Genomics, Transcriptomics, and Statistical/Machine Learning to Enhance Understanding of Methanotrophic Microbial Metabolism in Isolates and Communities.

Janet Matsen

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Mary Lidstrom, Chair

David Beck, Chair

David Baker

François Baneyx

Program Authorized to Offer Degree:
Chemical Engineering

University of Washington

Abstract

Genomics, Transcriptomics, and Statistical/Machine Learning to Enhance Understanding of Methanotrophic Microbial Metabolism in Isolates and Communities.

Janet Matsen

Co-Chairs of the Supervisory Committee:
Professor Mary Lidstrom
Chemical Engineering, Microbiology

Research Assistant Professor David Beck
Chemical Engineering

Engineered microbes will play a key role in the transition from fossil fuel derived chemicals to sustainable chemicals. Successful metabolic reenginnering requires deep understanding of microbial physiology, bioinformatics, and data science. This thesis utilizes all three to study the metabolism of methane and methanol-utilizing microbes. Both pure cultures (Chapter 2) and complex natural (Chapter 3) communities are studied. The potential to leverage statistical and machine learning for large meta-omics datasets is also explored (Chapter 4). Overall, we are able to make strong conclusions when high-quality isolate genomes are available, however, these inferences are much more difficult in the case for complex microbial communities with unknown underlying genomic composition.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

AMI	Amazon machine image. A master image for the creation of virtual servers (known as EC2 instances) in the Amazon Web Services (AWS) environment.
ANI	Average nucleotide identity, a measure of similarity between organisms.
assembly	The process of, or result of inferring the original genome sequences that produced sequencing reads.
AWS	Amazon Web Services, the leading cloud computing platform.
AWS instance	A single computer rented from Amazon Web Services, with user-selected performance characteristics.
bin	A collection of contigs that approximates one organism, or a group of closely related strains.
binning	The process of clustering contigs into bins.
BWA	Burrows-Wheeler Aligner, a tool for mapping reads to reference DNA sequences.
CCA	Canonical correlation analysis, a statistical learning technique.
CheckM	A computational tool used to assess genome bin completeness and contamination.
contig	A contiguous DNA sequence. In this case, they are the longer DNA stretches identified from assembling Illumina reads.
C1	Single-carbon compounds such as methane, methanol, and formate.
coverage	The average number of times a base of a genome (or genome fragment) is sequenced.
cross-validation	A model validation technique that loops over subsets of the data, and leads to an assessment of how the results of a model will generalize to an independent dataset.

d	The dimensionality of the data. In this thesis, it is usually the number of genes.
EC2	A virtual computing environment within Amazon Web Services.
EFS	Amazon's Elastic File System, a scalable file storage service.
Elviz	A Joint Genome Analysis toolchain for metagenomics and metatranscriptomics.
FASTA	A biology sequence data format, which can contain DNA, RNA, or protein sequences for different units of information such as genes, contigs, or whole genomes. Individual FASTA files are referred to as `.fasta` files herein.
fastq	A `.fasta`-like file with associated sequence quality information. This is the most raw output from the sequencing platform handled in this thesis.
feature	A single measurable property of a phenomenon being observed, such as one gene's expression level.
Gaussian distribution	Normal distribution; a commonly used symmetric distribution.
GC	Strictly speaking, G and C are each DNA bases. When used in terms like "high GC", it indicates high fraction of G, C nucleotides over A, T nucleotides.
genomics	The collective characterization and quantification of pools of DNA molecules
genome bin	A collection of contigs used to approximate a single organism's genome. Often referred to as "bin".
hyperparameter	Parameters in a machine learning model that express higher-level properties of the model such as the model complexity or the learning rate.
i.i.d.	Independent and identically distributed. Each random variable has the same probability distribution as the others and all are mutually independent.
isolate	A single microbial strain that has been isolated and propagated in pure culture.

JGI	The Joint Genome Institute, a U.S. Department of Energy laboratory which provided sequencing services for the metagenomics work of this thesis.
kb	Kilobase. A unit of DNA (or RNA) length, corresponding to 1000 base pairs.
L1 regularization	A method to limit model complexity by penalizing the absolute value of model coefficients. Leads to sparse models (truly zero coefficients), rather than coefficients with small magnitudes (e.g. 0.00001).
mapping	Alignment of sequencing reads to target DNA such as genome or contig sequences.
machine learning	Models that have the ability to learn without being explicitly programmed.
MetaBAT	A binning tool.
metagenome	DNA recovered directly from a mixed microbial community, rather than an isolated strain.
metatranscriptome	RNA recovered directly from a mixed microbial community, rather than an isolated strain.
methanotroph	A microbe which can use methane as its sole carbon and energy source.
methylotroph	A microbe which can use reduced C1 compounds such as methanol as its sole carbon and energy source.
methane monooxygenase	(MMO) A methanotroph enzyme that can oxidize the C-H bond in preparation for assimilation or energy production.
multilocus mapped reads	(MMR) Reads that map more than one place equally well, causing uncertainty about their true origin.
non-methanotrophic methylotroph	While all methanotrophs are methylotrophs, not all methylotrophs can utilize methane. Such methylotrophs are often described as non-methanotrophic methylotrophs.
MyCC	A binning tool.
N	The number of samples available for machine learning, e.g. 88 for 88 metatranscriptomes.
OB3b	*Methylosinus trichosporium* OB3b, a RuMP cycle methanotroph, that is the subject of study in Chapter 2

omics	Refers to fields of study ending in *-omics*, such as genomics and transcriptomics, which quantify pools of biological molecules.
partial correlation	A measure of the strength and direction of a linear relationship between two continuous variables whilst controlling for the effect of one or more other continuous variables.
pcor	Abbreviation for the partial correlation matrix.
PhyloPhlAn	A computational tool that can assign taxonomy to metagenome bins.
Prokka	A computational tool for annotating genes on contigs.
read	a single short DNA sequence of a larger high-throughput sequencing dataset.
regularization	A technique to limit model complexity and prevent over-fitting to training data.
RPKM	Reads per kilobase per kilobase mapped.
shrinkage	Nearly sononomous with regularization; a method for limiting the number of parameters in a model.
sparse	Data or computational models with many zero values or coefficients.
taxa	A taxonomic category, such as a genus or species.
taxonomy	A scheme of classification for organisms.
tetranucleotide frequency	The frequency of 4-nucleotide sequences in DNA, which is conserved for an organism.
TPM	Transcripts Per Kilobase Million, an alternative to RPKM. Reads per kilobase is calculated first, then these totals are normalized by dividing by the sum and 1,000,000.
training set	Data used to train a machine learning model.
transcriptomics	The collective characterization and quantification of pools of RNA molecules (transcripts).
16S	A ribosomal protein, whose DNA sequence is commonly used for evolutionary inference.

ACKNOWLEDGMENTS

I would like to express immense gratitude to Mary Lidstrom, who taught me so much about biology and metabolic engineering, then supported my transition to computational biology and data science. Mary was an incredible mentor for every phase of my PhD work, providing brilliant ideas on unfamiliar new data both before and after my pivot to computational science. Every interaction was positive, every meeting productive, and she was incredibly accessible while simultaneously serving as the Vice Provost of Research for the University of Washington. It was always clear that my and other lab members' best interests for scientific and professional development were her top priority. Mary will forever remain one of my top role models for science and leadership.

I also would like to thank David Beck for his computational mentorship throughout my PhD, and particularly for the last two years when we worked so closely together. His incredible breadth and depth of knowledge, and his eagerness to share it with students are unparalleled. It was Dave's enthusiasm for data science that led me to the Advanced Data Science Option, which has already had positive impacts on my career.

I am also thankful for the dozens of people who lent support and collaborated on ambitious wet lab experiments, most of which are not described in this thesis. I thank the Formolase project team, particularly Amanda Smith, David Baker, Adam Wargacki, and Justin Siegel. I thank the numerous Lidstrom Lab members for support around microbiology, molecular biology, and omics techniques. I thank Marina Kalyuzhnaya for initial mentorship when I joined the lab, and Ludmila Chistoserdova for regular advice concerning methylotrophy.

I thank my family for guiding me toward engineering at a young age, and the support through these 25 years of education.

I also thank the community around the Data Science Studio and Advanced Data Science Option at UW.

Finally, I thank my committee members for support throughout a particularly diverse PhD experience.

DEDICATION

I dedicate this to my unbelievably cherished husband, Erick. You inspired me to embark on this journey, helped keep the productivity tempo high, and contributed so greatly to my happiness. I will fondly remember programming side by side at 5AM most mornings, regular discussions about computation and data science, and all the nerdy jokes we shared forever.

Chapter 1

INTRODUCTION

I came to the Lidstrom Lab eager to learn the art of engineering microbes, with understanding that the future of sustainable chemical production will feature such processes. During my PhD, I had the opportunity to do significant amounts of challenging wet-lab biology, much of which is not included in the scope of this dissertation. While working in strain design, engineering, enzyme design, enzyme evolution, and fermentation, it became clear to me that our ability to achieve the potential of our field relies on robotics, high-throughput screening, and advanced analytics. With Mary's incredible support, I redirected my focus to computational biology and data science. Along the way, I completed data science coursework and was granted the first Advanced Data Science Option from the University of Washington. Again with Mary's generous support, I was able to do two computational internships at two leading biotech startups in our field. Moving forward as a biology-focused data scientist, I am so grateful for the understanding of microbial physiology and biological experiments that the Lidstrom Lab provided me.

This thesis focuses on genomic and transcriptomic studies of pure cultures and mixed microbial populations, as both an experimentalist (Chapter 2), and fully computational data scientist (Chapters 3, 4). For information about the wet-lab work regarding the Formolase Pathway [1], development of SIP3-4 as a model organism, and development of a high-throughput formaldehyde-production enzyme screen, please see my general exam (https://github.com/JanetMatsen/thesis/blob/master/documents/2015_Matsen_General_Exam.pdf).

1.1 Methylotrophy and Methanotrophy

The focus of the Lidstrom lab is understanding and manipulating methylotrophs, which are microorganisms that grow using reduced single-carbon compounds such as methane, methanol, methylamine and formate as their sole carbon source and energy source [2, 3]. Methanotrophs are methylotrophs that have extra metabolic modules enabling them to enzymatically convert methane to methanol [4]. The energetics of activating these highly reduced single carbon compounds and the unique challenge of assimilating the activated products requires special metabolic pathways not present in other microbes. Furthermore, understanding of both methylotrophs and methanotrophs is increasingly important given their role in mitigating these greenhouse gases, and their biotechnological potential for converting inexpensive single-carbon compounds to economically valuable multi-carbon compounds.

1.2 Tools for Metabolism Studies

The Lidstrom Lab uses many approaches to observe and alter microbial metabolic networks, allowing deeper understanding of the complex network and the potential to perturb it for human good. Metabolism can be observed at many levels; at each level there are distinct sets of techniques. The most commonly quantified entities are DNA pools, RNA pools, metabolite concentrations, and carbon flux. DNA is used to identify metabolic potential of a single organism or population, in terms of what genes are available. RNA is used to identify which of these genes microbes actually express, and under which conditions. Metabolite concentrations and flux measurements indicate the effects of gene expression on the chemical environment of the cell.

Each type of study corresponds to a different scientific discipline of ”omics”, the collective characterization and quantification of pools of biological molecules. The study of genomes is called ”genomics”, while the study of transcripts (RNA) is called ”transcriptomics”. While the Lidstrom Lab combines all of these observational tools

and a variety of pathway-altering tools, this thesis focuses on analysis of genomic and transcriptomic datasets obtained via high-throughput sequencing.

Genomics and transcriptomics can further be classified by whether they address a single organism's nucleic acids or those from a collection of organisms. When genomics and transcriptomics are used to study populations rather than single organism (pure) cultures, the prefix "meta" is added, yielding "metagenomics" and "metatranscriptomics". Metagenomics can imply different types of sequencing, which can be coarsely divided into deeper sequencing of only 16S ribosomal DNA to profile taxa abundance, or more broad sequencing of the entire DNA pool in a community. This dissertation will focus entirely on the latter, termed "shotgun" metagenomics to clarify that specific types of DNA sequences such as 16S or other marker genes are not targeted for sequencing. This un-targeted sampling allows discovery of the genes available to the population of organisms, and the potential to identify genomes of un-cultured organisms.

1.3 Machine Learning

'Omics studies usually stop after tabulating and describing the measurements made. For metagenomics and transcriptomics, this corresponds to descriptions of the genetic material present and the expression level of the predicted genes. There can, however, be much richer descriptions of datasets if explored with the appropriate computational techniques. Chapter 4 takes that extra step, and applies machine learning techniques to extract experimentally testable hypotheses from complex meta-'omics data.

1.4 Road Map for Thesis

The broad theme of this dissertation is use of high-throughput sequencing to understand microbial metabolism. It begins with the simplest and most traditional type of study, using a wild-type lab isolate, and transitions to the increased biological and analytical complexity of natural sediment communities. For both systems, it aims to identify

which metabolic pathways organism(s) use, given species-level and community-level metabolic redundancy. For the case of complex communities, additional computational and machine learning techniques are used to extract higher-order function from the communities.

Chapter 2 describes how carbon flows through the metabolic network of a model methanotroph. This thesis focuses on the the transcriptomics half [5] of a two-part study including metabolomics [6] of the same culture. Pure cultures allow high-confidence elucidation of metabolic pathways, but may not represent the often un-cultivatable wild relatives.

Interest in how methane is oxidized in nature led to Chapter 3, which explores metabolism in species-rich methanotroph-enriched natural communities. It aims to clarify how diverse populations of methanotrophs and auxiliary microbes oxidize methane in nature: Which microbes are important in natural communities? Are the lab isolates (e.g. Chapter 2) good representatives of the wild variants? Which metabolic pathways do individual microbes use? What metabolic interactions occur between species in nature? Can the presence of one organism be linked to a shift in metabolic strategy of another?

This work is part of a rapidly developing field of community 'omics. The newness of the field and corresponding lack of standards, combined with a plethora of tools to navigate requires thoughtful method selection, custom definitions of efficacy, and consideration of trade-offs between seemingly redundant tools. For this thesis, special attention is paid to monitoring how well the results reflect the ground-truth raw sequences representing each sample.

Chapter 3 approximates the question of "who is there", and highlights the importance of oxygen availability, a key environmental variable, in determining community composition, and community gene expression.

Chapter 4 builds on the abundance and gene-expression results of Chapter 3 by applying machine learning to identify patterns not evident to the human eye. It

highlights portions of the machine learning literature that can be leveraged for analysis of communities, given that the data in this field has a much larger number of features and smaller number of samples than is common in typical machine learning applications. These methods can provide testable hypotheses about between-species interactions to enhance understanding of community metabolism.

Chapter 2

GLOBAL MOLECULAR ANALYSES OF METHANE METABOLISM IN METHANOTROPHIC ALPHAPROTEOBACTERIUM, *METHYLOSINUS TRICHOSPORIUM* OB3B. PART I: TRANSCRIPTOMIC STUDY

2.1 Abstract

Methane utilizing bacteria (methanotrophs) are important in both environmental and biotechnological applications, due to their ability to convert methane to multicarbon compounds. However, systems-level studies of methane metabolism have not been carried out in methanotrophs. In this work we have integrated genomic and transcriptomic information to provide an overview of central metabolic pathways for methane utilization in *Methylosinus trichosporium* OB3b, a model alphaproteobacterial methanotroph. Particulate methane monooxygenase, PQQ-dependent methanol dehydrogenase, the H_4MPT-pathway, and NAD-dependent formate dehydrogenase are involved in methane oxidation to CO_2. All genes essential for operation of the serine cycle, the ethylmalonyl-CoA (EMC) pathway, and the citric acid (TCA) cycle were expressed. PEP-pyruvate-oxaloacetate interconversions may have a function in regulation and balancing carbon between the serine cycle and the EMC pathway. A set of transaminases may contribute to carbon partitioning between the pathways. Metabolic pathways for acquisition and/or assimilation of nitrogen and iron are discussed.

Note: this chapter is published as an article [5], alongside a complementary metabolomics article [6]. See the original paper (http://journal.frontiersin.org/article/10.3389/fmicb.2013.00040/full) for supplementary material.

2.2 Introduction

Aerobic methanotrophic bacteria (methanotrophs) are a highly specialized group of microbes utilizing methane as a sole source of carbon and energy [7, 8]. As the recognition of methane's impact on global climate change increases, a multitude of research activities have been directed toward understanding the natural mechanisms for reducing methane emissions, including consumption by methanotrophs. The number of described microbial species capable of methane oxidation has recently expanded dramatically. A number of novel methanotrophic phyla have been isolated and described in the past few years, including new members of the Alpha- and Gammaproteobacteria, and Verrucomicrobia [9, 3, 8]. Several genomes of methanotrophic bacteria have been sequenced opening new dimensions in characterization of methane metabolism [10, 11, 12, 13, 14, 15, 16]. Initial genome-based reconstructions of methane metabolism in methanotrophic proteobacteria and Verrucomicrobia have been performed [10, 17, 12, 18].

Methylosinus trichosporium OB3b, an obligate alphaproteobacterial methanotroph, has served as a model system for years (first described in [19]). Research on both fundamental and biotechnological aspects of methanotrophy in *M. trichosporium* OB3b has been carried out with applications involving cometabolism of contaminants [20, 21, 22], epoxidation of propene [23], and synthesis of polyhydroxybutyrate (PHB) [24, 25]. *M. trichosporium* OB3b possesses two systems for methane oxidation, a particulate methane monooxygenase (pMMO), expressed under high biomass/copper ratios, and a soluble methane monooxygenase (sMMO) which is expressed at low copper conditions [26, 27]. It has been shown that the strain is capable of fixing nitrogen [28, 29]. Although significant progress has been made in the understanding of primary methane and methanol oxidation pathways in this model bacterium, little work has been carried out on carbon assimilation by *M. trichosporium* OB3b. The reconstruction of the core metabolic pathways for alphaproteobacterial methanotrophs has been primarily based on a restricted set of enzymatic studies and extrapolations relying on similarity to non-

methane utilizing methylotrophs [30, 31]. A draft genome of *M. trichosporium* OB3b has recently been generated [14]. This genetic blueprint provides an essential background for revisiting the established model of methanotrophy in Alphaproteobacteria using modern system-level approaches. For this research, we integrated heterogeneous multi-scale genomic, transcriptomic, and metabolomic data to redefine the metabolic framework of C_1-utilization in *M. trichosporium* OB3b grown in batch culture under copper, oxygen, and iron sufficiency on methane and nitrate as the sources of carbon and nitrogen, respectively. In this part of our work we present transcriptomic-based analysis of the methanotrophic metabolic network. Metabolomic and 13C-labeling studies are presented in a follow-up paper [6].

2.3 Results and Discussion

2.3.1 Gene Expression Studies

Gene expression studies were carried out with *M. trichosporium* OB3b cultures grown on methane at N (10 mM), Cu (9 μM), and Fe (9 μM) sufficiency conditions. The maximum specific growth rate of *M. trichosporium* OB3b in shake flasks during the exponential growth phase was $\mu = 0.038 \pm 0.004$ h^{-1}. The methane consumption rate during the period of maximum growth rate was 8.95 mmol of CH_4 h^{-1} L $culture^{-1}$ ($OD_{600} = 1$).

All experiments were performed with at least two biological replicates. RNA samples were prepared as described in the Section "Materials and Methods." Illumina sequencing for two biological replicates (BR1 and BR2) returned 28 and 29 million 36-bp reads. The Burrows-Wheeler Aligner (BWA, [32]) aligned 98% of the reads to the *M. trichosporium* OB3b genome annotated by MaGe [33] using the default parameters for small genomes. Reads per kilobase of coding sequence per million (reads) mapped (RPKM) [34] was calculated to compare gene expression within and across replicates, and no further normalization (other than RPKM) was applied. The samples were in good agreement with each other, with per gene coding sequence RPKM correlations of 0.959 and 0.989 for

the Pearson and Spearman correlations, respectively. In total, 4,762 of 4,812 ORFs (CDS, tRNA, and rRNA predicted from the draft genome) were detected. Based on relative expression, genes (omitting rRNAs) could be grouped into six major expression categories (Table 2.1): *very high* (RPKM $\geq$ 15,000), *high* (RPKM $\geq$ 1,500), *moderate* (1,500 $>$ RPKM $\geq$ 500), *modest* (500 $>$ RPKM $\geq$ 250), *low* (250 $>$ RPKM $\geq$ 150) *very low* (150 $>$ RPKM $\geq$ 15), and *not expressed* (RPKM $<$ 15). The majority of genes fell into *low/very low* expression categories (74%). About 14% of genes displayed *moderate/modest* expression and only a small fraction of the genome showed *very high/high* expression (2.7%).

Table 2.1: Classification of gene expression level based on replicate averaged RPKMs.

Description of expression level	**RPKM range**	**% of ORFs**	**Number of ORFs**
Very high	>15,000	0.23	11
High	1,500–15,000	2.49	120
Moderate	500–1,500	5.30	255
Modest	250–500	8.61	414
Low	50–250	40.41	1,944
Very low	15–50	23.70	1,140
Not expressed	<15	19.27	927

In order to determine whether the draft genome of the strain is missing some functional genes, we performed *de novo* assembly of the transcriptome. Using this approach, a total of 173 genes that are not present in the genome sequence, but have homologs in the non-redundant database were detected. Among those are key subunits of succinate dehydrogenase (*sdhABCD*), 2-oxoglutarate dehydrogenase (E2), and nitric oxide reductase (*norB*) (Table A.1). The *de novo* transcriptome assembly provides additional information for highly expressed genes and it was used for verification of some metabolic functions that were predicted by enzymatic studies but were not detected in the draft genome assembly (see below).

In addition, the reads obtained from RNA-seq were aligned to the reference genome in order to identify transcription boundaries and transcription start sites for the most highly expressed genes, including the *pmoCAB* operons, *mxaFJGI* operon, *fae1*, *pqqA*, and key genes of the serine cycle (Table A.2). Gene expression data were used to reconstruct central metabolic pathways in *M. trichosporium* OB3b (Table 2.2; Figure 2.1; Table A.2). Core functions are described below.

