

Georgia State University

ScholarWorks @ Georgia State University

Computer Science Dissertations

Department of Computer Science

5-1-2023

Expectation Maximization Methods for Metabolic Pathway Analysis

Filipp Rondel

Follow this and additional works at: https://scholarworks.gsu.edu/cs_diss

Recommended Citation

Rondel, Filipp, "Expectation Maximization Methods for Metabolic Pathway Analysis." Dissertation, Georgia State University, 2023.

doi: <https://doi.org/10.57709/35365476>

This Dissertation is brought to you for free and open access by the Department of Computer Science at ScholarWorks @ Georgia State University. It has been accepted for inclusion in Computer Science Dissertations by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact scholarworks@gsu.edu.

Expectation Maximization Methods for Metabolic Pathway Analysis

by

Filipp Rondel

Under the Direction of Alexander Zelikovsky, Ph.D.

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2023

ABSTRACT

Metabolic pathways are a series of enzyme-mediated reactions that result in the transformation of substances from one form to another. While methods for studying metabolic pathways are constantly improving, analyzing these pathways can be challenging. To accurately predict metabolic pathway activity, it is essential to understand and quantify the relative involvement of enzymes in these pathways. In my dissertation, I propose a novel method based on the maximum likelihood Expectation-Maximization (EM) algorithm to estimate metabolic pathway activity levels using enzyme participation as a latent variable. This improved maximum likelihood model will be used to conduct downstream analysis of metabolic pathway expression, which will be estimated from RNA-Seq samples obtained from rodents and a planktonic microbial community.

INDEX WORDS: Expectation Maximization, Pathway activity level, Enzyme expression, Enzyme participation in pathways

Copyright by
Filipp Rondel
2023

Expectation Maximization Methods for Metabolic Pathway Analysis

by

Filipp Rondel

Committee Chair: Alex Zelikovsky

Committee: Pavel Skums

Murray Patterson

Artem Rogovskyy

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

May 2023

DEDICATION

To my grandfather who always believed in me, to my grandmother who always understood me, to my parents who encouraged me, and to my lovely wife that supported me every step of the way.

ACKNOWLEDGMENTS

I would like to express my gratitude to Dr. Alex Zelikovsky, my scientific advisor, for providing me with careful guidance and support throughout my Ph.D. journey. His mentorship has not only helped me grow as a professional but also as a collaborator and an individual. I would like to extend my appreciation to Dr. Pavel Skums, Dr. Murray Patterson, Dr. Artem Rogovsky, Dr. Frank Stewart, Dr. Bogdan Pasaniuc, Dr. Ion Mandoiu and Dr. Igor Mandric for their invaluable advises and guidance. I had the pleasure of collaborating with my lab colleagues and peers, and I would like to express my gratitude to Dr. Andrew Melnyk, Dr. Viachaslau Tsyvina, Dr. Mark Grinshpon, Dr. Sergey Knyazev, Dr. Kiril Kuzmin, Roya Hosseini, Hafsa Farooq, Fatemeh Mohebbi, Akshay Juyal, Alina Nemira, Bikram Sahoo, and the master's and undergraduate students from Georgia State University, Georgia Tech, and UCLA who worked with me on various projects. Finally, I would like to express my appreciation to all the friends I made in the Computer Science department, who have accompanied me throughout this long journey.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	v
LIST OF TABLES	viii
LIST OF FIGURES	x
1 INTRODUCTION	1
1.1 Metabolic pathways and enzyme participation	2
1.2 Problem formulations	3
1.3 Contributions	4
1.4 Refereed Journal Articles	5
1.5 Refereed Articles in Conference Proceedings	5
1.6 Books	5
2 PIPELINE FOR ANALYZING ACTIVITY OF METABOLIC PATHWAYS IN PLANK- TONIC COMMUNITIES USING METATRANSCRIPTOMIC DATA	7
2.1 Methods	9
<i>2.1.1 Direct EM for inferring athway activity levels</i>	11
<i>2.1.2 EM for enzyme expression and pathway activity level estimation</i>	15
2.2 Datasets	19
2.3 Results	23
<i>2.3.1 Enzyme participation coefficients</i>	24
<i>2.3.2 Correlation of pathway activity levels with environmental parameters</i>	24
<i>2.3.3 Cyclic changes of enzyme expressions and pathway activities</i>	26
2.4 Discussion	28

3	ASSESSING THE LEVELS OF ENZYME EXPRESSION AND METABOLIC PATHWAY ACTIVITY IN MICE, BOTH INFECTED AND UNINFECTED WITH BORRELIA BURGDORFERI	31
3.1	Methods	33
3.1.1	<i>Pipeline for estimating metabolic pathway activity of C3H and P. leucopus</i>	33
3.1.2	<i>Mapping between genes, enzymes and pathways for C3H and P. leucopus</i>	36
3.1.3	<i>Enzyme grouping</i>	36
3.1.4	<i>Feedback loop for pathway activity level estimation</i>	39
3.2	Datasets	40
3.2.1	<i>Bacterial inoculum</i>	40
3.2.2	<i>Rodent infection</i>	40
3.2.3	<i>RNA sequencing</i>	41
3.3	Results	42
3.4	Conclusions	44
4	EMPATHWAYS2: ESTIMATION OF ENZYME EXPRESSION AND METABOLIC PATHWAY ACTIVITY USING RNS-SEQ READS	46
4.1	Introduction	46
4.2	Materials	48
4.2.1	<i>Software and Data</i>	48
4.3	Methods	50
4.3.1	<i>Preparation and Quality Assessment</i>	50
4.3.2	<i>Align RNA-seq read</i>	51
4.3.3	<i>Produce gene expression data using IsoEM2</i>	52
4.3.4	<i>Calculate enzyme and metabolic pathway expression with EMPathways2</i>	52
4.4	Notes	54
	REFERENCES	56

LIST OF TABLES

Table 2.1	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 2m and 18m.	19
Table 2.2	Enzyme participation levels for all enzymes across all data points for 2m depth in the metabolic pathway ec00620.	22
Table 2.3	Enzyme participation levels for all enzymes across all data points for 2m depth in the metabolic pathway ec00561.	23
Table 2.4	1. The number of enzymes significantly correlated with each of 6 environmental parameters and their linear combination (via multiple linear regression (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.	24
Table 2.5	Global Loop EM. 1. The number of pathways significantly correlated with each of 6 environmental parameters and correlated via 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.	26
Table 2.6	Direct EM. Similarly to Table 2.5 this table presents the results of the statistical validation, the only difference is the Direct EM from contigs to pathway activity being used here.	26
Table 2.7	Enzyme participation levels for all enzymes across all data points for 2m depth in the metabolic pathway ec00020. Two rightmost columns are means and standard deviations of enzyme participation levels.	30
Table 3.1	A pair of individually unstable enzymes that are stable when summed into a group.	37
Table 3.2	Three triplets and one quadruplet of collapsed enzymes.	38
Table 3.3	<i>C3H</i> pathways with significant different activity level across infected and uninfected groups.	43
Table 3.4	<i>C3H</i> pathways with slightly different activity level across infected and uninfected groups.	43

Table 3.5	<i>P. leucopus</i> pathways with significant different activity level across infected and uninfected groups.	44
Table 3.6	<i>P. leucopus</i> pathways with slightly different activity level across infected and uninfected groups.	44
Table 3.7	The enzyme expression coefficients and relative standard deviations (<i>%RSD</i>) for the enzyme participation coefficients in pathway ec00620.	45
Table 4.1	Software and URLs	49

LIST OF FIGURES

Figure 2.1	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 ⁴⁸ . 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.	9
Figure 2.2	Direct EM estimates pathway activity based on contig frequencies.	12
Figure 2.3	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 EM's, The pathway activity level and enzyme participation, are integrated into a single global Loop which infers pathway activity.	17
Figure 2.4	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).	20
Figure 2.5	Correlations between enzyme expressions for 3 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 3 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).	28
Figure 3.1	Full pipeline for metabolic pathway analysis for rodent samples. The RNA-Seq data obtained from rodents are sequenced, then raw reads are mapped into genes. The genes containing obtained contigs are further mapped into the enzyme-pathway database. Gene expression is obtained using IsoEM2 ⁴⁹ . Then we estimate estimate enzyme expression using gene expression. Finally, the the pathway activity level and enzyme participation coefficients are estimated in the feedback loop.	34
Figure 3.2	EMPathways2 pipeline for metabolic pathway analysis for rodent samples. The RNA-Seq data obtained from rodents are sequenced, then raw reads are mapped into genes. The genes containing obtained contigs are further mapped into the enzyme-pathway database. Gene expression is obtained using IsoEM2 ⁴⁹ . Then we estimate estimate enzyme expression using gene expression. Finally, the the pathway activity level and enzyme participation coefficients are estimated in the feedback loop.	35

Figure 3.3 (A) Enzymes that cannot be distinguished from each other must be treated as groups. (B) Enzymes that are unstable are collapsed into a single enzyme with the lowest EC nomenclature number. 38

CHAPTER 1

INTRODUCTION

The term metabolism is derived from the Greek word - metabolē meaning "to change". Metabolism involves numerous biochemical processes that occur continually within an organism to sustain life. These processes are complex and essential, involving the combination of substrates, proteins, and other molecules to produce, release, and regulate energy. This energy then serves as fuel for all living cells. Even when an organism is at rest, metabolism remains active and ongoing, providing energy for basic functions such as respiration, circulation, digestion, growth, homeostasis, and other essential processes.

Understanding metabolism is key to understanding life as we know it. It has been a subject of fascination with scientists for over 150 years¹. However, the majority of progress occurred in the last few decades mostly due to the advancements in sequencing and computational technologies. Transcriptome RNA sequencing (RNA-Seq) has recently emerged as an accurate and robust tool for expression pattern analysis of genes due to its extensive genomic range, high reproducibility, and superior evaluation for expression levels^{56,51,45}. At increasingly reduced cost, RNA-Seq has become a routine technique for gene expression analysis^{53,17,69,59,60}. In short, expressing a gene means manufacturing its corresponding protein. Enzymes are a kind of specialized protein that catalyze a biochemical reaction. Furthermore, particular groups of such enzymes can catalyze a series of consecutive reactions, known as metabolic pathways, which break down and/or produce complex biological molecules in order to regulate energy.

Metabolic pathways are distinct, organized parts of metabolism. Metabolic pathway activity

quantification is crucial to understanding metabolism. Despite metabolic pathways having been studied for decades, quantification has only become possible in the last decade^{76,58,25}. In this dissertation, I propose a method for a more accurate estimation of metabolic pathway activity levels using an individual enzyme's participation in metabolic pathways as a latent variable that is computed using a maximum likelihood model based on gene expression from RNA-seq data.

1.1 Metabolic pathways and enzyme participation

Metabolic pathways are linked sequences of biochemical reactions that occur within a living cell. These sequences of biochemical reactions are connected to each other by their intermediate products - the metabolites of one reaction are the substrates for the next. However, the metabolites do not simply react with each other on their own, the chemical reactions also need biological catalysts. Metabolic pathways heavily depend upon enzymes to catalyze individual steps of the series of reactions. Enzymes are proteins that function as biocatalysts, which accelerate the reactions by lowering the activation energy^{5,41,62}. Most metabolic processes in any living cell require enzyme catalysis to occur fast enough in order to sustain life^{72,12,42,26}.

While all enzymes in a given metabolic pathway are necessary in order for the metabolic reactions to occur, a number of enzymes are shared between multiple metabolic pathways. Despite the exact same structure, the functions of individual enzymes may differ across various pathways. For example, the same enzyme may be activated or inhibited purely depending on the concentration and the type of substrate available in the cell. The presence of the enzyme alone might not necessarily indicate the activity of a certain metabolic pathway. In some cases, it becomes challenging

to predict which metabolic pathway said enzyme may be expressing for. However, individual enzyme's expression and its relation to metabolic pathways may be used to compute its relative importance to a metabolic pathways activity. Latter allows any enzyme's relative participation in various pathways to be measured using a participation coefficient.

First, I discuss using static enzyme participation coefficient to improve the Expectation-Maximization (EM) based algorithm to infer metabolic pathway activity. Then, I explore using maximum likelihood EM based model to infer every enzyme's participation coefficient for select metabolic pathways present in microbial community samples, as well as multiple contrasting groups of house and white-footed mice. Finally, I compare both metabolic pathway inference model's accuracy and discuss challenges related to identifying specific enzyme's participation in particular metabolic pathways.

