineering, University of Virginia, Charlottesville, VA, USA
2. Department of Medicine, Division of Infectious Diseases & International Health, University of Virginia, Charlottesville, VA, USA
3. Department of Biochemistry & Molecular Genetics, University of Virginia, Charlottesville, VA, USA

Abstract

The metabolic responses of bacteria to dynamic extracellular conditions drives not only the behavior of single species, but also entire communities of microbes. Over the last decade, genome-scale metabolic network reconstructions have assisted in our appreciation of important metabolic determinants of bacterial physiology. These network models have been a powerful force in understanding the metabolic capacity that species may utilize in order to succeed in an environment. However, previous approaches to use these platforms in concert with omics data to better characterize experimental systems have met challenges due to assumptions necessary by the various integration platforms or due to large input data requirements. With these challenges in mind, we developed RIPTiDe (Reaction Inclusion by Parsimony and Transcript Distribution) which uses both parsimony of overall flux and transcriptomic abundances to identify the most cost-effective usage of metabolism that also best reflects the cell's investments into transcription. This approach is a shift away from previous

omic integration algorithms which sought to strictly maximize concordance in reaction usage with the given transcriptome. Utilizing a metabolic network reconstruction for the model organism *Escherichia coli* str. K-12 substr. MG1655, we found that RIPTiDe correctly identifies context-specific metabolic pathway activity without supervision. We also assessed the application of RIPTiDe to *in vivo* metatranscriptomic data where *E. coli* was present at high abundances, and found that our approach also correctly predicts metabolic behaviors of host-associated bacteria with high accuracy. In the context of human health, metabolic changes in bacteria have downstream impacts at multiple levels including the expression of virulence factors from pathogens or inducing disease-associated dysbiosis across the microbiota. Toward this point, RIPTiDe has far reaching potential for understanding context-specific bacterial metabolism within complex communities.

Author Summary

Transcriptomic analyses of bacteria have become instrumental to our understanding of their responses to changes in their environment. While traditional analyses have been informative, leveraging these datasets within genome-scale metabolic network reconstructions (GENREs) can provide greatly improved context for shifts in pathway utilization and downstream/upstream ramifications for changes in metabolic regulation. Previous techniques for GENRE transcript integration have focused on creating maximum consensus with the input datasets. However, these approaches have collectively performed poorly for metabolic predictions even compared to transcript-agnostic approaches of flux minimization (pFBA) that identifies the most efficient patterns of metabolism given certain growth constraints. Our new method, RIPTiDe, combines these concepts and utilizes overall minimization of flux in conjunction with transcriptomic analysis to identify the most energy efficient pathways to achieve growth that include more highly transcribed enzymes. Utilizing a well-studied GENRE from *Escherichia coli*, we demonstrate that this new approach correctly predicts patterns of metabolism under a variety of both *in vitro* and *in vivo* growth conditions. Furthermore, RIPTiDe requires a low level of manual intervention which leads to reduced bias in predictions. This platform

could be important for revealing context-specific bacterial phenotypes in line with governing principles of adaptive evolution, that drive disease manifestation or interactions between microbes.

Introduction

Metabolic plasticity is critical for bacterial survival in changing environments, and fundamental to stable microbial communities [1,2]. In the context of human health, it has been shown that certain pathogens adapt their metabolism to their current environment to most effectively colonize a new host [3,4]. By fully understanding these shifts and their implications it may provide opportunities for novel therapeutic strategies. However, describing these situational differences in overall metabolism of an organism have remained difficult to quantify. While genomic analyses of these species reveal catalogues of metabolic capabilities for a species, they do not provide information toward the inter-connected networks of metabolism that invariably occur within the cell. One approach that addresses these limitations is computationally combining genomic data with known biochemical constraints for predicting/modeling cellular physiology with Genome-Scale metabolic Network Reconstructions (GENREs). A GENRE is composed of the collection of genes and metabolic reactions associated with the species of interest, representing the totality of known metabolic function that organism is able to employ. This functionality can then be formalized with a mathematical framework and constrained by known biological parameters to allow for simulation of metabolic processes. These powerful discovery platforms have allowed for guided genetic engineering efforts, directed hypothesis generation for downstream laboratory testing, and deconvoluted metabolic responses of bacteria to antibiotic stress [5,6].

GENREs have also become strong platforms for providing context to multiple forms of high-throughput omics data, allowing for the study of whole-network changes in metabolism due to upstream/downstream shifts in enzyme activity [7]. To this end, advances over the last decade in RNA-Seq technologies have indeed enabled whole bacterial community metatranscriptomic analysis [7]. Conversely, changes identified through other means occasionally can be difficult to attribute to a single organism from complex biological samples (i.e. untargeted metabolomics). However, shotgun

transcriptomic analysis followed by strict read mapping protocols allows for the characterization of individual groups of species within these communities of organisms [8,9]. These large datasets provide informative possibilities for analyses; however holistic (top down) analyses are limited in their ability to make any mechanistic claims outside of correlative results. Against the backdrop of the reductionist (bottom up) approach of GENREs, their combined informative capacity allows for mechanistic insight into the metabolic status and ecological interactions within the microbiome [10].

Previous approaches at integrating transcriptomic data into GENREs have primarily focused on identifying reactions associated with arbitrary cutoffs of gene transcription, then maximizing flux through those reactions with many transcripts or minimizing flux through those reactions with few transcripts [11–16]. Recently, parsimonious Flux Balance Analysis (pFBA) was developed, which seeks to minimize the flux associated with each reaction in the model, while maintaining optimum flux through the objective [17]. In essence, this approach identifies the least biologically “expensive” usage of an organism’s metabolism to achieve high growth rates. This method is in line with the concept that evolutionary pressures have selected for metabolic states in microbes with minimized cellular cost that maximize growth rate under a variety of environmental conditions [18]. Surprisingly, pFBA alone was found to be able to largely outperform other methods in predicting metabolic network behavior under various growth conditions despite not utilizing any transcriptomic data for these predictions [19]. These studies demonstrate that contextualizing data in a truly biological meaningful way has proven difficult, highlighting large gaps in our understanding of metabolic regulatory mechanisms. Moreover, recent studies on *E. coli* grown in defined *in vitro* conditions have shown that levels of transcript were only predictive for roughly half of the concentrations of the corresponding enzymes in the cell [20,21]. Despite this shortcoming, transcript abundances remain an indication of cellular investment into a given metabolic strategy. Furthermore, as the expression of a functional enzyme is not as simple as an on-or-off scenario, we have also implemented an unsupervised procedure to assign weights to reactions based on the distribution inherent to RNA-Seq data [22]. As arbitrary thresholds have been shown to be biased [23], we chose calculate continuous values for each gene along the transcript

abundance distribution. This approach is unbiased by depth of sequencing across datasets, and it restricts the utility of low-transcription reactions to an optimal pFBA solution yet does not entirely prohibit it. With both of these concepts in mind, we have leveraged the insights gained by pFBA, integrating transcript abundance as a weighting metric to direct the most parsimonious flux solutions toward a state of higher fidelity to *in situ* biology.

We call this combined method RIPTiDe, or Reaction Inclusion by Parsimony and Transcript Distribution. RIPTiDe was then tested on the most well curated GENRE of the highly studied model organism *E. coli* str. K-12 substr. MG1655. Transcriptomic data from multiple distinct *in vitro* growth conditions were compared against each other as well as the base implementation of pFBA. RIPTiDe was able to accurately discern context-specific phenotypes including growth rates that closely match experimentally measured values under the same conditions, as well as gene essentiality predictions in metabolic pathways relevant to the corresponding media. The platform was subsequently tested utilizing *in vivo* metatranscriptomic sequence data from clindamycin-treated mouse cecal content where *E. coli* is the dominant member of the bacterial community [8]. When contrasted against results from the previous analysis, the *in vivo* RIPTiDe-contextualized model grew significantly slower than in aerobic rich media and more central metabolites were produced intracellularly when in competition with other bacterial species. These combined analyses validated our methodology by correctly uncovering context-specific biological trends, and previously unappreciated host-associated behaviors of *E. coli*. Furthermore, this technique could present a powerful tool for the study of the microbiota through the simultaneous unsupervised contextualization of omics data from multiple organisms within the same community.

Results and Discussion

Reaction Inclusion by Parsimony and Transcript Distribution (RIPTiDe)

RIPTiDe relies on bi-level optimization guided by transcript abundances in order to identify the most efficient usage of metabolism that also reflects an organism's transcriptional state. The required input for RIPTiDe includes a metabolic network reconstruction with an assigned objective and

gene-associated transcript abundances. A list of known cellular tasks that should be included/excluded may also be provided if *a priori* knowledge exists about the organism's behavior in the environment of interest. Transcript mapping is reduced to only genes that appear in the genome-scale network reconstruction of interest, therefore not skewing the distribution with data that is uninformative to the current analysis.

