

# Pipeline for Analyzing Activity of Metabolic Pathways in Planktonic Communities Using Metatranscriptomic Data

FILIPP MARTIN RONDEL,<sup>1,i</sup> ROYA HOSSEINI,<sup>1</sup> BIKRAM SAHOO,<sup>1</sup> SERGEY KNYAZEVA,<sup>1</sup>  
IGOR MANDRIC,<sup>1</sup> FRANK STEWART,<sup>2</sup> ION I. MĂNDOIU,<sup>3</sup> BOGDAN PASANIUC,<sup>4</sup>  
YURI POROZOV,<sup>5,6</sup> and ALEXANDER ZELIKOVSKY<sup>1,5,ii</sup>

## ABSTRACT

In this article, we present our novel pipeline for analysis of metabolic activity using a microbial community's metatranscriptome sequence data set for validation. Our method is based on expectation-maximization (EM) algorithm and provides enzyme expression and pathway activity levels. Further expanding our analysis, we consider individual enzymatic activity and compute enzyme participation coefficients to approximate the metabolic pathway activity more accurately. We apply our EM pathways pipeline to a metatranscriptomic data set of a plankton community from surface waters of the Northern Gulf of Mexico. The data set consists of RNA-seq data and respective environmental parameters, which were sampled at two depths, six times a day over multiple 24-hour cycles. Furthermore, we discuss microbial dependence on day–night cycle within our findings based on a three-way correlation of the enzyme expression during antipodal times—midnight and noon. We show that the enzyme participation levels strongly affect the metabolic activity estimates: that is, marginal and multiple linear regression of enzymatic and metabolic pathway activity correlated significantly with the recorded environmental parameters. Our analysis statistically validates that EM-based methods produce meaningful results, as our method confirms statistically significant dependence of metabolic pathway activity on the environmental parameters, such as salinity, temperature, brightness, and a few others.

**Keywords:** enzyme expression, metatranscriptome, microbial community, NGS, pathway activity level.

---

<sup>1</sup>Department of Computer Science, Georgia State University, Atlanta, Georgia, USA.

<sup>2</sup>Department of Microbiology and Immunology, Montana State University, Bozeman, Montana, USA.

<sup>3</sup>Computer Science & Engineering Department, University of Connecticut, Storrs, Connecticut, USA.

<sup>4</sup>Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, USA.

<sup>5</sup>World-Class Research Center “Digital biodesign and personalized healthcare,” I.M. Sechenov First Moscow State Medical University, Moscow, Russia.

<sup>6</sup>Department of Computational Biology, Sirius University of Science and Technology, Sochi, Russia.

<sup>i</sup>ORCID ID (<https://orcid.org/0000-0001-6278-1074>).

<sup>ii</sup>ORCID ID (<https://orcid.org/0000-0003-4424-4691>).

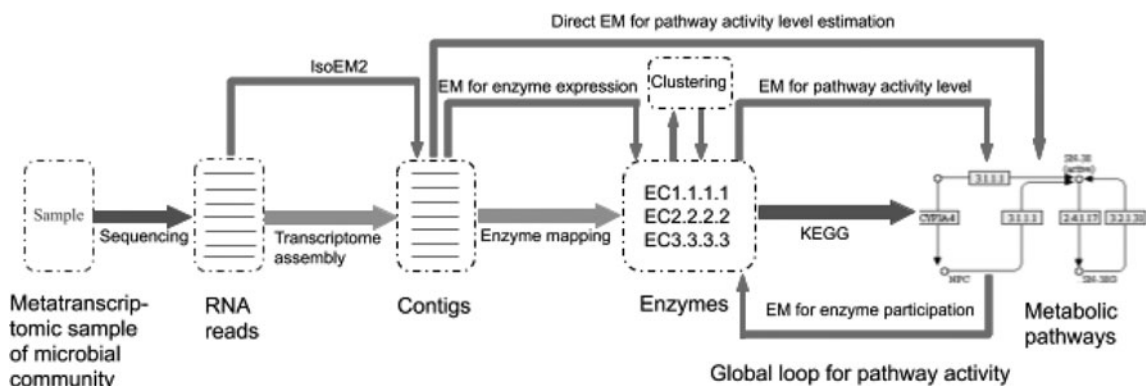
## 1. INTRODUCTION

CALCULATING THE FUNCTIONAL ACTIVITY AND INTERACTION of metabolic pathways in microbial communities is essential for understanding ecological and biochemical contributions of microorganisms. Despite many advances in using RNA-seq to understand individual contributions of organisms, it remains challenging to quantify how the expression of individual enzymes contributes to the activity of multienzyme metabolic pathways. In this study, we analyze time-series metatranscriptomic data to generate enzyme expression and metabolic pathway activity levels, as well as calculate individual contributions of enzymes to metabolic pathways. (Subramanian et al., 2005; Efron and Tibshirani, 2007; Mitrea et al., 2013; Shen et al., 2019).

Even though advances in high-throughput sequencing have aided the exploration of RNA sequencing data, it is often challenging to disentangle community-level data (Tarca et al., 2012; Donato et al., 2013; Mitrea et al., 2013), notably as existing pathway analysis tools (e.g., MEGAN4, MetaPathways, MinPath) often yield variable conclusions about the activity of pathways based on RNA data (Ye and Doak, 2009; Huson et al., 2011; Sharon et al., 2011; Konwar et al., 2013). We developed a workflow that uses a maximum likelihood-based model, annotations provided by KEGG (Kanehisa, 2000), as well as MAP platform (Huntemann et al., 2016) that predicts genes expressed in samples, while also provides information about gene classification into orthology groups (Fig. 1) to estimate transcript frequency, enzyme expression, enzyme participation in pathways, and metabolic pathway activity. In this article, we test this model using metatranscriptomic data from a marine microbial community sampled during both day and night, therefore likely exhibiting predictable variation in community transcription patterns. The data span multiple time points with different environmental parameters to elucidate the complex metabolic pathway activity in the microbial community, generally challenging to mimic in a laboratory environment.

The proposed methodology is the first to use a likelihood model to infer the pathway activity using an enzyme's expression and participation coefficient. First, we filtered the microbial community-specific metabolic pathways from the KEGG database and merged the expression of enzymes sharing the same contigs and having sequence homologs. We implemented a novel expectation-maximization (EM) algorithm to estimate the enzyme participation level in each pathway and then used these estimations for more accurate predictions of pathway activity. Increased correlation between estimated metabolic pathway activity and environmental parameters validated our approach. Our contributions include the following:

- A direct EM-based algorithm estimating pathway activity levels based on metatranscriptomic read data
- An EM-based algorithm for estimating enzyme expression
- A novel EM-based algorithm for estimating metabolic pathway activity levels using estimation of enzyme participation level in each pathway
- Validation of estimated enzyme expression and pathway activity as well as their dependency on the environmental parameters.



