Full Title: Parsimonious transcript data integration improves context-specific predictions of bacterial metabolism in complex environments

Short Title: Parsimonious Transcript Integration into Genome-scale Network Models

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 fundamental shift 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

82 platforms have allowed for guided genetic engineering efforts, directed hypothesis generation for
83 downstream laboratory testing, and deconvoluted metabolic responses of bacteria to antibiotic stress
84 [5,6].

85 GENREs have also become strong platforms for providing context to multiple forms of high-
86 throughput omics data, allowing for the study of whole-network changes in metabolism due to
87 upstream/downstream shifts in enzyme activity [7]. To this end, advances over the last decade in RNA-
88 Seq technologies have indeed enabled whole bacterial community metatranscriptomic analysis [7].
89 Conversely, changes identified through other means occasionally can be difficult to attribute to a single
90 organism from complex biological samples (i.e. untargeted metabolomics). However, shotgun
91 transcriptomic analysis followed by strict read mapping protocols allows for the characterization of
92 individual groups of species within these communities of organisms [8,9]. These large datasets provide
93 informative possibilities for analyses; however holistic (top down) analyses are limited in their ability to
94 make any mechanistic claims outside of correlative results. Against the backdrop of the reductionist
95 (bottom up) approach of GENREs, their combined informative capacity allows for mechanistic insight
96 into the metabolic status and ecological interactions within the microbiome [10].

97 Previous approaches at integrating transcriptomic data into GENREs have primarily focused on
98 identifying reactions associated with arbitrary cutoffs of gene transcription, then maximizing flux through
99 those reactions with many transcripts or minimizing flux through those reactions with few transcripts
100 [11–13]. Recently, parsimonious Flux Balance Analysis (pFBA) was developed, which seeks to
101 minimize the flux associated with each reaction in the model, while maintaining optimum flux through
102 the objective [14]. In essence, this approach identifies the least biologically “expensive” usage of an
103 organism’s metabolism to achieve high growth rates. This method is in line with the concept that
104 evolutionary pressures have selected for metabolic states in microbes with minimized cellular cost that
105 maximize growth rate under a variety of environmental conditions [50]. Surprisingly, pFBA alone was
106 found to be able to largely outperform other methods in predicting metabolic network behavior under
107 various growth conditions despite not utilizing any transcriptomic data for these predictions [15]. These
108 studies demonstrate that contextualizing data in a truly biological meaningful way has proven difficult,

109 highlighting large gaps in our understanding of metabolic regulatory mechanisms. Moreover, recent
110 studies on *E. coli* grown in defined *in vitro* conditions have shown that levels of transcript were only
111 predictive for roughly half of the concentrations of the corresponding enzymes in the cell [16,17].
112 Despite this shortcoming, transcript abundances remain an indication of cellular investment into a given
113 metabolic strategy. Furthermore, as expression of a functional enzyme is not as simple as an on-or-off
114 scenario, we have also implemented an unsupervised procedure to discretize levels of transcription
115 based on the distribution inherent to RNA-Seq data [18]. This approach is unbiased by depth of
116 sequencing across datasets, and it restricts the utility of low-transcription reactions to an optimal pFBA
117 solution yet does not entirely prohibit it. With both of these concepts in mind, we have leveraged the
118 insights gained by pFBA, integrating transcript abundance as a weighting metric to direct the most
119 parsimonious solution toward a state of higher fidelity to *in situ* biology.

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

136 **Results and Discussion**

137 **Reaction Inclusion by Parsimony and Transcript Distribution (RIPTiDe)**

138 RIPTiDe relies on bi-level optimization guided by transcript abundances in order to identify the
139 most efficient usage of metabolism that also reflects an organism's transcriptional state. The required
140 input for RIPTiDe includes a metabolic network reconstruction with an assigned objective and gene-
141 associated transcript abundances. A list of known cellular tasks that should be included/excluded may
142 also be provided if *a priori* knowledge exists about the organism's behavior in the environment of
143 interest. To generate distinct and reproducible delineations of transcription levels for genes, we utilized
144 a quantile approach of the negative-binomial transcript distribution inherent in RNA-Seq data (Fig. 1A)
145 [18]. Transcript mapping is reduced to only genes that appear in the genome-scale network
146 reconstruction of interest, therefore not skewing the distribution with data that is uninformative to the
147 current analysis. Breaking down levels of transcription into more discrete categories of activity is similar
148 to previous approaches [13]. However, in the case of RIPTiDe higher transcript abundance
149 corresponds to a higher probability of being included in the final metabolic model. This approach
150 removes the need for different arbitrary cutoffs between datasets and reduces the need for manual
151 intervention in the generation of functional context-specific models.

152 As with pFBA, reversible reactions are split into pairs of irreversible reactions carrying non-
153 negative flux. In this way, we were able to minimize overall flux subject to the transcript-guided linear
154 coefficients added by RIPTiDe. Within the transcript abundance distribution, progressively smaller
155 linear coefficients are assigned to genes with increased transcription at each cutoff (Fig. 1A).
156 Coefficients are then subsequently transferred to the corresponding metabolic reactions for each gene.
157 The magnitude of linear coefficient drives the amount of flux carried by each reaction. Larger
158 coefficients correspond to the reaction receiving larger flux during optimal growth, which would
159 therefore be favored less when minimizing the overall flux through the objective function. Reactions
160 without genes are assigned the median score to not be incorrectly penalized, nor overly biased to their
161 inclusion. Importantly, this approach is equally tractable between different sequencing depths as well as
162 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 [19]. 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 [20,21]. 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. 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. Optimal flux through the objective function is set as a constraint, allowing ≧80% of the optimum
2. Linear coefficients are assigned as constraints to each reaction based on transcriptomic abundance of the associated gene-enzyme assignments. In the 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 previous approaches, models produced by RIPTiDe are able to achieve flux through their given objectives by default without any additional gap filling. This is critical as the principle of optimum cellular

190 economy with respect to the transcript data is never broken throughout the contextualization process.
191 For a more detailed description of the algorithm, refer to the Methods section.

192 **Application of RIPTiDe to an Example Model of Bacterial Metabolism**

193 To test RIPTiDe, we created a basic metabolic network reconstruction with two compartments
194 and generation of ATP as the cellular objective function. The model has two separate means of
195 accomplishing this goal; either through glycolysis where glucose is the primary substrate, or Stickland
196 fermentation which requires pairs of amino acids in coupled redox reactions to create energy [22]. In
197 nature, fewer ATP are generated per unit of substrate in Stickland fermentation compared to glycolysis
198 [23]. This phenomenon was replicated in the example model as fewer reactions are required to
199 generate ATP from glucose, making this pathway more energetically favorable (Table S1). Utilizing the
200 base implementation of pFBA available within COBRApy to quantify which pathway was the more
201 parsimonious solution, we ensured that glycolysis was actually the most efficient (fewest enzymes
202 used) for optimal ATP generation (Fig. 2A). Critically, this ATP generation was performed under the
203 conditions where all possible extracellular substrates were provided in excess to the model as to not
204 bias results only on the basis of resource availability.

205 As a means to determine the context specificity of RIPTiDe output models, we then generated
206 two distinct simulated transcript abundance distributions that each reflected elevated transcription of a
207 single pathway (Table S1). In Fig. 2B, RIPTiDe was provided with a transcriptome in which genes for
208 glycolysis and related reactions were more highly transcribed. The contextualized model in this case
209 also utilized glycolysis as the only means of generating ATP, directly matching those results obtained
210 from pFBA. Alternatively, as illustrated in Fig. 2C, when genes associated with Stickland fermentation
211 and peptide substrate transporters were more highly transcribed, RIPTiDe was able to discern that the
212 less directly parsimonious pathway should be utilized instead when provided sufficient transcriptomic
213 evidence. In each case, reactions related to the opposite pathway were correctly pruned while
214 generating the final context-specific model. Furthermore, flux samples for the cellular objective from
215 each contextualized example model revealed that significantly more ATP could be produced when
216 employing glycolysis (Table S1), further supporting the energetic favorability of this pathway. These

217 combined results confirmed that RIPTiDe was indeed able to use a less parsimonious pathway when
218 necessary to better reflect the associated transcriptomic dataset.