To generate reactions coefficients that closely reflect the normalized transcript abundance [24] for their associated genes, we calculated the ratio of transcription for each gene with the maximum abundance in the dataset (Fig. 1A) [22]. This process results in a distribution of linear coefficients that is >0.0 and ≦ 1.0 and as a result, higher transcript abundance corresponds to a higher probability of being utilized in the final metabolic model. Additionally, in the cases where multiple genes mediate the same reaction (i.e. multimers), then the lowest coefficient (highest transcript abundance) among the group is chosen as it reflects in what processes the cell has invested energy. Other previous platforms [14,16] have sought to incorporate parsimonious flux analysis into GENRE omics integration, however these methods did not include any steps to explore distinct forms of metabolism that are still able to occur given the distinct constraints that each algorithm applies to models. Altogether, our approach removes the need for different arbitrary cutoffs between datasets and reduces the requirement for manual intervention in the generation of functional context-specific models.

As with pFBA, reversible reactions are split into pairs of irreversible reactions carrying non-negative flux. In this way, we were able to minimize overall flux subject to the transcript-guided linear coefficients added by RIPTiDe. Within the transcript abundance distribution, progressively smaller linear coefficients are assigned to genes with increased transcription at each cutoff (Fig. 1A). Coefficients are then subsequently transferred to the corresponding metabolic reactions for each gene. The magnitude of linear coefficient drives the amount of flux carried by each reaction. Larger coefficients correspond to the reaction receiving larger flux during optimal growth, which would therefore be favored less when minimizing the overall flux through the objective function. Reactions without genes are assigned the median score to not be incorrectly penalized, nor overly biased to their

inclusion. Importantly, this approach is equally tractable between different sequencing depths as well as both transcriptomes sequenced from pure culture or extracted from a metatranscriptome.

The scope of feasible solution states in a GENRE typically out-scales the number of realistic metabolic states of the cell by a large magnitude [25]. Placing further constraints on a model that reflects likely metabolism allows for improved predictions, but only partially mitigates the potential for biologically meaningless solutions to be reported. Sampling of feasible flux distributions reveals how the applied constraints limit the possible metabolic states and markedly reduce solution space [26,27]. Integrating a flux sampling technique into RIPTiDe, we were able to further leverage our transcript-based constraints to maximize the context-specificity of resultant models. Following assignment of linear coefficients to metabolic reactions, we proceeded through the linear optimization steps at the core of RIPTiDe as illustrated in Fig. 1B. In order to be quickly and uniformly sample this highly constrained solution volume, we utilized the sampler gapsplit which is tuned to maximize efficiency per sample under this type of condition [28]. Transcript abundances were used to guide the bi-level optimization and flux sampling in order to identify the most economic, likely usage of metabolism with the given cellular investment in transcription. A more detailed description of the steps in RIPTiDe are as follows:

1. Near optimal flux through the objective function is set as a constraint (≧80%, Table S1).
2. Linear coefficients are assigned as constraints to each reaction based on transcriptomic abundance of the associated gene-enzyme assignments. In cases where multiple gene products contribute to the same reaction, the largest weight among them is selected.
3. Total sum of fluxes in the model is minimized subject to previously outlined constraints.
4. Standard flux balance analysis then identifies inactivated subnetworks (zero flux) which are subsequently removed from the model. Orphaned genes and metabolites are then pruned.
5. Inverse linear coefficients within the original provided range are calculated and assigned for each reaction, then flux sampling is performed on the constrained and reduced model to find most likely pathway utilization patterns.

The resultant pruned model has a dramatically more focused, biologically feasible, metabolic solution space that has been adequately sampled for feasible metabolic states (Fig. 1C). Unlike most many previous approaches, models produced by RIPTiDe are able to achieve flux through their given objectives by default without any reintroduction of deactivated reactions. This is critical as the principle of optimum cellular economy with respect to the transcript data is never broken throughout the contextualization process. Furthermore, we have also integrated functionality to instead utilize one or more metabolic tasks as constraints which must achieve positive flux in order to accommodate larger-scale models where a single defined cellular objective may be a biologically invalid assumption (i.e. human tissue). For a more detailed description of the algorithm, refer to the Methods section.

Application of RIPTiDe to an Example Model of Bacterial Metabolism

To test RIPTiDe, we created a small, flux-consistent metabolic network reconstruction with two compartments and generation of ATP (in arbitrary units) as the cellular objective function. The model has two separate means of accomplishing this goal; either through glycolysis where glucose is the primary substrate, or Stickland fermentation which requires pairs of amino acids in coupled redox reactions to create energy [29]. In nature, fewer ATP are generated per unit of substrate in Stickland fermentation compared to glycolysis [30]. This phenomenon was replicated in the example model as fewer reactions are required to generate ATP from glucose, making this pathway more energetically favorable (Table S2). Utilizing the base implementation of pFBA available within COBRApy to quantify which pathway was the more parsimonious solution, we ensured that glycolysis was actually the most efficient (fewest enzymes used) for optimal ATP generation (Fig. 2A). Critically, this ATP generation was performed under the conditions where all possible extracellular substrates were provided in excess to the model as to not bias results only on the basis of resource availability.

As a means to determine the context specificity of RIPTiDe output models, we then generated two distinct simulated transcript abundance distributions that each reflected elevated transcription of a single pathway (Table S2). In Fig. 2B, RIPTiDe was provided with a transcriptome in which genes for glycolysis and related reactions were more highly transcribed. The contextualized model in this case

also utilized glycolysis as the only means of generating ATP, directly matching those results obtained from pFBA. Alternatively, as illustrated in Fig. 2C, when genes associated with Stickland fermentation and peptide substrate transporters were more highly transcribed, RIPTiDe was able to discern that the less directly parsimonious pathway should be utilized instead when provided sufficient transcriptomic evidence. In each case, reactions related to the opposite pathway were correctly pruned while generating the final context-specific model. Furthermore, flux samples for the cellular objective from each contextualized example model revealed that significantly more ATP could be produced when employing glycolysis (Table S2), further supporting the energetic favorability of this pathway. These combined results confirmed that RIPTiDe was indeed able to use a less parsimonious pathway when necessary to better reflect the associated transcriptomic dataset.

RIPTiDe with a Curated Model of *E. coli* Metabolism

To perform a thorough evaluation of our approach, we utilized the genome-scale metabolic network reconstruction iJO1366, for the model organism *E. coli* strain K-12 substr. MG1655 [31]. This GENRE has been the subject of extensive manual curation and experimental validation, making it the best possible candidate for examining new approaches in constraint-based modeling [32]. Leveraging these years of research, we utilized this well described system to determine the biological validity of predicted context-specific differences returned by RIPTiDe. To accomplish this goal, we performed a comparative analysis using 3 separate transcriptomic sequencing datasets, each derived from well characterized *in vitro* media conditions including LB aerobic liquid rich media, M9+glucose aerobic liquid minimal media, and M9+glucose anaerobic liquid minimal media. Each dataset was collected at approximately the same mid-exponential phase of growth [33,34]. These data were chosen due to the large amount of characterization each has received over time, allowing our predictions to be more readily supported compared to assessing metabolic differences possibly seen across numerous less well explored conditions. Additionally, *E. coli* is a facultative anaerobe with known differences in metabolism between these lifestyles that we hoped to capture with RIPTiDe [35]. For all subsequent work, we took an unsupervised approach by allowing for unpenalized flow of all extracellular

metabolites in and out of the system (exchange reaction bounds set to ±1000). This approach allowed for inferences about the most likely growth substrates for a contextualized GENRE since only those that carry flux following RIPTiDe remain.

We first set a performance baseline with three commonly utilized platforms for integration of transcriptomic data with GENREs; GIMME, iMAT, and CORDA [11–13]. These algorithms require abundance cutoffs for selectively deactivating reactions associated with genes that were transcribed below the desired level and as such, we utilized the recommended values published alongside each method. The python implementations of both GIMME and iMAT initially returned models that were unable to achieve flux through the biomass synthesis reaction in any condition, and required reintroduction of metabolic reactions establish the ability of the models to grow (Table S3). The resultant models were no longer concordant with the input transcriptomes, reintroducing 138 and 104 reactions respectively associated with genes that fall below the defined transcription cutoff (Table S3). CORDA did produce a model that was capable of growth, however 1001 out of 2583 reactions were included in the resulting model which does not substantially narrow the possible strategies for growth to a level that can be studied easily. These results in combination with the previously mentioned performance of pFBA alone (without omic data, only maximized parsimony of flux) highlight the advantages of RIPTiDe over other established methods of transcriptomic data integration.