**FIG. 1.** Pipeline of metabolic pathway analysis for a microbial community sample. The metatranscriptomic data obtained from microbial community samples are sequenced, and raw reads are assembled into contigs. The genes containing obtained contigs are further mapped into the enzyme-pathway database. Contig frequencies are obtained using IsoEM2 (Mandric et al., 2017a). The direct EM estimates pathway activity levels using directly contig frequencies. Alternatively, we first estimate the enzyme expressions, then cluster enzymes, and simultaneously estimate enzyme participation in each pathway and pathway activity levels. EM, expectation-maximization.

The rest of the article is organized as follows. In the next section we describe the pipeline of our software framework and several EM-based algorithms for estimating enzyme expression and metabolic pathway activity in microbial communities. Then we describe our data sets, including sequencing data, and extraction of metabolic enzymes and pathways. Finally, our results statistically validate the proposed pipeline.

## 2. METHODS

We first describe the pipeline containing the previous version of our software and an alternative flow with three new EM algorithms. Then each of these three new EMs are described separately and the global loop for pathway activity level estimation concludes description of our software.

In this section, we describe the procedure of inferring metabolic pathway activity levels from RNA-Seq data for microbiome communities. We also apply differential pathway activity level analysis similar to the nonparametric statistical approach described in Al Seesi et al. (2014), which was successfully applied for gene differential expression.

This article proposes to enhance the pipeline proposed in Mandric et al. (2017b) (Fig. 1) with the inference of enzyme expressions and enzyme participation levels in metabolic pathway repeatedly applying the maximum likelihood model. These models are resolved using the EM algorithm. The proposed inferences are highlighted in red (Fig. 1). The first step is to estimate the abundances of the assembled contigs. The abundances can be inferred by any RNA-seq quantification tool, but we suggest using IsoEM (Mandric et al., 2017a) since it is sufficiently fast to handle Illumina Hiseq data and more accurate than Kallisto (Bray et al., 2016). We propose to estimate the enzyme expressions based on contig abundances and mapping of contigs onto enzymes (EM for enzyme expression in Fig. 1). The EM for pathway activity levels is based on inferred enzyme expressions and metabolic pathway annotation. Each enzyme is initially assigned a participation level of  $1/|w|$ , where  $|w|$  is the total amount of enzymes in the pathway  $w$ . The *Global loop for pathway activity* updates the enzyme participation level by fitting expected enzyme expressions to the expressions estimated by *EM for enzyme expression*. The *Global Loop for Pathway Activity* replaces *Direct EM for pathway activity level estimation* proposed in Mandric et al. (2017b), which directly estimates pathway activity from contig abundances, bypassing enzyme expression and participation coefficients.

### 2.1. MAP

We obtained a preliminary annotation of RNA-seq data using the DOE-JGI Metagenome Annotation Pipeline (MAP v.4; JGI portal) (Huntemann et al., 2016). MAP consists of feature prediction, including identification of protein-coding genes. First, the MEGAHIT metagenome assembler is used to assemble RNA-Seq reads into scaffolds. Second, several software suites (GeneMark.hmm, MetaGeneAnnotator, Prodigal, and FragGeneScan) are used to predict genes on assembled scaffolds. The MAP pipeline uses enzyme commission (EC) numbers to annotate genes, which is a required input in model. The annotations are obtained through homology searches (using USEARCH), within a nonredundant proteins-sequence database (maxhits=50,  $e$ -value=0.1), where each protein is assigned to a KEGG Orthology (KO) group. The top five hits for each KO, with the condition that the identity score is at least 30% and 70% of the protein length is matched, are used. The KO IDs are translated into EC numbers using KEGG KO to EC mapping.

### 2.2. Direct EM for inferring pathway activity levels

We first estimate the frequencies of the assembled contigs using IsoEM2 (Mandric et al., 2017a), as this method is almost as fast and more accurate than Kallisto (Bray et al., 2016). Then we need to estimate the frequencies of enzymes based on contig frequencies and in turn use them to infer metabolic pathway activity levels. These steps can be also integrated into a single *direct EM* that directly infers pathway activity levels from contig frequencies.

**2.2.1. EM approach.** Let  $w$  be a pathway that is considered to be a set of enzymes. Traditionally, pathway maps are drawn as graphs with EC number nodes. EC numbers have been widely used as a primary identifier for reconstructing the metabolic pathway from the complete genome. A more recent attempt to reconcile metabolic pathways with nonmetabolic ones resulted in introduction of the so-called KO. As in this article we are only interested in quantifying the activity of metabolic pathways, our primary goal of interest will be considering EC numbers and their contribution to pathway activity levels. We will, therefore, refer to the pathway  $w$  as a set of EC numbers as the signature describing the biochemical activity occurring in a given microbial/viral community. A well-known fact is that different EC numbers may take part in multiple pathways. Therefore, it is a challenging task to quantify the activity of each pathway in the condition of uncertainty of whether enzymes belonging to a particular EC number participate in one particular metabolic pathway and not in another one.

Hereunder we present an elaborated continuous maximum likelihood model based on contig abundances.

Let  $T$  be a random variable with values from the set of observed transcripts/contigs, and let  $W$  be a random variable whose values belong to the set of relevant metabolic pathways (Fig. 2). The probability of observing a contig  $t$  is given by the following formula:  $P(T=t) = \sum_{w \in W} f_w P(T=t | W=w)$ , where  $f_w$  stands for the frequency of the pathway  $w$ , which will be also referred as the *activity level* of  $w$ . We are interested in computing the distribution of frequencies on the set of pathways:  $f_W = (f_{w_1}, f_{w_2}, \dots, f_{w_{|W|}})$ . Thus, in our model we adopt the following likelihood function:

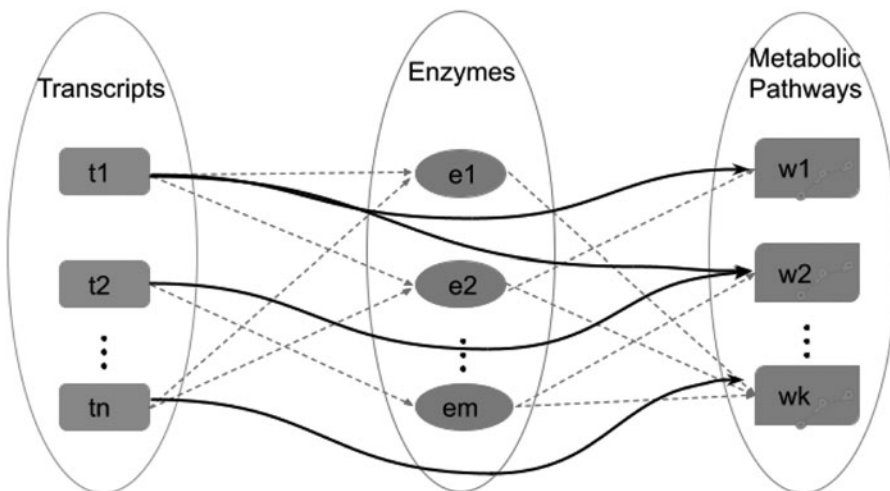
$$L(f_W) = \prod_{t \in T} \left( \sum_{w \in W} f_w P(T=t | W=w) \right)^{a_t},$$

where  $a_t$  denotes the abundance of  $t$  estimated by IsoEM2. The corresponding log-likelihood is

$$l(f_W) = \sum_{t \in T} a_t \log \left( \sum_{w \in W} f_w P(T=t | W=w) \right).$$