1.2 Problem formulations

This dissertation addresses the following problems:

- Given RNA-Seq reads from biological samples

Estimate

(i) Gene expression

(ii) Enzyme expression

(iii) Metabolic pathway activity level

- Given:

- (i) Enzyme expression
- (ii) Metabolic pathway activity level

Estimate enzyme-in-pathway participation coefficient

1.3 Contributions

This dissertation discusses the following contributions:

- Exploring static enzyme-in-metabolic-pathway participation coefficients.
- Designing a novel EMPathways algorithm that uses enzyme participation coefficients to more accurately infer metabolic pathway activity. The algorithm uses Expectation-Maximization based methods to estimate enzyme participation coefficients in each metabolic pathway. This approach allows to predict and group enzymes with similar functions as well as leverage groups of enzymes to improve accuracy and stability of inferred pathways.
- Performing a differential analysis of metabolic pathway activity from RNA-Seq data sampled from a microbial community.
- Discussing validation of estimated enzyme expression and pathway activity as well as their dependency on the environmental parameters.
- Analyzing of pathway activity levels for infected and uninfected house and white-footed mice.

- Comparing metabolic pathway expression estimate calculated using static enzyme participation as opposed to participation coefficients inferred using a maximum likelihood EM based model.

1.4 Refereed Journal Articles

2. **F. Rondel**, R. Hosseini, H. Farooq, B. Bello, A. Juyal, S. Knyazev, B. Pasaniuc, S. Mangul, A. S. Rogovskyy, A. Zelikovsky, "Estimating enzyme expression and metabolic pathway activity in *Borrelia*-infected and uninfected mice." **Biomolecules** (invited)
1. **F. Rondel**, R. Hosseini, B. Sahoo, S. Knyazev, I. Mandric, F. Stewart, I. I. Măndoiu, B. Pasaniuc, Y. Porozov, A. Zelikovsky, "Pipeline for Analyzing Activity of Metabolic Pathways in Planktonic Communities Using Metatranscriptomic Data," **Journal of Computational Biology** **28(8)**: 1-14, 2021 doi: 10.1089/cmb.2021.0053

1.5 Refereed Articles in Conference Proceedings

1. **F. Rondel**, R. Hosseini, B. Sahoo, S. Knyazev, I. Mandric, F. Stewart, I. I. Măndoiu, B. Pasaniuc, A. Zelikovsky, "Estimating Enzyme Participation in Metabolic Pathways for Microbial Communities from RNA-Seq Data," Proc. of International Symposium on Bioinformatics Research & Applications (ISBRA), 2020, Lecture Notes in Bioinformatics 12304, 335-343

1.6 Books

1. **F. Rondel**, H. Farooq, M. Grinshpon, R. Hosseini, A. Zelikovsky, "EMPathways 2: Expectation Maximization Methods for Metabolic Pathway Analysis," Deciphering metabolic

pathways through metatranscriptomic data analysis, Transcriptome Data Analysis, Methods in Molecular Biology, Springer Nature 2023 (invited)

Invited Talks

3. **F. Rondel** and A. Zelikovsky Estimating enzyme expression and metabolic pathway activity in mice 17th International Symposium on Bioinformatics Research and Applications (ISBRA), 2022
2. **F. Rondel** and A. Zelikovsky Estimating enzyme participation in metabolic pathways for microbial communities from RNA-seq data 16th International Symposium on Bioinformatics Research and Applications (ISBRA), 2021
1. **F. Rondel** and A. Zelikovsky Analysis of metabolic pathway activity in planktonic communities 9th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS), 2019

CHAPTER 2

PIPELINE FOR ANALYZING ACTIVITY OF METABOLIC PATHWAYS IN PLANKTONIC COMMUNITIES USING METATRANSCRIPTOMIC DATA

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 multi-enzyme metabolic pathways^{43,21,13}. 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.^{70,19,54,66} Even though advances in high-throughput sequencing have aided the exploration of RNA sequencing data, it is often challenging to disentangle community-level data^{54,71,16}, notably as existing pathway analysis tools (e.g., MEGAN4, MetaPathways, MinPath) often yield variable conclusions about the activity of pathways based on RNA data^{32,40,75,65}. We developed a workflow that uses a Maximum Likelihood-based model, annotations provided by KEGG³⁶, as well as MAP platform³⁰ which predicts genes expressed in samples, while also provides information about gene classification into orthology groups (see Figure 3.2) to estimate transcript frequency, enzyme expression, enzyme participation in pathways, and metabolic pathway activity. In this paper, 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 envi-

ronment.

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 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. My 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.

The rest of the paper 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 datasets including sequencing data, and extraction of metabolic enzymes and pathways. Finally our results statistically validate the proposed pipeline.

2.1 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 non-parametric statistical approach described in² which was successfully applied for gene differential expression.

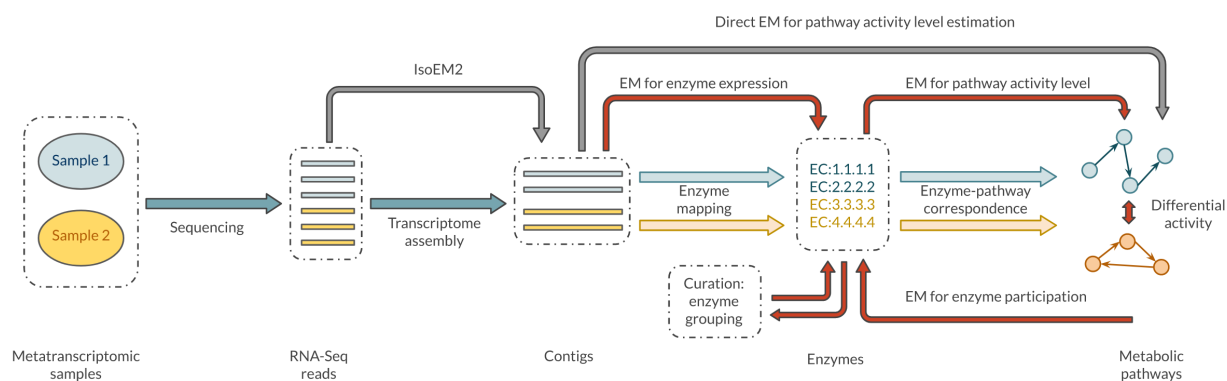


Figure 2.1: Pipeline of metabolic pathway analysis for a microbial community sample. The metatranscriptomic data obtained from microbial community samples are sequenced, and raw reads are assembled into contigs. The genes containing obtained contigs are further mapped into the enzyme-pathway database. Contig frequencies are obtained using IsoEM2⁴⁸. 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.

This paper proposes to enhance the pipeline proposed in⁴⁹ (see Figure 3.2) 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 Expectation-Maximization

(EM) algorithm. The proposed inferences are highlighted in red (see Fig. 3.2). 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 IsoEM⁴⁸ since it is sufficiently fast to handle Illumina Hiseq data and more accurate than Kallisto⁸. We propose to estimate the enzyme expressions based on contig abundances and mapping of contigs onto enzymes (EM for enzyme expression in Fig. 3.2). 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 by Mandric⁴⁹ which directly estimates pathway activity from contig abundances, bypassing enzyme expression and participation coefficients.

MAP We obtained a preliminary annotation of RNA-seq data using the DOE-JGI Metagenome Annotation Pipeline (MAP v.4) (JGI portal)³⁰. MAP consists of feature prediction, including identification of protein-coding genes. Firstly, the MEGAHIT⁴⁴ metagenome assembler is used to assemble RNA-Seq reads into scaffolds. Secondly, several software suites (GeneMark.hmm, MetaGeneAnnotator, Prodigal, FragGeneScan) are used to predict genes on assembled scaffolds^{46,61,33,57}. The MAP pipeline uses EC numbers to annotate genes, which is a required input in model. The annotations are obtained via homology searches (using USEARCH)³, within a non-redundant proteins-sequence database (maxhits=50, e-value=0.1), where each protein is assigned to a KEGG Orthology group (KO). The top 5 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.

2.1.1 Direct EM for inferring pathway activity levels

We first estimate the frequencies of the assembled contigs using IsoEM2⁴⁸, as this method is almost as fast and more accurate than Kallisto⁸. 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.

Expectation-Maximization approach Let w be a pathway that is considered to be a set of enzymes. Traditionally, pathway maps are drawn as graphs with Enzyme Commission number nodes. Enzyme Commission numbers (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 non-metabolic ones resulted in introduction of the so-called KEGG Orthology. As in this paper 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 pathways 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.

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

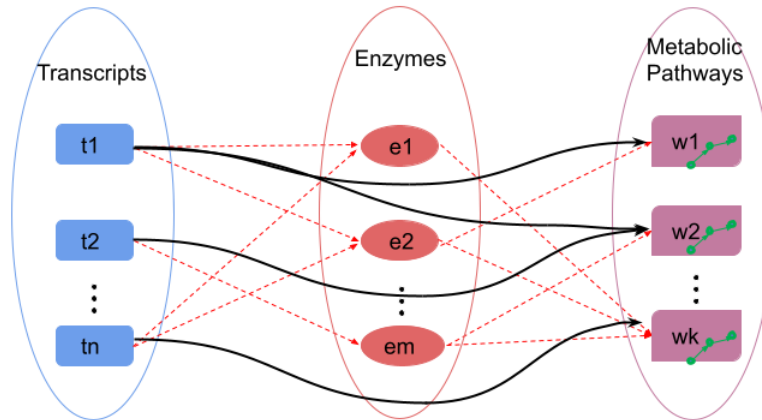


Figure 2.2: Direct EM estimates pathway activity based on contig frequencies.

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 (see Fig. 2.2). 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:

$$\begin{aligned} P(T = t|W = w) &= \sum_{e,t \in e} P(T = t, E = e|W = w) \\ &= \sum_{e:t \in e} P(E = e|W = w) \cdot P(T = t|E = e) \end{aligned} \quad (2.1)$$

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

$$P(E = e|W = w) = p_{ew} = \begin{cases} \frac{1}{|w|}, & \text{if } e \in w \\ 0, & \text{otherwise} \end{cases} \quad (2.2)$$

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

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

Further, 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 α_{tw} are computed from the corresponding tripartite

graph (see Figure 2.2). 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 algorithm¹⁴ 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 criteria 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:

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

The algorithm halts when the new estimates are “close” to the ones from the previous step:

$$\|f_W^{new} - f_W\| \leq \epsilon, \text{ where } \epsilon \ll 1$$

2.1.2 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: $P(T = t) = \sum_{w \in W} f_w P(T = t \mid E = e)$, 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. Following⁴⁹ we estimate the probability of contig t coming from enzyme e as follows:

$$P(T = t | E = e) = p_{te} = \frac{b_{te}}{\sum_{t' \in e} b_{t'e}} \quad (2.3)$$

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 criteria 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:

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

The algorithm halts when the change in estimates between iterations is small enough: $\|f_E^{new} -$

$f_E || \leq \epsilon$, where $\epsilon \ll 1$

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 algorithm above. The only difference is that instead of equation (2.3) we use the uniform probability distribution over the set of enzymes/enzyme groups participating in each pathway (see (2.2)).

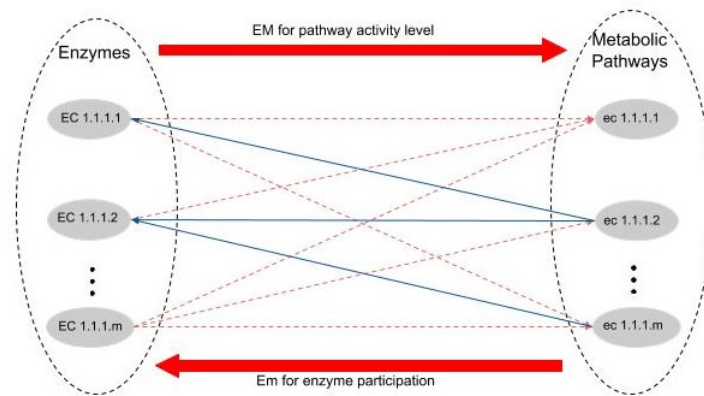


Figure 2.3: Global Loop for pathway activity consists of alternative execution of the EM for pathway activity level and the EM for enzyme participation level. Together the two EM's, The pathway activity level and enzyme participation, are integrated into a single global Loop which infers pathway activity.