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

220 To perform a thorough evaluation of our approach, we utilized the genome-scale metabolic
221 network reconstruction iJO1366, for the model organism *E. coli* strain K-12 substr. MG1655 [24]. This
222 GENRE has been the subject of extensive manual curation and experimental validation, making it the
223 best possible candidate for examining new approaches in constraint-based modeling [25]. Leveraging
224 these years of research, we utilized this well described system to determine the biological validity of
225 predicted context-specific differences returned by RIPTiDe. To accomplish this goal, we performed a
226 comparative analysis using 3 separate transcriptomic sequencing datasets, each derived from well
227 characterized *in vitro* media conditions including LB aerobic liquid rich media, M9+glucose aerobic
228 liquid minimal media, and M9+glucose anaerobic liquid minimal media. Each dataset was collected at
229 approximately the same mid-exponential phase of growth [26,27]. These data were chosen due to the
230 large amount of characterization each has received over time, allowing our predictions to be more
231 readily supported compared to assessing metabolic differences possibly seen across numerous less
232 well explored conditions. Additionally, *E. coli* is a facultative anaerobe with known differences in
233 metabolism between these lifestyles that we hoped to capture with RIPTiDe [28]. For all subsequent
234 work, we took an unsupervised approach by allowing for unpenalized flow of all extracellular
235 metabolites in and out of the system (exchange reaction bounds set to ±1000). This approach allowed
236 for inferences about the most likely growth substrates for a contextualized GENRE since only those that
237 carry flux following RIPTiDe remain.

238 We first set a performance baseline with two commonly utilized platforms for integration of
239 transcriptomic data with GENREs, GIMME and iMAT [11,12]. Both algorithms require a lower
240 abundance cutoff for selectively deactivating reactions associated with genes that were transcribed
241 below the desired level. As such, we utilized the Q50 value at the bottom of the default RIPTiDe
242 abundance distribution settings as this setting for GIMME and iMAT. In addition to the lower cutoff,
243 iMAT requires an abundance threshold for genes that have very high transcription. For this iMAT

244 threshold we used the highest default category from RIPTiDe of transcripts with an abundance above
245 Q87.5. Both approaches initially returned models that were unable to achieve flux through the biomass
246 synthesis reaction in any condition, and required additional gap filling to re-establish the ability of the
247 models to grow (Table S2). However, after this gap-filling step the resulting models were no longer
248 concordant with the input transcriptomes, reintroducing 138 and 104 reactions respectively associated
249 with genes that fall below the defined transcription Q50 cutoff (Table S2). These results in combination
250 with the previously mentioned performance of pFBA alone highlight the advantages of RIPTiDe over
251 other established methods of transcriptomic data integration.

252 **Context-specific *Escherichia coli* Biology Uncovered by RIPTiDe**

253 We next performed RIPTiDe-based analyses of iJO1366, utilizing the same transcriptome data
254 described previously. As a standard of comparison, we also implemented a more standard version of
255 the pFBA algorithm where the sum of fluxes is minimized while maintaining a maximum objective flux,
256 to establish the most "economic" overall flux distribution that achieves maximum growth. Different
257 topologies of reaction and metabolite inclusion were found among each of the contextualized models;
258 however none were completely unique and a core set of metabolic reactions that were necessary
259 across all growth conditions was shared (Fig. 3A). These distinctions in network topology were further
260 crystallized by conditional gene essentiality between the RIPTiDe-contextualized models (necessary for
261 at least 1% of optimum growth). Although most essential genes were shared by all four models (187
262 genes, Table S3), those essential genes that were context-specific primarily related to pathways of
263 biological significance to their respective media conditions (Fig. 3B). As a corollary to creating a smaller
264 metabolic network than those using transcriptomes contextualized by RIPTiDe (by an average of
265 9.12%), fewer genes were distinctly essential in the model resulting from standard pFBA (5 genes)
266 compared to any of the RIPTiDe-contextualized *in vitro* transcriptomes (8, 15, and 20 genes
267 respectively; Table S3). Interestingly, the contextualized models on average shared larger ratios of
268 components (~7.07% reactions and ~1.78% metabolites) with each other than with pFBA, supporting
269 that RIPTiDe selects more biologically relevant patterns of metabolism than pFBA alone (Fig. 3A).
270 Collectively, these results indicate that our algorithm correctly identifies context-specific network

271 topologies given an associated transcriptome.

272 We also calculated the ranges of relative doubling time for each of the models from the flux
273 samples associated with the biomass objective function [29]. The base model with no additional flux
274 sum minimization constraints displayed much more variation in the amount of flux that is possible
275 through the biomass reaction compared to any of the other, more constrained models. It is important to
276 note that achieving higher flux values, as in the unconstrained model, is still possible in the constrained
277 and contextualized models however the number of specific distributions that reach these levels have a
278 very low probability of ever appearing and would likely only do so after many more sampled flux
279 distributions. This outcome speaks to the truly context-specific nature of the output models that better
280 reflect actual biological trends. Interestingly, growth from the pFBA-generated model closely mirrored
281 those from the RIPTiDe-contextualized model in LB rich media (Fig. 3C). This similarity signified that
282 rapid doubling time in rich media identified by RIPTiDe is analogous to the most mathematically optimal
283 growth solution, which was in line with many predictions in the original publication of pFBA [14]. We
284 went on to compare experimentally measured doubling times of *E. coli* K-12 in liquid LB [35], and found
285 that the most frequent doubling times for both RIPTiDe and pFBA were slower than those measured in
286 the laboratory. Another interesting finding was that the computed doubling times for contextualized
287 models from both aerobic & anaerobic M9+glucose minimal medias were significantly slower than in
288 rich media and analogous to previously measured experimental values (Fig. 3C) [30]. These results
289 suggested the RIPTiDe-contextualized GENREs not only topologically reflected known biological
290 trends, but also phenotypically mimicked experimentally measured behaviors.

291 We subsequently focused our analysis on assessing known differences in core metabolism
292 across transcriptomes in RIPTiDe-contextualized models. Growth conditions are known to affect which
293 forms of metabolism are possible or favored at a given time, driving the cell to utilize alternative
294 pathways for replenishing it's cellular currency. One such anabolic circuit is the pentose phosphate
295 pathway, which is a critical component of core metabolism for most life that combats oxidative stress as
296 well as maintains carbon homeostasis through provision of NADPH and 5-carbon sugars [31]. The
297 pathway consists of two phases, oxidative (NADPH generating) and non-oxidative (pentose

298 generating), and has been previously shown to be differentially activated in wildtype *E. coli* during
299 aerobic growth in M9 minimal media (favors oxidative) and LB rich media (favors non-oxidative) [32].
300 Strikingly, the RIPTiDe-contextualized model in aerobic M9 minimal media indeed only contained the
301 oxidative component of the pentose phosphate pathway (Fig. 4A). Alternatively, only the non-oxidative
302 component was present in the model resulting from growth in LB, instead generating more of its
303 NADPH from the TCA cycle (Fig. 4B). It is important to note also that in both growth conditions a
304 primary source of NADPH was glycerol-3-phosphate dehydrogenase, with significantly more activity
305 captured in flux samples from contextualized M9 growth (Fig. S1A). This reaction is a major contributor
306 to the bacterial electron transport chain and an essential part of energy generation in all known growth
307 conditions [33]. These results signified that RIPTiDe is able to capture previously measured, context-
308 specific behaviors with only transcriptomic characterization.