To each transcript we associate a set of EC numbers. Namely, transcripts are aligned to a protein database and the set of all EC numbers  $E$  corresponding to the matching proteins is retrieved. In general, more than one EC number is associated with every transcript (otherwise stated,  $|E| \geq 1$ ). We apply the law of total probability to decompose further each term  $P(T=t | W=w)$  participating in the log-likelihood:

$$\begin{aligned} P(T=t | W=w) &= \sum_{e, t \in e} P(T=t, E=e | W=w) \\ &= \sum_{e: t \in e} P(E=e | W=w) \cdot P(T=t | E=e) \end{aligned} \quad (1)$$



**FIG. 2.** Direct EM estimates pathway activity based on contig frequencies.

We use the uniform probability distribution over the set of EC numbers participating in each pathway. This means the following:

$$P(E=e|W=w)=p_{ew}=\begin{cases} \frac{1}{|w|}, & \text{if } e \in w \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Therefore, each probability term from the log-likelihood function may be written in the following form:

$$P(T=t|W=w)=\frac{1}{|w|} \cdot \sum_{e:t \in e, e \in w} P(T=t|E=e).$$

Furthermore, the log-likelihood is transformed into the following:

$$l(f_W)=\sum_{t \in T} a_t \log \left( \sum_{w \in W} f_w \cdot \left( \frac{1}{|w|} \cdot \sum_{e:t \in e, e \in w} P(T=t|E=e) \right) \right).$$

Finally,

$$l(f_W)=\sum_{t \in T} a_t \log \left( \sum_{w \in W} \frac{f_w}{|w|} \cdot \sum_{e:t \in e, e \in w} p_{te} \right),$$

where

$$p_{te}=P(T=t|E=e)=\frac{b_{te}}{\sum_{t' \in e} b_{t'e}}.$$

In the last formula,  $b_t$  are the bit-scores obtained from the alignment of assembled transcripts to the proteins of EC number  $e$ . We use the bit-score measure as the degree of reliability of each alignment. In other words, the probability of assigning a transcript  $t$  to an EC number  $e$  is proportional to the bit-score of the alignment  $(t, e)$ . Finally, we obtain

$$l(f_W)=\sum_{t \in T} a_t \log \left( \sum_{w \in W} \alpha_{tw} f_w \right),$$

where

$$\alpha_{tw}=\frac{1}{|w|} \cdot \sum_{e:t \in e, e \in w} p_{te}.$$

In the log-likelihood function  $l(f_W)$  the values  $a_t$  are obtained by running IsoEM2 (or any other tool for transcript quantification). The values  $\alpha_{tw}$  are computed from the corresponding tripartite graph (Fig. 2). The only values to be determined are  $f_w$ . We aim at finding the values  $f_w$ , which maximize the log-likelihood  $l(f_W)$ .

We apply the EM-type algorithm (Dempster et al., 1977) for determining the values  $f_w$ . We initialize each of the abundance estimates for each pathway with a random number  $f_w \in [0, 1]$ ,  $w \in W$ . Then, we iterate the following two steps until a convergence criterion is satisfied:

*The E-step.* We first compute the expected number of reads  $n_w$  emitted by each pathway  $w$  through the following formula:

$$n_w=\sum_{t \in T} a_t \cdot \frac{\alpha_{tw} f_w}{\sum_{w' \in W} \alpha_{tw'} f_{w'}}.$$

*The M-step.* The new estimates are provided based on a standard maximization EM step:

$$f_w^{\text{new}}=\frac{n_w}{\sum_{w' \in W} n_{w'}}.$$

The algorithm halts when the new estimates are ‘‘close’’ to the ones from the previous step:  $\|f_w^{\text{new}} - f_w\| \leq \varepsilon$ , where  $\varepsilon \ll 1$ .

### 2.3. EM for enzyme expression and pathway activity level estimation

Let  $T$  be a random variable with values from the set of observed contigs, and let  $E$  be a random variable whose values belong to the set of relevant metabolic enzymes from the KEGG database. The probability of observing a contig  $t$  is given by the following formula:  $P(T=t) = \sum_{w \in W} f_w P(T=t | E=e)$ , where  $f_e$  stands for the expression of the relevant metabolic enzyme  $e$ . Thus, in our model we adopt the following likelihood function:

$$L(f_e) = \prod_{t \in T} \left( \sum_{e \in E} f_e P(T=t | E=e) \right)^{a_t},$$

where  $a_t$  denotes the abundance of  $t$  estimated by IsoEM2. Following Mandric et al. (2017b) we estimate the probability of contig  $t$  coming from enzyme  $e$  as follows:

$$P(T=t|E=e) = p_{te} = \frac{b_{te}}{\sum_{t' \in e} b_{t'e}}, \quad (3)$$

where  $b_{te}$  is the best bit-score obtained from the alignment of  $t$  to the protein that have a function of the enzyme  $e$ .

The details of the EM for enzyme expression are as follows. We initialize estimates for each enzyme with a random number  $f_e \in [0, 1]$ ,  $e \in E$ . Then, we iterate the following two steps until a convergence criterion is satisfied:

*The E-step.* We first compute the expected number of reads  $n_e$  emitted by each enzyme  $e$  through the following formula:

$$n_e = \sum_{t \in T} a_t \cdot \frac{p_{te} f_e}{\sum_{e' \in E} p_{te'} f_{e'}}.$$

*The M-step.* The new estimates are provided based on a standard normalization step:

$$f_e^{\text{new}} = \frac{n_e}{\sum_{e' \in E} n_{e'}}.$$

The algorithm halts when the change in estimates between iterations is small enough:  $\|f_E^{\text{new}} - f_E\| \leq \varepsilon$ , where  $\varepsilon \ll 1$ .

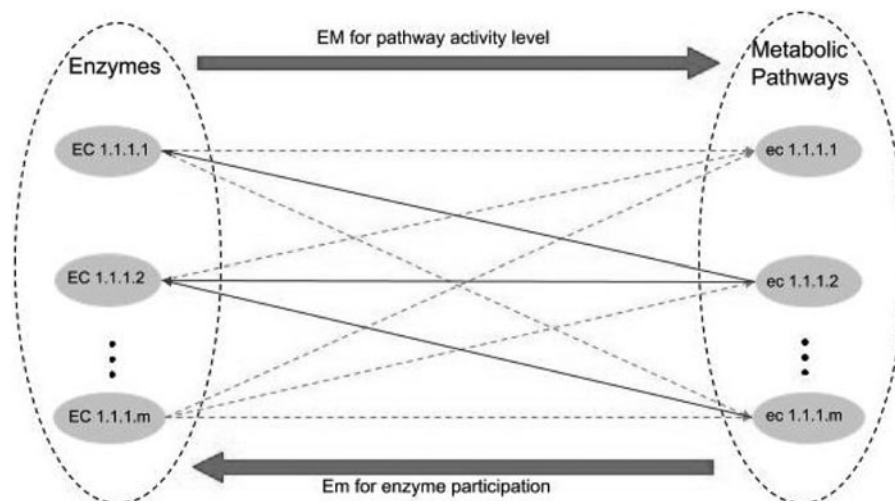
The EM algorithm for estimating pathway activity levels  $f_w = \{f_w | w \in W\}$  based on frequencies of enzymes  $f_E = \{f_e | e \in E\}$  is similar to the EM aforementioned algorithm. The only difference is that instead of Equation (3) we use the uniform probability distribution over the set of enzymes/enzyme groups participating in each pathway [see Eq. (2)].