The initial estimate (2.2) 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 below estimates pathway activity levels Steps (1-3) and then checks how well the 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 2.2.
2. According to (2.2), 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 $f_e^{exp} = \sum_{w \in W} p_{ew} f_w$ If expected and observed enzymes frequencies are close to each other: $\|f_{e \in E}^{exp} - f_{e \in E}\| = \sum_{e \in E} (f_e^{exp} - f_e)^2 < \epsilon \ll 1$, then exit, i.e. go to step 7.
5. Find better fitted p'_{ew} 's by using the following EM algorithm:

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

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

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

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

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

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

6. For each $e \in E, w \in W$, update $p_{ew} \leftarrow p'_{ew}$ and go to step 3
7. Output $\{f_w | w \in W\}$ and $\{p_{ew} | e \in E, w \in W\}$

2.2 Datasets

Samples. The dataset that we used to validate our EM model is a metatranscriptomic dataset 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 2 depths - 2 meters and 18 meters, every 4 hours throughout 48 hours totaling in 13 samples per depth. Six environmental parameters - including PAR (photosynthetic active radiation) and seawater dissolved oxygen concentration, density, salinity, temperature, and chlorophyll concentration were measured for each sample. All datasets are publicly available through the JGI Genomes Online (GOLD) database via GOLD ID Gs0110190. Out of 26 samples four samples (Day1, 12:00, 18m; Day 2, 20:00, 2m; Day 3, 08:00, 2m; Day 3, 12:00, 18m) were discarded as they did not contain enough reads to assemble transcripts for our pipeline (see Table 2.1).

Samples						
Depth	18 meters			2 meters		
Time \ Day	Day 1	Day 2	Day 3	Day 1	Day 2	Day 3
00:00		✓	✓		✓	✓
04:00		✓	✓		✓	✓
08:00		✓	✓		✓	✗
12:00	✗	✓	✓	✓	✓	✗
16:00	✓	✓		✓	✓	
20:00	✓	✓		✓	✗	

Table 2.1: 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 2m and 18m.

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^{29,23,35}. 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.

Enzymes	Run 1	Run 2	Run 3	Run 4	Run 5
a) 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



ORTHOLOGY: K16055

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

Figure 2.4: 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).

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, while the sum of the expression

of all enzymes in the group does not change (see Fig. 2.4 top). For example, five different EM runs converge to different expression of enzymes EC:3.1.3.12 and EC:2.4.1.15 while 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 above method, we obtain expressions of 1446 enzymes and enzyme groups for the metabolic pathway activity analysis.

ec00620	D1:12	D1:16	D1:20	D2:00	D2:04	D2:08	D2:12	D2:16	D3:00	D3:04	D3:12	AVE	STD
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	0	1.45	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

Table 2.2: Enzyme participation levels for all enzymes across all data points for 2m depth in the metabolic pathway ec00620.

ec00561	D1:12	D1:16	D1:20	D2:00	D2:04	D2:08	D2:12	D2:16	D3:00	D3:04	D3:12	AVE	STD
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	38.90	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	0.00	119.22	122.29	0.00	115.50	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	0.00	90.59	81.77	59.46	69.55	68.97	76.55	59.72	69.07	75.58	0.00	72.36	9.46
EC:1.1.1.156	90.96	0.00	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	0.00	0.00	0.00	0.00	0.00	0.00	57.87	52.78	0.00	0.00	0.00	55.33	2.54
EC:2.3.1.15	50.80	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	50.80	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

Table 2.3: Enzyme participation levels for all enzymes across all data points for 2m depth in the metabolic pathway ec00561.

2.3 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.

2.3.1 Enzyme participation coefficients

We estimate the participation level of each enzyme in each pathway separately for each data point. Table 2.7 presents the participation level of all expressed enzymes in the pathway ec00020. Similar table can be found for ec00561 in Table 2.3. We can see that the participation level does not significantly change from one data point to another, i.e. 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 0. This means that we need to take in account only data points with non-zero participation levels when computing mean and standard deviation over all data points.

	Salinity	Temp	Oxygen	Chl	PAR	Density	MLR
1. # 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

Table 2.4: 1. The number of enzymes significantly correlated with each of 6 environmental parameters and their linear combination (via multiple linear regression (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.

2.3.2 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-meter depth (see Table 2.4). In the row 2 we give 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, i.e. EC 1.2.1.59. From Table 2.4 we see that most parameters do not correlate well with enzymes, except perhaps PAR.

Table 2.5 is the same as Table 2.4 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^{28,37,10}. 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 oxygen²⁹. Additionally, 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 ATP, crucial for microbial metabolism^{23,35}.

	Salinity	Temp	Oxygen	Chl	PAR	Density	MLR
1. # pathways	31	22	19	18	14	30	22
2. 95% CI	1-8	0-8	0-6	0-6	0-6	1-8	0-7
3. Pathway	ec00071	ec00195	ec00622	ec00460	ec00360	ec00071	ec00626

Table 2.5: **Global Loop EM.** 1. The number of pathways significantly correlated with each of 6 environmental parameters and correlated via 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.

Table 2.6 is the same as Table 2.5. 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. While 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.

	Salinity	Temp	Oxygen	Chl	PAR	Density	MLR
1. # pathways	5	14	5	8	1	4	10
2. 95% CI	1-10	1-11	1-8	0-7	0-6	1-8	0-8
3. Pathway	ec00364	ec00310	ec00281	ec00281	ec00740	ec00623	ec00623

Table 2.6: **Direct EM.** Similarly to Table 2.5 this table presents the results of the statistical validation, the only difference is the Direct EM from contigs to pathway activity being used here.

2.3.3 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 respectively mid-days and mid-nights which 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 2.5.(a) (respectively, Figure 2.5.(b)) shows the correlation between enzyme expressions in 3 time points at the depth of 2m (respectively, 18 m). Similarly, Fig.2.5.c, d show the correlations between pathway activity levels. For the enzyme expressions and the pathway activity levels, the correlation between midnight samples (24 hours 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 pathways activity levels.

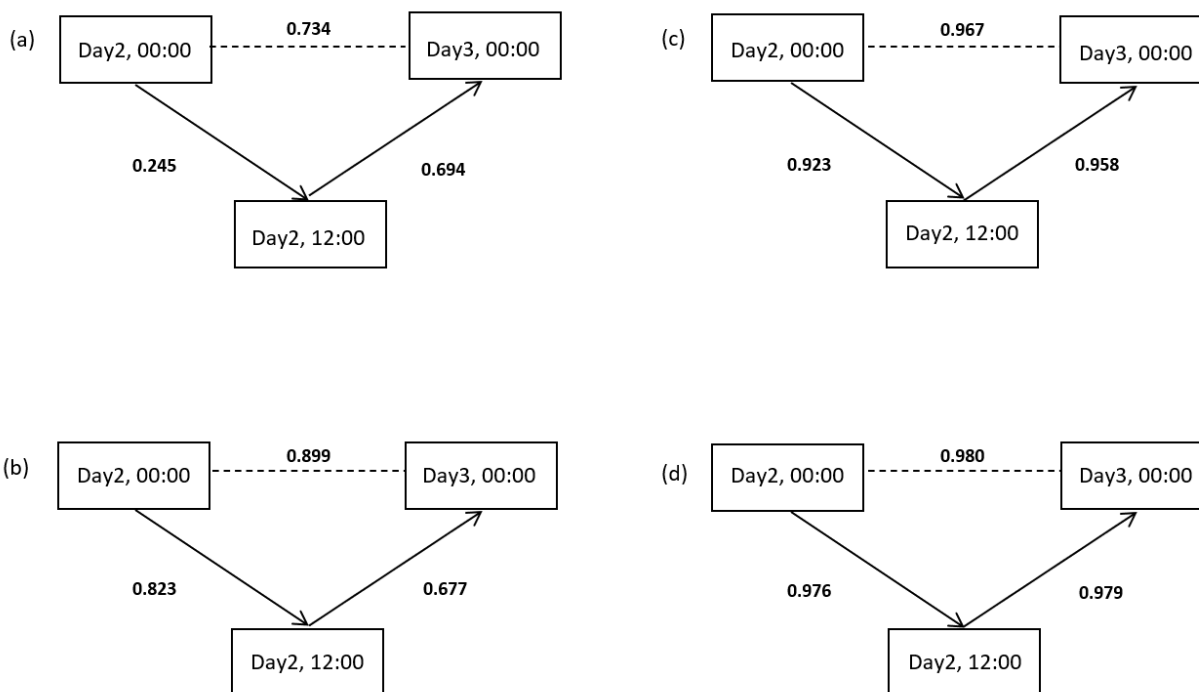


Figure 2.5: Correlations between enzyme expressions for 3 time points (time 00:00 of the day 2, 00:00 of the day 3, and 12:00 of the day 2) at 2 m-depth (a) and, respectively, at 18 m depth (b). Correlations between pathway activity levels for 3 time points (time 00:00 of day 2, 00:00 day 3, and 12:00 of day 2) at 2 m-depth (c) and, respectively, at 18 m depth (d).

2.4 Discussion

This paper 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 3-way metabolic pathway expression correlation across 4 groups of samples shows that the metabolic activity at depth of 2 meters 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, e.g. 31 out of 61 metabolic pathways significantly correlate with salinity.

Supplementary Materials

ec00020	D1:12	D1:16	D1:20	D2:00	D2:04	D2:08	D2:12	D2:16	D3:00	D3:04	D3:12	AVE	STD
EC:1.2.4.1	12.82	21.68	20.64	33.71	35.76	30.38	21.78	23.71	32.40	28.07	21.98	25.72	6.60
EC:1.2.7.1	0.51	6.18	15.43	6.69	4.97	9.32	13.14	9.61	7.87	12.95	2.54	8.11	4.37
EC:1.2.7.3	13.99	21.46	20.32	26.74	28.96	24.87	21.26	22.22	27.08	24.44	26.70	23.46	4.02
EC:1.8.1.4	7.61	12.92	11.24	16.94	16.65	14.39	12.93	16.92	19.16	14.03	22.16	15.00	3.78
EC:2.3.1.12	12.82	21.68	20.64	33.71	35.76	30.38	21.78	23.71	32.40	28.07	21.98	25.72	6.60
EC:4.1.1.32	12.82	21.68	20.64	33.71	35.76	30.38	21.78	23.71	32.40	28.07	21.98	25.72	6.60
EC:4.1.1.49	14.78	23.66	23.38	32.19	36.13	37.34	26.62	28.41	35.90	33.66	25.61	28.88	6.60
EC:1.1.1.37	18.14	19.76	26.62	17.90	18.93	30.78	20.27	20.43	22.97	22.13	44.21	23.83	7.43
EC:1.1.1.41	72.88	72.85	70.78	71.20	68.42	38.66	45.68	60.11	62.77	61.29	27.09	59.25	14.74
EC:1.1.1.42	19.96	24.06	22.58	21.52	23.68	19.95	22.48	22.32	22.95	21.92	42.38	23.98	5.95
EC:1.1.5.4	0.00	0.00	0.00	29.35	0.00	0.00	0.00	20.53	0.00	0.00	0.00	24.94	4.41
EC:1.2.4.2	10.10	13.02	10.76	11.91	10.91	11.72	12.75	14.08	14.74	10.13	25.75	13.26	4.21
EC:1.3.5.1	21.35	27.74	28.74	34.65	39.51	30.74	29.40	29.56	36.38	33.32	46.73	32.56	6.43
EC:2.3.1.61	10.10	13.02	10.76	11.91	10.91	11.72	12.75	14.08	14.74	10.13	25.75	13.26	4.21
EC:2.3.3.1	86.31	41.26	66.16	28.14	39.20	260.41	208.96	93.27	70.39	107.86	96.40	99.85	68.92
EC:2.3.3.8	19.96	24.06	22.58	21.52	23.68	19.95	22.48	22.32	22.95	21.92	42.38	23.98	5.95
EC:4.2.1.2	14.54	18.81	19.68	23.77	28.00	20.30	19.67	20.16	24.74	22.70	32.79	22.29	4.72
EC:4.2.1.3	33.31	29.83	34.13	23.43	28.96	41.10	44.43	37.46	35.39	38.11	69.02	37.74	11.35
EC:6.2.1.4	19.96	24.06	22.58	21.52	23.68	19.95	22.48	22.32	22.95	21.92	42.38	23.98	5.95
EC:6.4.1.1	14.54	18.81	19.68	23.77	28.00	20.30	19.67	20.16	24.74	22.70	32.79	22.29	4.72

Table 2.7: Enzyme participation levels for all enzymes across all data points for 2m depth in the metabolic pathway ec00020. Two rightmost columns are means and standard deviations of enzyme participation levels.

CHAPTER 3

ASSESSING THE LEVELS OF ENZYME EXPRESSION AND METABOLIC PATHWAY ACTIVITY IN MICE, BOTH INFECTED AND UNINFECTED WITH *BORRELIA BURGENDORFERI*

Mice have been widely used as experimental subjects in immunology, and studying their immune responses has provided valuable insights into the human immune system^{55,68,7,52}. Since mice and humans share a significant portion of their protein-coding gene sequences, analyzing the metabolic pathways of mice with different immune responses is crucial for gaining a deeper understanding of the human immune system^{50,39,73,31,47,24}. To comprehend the biochemical and metabolic changes that may occur in humans during stress or disease, it is vital to measure the functional activity, enrichment, and interaction of metabolic pathways in rodent groups with opposing health conditions^{18,4,27}. However, quantifying the contribution of individual enzymes to the activity of multi-enzyme metabolic pathways remains a challenging task, despite many advances in using biomolecules such as DNA, RNA, and enzymes. To address this issue, this study analyzes differentially active metabolic pathways from RNA sequencing data to generate an efficient model for understanding metabolic pathway activity changes^{70,19,54,66}.