Context-specific *Escherichia coli* Biology Uncovered by RIPTiDe

We next performed RIPTiDe-based analyses of iJO1366, utilizing the same transcriptome data described previously. As a standard of comparison, we also implemented a more standard version of the pFBA algorithm where the sum of fluxes is minimized while maintaining a maximum objective flux, to establish the most “economic” overall flux distribution that achieves maximum growth. Different topologies of reaction and metabolite inclusion were found among each of the contextualized models; however none were completely unique and a core set of metabolic reactions that were necessary across all growth conditions was shared (Fig. 3A). These distinctions in network topology were further crystallized by conditional gene essentiality between the RIPTiDe-contextualized models (necessary for

at least 1% of optimum growth). Although most essential genes were shared by all four models (187 genes, Table S4), those essential genes that were context-specific primarily related to pathways of biological significance to their respective media conditions (Fig. 3B). As a corollary to creating a smaller metabolic network than those using transcriptomes contextualized by RIPTiDe (by an average of 9.12%), fewer genes were distinctly essential in the model resulting from standard pFBA (5 genes) compared to any of the RIPTiDe-contextualized *in vitro* transcriptomes (8, 15, and 20 genes respectively; Table S4). Interestingly, the contextualized models on average shared larger ratios of components (~7.07% reactions and ~1.78% metabolites) with each other than with pFBA, supporting that RIPTiDe selects more biologically relevant patterns of metabolism than pFBA alone (Fig. 3A). Collectively, these results indicate that our algorithm correctly identifies context-specific network topologies given an associated transcriptome.

We also calculated the ranges of relative doubling time for each of the models from the flux samples associated with the biomass objective function [36]. The base model with no additional flux sum minimization constraints displayed much more variation in the amount of flux that is possible through the biomass reaction compared to any of the other, more constrained models. It is important to note that achieving higher flux values, as in the unconstrained model, is still possible in the constrained and contextualized models however the number of specific distributions that reach these levels have a very low probability of ever appearing and would likely only do so after many more sampled flux distributions. This outcome speaks to the truly context-specific nature of the output models that better reflect actual biological trends. Interestingly, growth from the pFBA-generated model closely mirrored those from the RIPTiDe-contextualized model in LB rich media (Fig. 3C). This similarity signified that rapid doubling time in rich media identified by RIPTiDe is analogous to the most mathematically optimal growth solution, which was in line with many predictions in the original publication of pFBA [17]. We went on to compare experimentally measured doubling times of *E. coli* K-12 in liquid LB [35], and found that the most frequent doubling times for both RIPTiDe and pFBA were slower than those measured in the laboratory. Another interesting finding was that the computed doubling times for contextualized

models from both aerobic & anaerobic M9+glucose minimal media were significantly slower than in rich media and analogous to previously measured experimental values (Fig. 3C) [37]. These results suggested the RIPTiDe-contextualized GENREs not only topologically reflected known biological trends, but also phenotypically mimicked experimentally measured behaviors.

We subsequently focused our analysis on assessing known differences in core metabolism across transcriptomes in RIPTiDe-contextualized models. Growth conditions are known to affect which forms of metabolism are possible or favored at a given time, driving the cell to utilize alternative pathways for replenishing it's cellular currency. One such anabolic circuit is the pentose phosphate pathway, which is a critical component of core metabolism for most life that combats oxidative stress as well as maintains carbon homeostasis through provision of NADPH and 5-carbon sugars [38]. The pathway consists of two phases, oxidative (NADPH generating) and non-oxidative (pentose generating), and has been previously shown to be differentially activated in wildtype *E. coli* during aerobic growth in M9 minimal media (favors oxidative) and LB rich media (favors non-oxidative) [39]. Strikingly, the RIPTiDe-contextualized model in aerobic M9 minimal media indeed only contained the oxidative component of the pentose phosphate pathway (Fig. 4A). Alternatively, only the non-oxidative component was present in the model resulting from growth in LB, instead generating more of its NADPH from the TCA cycle (Fig. 4B). It is important to note also that in both growth conditions a primary source of NADPH was glycerol-3-phosphate dehydrogenase, with significantly more activity captured in flux samples from contextualized M9 growth (Fig. S1A). This reaction is a major contributor to the bacterial electron transport chain and an essential part of energy generation in all known growth conditions [40]. These results signified that RIPTiDe is able to capture previously measured, context-specific behaviors with only transcriptomic characterization.

Application of RIPTiDe With *In vivo* Metatranscriptomic Data

Finally, we sought to test the utility of RIPTiDe for capturing metabolic behavior of bacteria from within host-associated communities of microorganisms. Computational analysis of a discrete bacterial species within a complex environment as a member of a larger microbial community has been

historically difficult [41]. To address this problem, we utilized *in vivo* metatranscriptomic abundance data collected from the cecum of wildtype C57Bl6 mice. These data were chosen because the specific antibiotic pretreatment (intraperitoneal injection of clindamycin) resulted in a bacterial community composed of >90% *E. coli* [8]. To begin the analysis, we first made certain that the *in vivo* transcript abundance distribution reflected the same negative binomial type as those derived from *in vitro* transcriptomes mapped to the same *E. coli* K-12 MG1655 genome (Fig. 5A). This result demonstrated that with sufficient sequencing depth, a species with a GENRE may be analyzed with RIPTiDe using data collected from among a diverse collection of microbes. After contextualization of this data we first found that calculated growth rates from biomass flux samples fell within expected ranges of those experimentally observed *in vivo* (Fig. S1B) [42]. Additionally, the context-specific flux values during sampling were correlated with assigned linear coefficients based on transcript abundances (Fig. 5A & Fig. S3). This result supported that the transcriptomic data was indeed informative, but not overly deterministic, of optimal metabolic states through RIPTiDe. We then assessed how distinctive the flux distributions were of those metabolic reactions which were shared between context-specific models. Cross-referencing each contextualized model to identify shared reactions, we then performed principle component analysis using Bray-Curtis dissimilarity for those flux samples corresponding with this core metabolic subnetwork (Fig. 5B). This analysis revealed that although numerous reactions and metabolites are shared between context-specific models, they are utilized significantly differently to manage growth across media conditions and supported differential metabolism is likely necessary to cope with distinct environments.

Next, we contrasted the RIPTiDe models generated using *in vivo* or LB rich media transcriptomes. Aside from one environment being highly aerobic and the other largely anaerobic, previous research on *E. coli* physiology during growth in liquid LB concluded that the bacterium utilizes amino acids that are simple to metabolize for growth early on, then switches to other substrates in a diauxic-like growth pattern. LB broth has also been shown to have only modest levels of usable carbohydrate growth substrates, increasing the importance of amino acid catabolism for *E. coli* [43].

Additionally, *in vivo* amino acids and simple sugars are contested resources in the cecum as the majority are already absorbed by the host or used by other microbes upstream in the GI tract, making these growth conditions extremely different [44]. As an analogy to differential growth substrates, we looked to assess the presence of context-specific exchange reactions. We identified 7 exchange reactions that were present in the LB-associated model but not *in vivo*, and 5 of these were peptides or peptide conjugates (Fig. 5C). Carbohydrate growth substrates have been demonstrated to be at very low concentrations in LB media; *E. coli* therefore primarily utilizes amino acids to generate energy during early exponential phase [43]. Indeed, upon assessing the measured changes in the *in vivo* concentrations of the highlighted metabolites they were reduced in the context where *E. coli* dominated the environment, possibly due to consumption by the microbe (Fig. S2). Additionally, the exchange for glucose-6-phosphate was also only found in LB-contextualized models, reinforcing that lowly abundant sugar phosphates are utilized secondarily for growth by *E. coli* compared to amino acid substrates in LB media [43]. Interestingly, fewer exchange reactions associated with the *in vivo* context may also imply that a greater amount of metabolite substrates must be produced intracellularly instead of imported from the environment. Continuing this analysis now with shared intracellular reactions we utilized Random Forest on the sampled flux distributions for each reaction to determine those metabolic functionalities that most distinguish the two models (Fig. 5D). Among other findings, machine learning revealed consistently high flux ranges for flavodoxin reductase in aerobic LB culture, but not *in vivo*. This enzyme is important for dealing with oxidative stress which would be much more prevalent in the broth culture than in the mostly anoxic environment of the mammalian gut [45]. Altogether, our results support that RIPTiDe is also a strong platform for the analysis of bacterial GENRE behavior in the context of complex microbial communities through the use of metatranscriptomic data.

Conclusions

In order to survive in a given environment, an organism must express the appropriate forms of metabolism to cope with changing conditions. Production of transcript and protein is metabolically expensive, and over-production likely puts the cell at risk if further perturbation is placed on the system

[46]. Although recent findings suggest that levels of transcript for a given enzyme product do not always directly correlate with the amount of functional protein present in the cell at a given time [20], these values may be considered a signature of cellular “investment” and therefore may still be useful in interrogation of metabolism. Furthermore, understanding the quantity of even a small number of functional enzymes remains extremely challenging, which negatively impacts our ability to grasp which and why resources are used by the cell [47]. We should then consider transcriptomic abundance data as signatures of cellular energy expenditure where high transcript levels for a particular gene are an indication of large cellular investment related to that particular function, but are not particularly deterministic of the associated phenotype. Parsimonious Flux Balance Analysis approaches the idea of maximizing cellular economy, and we have built on this idea but employing RNA-Seq data to inform where the cell is most likely spending its energy.