The initial estimate [Eq. (2)] of the participation level of enzyme  $e$  in the pathway  $w$  can be very far from reality.

More accurate estimates of the enzyme participation levels can lead to more accurate estimates for the pathway activity levels. Enzymes are represented by their ortholog groups  $w = \{p_1, \dots, p_k\}$ . Since an ortholog group can have multiple functions and participate in multiple pathways, the pathways can be viewed as a family of subsets  $W$  of the set of all ortholog groups  $P$ . The algorithm hereunder (Fig. 3) estimates pathway activity levels Steps (1–3) and then checks how well the computed activities  $f_w$ 's fit the enzyme expressions [Step (4)]. If the fit is not good enough, then EM-based algorithm is applied to update the enzyme participation levels  $p_{ew}$ 's [Steps (5–6)] and then  $f_w$ 's are recomputed according to updated  $p_{ew}$ 's in Step (3).

1. Find expression  $f(e)$  of each enzyme  $e$  running EM from Section 2.3.
2. According to Equation (2), initialize  $p_{ew} = \frac{1}{|w|}$  for  $e \in w$  and  $p_{ew} = 0$ , otherwise.
3. Find activity levels  $f_w$  for each pathway  $w \in W$  running EM from Section 2.3.
4. Find expected frequency of each enzyme  $e$  according to formula  $f_e^{\text{exp}} = \sum_{w \in W} p_{ew} f_w$ . If expected and observed enzymes frequencies are close to each other:  $\|f_{e \in E}^{\text{exp}} - f_{e \in E}\| = \sum_{e \in E} (f_e^{\text{exp}} - f_e)^2 < \varepsilon \ll 1$ , then exit, that is, go to Step (7).
5. Find better fitted  $p'_{ew}$ 's by using the following EM algorithm:

*The E-step.* Compute expected  $p_{ew}^{\text{exp}}$ 's that will make  $f_e = f_e^{\text{exp}}$  for each  $e \in E, w \in W$ ,



**FIG. 3.** Global Loop for pathway activity consists of alternative execution of the EM for pathway activity level and the EM for enzyme participation level. Together the two EMs, the pathway activity level and enzyme participation, are integrated into a single global loop that infers pathway activity.

$$P_{ew}^{\text{exp}} = p_{ew} \times \frac{f_e}{f_e^{\text{exp}}}.$$

*The M-step.* Provide the new estimates by normalization for each  $e \in E$ ,  $w \in W$ ,

$$p_{ew}^{\text{new}} = \frac{P_{ew}^{\text{exp}}}{\sum_{e \in E} P_{ew}^{\text{exp}}}.$$

The algorithm halts when the change in estimates between iterations is small enough:

$$\|p^{\text{new}} - p\| = \sum_{e \in E, w \in W} (p_{ew}^{\text{new}} - p_{ew})^2 \leq \varepsilon \ll 1.$$

6. For each  $e \in E$ ,  $w \in W$ , update  $p_{ew} \leftarrow p_{ew}^{\text{new}}$  and go to Step (3)
7. Output  $\{f_w | w \in W\}$  and  $\{p_{ew} | e \in E, w \in W\}$ .

### 3. DATA SETS

#### 3.1. Samples

The data set that we used to validate our EM model is a metatranscriptomic data set of a bacterioplankton community from surface waters of the Northern Gulf of Mexico. The RNA-seq data and respective environmental parameters were sampled in July 2015 at two depths—2 and 18 m, every 4 hours throughout 48 hours totaling in 13 samples per depth. Six environmental parameters—including photosynthetic active radiation (PAR) and seawater dissolved oxygen concentration, density, salinity, temperature, and chlorophyll concentration were measured for each sample. All data sets are publicly available through the JGI Genomes Online (GOLD) database through GOLD ID Gs0110190. Out of 26 samples four samples (Day 1, 12:00, 18 m; Day 2, 20:00, 2 m; Day 3, 08:00, 2 m; Day 3, 12:00, 18 m) were discarded as they did not contain enough reads to assemble transcripts for our pipeline (Table 1).

TABLE 1. THE 26 RNA-SEQ SAMPLES OF MICROBIAL COMMUNITIES DRAWN FROM THE NORTHERN LOUISIANA SHELF DURING CONTRASTING LIGHT AND DARK CONDITIONS DURING 3 CONSECUTIVE DAYS AT TWO DEPTHS 2 AND 18 M

<i>Sample</i>						
<i>Depth Time/day</i>	<i>18 m</i>			<i>2 m</i>		
	<i>Day 1</i>	<i>Day 2</i>	<i>Day 3</i>	<i>Day 1</i>	<i>Day 2</i>	<i>Day 3</i>
00:00		✓	✓		✓	✓
04:00		✓	✓		✓	✓
08:00		✓	✓		✓	×
12:00	×	✓	✓	✓	✓	×
16:00	✓	✓		✓	✓	
20:00	✓	✓		✓	×	

### 3.2. Microbial-specific metabolic pathway identification

Using KEGG database we extracted metabolic pathways that play a significant role in microbial communities, which is confirmed by literature referenced in PubMed (Hu and Holden, 2006; Gago et al., 2011; Janßen and Steinbüchel, 2014). We removed from consideration the high-level metabolic pathways, including ec01100, ec01110, ec01120, and ec01130. In the end, we extracted 69 microorganism-relevant pathways out of 152 metabolic pathways.


### 3.3. Enzyme identification and clustering

We restrict ourselves to enzymes that belong to microbial metabolic pathways and remove the unlikely enzyme matches. The RNA-seq coverage of may be not deep enough to distinguish genes sharing long common segments. Any contig matching one of such genes and corresponding enzymes will match another one. Therefore, we can estimate only total expression of a group of such indistinguishable enzymes rather than each of them individually. For detecting such groups of enzymes, we use an essential property that the individual enzyme expression can vary across randomly initialized EM runs, whereas the sum of the expression of all enzymes in the group does not change (Fig. 4 top). For example, five different EM runs converge to different expression of enzymes EC:3.1.3.12 and EC:2.4.1.15, whereas the sum of expressions is constant. We clustered the enzymes from the same group and rerun EM to get an accurate and stable expression of enzymes and enzymes groups. After applying the aforementioned method, we obtain expressions of 1446 enzymes and enzyme groups for the metabolic pathway activity analysis.