Although RNA-Seq data exploration has been aided by advances in high-throughput sequencing, analyzing metabolic pathway activity changes in organisms with varying health conditions remains difficult. Existing pathway analysis tools often yield variable conclusions about the activity of pathways based on RNA data^{32,40,75,65}. To overcome these challenges, we developed a workflow that employs a Maximum Likelihood-based model and annotations based on the KEGG³⁶ database to estimate transcript frequency, enzyme expression, enzyme participation in pathways,

and metabolic pathway activity in microbial communities^{63,64}.

In this paper, we test this model using transcriptomic data from mice infected with *Borrelia burgdorferi*, an agent of Lyme disease, and their uninfected controls. The data describes the infected as well as the uninfected groups of two rodent species - *C3H*, a laboratory mouse strain and *Peromyscus leucopus* to elucidate the complex metabolic pathway activity changes between rodents with inherent tolerance to *B. burgdorferi* infection (*P. leucopus* mice) and those that develop Lyme disease (a laboratory strain of *C3H* mice). The proposed methodology is to use a maximum likelihood estimate to infer the pathway activity considering an enzyme's participation. First, we filtered mouse specific metabolic pathways from the KEGG database and merged the expression of enzymes represented by the same group of genes. We adjusted our EM algorithm based pipeline and improved it using enzyme participation level in each pathway and then used these estimations for more accurate predictions of pathway activity⁶⁴. Our contributions include:

- (a) estimation metabolic enzyme expression, find groups of rodents' enzymes that are represented by the same group of genes
- (b) estimation enzyme-in-pathways coefficients and demonstrate that they are more stable than for microbial communities in⁶⁴. Also we show that these coefficients do not significantly vary across species and infected and uninfected mice
- (c) differential analysis of metabolic pathway activity in *P. leucopus* and a laboratory strain of *C3H* across species uninfected and infected with *B. burgdorferi*

The remaining sections of the paper are structured as follows. In the subsequent section, we

present the pipeline of our software framework, along with several EM-based algorithms that are utilized to estimate enzyme expression and metabolic pathway activity among multiple types of rodents. Next, we provide a detailed account of our data, including sequencing data and the extraction of metabolic enzymes and pathways. Finally, we employ our findings to perform a statistical validation of the proposed pipeline.

3.1 Methods

Previously we created a pipeline for estimating metabolic pathway activity levels in a microbial community⁶⁴. We investigated the variation in pathway activity within a microbial community under different conditions. Analyzing a metagenomic community can be challenging due to the diversity and abundance of species present in the samples, making it difficult to interpret the results accurately. This complexity arises from the fact that metagenomic communities consist of a variety of organisms with distinct genetic backgrounds, which interact in complex ways to drive community-level functions and dynamics. Additionally, the genetic material in a metagenome is often fragmented, making it harder to identify and analyze specific pathways and enzymes.

3.1.1 Pipeline for estimating metabolic pathway activity of C3H and P. leucopus

Below we describe our novel metabolic pathway activity pipeline *EMPathways2* (see Fig 3.2) that is used for estimating pathway activity in mice. These models are resolved using the EM algorithm. The proposed inferences are highlighted in red (see Fig. 1).

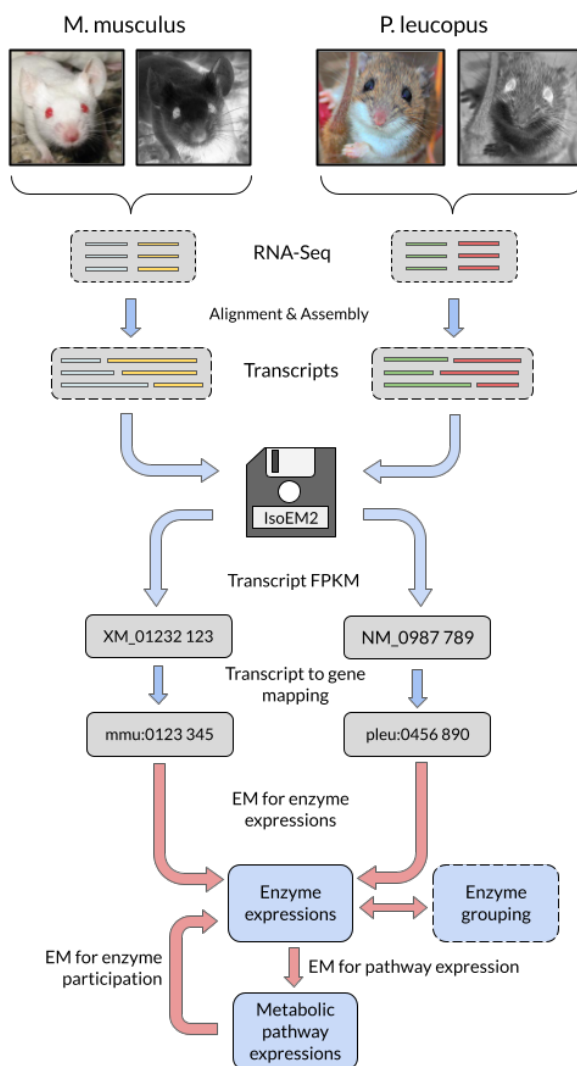


Figure 3.1: Full pipeline for metabolic pathway analysis for rodent samples. The RNA-Seq data obtained from rodents are sequenced, then raw reads are mapped into genes. The genes containing obtained contigs are further mapped into the enzyme-pathway database. Gene expression is obtained using IsoEM2⁴⁹. Then we estimate enzyme expression using gene expression. Finally, the pathway activity level and enzyme participation coefficients are estimated in the feedback loop.

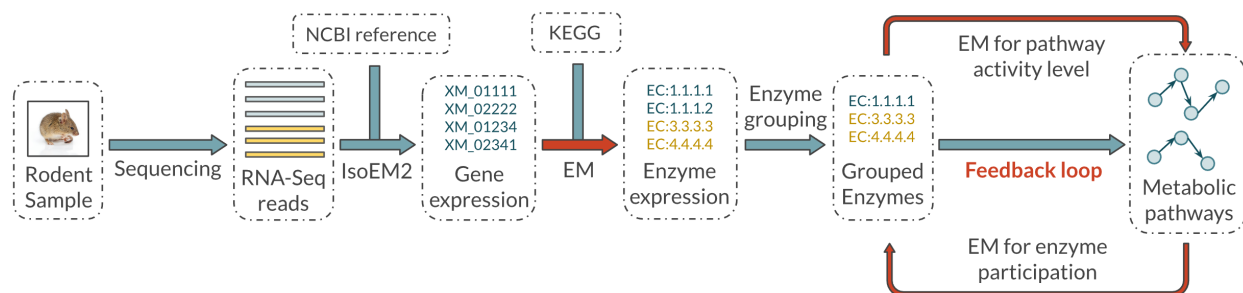


Figure 3.2: EMPathways2 pipeline for metabolic pathway analysis for rodent samples. The RNA-Seq data obtained from rodents are sequenced, then raw reads are mapped into genes. The genes containing obtained contigs are further mapped into the enzyme-pathway database. Gene expression is obtained using IsoEM2⁴⁹. Then we estimate enzyme expression using gene expression. Finally, the pathway activity level and enzyme participation coefficients are estimated in the feedback loop.

The entire pipeline *EMPathways2* consists of the following five steps:

- The first step is the collection of samples from infected and uninfected rodent groups, which then get sequenced.
- RNA-Seq reads are mapped into reference transcriptomes of *C3H* and *P. leucopus* collected from NCBI reference database. The mapped reads were used by IsoEM2 to generate gene expression data⁴⁹.
- We use KEGG to establish the many-to-many correspondence between genes and enzymes (see Sec. 3.1.2). Then, using EM we estimate enzyme expressions based on gene expression (see first red arrow in Fig. 1).
- Unstable enzymes that converge inconsistently were identified, grouped, and collapsed (see Sec. 3.1.3).

- *The Feedback loop* is based on inferred enzyme expressions and metabolic pathway annotation. It simultaneously estimates enzyme participation coefficients and metabolic pathways activity levels (see Sec. 3.1.4).

3.1.2 Mapping between genes, enzymes and pathways for C3H and P. leucopus

KEGG metabolic pathway database has information on all metabolic pathways that occur in the living organisms. However, the scope of *EMPathways2* is to analyze metabolic pathways in rodents. We concentrate on 152 metabolic pathways and 2386 enzymes that play a significant role in mouse metabolism which is confirmed by literature referenced in PUBMED.

In order to compute metabolic pathway activity levels *EMPathways2* requires an input in a form of a correspondence between genes and enzymes as well as a dictionary of enzymes participating in metabolic pathways. Gene-enzyme as well as enzyme-pathway mappings were extracted from NCBI Entrez database⁹ for molecular biology as well as KEGG pathway database respectively and which provides consolidated access to nucleotide, protein sequence, gene-centered and genomic mapping data. We used KEGG's and NCBI's API to collect raw data allowing us to produce a correspondence of genes to enzymes and enzymes to metabolic pathways. We used the collected data to create sets of genes participating in production of every enzyme, as well as sets of enzymes required for functional activity of every metabolic pathway.

3.1.3 Enzyme grouping

There is a many-to-many correspondence between genes and enzymes which may pose challenges to computing enzymes expression. To approach this challenge we use a maximum likelihood

EM model to infer enzyme expression from gene expression which converges consistently in vast majority of cases. However, there are enzymes that share some genes as well as enzymes whose genes are entirely a subset of genes used for production of another enzyme. In some of those cases EM struggles to discern one such enzyme from its genetic relatives and in turn converges inconsistently from one run to another. The enzymes that fail to converge consistently are labeled unstable and grouped into clusters whose expression as a single entity converges consistently after every EM iteration.

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

Table 3.1: A pair of individually unstable enzymes that are stable when summed into a group.

After running a few iterations of gene-enzyme EM we observe clusters of enzymes whose expression varies individually, but is stable in groups. The unstable enzymes individual expressions vary from one run to another. However, summed together always converge to the same expression in every run.(see Table 3.1). This instability makes such groups of enzymes indistinguishable to our algorithm. To establish the groups accurately we run EM and produce enzyme expression values for every enzyme. We establish clusters by evaluating the grouped enzyme expressions which do not converge consistently individually, but the sum of their expressions always converges to the same value. As a result such enzymes must be treated as single entities. After all unstable enzyme groups are found, we collapse them into one (see Figure 3.3 (A)). The groups are collapsed to a single enzyme with the lowest EC number nomenclature. The collapsed group enzyme is then used to compute metabolic pathway expressions of all related pathways (see Figure 3.3 (B)).

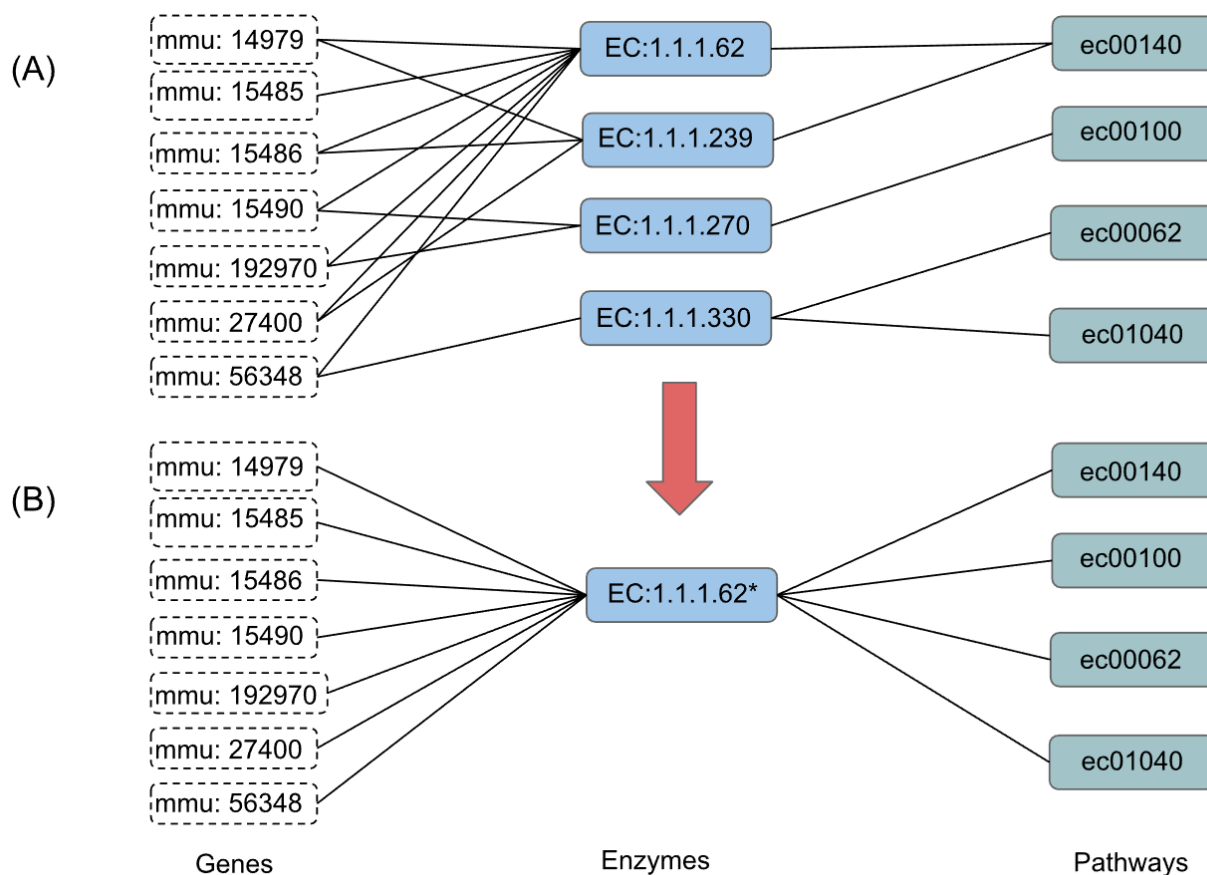


Figure 3.3: (A) Enzymes that cannot be distinguished from each other must be treated as groups. (B) Enzymes that are unstable are collapsed into a single enzyme with the lowest EC nomenclature number.