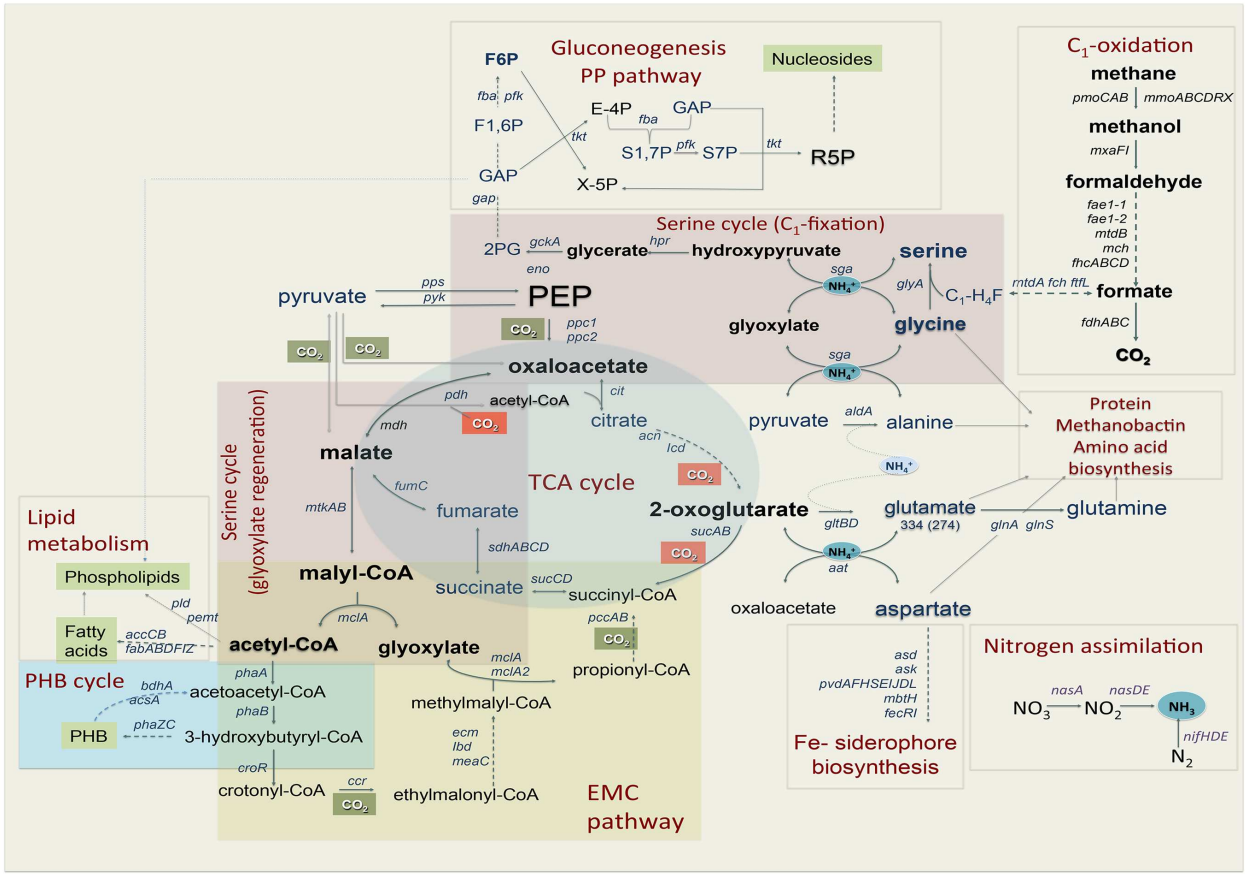


Figure 2.1: Central metabolism of *Methylosinus trichosporium* OB3b grown on methane as sole source of energy and carbon as deduced from the genome sequences and transcriptomic studies. Font size of the gene name indicates the expression level.

Table 2.2: Gene expression profile in methane-grown cells of *M. trichosporium* OB3b

Gene ID	Predicted function	Gene	Replicate 1	Replicate 2
METHANE AND METHANOL OXIDATION				
METTOv1_1270003	Particulate methane monooxygenase subunit C	*pmoC*	123026	127241
METTOv1_1270002	Particulate methane monooxygenase subunit A	*pmoA*	37102	31813
METTOv1_1270001	Particulate methane monooxygenase subunit B	*pmoB*	27371	22917
METTOv1_310040	Particulate methane monooxygenase subunit C2	*pmoC2*	532	492
METTOv1_50081	Soluble methane monooxygenase alpha subunit	*mmoX*	9	8
METTOv1_50082	Soluble methane monooxygenase beta subunit	*mmoY*	13	9
METTOv1_50084	Soluble methane monooxygenase gamma subunit	*mmoZ*	20	19
METTOv1_240014	PQQ-dependent methanol dehydrogenase	*mxaF*	15313	13760
METTOv1_240011	PQQ-dependent methanol dehydrogenase	*mxaI*	24552	28474
METTOv1_240012	Cytochrome *c* class I	*mxaG*	5712	6117
METTOv1_240013	Extracellular solute-binding protein family 3	*mxaJ*	1942	1838
METTOv1_240001	Putative methanol utilization control sensor protein	*mxaY*	36	41
METTOv1_240002	Putative two-component response regulator	*mxaB*	303	317
METTOv1_240003	MxaH protein, involved in methanol oxidation	*mxaH*	399	391
METTOv1_240004	MxaD protein, involved in methanol oxidation	*mxaD*	1137	1077
METTOv1_240005	von Willebrand factor type A, involved in methanol oxidation	*mxaL*	191	201
METTOv1_240006	Protein of unknown function, involved in methanol oxidation	*mxaK*	124	132
METTOv1_240007	von Willebrand factor type A, involved in methanol oxidation	*mxaC*	144	141
METTOv1_240008	MxaA protein, involved in methanol oxidation	*mxaA*	137	127
METTOv1_240009	MxaS protein, involved in methanol oxidation	*mxaS*	202	167
METTOv1_240010	ATPase, involved in methanol oxidation	*mxaR*	563	538
METTOv1_110056	Coenzyme PQQ biosynthesis protein A	*pqqA*	11857	13927
METTOv1_160001	Coenzyme PQQ biosynthesis protein E	*pqqE*	166	161
METTOv1_160002	Coenzyme PQQ biosynthesis protein PqqC/D	*pqqC/D*	372	344
METTOv1_160003	Coenzyme PQQ biosynthesis protein B	*pqqB*	306	313
METTOv1_20046	Coenzyme PQQ biosynthesis protein F	*pqqF*	183	185
METTOv1_20047	Coenzyme PQQ biosynthesis protein G	*pqqG*	157	142
METTOv1_610028	Aldehyde dehydrogenase	*aldh*	37	37
METTOv1_290006	Aldehyde oxidase	*aor*	45	38
METTOv1_100046	Aldehyde dehydrogenase	*aldh-F7*	7	9
FORMALDEHYDE OXIDATION				
METTOv1_40010	Methenyltetrahydromethanopterin cyclohydrolase	*mch*	393	312
METTOv1_40011	Tetrahydromethanopterin-linked C1 transfer pathway protein. Orf5	*orf5*	128	111
METTOv1_40012	Tetrahydromethanopterin-linked C1 transfer pathway protein, Orf7	*orf7*	73	72
METTOv1_40013	Formaldehyde activating enzyme	*fae1*	24353	24787
METTOv1_40014	Formaldehyde activating enzyme	*fae1-2*	4024	3676
METTOv1_840013	Formaldehyde activating enzyme homolog	*fae2*	535	581
METTOv1_40015	Tetrahydromethanopterin-linked C1 transfer pathway protein	*orf17*	38	45
METTOv1_110058	Tetrahydromethanopterin formyltransferase, subunit C	*fhcC*	535	453
METTOv1_110059	Tetrahydromethanopterin formyltransferase, subunit D	*fhcD*	496	470
METTOv1_110060	Tetrahydromethanopterin formyltransferase, subunit A	*fhcA*	591	546
METTOv1_110061	Tetrahydromethanopterin formyltransferase, subunit B	*fhcB*	620	570
METTOv1_560001	Tetrahydromethanopterin -linked C1 transfer pathway protein	*orf9*	172	167
METTOv1_560002	Methylenetetrahydrofolate dehydrogenase (NAD)	*mtdB*	688	607
METTOv1_440045	Ribofuranosylaminobenzene 5^0-phosphate synthase	*mptG*	94	80
FORMATE OXIDATION				
METTOv1_630016	Transcriptional regulator, LysR family	*fdsR*	52	39
METTOv1_630017	NAD-linked formate dehydrogenase, subunit G	*fdsG*	672	608
METTOv1_630018	NAD-linked formate dehydrogenase, subunit B	*fdsB*	585	531
METTOv1_630019	NAD-linked formate dehydrogenase, subunit A	*fdsA*	593	554
METTOv1_370001	Formate dehydrogenase family accessory protein	*fdsC*	210	199

(Continued)

Table 2 | Continued

Gene ID	Predicted function	Gene	Replicate 1	Replicate 2
METTOv1_370002	NAD-linked formate dehydrogenase, subunit D	*fdsD*	368	312
METTOv1_220028	NAD-linked formate dehydrogenase, subunit A	*fdhA2*	9	7
C1-ASSIMILATION:SERINE CYCLE				
METTOv1_130002	Phosphoenolpyruvate carboxylase	*ppc2*	104	89
METTOv1_400011	Glycerate kinase	*gckA*	229	211
METTOv1_400012	Conserved protein of unknown function	*orf1*	626	708
METTOv1_400013	Malyl-CoA lyase/beta-methylmalyl-CoA lyase	*mclA*	1713	1615
METTOv1_400014	Phosphoenolpyruvate carboxylase	*ppc1*	141	139
METTOv1_400015	Malate thiokinase, small subunit	*mtkB*	516	485
METTOv1_400016	Malate thiokinase, large subunit	*mtkA*	534	455
METTOv1_400017	Methenyltetrahydrofolate cyclohydrolase	*fch*	355	281
METTOv1_400018	NADP-dependent methylenetetrahydrofolate dehydrogenase	*mtdA*	281	243
METTOv1_400019	2-Hydroxyacid dehydrogenase NAD-binding	*hprA*	375	348
METTOv1_400020	Serine-glyoxylate transaminase	*sga*	1840	1969
METTOv1_400021	Formate-tetrahydrofolate ligase	*ftfL*	448	412
METTOv1_670019	Serine hydroxymethyltransferase	*glyA*	1342	1197
METTOv1_20135	Enolase	*eno*	432	408
C1-ASSIMILATION:EMP PATHWAY AND PHB CYCLE				
METTOv1_100079	Acetyl-CoA acetyltransferase	*phaA*	597	561
METTOv1_100080	Acetoacetyl-CoA reductase	*phaB*	1160	1060
METTOv1_50006	Crotonase	*croR*	235	275
METTOv1_110068	Crotonyl-CoA reductase	*ccr*	577	523
METTOv1_60013	Ethylmalonyl-CoA mutase	*ecm*	187	162
METTOv1_510010	Methylsuccinyl-CoA dehydrogenase	*ibd*	309	295
METTOv1_110043	Mesaconyl-CoA hydratase	*meaC*	341	317
METTOv1_30129	Methylmalonyl-CoA epimerase	*epm*	428	394
METTOv1_220010	Malyl-CoA lyase/beta-Methylmalyl-CoA lyase	*mclA2*	137	135
METTOv1_200020	Acetyl/propionyl-CoA carboxylase	*ppcA*	353	310
METTOv1_220035	Propionyl-CoA carboxylase	*ppcB*	472	455
METTOv1_50067	Methylmalonyl-CoA mutase, large subunit	*mcmA*	201	188
METTOv1_10062	Methylmalonyl-CoA mutase small subunit B	*mcmB*	144	144
METTOv1_270063	3-Hydroxybutyrate dehydrogenase	*bdhA*	235	232
METTOv1_130047	Poly-beta-hydroxybutyrate polymerase	*phaC*	30	30
METTOv1_200042	Acetoacetate decarboxylase	*aad*	123	114
METTOv1_200022	Acetoacetyl-coenzyme A synthetase	*aas*	116	114
METTOv1_630008	Polyhydroxyalkanoate depolymerase	*phaZ*	237	263
C1-ASSIMILATION:TCA CYCLE				
METTOv1_360040	Malate dehydrogenase	*mdh*	539	473
METTOv1_360041	Succinyl-CoA synthetase, beta subunit	*sucC*	660	631
METTOv1_510003	Succinyl-CoA synthetase, alpha subunit	*sucD*	1198	1135
METTOv1_510002	2-Oxoglutarate dehydrogenase E1	*sucA*	236	237
METTOv1_370050	2-Oxoglutarate dehydrogenase E2	*sucB*	191	181
METTOv1_80046	Succinate:ubiquinone oxidoreductase	*sdhB*	327	348
METTOv1_80046	Succinate:ubiquinone oxidoreductase	*sdhA*	311	299
METTOv1_80051	Succinate:ubiquinone oxidoreductase, cytochrome b556 subunit	*sdhC*	318	329
METTOv1_40061	Fumarate hydratase	*fum*	196	185
METTOv1_1080004	2-Oxoacid ferredoxin oxidoreductase	*ofr*	79	77
INTERMEDIARY METABOLISM AND ANAPLEROTIC CO_2-FIXATION				
METTOv1_70038	Phosphoenolpyruvate synthase	*pps*	28	26
METTOv1_120036	Pyruvate carboxylase	*pcx*	145	140
METTOv1_830002	Acetyl-coenzyme A carboxylase subunit beta	*accD*	246	228
METTOv1_380021	Acetyl-CoA carboxylase subunit alpha	*accA*	211	204

(Continued)

Table 2 | Continued

Gene ID	Predicted function	Gene	Replicate 1	Replicate 2
METTOv1_130018	Acetyl-CoA carboxylase, biotin carboxyl carrier protein	*accB*	307	276
METTOv1_150014	Pyruvate kinase	*pyk1*	245	224
METTOv1_340039	Pyruvate dehydrogenase (acetyl-transferring) E1	*pdhA*	181	175
METTOv1_340041	Pyruvate dehydrogenase subunit beta	*pdhB*	160	158
METTOv1_340042	Pyruvate dehydrogenase	*pdhC*	101	106
METTOv1_350050	Pyruvate phosphate dikinase	*pdk*	50	52
METTOv1_80025	Malic enzyme	*mae*	107	106
METTOv1_680013	Phosphoglycerate mutase	*gpmA*	136	135
METTOv1_100061	Phosphoglycerate mutase (modular protein)	*pgm*	101	99
METTOv1_10180	Phosphoglycerate mutase (modular protein)	*pgm*	72	65
METTOv1_280049	Phosphoglycerate kinase	*pgk*	199	208
METTOv1_280047	Glyceraldehyde-3-phosphate dehydrogenase	*gpd*	391	407
METTOv1_620016	Ribokinase	*rik*	140	144
METTOv1_620017	Phosphoribulokinase	*prk*	173	164
METTOv1_620018	Transketolase	*tkl*	147	123
METTOv1_620019	Fructose-bisphosphate aldolase, class II	*fba*	427	461
METTOv1_220030	6-Phosphofructokinase	*pfk*	243	239
METTOv1_200031	Fructose 1,6-bisphosphatase II	*glp*	68	56
METTOv1_550029	Glucose-6-phosphate isomerase	*pgi*	88	84
METTOv1_620022	Ribulose-phosphate 3-epimerase	*rpe*	51	53
NITROGEN, Cu, Fe METABOLISM				
METTOv1_310019	Nitrate transporter component	*nrtA*	151	154
METTOv1_310020	Nitrite reductase (NAD(P)H), large subunit	*nasB*	1762	1562
METTOv1_310021	Nitrite reductase (NAD(P)H), small subunit	*nasD*	726	610
METTOv1_310022	Nitrate reductase, large subunit	*nasA*	646	597
METTOv1_130049	Ammonium transporter	*amtB*	2029	1766
METTOv1_300058	Glutamate synthase large subunit (NADPH/GOGAT)	*gltB*	199	183
METTOv1_300033	Glutamate synthase small subunit (NADPH/GOGAT)	*gltD*	431	401
METTOv1_190023	Glutamate dehydrogenase	*gdh*	2	2
METTOv1_200046	Glutamate-ammonia ligase	*glnS*	1413	1375
METTOv1_200047	Nitrogen regulatory protein P-II	*glnK*	2400	2277
METTOv1_200048	Glutamine synthetase, type I	*glnA*	2452	2321
METTOv1_280018	Alanine dehydrogenase	*aldA*	29	36
METTOv1_80043	Phosphoserine aminotransferase	*serC*	415	364
METTOv1_560023	Cytochrome c^{Q}alpha	*cycA*	336	370
METTOv1_230076	Putative oxygenase		6	8
METTOv1_230077	Hydroxylamine reductase	*hcp*	21	16
METTOv1_230078	Putative transcriptional regulator	*nsrR*	38	35
METTOv1_730005	Putative FecR iron sensor protein	*fecR*	48	58
METTOv1_730006	Putative TonB-dependent receptor protein	*tonB*	382	383
METTOv1_CDS4222756D	Methanobactin precursor	*Mb*	1439	2177
METTOv1_730007	Putative lyase		306	384
METTOv1_730008	Conserved protein of unknown function	*hp*	170	175
METTOv1_730009	Conserved protein of unknown function	*hp*	116	143
METTOv1_660011	l-Ornithine 5-monooxygenase	*pvdA1*	1410	1450
METTOv1_760004	Putative hydroxy-l-ornithine formylase	*pvdF*	2255	2676
METTOv1_760006	l-Ornithine 5-monooxygenase	*pvdA2*	979	1116
METTOv1_760007	Diaminobutyrate-2-oxoglutarate aminotransferase	*pvdH*	720	770
METTOv1_760008	Sigma-24 (FecI-like)	*pvdS*	1260	1522
METTOv1_760009	Putative pyoverdine ABC export system, permease	*pvdE*	532	500
METTOv1_760010	TonB-dependent siderophore receptor	*fpvA*	2197	2174
METTOv1_760011	FecR-like protein	*fecR*	330	321

(Continued)

Table 2 | Continued

Gene ID	Predicted function	Gene	Replicate 1	Replicate 2
METTOv1_760012	FecI-family sigma factor	*fecI*	922	951
METTOv1_870003	Ferribactin synthase	*pvdL*	433	441
METTOv1_870004	Pyoverdine biosynthesis regulatory protein-TauD/TfdA family protein		932	1039
METTOv1_870005	Pyoverdine synthetase, thioesterase component	*pvdG*	1542	1473
METTOv1_870006	Integral components of bacterial non-ribosomal peptide synthetases	*MbtH*	4176	5481
METTOv1_1220001	Putative pyoverdine sidechain peptide synthetase IV, d-Asp-l-Ser component	*pvdI/J*	480	564
METTOv1_1220002	Putative non-ribosomal peptide synthase	*pvdJ/D*	379	396

Values represent reads per kilobase of coding sequence per million (reads) mapped (RPKM).

2.3.2 C_1-Oxidation: Methane-To-Methanol

It has been previously demonstrated that *M. trichosporium* OB3b possesses two types of methane oxidation enzymes: pMMO and sMMO. The expression of the enzymes is determined by copper availability; sMMO is dominant in copper-limited environments while pMMO dominates under copper sufficiency [26, 27]. Structures of both enzymes are available [35, 36]. In this study, *M. trichosporium* OB3b was grown at a copper concentration that has been shown to be sufficient to suppress the expression of sMMO [37, 38, 39, 40, 41]. Indeed, virtually no expression of the sMMO gene cluster (*mmoXYBZC*) was observed. In contrast, the *pmoCAB* genes were the most highly expressed in the transcriptome, representing about 14% of all reads mapped to the coding regions (Table 2.2). It has previously been shown that PMMO in *M. trichosporium* OB3b is encoded by two copies of the *pmoCAB* operon that appear to be identical [42]. The current genome assembly failed to resolve these closely related duplicated regions. The *pmoCAB* genes were found within one relatively short contig, which includes 320 bp upstream from *pmoC*, and about 66 bp downstream from *pmoB*. It is possible that in the genome assembly, the *pmo* contig represents only those parts of the duplicated regions that are highly similar. Thus, it was not possible to determine relative expression of the two operons with the transcriptomic data.

Previous attempts to identify transcriptional starts of the *pmoCAB* operons in *M. trichosporium* OB3b using a conventional primer extension approach were not successful

[42]. The RNA-seq data were used for identification of transcriptional starts for the *pmoCAB* operons. Because the published METTOv1 genome did not contain a complete *pmoCAB* cluster, a separate alignment run was performed using a previously published sequence as the scaffold [43]. For this sequence, two possible transcriptional start sites were identified. It is not known whether these reflect the same start sites of both operons, different start sites for each, or expression of only one operon with two start sites. The position -274nt (A) from the translational start of the *pmoC* gene was predicted as the most prominent start of transcription of the operon (Figure A.1). Putative σ^{70}-like -10 and -35 regions could be identified upstream of the predicted start (Table A.3). The structure of the putative promoter region from *M. trichosporium* OB3b shows significant similarity to a *pmoCAB* promoter region previously identified in *Methylocystis* sp. M [42]. Another potential transcriptional start is at position -324 from the translational start of the *pmoC* gene. It should be noted that the region between the two predicted start sites was also covered with relatively high count (region between -324 and -274nt with respect to the translational start of *pmoC*). No putative promoter sequences were found upstream of position 324.

The genome predicts an additional copy of the *pmoC* gene by itself (*pmoC2*, *METTOv1_310040*), which can be distinguished from the other *pmoC* genes in the transcriptomics data due to sequence divergence. It has previously been demonstrated that additional copies of *pmoC* are essential for methanotrophic growth in other strains [44, 45]. It has also been shown that the homologous *amoC* (additional lone copy of *amoC* in ammonia-oxidizing bacterium *Nitrosomonas europaea*) plays role in cell recovery from ammonium starvation [46]. However the functional role of PmoC is not known. The relative expression of *pmoC2* was approximately 450-fold less than the expression of the two *pmoC* genes from the *pmo*-operons (Table 2.2). Low relative expression of the *pmoC* homolog may suggest a role in regulation or sensing rather than catalytic activity.

2.3.3 C_1-Oxidation: Methanol-To-Formaldehyde

The product of methane oxidation (i.e., methanol) is converted to formaldehyde by a PQQ-dependent methanol dehydrogenase (MDH) [2, 47, 48, 49]. The enzyme has been previously purified from *M. trichosporium* OB3b and well characterized [48]. MDH is a hetero-tetrameric enzyme encoded by *mxaF* and *mxaI*. The activity of the enzyme *in vivo* requires cytochrome c_L (*mxaG*) and a number of chaperones, regulators, and enzymes, including genes required for Ca^{2+} insertion [49, 47]. Most of the genes essential for this methanol conversion step in *M. trichosporium* OB3b are organized in one large operon in an order similar to that found in other methylotrophs (Figure A.2A). The first four genes of the operon (*mxaFJGI*), encoding the two subunits of the MDH, the associated cytochrome, and a gene of unknown function (*mxaJ*) were detected at relatively high RPKM counts. The relative expression of genes downstream from *mxaI*, including those for chaperones, regulators, and Ca^{2+} insertion functions drops by 10- to 50-fold (Table 2.2). The overall mapping pattern of the *mxa*-cluster is as follows: *mxaFJGI* (highly expressed), *mxaD* (moderate expression), *mxaRSACKL*, *mxaB* (low expression), and *mxaY* (very low expression). It remains to be elucidated if the *mxaRSACKL* transcripts arise from the same start as *mxaF* and are attenuated by some transcriptional or post-transcriptional mechanism, or whether separate, lower expression promoter(s) is/are present. Orientations and/or the expression patterns of *mxaD*, *mxaB*, and *mxaY*, suggest that they are not part of the major *mxaF*-operon (Figure A.2A) and most likely have independent regulatory/promoter regions. According to RNA-seq mapping data, a putative transcriptional start of the *mxaFJGI* operon is predicted at position -164 from the predicted translational start (Figure A.3). Just upstream from the predicted transcriptional start, putative σ^{70}-like -10 and -35 sequences were identified (Table A.3).