309 **Application of RIPTiDe With *In vivo* Metatranscriptomic Data**

310 Finally, we sought to test the utility of RIPTiDe for capturing metabolic behavior of bacteria from
311 within host-associated communities of microorganisms. Computational analysis of a discrete bacterial
312 species within a complex environment as a member of a larger microbial community has been
313 historically difficult [51]. To address this problem, we utilized *in vivo* metatranscriptomic abundance data
314 collected from the cecum of wildtype C57Bl6 mice. These data were chosen because the specific
315 antibiotic pretreatment (intraperitoneal injection of clindamycin) resulted in a bacterial community
316 composed of >90% *E. coli [8]*. To begin the analysis, we first made certain that the *in vivo* transcript
317 abundance distribution reflected the same negative binomial type as those derived from *in vitro*
318 transcriptomes mapped to the same *E. coli* K-12 MG1655 genome (Fig. 5A). This result demonstrated
319 that with sufficient sequencing depth, a species with a GENRE may be analyzed with RIPTiDe using
320 data collected from among a diverse collection of microbes. After contextualization of this data we first
321 found that calculated growth rates from biomass flux samples fell within expected ranges of those
322 experimentally observed *in vivo* (Fig. S1B) [34]. We then assessed how distinctive the flux distributions
323 were of those metabolic reactions which were shared between context-specific models. Cross-
324 referencing each contextualized model to identify shared reactions, we then performed principle

325 component analysis using Bray-Curtis dissimilarity for those flux samples corresponding with this core
326 metabolic subnetwork (Fig. 5B). This analysis revealed that although numerous reactions and
327 metabolites are shared between context-specific models, they are utilized significantly differently to
328 manage growth across media conditions and supported differential metabolism is likely necessary to
329 cope with distinct environments.

330 Next, we contrasted the RIPTiDe models generated using *in vivo* or LB rich media
331 transcriptomes. Aside from one environment being highly aerobic and the other largely anaerobic,
332 previous research on *E. coli* physiology during growth in liquid LB concluded that the bacterium utilizes
333 amino acids that are simple to metabolize for growth early on, then switches to other substrates in a
334 diauxic-like growth pattern. LB broth has also been shown to have only modest levels of useable
335 carbohydrate growth substrates, increasing the importance of amino acid catabolism for *E. coli* [35].
336 Additionally, *in vivo* amino acids and simple sugars are contested resources in the cecum as the
337 majority are already absorbed by the host or used by other microbes upstream in the GI tract, making
338 these growth conditions extremely different [36]. As an analogy to differential growth substrates, we
339 looked to assess the presence of context-specific exchange reactions. We identified 7 exchange
340 reactions that were present in the LB-associated model but not *in vivo*, and 5 of these were peptides or
341 peptide conjugates (Fig. 5C). Carbohydrate growth substrates have been demonstrated to be at very
342 low concentrations in LB media; *E. coli* therefore primarily utilizes amino acids to generate energy
343 during early exponential phase [35]. Indeed, upon assessing the measured changes in the *in vivo*
344 concentrations of the highlighted metabolites they were reduced in the context where *E. coli* dominated
345 the environment, possibly due to consumption by the microbe (Fig. S2). Additionally, the exchange for
346 glucose-6-phosphate was also only found in LB-contextualized models, reinforcing that lowly abundant
347 sugar phosphates are utilized secondarily for growth by *E. coli* compared to amino acid substrates in
348 LB media [35]. Interestingly, fewer exchange reactions associated with the *in vivo* context may also
349 imply that a greater amount of metabolite substrates must be produced intracellularly instead of
350 imported from the environment. Continuing this analysis now with shared intracellular reactions we
351 utilized Random Forest on the sampled flux distributions for each reactions to determine those

352 metabolic functionalities that most distinguish the two models (Fig. 5D). Among other findings, machine
353 learning revealed consistently high flux ranges for flavodoxin reductase in aerobic LB culture, but not *in*
354 *vivo*. This enzyme is important for dealing with oxidative stress which would be much more prevalent in
355 the broth culture than in the mostly anoxic environment of the mammalian gut [37]. Altogether, our
356 results support that RIPTiDe is also a strong platform for the analysis of bacterial GENRE behavior in
357 the context of complex microbial communities through the use of metatranscriptomic data.

358 **Conclusions**

359 In order to survive in a given environment, an organism must express the appropriate forms of
360 metabolism to cope with changing conditions. Production of transcript and proteins is metabolically
361 expensive, and over-production likely puts the cell at risk if further perturbation is placed on the system
362 [38]. Although recent findings suggest that levels of transcript for a given enzyme product do not always
363 directly correlate with the amount of functional protein present in the cell at a given time [16], these
364 values may be considered a signature of cellular "investment" and therefore may still be useful in
365 interrogation of metabolism. Furthermore, understanding the quantity of even a small number of
366 functional enzymes remains extremely challenging, which negatively impacts our ability to grasp which
367 and why resources are used by the cell [39]. We should then consider transcriptomic abundance data
368 as signatures of cellular energy expenditure where high transcript levels for a particular gene are an
369 indication of large cellular investment related to that particular function, but are not particularly
370 deterministic of the associated phenotype. Parsimonious Flux Balance Analysis approaches the idea of
371 maximizing cellular economy, and we have built on this idea but employing RNA-Seq data to inform
372 where the cell is most likely spending its energy.

373 We have demonstrated empirically that our algorithm correctly calculated both *in vitro* and *in*
374 *vivo* metabolic phenotypes with reasonably high accuracy. In addition to these predictions, our method
375 does not require prior knowledge of active metabolic subnetworks which can preclude work with
376 understudied or uncultured species that cannot or have not been grown or characterized in isolation.
377 Some possible limitations of RIPTiDe do exist however. First, as RIPTiDe relies on the distribution
378 associated with shotgun transcriptomic sequencing, it is less suited for other types of omic data in its

379 current form [40]. RNA-Seq reads are easily attributable to single species within the microbiota due to
380 strict genomic mapping techniques and are measured within a single sequencing assay, however with
381 sufficient consideration to reaction weights given other data types RIPTiDe's functionality could be
382 expanded. Second, by condensing to such a small metabolic network "solution space", these models
383 may have a risk of overfitting. To counteract this, we provided options in the python implementation of
384 the method to shift the quantile transcript abundance thresholds as well as for reducing the degree of
385 weighting for reactions as needed. In spite of these concerns, we have shown our algorithm to be a
386 strong approach for contextualization of genome-scale metabolic network reconstructions for both *in*
387 *vitro* and *in vivo* conditions, producing models that closely mirror known biological phenotypes. The
388 implementation of RIPTiDe rewards divergence from the most numerically parsimonious state to one
389 that is more reflective of cellular investment, and in turn more illustrative of *in situ* biology. In summary,
390 RIPTiDe approaches transcriptomic data integration uniquely from previous techniques and provides a
391 more biologically relevant platform for considering the cellular behavior in complex environments.