In total we found and collapsed 59 pairs, three triplets and one quadruple of indistinguishable enzymes. Fig 3.2 gives the list of triplets and a quadruplet found in mice. We have compared the list of collapsed enzymes for microbial communities found in⁶⁴ with the list of collapsed enzymes in rodents. We found out that there are 28 pairs common for these two data sets.

Triplet1	EC:1.1.1.51	EC:1.1.1.213	EC:1.1.1.188	
Triplet2	EC:6.3.4.13	EC:6.3.3.1	EC:2.1.2.2	
Triplet3	EC:2.1.3.2	EC:6.3.5.5	EC:3.5.2.3	
Quadruplets	EC:6.3.4.9	EC:6.3.4.10	EC:6.3.4.11	EC:6.3.4.15

Table 3.2: Three triplets and one quadruplet of collapsed enzymes.

3.1.4 Feedback loop for pathway activity level estimation

Each enzyme is initially assigned a participation coefficient of $1/|w|$, where $|w|$ is the total amount of enzymes in the pathway w . The *Feedback loop for pathway activity* updates the enzyme participation level by fitting expected enzyme expressions to the expressions estimated by *EM for enzyme expression*.

The initial estimate of the participation level of an enzyme e in a pathway w may be far from accurate. However, more accurate estimates of enzyme participation can lead to more accurate estimates for the pathway activity levels. Our algorithm first estimates enzyme expression from gene expression using the *EM for enzyme expression*. The **E-step** and **M-step** are ran in order to compute expected expression and compare it to the new estimate respectively. After computing enzyme expressions we then filter out enzymes with stable expressions and perform enzyme grouping on enzymes with unstable expressions. Pathway activity levels are in turn computed using *EM for pathway activity level*.

Following we estimate how well the computed activities f_w 's fit the enzyme expressions using the *EM for enzyme participation* depicted in Figure 3.2.

Together, EM for enzyme participation and EM for pathway activity level make up the *Feedback loop* for pathway activity level estimation. If the fit is not good enough, then the *Feedback loop for pathway activity level* is applied to update the enzyme participation levels p_{ew} 's with the *EM for enzyme participation* and then f_w 's are recomputed according to updated p_{ew} 's.

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

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

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

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

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

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

3.2 Datasets

3.2.1 Bacterial inoculum

Borrelia burgdorferi strain 297 (*B. burgdorferi* 297) was propagated in Barbour-Stoener-Kelly II medium supplemented with 6% rabbit serum (referred to here as BSK-II medium; Gemini Bio-Products, USA) under 2.5% CO₂ at 35°C. To prevent bacterial and fungal contamination of mouse tissue cultures, BSK-II medium was also supplemented with an antimicrobial cocktail (0.02 mg/mL phosphomycin, 0.05 mg/mL rifampicin, and 2.5 mg/mL amphotericin B).

3.2.2 Rodent infection

Six male C3H/HeJ (C3H) mice of 4-6 weeks of age (The Jackson Laboratory, USA) and 6 *P. leucopus* mice of 5-8 weeks (The Peromyscus Genetic Stock Center, the University of South Carolina,

USA) were divided into groups of 3 animals each²². Three C3H mice and 3 *P. leucopus* mice were subcutaneously (s.c.) inoculated in the shoulder region with 1×10^5 spirochetes of *B. burgdorferi* 297 per animal (100 μ L inoculum). The other six mice were s.c. inoculated with 100 μ L of sterile saline (the uninfected control groups). The mouse infection was confirmed by culturing 50 μ L of blood sampled from each infected mouse via maxillary bleed at day 7 post infection (pi) in BSK-II medium. Long-term infection was also confirmed by culturing other mouse tissues (ear pinnae, bladder, tibiotarsal joint, heart) harvested at day 70 pi and cultured in BSK-II as described⁶. The presence or absence of viable spirochetes was verified by weekly observing cultures via dark-field microscopy over a four-week period. At day 70 pi, harvested spleens were immediately preserved in Invitrogen RNAlater stabilization solution (Thermo Fisher Scientific, USA) and frozen at -80°C until further analysis.

3.2.3 RNA sequencing

Total RNA from spleens harvested from 6 infected and 6 uninfected mice was individually isolated by utilizing QIAGEN RNeasy Mini Kit (QIAGEN, USA) according to manufacturer's instruction. The concentration and quality of RNA were determined, respectively, via Qubit, Broad Range fluorometric assay (Thermo Fisher Scientific) and Agilent TapeStation 2200 system on the RNA screen tape (Agilent Technologies, USA). RNA was normalized to 80 ng/ μ L prior to utilizing the Illumina TruSeq Stranded mRNA LS library preparation kit. After individually constructed libraries were barcoded, the quality of libraries was assessed by using Agilent TapeStation 2200 D1000 DNA screen tape (Agilent Technologies). The libraries were normalized and pooled in equimolar concentration and then sequenced using Illumina NextSeq 500 75 cycle High Output kit at the Texas

A&M Institute for Genomics Sciences and Society (Texas A&M University, USA). As a result, approximately 400 million 75 base-pair, single-end sequencing reads were produced. The Illumina BaseSpace (basespace.illumina.com) was utilized to generate FASTQ data and demultiplex sequencing reads.

3.3 Results

We have applied the proposed pipeline *EMPathways2* to rodent RNA-Seq data. For each group of rodents, we compute the mean and the standard deviation for each pathway activity level. We categorize a metabolic pathway as having significantly (resp. slightly) different activity across conditions if its standard deviation intervals do not intersect (resp. its standard deviation intervals intersect but do not contain each other means) for different conditions. Note that if a metabolic pathway has significantly (resp. slightly) different activity, then the probability that the activity is the same is below 0.25% (resp. 5%).

The list of metabolic pathways with significantly different activity across infected/uninfected *C3H* (resp. *P. leucopus*) are in Tables 3.3,3.5. We found that four *C3H* metabolic pathways are expressed with differing activity levels. E.g. caffeine metabolism has a significant difference in its activity levels between the infected and uninfected groups. Note that the number of metabolic pathways of *P. leucopus* significantly affected by the infection is much higher than for *C3H* that can explain why *C3H* get sick after infection while *P. leucopus* do not show any symptoms.

The list of metabolic pathways with slightly different activity across infected/uninfected *C3H* (resp. *P. leucopus*) are in Tables 3.4,3.6. Note that the lists of these pathways are very different for

Pathway Name	ID	Infected Mice Mean \pm Std	Uninfected Mice Mean \pm Std
Caffeine metabolism	ec00232	84.48 \pm 1.069	82.888 \pm 0.357
Mucin type O-glycan biosynthesis	ec00512	0.873 \pm 0.666	2.205 \pm 0.656
Pentose & glucuronate interconversions	ec00040	273.774 \pm 0.896	269.624 \pm 1.82
Thiamine metabolism	ec00730	49.922 \pm 0.297	59.741 \pm 0.205

Table 3.3: *C3H* pathways with significant different activity level across infected and uninfected groups.

Pathway Name	ID	Infected Mice Mean \pm Std	Uninfected Mice Mean \pm Std
Ascorbate and aldarate metabolism	ec00053	139.789 \pm 0.958	142.04 \pm 1.581
Drug metabolism - cytochrome P450	ec00982	104.598 \pm 0.85	105.261 \pm 0.518
Glycine, serine and threonine metabolism	ec00260	50.586 \pm 0.807	48.544 \pm 2.094
Glycosaminoglycan degradation	ec00531	78.611 \pm 0.568	77.616 \pm 1.778
Glycosphingolipid biosynthesis-globo & isoglobo series	ec00603	198.785 \pm 8.711	202.718 \pm 1.443
Selenocompound metabolism	ec00450	141.024 \pm 23.292	159.357 \pm 1.326
Amino sugar and nucleotide sugar metabolism	ec00520	105.101 \pm 0.287	104.142 \pm 1.246
Arginine and proline metabolism	ec00330	102.133 \pm 0.884	100.602 \pm 0.933
Citrate cycle	ec00020	116.843 \pm 12.089	124.87 \pm 0.702
Fatty acid biosynthesis	ec00061	303.491 \pm 5.538	307.308 \pm 0.489
Fatty acid elongation	ec00062	67.066 \pm 8.073	71.807 \pm 0.022
Folate biosynthesis	ec00790	302.951 \pm 9.635	287.446 \pm 9.319
Glycolysis	ec00010	145.131 \pm 6.6	138.049 \pm 11.634
Lysine degradation	ec00310	13.663 \pm 3.617	8.986 \pm 3.171
Mannose type O-glycan biosynthesis	ec00515	136.003 \pm 20.316	152.586 \pm 6.335
Metabolism of xenobiotics by cytochrome P450	ec00980	69.32 \pm 0.17	68.617 \pm 0.827
N-Glycan biosynthesis	ec00510	221.444 \pm 2.738	227.992 \pm 5.498
O-glycan biosynthesis	ec00514	162.416 \pm 1.829	155.666 \pm 8.056
Other glycan degradation	ec00511	177.914 \pm 1.182	175.957 \pm 4.27
Pantothenate and CoA biosynthesis	ec00770	24.598 \pm 8.195	27.777 \pm 0.525
Pentose phosphate	ec00030	102.537 \pm 0.314	97.822 \pm 9.699
Propanoate metabolism	ec00640	212.329 \pm 1.465	201.314 \pm 9.563
Pyrimidine metabolism	ec00240	172.185 \pm 4.223	181.393 \pm 6.125
Sulfur metabolism	ec00920	33.591 \pm 7.459	36.213 \pm 2.483

Table 3.4: *C3H* pathways with slightly different activity level across infected and uninfected groups.

different mouse species.

Finally, we check how stable are the enzyme participation coefficients across different mice species (see Table 3.7). Note that the average relative standard deviation (RSD) for *C3H* is 2.7% in contrast to much higher RSD for 8.9% for *P. leucopus*. That can be caused by that fact that lab mice *C3H* are genetically more similar to each other than the wild mice *P. leucopus*. Note that the average RSD for enzyme participation coefficients in the microbial community for the same metabolic pathway (ec00620) is 34.8% which is significantly higher (see⁶⁴) than RSD for mice.