The genome of *M. trichosporium* OB3b contains the following three homologs of the large subunit of the MDH: *xoxF1*, *xoxF2*, and *xoxF3*. Relative expression of all *xoxF*-homologs is very low. The most highly expressed *xox*-homolog (*xoxF1*) showed

only about 2% of the *mxaF* expression. The function of the *xox*-gene products has not been studied in *M. trichosporium* OB3b. In the non-methanotrophic methylotroph *Methylobacterium extorquens* AM1, it has been shown that *xoxF* may display methanol-oxidizing activity [50], and can contribute to the complex regulation of *mxa*-genes [51]. Furthermore, there are suggestions that *xoxF* may play a role in formaldehyde oxidation [52]. The low expression of all *xoxF*-homologs in *M. trichosporium* OB3b compared to *mxaFI* or H_4MTP-linked pathway genes suggests that *xox*-genes may have no or a minor contribution to methanol oxidation in *M. trichosporium* OB3b under the tested growth conditions. However, our data do not rule out the possibility that one or more of the *xoxF* gene products are involved in regulation, either of methanol or formaldehyde oxidation.

Pyrroloquinoline quinone (PQQ) biosynthesis is another function essential for operation of the primary methanol oxidation system [53, 47]. A total of six *pqq* genes appear to be present in the *M. trichosporium* OB3b genome in two clusters: *pqqBCDE* and *pqqFG*. Moderate expression of both clusters was observed (Table 2.2). No gene for the small PQQ precursor (PqqA) is predicted in the current version of the genome. Our manual review of the sequences revealed a fragment within the METTOv1_110055 - METTOv1_110057 gene locus (positions 1424678 - 1424755 of current version of the genome) with high sequence identity [83% nucleic acid (NA) identity and 96% amino acid (AA) similarity] to the *pqqA* sequence from *Methylobacterium* spp (Figures 2.2A,B). Transcript mapping data indicated that only the *pqqA*-like region of the locus is highly expressed (Figure 2.2C). The relative expression of the putative PQQ precursor gene is comparable to the high expression of the *mxaFI* genes. The rest of the genes involved in PQQ biosynthesis showed modest to low expression (Figure A.2B).

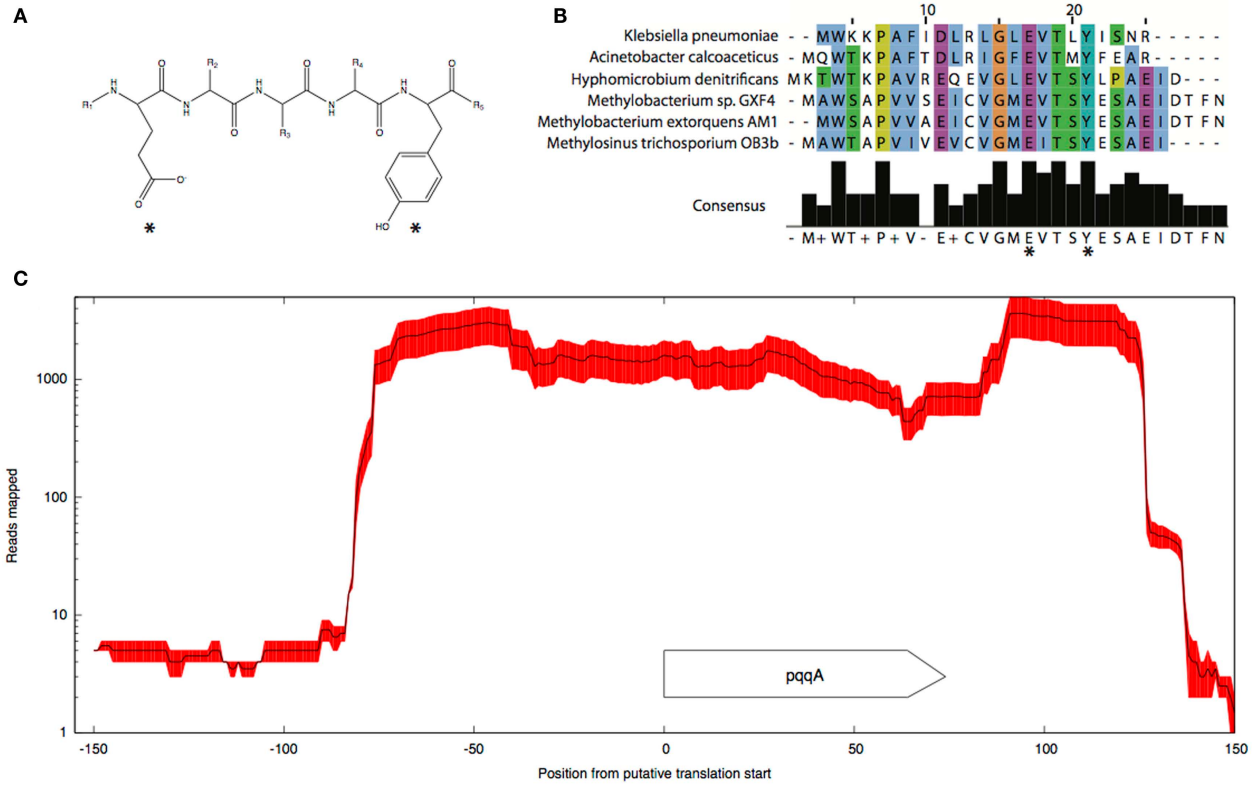


Figure 2.2: Predicted structure (A) and alignment of the putative *pqqA* peptide (B) from *M. trichosporium* OB3b and *pqqA* peptides from Methylobacterium sp. GXF4, *M. extorquens* AM1, *Hyphomicrobium denitrificans* ATCC51888, *K. pneumoniae*, and *A. calcoaceticus*. (C) Mapping.

2.3.4 C_1-Oxidation: Formaldehyde-To-Formate

Previous enzymatic studies predict three possible pathways for formaldehyde oxidation: (1) direct oxidation through dye-linked heme-containing formaldehyde dehydrogenase [54], (2) H_4folate-, and (3) H_4MTP-mediated C_1 transfers [55, 25]. Contrary to enzymatic studies, BLAST searches of the draft genome of *M. trichosporium* OB3b did not reveal any obvious system that could be attributed to heme-containing formaldehyde oxidation. Three broad-specificity aldehyde-detoxification systems, including two NAD-dependent aldehyde dehydrogenases (Aldh-F7 METTOv1_100046 and Aldh, METTOv1_610028) and one aldehyde oxidase (Aor, METTOv1_290006) were predicted in the genome. Two of

them, *aldh* and *aor* show low expression (Table 2.2), while *aldh-F7* was barely detected in the transcriptome. None of the putative genes identified by *de novo* transcriptome assembly could be readily attributed to any dye-linked aldehyde dehydrogenases [56]. Thus, even if the enzyme is present in the genome, its expression during growth of the strain on methane must be low.

For years it has been assumed that methylene H_4F is formed as a result of the spontaneous (non-enzymatic) condensation of formaldehyde with H_4F [57]. It has recently been demonstrated that formate, rather than formaldehyde serves as an entry substrate for assimilation in serine cycle methylotrophs [58]. With this metabolic arrangement, the H_4Folate C_1 transfer could be considered as a part of the assimilatory network that converts formate into methylene H_4F. In *M. trichosporium* OB3b genes encoding all three steps of the H_4Folate pathway converting methylene-H_4F to formate (formyl-H_4F ligase, *ftfL*, methenyl-H_4F cyclohydrolase, *fch*, and methylene-H_4F dehydrogenase *mtdA*) were co-localized and co-transcribed with genes encoding the serine cycle enzymes (Figure A.4). While the assimilatory function of the pathway is more apparent, it is still possible that key enzymes of the pathway contribute to formaldehyde oxidation in *M. trichosporium* OB3b.

The tetrahydromethanopterin (H_4MTP) pathway was proposed to be the key pathway for formaldehyde oxidation/detoxification in alphaproteobacterial methylotrophs [59]. It was speculated that this pathway contributes to formaldehyde oxidation in methane utilizing proteobacteria [55]. Nineteen genes encoding enzymes and genes for tetrahydromethanopterin biosynthesis were identified in the *M. trichosporium* OB3b genome. These genes were not clustered together in the genome, but formed five different gene islands: (1) *mch-orf5-orf7-fae1/1-fae1/2-orf17* (Figure A.2C); (2) *orf19-orf20-afpA-orf21-orf22*; (3) *fhcCDAB*; (4) *orf9-mtdB*; and (5) *pcbD-mptG*. No homologs of *orfY*, *dmrA*, and *pabB* genes, which are commonly associated with tetrahydromethanopterin biosynthesis in methylotrophs [60, 61, 62, 63], were found in the genome or detected in transcriptome.

Formaldehyde activating enzyme (FAE) is the first enzyme of the H_4MTP-pathway,

and has been shown to catalyze the condensation of formaldehyde and H_4MPT to form methylene-H_4MPT [64]. The draft genome of *M. trichosporium* OB3b predicts three homologs of Fae; two of them (*fae1/1* and *fae1/2*) share a high degree of identity (NA 82.2%) and are co-localized in the genome (Figure A.2C). Though both *fae* genes are clustered with four other genes involved in the H_4MTP-pathway, they are expressed in dramatically different patterns. The abundance of *fae1-1* transcripts was almost 40-fold higher than the abundance of any other gene in the cluster, except *fae1-2* (Table 2.2; Figure A.2C). The relative abundance of the second homolog (*fae1-2*) was one fifth of that observed for *fae1-1*, and *fae1-2* was the second highest expressed gene in the pathway. Mapping data indicate that *fae1-1* and *fae1-2* are most likely co-transcribed (Figure A.5). RNA-Seq mapping data suggest two putative transcriptional starts at the positions -215 (with σ^{70}-like -10 and -35 sequences upstream) and -105 (with a conserved "AATGGTT" sequence in the -35 region) upstream from the *fae1-1* translational start (Figure A.5; Table A.3).

The third homolog of Fae (*fae2*) demonstrates moderate expression. The rest of the genes encoding key enzymes of the pathway (*mtdB*, methylene-H_4MPT dehydrogenase; *mch*, methenyl-H_4MPT cyclohydrolase; *fhcABCD*, formyltransferase/hydrolase) have a similarly moderate expression. The relative abundance of genes encoding key enzymes of the pathway were 5- to 10-fold higher than those involved in cofactor biosynthesis. Overall, transcriptomic data indicate that the H_4MTP-pathway serves as the key pathway for formaldehyde oxidation in *M. trichosporium* OB3b.

2.3.5 Formate Oxidation

Formate is oxidized to CO_2 by a NAD-dependent formate dehydrogenase in most, if not all, methanotrophs [2]. It has been suggested that most of the reducing power required for methane metabolism is produced by formaldehyde oxidation to formate and then to CO_2 [7]. It has been speculated that in microbes with a functional serine pathway, formate serves as a key branch point between assimilation and catabolism [65]. NAD-

dependent formate dehydrogenase from *M. trichosporium* OB3b has been purified and characterized [66]. The enzyme is composed of four subunit types and contained flavin, iron, and molybdenum [66]. The genome of *M. trichosporium* OB3b predicts one NAD-dependent molybdenum-containing formate dehydrogenase encoded by *fdsABGCD* and an additional single copy of the alpha subunit (*fdhA*). The two genes *fdsA* and *fdhA* share 81% identity. Only one of them, *fdsA* as well as the rest of the *fds* cluster genes were expressed in the transcriptome (Table 2.2).

2.3.6 C_1-Assimilation: Serine Cycle

It has been previously suggested that the serine cycle is the major pathway for C_1-assimilation in *M. trichosporium* OB3b [31]. All genetic elements essential for operation of the cycle were predicted in the genome and cluster together [14]. However, genes of the pathway show deferent levels of expression (Table 2.2). While *sga*, *glyA*, and *mclA* have high expression levels, the rest of the genes involved in the pathway show modest expression (Table 2.2; Figure A.4). In addition to the serine-glyoxylate aminotransferase (*sga*), a key aminotransferase in the central metabolism of serine cycle microbes, moderate levels of expression were observed for two other aminotransferases, phosphoserine aminotransferase, and aspartate aminotransferase (Table 2.2).

The genome of *M. trichosporium* OB3b encodes two copies of phosphoenolpyruvate (PEP) carboxylase (*ppc1* and *ppc2*). The two enzymes are only distantly related to each other and share 33% identity at the amino acid level. One of them, Ppc1, clusters with PEP carboxylases usually found in bacteria possessing the serine cycle for C_1-assimilation (Figure 2.3). Serine cycle Ppcs belong to a "*non-regulated*" group of PEP carboxylases [2]. The activity of these enzymes is not controlled by intermediates of the TCA cycle or glycolysis/gluconeogenesis [67]. The second homolog of Ppc (*ppc2*) clusters with anaplerotic "*regulated*" PEP carboxylases, which are controlled by a variety of metabolic effectors [68, 69]. Both *ppc1* and *ppc2* transcripts demonstrate comparable

levels of abundance in this study (Table 2.2). The sequences of the two genes were further investigated in an attempt to better understand the rationale for the enzymatic redundancy at the PEP to oxaloacetate conversion step. We used multiple sequence alignments of Ppc1, Ppc2, and other characterized PEP carboxylases and homology models (not shown) built from tensed and relaxed state crystal structures [70] to investigate the predicted allosteric regulation sites of these two enzymes. The alignment shows that only the catalytic elements, such as PEP-binding-site residues, are conserved in both proteins (Ppc1 and Ppc2). However, sequence features required for the allosteric regulation of the enzyme activity show several structural differences. The majority of the characterized bacterial PEP carboxylases are activated by acetyl-CoA, FBP, long-chain fatty acids, and pGp. Inhibition occurs in the presence of aspartate and L-malate. In the case of Ppc1 from *M. trichosporium* OB3b, two of the four highly conserved polar amino acids that bind allosteric inhibitors (e.g., aspartate, malate) were hydrophobic: L805 and A912 (K and N in *E. coli* respectively) suggesting alternate inhibitors or a lack of sensitivity to L-malate and aspartate. The activator-binding residues were conserved except for a R159 instead of K. By contrast, Ppc2 was well conserved relative to the well characterized PEP carboxylases and for those with structures, only minor rearrangements of the monomeric interfaces were predicted. It is tempting to speculate that the presence of two functionally identical but differently regulated enzymatic systems in *M. trichosporium* OB3b evolved as a way to control flux through PEP-oxaloacetate in response to levels of the serine cycle and EMC pathway intermediates. The flux is never completely blocked, due to the insensitivity of Ppc1 to the metabolic state of the cell. Increases in the intracellular levels of acetyl-CoA, aspartate, or malate (as a result of saturation of the downstream EMC pathway and the TCA cycle) can reduce the flux through the PEP-oxaloacetate presumably twofold via allosteric inhibition and lack of activation of Ppc2 activity. In this case, C_1-carbon assimilated via the serine cycle is re-directed to gluconeogenesis or converted into pyruvate.

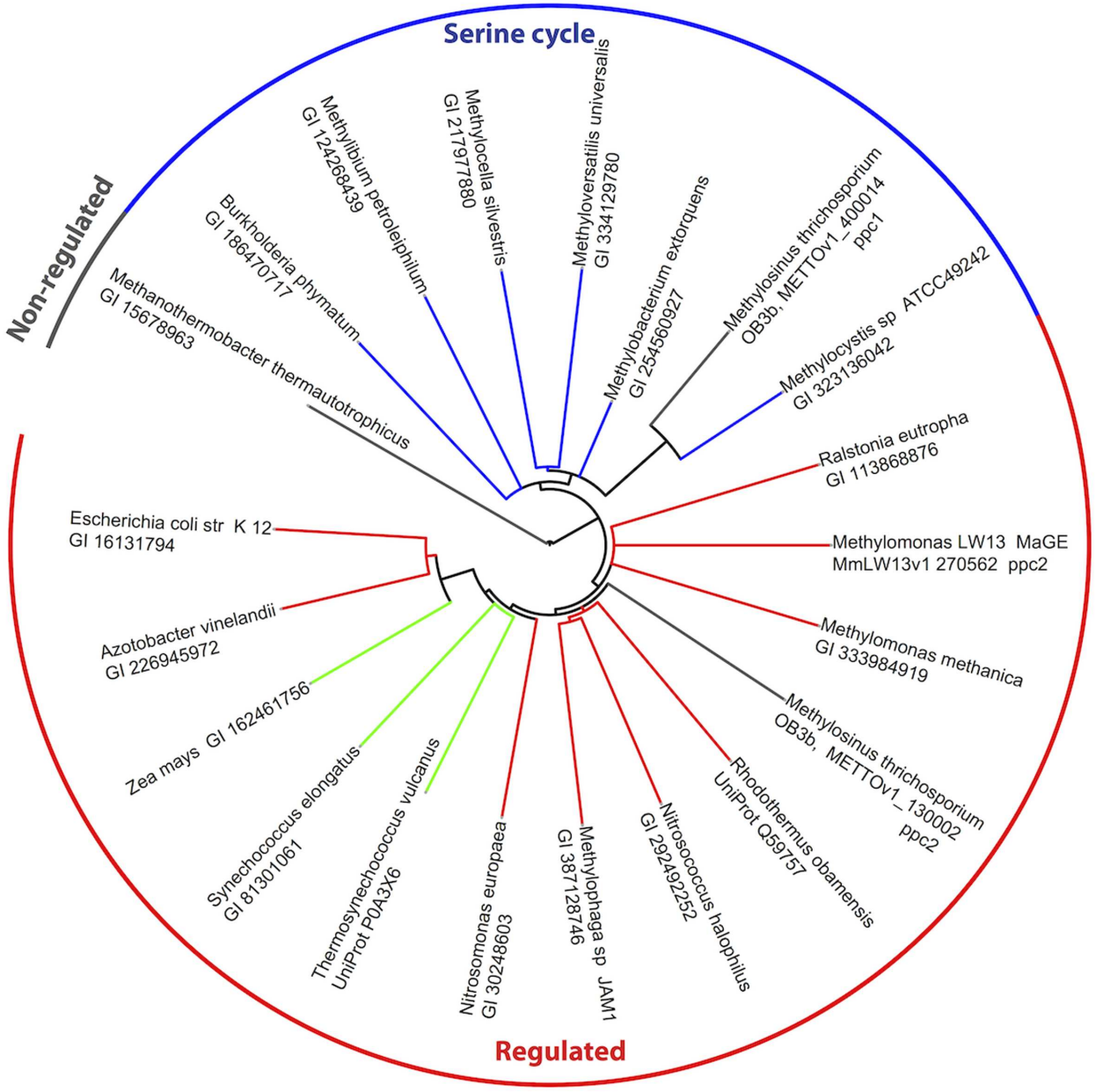


Figure 2.3: Phylogenetic tree of phosphoenolpyruvate carboxylases. Sequence identifiers follow the source organism label. Sequences were aligned with MUSCLE v3.8.31 [71] and the tree created with ClustalW2 2.0.12 [72] and rendered with iTOL [73].

Regeneration of glyoxylate is an essential part of the serine cycle [2, 74, 75]. Like other obligate methanotrophic bacteria, *M. trichosporium* OB3b lacks isocitrate lyase,

a key enzyme of the glyoxylate shunt [9]. Homologs of enzymes involved in the ethylmalonyl-CoA (EMC) pathway, an alternative route for glyoxylate regeneration [75], were identified in the draft genomes of *M. trichosporium* OB3b and another obligate methanotroph *Methylocystis* sp. [14, 15]. However, a functional EMC pathway has not yet been demonstrated in methanotrophs. Furthermore, the recent investigation of PHB-metabolism in *Methylocysctis parvus* OBBP, an alphaproteobacterial methanotroph, suggested that this metabolic module can not supply C_2 (glyoxylate) units for biosynthesis [76]. As the initial steps of the EMC pathway are shared with PHB biosynthesis [acetyl-CoA acetyltransferase (*phaA*) and acetoacetyl-CoA reductase (*phaB*), see Figure 2.1] in context with the data presented by Pieja et al. [76] call into question the operation of the EMC pathway in methanotrophs.

We found that genes encoding the initial steps of the PHB-synthesis (*phaA*/*phaB*) show moderate levels of expression. As could be expected for cells in early-mid exponential growth, the expression of the gene encoding PHB synthase (*phaC*) was low. The data are consistent with previous observations of high activity of PhaA and PhaB and low activity of PHB synthase (PhaC) in exponentially grown cells of *M. trichosporium* OB3b [24, 25]. The PHB-degradation pathway genes, including 3-hydroxybutyrate dehydrogenase, acetoacetate decarboxylase, and acetoacetyl-coenzymeA synthetase, show modest expression levels (Table 2.2).

All homologs of the EMC pathway enzymes were expressed in *M. trichosporium* OB3b during growth on methane (Table 2.2). Several putative acetyl/propionyl-CoA carboxylases are predicted in the genome, however only METTOv1_200020 (putative *ppcA*) and METTOv1_220035 (putative *ppcB*) were expressed. Furthermore, the PpcAB and crotonyl-CoA reductase (*ccr*) genes display the highest level of expression among all CO_2-fixing enzymes in the *M. trichosporium* OB3b transcriptome. Thus, the transcriptional profile of *M. trichosporium* OB3b indicates the methanotroph may possess an active EMC pathway.

2.3.7 C_1-Assimilation: TCA Cycle and Anaplerotic CO_2-Fixation

All previous enzymatic studies predict that the tricarboxylic acid cycle (TCA cycle) in alphaproteobacterial methanotrophs is complete [9]. However, the functional role of this metabolic pathway in methanotrophs is not fully understood. It has been suggested that the main role of the TCA cycle in the methanotrophs is carbon assimilation rather than energy production, due to low enzyme activity and lack of pyruvate dehydrogenase [77, 2, 78, 9]. However, labeling studies on acetate and pyruvate utilization have predicted the presence of a catabolic TCA cycle in type II methanotrophs [79, 80]. *In silico* genome analysis indicates that *M. trichosporium* OB3b contains predicted genes for all key enzymes of the TCA cycle and pyruvate dehydrogenase (*pdh*). All of these genes were expressed (Table 2.2). These steps of the TCA cycle are shared between the EMC pathway and the serine cycle (Figure 2.1). *De novo* transcriptome assembly indicated that *M. trichosporium* OB3b possesses an additional homolog of succinate:ubiquinone oxidoreductase (*sucABCD*, Table A.1). Genes encoding the succinate:ubiquinone oxidoreductase and succinyl-CoA synthase (*sucCD*) are among the most highly expressed TCA cycle functions. The reductive branch of the pathway (including genes for citrate synthase, aconitase, isocitrate dehydrogenase, and 2-ketoglutarate dehydrogenase) displays moderate-to-low expression. Low expression of *pdh* genes is consistent with the previous enzymatic studies that show low/no activity of pyruvate dehydrogenase [77].

It has been shown that the CO_2-fixation potential is maximal during early stages of logarithmic growth [37, 81]. However, data on carboxylation system(s) in *M. trichosporium* OB3b are controversial. Most previous enzymatic studies predict that the PEP carboxylase (Ppc), a key enzyme of the serine cycle, is the major entry point for CO_2 in alphaproteobacterial methanotrophs [78]. On the other hand Naguib ([82], 1979) has shown that *M. trichosporium* OB3b possesses different carboxylation systems, including membrane bound and cytoplasmic enzymes. It could be predicted that the

EMC pathway also contributes to CO_2 assimilation. *In silico* analysis of the genome sequence also revealed that in addition to the CO_2-fixing functions described above, genes for NAD(P)-dependent malic enzyme (*mae*), acetyl-CoA carboxylase (*accABD*), phosphoribosyl aminoimidazole carboxylase, pyruvate carboxylase (*pcx*), and a putative 2-oxoacid ferredoxin oxidoreductase are all present. All of these genes were expressed (Table 2.2).