We have demonstrated empirically that our algorithm correctly calculated both *in vitro* and *in vivo* metabolic phenotypes with reasonably high accuracy. In addition to these predictions, our method does not require prior knowledge of active metabolic subnetworks which can preclude work with understudied or uncultured species that cannot or have not been grown or characterized in isolation. Some possible limitations of RIPTiDe do exist however. First, as RIPTiDe relies on the distribution associated with shotgun transcriptomic sequencing, it is less suited for other types of omic data in its current form [48]. RNA-Seq reads are easily attributable to single species within the microbiota due to strict genomic mapping techniques and are measured within a single sequencing assay, however with sufficient consideration to reaction weights given other data types RIPTiDe’s functionality could be expanded. Second, by condensing to such a small metabolic network “solution space”, these models may have a risk of overfitting. To counteract this, we provided options for manually adding a minimum threshold that coefficients are allowed as well as providing core reactions that must be included in a context-specific model. In spite of these concerns, we have shown our algorithm to be a strong approach for contextualization of genome-scale metabolic network reconstructions for both *in vitro* and *in vivo* conditions, producing models that closely mirror known biological phenotypes. This technique

may also prove effective for generating tissue-specific models in GENREs from larger eukaryotes. The implementation of RIPTiDe rewards divergence from the most numerically parsimonious state to one that is more reflective of cellular investment, and in turn more illustrative of *in situ* biology. In summary, RIPTiDe approaches transcriptomic data integration uniquely from previous techniques and provides a more biologically relevant platform for considering the cellular behavior in complex environments.

Methods

Parsimonious Flux Balance Analysis Adaptation

RIPTiDe was implemented within COBRApy [49] and is freely available with installation instructions at https://github.com/mjenior/riptide. To save on computational time, which can be intensive during flux sampling object generation, flux variability analysis was utilized first to identify and remove all blocked reactions (those reactions which do not carry flux under any circumstances). At this point, optimal flux through the objective function was set as a constraint with a 20% range of allowed variation (based on 80% fraction of optimal objective function value). Following sensitivity testing, this level was found to result in models which best achieved high growth rates while reaching acceptable levels of concordance with input transcriptomes (Table S1). As previously described, remaining reactions were then assigned a linear coefficient based on the user provided transcriptomic abundance data [50].

$$C_g = \frac{T_g}{T_{max}}$$

For reaction weight C_g (linear coefficient) assignments, the reads-per-million normalized transcript abundance of gene T_g within the sorted (smallest to largest) transcript abundance distribution is divided by the largest abundance T_{max} this distribution. These C_g values are assembled into a sorted (largest to smallest) distribution of linear coefficients. Each gene in the sorted transcript abundance distribution is then assigned the corresponding linear coefficient from this inverted coefficient distribution. Reactions in the model without genes are assigned the median value from the coefficient distribution.

$$min \sum^{r} v_{irrev} \cdot w_r$$

$$s.t.\ \ max\ v_{objective} = v_{objective,\ lb}$$

$$S_{irrev} \cdot v_{irrev} = 0$$

$$0 \leq v_{irrev} \leq v_{max}$$

As in parsimonious-FBA, each reversible reaction is separated into individual forward and reverse reactions v_{irrev}. Linear coefficients w_r for each vector can now integrated and an objective function is then defined to minimize total of flux across the entire model. These steps are all performed subject to near optimal biomass flux ($\max v_{objective}$) and mass-balance constraints. Optimization (standard FBA) is performed and inactive reactions are then removed from the model. The model is then checked for orphaned genes and metabolites, which are subsequently removed as well.

Flux Sampling

In order to adequately explore intracellular flux through the remaining reactions, previous coefficients are reassigned to remaining reactions using the unsorted linear coefficient distribution (assigned directly corresponding to transcript abundance). All reactions with coefficients are then assembled into a single expression which is set as a constraint for the overall solution space of all reactions with symengine (https://github.com/symengine/symengine):

$$max \sum^{r} v_{irrev} \cdot w_r$$

The new coefficients are then assigned to the corresponding irreversible reactions v_{irrev}. Flux sampling was then performed with the gapsplit sampler using standard parameters and 500 samples (https://github.com/jensenlab/gapsplit) [28]. Flux variability analysis for the constrained models is also performed to establish flux ranges for each reaction [51]. An overall concordance score is then calculated by finding the Spearman correlation coefficient and associated *p*-value between median absolute flux and their associated linear coefficient during flux sampling for all reactions, then multiplying the resultant *rho* value by 100.

Example Model Construction

The example model used for initial RIPTiDe testing consists of 16 reactions (7 exchanges, 7 transporters, and 2 metabolic) and 14 metabolites across 2 compartments (extracellular and intracellular). The intracellular demand reaction for ATP generation was utilized as the objective. Simulated transcriptomes are shown in Table S2 and an SBML file (example_GENRE.sbml) for the model is also available at github.com/mjenior/Jenior_RIPTiDe_2019.

Data Sources and Processing

Fastq files associated with each study were accessed from the NCBI Sequence Read Archive (SRA: PRJNA313438, SRA: PRJNA212627, & SRA: PRJNA415307). Raw cDNA reads were quality-trimmed with Sickle [52] at ≧Q30 and subsequently mapped to *Escherichia coli* K-12 MG1655 genes (EMBL-EBI on Jan 28, 2019) using Bowtie2 [53]. MarkDuplicates was then used to remove optical and PCR duplicates (broadinstitute.github.io/picard/). The remaining mappings were converted to idxstats format using SAMtools [54], and the read counts per gene were tabulated. Discordant pair mappings were discarded, and counts were then normalized to read length and gene length to give a per-base report of gene coverage. If replicate sequencing was performed, median transcription values for each gene were then calculated. The data were then converted to a python dictionary of gene keys and associated numeric transcript abundances for use with RIPTiDe. *In vivo* untargeted metabolomic data were generated for a previous study [8], and accessed from github.com/SchlossLab/Jenior_Metatranscriptomics_mSphere_2018/. The current version of iJO1366 (GENRE for *E. coli* K-12 MG1655) was accessed from the BiGG Model database [55] on 09-26-2018.

Statistical Methods

All statistical analysis was performed using R (v.3.2.0). Significant differences between flux distributions and metabolite concentrations were determined by Wilcoxon signed-rank test with Benjamini-Hochberg correction. Principle Component Analysis of sampled flux distributions was performed using the vegan package [56], and machine-learning was accomplished with the base implementation of Random Forest in R [57].

Data and Software Availability

As previously stated, RIPTiDe is freely available for download at https://github.com/mjenior/riptide and also as a full package from the Python Package Index (https://pypi.org/project/riptide/). Complete analysis associated with this study is available in a Jupyter notebook hosted on a GitHub repository along with all other data and code associated with this study (https://github.com/mjenior/Jenior_RIPTiDe_2019).

Acknowledgements

The authors would like to thank Paul Jensen for the suggestion of their gapsplit flux sampling algorithm. The authors also appreciate Jhansi Leslie, Greg Medlock, Laura Dunphy, Kristopher Rawls, and Maureen Carey for numerous discussions on the utility, execution, and biological validity of the core algorithm.