392 **Methods**

393 **Parsimonious Flux Balance Analysis Adaptation**

394 RIPTiDe was implemented within COBRApy [41] and is freely available with installation
395 instructions at https://github.com/mjenior/riptide. To save on computational time, which can be intensive
396 during flux sampling object generation, flux variability analysis was utilized first to identify and remove
397 all blocked reactions (those reactions which do not carry flux under any circumstances). At this point,
398 optimal flux through the objective function was set as a constraint with a 20% range of allowed
399 variation. As previously described, remaining reactions were then assigned a linear coefficient based
400 on the user provided transcriptomic abundance data [42]. For reaction weight (linear coefficient)
401 assignments, the position of transcript abundance level for gene x within the sorted transcript
402 abundance distribution T_d that corresponds to the percentile of interest p, resulting in abundance cutoff
403 q_p for that quantile. The correct linear coefficient is then taken from the sorted coefficient distribution C,
404 providing the weight w_g factor for all reactions associated with gene g. Finally, linear coefficient w_r is
405 assigned to each reaction in the model based on the gene-reaction lookup table R_g. Reactions in the

model without genes are assigned the median value provided in C.

$$
\begin{aligned}
&min \sum^{r} v_{irrev} \cdot w_r \\
&s.t. \quad max\, v_{biomass} = v_{biomass,lb} \\
&\qquad S_{irrev} \cdot v_{irrev} = 0 \\
&\qquad 0 \leq v_{irrev} \cdot w_r \leq v_{max} \cdot w_r
\end{aligned}
$$

As in pFBA, 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 ($\max v_{biomass}$) and mass-balance constraints. Optimization (standard FBA) is performed and inactive reactions are removed. Models were then checked for orphaned genes and metabolites, which are subsequently removed as well.

Flux Sampling

In order to maximize intracellular flux through the remaining reactions, previous coefficients were converted to their inverse within the supplied range:

$$
\begin{aligned}
&r_w = min(C) + max(C) \\
&w_{rx} = \{ w_r : r_w - w_r \}
\end{aligned}
$$

The total range of reaction weights r_w is determined, then the difference with the current linear coefficient for each reaction w_{rx} is assigned in place of the original. 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):

$$
max \sum^{r} v_{irrev} \cdot w_{rx}
$$

The new coefficients are then assigned to the corresponding irreversible reactions v_{irrev}. Flux sampling was then performed with the Artificial Centering Hit-and-Run sampler with standard parameters and 10,000 samples [21]. If the constrained solution space is now too narrow for flux sampling to correctly initiate which could occur only in very small network reconstructions, flux variability analysis is instead implemented to establish flux ranges for each reaction [43].

Example Model Construction

429 The example model used for initial RIPTiDe testing consists of 16 reactions (7 exchanges, 7
430 transporters, and 2 metabolic) and 14 metabolites across 2 compartments (extracellular and
431 intracellular). The intracellular demand reaction for ATP generation was utilized as the objective.
432 Simulated transcriptomes are shown in Table S1 and an SBML file (example_GENRE.sbml) for the
433 model is also available at github.com/mjenior/Jenior_RIPTiDe_2019.

434 **Data Sources and Processing**

435 Fastq files associated with each study were accessed from the NCBI Sequence Read Archive
436 (SRA: PRJNA313438, SRA: PRJNA212627, & SRA: PRJNA415307). Raw cDNA reads were quality-
437 trimmed with Sickle [44] at ≧Q30 and subsequently mapped to *Escherichia coli* K-12 MG1655
438 genes (EMBL-EBI on Jan 28, 2019) using Bowtie2 [45]. MarkDuplicates was then used to remove
439 optical and PCR duplicates (broadinstitute.github.io/picard/). The remaining mappings were converted
440 to idxstats format using SAMtools [46], and the read counts per gene were tabulated. Discordant pair
441 mappings were discarded, and counts were then normalized to read length and gene length to give a
442 per-base report of gene coverage. If replicate sequencing was performed, median transcription values
443 for each gene were then calculated. The data were then converted to a python dictionary of gene keys
444 and associated numeric transcript abundances for use with RIPTiDe. *In vivo* untargeted metabolomic
445 data were generated for a previous study [8], and accessed from
446 github.com/SchlossLab/Jenior_Metatranscriptomics_mSphere_2018/. The current version of iJO1366
447 (GENRE for *E. coli* K-12 MG1655) was accessed from the BiGG Model database [47] on 09-26-2018.

448 **Statistical Methods**

449 All statistical analysis was performed using R (v.3.2.0). Significant differences between flux
450 distributions and metabolite concentrations were determined by Wilcoxon signed-rank test with
451 Benjamini-Hochberg correction. Principle Component Analysis of sampled flux distributions was
452 performed using the vegan package [48], and machine-learning was accomplished with the base
453 implementation of Random Forest in R [49].

454 **Data and Software Availability**

455 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. 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).

Author Contributions

M.L.J. conceived of the study, authored code, performed analysis, and drafted the manuscript. T.J.M. contributed to algorithm development and edited the manuscript. J.P. contributed to algorithm development and edited the manuscript.

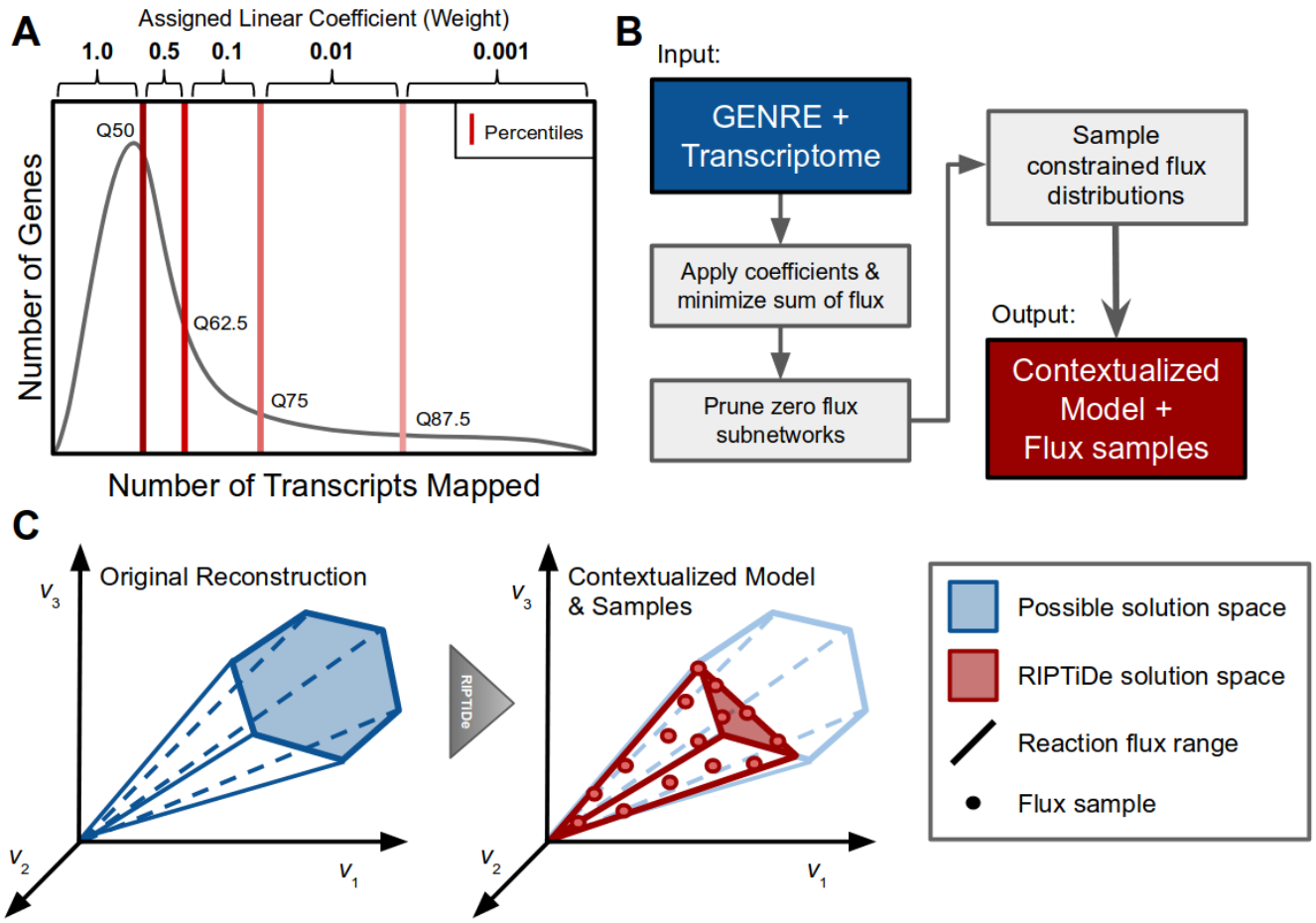
Acknowledgements

The authors would like to thank 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. Funding for this project was provided by a grant from the U.S. National Institute of Health (1R01AT010253-01).