**a**

Enzymes	Run 1	Run 2	Run 3	Run 4	Run 5
EC:3.1.3.12	0.054	0.311	0.251	0.317	0.12
EC:2.4.1.15	0.404	0.147	0.207	0.141	0.338
Sum	0.458	0.458	0.458	0.458	0.458

**b**



ORTHOLOGY: K16055

<b>Entry</b>	K16055	KO
<b>Name</b>	TPS	
<b>Definition</b>	trehalose 6-phosphate synthase/phosphatase [EC:2.4.1.15 3.1.3.12]	
<b>Pathway</b>	ko00500	Starch and sucrose metabolism
	ko01100	Metabolic pathways
	ko01110	Biosynthesis of secondary metabolites

FIG. 4. Clustering enzymes. Over multiple runs the enzyme expressions of EC:3.1.3.12 and EC:2.4.1.15 are changing from one run to another, but the sum converges to the same overall stable group expression (a). Using KEGG we were able to verify that the two enzymes in fact belong to the same orthology (b). EC, enzyme commission.



## 4. RESULTS

Our results consist of empirical and statistical validation of estimated enzyme expression, enzyme participation levels, and pathway activity level estimations. We first analyze the stability of enzyme participation levels and then check how many enzyme expressions and pathway activities correlate with environmental parameters.

### 4.1. Enzyme participation coefficients

We estimate the participation level of each enzyme in each pathway separately for each data point. Supplementary Table S1 presents the participation level of all expressed enzymes in the pathway ec00020.

TABLE 2. ENZYME PARTICIPATION LEVELS FOR ALL ENZYMES ACROSS ALL DATA POINTS FOR 2 M DEPTH IN THE METABOLIC PATHWAY ec00620

<i>ec00620</i>	<i>D1:12</i>	<i>D1:16</i>	<i>D1:20</i>	<i>D2:00</i>	<i>D2:04</i>	<i>D2:08</i>	<i>D2:12</i>	<i>D2:16</i>	<i>D3:00</i>	<i>D3:04</i>	<i>D3:12</i>	<i>AVE</i>	<i>STD</i>
EC:1.1.1.27	42.95	35.51	0	33.42	32.4	44.44	33.24	29.38	36.59	40.64	0	36.51	4.83
EC:1.2.1.3	24.76	18.14	16.58	7.76	8.41	18.7	18.99	13.19	11.15	18.12	62.69	19.86	14.38
EC:1.2.4.1	38.37	42.02	37.39	44.65	45.06	44.91	40.53	37.73	44.44	48.06	44.5	42.51	3.38
EC:1.2.7.1	1.52	11.99	27.96	8.86	6.26	13.78	24.45	15.29	10.8	22.17	5.14	13.47	8
EC:1.2.7.3	41.86	41.6	36.82	35.42	36.49	36.77	39.56	35.36	37.14	41.85	54.06	39.72	5.13
EC:1.8.1.4	22.78	25.05	20.36	22.44	20.98	21.27	24.06	26.92	26.28	24.03	44.86	25.37	6.49
EC:2.3.1.12	38.37	42.02	37.39	44.65	45.06	44.91	40.53	37.73	44.44	48.06	44.5	42.51	3.38
EC:2.7.1.40	35.64	28.45	28.35	26.37	22.74	31.44	34.27	25.96	28.45	32.93	43.66	30.75	5.49
EC:4.1.1.32	38.37	42.02	37.39	44.65	45.06	44.91	40.53	37.73	44.44	48.06	44.5	42.51	3.38
EC:4.1.1.49	44.22	45.88	42.37	42.64	45.52	55.21	49.52	45.22	49.23	57.65	51.86	48.12	4.83
EC:6.2.1.1	46.31	40.16	47.24	23.62	23.27	45.55	38.32	33.7	30.74	36.18	65.96	39.19	11.6
EC:6.2.1.13	0	0	0	0	0	0	0	35.26	0	0	0	35.26	0
EC:1.1.1.37	54.26	38.32	48.23	23.71	23.86	45.51	37.71	32.51	31.51	37.9	89.51	42.09	17.51
EC:1.1.5.4	0	0	0	38.88	0	0	0	32.66	0	0	0	35.77	3.11
EC:1.3.5.1	63.87	53.78	52.08	45.91	49.78	45.44	54.7	47.05	49.9	57.05	94.61	55.83	13.3
EC:4.2.1.2	43.51	36.47	35.65	31.49	35.27	30.01	36.6	32.08	33.94	38.87	66.38	38.21	9.59
EC:6.4.1.1	43.51	36.47	35.65	31.49	35.27	30.01	36.6	32.08	33.94	38.87	66.38	38.21	9.59
EC:6.4.1.2	30.87	37.69	42.5	33.54	36.18	45.06	46.21	46.7	44.16	52.04	100.02	46.82	17.87
EC:2.3.1.9	19.22	23.39	18.19	15.1	15.74	17.95	21.76	19.98	16.93	23.02	70.94	23.84	15.13
EC:1.1.1.79	26.73	28	26.2	27.73	31.77	30.16	32.11	38.07	37.46	38.88	0	31.71	4.61
EC:2.3.3.13	34.2	29.92	31.57	35.2	43.16	30.56	32.52	29.24	36.83	40.3	54.11	36.15	7.04
EC:1.2.1.10	0	0	0	0	0	0	0	1.45	0	0	0	1.45	0
EC:2.3.1.8	48.03	37.41	48.06	21.72	21.95	41.09	36.71	31.92	28.92	34.4	0	35.02	8.83
EC:2.7.2.1	36.84	31.58	39.38	19.17	18.38	30.02	27.06	24.59	22.92	26.26	0	27.62	6.59
EC:1.1.1.28	0	29.92	0	35.2	0	30.56	32.52	29.24	0	0	54.11	35.26	8.66
EC:1.1.1.38	34.2	29.92	31.57	35.2	43.16	30.56	32.52	29.24	36.83	40.3	0	34.35	4.37
EC:1.1.1.39	0	36.15	39.18	0	44.34	0	0	0	0	0	0	39.89	3.38
EC:1.1.1.40	43.25	36.15	39.18	35.34	44.34	42.6	44.95	39.4	44.41	54.22	0	42.38	5.13
EC:1.1.2.3	34.2	29.92	31.57	35.2	43.16	30.56	32.52	29.24	36.83	0	54.11	35.73	7.26
EC:1.1.2.4	34.2	29.92	31.57	35.2	43.16	30.56	32.52	29.24	36.83	40.3	0	34.35	4.37
EC:1.2.1.21	0	0	50.87	34.5	0	0	0	50.81	0	0	0	45.39	7.7
EC:2.3.1.54	0	29.92	31.57	35.2	0	30.56	32.52	0	36.83	40.3	0	33.84	3.5
EC:2.3.3.9	62.26	43.35	50.87	34.5	44.64	63.03	64.99	50.81	54.47	69.31	95.72	57.63	15.67
EC:2.7.9.1	46.02	37.73	37.8	30.5	35.23	34.88	41.95	35.93	36.27	43.75	0	38.01	4.4
EC:2.7.9.2	36.84	31.58	39.38	19.17	18.38	30.02	27.06	24.59	22.92	26.26	59.92	30.56	11.22
EC:2.8.3.1	10	4.42	2.61	3.21	2.68	4.52	9.71	5.34	9.18	8.75	0	6.04	2.88
EC:3.1.2.6	34.2	29.92	31.57	35.2	43.16	30.56	32.52	29.24	36.83	40.3	0	34.35	4.37
EC:3.6.1.7	15.66	0	0	4.01	5.97	8.11	0	6.07	7.51	7.24	0	7.79	3.44
EC:4.1.1.31	40.32	33.98	42.49	20.31	19.85	34.12	30.68	27.67	25.21	29.6	0	30.42	7.21
EC:4.2.1.130	34.2	29.92	31.57	35.2	43.16	30.56	0	29.24	36.83	0	0	33.83	4.33
EC:4.4.1.5	34.2	29.92	31.57	35.2	43.16	30.56	32.52	29.24	36.83	40.3	0	34.35	4.37