Figures and Legends

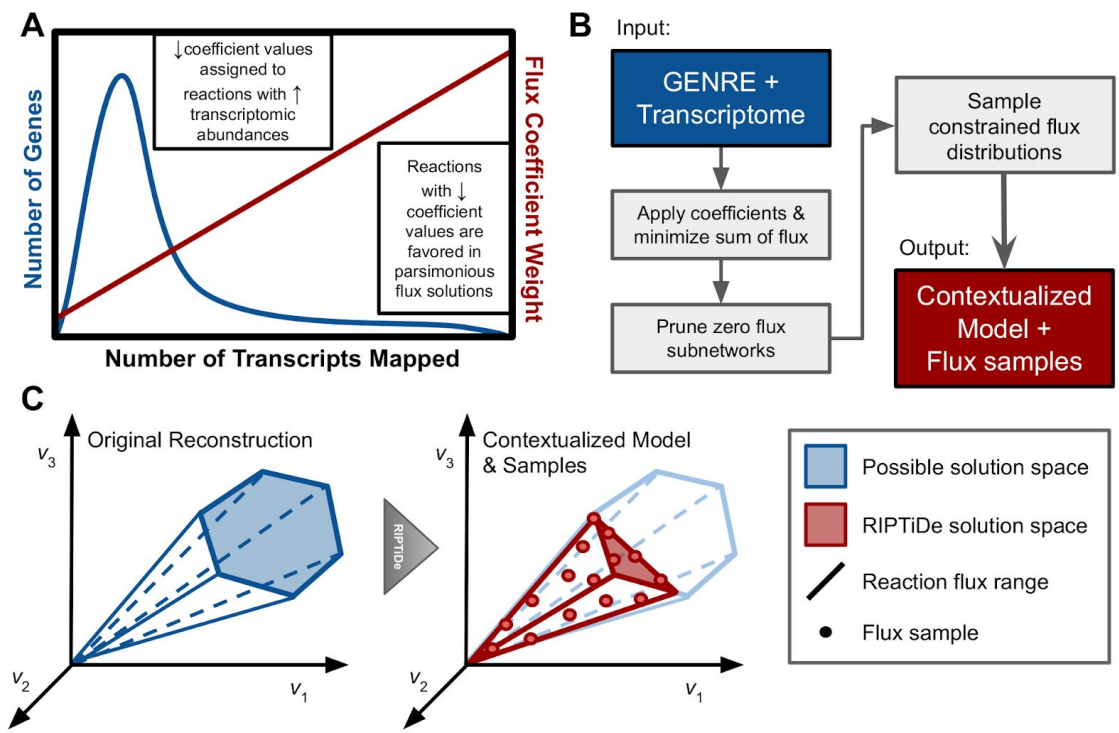

Figure 1 | Visual representation of RIPTiDe workflow. (A) Example transcript and resultant reaction weight distributions with calculated by RIPTiDe. Reactions for genes that recruit greater abundances of transcript are assigned smaller linear coefficients which in turn result in higher likelihood of usage in an overall flux minimization objective. **(B)** Schematic outlining computation performed during each step of RIPTiDe. The platform requires two input data structures: a genome-scale network reconstruction and transcript abundances associated with genes. During initialization totally blocked reactions are identified and removed, and transcript abundances for genes are transferred to their corresponding reaction allowing coefficients to be assigned. Each reversible reaction is now also separated into pairs of irreversible reactions. Applying principles of parsimonious FBA to the constrained model, a minimum sum of fluxes optimization is performed with respect to a predefined minimum flux through the cellular objective. Reactions that no longer carry flux in this state are pruned, and flux sampling is performed on the intermediate model to determine context-specific bounds for the remaining reactions. **(C)** Pictorial representation of the impact of RIPTiDe on the original reconstruction, converting it to a context-specific model of metabolism with as parsimonious a possible metabolic solution space with respect to the given transcriptomic data.

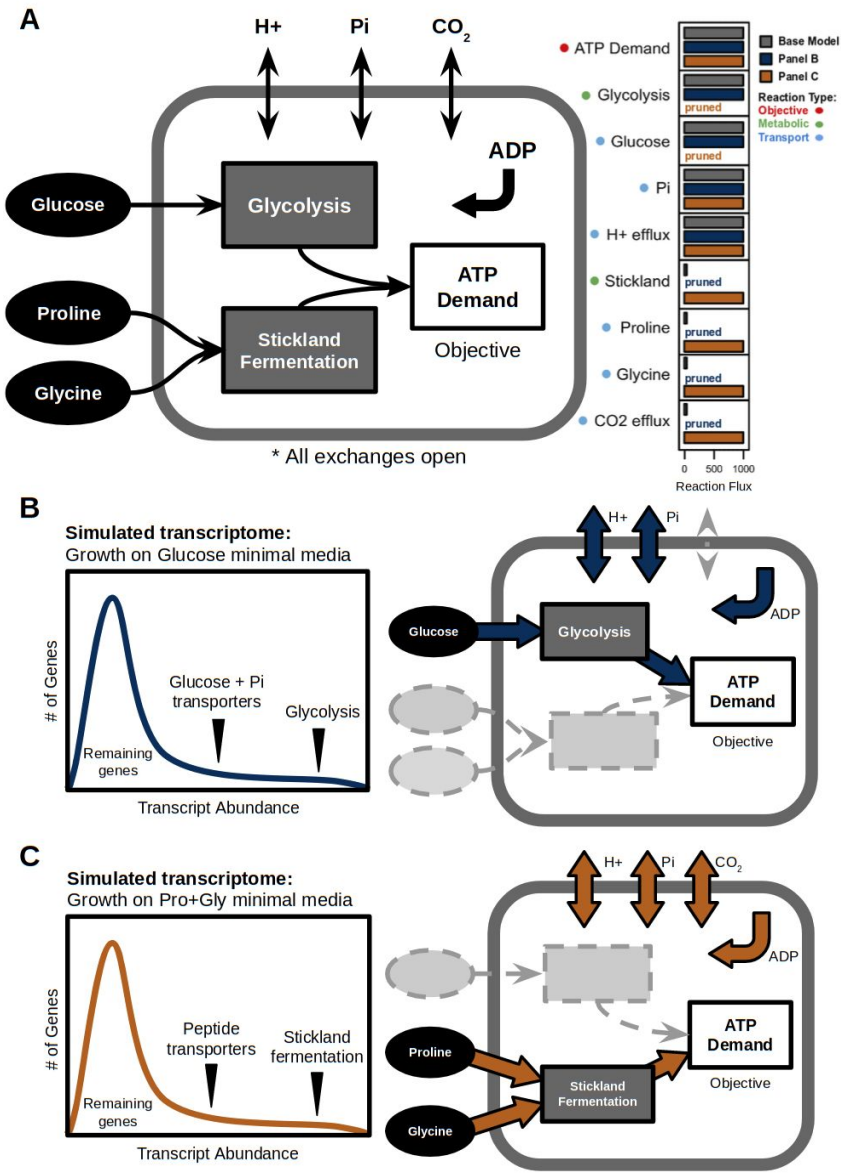

Figure 2 | Example model of bacterial metabolism supports the utility of RIPTiDe for identifying most likely context-specific strategies. (A) General topology of simplified GENRE that consists of 16 reactions and 14 metabolites in total (Table S2). The objective of ATP generation can be achieved through two separate means: either the catabolism of glucose via glycolysis or through the paired fermentation of proline and glycine. Glycolysis is the more parsimonious pathway for generating ATP, which is reflected in the pFBA solution fluxes to the right of the diagram. **(B)** RIPTiDe is able to correctly identify the most likely route of flux when provided a transcriptome. **(C)** When provided with transcriptomic evidence, RIPTiDe is able to identify a less parsimonious, but more concordant solution.

Despite very low transcription, RIPTiDe still identifies the necessity for carbon dioxide efflux during fermentation to maintain mass balance.

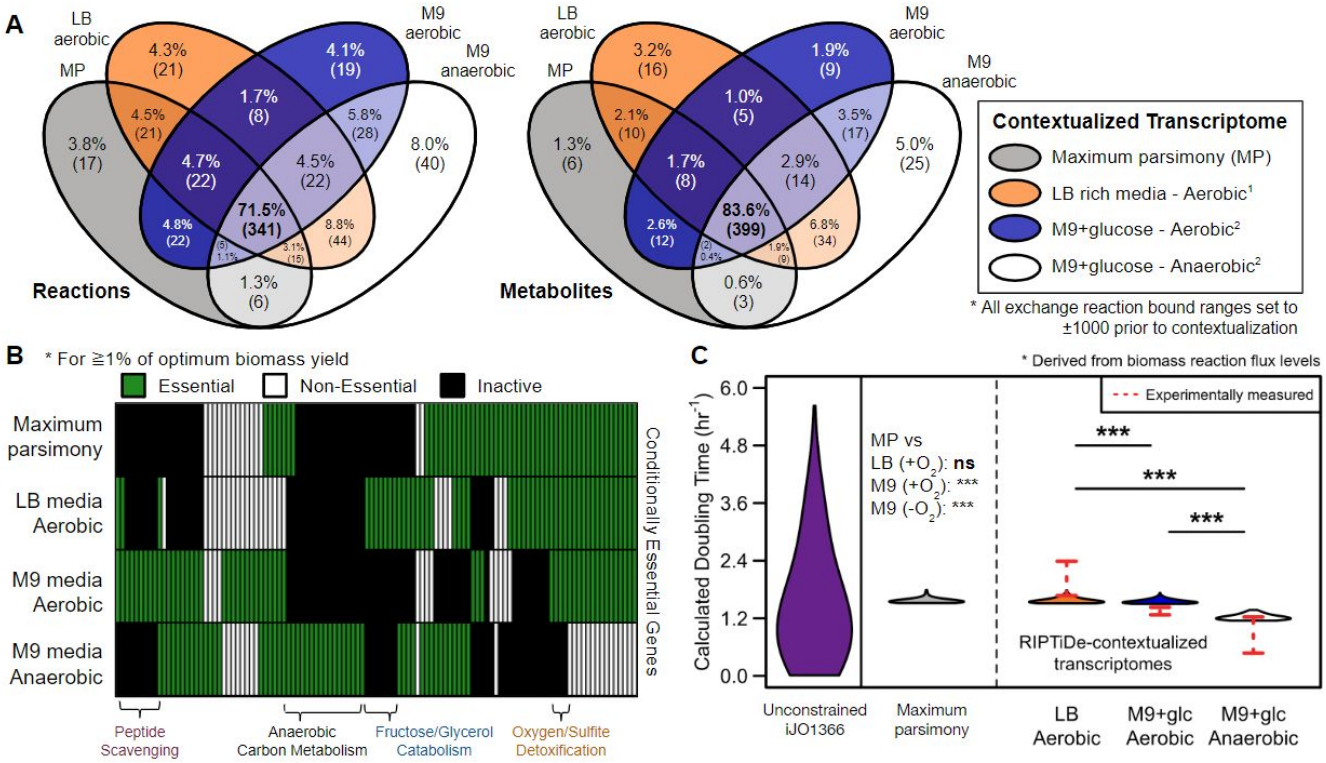

Figure 3 | RIPTiDe identifies established biological differences in *E. coli* str. K-12 substr. MG1655 across *in vitro* growth conditions. Without increased user supervision, RIPTiDe correctly predicts behavior and context-specific pathways for *E. coli* str. K-12 substr. MG1655, while simultaneously producing functional models that can be characterized phenotypically. **(A)** Comparison of included network topology among RIPTiDe-contextualized models, also contrasted against maximally parsimonious flux distributions. The majority of reactions and metabolites are conserved across models within a central set of pathways (71.5% and 83.6% respectively). **(B)** Comparison of essentiality for conserved genes across pruned RIPTiDe models. Hierarchical clustering reveals context-specific pathway essentially, labeled across the bottom axis, based on the environment in which the bacterium is growing. **(C)** Calculated doubling times from 500 biomass reaction flux samples. Significance determined by Wilcoxon signed rank test with Benjamini Hochberg correction (***, *p*-value

≦ 0.001; ns, *p*-value > 0.05). Red lines indicate experimental measurements for corresponding *in vitro* media conditions.