Competing Interests

The authors declare that no competing interests exist.

483 **Figures and Legends**



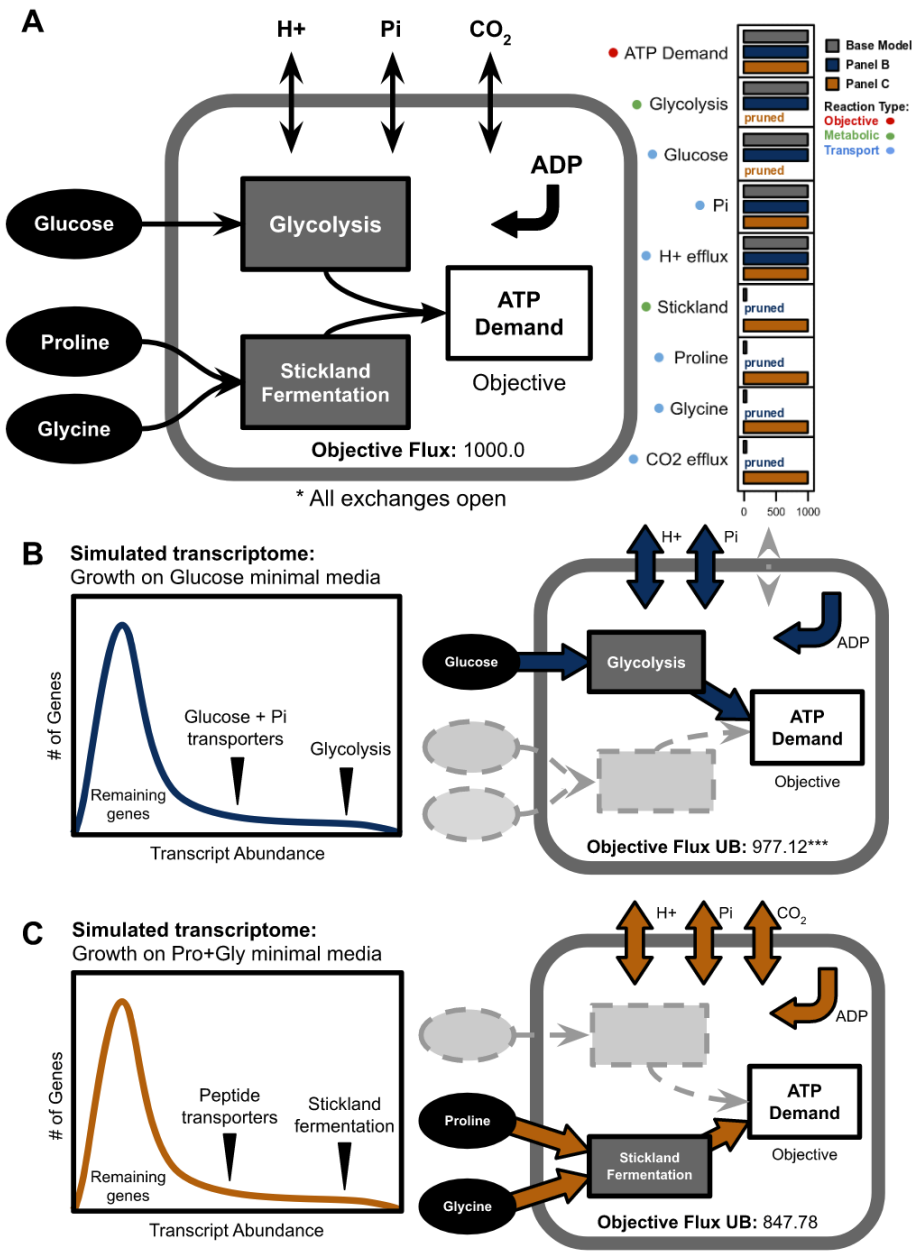
484

485 **Figure 1 | Visual representation of RIPTiDe workflow. (A)** Example transcript distribution, with
486 default quantile divisions shown next to each red line. The resulting abundance ranges are then
487 assigned a linear coefficient, shown along the top, to be used during subsequent analysis. **(B)**
488 Schematic outlining computation performed during each step of RIPTiDe. The platform requires two
489 input data structures: a genome-scale network reconstruction and transcript abundances associated
490 with genes. During initialization totally blocked reactions are identified and removed, and transcript
491 abundances for genes are transferred to their corresponding reaction allowing coefficients to be
492 assigned. Each reversible reaction is now also separated into pairs of irreversible reactions. Applying
493 principles of parsimonious FBA to the constrained model, a minimum sum of fluxes optimization is
494 performed with respect to a predefined minimum flux through the cellular objective. Reactions that no
495 longer carry flux in this state are pruned, and flux sampling is performed on the intermediate model to
496 determine context-specific bounds for the remaining reactions. **(C)** Pictorial representation of the impact

497 of RIPTiDe on the original reconstruction, converting it to a context-specific model of metabolism with

498 as parsimonious a possible metabolic solution space with respect to the given transcriptomic data.

499



500

501 **Figure 2 | Example model of bacterial metabolism supports the utility of RIPTiDe for identifying**

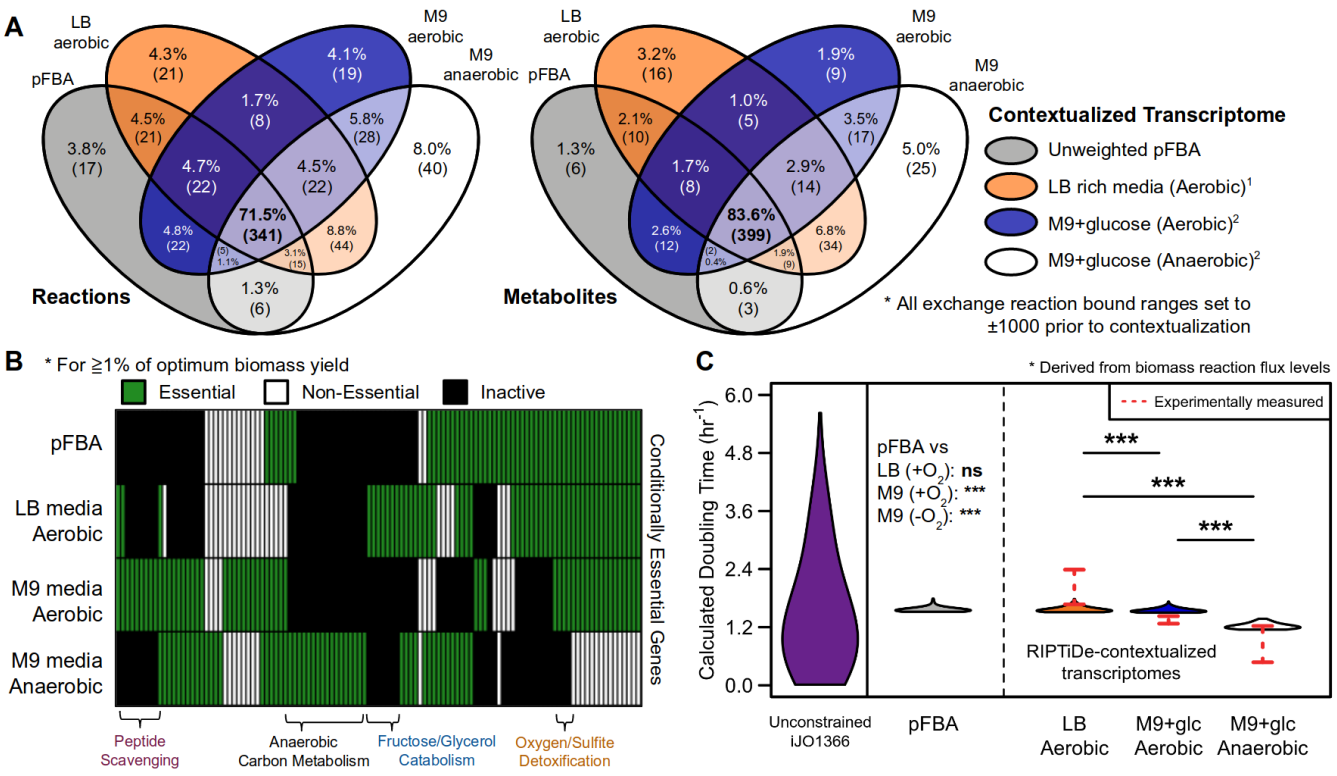
502 **most likely context-specific strategies. (A)** General topology of simplified GENRE that consists of 16