Pathway Name	ID	Infected Mice Mean \pm Std	Uninfected Mice Mean \pm Std
Arginine and proline metabolism	ec00330	108.443 \pm 3.567	103.845 \pm 1.015
D-Amino acid metabolism	ec00470	218.092 \pm 0.626	206.601 \pm 7.797
Glycerophospholipid metabolism	ec00564	78.228 \pm 0.336	77.621 \pm 0.172
Glycine, serine and threonine metabolism	ec00260	49.423 \pm 0.728	47.543 \pm 0.119
One carbon pool by folate	ec00670	66.566 \pm 0.204	67.377 \pm 0.301
Selenocompound metabolism	ec00450	103.557 \pm 25.685	137.99 \pm 8.249
Starch and sucrose metabolism	ec00500	64.353 \pm 1.33	66.401 \pm 0.433
Tryptophan metabolism	ec00380	98.223 \pm 0.896	102.88 \pm 0.892
ascorbate and aldarate metabolism	ec00780	24.271 \pm 0.578	25.417 \pm 0.049
Ascorbate and aldarate metabolism	ec00053	131.871 \pm 1.17	136.458 \pm 0.912
Citrate cycle	ec00020	116.276 \pm 10.912	128.679 \pm 0.663
Glycosaminoglycan biosynthesis-heparan sulfate/heparin	ec00534	85.392 \pm 1.203	90.012 \pm 1.656
Glycosaminoglycan biosynthesis-keratan sulfate	ec00533	351.816 \pm 1.994	342.511 \pm 1.023
Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	ec00563	348.609 \pm 1.349	353.073 \pm 1.766
Linoleic acid metabolism	ec00591	440.035 \pm 10.893	423.801 \pm 1.7
Other glycan degradation	ec00511	164.744 \pm 2.361	135.58 \pm 0.722
Pentose phosphate	ec00030	103.646 \pm 0.475	104.649 \pm 0.247
Pyrimidine metabolism	ec00240	167.062 \pm 0.407	179.749 \pm 11.62
Valine, leucine and isoleucine biosynthesis	ec00290	77.081 \pm 2.466	83.37 \pm 2.5
Valine, leucine and isoleucine degradation	ec00280	113.366 \pm 4.269	103.142 \pm 5.56
Vitamin B6 metabolism	ec00750	56.675 \pm 0.557	52.601 \pm 0.395

Table 3.5: *P. leucopus* pathways with significant different activity level across infected and uninfected groups.

Pathway Name	ID	Infected Mice Mean \pm Std	Uninfected Mice Mean \pm Std
Amino sugar and nucleotide sugar metabolism	ec00520	104.8 \pm 1.365	102.262 \pm 2.796
Arachidonic acid metabolism	ec00590	163.557 \pm 0.317	162.903 \pm 1.17
Nitrogen metabolism	ec00910	102.949 \pm 0.324	101.743 \pm 0.897
Folate biosynthesis	ec00790	314.768 \pm 6.619	307.406 \pm 1.934
Fructose and mannose metabolism	ec00051	30.991 \pm 0.403	30.493 \pm 0.193
Glutathione metabolism	ec00480	45.435 \pm 0.73	44.655 \pm 0.569
Glycosphingolipid biosynthesis-lacto & neolacto series	ec00601	29.256 \pm 6.267	41.326 \pm 6.14
Glyoxylate and dicarboxylate metabolism	ec00630	108.993 \pm 11.861	120.784 \pm 8.979
Inositol phosphate metabolism	ec00562	39.927 \pm 0.154	39.575 \pm 0.588
Porphyrin metabolism	ec00860	278.62 \pm 1.556	275.075 \pm 6.258
Riboflavin metabolism	ec00740	117.214 \pm 8.465	105.181 \pm 5.623
Steroid hormone biosynthesis	ec00140	131.347 \pm 2.431	132.832 \pm 0.644
Thiamine metabolism	ec00730	58.599 \pm 0.158	58.15 \pm 0.715
Tyrosine metabolism	ec00350	70.298 \pm 2.634	66.036 \pm 2.207
Ubiquinone and other terpenoid-quinone biosynthesis	ec00130	194.363 \pm 4.996	201.709 \pm 5.371

Table 3.6: *P. leucopus* pathways with slightly different activity level across infected and uninfected groups.

3.4 Conclusions

In this paper we propose an improved maximum likelihood-based pipeline for the estimation of metabolic pathway activity in mice using the KEGG pathway database. Specifically, the proposed approach uses EM-based algorithms to estimate enzyme expression, enzyme participation levels

ec00620	Infected C3H			Uninfected C3H			%RSD	Infected P. leucopus			Uninfected P. leucopus			%RSD
	EC:1.1.1.1	.110	.107	.113	.106	.109		.112	2.501	.054	.061	.049	.051	
EC:1.5.8.3	.027	.025	.026	.026	.026	.026	2.433	.035	.031	.038	.033	.035	.041	10.039
EC:3.1.3.3	.027	.025	.026	.026	.026	.026	2.433	.035	.031	.038	.033	.035	.041	10.039
EC:2.1.2.10	.034	.034	.035	.035	.034	.034	1.504	.028	.030	.027	.027	.025	.025	7.027
EC:5.1.1.18	.032	.038	.033	.037	.034	.031	8.157	.013	.016	.011	.015	.012	.012	14.740
EC:1.4.3.21	.050	.055	.055	.054	.054	.051	4.019	.028	.029	.019	.027	.025	.024	14.269
EC:2.6.1.52	.059	.058	.060	.059	.061	.060	1.763	.047	.050	.042	.043	.041	.040	8.826
EC:2.1.1.20	.027	.025	.026	.026	.026	.026	2.433	.035	.031	.038	.033	.035	.041	10.039
EC:2.7.1.165	.095	.086	.087	.088	.087	.088	3.696	.077	.067	.074	.061	.060	.059	11.586
EC:1.5.3.1	.027	.025	.026	.026	.026	.026	2.433	.035	.031	.038	.033	.035	.041	10.039
EC:2.3.1.29	.027	.025	.026	.026	.026	.026	2.433	.035	.031	.038	.033	.035	.041	10.039
EC:4.1.2.48	.027	.025	.026	.026	.026	.026	2.433	.035	.031	.038	.033	.035	.041	10.039
EC:1.1.99.1	.027	.025	.026	.026	.026	.026	2.433	.035	.031	.038	.033	.035	.041	10.039
EC:2.3.1.37	.055	.052	.055	.055	.056	.055	2.499	.072	.066	.074	.067	.070	.077	5.909
EC:2.1.2.1	.051	.051	.052	.052	.052	.050	1.591	.038	.041	.035	.036	.034	.033	8.093
EC:1.1.1.95	.050	.048	.050	.050	.050	.050	1.644	.050	.050	.048	.048	.046	.049	3.127
EC:1.1.1.103	.027	.025	.026	.026	.026	.026	2.433	.035	.031	.038	.033	.035	.041	10.039
EC:2.1.4.1	.049	.047	.049	.049	.049	.050	2.013	.049	.049	.044	.047	.046	.051	5.252
EC:4.2.1.22	.050	.048	.050	.050	.050	.050	1.644	.050	.050	.048	.048	.046	.049	3.127
EC:4.4.1.1	.061	.059	.061	.061	.060	.060	1.353	.047	.050	.043	.045	.042	.042	7.112
EC:4.3.1.17	.050	.048	.050	.050	.050	.050	1.644	.050	.050	.048	.048	.046	.049	3.127
EC:1.4.3.4	.062	.070	.064	.070	.067	.068	4.864	.028	.033	.023	.028	.027	.026	11.895
EC:1.4.3.3	.046	.053	.047	.052	.049	.045	6.711	.019	.023	.015	.021	.017	.017	15.771
EC:1.8.1.4	.088	.090	.089	.091	.090	.090	1.152	.042	.049	.034	.043	.039	.040	12.040
EC:2.1.1.5	.050	.048	.050	.050	.050	.050	1.644	.050	.050	.048	.048	.046	.049	3.127
EC:2.1.1.2	.049	.047	.049	.049	.049	.050	2.013	.049	.049	.044	.047	.046	.051	5.252

Table 3.7: The enzyme expression coefficients and relative standard deviations (%RSD) for the enzyme participation coefficients in pathway ec00620.

in pathways, and metabolic pathway activity.

The proposed metabolic pathway analysis was applied to the RNA-Seq data from 12 mice samples collected from *C3H* and *P. leucopus* with half them infected by *B. burgdorferi* 297.

The key findings of the study are as follows:

- The enzyme expression and metabolic pathway activity levels are significantly more stable when considering enzyme participation coefficients.
- The infection affects metabolism of both mice while for *P. leucopus*, the affect is more significant than for *C3H*.
- The enzymes participation coefficients vary insignificantly for *C3H* in contrast to higher variation for *P. leucopus* and much higher variation for microbial communities.

CHAPTER 4

EMPATHWAYS2: ESTIMATION OF ENZYME EXPRESSION AND METABOLIC PATHWAY ACTIVITY USING RNS-SEQ READS

In this chapter, I outline an approach to analyze metatranscriptomic data, focusing on the assessment of differential enzyme expression and metabolic pathway activities using a novel bioinformatics software tool, EMPathways2. The analysis pipeline commences with raw data originating from a sequencer and concludes with an output of enzyme expressions and an estimate of metabolic pathway activities.

The initial step involves aligning specific transcriptomes assembled from RNA-Seq data using Bowtie2, followed by gene expression data acquisition with IsoEM2. Subsequently, the pipeline proceeds to quality assessment and preprocessing of the input data, ensuring accurate estimates of enzymes and their differential regulation. Upon completion of the preprocessing stage, EMPathways2 is employed to decipher the intricate relationships between genes, enzymes, and pathways.

An online repository containing sample data has been made available, alongside custom Python scripts designed to modify the output of the programs within the pipeline for diverse downstream analyses. This chapter highlights the technical aspects and practical applications of using EMPathways2, which facilitates the advancement of transcriptome data analysis and contributes to a deeper understanding of the complex regulatory mechanisms underlying living systems.

4.1 Introduction

Understanding metabolic pathways is crucial for elucidating the complex regulatory mechanisms underlying cellular functions, as these pathways represent the interconnected series of biochemical

reactions that maintain and modulate the dynamic balance of a living system. In this chapter, we present a comprehensive guide to EMPathways2, a novel bioinformatics software tool designed to unravel the mysteries of metabolic pathways by analyzing metatranscriptomic data and quantifying differential enzyme expression.

Metabolic pathways play a vital role in cellular processes such as energy production, biosynthesis of biomolecules, and detoxification. Accurate assessment of differential expression in these pathways provides invaluable insights into an organism's response to environmental changes, gene regulation, and potential therapeutic targets for various diseases. Traditional methods, such as Gene Set Enrichment Analysis (GSEA)⁶⁷, have been widely employed to estimate gene expression and activities under multiple conditions^{11,20,34}. However, GSEA has limitations in its ability to estimate activity levels⁷⁴, offering only a binary indication of gene presence or absence. Metabolic pathways play a vital role in cellular processes such as energy production, biosynthesis of biomolecules, and detoxification. Accurate assessment of differential expression in these pathways provides invaluable insights into an organism's response to environmental changes, gene regulation, and potential therapeutic targets for various diseases. Traditional methods, such as Gene Set Enrichment Analysis (GSEA)⁶⁷, have been widely employed to estimate gene expression and activities under multiple conditions^{11,20,34}. However, GSEA has limitations in its ability to estimate activity levels, offering only a binary indication of gene presence or absence.

In this protocol, we describe a pipeline titled EMPathways2, which has been developed to address such limitations and offer a more comprehensive view of metabolic pathway activities. In contrast to GSEA, EMPathways2 enables biologists to calculate metabolic pathway activity levels

and infer enzyme expression across multiple conditions with greater accuracy and detail. This enhanced capacity allows researchers to delve deeper into the functional roles of genes and their products within the context of the entire metabolic network, ultimately facilitating a more thorough understanding of the intricate relationships between genes, and enzymes.

In this chapter, we will provide a step-by-step guide on using EMPathways2 for metatranscriptomic data analysis and discuss the technical aspects of the software. By the end of this chapter, readers will gain a solid understanding of the advantages EMPathways2 offers over traditional methods like GSEA and how this powerful tool can be employed to advance the field of transcriptome data analysis.

4.2 Materials

4.2.1 Software and Data

Before starting the tasks outlined in each section of this protocol, a few initial steps need to be accomplished. These involve acquiring the data from the designated repository and installing the required software listed in Table 1. To use the tools in this pipeline, a GNU/Linux command line environment is necessary, which can be achieved via a standalone installation or a virtual machine (VM) as long as adequate resources are allocated to the VM. For Windows users, the Windows Subsystem for Linux is a viable option, as it offers a complete GNU/Linux command line environment within Windows without the need for virtualization and resource allocation.

Common distributions like Ubuntu and Debian might already include many essential tools in their repositories for effortless installation. However, these versions could be outdated, potentially

leading to workflow complications. Although several programs are natively available for Mac OS X, users might have to compile them from source packages. In contrast, binaries are typically accessible for GNU/Linux.

Table 1 contains a general guide on software installation and modifying the \$PATH. Nonetheless, the steps needed for compiling software from source can differ among programs. It is advised that users download the most recent available binary and add it to the /usr/local/bin directory whenever possible.

The commands featured in this pipeline are designed for a standard desktop with 4 CPU cores and 8 GB of RAM. Handling larger datasets might require more RAM, and users with extra cores and memory can modify the relevant parameters by referring to the manual for each particular program.