2.3.8 Glycolysis/Gluconeogenesis and Pentose-Phosphate Pathways

The absence of enzymatic activity for the initial steps of the gluconeogenic pathway including pyruvate-PEP or oxaloacetate-PEP conversions was one of the most common explanations for the inability of alphaproteobacterial methanotrophic bacteria to grow on poly-carbon compounds such as pyruvate or acetate [83, 78]. No homolog of PEP-carboxykinase was found in the *M. trichosporium* OB3b genome. However, contrary to expectations based on enzymatic inferences, a set of pyruvate-acetyl-CoA, pyruvate-PEP, and pyruvate-malate interconversions could be predicted from the genome annotation. Homologs of PEP synthase (*pps*), pyruvate kinase (*pyk1* and *pyk2*), and pyruvate phosphate dikinase were detected. The relative abundances of *pyk1*, *pps*, and *pdk* transcripts were low, but a second pyruvate kinase (*pyk2*) displayed modest expression (Table 2.2).

During growth on C_1 compounds (methane or methanol), gluconeogenesis starts with conversion of 2-phosphoglycerate into 3-phosphoglycerate by phosphoglycerate mutase (*pgm*). This metabolic step represents the main branch point of the serine cycle. Four homologs of *pgm* were identified in the *M. trichosporium* OB3b genome. Three of them (METTOv1_10180, METTOv1_100061, and METTOv1_680013) were expressed at tested conditions. Homologs of the genes for the rest of the enzymes in the pathway were detected in the genome, mostly in single copies. All gene transcripts were observed in the RNA-Seq data (Table 2.2).

The genome analysis suggests that the pentose-phosphate pathway (PPP) is incomplete in *M. trichosporium* OB3b. Glucose-6-phosphate dehydrogenase, gluconolactonase and phosphogluconate dehydrogenase (oxidative PPP), and transaldolase or sedoheptulose bisphosphatase (non-oxidative PPP), are missing in the genome. In addition, no homologs of the genes were detected among *de novo* assembled transcripts. The lack of the oxidative branch of the PPP is consistent with previous enzymatic data and the inability of alphaproteobacterial methanotrophs to utilize sugars. However if the non-oxidative PPP operates as a route for generation of ribose-5-phosphate for the synthesis of nucleotides, an unknown enzyme must be involved in the sedoheptulose-phosphate interconversion. One possible system is a pyrophosphate-dependent phosphofructokinase (Pfk). It has been shown that Pfks from methanotrophic bacteria have surprisingly high affinity for sedoheptulose phosphate, and can catalyze the conversion of sedoheptulose-1,7-bisphosphate to sedoheptulose-7-phosphate [84, 85]. It is possible that Pfk contributes to sedoheptulose-1,7-bisphosphate conversion in *M. trichosporium* OB3b.

2.3.9 Lipid Metabolism

Methane oxidation via particulate methane monooxygenase is linked to formation of extensive intracellular membranes. It has been shown that lipid/biomass content of *M. trichosporium* OB3b cells grown on methane is 9.2% and that phospholipids represent a significant fraction of membrane lipids (83.4%) [86, 87]. Concurrent with previous observation, genes essential for biosynthesis of major fatty acids (stearic, oleic, and palmitoleic acids) and phospholipids (including phosphotidylcholine, phosphatidylglycerol, phosphatidylserine, and phosphatidylethanolamine) show moderate level of expression (Table A.2).

2.3.10 Nitrogen, Copper, and Iron Metabolism

The pathways of nitrogen assimilation have been studied in a number of obligate methanotrophic species including *M. trichosporium* OB3b. Methanotrophs are able to grow with ammonia, nitrate, and molecular nitrogen as N-sources [19, 88, 89, 90, 91, 92]. No activities of alanine or glutamate dehydrogenases were detected in cell extracts of *M. trichosporium* OB3b grown on any source of nitrogen [88, 89, 90]. It has been concluded that ammonia was assimilated exclusively via the glutamine synthetase/glutamate synthase pathway [90].

In this study, cells of *M. trichosporium* OB3b were grown using nitrate as the N-source. Despite the presence of an exogenous source of nitrogen, some (very low) expression of the nitrogenase gene cluster was observed. Relative expression of nitrogenase structural genes (*nifHDK*) was about four to five times higher than expression of the chaperone and cofactor biosynthesis genes (Table A.2). High expression of genes involved in assimilatory nitrate reduction, including the nitrate transporter (*nrtA*), nitrate reductase (*nasA*), and nitrite reductase (*nasDE*) was detected. Interestingly, moderate expression of a putative ammonium transporter (METTOv1_130049) was detected, although a gene cluster with putative involvement in hydroxylamine detoxification (METTOv1_230076-78), which should only be needed under conditions of high ammonium concentration, showed low expression (Table 2.2). The gene encoding cytochrome c'-alpha (METTOv1_560023), a protein implicated in NOx detoxification, was moderately expressed. Homologs of alanine dehydrogenase (METTOv1_280018) and glutamate dehydrogenase (METTOv1_190023) were identified; however, neither gene was expressed (Table 2.2). High expression of glutamate synthase (both NADH and Fd-dependent), glutamate-ammonia ligase (METTOv1_200046), and Type I glutamine synthetase (METTOv1_200048) was observed. Based on these transcriptomic studies and genome analysis, the only pathway for alanine biosynthesis is transamination of pyruvate. The most likely enzymatic system for this conversion is the serine-glyoxylate

aminotransferase (Sga), which is known to catalyze serine-pyruvate transamination [93]. It has been shown that alanine may serve as alternative substrate for SGA in methylotrophs [94].

Copper is an important microelement in the physiology of methanotrophic bacteria possessing pMMO [2]. Methanotrophic bacteria synthetize methanobactin (Mb), a copper-chelating compound [95, 96, 27]. It has been suggested that Mb provides copper for the regulation and activity of methane oxidation machinery in methanotrophs [97, 27]. It has been shown that Mb is a peptide-derived molecule. A gene encoding the Mb precursor in *M. trichosporium* OB3b has recently been identified [98]. We found that the Mb-gene is among the top 5% most abundant transcripts despite the fact that the culture in our experiments was grown with sufficient Cu (Table 2.2). It is not known how Mb is synthesized or cleaved, however it has been suggested that genes downstream of *mb* are involved [98]. These genes were also expressed, but the expression level was low.

Iron is another essential metal in C_1-metabolism. It has been previously observed that *M. trichosporium* OB3b can produce a Fe-chelating compound [99]; however the siderophore structure and pathways for its biosynthesis remain to be discovered. The production of a fluorescent compound was observed on plates at tested growth conditions (data not shown). Our transcriptomic studies revealed relatively high expression of genes homologous to those involved in pyoverdine (*pvd*) I biosynthesis, excretion, uptake, and regulation (Table 2.2), making this a possibility for a siderophore. All four essential non-ribosomal peptide synthetases (*pvdLIJD*) were identified. Unfortunately, the *pvd* genes are represented in fragments in the current genome assembly, making it impossible to predict the order of the amino acids in the peptide product.

2.4 Conclusion

In this work we performed genomic- and transcriptomic-based reconstruction of the central metabolic pathways in *Methylosinus trichosporium* OB3b grown on methane as a sole source of carbon and energy. The overview of the methane metabolism is

summarized in Figure 2.1. While some metabolic functions correlate well with previous enzymatic and genetic studies, several novel functions were detected and characterized. The major outcomes of our work are listed below:

1. Exceptionally high expression of pMMO in comparison to other central pathway functions (such as methanol or formaldehyde oxidation) implies a relatively low turn over at the first step of methane conversion.

2. We propose that *M. trichosporium* OB3b uses the EMC variant of the serine cycle for carbon assimilation. In addition to carbon fixing reactions of the EMC-serine cycle, a number of carboxylation reactions are predicted. The role of CO_2 fixation during methanotrophic growth has further been explored by Yang et al. [6].

3. The diversity of predicted reactions at the PEP-pyruvate-oxaloacetate node suggests that metabolic interconversions may play an important role in the distribution of carbon flux between the serine cycle, EMC pathway, and TCA cycle. In *M. trichosporium* OB3b the PEP-oxaloacetate conversion is predicted to be performed by two enzymatic systems under different metabolic control. Increases in the intracellular pools of malate, aspartate, and acetyl-CoA could activate flow of C_1-derived carbon into gluconeogenesis and/or pyruvate. Our results indicate that multiple PEP-pyruvate conversion reactions may be taking place in the strain during growth on methane as a way to regenerate energy and to provide pyruvate for biosynthesis. Due to the lack of PEP-carboxykinase, the PEP synthesis from C_4 compounds is also possible and could be achieved via decarboxylation of malate (MalE). Two reactions are predicted for PEP synthesis from pyruvate, however both of them seem to be of minor importance during growth on methane.

4. A number of transamination reactions contribute to carbon partitioning and nitrogen assimilation. It has been predicted that the growth of majority of alphaproteobacterial methylotrophic bacteria is NAD(P)H-limited due to the high NADH-requirements for formaldehyde assimilation via serine cycle [100]. Biosynthesis of key amino acids (such as alanine, glutamate and aspartate) via transamination seems to be a rational metabolic compensation to NAD(P)H-limitation.

5. While copper acquisition is quite well characterized in *M. trichosporium* OB3b, relatively little is known about iron uptake systems. Transcriptomic data provide initial evidence for siderophore production in this methanotroph.

2.5 Materials and Methods

2.5.1 Strain and Cultivation Conditions

Methylosinus trichosporium strain OB3b was kindly provided by Dr. Lisa Stein. The culture was grown in 250 mL glass bottles on modified NMS medium [19] containing (per liter of distilled water): 1 g·KNO_3, 1 g·$MgSO_4 \cdot 7\,H_2O$, 0.134 g·$CaCl_2 \cdot 2\,H_2O$, 0.25 g·KH_2-PO, 0.7 g·$Na_2HPO_4 \cdot 12\,H_2O$, and 2 mL of trace elements solution. The trace elements solution contained 0.5 g·Na_2-EDTA, 1.0 g·$FeSO_4 \cdot 7\,H_2O$, 0.75 g·$Fe-EDTA$, 0.8 g·$ZnSO_4 \cdot 7\,H_2O$, 0.005 g·$MnCl_2 \cdot 4\,H_2O$, 0.03 g·H_3BO_3, 0.05 g·$CoCl_2 \cdot 6\,H_2O$, 0.4 g·$Cu-EDTA$, 0.6 g·$CuCl_2 \cdot 2\,H_2O$, 0.002 g·$NiCl_2 \cdot 6\,H_2O$, and 0.05 g·$Na_2MoO_4 \cdot 2\,H_2O$. The bottles were sealed with rubber stoppers and aluminum caps, then 50 mL of methane was added to the 200 mL headspace. Bottles were shaken at 250 RPM at 30^{o}C for 1-4 days.

2.5.2 Growth Parameters and Methane Consumption Rate Measurements

Methane consumption rates and cell density (OD_{600}) were measured in triplicate as cultures grew. Methane measurements were made on a Shimadzu Gas Chromatograph GC-14A, using the FID detection with helium as the carrier gas. Concentrations were deduced from standard curves. OD_{600} was measured on a Beckman DU® 640B spectrophotometer in plastic 1.5 mL cuvettes with a 1 cm path length.

2.5.3 RNA-seq

Two replicate cultures were grown to mid exponential phase (OD_{600} 0.29 $\pm$ 0.01) for approximately 24 h, then collected by pouring 45 mL of culture into 50 mL tubes containing 5 mL of *stop solution* comprised of 5% water-equilibrated phenol, pH 6.6

(Sigma; St. Louis, MO, USA), and 95% ethanol (200 Proof; Deacon Labs, Inc., King of Prussia, PA, USA). The cells were collected by centrifugation at 4,300 $\times g$ at 4^{o}C for 10 min. The resultant pellet was re-suspended in 0.75 mL of extraction buffer [2.5% CTAB (Sigma; St. Louis, MO, USA), 0.7 M NaCl, and 0.075 M pH 7.6 phosphate buffer] and transferred to a 2 mL sterilized screw-cap tube containing 0.75 mL of phenol:chloroform:isoamylic alcohol with a volume ratio of 25:24:1 (Ambion®; Austin, TX, USA), 0.5 g of 0.1 mm silica beads (Biospec products; Bartlesville, OK, USA), 0.2% SDS (Ambion®; Austin, TX, USA), and 0.2% lauryl sarkosine (Sigma; St. Louis, MO, USA). The mixtures were homogenized in a bead beater (Mini-Beadbeater; Biospec Products; Bartlesville, OK, USA) for 2 min (75% of the maximum power). The resulting slurry was centrifuged for 5 min at 4^{o}C and 20,800$\times g$. The aqueous layer was transferred to a fresh tube containing 0.75 mL of chloroform:isoamylic alcohol with a volumetric ratio of 24:1 (Sigma; St. Louis, MO, USA) and centrifuged again for 5 min at 4^{o}C and 20,800$\times g$ to remove dissolved phenol. The aqueous phase was transferred to a new tube. $MgCl_2$ (final concentration 3 mM), sodium acetate (10 mM, pH 5.5), and 0.8 mL icecold isopropanol were added. Nucleic acids were transferred to -80^{o}C for overnight precipitation. Precipitated samples were centrifuged for 45 min at 4^{o}C and 14,000 RPM (20,800$\times g$), washed with 0.5 mL of 75% ethanol (made from 200 proof; Deacon Labs, Inc., King of Prussia, PA, USA), and dried for 15 min at room temperature.

An RNeasy Mini Kit (Qiagen©; Venlo, Netherlands) with two types of DNA digestions was used to isolate the mRNA. Initially, the DNA/RNA pellet was re-suspended in 80 mL of a DNase I (RNase-free) mixture (Ambion®; Austin, TX, USA) and incubated for 30 min at 37^{o}C. Then, the samples were purified on RNeasy Mini Kit columns as described in the RNA cleanup section of the manual, including the optional on-column DNAse digestion. The MICROBExpress™ (Ambion®; Austin, TX, USA) kit was applied to each sample to reduce the rRNA concentration and increase the mRNA sequencing depth.

The sample quality was monitored with three techniques: (1) by electrophoresis

in TAE buffer in 1% agarose gels (2) using an Agilent 2100 Bioanalyzer with Agilent RNA 6000 Nano-kit as suggested by the manufacturer, and (3) by real-time reverse-transcriptase PCR (RT-RT PCR) with 16S rRNA (27F/536R) and *pmoA*-specific [29] primers.

2.5.4 Transcript Sequencing, Alignment, and Mapping

Enriched RNA samples (i.e., two biological replicates) were submitted to the University of Washingtons High-Throughput Sequencing Solutions Center on dry ice for single-read Illumina® sequencing (Department of Genome Sciences, University of Washington). The replicates were aligned to the reference genome using BWA under default parameters [32]. The "METTOv1" genome sequence was downloaded from MaGE [33]. The single large pseudo scaffold distributed by MaGE was split into 187 separate contigs at each stretch of N bases. In addition, chimeric contigs from the assembly and low quality gene calls were removed (Table A.5). The summary of RNA-seq (Illumina) reads can be found in Table A.4. The METTOv1 genome did not have a complete *pmoCAB* cluster suitable for alignment. In order to include *pmoCAB*, a separate alignment run was performed with the previously published sequence of this gene cluster [43]. After the alignment with BWA, SAM tools was used to generate a *pileup* file that was loaded into a MySQL database for normalization from Reads Per Kilobase of gene per Million mapped reads (RPKM) to coding sequences [34].

2.5.5 Transcription Site Mapping and Transcriptome Based Gene Assembly

The reads mapped at each base position in the genomic scaffolds generated from the *pileup* were manually examined to identify putative transcription starts and stops. Briefly, reads mapped per base data for the two replicates were plotted on a log scale. The boundary of a rapid transition from near zero reads mapped to 10, 100 RPKM or more that was upstream of a gene start was designated a transcription start. Stops were similarly identified as

a rapid transition to low numbers of reads mapped downstream of a gene termination codon.

Given the fragmented nature of the *M. trichosporium* OB3b genome, we performed *de novo* assembly of the RNA-Seq reads, in an attempt to identify transcripts whose genomic sequences were incomplete. The assembly was performed with Velvet 1.2.06 [101] and Oases 0.2.08 [102]. The oases pipeline tool distributed with Oases was used to survey assemblies across the range of odd k-mers from 17 to 35 where the minimum fragment length was set to 100 bp. The final merged assembly from the pipeline tool was stripped down to the highest confidence transcript for each locus with confidence ties resolved by taking the longest sequence. The high confidence assembled transcripts were aligned to the *M. trichosporium* OB3b scaffolds using BLASTn. Transcripts without significant matches were aligned with BLASTx to the protein non-redundant database, as retrieved on January 13, 2012.

Chapter 3

MICROBIAL COMMUNITY ANALYSIS: METAGENOMICS AND METATRANSCRIPTOMICS

3.1 Abstract

Methane is one of the most potent and common greenhouse gases. In nature, specialized bacteria called methanotrophs are able to consume methane as their only carbon and energy source [8]. Deeper understanding of these microbes sheds light on a significant natural green house gas remediation system. The sediment of Lake Washington has served as a model ecosystem for methanotrophy studies [103, 104, 105, 106, 107, 108] for decades. Recent advances in high-throughput sequencing technology provided opportunities to observe the complex communities supported by methanotrophic methane oxidation in a more natural state than has been possible previously. This study analyzes serial propagations of Lake Washington sediment incubated with methane as the only carbon and energy source, from both a metagenomics and metatranscriptomics perspective. Inference of which microbes dominate the samples, and what metabolic functions are active are discovered. The importance of O_2 as a factor in community composition and function is highlighted.

3.2 Introduction

3.2.1 Lake Washington Methane Cycling Studies

Methanotrophs are concentrated at the transition from the aerobic zone to the anaerobic zone in lake sediments. This interface corresponds to both availability of methane, which rises from below, and O_2 which diffuses from above [109, 110, 111]. This microenvironment is rich in methanotrophs, methylotrophs, and other species. When

natural sediment samples are incubated in the lab with methane as the only carbon and energy source, a high abundance of non-methanotrophs is often supported [112]. Many of the abundant non-methanotrophs are methylotrophs, presumably consuming methanol, an intermediate of methane metabolism [113]. There are also many species which can only grow using multi-carbon compounds, and therefore must be consuming excreted organic multi-carbon compounds, or cell biomass. Trends in the types of communities that form in these sediment incubations have been observed [112]. Some species pairs appear to occur more than others and have been proposed to interact [113].

Early studies addressed these questions by sampling specific types of DNA from Lake Washington sediment (e.g. [103, 105, 114]), and studying isolates for which genome sequences were not available [111, 115, 116]. Later, significant effort was put toward isolation and sequencing of more species known to be active in Lake Washington sediment [106, 117, 107, 108] to better understand these communities.

Awareness that not all functionally important microbes can be isolated [118, 119], and that microbes probably behave differently when growing in communities due to influence of other microbes [120] motivates study of these microbes in their natural community compositions. The omics studies used to observe pure cultures (e.g. Chapter 2) can be applied to communities, though analysis becomes much more challenging. When omics methods are applied to mixed populations of organisms, the prefix ”meta” is applied, resulting in terms such as ”metatranscriptomics”. Correspondingly the term ”metagenomics” is used to describe sequencing the DNA of the community to allow inference of the abundant taxa and their genetic composition. Having the paired metagenomic and metatranscriptomic data allows estimation of gene expression for those microbes, by mapping sequences to the reference DNA identified from the metagenomes.

Previous metagenomics and metatranscriptomics studies of Lake Washington sediment have highlighted dominance of the methanotrophic family *Methylococcaceae* [121, 122, 112, 123] (Figure 3.1). These methanotrophs provide substrates for non-methanotrophic methylotrophs to grow [121]. The particular methanotroph species

and methylotroph species that dominate the sample are known to be influenced by O_2 availability [123]. In previous studies, it was noted that low O_2 tensions select for *Methylotenera/Methylobacter* partnerships whereas high O_2 tensions select for *Methylophilus/Methylosarcina* partnerships [123]. There is also a revolving cast of (mostly) non-methylotrophs including Burkholderiales and Bacteroidetes [124, 122]. Better understanding of the metabolic roles that each of these species play, and why certain partnerships tend to form would provide insight into this greenhouse-gas mitigating microbial community. Such insights are important for predicting future changes in new methane emissions to the atmosphere.

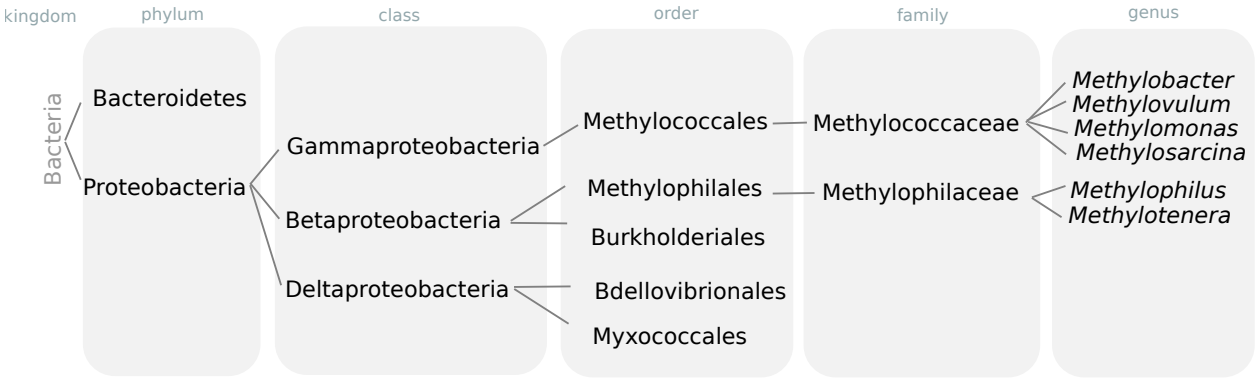


Figure 3.1: Taxonomy of microbes that often appear in methane incubations of Lake Washington sediment. The family *Methylococcaceae* is methanotrophic. The family *Methylophilaceae* includes non-methanotrophic methylotrophs. Burkholderiales grow on multi-carbon compounds, but in addition, some are methylotrophs.

Goals of this study

This study aims to identify the major methanotrophic, methylotrophic, and other microbial species that together enable methane consumption in Lake Washington, as well as to deduce the contribution of each taxa to the community metabolism. Identification of which methanotrophs dominate in natural communities provides insights into how well past isolate studies reflect the true drivers of methane oxidation. Identification of

which metabolic pathways are expressed by these methanotrophs and the accompanying non-methanotrophs informs the mechanism of methane oxidation in this natural system. Understanding the contribution of each microbe to the community metabolism generates hypotheses about genetic factors in methanotroph/methylotroph partnerships.

The experimental design chosen by Maria Hernandez and Ludmila ("Mila") Chistoserdova to answer these questions is shown in Figure 3.2.