503 reactions and 14 metabolites in total (Table S1). The objective of ATP generation can be achieved

504 through two separate means: either the catabolism of glucose via glycolysis or through the paired

505 fermentation of proline and glycine. Glycolysis is the more parsimonious pathway for generating ATP,

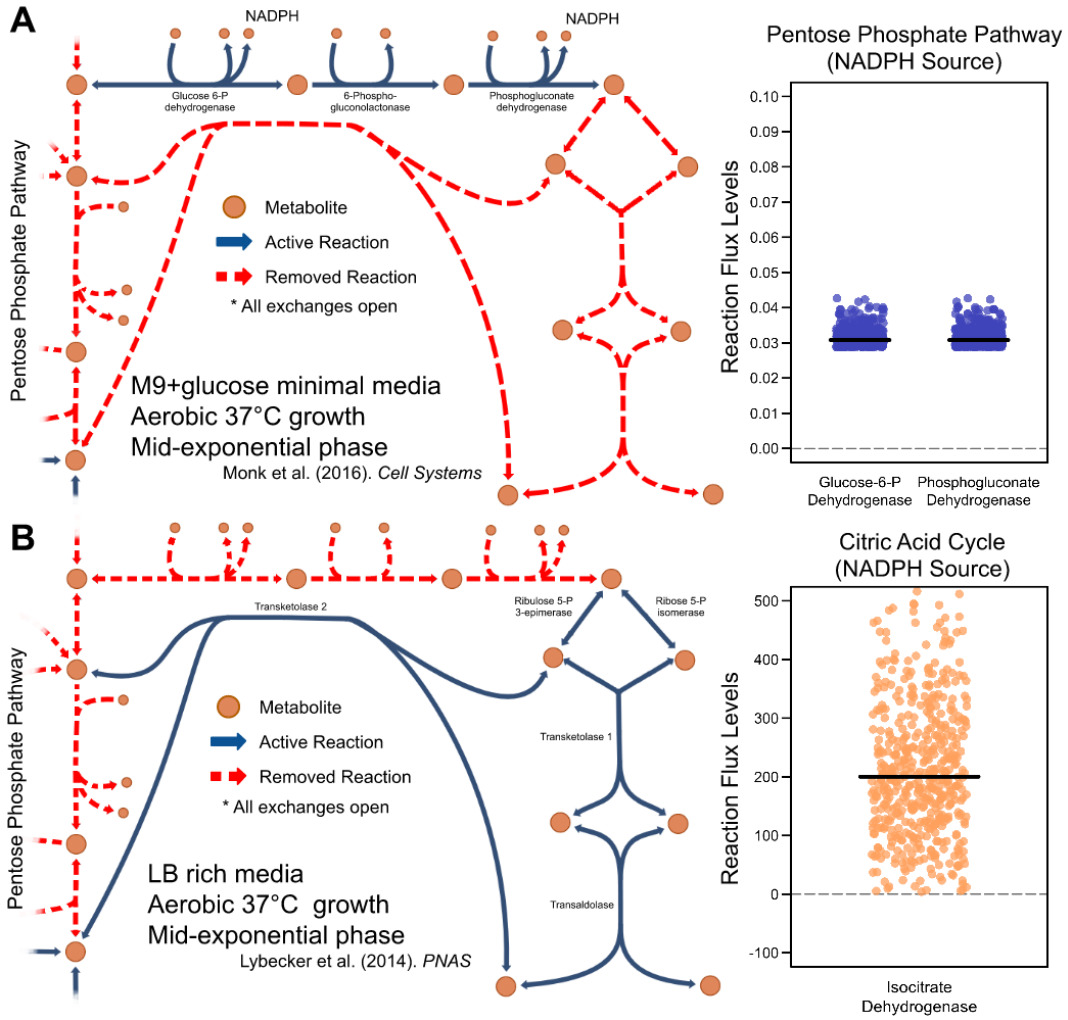
506 which is reflected in the pFBA solution fluxes to the right of the diagram. **(B)** RIPTiDe is able to correctly
507 identify the most likely route of flux when provided a transcriptome. **(C)** When provided with
508 transcriptomic evidence, RIPTiDe is able to identify a less parsimonious, but more concordant solution.
509 Despite very low transcription, RIPTiDe still identifies the necessity for carbon dioxide efflux during
510 fermentation to maintain mass balance. Significant differences in cellular objective flux samples were
511 determined by Wilcoxon signed-rank test with Benjamini-Hochberg correction (***, p-value ≦ 0.001).
512



513

514 **Figure 3 | RIPTiDe identifies established biological differences in *E. coli* str. K-12 substr.**
515 **MG1655 across *in vitro* growth conditions**. Without increased user supervision, our algorithm
516 correctly predicts behavior and context-specific pathways for *E. coli* str. K-12 substr. MG1655, while
517 simultaneously producing functional models that can be characterized phenotypically. **(A)** Comparison
518 of included network topology among contextualized models, also contrasted against that standard
519 implementation of parsimonious FBA (pFBA). The majority of reactions and metabolites are conserved
520 across models within a central set of pathways (71.5% and 83.6% respectively). **(B)** Comparison of
521 essentiality for conserved genes across pruned RIPTiDe models. Hierarchical clustering reveals

522 context-specific pathway essentially, labeled across the bottom axis, based on the environment in
523 which the bacterium is growing. **(C)** Calculated doubling times from 500 biomass reaction flux samples.
524 Significance determined by Wilcoxon signed rank test with Benjamini Hochberg correction (***, *p*-value
525 ≦ 0.001; ns, *p*-value > 0.05). Red lines indicate experimental measurements for corresponding *in*
526 *vitro* media conditions.
527



528

529 **Figure 4 | Cellular sources of NADPH in RIPTiDe contextualized models reflect known biological**
530 **differences in pentose phosphate pathway activity**. In both panels are shown: Reactions included
531 from the pentose phosphate pathway in each context-specific model, pruned reactions from this
532 pathway in both contexts, the distribution of transcript abundances in each dataset, and the flux
533 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).