Software	URL
DOE-JGI	https://jgi.doe.gov/
STAR	https://github.com/alexdobin/STAR
HISAT2	http://daehwankimlab.github.io/hisat2/
KEGG	https://www.kegg.jp/
IsoEM2	https://github.com/mandricigor/isoem2
SAM Format Specification	https://samtools.github.io/hts-specs/SAMv1.pdf
Windows Subsystem for Linux	https://docs.microsoft.com/en-us/windows/wsl/install-win10
Compiling Software on Linux	https://itsfoss.com/install-software-from-source-code/

Table 4.1: Software and URLs

This estimation of enzyme expression and pathways activity level requires procuring the meta-

transcriptomic dataset of a plankton community of the surface waters of the Northern Gulf from the repository. RNA-seq data was sampled along with six different environmental parameters and samples were gathered from two different depths, 2 meters and 18 meters. An initial annotation of RNA-seq data was acquired through the the DOE-JGI Metagenome Annotation Pipeline (MAP v.4; JGI portal)³⁰. It is available in the JGI Genome Online database through GOLD ID Gs0110.

4.3 Methods

4.3.1 Preparation and Quality Assessment

The sample dataset used to run EMPathways2 is a metatranscriptomic dataset of a plankton community from the surface waters of the Northern Gulf of Mexico. The RNA-seq data and respective environmental parameters were sampled in July 2015 at 2 depths - 2 meters and 18 meters, every 4 hours throughout 48 hours totaling in 13 samples per depth. Six environmental parameters - including PAR (photosynthetic active radiation) and seawater dissolved oxygen concentration, density, salinity, temperature, and chlorophyll concentration were measured for each sample. All datasets are publicly available through the JGI Genomes Online (GOLD) database via GOLD ID Gs0110190. Out of 26 samples, four samples (Day1, 12:00, 18m; Day 2, 20:00, 2m; Day 3, 08:00, 2m; Day 3, 12:00, 18m) were discarded as they did not contain enough reads to assemble transcripts for our pipeline (see Table 2.1).

The preliminary annotation of RNA-seq data was obtained using the DOE-JGI Metagenome Annotation Pipeline (MAP v.4) (JGI portal). The MAP pipeline comprises feature prediction, including the identification of protein-coding genes. Initially, the MEGAHIT metagenome assembler

assembles RNA-Seq reads into scaffolds. Subsequently, various software suites (GeneMark.hmm, MetaGeneAnnotator, Prodigal, and FragGeneScan) predict genes on the assembled scaffolds. The MAP pipeline employs enzyme commission (EC) numbers to annotate genes, which is a necessary input for the model. These annotations are acquired through homology searches (using USEARCH) within a nonredundant protein-sequence database (maxhits = 50, e-value = 0.1), where each protein is assigned to a KEGG Orthology (KO) group. For each KO, the top five hits, provided that the identity score is at least 30% and 70% of the protein length is matched, are utilized. The KO IDs are then converted into EC numbers using the KEGG KO to EC mapping.

4.3.2 Align RNA-seq read

The first step in our pipeline involves aligning RNA-seq reads to a reference genome using an aligner tool such as HISAT2 or STAR^{38,15}. For this particular analysis we picked STAR. The command for this step is as follows:

```
$ mkdir STAR_Output  
  
$ STAR --runThreadN NumberOfThreads --genomeDir STAR_Genome_Index  
--readFilesCommand zcat --readFilesIn Read1.fq.gz Read2.fq.gz  
--outFileNamePrefix STAR_Output/
```

In this command, `NumberOfThreads` specifies the number of threads for parallel processing, while `STAR_Genome_Index` refers to the directory containing the genome index files.

`Read1.fq.gz` and `Read2.fq.gz` denote the input read files in FASTQ format, which can be compressed using gzip. The output files will be saved in the `STAR_Output` directory.

4.3.3 Produce gene expression data using IsoEM2

After the alignment step, the IsoEM2 is used to estimate gene expression levels from the aligned reads. Before running IsoEM2, input files need to be prepared, which include a set of known isoforms in GTF format and a file with aligned reads in SAM format. To ensure proper analysis, the aligned reads must be sorted by read name. If you are uncertain whether your reads are sorted, you can run the following command to sort the file:

```
sort -k 1,1 aligned_reads.sam > aligned_reads_sorted.sam
```

Once the files are prepared, you can execute IsoEM2 using the command line as follows:

```
isoem2 -G genes.gtf -m 200 -d 20 aligned_reads_sorted.sam
```

In the command above, `gene_annotation.gtf` represents the gene annotation file in GTF format, and `manifest.txt` is a file containing the paths to the aligned reads in SAM/BAM format produced by the STAR aligner. The output files will be saved in the specified output directory.

4.3.4 Calculate enzyme and metabolic pathway expression with EMPathways2

Finally, the EMPathways2 tool calculates enzyme and metabolic pathway expression based on the gene expression data obtained from IsoEM2. The command for this process is:

```
$ python empathways2.py -ge gene_enzyme_file  
-epd enzyme_pathway_dictionary -gexp gene_expression_file  
-eo enzyme_output -po pathway_output  
--theta_e 0.0001 --theta_p 0.0001
```

In this command, `gene_enzyme_file` is the input file containing gene-enzyme relationships, and `enzyme_pathway_dictionary` refers to the Enzyme Pathway Dictionary (EPD) file. The `gene_expression_file` contains the gene expression data, such as FPKM values, derived from IsoEM output. The optional `-eo` and `-po` flags allow users to specify output filenames for enzymes and pathways, respectively. The `--theta_e` and `--theta_p` parameters represent the convergence thresholds for enzyme and pathway calculations, with default values set to 0.000001.

EMPathways2 pipeline generates two output files as a result of the analysis: the enzyme expression output file and the pathway activity output file. These files provide valuable information on the enzyme expression levels and the metabolic pathway activity levels, respectively. The enzyme expression output file contains enzyme expression levels for each enzyme in the dataset. Each line in the file represents the expression level of a single enzyme, formatted as `enzyme_id:expression`. This file enables researchers to examine the expression levels of individual enzymes and identify those that may play crucial roles in the biological processes being studied. The pathway activity output file provides information on the activity levels of metabolic pathways. Each line in this file represents the activity level of a metabolic pathway, formatted as `pathway_id:activity_level`. By analyzing these pathway activity levels, researchers can gain insights into the overall metabolic landscape of the organisms under investigation and identify key pathways that may be involved in the response to specific conditions or stimuli.

Both the enzyme and pathway IDs use KEGG nomenclature, which offers several benefits. Firstly, KEGG provides accurate and experimentally verified mappings, ensuring reliable results in the analysis. Secondly, the use of KEGG streamlines the analysis process by integrating well-

established biological knowledge, allowing for easier interpretation of the results. Furthermore, KEGG nomenclature facilitates comparison and integration with other studies that utilize KEGG data, promoting consistency and reproducibility in the research community. Lastly, relying on KEGG mappings saves time and resources compared to creating custom mappings, allowing researchers to focus on the biological interpretation and potential implications of their findings.

By utilizing the EMPathways2 pipeline and its output files, researchers can gain valuable insights into the enzyme expression and metabolic pathway activity levels in their datasets, paving the way for a deeper understanding of the complex interactions and regulatory mechanisms that govern biological processes.

4.4 Notes

EMPathways2 pipeline is designed to be highly adaptable and user-friendly, making it compatible with a wide range of RNA-Seq analyses. Although no software can replace the expertise of a skilled bioinformatician, this pipeline serves as an excellent starting point for researchers who are new to RNA-Seq analysis or those who want to delve deeper into the computational aspects of their work. EMPathways2 also provides a solid foundation for those interested in building their own customized pipelines.

We have optimized the pipeline for typical total mRNA-derived RNA-Seq datasets. However, it is important to recognize that certain situations may necessitate adjustments to the default parameters. EMPathways2 is particularly valuable for researchers working with organisms lacking established genomic references, a challenge frequently encountered in plant research.

Ultimately, the quality of the results produced by any pipeline or bioinformatician is contingent upon the accuracy of the experimental data. Researchers must carefully consider the experimental conditions being compared and employ appropriate pipelines for reliable and robust conclusions.

REFERENCES

1. *Essays Biochem*, volume 64 (4): 607–647. 2020.
2. S. Al Seesi, Y. T. Tiagueu, A. Zelikovsky, and I. I. Măndoiu. Bootstrap-based differential gene expression analysis for RNA-Seq data with and without replicates. *BMC Genomics*, 15 Suppl 8:S2, Nov. 2014.
3. T. Alloui, I. Boussebough, A. Chaoui, A. Z. Nouar, and M. C. Chettah. Usearch: A meta search engine based on a new result merging strategy. In *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, volume 01, pages 531–536, Nov. 2015.
4. I. C. Arnold, N. Dehzad, S. Reuter, H. Martin, B. Becher, C. Taube, and A. Müller. Helicobacter pylori infection prevents allergic asthma in mouse models through the induction of regulatory T cells. *J. Clin. Invest.*, 121(8):3088–3093, Aug. 2011.
5. A. Arsalan and H. Younus. Enzymes and nanoparticles: Modulation of enzymatic activity via nanoparticles. *Int. J. Biol. Macromol.*, 118(Pt B):1833–1847, Oct. 2018.
6. M. Batool, A. E. Hillhouse, Y. Ionov, K. J. Kochan, F. Mohebbi, G. Stoica, D. W. Threadgill, A. Zelikovsky, S. D. Waghela, D. J. Wiener, and A. S. Rogovskyy. New zealand white rabbits effectively clear borrelia burgdorferi B31 despite the bacterium's functional vls antigenic variation system. *Infect. Immun.*, 87(7), July 2019.
7. L. K. Beura, S. E. Hamilton, K. Bi, J. M. Schenkel, O. A. Odumade, K. A. Casey, E. A. Thompson, K. A. Fraser, P. C. Rosato, A. Filali-Mouhim, R. P. Sekaly, M. K. Jenkins, V. Vezys,

- W. N. Haining, S. C. Jameson, and D. Masopust. Normalizing the environment recapitulates adult human immune traits in laboratory mice. *Nature*, 532(7600):512–516, Apr. 2016.
8. N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter. Erratum: Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, 34(8):888, Aug. 2016.
 9. J. P. Buchmann and E. C. Holmes. Entrezpy: a python library to dynamically interact with the NCBI entrez databases. *Bioinformatics*, 35(21):4511–4514, Nov. 2019.
 10. C. Carvalho and M. Caramujo. The various roles of fatty acids, 2018.
 11. N. R. Clark and A. Ma’ayan. Introduction to statistical methods for analyzing large data sets: gene-set enrichment analysis. *Sci. Signal.*, 4(190):tr4, Sept. 2011.
 12. D. W. Deamer. The first living systems: a bioenergetic perspective. *Microbiol. Mol. Biol. Rev.*, 61(2):239–261, June 1997.
 13. W. C. DeLoache, Z. N. Russ, and J. E. Dueber. Towards repurposing the yeast peroxisome for compartmentalizing heterologous metabolic pathways. *Nat. Commun.*, 7:11152, Mar. 2016.
 14. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via theEMAlgorithm, 1977.
 15. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, Jan. 2013.
 16. M. Donato, Z. Xu, A. Tomoiaga, J. G. Granneman, R. G. Mackenzie, R. Bao, N. G. Than, P. H. Westfall, R. Romero, and S. Draghici. Analysis and correction of crosstalk effects in pathway analysis. *Genome Res.*, 23(11):1885–1893, Nov. 2013.