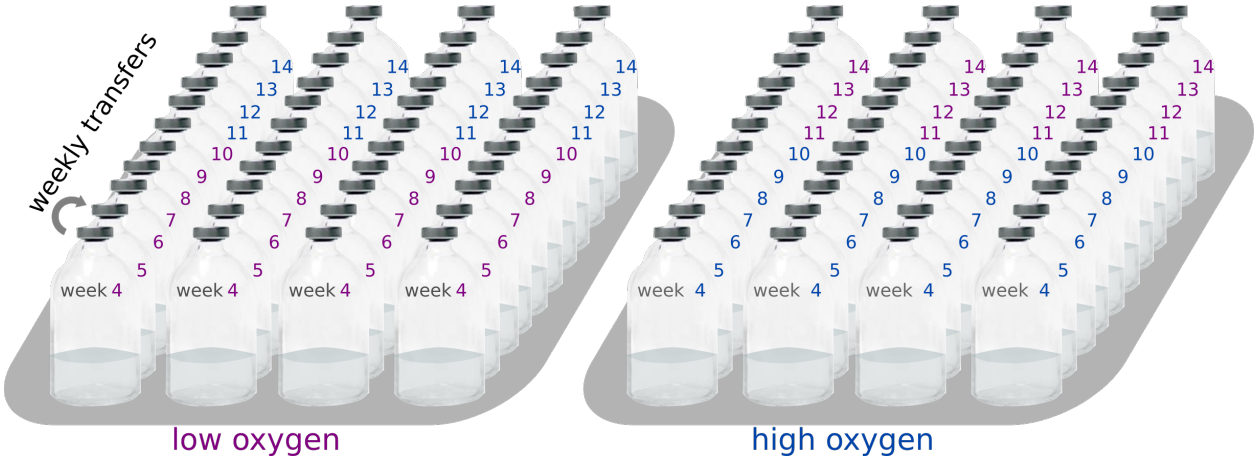


Figure 3.2: Experimental design for the data underlying Chapters 2 and 3. Sediment from Lake Washington was thawed from -80°C and cultured in 8 different bottles, half of which were provided with low O_2, and half of which were provided with high O_2 (see methods). Bottles were serially transferred for a total of 14 weeks. The last four bottles in each series are at the opposite O_2 condition of the original experimental design (indicated by bottle label color switch). Metagenomes and metatranscriptomes were obtained for weeks 4-14, resulting in 88 metagenomes and 88 metatranscriptomes. This was done by Dr. Maria Hernandez and Dr. Mila Chistoserdova.

This relatively simple experimental design, including high degree of replication for studies of this type, was chosen to provide statistical power despite differences across replicates and measurement noise. Variation across replicates can arise from stochasticity as communities rarefy, and are compounded by the noise-generating steps of nucleic acid extraction, ribosomal RNA removal (in the case of metatranscriptomics), library prep, and the sequencing process. The most important environmental variable identified in

previous studies, the O_2 availability, was modulated while holding all other variables constant. Availability of sequenced isolate strains from the same ecosystem enhances exploration of the dataset by providing some ground-truth. Lake Washington isolates can also be integrated into positive controls for many of the computational methods.

Though sequencing has become routine in most biological domains, this dataset is exceptional for several reasons. Having four replicates for each experimental condition is much greater than typical metagenomic studies. Furthermore, most metagenomics/metatranscriptomics studies are single time point snapshots, rather than time-series. Having 11 samples in each series leads to a total of 88 samples. For each of the 88 samples, untargeted metagenomes and metatranscriptomes were gathered. These pairs allow exploration of the community without the restriction of referencing the genomes of cultured strains. Ribosomal RNA was depleted, so the majority of the information in the transcriptomes derives from mRNA. In all the dataset totals to 9TB. My role was in this study was to analyze these extensive datasets.

3.2.2 Challenges and Strategies

Despite the size and replication of this dataset, answering the questions outlined above proved challenging. First, the study aims to answer who is there, without use of reference DNA for read mapping. Thus, the first task is to assemble reference DNA from the vast number of short sequences provided. This DNA then becomes the basis for answering both ”who is there”, and the reference for ”what are they doing”. Ideally, the assembly yields a small number of long sequences. More often than not, however, under-sampling of DNA or lack of punctuation in well-sampled (highly abundant) strains leads to fragmented reference DNA. Such fragmentation causes uncertainty and information loss to propagate through the rest of the analysis pipeline, so care must be taken when interpreting all downstream results (see Figure 3.4). This thesis addresses these challenges by developing numerous metrics for inference efficacy and propagation of uncertainty

(see Section 3.3.2).

Tool selection and evaluation is also challenging. As discussed in the introduction, metagenomics and metatranscriptomics of natural microbial communities is a rapidly developing field laden with methodological challenges. Numerous tools are available for each inference step. Some understanding of the underlying methods is necessary for prudent selections. Given that each tool has different efficacy for different datasets, several tools are usually tried and evaluated for each step. For each tool, there are numerous settings the authors encourage users to tune, leading to a combinatorial explosion of potential outcomes for each inference step. The nearly complete turnover of tools approximately every 5 years leads to very few benchmarking studies, and review papers that are out date within months of publication. Furthermore, benchmark studies comparing tools use a small number of datasets that may not behave like your own dataset. The approach for this thesis included consideration, testing, and tuning parameters of multiple software packages for each challenging step.

The field also lacks standards for assessing the efficacy of each inference step. This is in part due to differences in goals across metagenomics/metatranscriptomics studies: some aim for high confidence inference about a specific sub-population such as a novel taxa, whereas others aim to describe the sample more broadly. Many tools provide only limited insight into their efficacy on your particular dataset. Care must be taken by the investigators to select the right tools, connect them properly, and evaluate results critically. Often the trade-offs between tools are not evident until they are tested, and the data are explored with a critical eye. Success in this field requires iteration as these insights are developed. This thesis advocates for reporting more measures of information loss to shed light on the efficacy of each computational step.

Furthermore, the size of this dataset is both luxurious and challenging. Yes, the high degree of replication and number of samples per series is essential for identifying hypotheses in noisy data. However, many computational tools are not designed to scale to input data of this size, leading to a variety of failure modes. This work addressed the

size challenge by using large memory AWS instances for memory-intensive steps, and doing analyses on subsets of the data when appropriate.

The following sections describe these challenges in more detail, and approaches we took to provide a coarse description microbial abundances and metabolic contributions. Steps to zoom in on particular microbes to tell more detailed stories about a subset of the population are outlined as future directions.

3.3 Metagenomics and Metatranscriptomics

3.3.1 Introduction to Metagenomics and Metatranscriptomic Inference

Metagenomics and metatranscriptomics are umbrella terms for any project that uses sequencing of mixed populations of organisms. It implies untargeted sampling of the entire DNA or RNA pool, rather than selecting for specific types of sequences. This is in contrast to 16S surveys, which amplify DNA from a segment of the conserved bacterial 16S ribosome [125]. 16S DNA sequences are used as molecular clocks for large timescales, allowing inference of which bacterial taxa are present in samples. 16S studies do not allow the possibility to determine what genes each organism has available. Sometimes two organisms are nearly identical at the 16S level, but have substantial metabolic differences due to insertions, deletions, and transposons. The term ”metagenomics” is typically reserved for un-targeted sampling of the entire community DNA pool. Metagenomics affords the possibility of identifying the genetic composition of individual microbes, albeit after much greater analysis efforts. Prior studies of Lake Washington sediment provided a 16S-based perspectives of who is present [121, 123, 112], and carried out preliminary untargeted metagenomics [121, 112]. This thesis chapter focuses solely on untargeted metagenomics and metatranscriptomics, to address the more challenging questions of microbial genetic composition, and the associated gene expression levels.

Once shotgun metagenomic samples are in hand, there are many possible ways to infer which organisms are present and what they express. One approach is to assess

how the samples relate to isolate organisms, by mapping reads to the genomes of those isolates. High mapping rates to an isolate genome suggest the sample contained a large abundance of a similar organism. Given enough similarity, mapping of DNA and RNA reads to these genomes can inform the abundances and expression patterns for organisms in the samples. This strategy is only effective if the isolate genomes accurately represent strains in the community. Isolates are often poor proxies for natural microbes given that the vast majority of bacteria in the environment are not currently culturable [118, 119]. Even if the isolates are perfect proxies for some organisms, the set may be missing whole categories of other organisms. In this project, we carried out a preliminary study using isolate genomes as a reference for mapping metatranscriptome reads, but only a small fraction of the RNA was found to map to genes in the isolate genomes (Figure 3.8). We then pivoted to reference-genome free workflows, whereby the metagenomic reads were assembled into contigs that collectively represent the DNA of the higher abundance microbes. These contigs can be annotated to identify genes, and organized into genome bins intended to represent single strains or groups of highly similar strains.

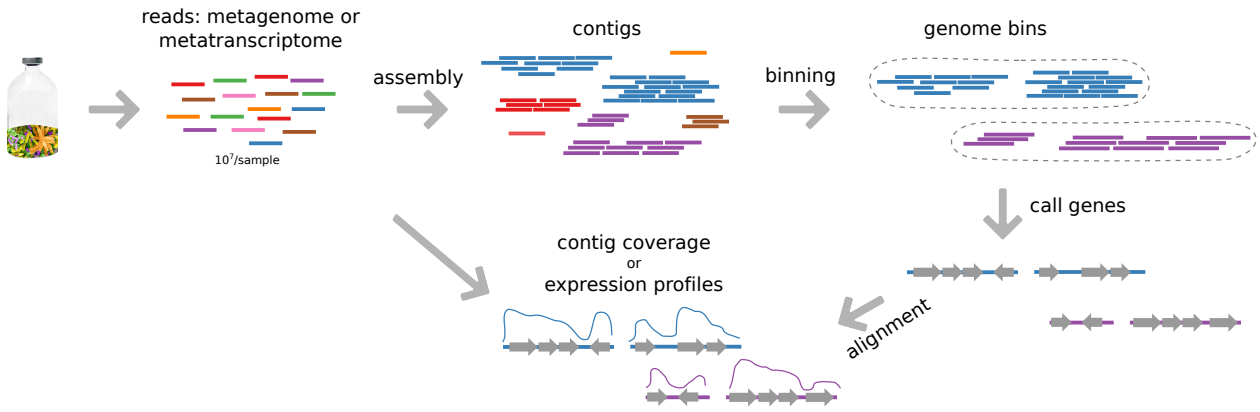


Figure 3.3: Graphical introduction to basic vocabulary and inference steps of metagenomics and metatranscriptomics workflows. Short colored lines = single reads, long colored lines = contigs, gray arrows on top of colored lines = gene calls, wiggly lines above contigs depict the coverage of an alignment.

A simplified representation of steps common to most shotgun metagenomics papers is depicted by Figure 3.3. The first step of reference-free metagenomics is assembly, wherein contigs are aligned and fragments are merged to infer the sequence from which the reads originated. Binning aims to identify contigs that in aggregate represent single organisms, or groups of highly related organisms [125]. Genes can be called on the contigs, either before or after binning, to reveal the genetic potential of the organisms.

3.3.2 Approach to Validation

The flow of information through metagenomics pipelines can be thought of as liquid flow through piping with leaks (Figure 3.4). In the process of aggregating information, each software step can lose information ("leak") along the way. Flow (read accountability) can be measured at the input and exit of pipes (represented by flow gauges), allowing identification of how well output data reflects the original samples, and at which step(s) information was lost.

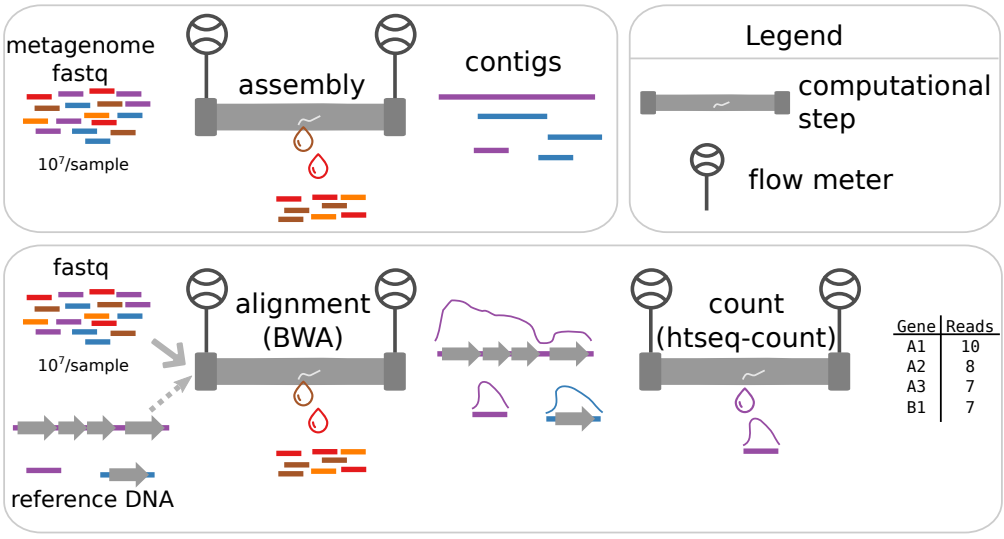


Figure 3.4: Cartoon representing the ability to deduce information loss at different steps of the metagenome/metatranscriptome work flow.

The pipe analysis framework illustrates the approach to guide process improvements: use multiple monitoring methods to determine which pipeline (computational) steps are associated with information loss, and estimate the upper-bound for improvements. These statistics also help clarify how effectively the results reflect the complexity of the original samples.

For example, assembly quality is checked by comparing the counts of all reads in each metagenome `.fastq` file to the number of reads that map back to the contigs identified (e.g. Figure B.2):

$$\text{Assembly quality:} \quad \frac{\texttt{.fastq} \text{ reads that map to contigs}}{\text{total reads in } \texttt{.fastq}}$$

Similarly, the ability of bins to represent the original metagenome `.fastq` files can be assessed by checking how many reads map to contigs that make it into bins (e.g. Figures 3.7, 3.12)::

$$\text{Binning efficacy:} \quad \frac{\texttt{.fastq} \text{ reads that map to contigs in bins}}{\text{total reads in } \texttt{.fastq}}$$

The upper-bound potential of binning can be assessed by comparing the ratio of <u>all</u> reads that map to contigs long enough to bin (e.g. Figures 3.7):

$$\text{Binning efficacy (upper bound):} \quad \frac{\texttt{.fastq} \text{ reads that map to contigs longer than binning length cutoff}}{\text{total reads in } \texttt{.fastq}}$$

The RNA-seq results can be measured by observing the fraction of reads that are mapped to genes (e.g. Figure 3.10):

$$\text{RNA-seq Information Retention:} \quad \frac{\texttt{.fastq} \text{ reads that map to genes}}{\text{total reads in } \texttt{.fastq}}$$

Differences across samples and trends across series and experimental conditions highlight

biological drivers of variation.

3.3.3 Importance of Validation

For papers where the goal is to describe one or several novel organism(s) with high confidence, the loss of information (reads) along the way may not be a significant concern. Often papers in this category do not clearly report estimates of the abundance of these organisms in the natural community. Other papers try to present a more holistic description of the community (e.g. [126]) by describing the diversity of organisms identified via binning.

In our case, it is crucial to quantify the fraction of the original sample that is described by the output of the computational analysis, given that we aim to describe the community as a whole. Low fractions suggest an incomplete portrait of the true microbial community. In most literature, statistics are usually not stated clearly in publications, allowing readers to over-estimate the importance of the described taxa in the ecosystem. This thesis advocates for publishing metrics such as those described in Section 3.3.2 in all future metagenomic studies.

Metagenomics/Metatranscriptomics Tool Selection & Challenges

Each computational step can be executed by several or sometimes dozens of competing software packages [127, 128]. Each tool can report basic statistics about efficacy based on inputs and outputs, but it it up to the researcher with larger-scope understanding of the project goals to write scripts to critically assess efficacy of the tools in aggregate. Results from one software package can give you better insight about how to evaluate the results of other packages. The results in this thesis are the product of such diligent iteration.

Another challenge of projects of this scope is the size of the data. As noted above, more data can lead to greater potential for extracting information from uncertain data, however, many tools are not designed to work on sequencing datasets as large as these. While some

tools fail quickly and with informative messages, other simply fail by running indefinitely without the tool either completing its task or aborting. When possible, the data were broken into subsets, processed separately, and re-joined upon completion. Other times (e.g. metagenome assembly), higher memory AWS instances were rented to support algorithm requirements.

This thesis addressed these challenges by testing multiple tools when appropriate, developing custom metrics to assess trade-offs between competing packages, and using faceted projections of the complex data to aid decision making. Our results in this regard are described below.

Assembly

Analysis of metagenomic and metatranscriptomic data usually begins with assembling metagenome reads into contigs (Fig 3.3), which are contiguous stretches of genomic sequence in which the order of bases is known to a high confidence level. The metagenomes can be aligned to these contigs to infer the abundance of organisms. The transcriptomes can be aligned to these same contigs to infer what genes are expressed. In principle, metagenomes and metatranscriptomes can be aligned to isolate genomes rather than contigs, however, in practice meta-omics experiments are usually used when the isolates available are known to be incomplete representatives of the microbial community under study. Even if a similar organism has been isolated, the investigation may be probing the diversity of similar organisms, or the influence of less abundant microbes on the function of the community. This is the case with our study.

The challenge of assembly scales with the complexity of the microbial community. High species diversity and strain-level heterogeneity add challenge to assembly [125, 128]. Communities with a small number of well-delineated species tend to assemble better, yielding shorter numbers of long contigs. In contrast, the high diversity of complex communities can lead to fragmented assemblies that have fewer long contigs and more

short contigs [125]. These shorter contigs are more challenging to analyze and can lead to an incomplete portrait of a community. Short contigs are more difficult to call genes on, and are also more difficult to bin because measures like GC and differential abundance are noisier for short contigs than long ones [127]. The most common measures of assembly efficacy are characteristics of the contigs themselves. How many are there? What is the distribution of contig lengths (e.g. N50)? This thesis advocates for additional measures that are rarely reported: the fraction of the total (unfiltered/untrimmed) original reads that map to the assembled contigs, and genome bins. These statistics provide readers with understanding of the abundance of the organism in the natural community.

3.3.4 Elviz Assemblies

The sequencing for many metagenomics projects, including this one, are done by the Joint Genome Institute (JGI). JGI provides some bioinformatics processing in addition to providing un-processed reads. The software they run assembles contigs for each sample individually, predicts genes, and calls taxonomy for contigs [129]. They also report the number of reads and average coverage for each contig. These measures are a great starting point for answering "who's there" in terms of taxa abundance. They do not include measures of how many reads failed to assemble. This chapter measured the fraction of reads that assembled across samples to address this concern (Figure B.2).

The contigs JGI provides are assembled for each sample, individually. For this project, that means there are 88 separate assemblies. Thus the contigs are not shared across samples, and comparison across samples is hindered. For example, it is difficult to infer whether a particular gene is differentially expressed across two samples if the contigs used as reference DNA are not the same across samples. More powerful analyses can be done when the DNA from all metagenomic samples are assembled together. This makes it possible to map DNA and RNA reads back to this shared DNA and thus to tabulate gene expression across samples. Such an assembly also leads to more powerful binning

(discussed below) by providing extra features to cluster with, specifically, the differential abundances of each contig across the samples [130]. Since the number of reads originating from a particular organism should be proportional to its abundance, the DNA coverage for a set of contigs originating from an organism should be similar in a particular sample, and correlated across samples. The per-sample nature of the JGI contigs guided decisions about how the data were used, and motivated generation of co-assembled contigs to represent the entirety of the dataset as discussed below.

Binning

Desire for gaining understanding at the organism level often leads investigators to approximate genomes as best they can from a set of metagenomic contigs. The phrase ”genome bin”, often called simply ”bin”, is used to describe a collection of contigs used to approximate the genome of a single organism. ”Bin” implies the collection is expected to be somewhat incomplete and potentially contaminated by DNA of other strains or species. Bins allow inference of which metabolic pathways are available, and can be used as a reference to map RNA for quantification of how these genes are expressed. Higher confidence claims would be possible upon isolation of the organism represented by the genome bin, but as noted above, laboratory isolates are often not obtainable either due to obligate partnerships with other organisms, or a variety of intolerances to laboratory conditions [119]. Binning can be done before or after annotating genes on the contigs, as very few methods use gene calls in the binning process.

The most common features used to bin contigs include GC content, tetranucleotide frequency, other oligonucleotide frequencies, and differential abundance across samples [127]. Most aim to be effective without needing reference genomes for alignment.

The ability to obtain complete and un-contaminated genome bins varies greatly from project to project. As noted above, more rarefied samples with low levels of strain heterogeneity are more amenable to binning [125, 128], as the contigs are more likely

to be long and few in number. Longer contigs have reduced noise for measurements of GC ratio, tetranucleotide frequency, and differential abundance [127]. This in turn leads to less ambiguity when sorting contigs into respective bins. Notably, most binning tools have been demonstrated on simple communities such as the premature infant gut [131]. Despite the added complexity of these samples relative to such classic data sets, draft bins were produced and steps to improve them are outlined.

The quality of bins can be determined using metrics of completeness, contamination, and efficacy of representing the sample. Completeness and contamination checks can be automated by use of CheckM [132], which uses the presence of marker genes in each bin as the basis for its approximations. The marker genes used by CheckM are different based on the approximate taxonomy CheckM infers for the bins. While this is certainly more effective than one set of genes used for all microbes, the reference sets available may not be good fits for the particular microbes under study. Furthermore, the choice of reference gene sets might be poor for messy bins that contain contigs from wildly different taxa. Some papers determine completeness and contamination by manual inspection using domain-specific marker gene sets. For papers published before CheckM, this manual check was the best option. Now CheckM is used in the majority of papers. It even allows the possibility of adding custom marker gene sets for CheckM to use.

The binning tools used herein combine sequence composition and coverage across multiple samples as inputs to machine learning algorithms that automatically cluster contigs into genomes. Several were tried, given that differences in their algorithms and output characteristics are seldomly and vaguely reported [127]. Metabat [133] uses k-medoid clustering of a distance measure composed of tetranucleotide frequency and abundance. MyCC [134] uses affinity propagation to cluster contigs, and uses the presence of marker genes to correct clusters. CONCOCT [135] uses a Gaussian mixture model fit with a variational Bayesian approximation, on features thinned out via principle components analysis (PCA). Almost every high-quality bin published in high-profile papers has used tools such as these, but the resulting bins are subsequently refined by

manual curation, with methods only vaguely described.

Mapping RNA

Metatranscriptomes can be mapped to contigs or bins. A key variable to consider is how to handle reads that map to multiple places (multilocus mapped reads). If a read maps equally well to two different contigs in a bin, should it be randomly assigned to one locus, or thrown away? If the study is aiming for differential expression type analysis, than the latter is better. Discarding ambiguous reads is the default of most alignment and counting tools. If the study aims to describe a holistic picture of expression in a sample, it may be worth guessing the source of such reads. Special care must be taken to account for added uncertainty after including ambiguously mapped reads.

3.4 Methods

Experimental

The culturing and sample preparation was done by Maria Hernandez, a visiting scholar. Sediment stored in an aqueous DMSO solution on a previous sampling trip was thawed from -80^{o}C and distributed into 250 mL bottles containing 100mL of NMS medium [19, 136], leaving 150mL of headspace (Figure 3.2). All bottles were air-tight, with head spaces containing 25% methane. For high O_2 samples, the head space was 75% air, while for the low O_2 samples, the head space was 15% air, with the remainder being N_2. Head space was renewed every weekday and once per weekend. Bottles were shaken at 250 RPM and held at 18^{o}C. The gaseous compositions for the days between transfers were not measured.