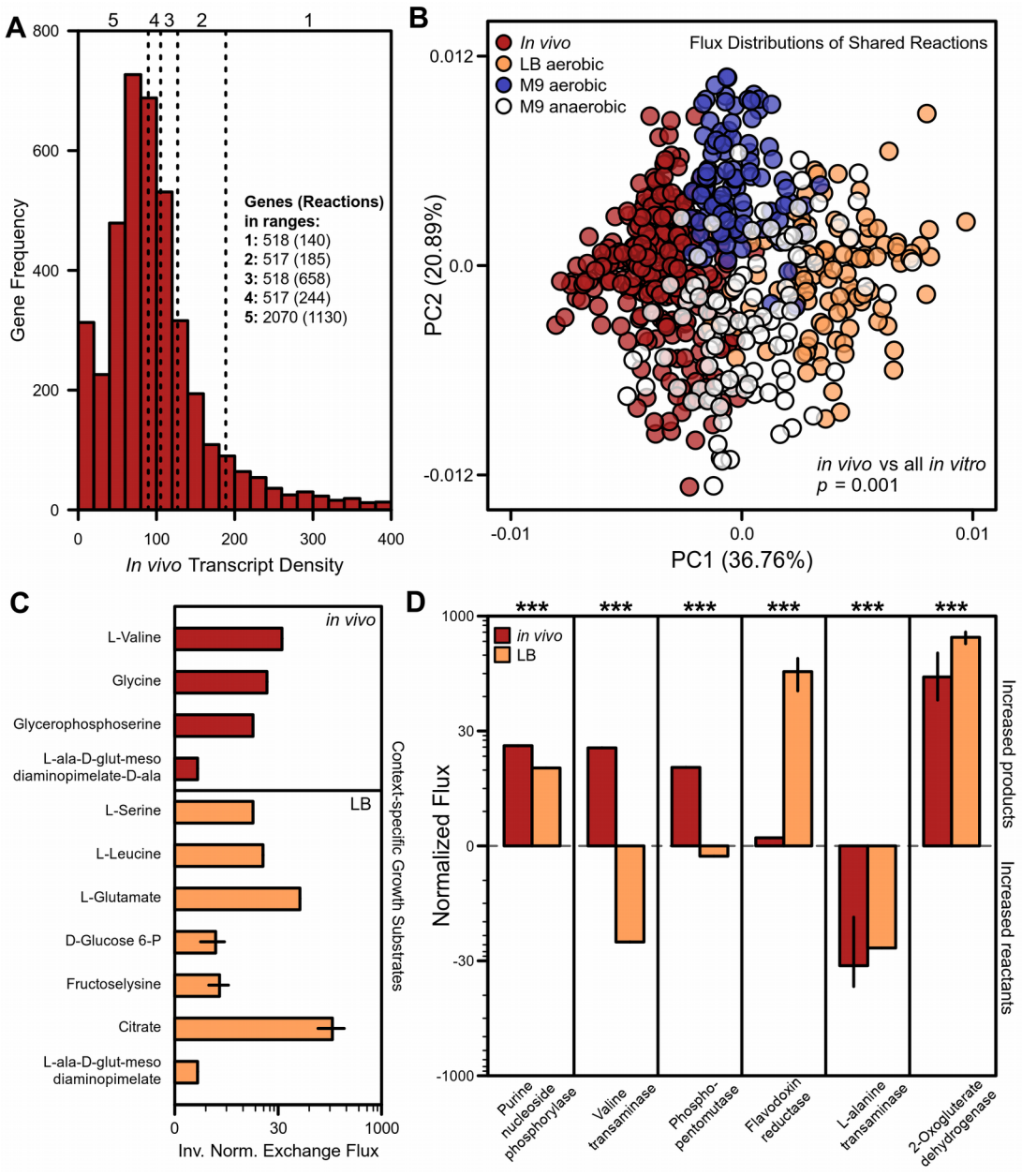
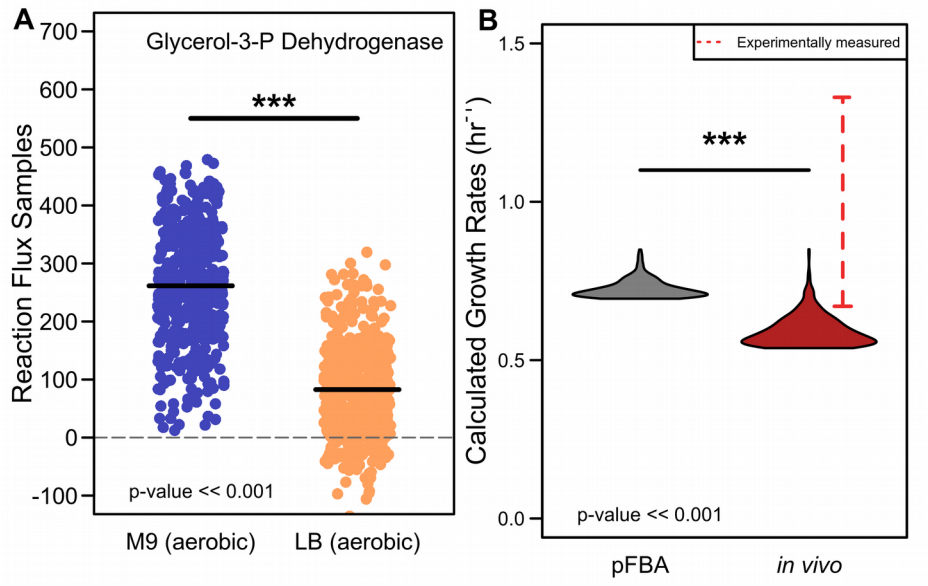


Figure 5 | RIPTiDe reveals differential host-associated, metabolism utilizing *in vivo* metatranscriptomic data. Transcriptomic reads attributable to *E. coli* were extracted from a metatranscriptomic dataset sequenced from the cecum of mice in which *E. coli* is the most highly abundant community member [8]. **(A)** *In vivo* transcript abundance of reads recruited to the *E. coli* K-12 MG-1655 genome. Quantile range bins for genes and their associated reactions are also shown. **(B)**

543 PCoA ordination of Bray-Curtis dissimilarities between flux distributions among reactions of each
544 contextualized iJO1366 model either from *in vivo* or LB rich media (aerobic, *in vitro*) conditions.
545 Significance was calculated by PERMANOVA. **(C)** Metabolite substrates of exchange reactions
546 exclusive to the contextualized models shown. Inverse, normalized flux was calculated by dividing each
547 flux sample by the associated flux through biomass in the same distribution, then multiplied by the
548 overall median flux through biomass across all conditions analyzed. Median and interquartile ranges
549 are displayed. **(D)** Results from Random Forest machine learning to identify the most differentiating
550 metabolic reactions between models. Top six reactions from this analysis are displayed. Median and
551 interquartile ranges are shown, and significant differences in computed doubling times were determined
552 by Wilcoxon signed-rank test with Benjamini-Hochberg correction (***, p-value $\leqq 0.001$).
553