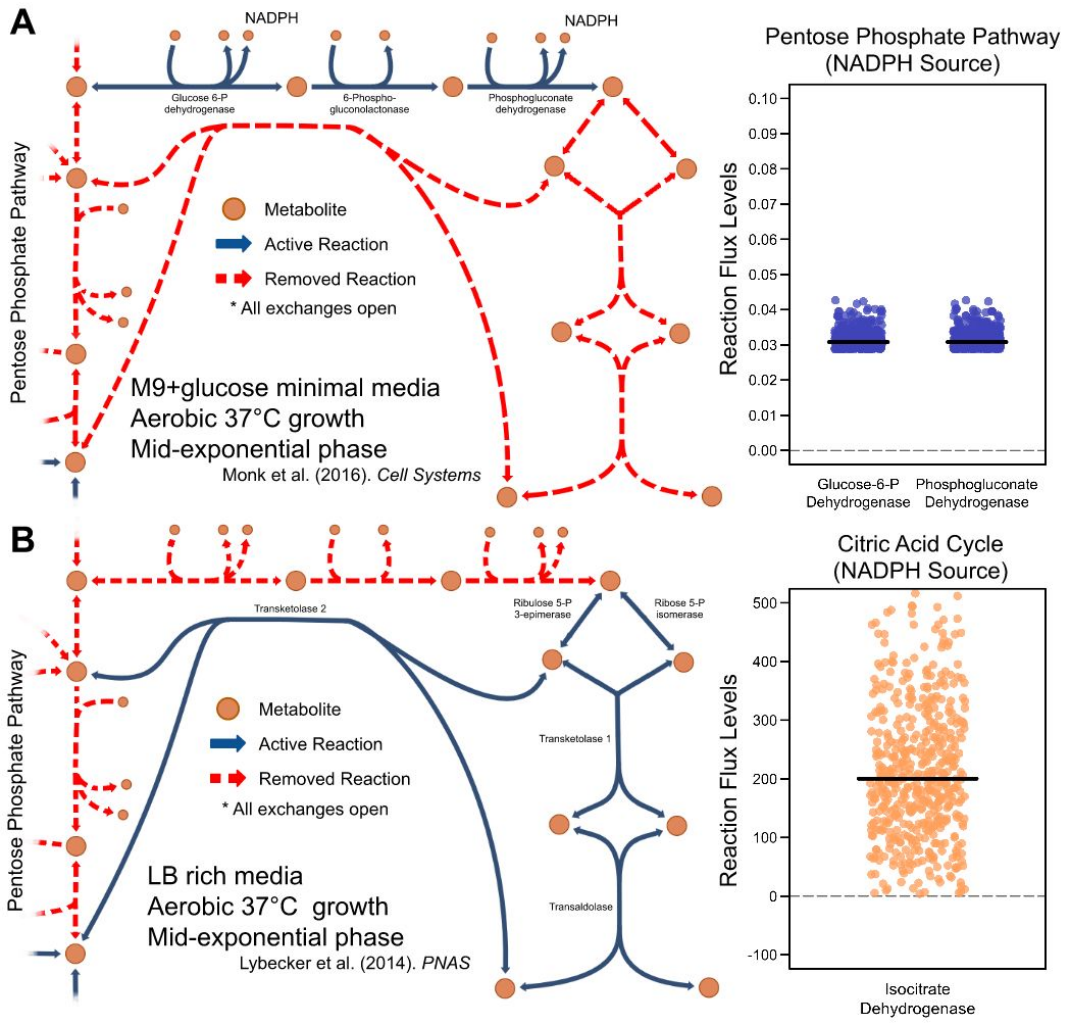

Figure 4 | Cellular sources of NADPH in RIPTiDe contextualized models reflect known biological differences in pentose phosphate pathway activity. In both panels are shown: Reactions included from the pentose phosphate pathway in each context-specific model, pruned reactions from this pathway in both contexts, the distribution of transcript abundances in each dataset, and the flux sampling results for NADPH producing reactions in each constrained model. Shown are RIPTiDe

results with contextualized transcriptomic abundances from *E. coli* K-12 MG1655 grown in **(A)** M9+glucose minimal media, and **(B)** LB rich media (using the iJO1366 GENRE).

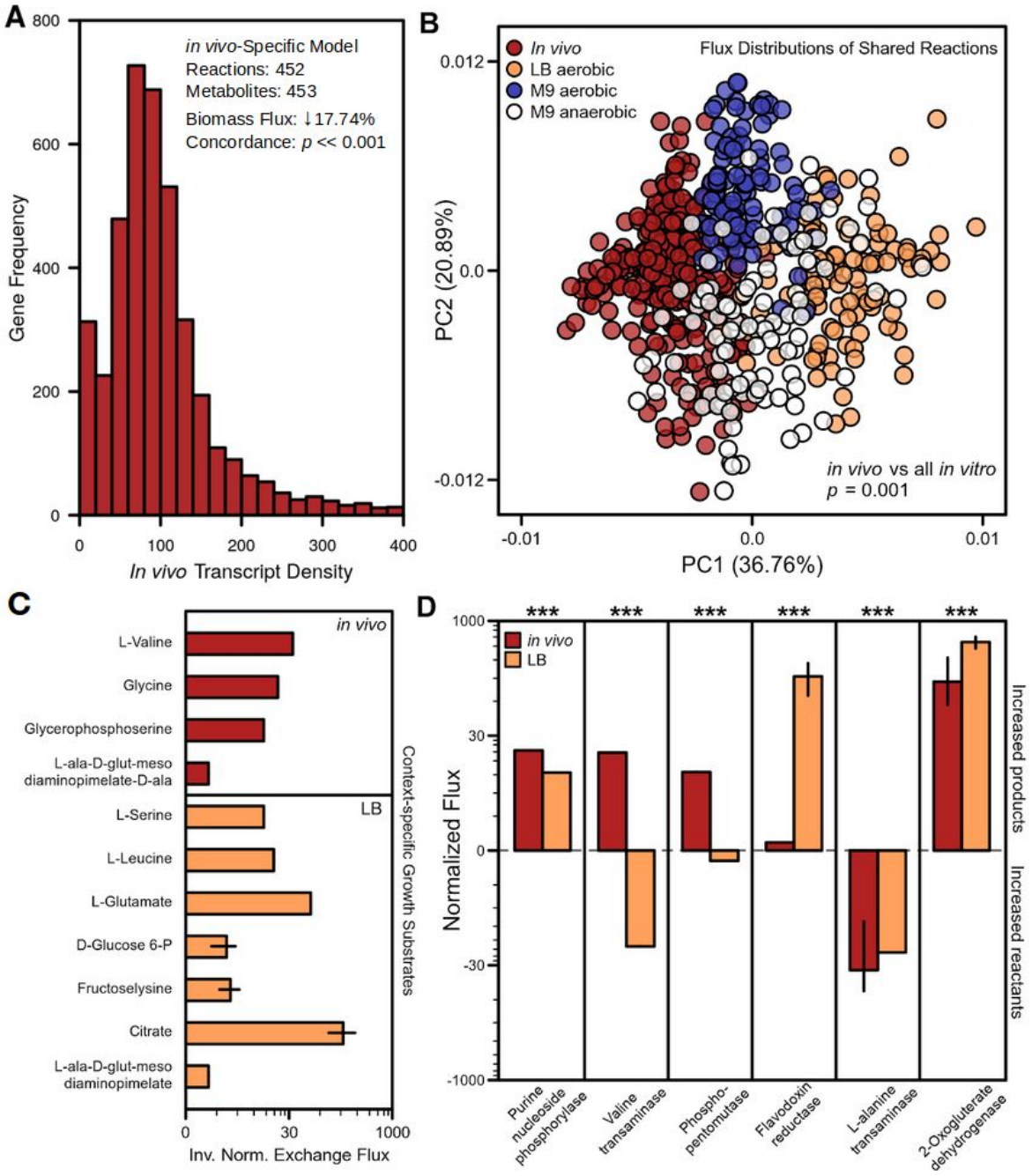

Figure 5 | RIPTiDe reveals differential host-associated, metabolism utilizing *in vivo* metatranscriptomic data. Transcriptomic reads attributable to *E. coli* were extracted from a metatranscriptomic dataset sequenced from the cecum of mice in which *E. coli* is the most highly

abundant community member [8]. **(A)** *In vivo* transcript abundance of reads recruited to the *E. coli* K-12 MG-1655 genome and metrics associated with the resulting context-specific model. **(B)** PCoA ordination of Bray-Curtis dissimilarities between flux distributions among reactions of each contextualized iJO1366 model either from *in vivo* or LB rich media (aerobic, *in vitro*) conditions. Significance was calculated by PERMANOVA. **(C)** Metabolite substrates of exchange reactions exclusive to the contextualized models shown. Inverse, normalized flux was calculated by dividing each flux sample by the associated flux through biomass in the same distribution, then multiplied by the overall median flux through biomass across all conditions analyzed. Median and interquartile ranges are displayed. **(D)** Results from Random Forest machine learning to identify the most differentiating metabolic reactions between models. Top six reactions from this analysis are displayed. Median and interquartile ranges are shown, and significant differences in computed doubling times were determined by Wilcoxon signed-rank test with Benjamini-Hochberg correction (***, *p*-value ≦ 0.001).