17. F. Dündar, L. Skrabanek, and P. Zumbo. Introduction to differential gene expression analysis using RNA-seq. *Applied Bioinformatics Core/Weill Cornell Medical College*, pages 1–67, 2015.
18. J. C. Edwards. The effects of trichinella spiralis infection on social interactions in mixed groups of infected and uninfected male mice. *Anim. Behav.*, 36(2):529–540, Apr. 1988.
19. B. Efron and R. Tibshirani. On testing the significance of sets of genes, 2007.
20. H. El-Saghire, H. Thierens, P. Monsieurs, A. Michaux, C. Vandevoorde, and S. Baatout. Gene set enrichment analysis highlights different gene expression profiles in whole blood samples x-irradiated with low and high doses. *Int. J. Radiat. Biol.*, 89(8):628–638, Aug. 2013.
21. S. P. France, L. J. Hepworth, N. J. Turner, and S. L. Flitsch. Constructing biocatalytic cascades: In vitro and in vivo approaches to de novo Multi-Enzyme pathways. *ACS Catal.*, 7(1):710–724, Jan. 2017.
22. A. M. Gaber, I. Mandric, C. Nitirahardjo, H. Piontkivska, A. E. Hillhouse, D. W. Threadgill, A. Zelikovsky, and A. S. Rogovskyy. Comparative transcriptome analysis of peromyscus leucopus and C3H mice infected with the lyme disease pathogen. *Front. Cell. Infect. Microbiol.*, 13, 2023.
23. G. Gago, L. Diacovich, A. Arabolaza, S.-C. Tsai, and H. Gramajo. Fatty acid biosynthesis in actinomycetes. *FEMS Microbiol. Rev.*, 35(3):475–497, May 2011.
24. N. C. Gassen, J. Hartmann, J. Zschocke, J. Stepan, K. Hafner, A. Zellner, T. Kirmeier, L. Kollmannsberger, K. V. Wagner, N. Dedic, G. Balsevich, J. M. Deussing, S. Kloiber, S. Lucae, F. Holsboer, M. Eder, M. Uhr, M. Ising, M. V. Schmidt, and T. Rein. Association of FKBP51

- with priming of autophagy pathways and mediation of antidepressant treatment response: evidence in cells, mice, and humans. *PLoS Med.*, 11(11):e1001755, Nov. 2014.
25. T. A. Gianoulis, J. Raes, P. V. Patel, R. Bjornson, J. O. Korbel, I. Letunic, T. Yamada, A. Paccanaro, L. J. Jensen, M. Snyder, P. Bork, and M. B. Gerstein. Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc. Natl. Acad. Sci. U. S. A.*, 106(5): 1374–1379, Feb. 2009.
 26. G. Hammes. *Enzyme Catalysis and Regulation*. Elsevier, Dec. 2012.
 27. C. Hayashi, C. V. Gudino, F. C. Gibson, 3rd, and C. A. Genco. Review: Pathogen-induced inflammation at sites distant from oral infection: bacterial persistence and induction of cell-specific innate immune inflammatory pathways. *Mol. Oral Microbiol.*, 25(5):305–316, Oct. 2010.
 28. S. M. Heinzemann, D. Chivall, D. M'Boule, D. Sinke-Schoen, L. Villanueva, J. S. Sinninghe Damsté, S. Schouten, and M. T. J. van der Meer. Comparison of the effect of salinity on the D/H ratio of fatty acids of heterotrophic and photoautotrophic microorganisms. *FEMS Microbiology Letters*, 362(10), 2015.
 29. Y. Hu and J. F. Holden. Citric acid cycle in the hyperthermophilic archaeon pyrobaculum islandicum grown autotrophically, heterotrophically, and mixotrophically with acetate. *J. Bacteriol.*, 188(12):4350–4355, 2006.
 30. M. Huntemann, N. N. Ivanova, K. Mavromatis, H. J. Tripp, D. Paez-Espino, K. Tennessen, K. Palaniappan, E. Szeto, M. Pillay, I.-M. A. Chen, A. Pati, T. Nielsen, V. M. Markowitz, and N. C. Kyrpides. The standard operating procedure of the DOE-JGI metagenome annotation

- pipeline (MAP v.4). *Stand. Genomic Sci.*, 11:17, Feb. 2016.
31. N. D. Huntington, C. A. J. Voshenrich, and J. P. Di Santo. Developmental pathways that generate natural-killer-cell diversity in mice and humans. *Nat. Rev. Immunol.*, 7(9):703–714, Sept. 2007.
 32. D. H. Huson, S. Mitra, H.-J. Ruscheweyh, N. Weber, and S. C. Schuster. Integrative analysis of environmental sequences using MEGAN4. *Genome Res.*, 21(9):1552–1560, Sept. 2011.
 33. D. Hyatt, G.-L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, and L. J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11:119, Mar. 2010.
 34. R. A. Irizarry, C. Wang, Y. Zhou, and T. P. Speed. Gene set enrichment analysis made simple. *Stat. Methods Med. Res.*, 18(6):565–575, Dec. 2009.
 35. H. J. Janßen and A. Steinbüchel. Fatty acid synthesis in escherichia coli and its applications towards the production of fatty acid based biofuels. *Biotechnol. Biofuels*, 7(1):7, Jan. 2014.
 36. M. Kanehisa. KEGG: Kyoto encyclopedia of genes and genomes, 2000.
 37. J. Z. Kaye. *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, 2004.
 38. D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, 37(8):907–915, Aug. 2019.
 39. S. K. Kim. Common aging pathways in worms, flies, mice and humans. *J. Exp. Biol.*, 210(Pt

- 9):1607–1612, May 2007.
40. K. M. Konwar, N. W. Hanson, A. P. Pagé, and S. J. Hallam. MetaPathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information. *BMC Bioinformatics*, 14:202, June 2013.
 41. A. Kumar, R. Gudiukaite, A. Gricajeva, M. Sadauskas, V. Malunavicius, H. Kamyab, S. Sharma, T. Sharma, and D. Pant. Microbial lipolytic enzymes – promising energy-efficient biocatalysts in bioremediation. *Energy*, 192(116674):116674, Feb. 2020.
 42. K. P. K. Lee, M. Dey, D. Neculai, C. Cao, T. E. Dever, and F. Sicheri. Structure of the dual enzyme ire1 reveals the basis for catalysis and regulation in nonconventional RNA splicing. *Cell*, 132(1):89–100, Jan. 2008.
 43. M. E. Lee, A. Aswani, A. S. Han, C. J. Tomlin, and J. E. Dueber. Expression-level optimization of a multi-enzyme pathway in the absence of a high-throughput assay. *Nucleic Acids Res.*, 41(22):10668–10678, Dec. 2013.
 44. D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*, 31(10):1674–1676, May 2015.
 45. R. Lister, B. D. Gregory, and J. R. Ecker. Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. *Curr. Opin. Plant Biol.*, 12(2):107–118, Apr. 2009.
 46. A. V. Lukashin and M. Borodovsky. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.*, 26(4):1107–1115, Feb. 1998.
 47. C. L. Mackall and R. E. Gress. Pathways of t-cell regeneration in mice and humans: implica-

- tions for bone marrow transplantation and immunotherapy. *Immunol. Rev.*, 157:61–72, June 1997.
48. I. Mandric, S. Knyazev, C. Padilla, F. Stewart, I. I. Măndoiu, and A. Zelikovsky. Metabolic analysis of metatranscriptomic data from planktonic communities. pages 396–402, 2017.
49. I. Mandric, Y. Temate-Tiagueu, T. Shcheglova, S. Al Seesi, A. Zelikovsky, and I. I. Măndoiu. Fast bootstrapping-based estimation of confidence intervals of expression levels and differential expression from RNA-Seq data. *Bioinformatics*, 33(20):3302–3304, Oct. 2017.
50. J. Margolin. Of mice, men, and the genome. *Genome Res.*, 10(10):1431–1432, Oct. 2000.
51. S. Marguerat, B. T. Wilhelm, and J. Bähler. Next-generation sequencing: applications beyond genomes. *Biochem. Soc. Trans.*, 36(Pt 5):1091–1096, Oct. 2008.
52. D. Masopust, C. P. Sivula, and S. C. Jameson. Of mice, dirty mice, and men: Using mice to understand human immunology. *J. Immunol.*, 199(2):383–388, July 2017.
53. E. Maza, P. Frasse, P. Senin, M. Bouzayen, and M. Zouine. Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments: A matter of relative size of studied transcriptomes. *Commun. Integr. Biol.*, 6(6):e25849, Nov. 2013.
54. C. Mitrea, Z. Taghavi, B. Bokanizad, S. Hanoudi, R. Tagett, M. Donato, C. Voichița, and S. Drăghici. Methods and approaches in the topology-based analysis of biological pathways, 2013.
55. D. E. Mosier, R. J. Gulizia, S. M. Baird, and D. B. Wilson. Transfer of a functional human immune system to mice with severe combined immunodeficiency. *Nature*, 335(6187):256–259, Sept. 1988.

56. M. A. Nawaz, C. Chen, F. Shireen, Z. Zheng, H. Sohail, M. Afzal, M. A. Ali, Z. Bie, and Y. Huang. Genome-wide expression profiling of leaves and roots of watermelon in response to low nitrogen. *BMC Genomics*, 19(1):456, June 2018.
57. H. Noguchi, T. Taniguchi, and T. Itoh. MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res.*, 15(6):387–396, Dec. 2008.
58. X. Peng, Z. Chen, F. Farshidfar, X. Xu, P. L. Lorenzi, Y. Wang, F. Cheng, L. Tan, K. Mojumdar, D. Du, Z. Ge, J. Li, G. V. Thomas, K. Birsoy, L. Liu, H. Zhang, Z. Zhao, C. Marchand, J. N. Weinstein, Cancer Genome Atlas Research Network, O. F. Bathe, and H. Liang. Molecular characterization and clinical relevance of metabolic expression subtypes in human cancers. *Cell Rep.*, 23(1):255–269.e4, Apr. 2018.
59. A. P. Rajkumar, P. Qvist, R. Lazarus, F. Lescai, J. Ju, M. Nyegaard, O. Mors, A. D. Børghlum, Q. Li, and J. H. Christensen. Experimental validation of methods for differential gene expression analysis and sample pooling in RNA-seq. *BMC Genomics*, 16(1):548, July 2015.
60. F. Rapaport, R. Khanin, Y. Liang, M. Pirun, A. Krek, P. Zumbo, C. E. Mason, N. D. Socci, and D. Betel. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.*, 14(9):R95, 2013.
61. M. Rho, H. Tang, and Y. Ye. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.*, 38(20):e191, Nov. 2010.
62. P. K. Robinson. Enzymes: principles and biotechnological applications. *Essays Biochem.*, 59: 1–41, 2015.

63. F. Rondel, R. Hosseini, B. Sahoo, S. Knyazev, I. Mandric, F. Stewart, I. I. Măndoiu, B. Pasaniuc, and A. Zelikovsky. Estimating enzyme participation in metabolic pathways for microbial communities from RNA-seq data. In *Bioinformatics Research and Applications*, pages 335–343. Springer International Publishing, 2020.
64. F. M. Rondel, R. Hosseini, B. Sahoo, S. Knyazev, I. Mandric, F. Stewart, I. I. Măndoiu, B. Pasaniuc, Y. Porozov, and A. Zelikovsky. Pipeline for analyzing activity of metabolic pathways in planktonic communities using metatranscriptomic data. *J. Comput. Biol.*, 28(8): 842–855, Aug. 2021.
65. I. Sharon, S. Bercovici, R. Y. Pinter, and T. Shlomi. Pathway-based functional analysis of metagenomes. *J. Comput. Biol.*, 18(3):495–505, Mar. 2011.
66. M. Shen, Q. Li, M. Ren, Y. Lin, J. Wang, L. Chen, T. Li, and J. Zhao. Trophic status is associated with community structure and metabolic potential of planktonic microbiota in plateau lakes. *Front. Microbiol.*, 10:2560, Nov. 2019.
67. J. Shi and M. G. Walker. Gene set enrichment analysis (GSEA) for interpreting gene expression profiles. *Curr. Bioinform.*, 2(2):133–137, 2007.
68. L. D. Shultz, M. A. Brehm, J. V. Garcia-Martinez, and D. L. Greiner. Humanized mice for immune system investigation: progress, promise and challenges. *Nat. Rev. Immunol.*, 12(11): 786–798, Nov. 2012.
69. A. Stupnikov, C. E. McInerney, K. I. Savage, S. A. McIntosh, F. Emmert-Streib, R. Kennedy, M. Salto-Tellez, K. M. Prise, and D. G. McArt. Robustness of differential gene expression analysis of RNA-seq. *Comput. Struct. Biotechnol. J.*, 19:3470–3481, May 2021.

70. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 102(43):15545–15550, Oct. 2005.
71. A. L. Tarca, S. Draghici, G. Bhatti, and R. Romero. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics*, 13:136, June 2012.
72. A. Warshel. Energetics of enzyme catalysis. *Proc. Natl. Acad. Sci. U. S. A.*, 75(11):5250–5254, Nov. 1978.
73. J. C. Wolters, J. Ciapaite, K. van Eunen, K. E. Niezen-Koning, A. Matton, R. J. Porte, P. Horvatovich, B. M. Bakker, R. Bischoff, and H. P. Permentier. Translational targeted proteomics profiling of mitochondrial energy metabolic pathways in mouse and human samples. *J. Proteome Res.*, 15(9):3204–3213, Sept. 2016.
74. G. Yaari, C. R. Bolen, J. Thakar, and S. H. Kleinstein. Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations. *Nucleic Acids Res.*, 41(18):e170, Oct. 2013.
75. Y. Ye and T. G. Doak. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput. Biol.*, 5(8):e1000465, Aug. 2009.
76. A. Zecchin, P. C. Stapor, J. Goveia, and P. Carmeliet. Metabolic pathway compartmentalization: an underappreciated opportunity? *Curr. Opin. Biotechnol.*, 34:73–81, Aug. 2015.