554 **Supplemental Figures**



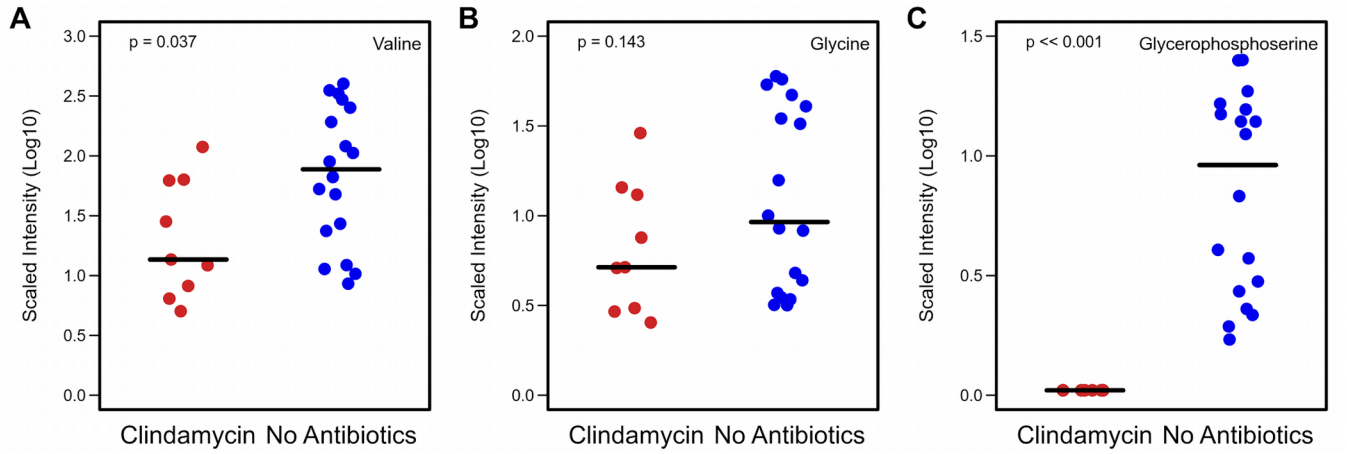
555

556 **Figure S1 | *In silico* growth predictions for *E. coli* K-12 MG1655 further supports the utility of**
557 **RIPTiDe. (A)** Glycerol-3-phosphate dehydrogenase activity of Glycerol-3-Phosphate Dehydrogenase in
558 M9 and LB. Alternative source of cellular NADPH that is conserved across both contextualized iJO1366
559 models. **(B)** Validation of RIPTiDe results from *in vivo* metatranscriptomic reads mapped to *E. coli* K-12
560 MG1655 genes. Computed *in vivo* doubling time from biomass reaction flux samples of contextualized
561 model. Red lines indicate previous experimentally measured in vivo growth rate ranges [34].

562 Computed *in vivo* range also significantly differs from pFBA analysis of iJO1366. Significant differences

563 were determined by Wilcoxon signed-rank test.

564



565

566 **Figure S2 | Relative concentrations of metabolites highlighted during *in vivo* analysis of iJO1366**

567 **with RIPTiDe.** Concentrations quantified with liquid chromatography mass-spectrometry previously [3].

568 Comparing the metabolite content in cecal content of mice pretreated with an intraperitoneal injection of

569 clindamycin against untreated control animals. **(A)** Valine, **(B)** Glycine, **(C)** and Glycerophosphoserine.

570 Significant differences were determined by Wilcoxon signed-rank test with Benjamini-Hochberg

571 correction.

572

573 **Table S1 | Structure and analysis for example metabolic reconstruction**

574 **Table S2 | Parameters and results from previous transcriptomic integration algorithms**

575 **Table S3 | Network topology and gene essentiality across RIPTiDe contextualized models**

576

577 **References**

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

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

582 3. Jenior ML, Leslie JL, Young VB, Schloss PD. *Clostridium difficile* colonizes alternative nutrient
583 niches during infection across distinct murine gut microbiomes. *mSystems*. 2017;
584 doi:10.1128/mSystems.00063-17.

585 4. Maksymiuk C, Balakrishnan A, Bryk R, Rhee KY, Nathan CF. E1 of α-ketoglutarate
586 dehydrogenase defends *Mycobacterium tuberculosis* against glutamate anaplerosis and
587 nitroxidative stress. *Proc Natl Acad Sci*. 2015; 112: E5834–E5843.

588 5. Oberhardt MA, Palsson BØ, Papin JA. Applications of genome-scale metabolic reconstructions.
589 *Mol Syst Biol*. 2009; 5: 320.

590 6. Dunphy LJ, Yen P, Papin JA. Integrated Experimental and Computational Analyses Reveal
591 Differential Metabolic Functionality in Antibiotic-Resistant Pseudomonas aeruginosa. *Cell Syst*.
592 2019; 8: 3–14.e3.

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

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

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

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

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

604 12. Zur H, Ruppin E, Shlomi T. iMAT: an integrative metabolic analysis tool. *Bioinformatics*. 2010;26:
605 3140–3142.

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

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

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

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

615 17. Hausser J, Mayo A, Alon U. Central dogma rates and the trade-off between precision and
616 economy. *Nature Comm*. 2018. doi:10.1101/276139

617 18. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*.
618 2010;11: R106.

619 19. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis?. *Nature Biotechnology*. 2010. pp.
620 245–248. doi:10.1038/nbt.1614.

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

623 21. Kaufman DE, Smith RL. Direction Choice for Accelerated Convergence in Hit-and-Run Sampling.

624 *Operations Research*. 1998. pp. 84–95. doi:10.1287/opre.46.1.84.

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

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

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

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

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

634 27. Monk JM, Koza A, Campodonico MA, Machado D, Seoane JM, Palsson BØ, Herrgård MJ, Feist
635 AM. Multi-omics Quantification of Species Variation of *Escherichia coli* Links Molecular Features
636 with Strain Phenotypes. *Cell Syst*. 2016; 3: 238–251.e12.

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

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

642 30. Murashko ON, Lin-Chao S. *Escherichia coli* responds to environmental changes using enolasic
643 degradosomes and stabilized DicF sRNA to alter cellular morphology. *Proc Natl Acad Sci*. 2017;
644 114: E8025–E8034.

645 31. Stincone A, Prigione A, Cramer T, Wamelink MMC, Campbell K, Cheung E, et al. The return of
646 metabolism: biochemistry and physiology of the pentose phosphate pathway. *Biological Reviews*.
647 2015. pp. 927–963. doi:10.1111/brv.12140.

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

651 33. Yeh JI, Chinte U, Du S. Structure of glycerol-3-phosphate dehydrogenase, an essential monotopic
652 membrane enzyme involved in respiration and metabolism. *Proc Natl Acad Sci*. 2008; 105: 3280–
653 3285.

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

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

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

660 37. Nakayama T, Yonekura SI, Yonei S, Zhang-Akiyama QM. *Escherichia coli* pyruvate:flavodoxin
661 oxidoreductase, YdbK - regulation of expression and biological roles in protection against oxidative
662 stress. *Genes Genet Syst*. 2013; 88: 175–188.

663 38. Kafri M, Metzl-Raz E, Jona G, Barkai N. The Cost of Protein Production. *Cell Reports*. 2016. pp.
664 22–31. doi:10.1016/j.celrep.2015.12.015.

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

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

669 41. Ebrahim A, Lerman JA, Palsson BØ, Hyduke DR. COBRApy: COnstraints-Based Reconstruction
670 and Analysis for Python. *BMC Syst Biol*. 2013; 7: 74.

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

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

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

677 45. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012; 9: 357–
678 359.

679 46. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The
680 Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009. pp. 2078–2079.

681 47. King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, Ebrahim A, Palsson BØ, Lewis NE.
682 BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic*
683 *Acids Res*. 2016; 44: D515–22.

684 48. Dixon P. VEGAN, a package of R functions for community ecology. Journal of Vegetation Science.
685 2003. p. 927. doi:10.1658/1100-9233.

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

687 50. Wortel MT, Bosdriesz E, Teusink B, Bruggeman FJ. Evolutionary pressures on microbial metabolic
688 strategies in the chemostat. *Scientific reports.* 2016; 6: 29503.

689 51. Stubbendieck RM, Vargas-Bautista C, Straight PD. Bacterial Communities: Interactions to Scale.
690 *Front Microbiol*. 2016;7:1234. doi:10.3389/fmicb.2016.01234.

691