AVE, average value; STD, standard deviation.

TABLE 3. ENZYME PARTICIPATION LEVELS FOR ALL ENZYMES ACROSS ALL DATA POINTS FOR 2 M DEPTH IN THE METABOLIC PATHWAY ec00561

<i>ec00561</i>	<i>D1:12</i>	<i>D1:16</i>	<i>D1:20</i>	<i>D2:00</i>	<i>D2:04</i>	<i>D2:08</i>	<i>D2:12</i>	<i>D2:16</i>	<i>D3:00</i>	<i>D3:04</i>	<i>D3:12</i>	<i>AVE</i>	<i>STD</i>
EC:1.1.1.2	56.43	52.43	43.04	46.68	51.86	41.29	65.25	39.34	46.54	37.81	0.00	48.07	8.10
EC:1.2.1.3	98.96	136.18	95.66	69.85	79.35	69.48	112.58	60.23	80.19	83.30	208.15	99.45	40.11
EC:1.1.1.21	61.63	62.54	48.77	46.64	50.41	39.57	55.37	34.04	49.16	40.73	77.37	51.47	11.72
EC:2.7.7.9	60.17	55.14	50.26	39.73	44.57	45.06	62.68	38.50	45.22	41.12	131.96	55.86	25.27
EC:3.2.1.22	47.41	47.43	38.91	41.32	42.99	35.23	0.00	30.73	41.31	38.9	0.00	40.47	5.07
EC:2.3.1.20	90.96	77.07	0.00	61.58	59.41	61.75	57.87	52.78	56.58	76.46	0.00	66.05	11.86
EC:2.7.1.31	94.47	131.20	99.82	119.04	122.46	0.00	0.00	0.000	119.22	122.29	0.00	115.5	12.28
EC:2.3.1.51	58.79	90.73	75.29	75.97	69.66	59.05	79.23	59.45	80.74	65.96	0.00	71.49	10.22
EC:3.13.1.1	0.000	90.59	81.77	59.46	69.55	68.97	76.55	59.72	69.07	75.58	0.000	72.36	9.46
EC:1.1.1.156	90.96	0.000	0.00	0.00	0.00	0.00	0.00	52.78	0.00	0.00	0.00	71.87	19.09
EC:1.1.1.6	0.000	0.00	0.00	0.00	0.000	0.000	57.87	52.78	0.00	0.00	0.00	55.33	2.54
EC:2.3.1.15	50.8	66.09	55.42	55.00	49.67	49.59	65.75	44.20	60.25	54.64	149.56	63.72	27.90
EC:2.3.1.22	90.96	77.07	63.98	61.58	59.41	0.00	0.00	52.78	0.00	0.00	0.00	67.63	12.72
EC:2.4.1.241	0.00	0.00	0.00	61.58	0.00	61.75	57.87	52.78	56.58	0.00	0.00	58.11	3.35
EC:2.4.1.315	0.00	0.00	0.00	61.58	59.41	0.00	0.00	0.00	0.00	0.00	0.00	60.49	1.08
EC:2.4.1.336	0.00	0.00	0.00	0.00	0.00	61.75	0.00	0.00	0.00	0.00	0.00	61.75	0.00
EC:2.4.1.337	0.00	0.00	0.00	0.00	59.41	0.00	0.00	0.00	0.00	0.00	0.00	59.41	0.00
EC:2.4.1.46	0.00	77.07	63.98	0.00	0.00	61.75	57.87	52.78	56.58	0.00	0.00	61.67	7.77
EC:2.7.1.107	50.8	66.09	55.42	55.00	49.67	49.59	65.75	44.20	60.25	54.64	0.00	55.14	6.77
EC:2.7.1.29	0.00	0.00	108.83	78.85	74.39	82.41	77.92	65.58	75.45	78.86	0.00	80.29	11.74
EC:2.7.1.30	90.96	77.07	63.98	61.58	59.41	61.75	57.87	52.78	56.58	76.46	0.00	65.84	11.27
EC:2.7.8.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	56.58	0.00	0.00	56.58	0.00
EC:3.1.1.23	0.00	0.00	63.98	61.58	59.41	61.75	57.87	52.78	56.58	76.46	0.00	61.30	6.59
EC:3.1.1.3	90.96	77.07	63.98	61.58	59.41	61.75	57.87	52.78	56.58	76.46	390.20	95.33	93.86
EC:3.1.1.34	0.00	0.00	63.98	0.00	0.00	0.00	0.00	0.00	0.00	76.46	0.00	70.22	6.24
EC:3.1.3.4	43.19	50.83	42.05	44.04	43.19	37.42	64.52	33.09	46.99	40.50	51.36	45.20	7.95
EC:3.1.3.81	0.00	0.00	0.00	0.00	0.00	49.59	0.00	0.00	0.00	54.64	0.00	52.12	2.53

Similar tables can be found for ec00620 in Table 2 and for ec00561 in Table 3. We can see that the participation level does not significantly change from one data point to another, that is, the standard deviation is significantly smaller than the mean for all enzymes. Note that if an enzyme is not expressed in a sample, then the participation is not defined and the participation level is reported as zero. This means that we need to take into account only data points with nonzero participation levels when computing mean and standard deviation over all data points.

#### 4.2. Correlation of pathway activity levels with environmental parameters

The goal of regression-based validation is to check our hypothesis that there exist enzymes and pathways whose expression and activity level variation across data points can be explained (i.e., correlate with) certain environmental parameters. For each environmental parameter, we check whether it significantly correlates ( $P < 5\%$ ) with each enzyme across 11 data points for the 2-m depth (Table 4). In row 2 we give

TABLE 4. CORRELATION OF ENZYME EXPRESSION WITH ENVIRONMENTAL PARAMETERS

	<i>Salinity</i>	<i>Temp</i>	<i>Oxygen</i>	<i>Chl</i>	<i>PAR</i>	<i>Density</i>	<i>MLR</i>
1. No. of enzymes	146	110	117	93	97	138	156
2. 95% CI	80–190	79–114	62–94	58–92	36–63	82–123	70–107
3. EC number	1.2.1.59	2.6.1.1	3.1.3.11	2.2.1.7	3.5.1.16	2.4.1.16	1.1.1.136

1. The number of enzymes significantly correlated with each of six environmental parameters and their linear combination (through MLR). 2. The number of enzymes strongly correlated with randomly permuted parameter values (95% CI). 3. The EC number of the metabolic enzyme, which is the most strongly correlated with the corresponding parameter.