Bottles were serially transferred for 14 weeks. Metagenomes and metatranscriptomes were obtained for all samples corresponding to weeks 4-14. Earlier weeks were omitted due to potential effects of residual DMSO, and interest in prioritizing sequencing of later, more-rarefied samples. The O_2 conditions (low/high) were switched for the last 4 bottles

in each series (see Figure 3.2).

Sequencing

This large dataset was made possible by a Joint Genome Institute sequencing award (JGI Proposal Id: 1601). JGI sequenced all samples on an Illumina platform to produce the 88 metagenomes and 88 metatranscriptomes. Ribosomal RNA was removed by JGI (Illumina Ribo-Zero rRNA Removal Kit) before sequencing the metatranscriptomes.

Elviz Analysis

Along with raw sequencing outputs (`fastq.gz` files) for each sample, Elviz data [129] were provided by JGI. These data included one assembly per metagenome, corresponding contig statistics, and transcriptome alignment statistics. In addition, they provided taxonomy for each contig, to some degree of certainty: some are only labeled to the kingdom level, whereas others are specified all the way to genus. The taxonomy, contig lengths, and number of reads mapped to each contig were used to infer the distribution of taxa in each sample. The metric used was the fraction of reads assigned to contigs with the specified taxonomy (code at https://github.com/JanetMatsen/elvizAnalysis/blob/master/abundance_utils.py). Read counts were not normalized by contig length prior to summation (see Section 3.5.2). The code checks that the taxas' abundances summed to 1 for each sample.

Tools were written to aggregate (a) based on taxonomy level, or (b) everything below a taxonomy level. Contigs with taxonomy not described to the specified level were grouped into "unknown & other". This framework was also used to summarize abundances at mixed taxonomy level (e.g. phylum Bacteroidetes, order Burkholderiales, family *Methylococcales*, and family *Methylophilales*), and plotted (code: https://github.com/JanetMatsen/elvizAnalysis/blob/master/abundance_plot_utils.py).

Statistics for the 88 samples were merged with information about the O_2, replicate,

and week values corresponding to each sample using Python and Pandas. Plotting was done with Matplotlib, Pandas, and Seaborn.

Computational resources

This computationally-intensive research was supported by an AWS research grant, which included $27,000 of compute resources and expert consulting on how to best utilize the vast services offered. Use of AWS has several advantages for projects of this type. It allows per-hour rental of machines with different specifications, which the user can choose based on algorithm needs. Furthermore, the ability to share machine images with other researchers greatly facilitates the potential reproducibility of research. Upon project completion, an Amazon Machine Image (AMI) encapsulating the entire data analysis pipeline will be provided so others can re-run our analyses.

The machine ("instance") type used in all methods to follow was an AWS EC2 `c4.8xlarge` instance, which has 36 cores and 60GB memory. This was sufficient for all work, other than the memory-intensive assembly step, as noted below. Instances could be turned on and off as needed, reducing compute costs. This is in contrast to sharing a similar machine with an entire research lab, which would have slowed down analysis and caused conflicts when certain computations used all of the compute resources, sometimes causing computer failure.

Each instance obtained data and wrote results to a shared file system. The 9 TB of data, scripts, and results were stored using the AWS Elastic File System (EFS). EFS is currently the most versatile and highest performance (most expensive) file storage available within AWS, which allows users to have multiple instances reading/writing to the same directory. This shared nature of the files allows the user to spin up a second computer instance as needed, and do read/write operations on the exact same set of files.

Parallelization of compute tasks was done by splitting jobs across a single large computer, or across separate EC2 instances. When tasks were parallelized across one

powerful computer, Gnu parallel or Python `pool.map` was used. When parallelizing across more, often weaker, computers, an AMI was made that mounts the EFS-stored data. These instances were usually started by hand through the console, though proof of principle studies using the Amazon Command Line Interface (CLI), and an Autoscaling/SQS service pair were explored to set the foundation of future work (see repository https://github.com/JanetMatsen/AWS_parallel_Cowsay) for the templates developed.

Map to isolate genomes

As a preliminary study, the metatranscriptomes were mapped to 55 genomes (Table 3.1) from strains isolated from Lake Washington by the Lidstrom Lab. The sequences were downloaded from NCBI, and concatenated together before mapping. Mapping to the multi-fasta rather than the individual genomes prevents double-counting of reads that map well to multiple loci. BWA-MEM [32] with default settings was used to map each transcriptome to the multi-fasta, and htseq-count [137] was used to tabulate expression estimates. Samtools [138] was used to evaluate mappings produced by BWA-MEM. Note: the default of BWA-MEM is to flag reads that map equally well to two or more loci as having quality score 0; htseq-count by default omits these counts from the table of reads per gene (discussed later).

Table 3.1: The 55 isolate genomes used.

Ancylobacter sp. FA202
Arthrobacter sp. 31Y
Arthrobacter sp. 35W
Arthrobacter sp. MA-N2
Bacillus sp. 37MA
Bacillus sp. 72

Bosea sp. 117
Flavobacterium sp. 83
Flavobacterium sp. Fl
Hoeflea sp. 108
Hyphomicrobium sp. 802
Hyphomicrobium sp. 99
Janthinobacterium sp. RA13
Methylobacter tundripaludum 21/22
Methylobacter tundripaludum 31/32
Methylobacterium sp. 10
Methylobacterium sp. 77
Methylobacterium sp. 88A
Methylocystis sp. LW5
Methylomonas sp. 11b
Methylomonas sp. LW13
Methylomonas sp. MK1
Methylophilaceae bacterium 11
Methylophilus sp. #1
Methylophilus sp. 42
Methylophilus sp. 5
Methylophilus sp. Q8
Methylopila sp. 73B
Methylopila sp. M107
Methylosarcina lacus LW14
Methylosinus sp. LW3
Methylosinus sp. LW4
Methylosinus sp. PW1
Methylotenera mobilis #13
Methylotenera mobilis JLW8

Methylotenera sp. 1P/1
Methylotenera sp. 73s
Methylotenera sp. G11
Methylotenera sp. L2L1
Methylotenera sp. N17
Methylotenera versatilis #7
Methylotenera versatilis 301,
Methylotenera versatilis 79
Methyloversatilis sp. FAM1
Methyloversatilis sp. RZ18-153
Methyloversatilis universalis FAM5
Methyloversatilis universalis Fam500
Methylovorus glucosetrophus SIP3-4
Mycobacterium sp. 141
Mycobacterium sp. 155
Paracoccus sp. N5
Pseudomonas sp. 11/12A
Xanthobacter sp. 126
Xanthobacter sp. 91
Xanthobacteraceae bacterium 501b

Assembly

Dave beck co-assembled the quality trimmed metagenomes with Megahit [139]. This memory-intensive computation was done on an AWS instance with 1 terabyte of memory. Assembly of multiple samples simultaneously provides a shared set of contigs to use when mapping DNA reads or RNA-seq reads from any of the 88 samples. This in turn allows for comparison of gene expression across samples. The fraction of reads represented by the assembly, in aggregate, was measured to verify assembly efficacy.

Gene calls

Genes were called on contigs with length greater than or equal to 1.5kb using Prokka [140] (scripts: Python script that calls Prokka, script that extracts contigs by size). This cutoff was selected because shorter contigs are not recommended for inclusion by most binning tools documentation. Additionally, shorter contigs are less likely to have genes identified on them, as genes are more likely to be incomplete due to a higher density of contig edges. Prokka did not tolerate input data of the scale of this dataset, so the input `.fasta` was broken into 5 files of approximately equal size (scripts). These were processed in parallel, using the `pool.map` function of Python. A script was written to merge the resulting annotated "general feature format" files (`.gff` files) into one representing all contigs $\geq$ 1.5kb.

Checking the ability of contigs to represent each metagenome

The loss of information resulting from omitting short contigs was assessed by counting the number of reads that map to contigs $\geq$ 1.5kb, and comparing that to the total number in the raw metagenome `fastq.gz` files (`.fastq` read counting script, analysis). This represents an upper bound in information content for the assembly step. Again, this used BWA-MEM [32], and a script written to count reads in `.fastq` files. Pandas (Python) was used to merge results together.

3.4.1 RNA-seq: mapping to contigs $\geq$ 1.5 kb

Like all other read mapping steps, transcriptomes were mapped with BWA-MEM (default settings), and per-gene results were summarized with htseq-count. Contigs $\geq$ 1.5kb were used as reference DNA. Metatranscriptomes were mapped to these contigs all at once, rather than to each bin individually in order to reduce over-counting. This does result in reads that map identically well to two places being thrown out, as mentioned earlier. Results from each sample were merged using a 60GB RAM AWS instance, rather than

SQL. The fraction of transcriptome reads was compared to the total number in the input `.fastq` files as a metric of information loss. See scripts and notebook.

Binning

Binning of contigs into genome-like bins was explored using three different packages. MetaBAT was used with default settings on contigs $\geq$ 1.5kb (script). As promised by the documentation, it scaled well to this large dataset.

MyCC was tested, but did not scale well to this large dataset. The large memory requirement of the underlying affinity propagation algorithm led to failures when default settings were used. At one time, MyCC wrote so many small files that the machine crashed due to running out of inodes. The authors suggested some alterations to the settings to reduce the computational requirements: `56mer, lt 0.4, st 50`. Use of their settings worked on the set of contigs longer than 2.5kb, but not the larger set of contigs longer than 1.5kb.

Concoct was tried initially, but bins were never obtained. This tool reports multiple potential clusters per conitg. The approach for resolving that conflicting information was not sorted out in the time frame of this project.

Average Nucleotide Identity

Average nucleotide identity (ANI) was used to compare bins. The underlying tool, provided by JGI [141] [1], calculates two-way global ANI measures and reports the fraction of each genome that aligned. The authors suggest removal of ribosomal RNA genes, which can have high ANI even for divergent organisms. They do not, however, provide options or suggest a tool for this step. Given that the ANI calculations intended only to give a rough idea of similarity between bins or bins and isolates, this step was not performed. The tool was used to assessed ANI between all pairs of isolate genomes and

[1] https://ani.jgi-psf.org/html/anicalculator.php

MetaBAT bins (script).

Additional bin characterization

Completeness and contamination were checked with CheckM [132] (scripts). CheckM uses marker genes from what it determines to be an evolutionarily similar microbe. The tool was benchmarked on isolates, to ensure it performed well on the key methylotrophic taxa under study. Custom marker genes specific to methylotrophs were not added, however, this is suggested as a possible future direction to assess bins.

Taxonomy was approximated using PhyloPhlan [142], and by ANI similarity with isolate genomes (scripts).

Code

The code described was developed with Git version control, and the full history is available on GitHub:

- **ElvizAnalysis**: `https://github.com/JanetMatsen/elvizAnalysis`
- **meta4** (assembly, binning, etc.): `https://github.com/BeckResearchLab/meta4`

3.5 Results and Discussion

3.5.1 Raw Reads

The cumulative number of raw and un-filtered reads is shown in Table 3.2. This collection of reads was used for all inferences described in this chapter. Also see Figure B.1 for plots showing specific values for each sample.

Table 3.2: Number of un-filtered reads: sample average and total across 88 samples.

	mean # reads per sample	total # reads
metagenomes	3.33e+07 ± 7.34e+06	2.93e+09
metatranscriptomes	4.24e+07 ± 6.22e+06	3.73e+09

3.5.2 Elviz Analysis

The best measure of which taxa dominate the samples comes from the Elviz [129] data. As described in the Methods section, the abundance measure was the sum of read counts across contigs labeled with the taxonomy of interest. Traditional metrics like RPKM [34] and TPM [143] are excellent for comparing coverage or expression across features or samples, but could not be used due to the need to sum values across contigs. To illustrate why lengths of contigs should not factor into the abundance measures, consider this example: a contig with length 10kb should contribute nearly identically to an equally expressed second 10kb stretch of DNA that for technical reasons was assembled into two 5kb stretches rather than one 10kb contig. Any scheme that controlled for contig length or was weighted by the statistic of each contig would skew the aggregated number down for metagenomic datasets such as this with many short, low-coverage contigs. The fact that these assemblies are dominated by short contigs means that RPKM and TPM would bias the aggregate statistics toward the statistics of the shorter and less reliable contigs. Similarly, median coverage was not used or the unreliable coverage numbers associated with short contigs would dominate the measure.

Figure 3.5 shows abundances of the four major taxonomic groups, Methylococcales, Methylophilales, Bacteroidetes, and Burkholderiales, according to the Elviz taxonomy calls for each contig. (Recall introductory Figure 3.1 for a hierarchical visualization of these taxonomies.) Their dominance is consistent with previous studies [121, 122, 112, 123, 124]. Figure 3.6 shows the breakdown of the methanotrophic and non-

methanotrophic methylotroph taxa in Figure 3.5. Notably, the path to rarefaction is not particularly consistent across samples. Like previous studies, the dominant methanotrophic order/genus was Methylococcales/*Methylococcaceae* and the dominant non-methanotrophic methylotrophic order/genus was Methylophilales/*Methylophilaceae*. Low O_2 was associated with a higher abundance of methanotrophs, and a lower proportion of methylotrophs than the high-O_2 samples. *Methylosarcina* is often found in high-O_2 samples, but rarely in low-O_2 samples. Two high O_2 samples are dominated by *Methylobacter*, but the others are not. The last four samples of each series are more similar to the previous samples in their series than to samples at the opposite O_2 tension.

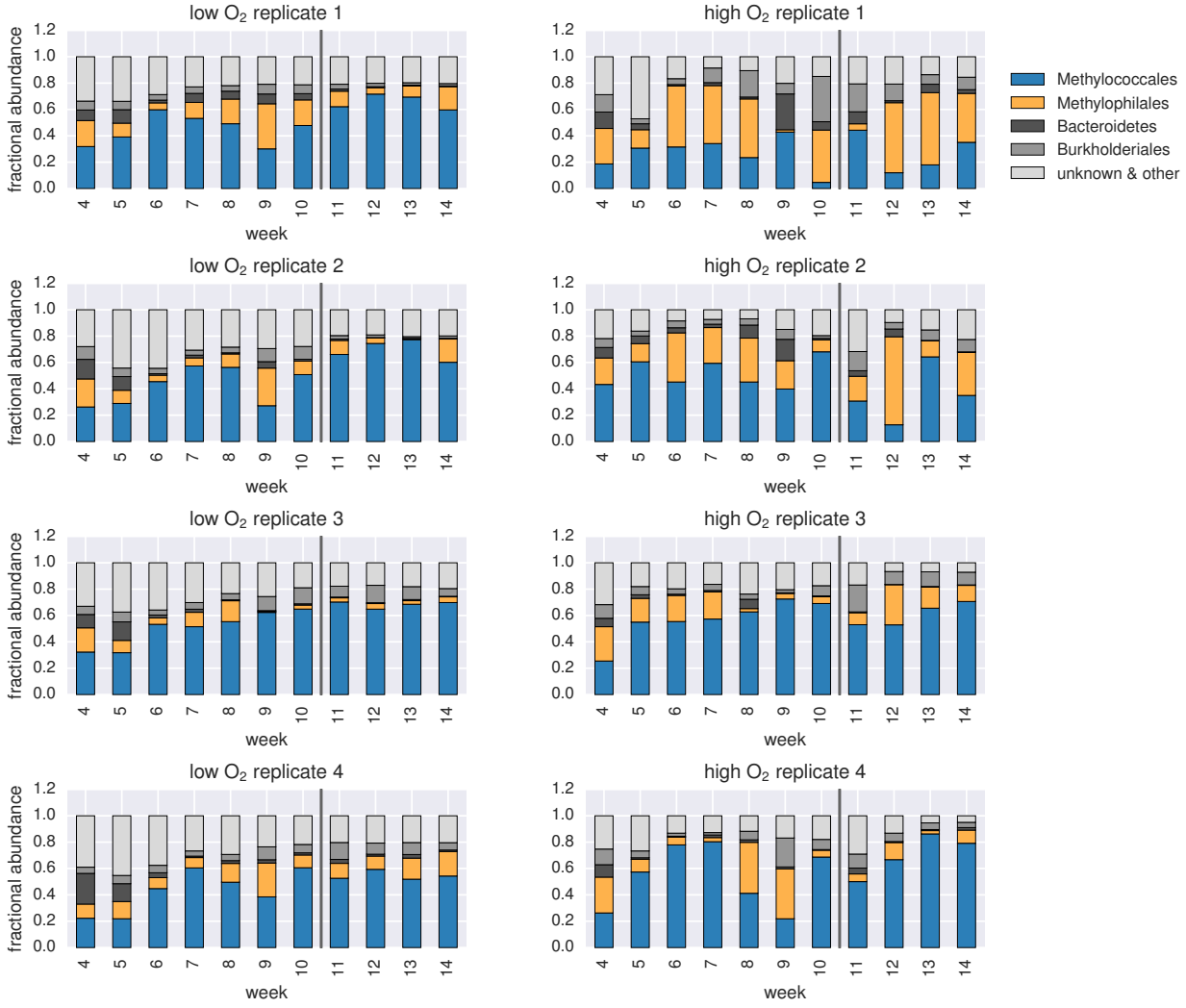


Figure 3.5: Fractional abundances of DNA from the four major taxonomic groups in Lake Washington sediment incubations: phylum Bacteroidetes, orders Methylococcales (methylotrophs), Methylophilales (non-methanotrophic methylotrophs), Burkholderiales. The vertical gray line indicates the week the O_2 conditions were reversed. These bars only reflect reads that mapped to contigs, according to the Elviz pipeline. For information about the fraction of reads not assigned to contigs, see Figure B.2.

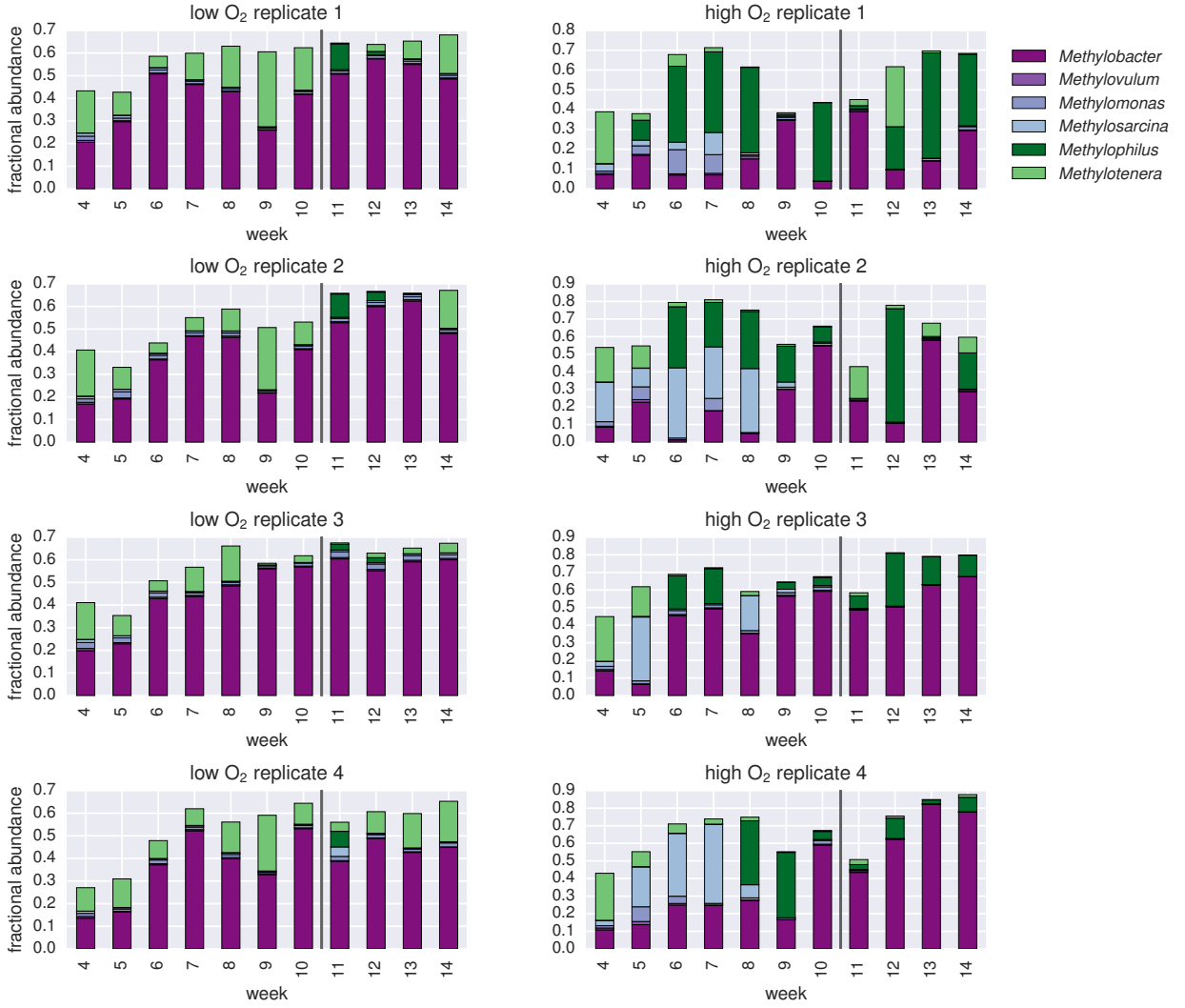


Figure 3.6: Fractional abundance of DNA from the dominant methylotrophic genera. Purple bars are methanotrophs of order Methylococcales, family *Methylococcaceae* (represents major taxa in blue bars of Figure 3.5). Green bars are non-methanotrophic methylotrophs of order Methylophilales, family *Methylophilaceae* (represents major taxa in yellow bars of Figure 3.5). See Figure 3.1 for hierarchical taxonomy reference. Vertical gray line indicates the week the O_2 conditions were reversed. Like Figure 3.5, these bars only reflect reads that mapped to contigs, according to the Elviz pipeline. For information about the fraction of reads not assigned to contigs, see Figure B.2.

Binning Potential: Elviz Contigs

The goal of this study is a comprehensive description of the samples, ideally in terms of genome bins. The potential for these Elviz contigs to paint a comprehensive picture of the microbial community was measured by summing all reads assigned to contigs long enough to bin, namely contigs $\geq$ 1.5kb. Figure 3.7 shows that while the entire set of contigs represents the samples well, the subset of contigs that are candidates for binning explain less than 20% of the reads obtained for many low O_2 samples. Another complication of binning these Elviz contigs, is that binning the contigs for each sample individually would lead to many redundant bins that need to be thinned. This process is prone to human bias, and can lead researchers to over-fit their bins to their quality measures and thus over-state the quality of their bins. Nonetheless, if a particular sample is enriched in an organism of interest, use of the contigs from that sample should be conisdered for future binning efforts.