Supplemental Figures

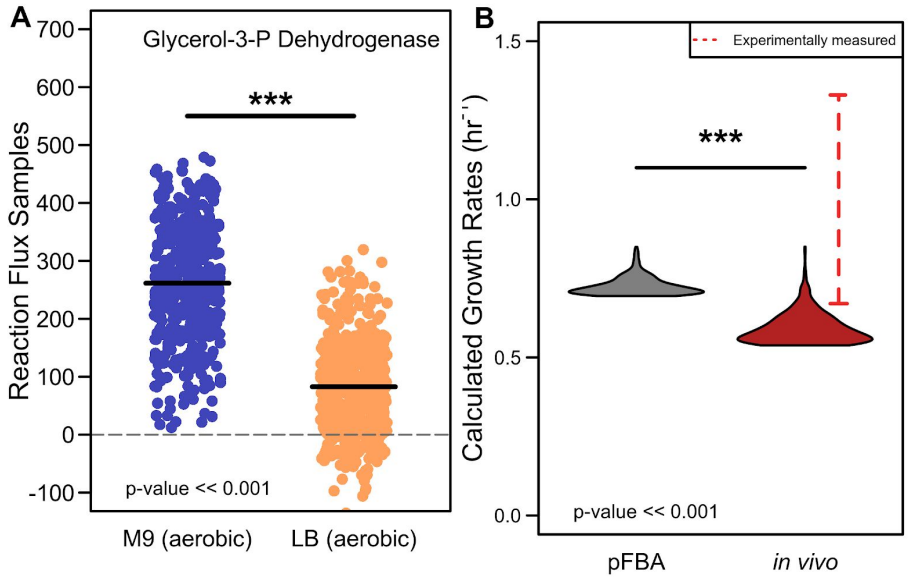

Figure S1 | *In silico* growth predictions for *E. coli* K-12 MG1655 further supports the utility of RIPTiDe. (A) Glycerol-3-phosphate dehydrogenase activity of Glycerol-3-Phosphate Dehydrogenase in M9 and LB. Alternative source of cellular NADPH that is conserved across both contextualized iJO1366

models. **(B)** Validation of RIPTiDe results from *in vivo* metatranscriptomic reads mapped to *E. coli* K-12 MG1655 genes. Computed *in vivo* doubling time from biomass reaction flux samples of contextualized model. Red lines indicate previous experimentally measured in vivo growth rate ranges [34]. Computed *in vivo* range also significantly differs from pFBA analysis of iJO1366. Significant differences were determined by Wilcoxon signed-rank test.

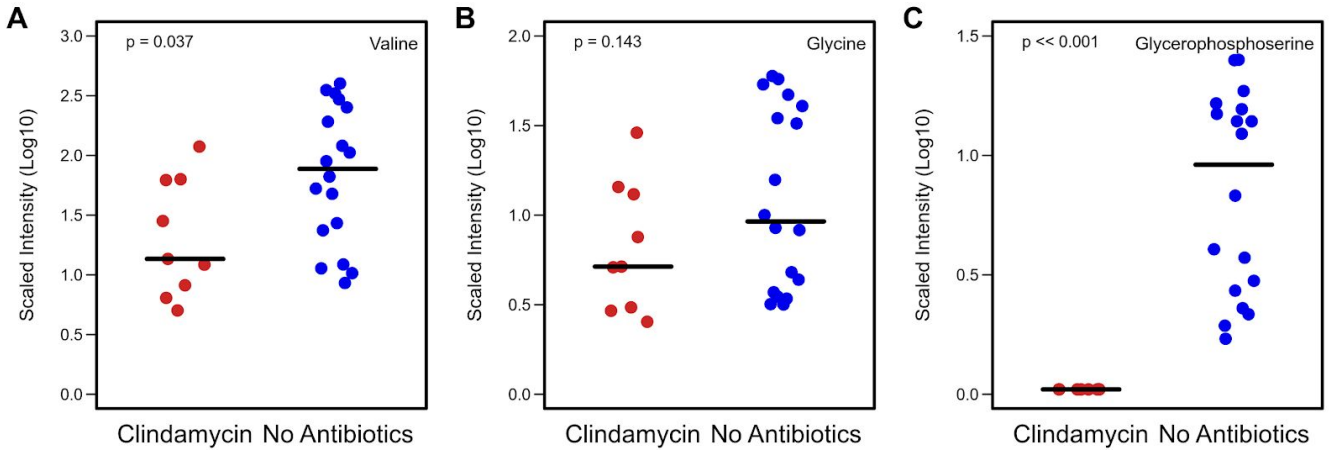

Figure S2 | Relative concentrations of metabolites highlighted during *in vivo* analysis of iJO1366 with RIPTiDe. Concentrations quantified with liquid chromatography mass-spectrometry previously [3]. Comparing the metabolite content in cecal content of mice pretreated with an intraperitoneal injection of clindamycin against untreated control animals. **(A)** Valine, **(B)** Glycine, **(C)** and Glycerophosphoserine. Significant differences were determined by Wilcoxon signed-rank test with Benjamini-Hochberg correction.

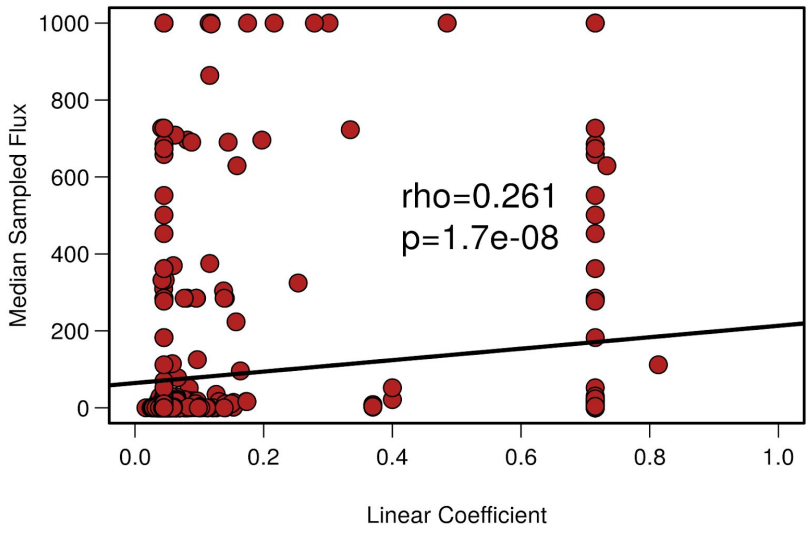

Figure S3 | Spearman correlation results between RIPTiDe linear coefficients from flux sampling and median values from sampled flux ranges from *in vivo* transcriptome. Assigned linear coefficients are significantly correlated with their associated reaction activities.

Table S1 | Sensitivity testing for minimum objective threshold percentage

Table S2 | Structure and analysis for example metabolic reconstruction

Table S3 | Parameters and results from previous transcriptomic integration algorithms

Table S4 | Network topology and gene essentiality across RIPTiDe contextualized models

References

1. Comte J, Fauteux L, Del Giorgio PA. Links between metabolic plasticity and functional redundancy in freshwater bacterioplankton communities. Front Microbiol. 2013;4: 112.

2. Güell O, Sagués F, Serrano MÁ. Essential plasticity and redundancy of metabolism unveiled by synthetic lethality analysis. PLoS Comput Biol. 2014;10: e1003637.

3. Jenior ML, Leslie JL, Young VB, Schloss PD. *Clostridium difficile* colonizes alternative nutrient niches during infection across distinct murine gut microbiomes. 2016. doi:10.1101/092304.

4. Maksymiuk C, Balakrishnan A, Bryk R, Rhee KY, Nathan CF. E1 of α-ketoglutarate dehydrogenase defends *Mycobacterium tuberculosis* against glutamate anaplerosis and nitroxidative stress. Proceedings of the National Academy of Sciences. 2015;112: E5834–E5843.

5. Oberhardt MA, Palsson BØ, Papin JA. Applications of genome-scale metabolic reconstructions.

Mol Syst Biol. 2009;5: 320.