MLR, multiple linear regression; PAR, photosynthetic active radiation.

TABLE 5. GLOBAL LOOP EXPECTATION-MAXIMIZATION

	<i>Salinity</i>	<i>Temp</i>	<i>Oxygen</i>	<i>Chl</i>	<i>PAR</i>	<i>Density</i>	<i>MLR</i>
1. No. of pathways	31	22	19	18	14	30	22
2. 95% CI	1–8	0–8	0–6	0–6	0–6	01-Aug	0–7
3. Pathway	ec00071	ec00195	ec00622	ec00460	ec00360	ec00071	ec00626

1. The number of pathways significantly correlated with each of six environmental parameters and correlated through multiple linear regression. 2. The number of pathways strongly correlated with randomly permuted parameter values (95% CI). 3. The EC number of the metabolic pathway, which is the most strongly correlated with the corresponding parameter.

95% CI for the number of significantly correlated enzymes with a randomly permuted parameter. Since the upper bound of 95% CI for salinity is 190 (row 2), we conclude that there is no evidence of enzymes significantly correlated with salinity. We also report the enzyme that correlates the most with salinity, that is, EC 1.2.1.59. From Table 4 we see that most parameters do not correlate well with enzymes, except perhaps PAR.

Table 5 is the same as Table 4 but reports correlation significance of pathway activities instead of enzyme expressions. In contrast to enzymes it is clear that the many metabolic pathways correlate with each environmental parameter and this correlation is not by chance. Indeed, pathway activity is supposed to be more stable than enzyme expression since generally metabolism is much less affected by the current. For each environmental parameter, we also cross-check the PubMed database whether the most correlated pathway is known to depend on this parameter. For instance, fatty acid degradation is well correlated with salinity, and several studies reported that fatty acid degradation is often altered by salinity at sea surface environments (Kaye, 2004; Heinzelmann et al., 2015; Carvalho and Caramujo, 2018). The citric acid pathway's role is to provide the energy required for the growth and division of microorganisms by breaking organic molecules in the presence of oxygen (Hu and Holden, 2006). In addition, it plays a central role in regulating other metabolic processes in microorganisms. The occurrence of fatty acid biosynthesis is diverse in the microbial community, which controls lipid homeostasis and biogenesis. Fatty acid biosynthesis supports the membrane biogenesis and controls the usages of adenosine triphosphate (ATP), crucial for microbial metabolism (Gago et al., 2011; Janßen and Steinbüchel, 2014).

Table 6 is the same as Table 5. The only exception for this table being Direct EM used to compute metabolic pathway activity directly from contigs, as opposed to Global Loop EM, which uses enzyme expression and enzyme participation coefficients to compute pathway activity. Although there is significant correlation between metabolic pathway activity and temperature, chlorophyll, as well as all environmental parameters bundled together, some other pathways may have correlated with the rest of the environmental parameters by chance. The statistical regression validation used to evaluate our model clearly demonstrates Global Loop EM's ability to calculate metabolic pathway activity more accurately than Direct EM.

#### 4.3. Cyclic changes of enzyme expressions and pathway activities

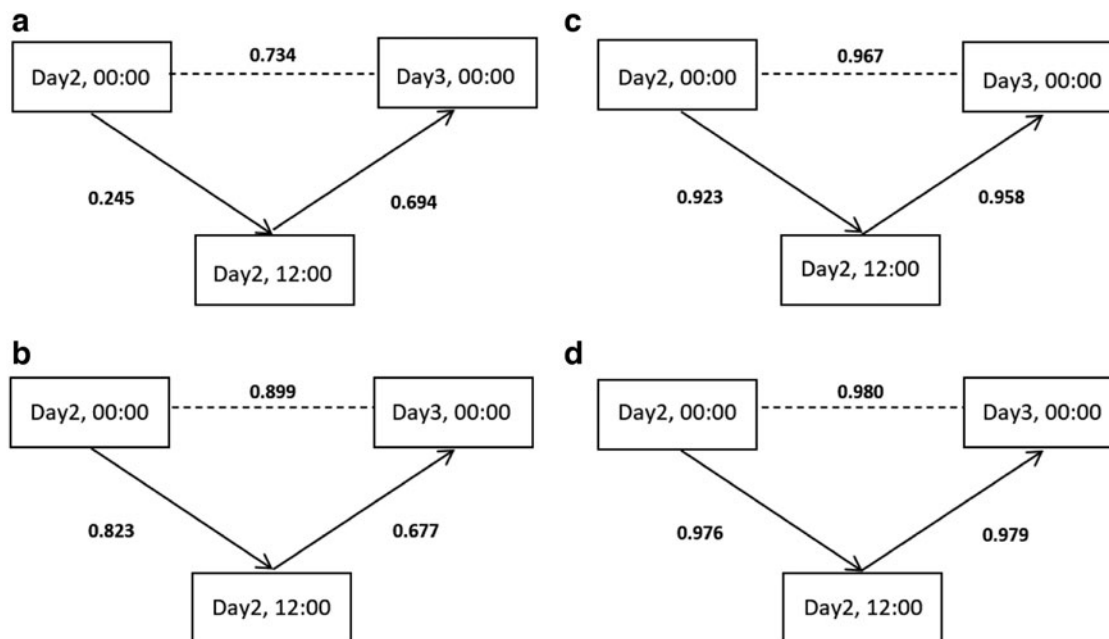
We hypothesize that we will be able to observe the cyclic changes in enzyme expression and pathway activity level during 36 hours from 00:00 am on day 2 until 12:00 am on day 3. The cyclic changes should manifest themselves as a higher similarity between two respective middays and midnights that

TABLE 6. DIRECT EXPECTATION-MAXIMIZATION

	<i>Salinity</i>	<i>Temp</i>	<i>Oxygen</i>	<i>Chl</i>	<i>PAR</i>	<i>Density</i>	<i>MLR</i>
1. No. of pathways	5	14	5	8	1	4	10
2. 95% CI	1–10	01–11	1–8	0–7	0–6	1–8	0–8
3. Pathway	ec00364	ec00310	ec00281	ec00281	ec00740	ec00623	ec00623

Similarly to Table 5 this table presents the results of the statistical validation, the only difference is the Direct EM from contigs to pathway activity being used here.

EM, expectation-maximization.