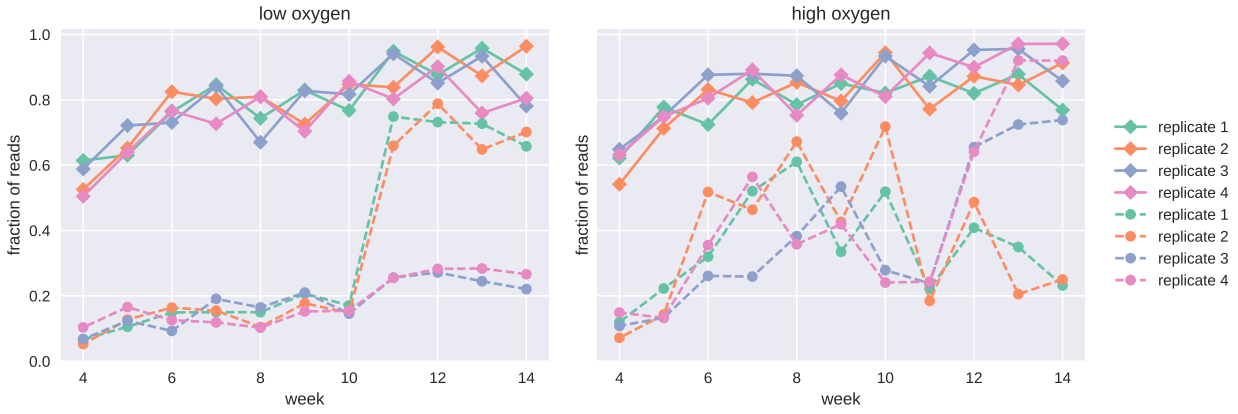


Figure 3.7: The upper-bound limit on success in binning Elviz contigs, as measured by the fraction of metagenome reads mapped to contigs. Lines are colored by replicate. Solid lines represent the fraction of reads in the raw `.fastq` file that map to contigs of any length. Dashed lines represent the fraction of reads in the raw `.fastq` file that map to contigs with length $\geq$ 1.5kb. Less than 20% of the raw fastq reads would be explained by aggressive binning of most low O_2 samples.

3.5.3 Map to Isolate Genomes

The next simplest approach for identifying which microbes are active in the samples, and what genes they express is to map the reads to the 55 isolate genomes (listed in Table 3.1). The potential efficacy of these results can be measured by quantifying the fraction of metagenomic and metatranscriptomic reads that align to these genomes (recall the pipe diagram, Figure 3.4). Figure 3.8 shows the fraction of the metatranscriptomes that can be explained using these reference genomes. The low accountability (mostly $<$15%) guarantees an incomplete portrait of the activity in these samples, so the data were not pursued further. Despite heroic Lidstrom Lab efforts to isolate strains, the set obtained thus far does not cover the complexity seen in real samples. This indicates missing taxa and/or dissimilarity between the isolated strains and the ones that dominate in nature. While this method of describing gene expression was not pursued farther, it provides a baseline fraction of RNA mapped for comparison of any subsequent methods.

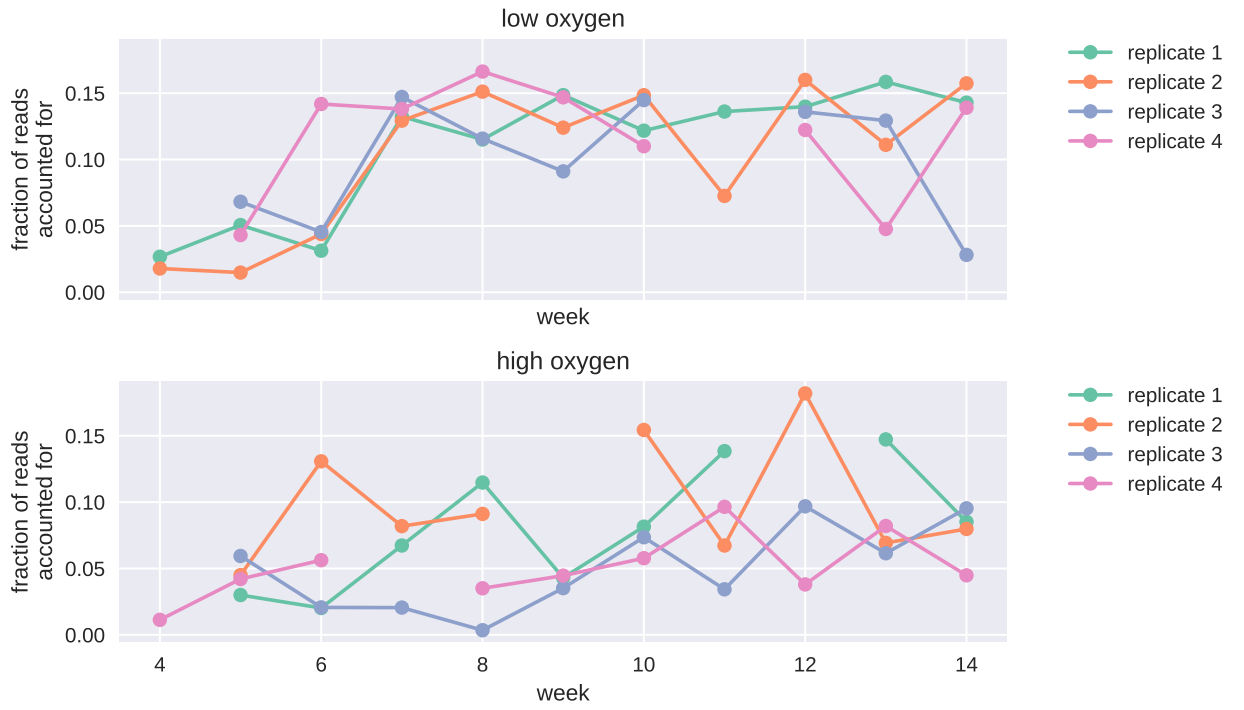


Figure 3.8: Measure of RNA-seq read accountability by isolate genomes. Each dot represents the total fraction of raw reads assigned to genes when mapped to a concatenation of 55 isolate genomes.

As noted in the Methods, all computational steps using BWA-MEM and Samtools use default handling of reads that map equally well to multiple places. This leads to omission of these multilocus mapped reads. Such handling is the best-practice approach for differential abundance analysis, which is the most common workflow for the tools used. It does, however, lead to under-estimation of gene expression levels and lower accountability of the original *.fastq* reads. Higher mapping fractions could be achieved by randomly assigning reads to one target from the set of best potential loci. Including these reads that have high mapping uncertainty makes it more difficult to be sure which genes/organisms are most influential in the samples. For example, there may be two genome bins that have very similar monooxygenase (MMO) genes [42] but substantial genomic differences. If reads map equally well to both sets of MMO genes, it is hard to

say which bin was more actively oxidizing methane in the sample.

3.5.4 Assembly

Given that the Elviz contigs and isolate genomes proved ineffective as reference DNA for comparative analysis of samples, a single set of contigs that could be used to assess all samples was developed. The metagenomes from every sample were pooled and assembled (by Dave Beck) into one set of contigs shared by all samples. This allows production of one table summarising reads mapped to different loci across all samples, and the potential to compare samples.

Assemblies which produce long contigs are desired, as longer contigs facilitate calling genes and clustering. Communities with extremely low complexity (just a few microbes co-existing) can yield long contigs with greater potential for being sorted into high-confidence genome bins. As complexity increases, reduced sequencing depth and homologous regions between organisms inhibit the formation of long contigs. The distribution of contigs obtained (Figure 3.9) does not suggest a simple community.

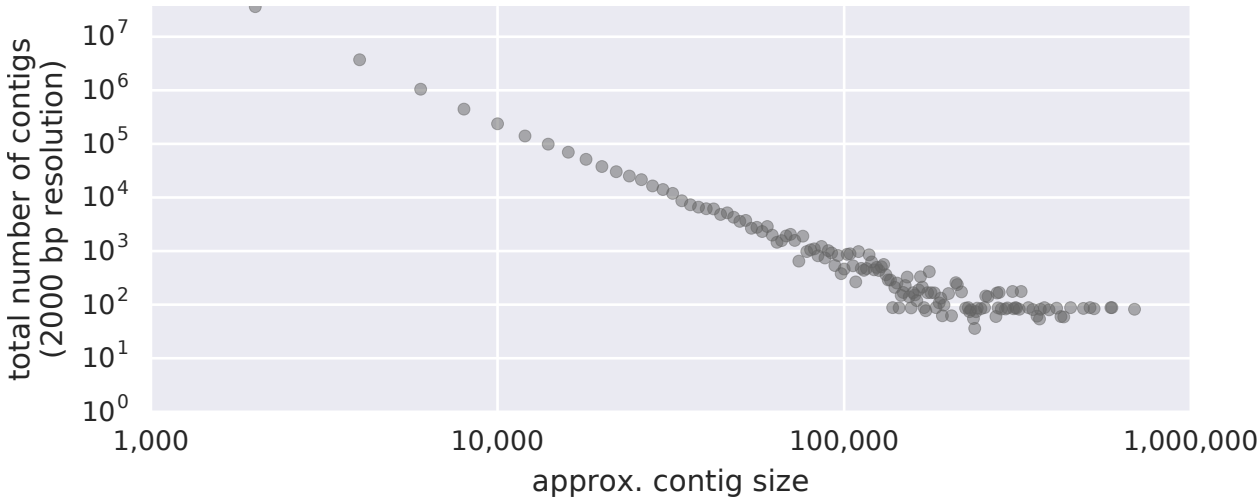


Figure 3.9: Distribution of sizes for contigs assembled by MEGAHIT (2,617,225 contigs in total).

3.5.5 *Annotation*

Genes were annotated (called) on contigs with length $\geq$ 1.5kb, in preparation for metatranscriptome studies (Table 3.3). See the supplementary attachments for the gene `.fasta` and corresponding `.gff` files.

Table 3.3: Annotation of MEGAHIT contigs.

	statistic
total # of genes	921,431
number of distinctly named genes	107,900

3.5.6 *Transcriptome Analysis*

Mappings of the 88 transcriptomes to the annotated contigs revealed which of the $\approx$0.9 million genes were strongly expressed. Table B.1 shows the names of the proteins associated with the highest total reads mapped, and how many copies of each gene contribute to each sum. See supplementary files (`.fasta`, `.gff`, and `.tsv`) for per-gene sequences, sequence information, and RNA-seq read counts by sample.

Figure 3.10 shows how well the called genes account for the transcripts seen in the raw sequence data. The green bars represent a high degree of certainty. Gray bars correspond to "no feature", which is presumably correlated to short contigs for which gene calls are less likely. Pink bars are labeled as having low alignment quality, but recall that BWA-MEM sets the quality score to zero when a read maps equally well to two different spots. It may be possible to improve (shrink) the gray bars via re-assembly of portions of the data (see Section 3.6.3). Similarly, recovery of reads that map equally well to different places can be brought out of the pink bars by adjusting the settings of htseq-count such that low quality reads are used (see Section 3.6.2). While this alteration would help with describing the big picture of which genes are expressed, it reduces certainty of read mappings and is

not appropriate for the analyses in the next chapter (Chapter 4).

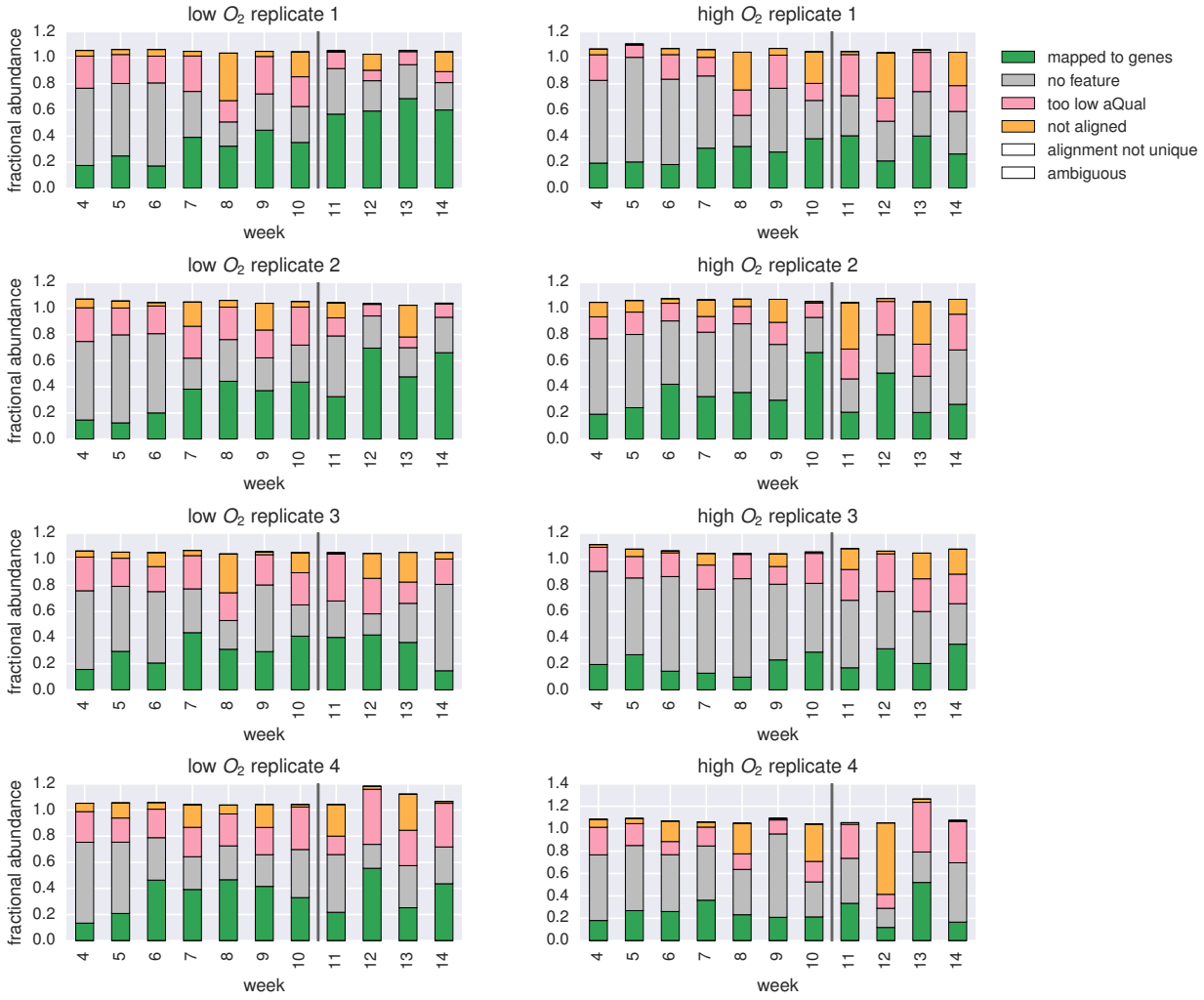


Figure 3.10: Efficacy of mapping RNA-seq to MEGAHIT contigs. Bar heights are fractions of reads with respect to the original `.fastq` files. Bar height totals are noticeably greater than one for samples with a larger fraction of reads that map equally well to multiple loci. Such multilocus mapped reads are included in the pink bar due to being assigned a quality score of zero by BWA-MEM.

Highest Expressed Proteins

The top genes include many methanotrophy genes, especially those encoding subunits of methane monooxygenase and methanol dehydrogenase (Table B.1). Other highly expressed methylotrophy genes encode products of the RuMP assimilatory pathway and formaldehyde oxidation, including 3-hexulose-6-phosphate synthase, transketolase, formaldehyde-activating enzyme, and 3-hexulose-6-phosphate isomerase. The presence of key methanotrophy and methylotrophy genes agrees with the Elviz portrait of methanotroph/methylotroph dominance in every sample. Other highly expressed genes allow speculation for metabolic roles of other organisms.

Two phage proteins are near the top of the list: Capsid protein (F protein), and Microvirus H protein (pilot protein) (Figures B.3, B.4). Phage blooms could contribute to the apparent stochasticity in community rarefaction.

Note that Table B.1 does not include normalization for gene length or sample sequencing depth. Before dividing by gene length, consistency of gene lengths for loci of the same gene product should be checked. It is possible that genes encoding product of the same name could be of varying length, depending on how Prokka handles gene fragments.

Expression of Genes by Locus

Which copies are expressed, and under which conditions? Breaking down the expression of these genes by locus and sample reveals which copies dominate. Many of the patterns appear to be O_2-dependent, and/or perturbed by the O_2 condition switch for the last four samples. For example, the patterns for which *pmoA* locus is expressed in each sample appears to be correlated with the O_2 tension in each sample (Figure 3.11). Exploration of these patterns and analysis of the genes taxonomy can shed light onto how O_2 influences community composition.

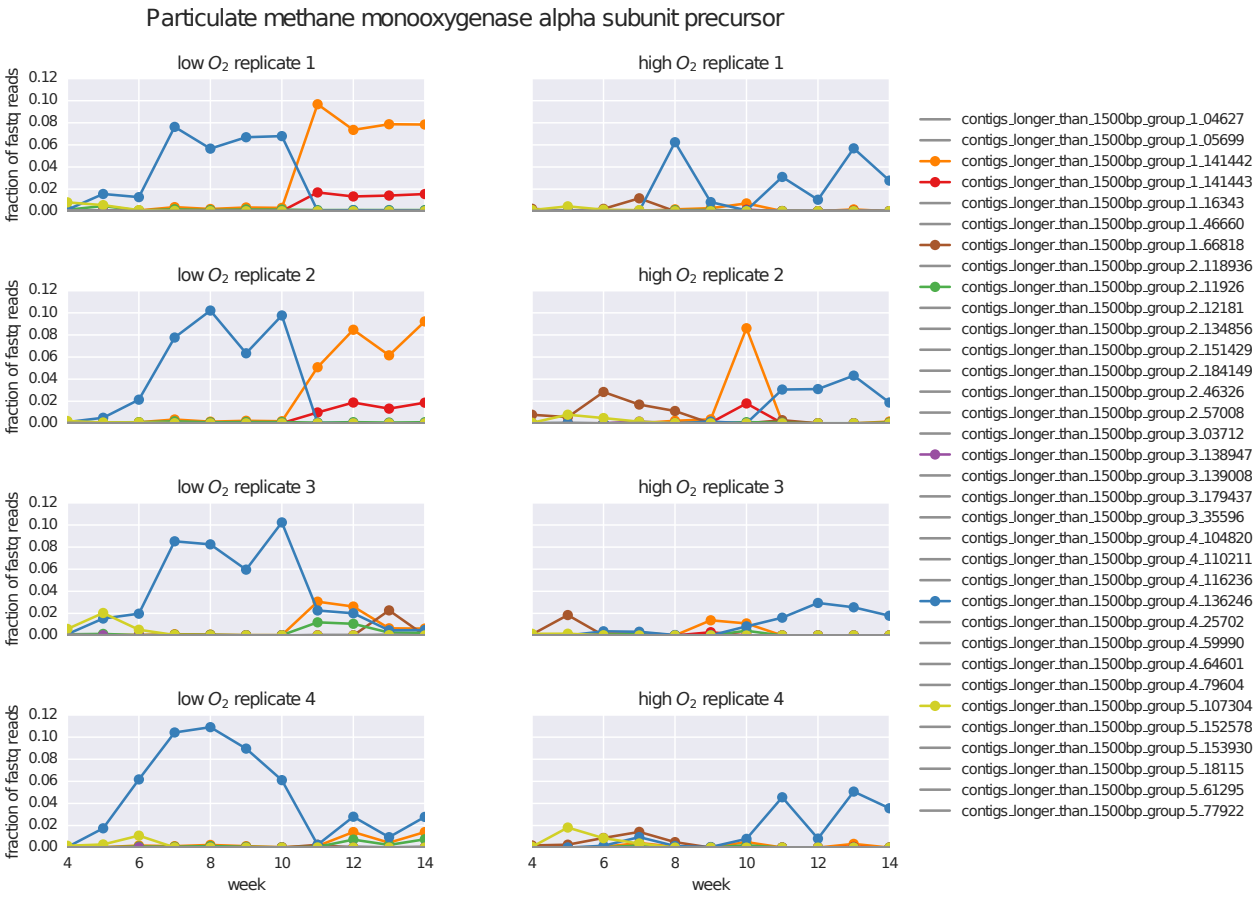


Figure 3.11: Methane monoxygenase alpha subunit locus expression profiles by sample. The legend shows many genes annotated as such, but only the highest expressed loci were plotted with color. The gray series have nearly zero expression and thus do not rise above the x-axis.

3.5.7 Binning of Co-assembled Contigs

Contigs of length $\geq$ 1.5kb were binned by MetaBAT and MyCC, two leading binning tools (Table 3.4). As mentioned, MyCC does not scale well to large data, so the full dataset including contigs as short as 1.5kb did not complete, even with a high performance AWS instance.

Table 3.4: Binning results: MetaBAT & MyCC.

	# bins	min contig size allowed
MetaBAT	330	1.5kb
MyCC	109	2.5kb

Figure 3.12 shows the fraction of the metagenomes included in the bins, as a measure of how effectively the bins describe the composition of organisms. The two sets of lines for the MetaBAT plots measure how well the bins represent the original `.fastq` files in total (dashed lines), and how well they did using only contigs longer than 2.5kb (to better compare with the MyCC results).

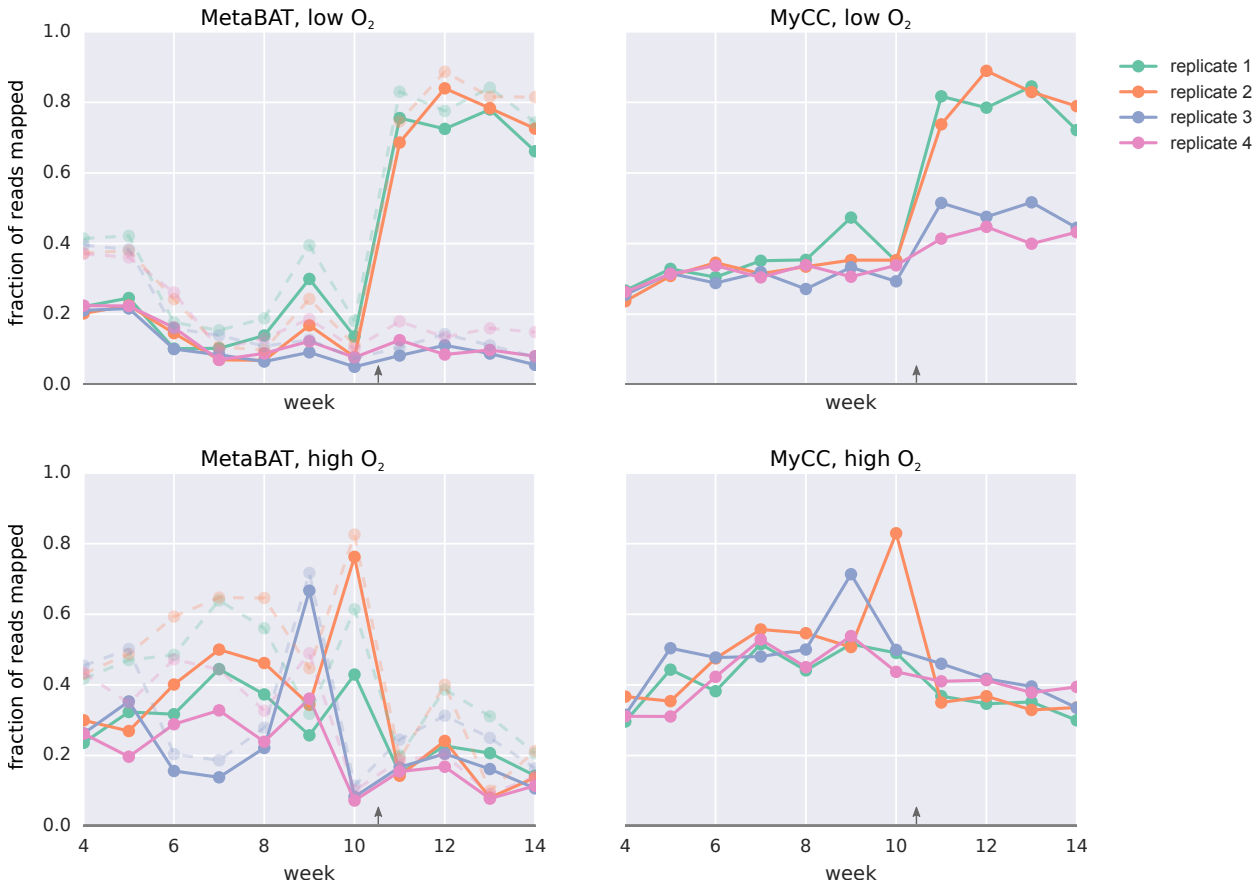


Figure 3.12: Fraction of reads in bins generated via MetaBAT (min contig size = 1.5kb, default settings) and MyCC (min contig size = 2.5kb, settings adjusted for memory issues). Points represent the fraction of reads that are associated with binned contigs for each sample. The solid MetaBAT lines represent the fraction of reads associated with contigs of length $\geq$ 2.5kb, as an appropriate comparison to the MyCC run. The dashed MetaBAT lines represent the full fraction of reads in the actual MetaBAT bins, which include contigs as short as 1.5kb. The arrows after week 10 in each plot indicate that all subsequent samples were treated with the opposite O_2 tension as the label indicates.