6. Dunphy LJ, Yen P, Papin JA. Integrated Experimental and Computational Analyses Reveal Differential Metabolic Functionality in Antibiotic-Resistant *Pseudomonas aeruginosa*. Cell Syst. 2019;8: 3–14.e3.

7. Bashiardes S, Zilberman-Schapira G, Elinav E. Use of Metatranscriptomics in Microbiome Research. Bioinform Biol Insights. 2016;10: BBI.S34610.

8. Jenior ML, Leslie JL, Young VB, Schloss PD. *Clostridium difficile* Alters the Structure and Metabolism of Distinct Cecal Microbiomes during Initial Infection To Promote Sustained Colonization. mSphere. 2018;3. doi:10.1128/msphere.00261-18.

9. Li M, Jain S, Baker BJ, Taylor C, Dick GJ. Novel hydrocarbon monooxygenase genes in the metatranscriptome of a natural deep-sea hydrocarbon plume. Environ Microbiol. 2014;16: 60–71.

10. Magnúsdóttir S, Thiele I. Modeling metabolism of the human gut microbiome. Curr Opin Biotechnol. 2018;51: 90–96.

11. Becker SA, Palsson BO. Context-Specific Metabolic Networks Are Consistent with Experiments. PLoS Comput Biol. 2008;4: e1000082.

12. Zur H, Ruppin E, Shlomi T. iMAT: an integrative metabolic analysis tool. Bioinformatics. 2010;26: 3140–3142.

13. Schultz A, Qutub AA. Reconstruction of Tissue-Specific Metabolic Networks Using CORDA. PLoS Comput Biol. 2016;12: e1004808.

14. Cho JS, Gu C, Han TH, Ryu JY, Lee SY. Reconstruction of context-specific genome-scale metabolic models using multiomics data to study metabolic rewiring. Current Opinion in Systems Biology. 2019. pp. 1–11. doi:10.1016/j.coisb.2019.02.009.

15. Wang Y, Eddy JA, Price ND. Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. BMC Syst Biol. 2012;6: 153.

16. Colijn C, Brandes A, Zucker J, Lun DS, Weiner B, Farhat MR, et al. Interpreting expression data with metabolic flux models: predicting *Mycobacterium tuberculosis* mycolic acid production. PLoS Comput Biol. 2009;5: e1000489.

17. Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, et al. Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. Mol Syst Biol. 2010;6: 390.

18. Wortel MT, Bosdriesz E, Teusink B, Bruggeman FJ. Evolutionary pressures on microbial metabolic strategies in the chemostat. Sci Rep. 2016;6: 29503.

19. Machado D, Herrgård M. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. PLoS Comput Biol. 2014;10: e1003580.

20. Guimaraes JC, Rocha M, Arkin AP. Transcript level and sequence determinants of protein abundance and noise in *Escherichia coli*. Nucleic Acids Res. 2014;42: 4791–4799.

21. Hausser J, Mayo A, Alon U. Central dogma rates and the trade-off between precision and

economy. 2018. doi:10.1101/276139.

22. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11: R106.

23. Opdam S, Richelle A, Kellman B, Li S, Zielinski DC, Lewis NE. A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models. Cell Syst. 2017;4: 318–329.e6.

24. Sullivan A, Purohit P, Freese NH, Pasha A, Esteban E, Waese J, et al. An “eFP-Seq Browser” for Visualizing and Exploring RNA-Seq Data. Plant J. 2019; doi:10.1111/tpj.14468.

25. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis?. Nature Biotechnology. 2010. pp. 245–248. doi:10.1038/nbt.1614.

26. Schellenberger J, Palsson BØ. Use of randomized sampling for analysis of metabolic networks. J Biol Chem. 2009;284: 5457–5461.

27. Kaufman DE, Smith RL. Direction Choice for Accelerated Convergence in Hit-and-Run Sampling. Operations Research. 1998. pp. 84–95. doi:10.1287/opre.46.1.84.

28. Keaty TC, Jensen PA. gapsplit: Efficient random sampling for non-convex constraint-based models. doi:10.1101/652917.

29. de Vladar HP. Amino acid fermentation at the origin of the genetic code. Biol Direct. 2012;7: 6.

30. Lacey RW. Basic medical microbiology (4th edition). Journal of Hospital Infection. 1992. pp. 135–136. doi:10.1016/0195-6701(92)90123-4.

31. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, et al. A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism--2011. Mol Syst Biol. 2011;7: 535.

32. Feist AM, Palsson BØ. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. Nat Biotechnol. 2008;26: 659–667.

33. Lybecker M, Zimmermann B, Bilusic I, Tukhtubaeva N, Schroeder R. The double-stranded transcriptome of *Escherichia coli*. Proc Natl Acad Sci U S A. 2014;111: 3134–3139.

34. Monk JM, Koza A, Campodonico MA, Machado D, Seoane JM, Palsson BO, et al. Multi-omics Quantification of Species Variation of *Escherichia coli* Links Molecular Features with Strain Phenotypes. Cell Syst. 2016;3: 238–251.e12.

35. Unden G, Becker S, Bongaerts J, Schirawski J, Six S. Oxygen regulated gene expression in facultatively anaerobic bacteria. Antonie Van Leeuwenhoek. 1994;66: 3–22.

36. Adadi R, Volkmer B, Milo R, Heinemann M, Shlomi T. Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. PLoS Comput Biol. 2012;8: e1002575.

37. Murashko ON, Lin-Chao S. responds to environmental changes using enolasic degradosomes and stabilized DicF sRNA to alter cellular morphology. Proc Natl Acad Sci U S A. 2017;114: E8025–E8034.

38. Stincone A, Prigione A, Cramer T, Wamelink MMC, Campbell K, Cheung E, et al. The return of metabolism: biochemistry and physiology of the pentose phosphate pathway. Biological Reviews.

2015. pp. 927–963. doi:10.1111/brv.12140.

39. Thakur CS, Brown ME, Sama JN, Jackson ME, Dayie TK. Growth of wildtype and mutant *E. coli* strains in minimal media for optimal production of nucleic acids for preparing labeled nucleotides. Appl Microbiol Biotechnol. 2010;88: 771–779.

40. Yeh JI, Chinte U, Du S. Structure of glycerol-3-phosphate dehydrogenase, an essential monotopic membrane enzyme involved in respiration and metabolism. Proc Natl Acad Sci U S A. 2008;105: 3280–3285.

41. Stubbendieck RM, Vargas-Bautista C, Straight PD. Bacterial Communities: Interactions to Scale. Frontiers in Microbiology. 2016. doi:10.3389/fmicb.2016.01234.

42. Poulsen LK, Licht TR, Rang C, Krogfelt KA, Molin S. Physiological state of *Escherichia coli* BJ4 growing in the large intestines of streptomycin-treated mice. J Bacteriol. 1995;177: 5840–5845.

43. Sezonov G, Joseleau-Petit D, D'Ari R. *Escherichia coli* Physiology in Luria-Bertani Broth. Journal of Bacteriology. 2007. pp. 8746–8749. doi:10.1128/jb.01368-07.

44. Suda M. Intestinal absorption of amino acids. Gastroenterologia Japonica. 1966. pp. 5–6. doi:10.1007/bf02775092.

45. Nakayama T, Yonekura S-I, Yonei S, Zhang-Akiyama Q-M. *Escherichia coli* pyruvate:flavodoxin oxidoreductase, YdbK - regulation of expression and biological roles in protection against oxidative stress. Genes Genet Syst. 2013;88: 175–188.

46. Kafri M, Metzl-Raz E, Jona G, Barkai N. The Cost of Protein Production. Cell Reports. 2016. pp. 22–31. doi:10.1016/j.celrep.2015.12.015.

47. Li G-W, Burkhardt D, Gross C, Weissman JS. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell. 2014;157: 624–635.

48. Fang Z, Martin J, Wang Z. Statistical methods for identifying differentially expressed genes in RNA-Seq experiments. Cell Biosci. 2012;2: 26.

49. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Syst Biol. 2013;7: 74.

50. X Y, Yang X, Hutson AD, Wang D. Nonparametric Estimation of Quantile and Quantile Density Function. Journal of Biometrics & Biostatistics. 2017. doi:10.4172/2155-6180.1000356.

51. Mahadevan R, Schilling CH. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab Eng. 2003;5: 264–276.

52. Joshi NA, Fass JN. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33). Available at https://github.com/najoshi/sickle.

53. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9: 357–359.

54. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009. pp. 2078–2079. doi:10.1093/bioinformatics/btp352.

55. King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, et al. BiGG Models: A platform for

integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. 2016;44: D515–22.

56. Dixon P. VEGAN, a package of R functions for community ecology. Journal of Vegetation Science. 2003. p. 927. doi:10.1658/1100-9233(2003)014[0927:vaporf]2.0.co;2.

57. Breiman L. Machine Learning. 2001. pp. 261–277. doi:10.1023/a:1017934522171.

51. Joshi NA, Fass JN. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33). Available at https://github.com/najoshi/sickle.