**FIG. 5.** Correlations between enzyme expressions for three time points (time 00:00 of the day 2, 00:00 of the day 3, and 12:00 of the day 2) at 2-m depth (a) and, respectively, at 18-m depth (b). Correlations between pathway activity levels for three time points (time 00:00 of day 2, 00:00 day 3, and 12:00 of day 2) at 2-m depth (c) and, respectively, at 18-m depth (d).

are 24 hours apart than the similarity between two data points that are 12 hours apart. We measure similarity between two data points by the correlation between all estimated enzyme expressions or, alternatively, all estimated pathway activity levels. Figure 5a (respectively, Fig. 5b) shows the correlation between enzyme expressions in three time points at the depth of 2 m (respectively, 18 m). Similarly, Figure 5c and d show the correlations between pathway activity levels. For the enzyme expressions and the pathway activity levels, the correlation between midnight samples (24-hour gap) is higher than the correlation between midnight and noon samples (just 12-hour gap). It is also important to notice that as expected pathway activity levels are more stable than enzyme expressions. Indeed, correlations between enzymes expression are significantly lower than correlations between pathway activity levels.

## 5. DISCUSSION

This article proposes a maximum likelihood model for the estimation of metabolic pathway activity in the microbial community using the KEGG pathway database. Specifically, the proposed approach uses an EM-based pipeline to estimate enzyme expression, enzyme participation levels in pathways, and metabolic pathway activity from metatranscriptomic data. The proposed metabolic pathway analysis was applied to the metatranscriptomic data of 26 samples collected with different environmental parameters. The key findings of the study are as follows:

- The participation levels of enzymes in pathways do not significantly vary across the data samples.
- The enzyme expression and metabolic pathway activities were validated using regression with each environmental parameter: salinity, temperature, oxygen, chlorophyll, and PAR.
- The three-way metabolic pathway expression correlation across four groups of samples shows that the metabolic activity at depth of 2 m during daytime is more closely related to the daytime activity the next day than either of the day samples related to the night sample's metabolic activity.
- In contrast to enzyme expressions, pathway activity levels significantly correlate with environmental parameters, for example, 31 out of 61 metabolic pathways significantly correlate with salinity.

## AUTHOR DISCLOSURE STATEMENT

The authors declare no competing financial interests.

## FUNDING INFORMATION

I.M., S.K., F.M.R., and A.Z. were partially supported from NSF Grants 1564899 and 16119110, and NIH grant 1R01EB025022-01, I.M. and S.K. were partially supported by GSU Molecular Basis of Disease Fellowship, I.I.M. was partially supported from NSF Grants 1564936 and 1618347, F.S. was partially supported by NSF Grants 1151698, 1558916, and 1564559, and Simons Foundation award 346253, Y.P. was partially supported by Ministry of Science and Higher Education of the Russian Federation Grant 075-15-2020-926.

## SUPPLEMENTARY MATERIAL

Supplementary Table S1

## REFERENCES

- Al Seesi, S., Tiagueu, Y.T., Zelikovsky, A., et al. 2014. Bootstrap-based differential gene expression analysis for RNA-Seq data with and without replicates. *BMC Genomics* 15 Suppl 8, S2.
- Bray, N.L., Pimentel, H., Melsted, P., et al. 2016. Erratum: Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34, 888.
- Carvalho, C., and Caramujo, M. 2018. The various roles of fatty acids. *Molecules* 23, 2583.
- Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977 Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B* 39, 1–22.
- Donato, M., Xu, Z., Tomoiaga, A., et al. 2013. Analysis and correction of crosstalk effects in pathway analysis. *Genome Res.* 23, 1885–1893.
- Efron, B., and Tibshirani, R. 2007. On testing the significance of sets of genes. *Ann. Appl. Stat.* 1, 107–129.
- Gago, G., Diacovich, L., Arabolaza, A., et al. 2011. Fatty acid biosynthesis in actinomycetes. *FEMS Microbiol. Rev.* 35, 475–497.
- Heinzelmann, S.M., Chivall, D., M'Boule, D., et al. 2015. Comparison of the effect of salinity on the D/H ratio of fatty acids of heterotrophic and photoautotrophic microorganisms. *FEMS Microbiol. Lett.* 362, fnv065.
- Hu, Y., and Holden, J.F. 2006. Citric acid cycle in the hyperthermophilic archaeon *Pyrobaculum islandicum* grown autotrophically, heterotrophically, and mixotrophically with acetate. *J. Bacteriol.* 188, 4350–4355.
- Huntmann, M., Ivanova, N.N., Mavromatis, K., et al. 2016. The standard operating procedure of the DOE-JGI metagenome annotation pipeline (MAP v.4). *Stand. Genomic Sci.* 11, 17.
- Huson, D.H., Mitra, S., Ruscheweyh, H.-J., et al. 2011. Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* 21, 1552–1560.
- Janßen, H.J., and Steinbüchel, A. 2014. Fatty acid synthesis in *Escherichia coli* and its applications towards the production of fatty acid based biofuels. *Biotechnol. Biofuels* 7, 7.
- Kanehisa, M. 2000. KEGG: Kyoto encyclopedia of genes and genomes. *NAR.* 28, 27–30.
- Kaye, J.Z. 2004. *Halomonas neptunia* sp. nov., *Halomonas sulfidaeris* sp. nov., *Halomonas axialensis* sp. nov. and *Halomonas hydrothermalis* sp. nov.: Halophilic bacteria isolated from deep-sea hydrothermal-vent environments.
- Konwar, K.M., Hanson, N.W., Pagé, A.P., et al. 2013. MetaPathways: A modular pipeline for constructing pathway/genome databases from environmental sequence information. *BMC Bioinformatics* 14, 202.
- Mandric, I., Knyazev, S., Padilla, C., et al. 2017a. Metabolic analysis of metatranscriptomic data from planktonic communities. Proceedings of International Symposium on Bioinformatics Research and Applications (ISBRA) 2017. *LNCIS* 10330. pp. 396–402.
- Mandric, I., Temate-Tiagueu, Y., Shcheglova, T., et al. 2017b. Fast bootstrapping-based estimation of confidence intervals of expression levels and differential expression from RNA-Seq data. *Bioinformatics* 33, 3302–3304.
- Mitreá, C., Taghavi, Z., Bokanizad, B., et al. 2013. Methods and approaches in the topology-based analysis of biological pathways. *Front. Physiol.* 4, 278.

- Sharon, I., Bercovici, S., Pinter, R.Y., et al. 2011. Pathway-based functional analysis of metagenomes. *J. Comput. Biol.* 18, 495–505.
- Shen, M., Li, Q., Ren, M., et al. 2019. Trophic status is associated with community structure and metabolic potential of planktonic microbiota in plateau lakes. *Front. Microbiol.* 10, 2560.
- Subramanian, A., Tamayo, P., Mootha, V.K., et al. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102, 15545–15550.
- Tarca, A.L., Draghici, S., Bhatti, G., et al. 2012. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics* 13, 136.
- Ye, Y., and Doak, T.G. 2009. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput. Biol.* 5, e1000465.

Address correspondence to:  
*Filipp Martin Rondel*  
*Department of Computer Science*  
*Georgia State University*  
*Atlanta, GA 30302-3965*  
*USA*

*E-mail:* frondel1@student.gsu.edu