MetaBAT with default settings produced bins with low sample accountability, whereas the MyCC bins represented a high fraction of the raw sample reads. This represents a trade-off between precision and recall, a balance inherent to binning regardless of the tool used.

One striking trend in the MetaBAT results is the difference in read accountability between replicates 1/2 and 3/4 at the low O_2 state. The large difference is associated with the proportion of reads associated with long contigs (Supplemental Figure B.5). The poorly represented samples are biased more toward short contigs. While noting that the last 4 data points in each series have the opposite O_2 condition as the series label, see that the high O_2 samples in all four subplots are better reflected by MetaBAT bins. The source of this conundrum has not yet been resolved, but could be due to a smear of organisms that were not amenable to assembly. The Elviz data points towards the genus *Methylobacter*, given its dominance in samples at the low-O_2 state. BLAST or other computational tools could be used to identify the taxonomy of contigs that are not binning well.

Figure 3.13 also shows that MyCC is more effective at representing the raw sequencing reads. Upon further investigation, it was revealed that MyCC binned >99% of all contigs supplied, including the shorter ($\approx$ 2.5kb) contigs which carry greater uncertainty. Aggressive binning is associated with low specificity, meaning inclusion of contigs that do not belong in bins.

Figure 3.13: MyCC and MetaBAT: fraction of contigs binned by length. With the settings used, MyCC binned $>$ 99% of contigs at all lengths. The one MyCC dot near y=0.6 had a $> 99\%$ binning rate as well, but is lower because the point includes contigs that were ineligible for binning (lengths below the 2.5kb cutoff).

Figure 3.14 reveals the consequence of the aggressive MyCC binning: high contamination scores according to CheckM. To test the predictive accuracy of CheckM on methylotrophic genome bins, a positive control test was used: CheckM was applied to the 55 genomes used in the isolate studies. While the marker lineages used were often surprising (e.g. Rhizobiales, Actinomycetales), the completeness and contamination statistics CheckM returned for these positive controls were nearly perfect (Table B.2). Thus future binning efforts should aim for nearly 100% completeness and less than a few percent contamination, even when CheckM uses the default (non-methylotrophic) marker lineages. CheckM determines strain heterogeneity as "the number of multi-copy marker which exceed a specified amino acid identity threshold". High measures of strain heterogeneity may be tolerated, as some of these isolates were estimated as having 100% strain heterogeneity, indicating the amino acid sequences are more divergent that CheckM expected (Table B.2).

Given the small fraction of the metagenome reads explained by MetaBAT (Figure 3.14), and high contamination scores of the MyCC bins (Figure 3.14), a new binning approach is required. See Section 3.6 below for suggested next steps in the binning workflow.

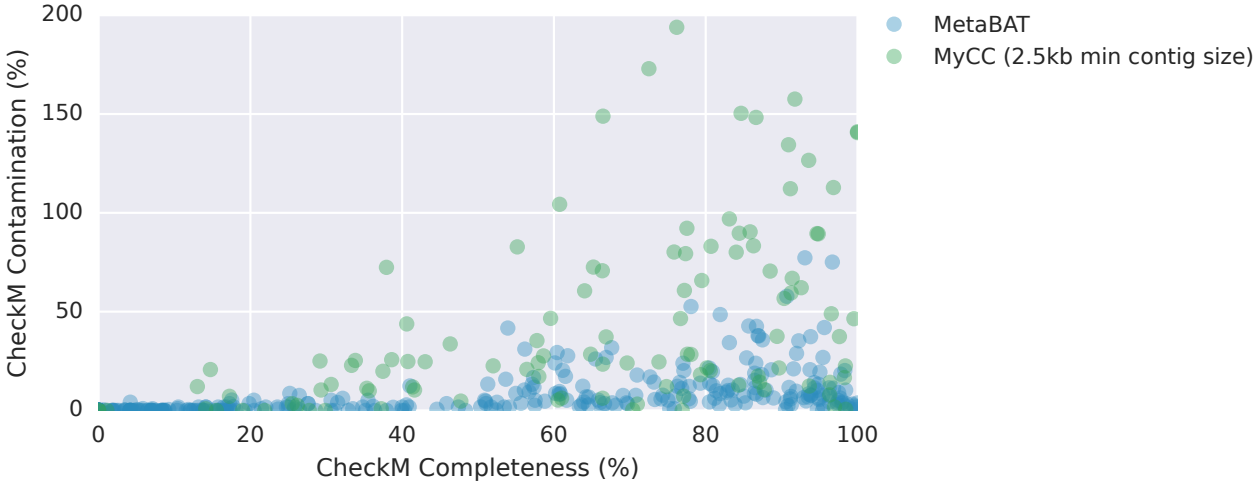


Figure 3.14: Completeness and contamination predictions for MetaBAT and MyCC bins, according to CheckM with default settings and marker gene sets.

3.5.8 Bin Taxonomy

Once final bins are in hand, taxonomic labels need to be assigned. This can be done by aggregating information from several sources, including the CheckM marker gene set (coarse label), PhyloPhlAn results (maximally specific), and average nucleotide identity (ANI, below).

PhyloPhlAn was applied to the MetaBAT bins, producing taxonomic labels that did not agree particularly well with the CheckM labels. This adds motivation for revisiting binning before describing the broad community composition in terms of genome bins.

3.5.9 Average Nucleotide Identity (ANI)

Average nucleotide identity (ANI) can also contribute evidence for taxonomic assignments, especially in cases such as this project when there are many environmental isolates from the exact ecosystem to use as references. The JGI ANI tool was used on all pairs of MetaBAT bins and isolate genomes, producing a large matrix of ANI values, a subset of which is shown in Figure 3.15.

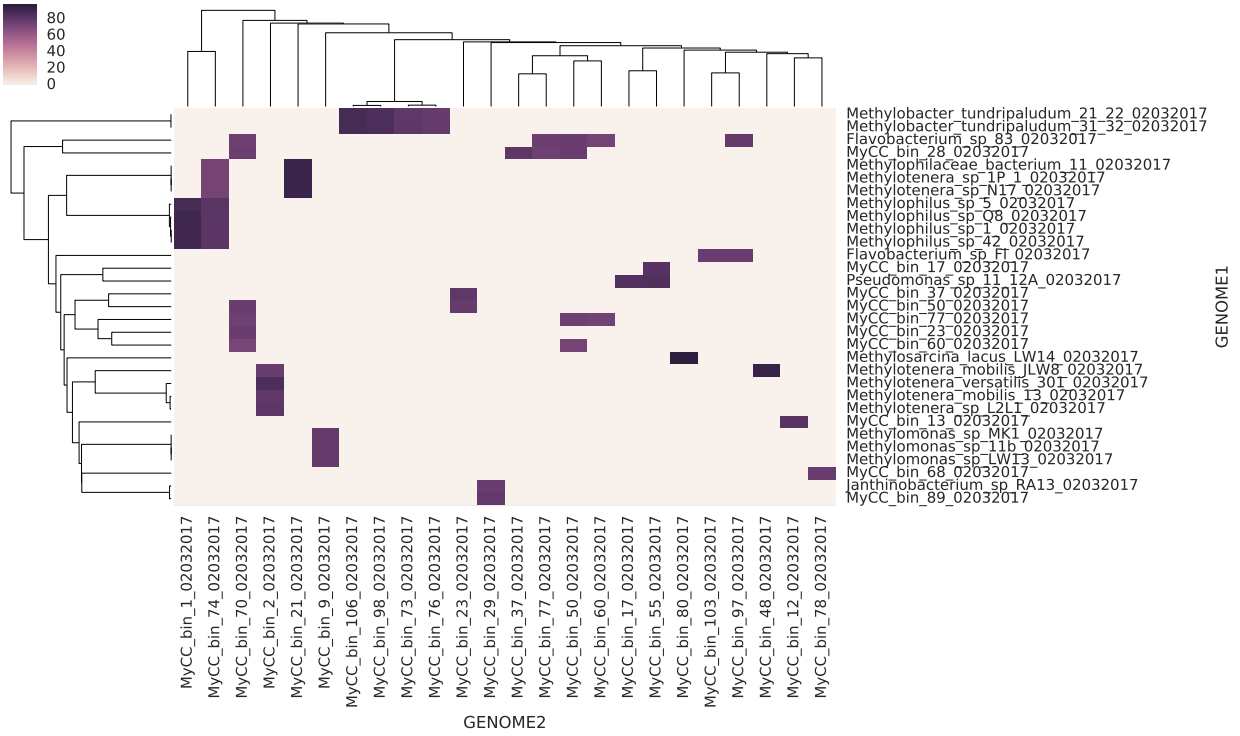


Figure 3.15: Demo use of ANI to infer taxonomy of genome bins. ANIs between all pairs of bins, and all isolate/bin pairs was computed, tabulated, and plotted to identify taxonomy of genome bins. This strategy, in conjunction with taxonomy calls by PhyloPhlAn [142] are recommended for taxonomic assignments when the final bins are identified. This clustermap was produced from raw ANI statistics via Seaborn (https://github.com/mwaskom/seaborn).

3.6 Future Directions

This analysis aimed to describe a holistic picture of the microbes in the communities, and assess how gene expression indicates metabolic pathways and interactions in the methane-oxidizing community. This chapter identified a starting point. Insights gained by iterating through pipeline steps have revealed possibilities to improving the analysis by modifying each step, as described in the following sections.

3.6.1 Goal Setting for Future Directions

This chapter highlights the metrics for measuring information loss at each step of the analysis pipeline (recall the pipeline analogy in Figure 3.4). Given that the current goal is to describe broad trends of microbial abundance and activity, setting specific goals for information retention in future work is essential. For example, substantial information in the metagenome `.fastq` files was lost along the path from raw reads to binned contigs. Before future binning work is done, the upper bound of binning success given the contigs at hand should be measured. Specifically, this upper bound (for metagenomes and metatranscriptomes) can be calculated by counting all reads that align to contigs longer than 1.5kb, as was done in Figure B.2 for the Elviz Contigs. If the fraction of reads mapping to these longer contigs is not sufficient for future publication, assembly can be revisited.

3.6.2 Investigate Multilocus Mapped Reads

If the aim is to describe the community composition and gene expression broadly, investigation of multilocus mapped reads is key. As mentioned, BWA-MEM flags reads that map equally well to two loci as having a quality score of zero. This means that any gene that appears twice in the reference contigs will have zero reads mapped to it. Similarly, two homologous genes that attract a similar set of reads will have underestimated gene expression.

This loss of multilocus mapped reads is expected to be a problem for highly conserved genes such as those encoding particulate methane monooxygenase (pMMO). Estimations of methanotroph abundances that rely on counting reads that map to various pMMO copies should have large associated uncertainty values. One possible way to handle these data is to have three statistics associated with read counts for each gene. The classically reported BWA-MEM read count will be preserved, but a second column could represent the number of additional reads that could be assigned to a gene if multilocus mapped

reads were spread evenly across their potential reference targets. To give that number context, it would be beneficial to have a third column informing how many genes were similar enough to have caused read mapping ambiguity.

3.6.3 Assembly: Next Steps

Like all steps of metagenomics, assembly in itself is an art. There is not one best assembly per dataset, as two different research goals could lead to two different ideal assemblies. For this thesis, the goal was a holistic portrait of which organisms are present in natural communities. While the assembly used throughout this chapter was appropriate for that question, there is potential to use additional assemblies to target particular organisms of interest.

New assemblies could be done with the aim of making contigs that bin a particular organism well. The enumerations of the taxa present in each sample, the gene expression profiles, and trends across samples will be explored to identify uncultivated taxa to target. For example, it is hypothesized that there is a ”smear” of similar uncultivated *Methylobacter* strains in these samples. If genes thought to be associated with different *Methylobacter* strains are found in particular samples, those could be re-assembled separately to get longer *Methylobacter* contigs and thus better bins. Another possibility is to iteratively bin, remove reads that map to those bins, and re-assemble the leftovers. These approaches may also lead to bins of important non-methylotrophs, whose metabolic role is least understood.

3.6.4 Binning: Next Steps

The analysis shown in this chapter highlights great opportunity to improve the bins and potentially allow for holistic descriptions of the communities. The easiest first step is to iteratively test the performance of MetaBAT with different binning parameters as the authors suggest. The resulting bins can be assessed, again by (a) CheckM completeness

and contamination scores, (b) fraction of metagenomes that are attracted to bins, and (c) the fraction of transcriptomes that are attracted to them. Special care can be taken to bin *Methylobacter* strains.

In addition, there are many tools that have not yet been tried. One promising tool from the Banfield Lab [144][2] just appeared on the Biology preprint server, bioRxiv. This tool uses an ensemble of binning algorithms (MetaBAT[133], MaxBin [145], CONCOCT [135] and tetraESOMs [146]). It counts the number of single-copy marker genes in candidate bins, removes good looking bins, and re-bins the remaining contigs. Given the broad efficacy of ensemble methods in machine learning, exploration of this tool is highly recommended.

3.6.5 Biology: Next Steps

The dataset prepared in this analysis allows investigation of many biological questions. Many of these questions are best answered with genome bins. For example, one topic of excitement in the field of methylotrophy pertains to two alternative methanol dehydrogenase systems. Recently it has been discovered that some organisms prefer to use lanthanum-dependent *xox* methanol dehydrogenases, rather than the classically studied *mdh* methanol dehydrogenases [147]. Though lanthanum was not intentionally supplied to the organisms in this study, it is possible that lanthanum was present in trace amounts. Identification of expression levels of these two methanol dehydrogenase alternatives in these samples will be explored soon.

Not all questions, however, require genome bins. Building taxonomic trees of the various signature genes such as methane monooxygenase can give higher resolution insight into community composition inferred in Figures 3.5, 3.6. There is also potential to explore signatures of phage in the existing contigs.

---

[2] http://biorxiv.org/content/early/2017/02/11/107789

3.7 Conclusions

In conclusion, this 88-metagenome, 88-metatranscriptome dataset was designed to allow identification of dominant taxa and gene expression in a natural methane oxidizing microbial community. The Elviz data were aggregated on taxonomic labels to illustrate the fractions of the metagenomes attributed to different taxa (see Figures 3.5, 3.6). Desire to bin contigs and compare gene expression across samples motivated co-assembly of all the metagenomes. These contigs were binned into a draft set of bins, and metrics were developed to assess whether the bins at hand adequately describe the community and its dynamics.

The contigs were annotated and the genetic content was explored. In total, 921,431 million genes were identified, encoding 107,900 different proteins. O_2-dependent trends in locus expression were identified, supporting evidence that O_2 is an important and only somewhat understood factor in community composition and function.

The biology inferred in this preliminary round of studies drove new ideas for pipeline improvement. Specific refinement strategies have been proposed, including targeting particular elusive taxa for binning. More high quality bins will shed light on the community dynamics for the major players. In addition, the ability to extract more information about community structure and dynamics is explored in Chapter 4.

Chapter 4

STATISTICAL/MACHINE LEARNING FOR METAGENOMICS AND METATRANSCRIPTOMICS

4.1 Abstract

The potential to apply statistical/machine learning to metatranscriptomics data is explored, using the data of Chapter 3. These data (N = 88 samples, d = 0.9 million measurements per sample) are too large to capture more than simple trends via manual inspections as is done in Chapter 3. While typical machine learning datasets have much larger sample sizes (N), select regularized statistical learning techniques are promising. This chapter focuses on two of them: canonical correlation analysis (CCA), and exploration of the approximated gene-expression partial correlation (pcor) matrix. Results generated from this chapter are intended to generate hypotheses that will be tested with wet-lab experiments. This chapter is left at an exploratory level; subsequent students will pursue the identified methods.

4.2 Introduction

4.2.1 Machine Learning for Metatranscriptomics

Essential Vocabulary

Before jumping in, this subsection describes key vocabulary for statistical/machine learning. "Sample" is used equivalently to biology, and the variable N is used to denote the sample size. In this experiment, N=88 for the 88 bottles that were sequenced. "Feature" describes a measured value that is supplied to an algorithm. A feature could be a raw measurement value (e.g. expression of gene A), or potentially some

abstracted version of one or more measurements such as log(expression), or (expression of A)*(expression of B). The variable d describes the number of features for each sample.

Machine learning is often used to make predictions, so there are terms to describe the different datasets used to train, validate, and test the models. Ideally a new dataset is split into training data, and test data. Training data are usually further split into chunks which can be iterated over to tune model ”hyperparameters”, a process called ”cross-validation”. One example of a hyperparameter is the ”regularization” strength, a measure of the penalty on large coefficients in models. These techniques allow tuning and evaluation of the model without compromising future predictive accuracy.

Considerations for Meta-omics Data

The learning techniques chosen for this chapter were highly influenced by the properties of the underlying data. These properties include sample independence, the samples size, compositional dependency, noise, high-dimensionality, and sparsity. Each is described to contextualize model selection and analysis, and to aid in future consideration of other statistical frameworks.

Non-i.i.d. Sampling

Machine learning techniques are generally applied to large datasets, where each data point is assumed to be independent and sampled from an identical data-generating processes (i.i.d.). For example, a die that is more likely to roll a 6 if a 6 was just rolled exemplifies a violation of i.i.d. Another violation example is a die that is more likely to roll a 6 when it is warm in the room. The time-series nature and differences in O_2 are analogous violations of i.i.d. assumptions for this dataset. This i.i.d. violation can be beneficial, if perturbations allow more signal in the measurements, but can compromise generalizability. Specifically, i.i.d. violations are more likely to produce models that are over-fit to the particular sample set and less likely to generalize to new data. The concern

about i.i.d. violations is set aside for this chapter given that wet-lab experiments will be designed to verify promising results.

Sample size

Machine learning is typically applied to datasets with sample sizes in the thousands or millions. Some classic example domains include prediction of housing prices, or recognition of hand-written digits [148]. Access to tens of thousands of sample points allows the underlying structure to be discovered despite noisy measurements. The data in this study have small N, limiting the number of frameworks that are appropriate. In particular, this small N is not suitable for generating predictive models.

In machine learning, predictive accuracy is tested by applying a model to un-seen data, and assessing the quality of the predictions. Thus, one must have enough data to set aside test data that is never touched until the model is to be published. The importance of not touching these test data until modeling is complete can not be over stated. Assessing model accuracy on the test set before model development is complete often leads to over-fitting to the training data because the researcher will make subsequent modeling choices with those fit statistics in mind. Given an N of only 88, predictive models cannot be produced. This is again consistent with the intention of this chapter to only use results for hypothesis generation.

Compositional Effects

A further challenge of this dataset is that metagenomics and metatranscriptomics data are fundamentally compositional in nature [149, 150]: counts are sampled from a pool and absolute measures are not available. As an illustration of the concern about compositional data, consider a hypothetical sample having only two genes: A and B. Let genes A and B each have 1 unit of expression in a sample. Then sampling of 60 reads leads (on average) to 30 reads counted per gene. Now consider a second sample where the expression of

B is doubled but that of A remains constant. When 60 reads are sampled again, only 20 will be associated with A and 40 will now be associated with B. It thus appears that the expression level of A has decreased, but this is an artifact of the increase of expression of B. In the two-gene example above, the compositional effect is severe, however, as the number of expressed genes increases, the compositional effect grows less strong.

When multiple organisms are present, equivalent compositional effects occur across species. Fluctuations in the abundance of one dominant organism can produce significant correlations for genes belonging to other microbial pairs that would be absent if comparing absolute measures of gene expression levels. As with the example above, these cross-species compositional effects diminish as the number of abundant organisms increases. In the limit of large community complexity, the compositional nature of metatranscriptomics/metagenomics data can be neglected [149]. This thesis assumes that the diversity of abundant organisms and number of expressed genes is high enough to find real biological signal above the effects of compositional artifacts.

Noise

Consideration of experimental noise is necessary for analysis of metagenomics/metatranscriptomics data, as there are many noise-generating steps. The first source of noise is biological variation. Different samples have different population evolution trajectories. These are often driven by unknown latent variables, and thus cannot be controlled for. There is also noise that is common to all sequencing datasets, such as amplification biases, sequencing biases, etc..

Additional noise is added when inferring the origins of reads. The reference DNA to which reads are mapped is known to be imperfect (see Figures 3.5, 3.10), so the tabulated RNA-seq measurements carry uncertainty. On top of that, there is ambiguity in how to count reads that map equally well to two or more places (Figure 3.6.2). This chapter explores the potential of finding signal in spite of these unavoidable layers of noise.

High Dimensionality

High dimensionality relative to the sample size can lead to over-fit models if model complexity is not limited [148]. For example, consider a housing price dataset with N=100 houses, and d=1,000 measurements including some presumably uninformative measurements such as the number of steps to the front door, and the last digit of street address. If only ten variables are allowed in the model, fitting is likely to select the most influential variables (e.g. square footage). However, if an arbitrarily large model is allowed, the model is likely to include the nonsense features. Thus it is essential to limit model complexity when d is larger than N. Regularization is a common technique that adds an additional term to the objective function to penalize large coefficients [148]. Regularization strategies are discussed below, and featured in all techniques highlighted in this chapter.

Sparsity & Heavy-tails

Many of the metatranscriptomics read count measurements for the 0.9 million genes are zero, indicating high sparsity. For the 754,836 genes found to have nonzero expression in at least one sample, only 10.6% of read count values are nonzero. To illustrate, Figure 4.1 shows the read counts for 150 randomly selected genes (columns) across samples (rows). White rectangles indicate zeros. All nonzero values are highlighted in color. Many algorithms are not designed to handle this level of sparsity. For example, models that use Gaussian (normal) distributions to represent the underlying data should not be assumed to handle sparsity, unless explicitly stated.

The distributions of counts are also heavy-tailed: large read counts are less common than small counts. Sparsity aside, these tails are problematic for many model classes, including those based on Gaussian distributions. The models selected in this chapter are reasonable choices given the sparsity and heavy tails, though even stronger adaptations of these models for high sparsity are likely available in the statistics literature if subsequent

students are open to programming their own algorithms.

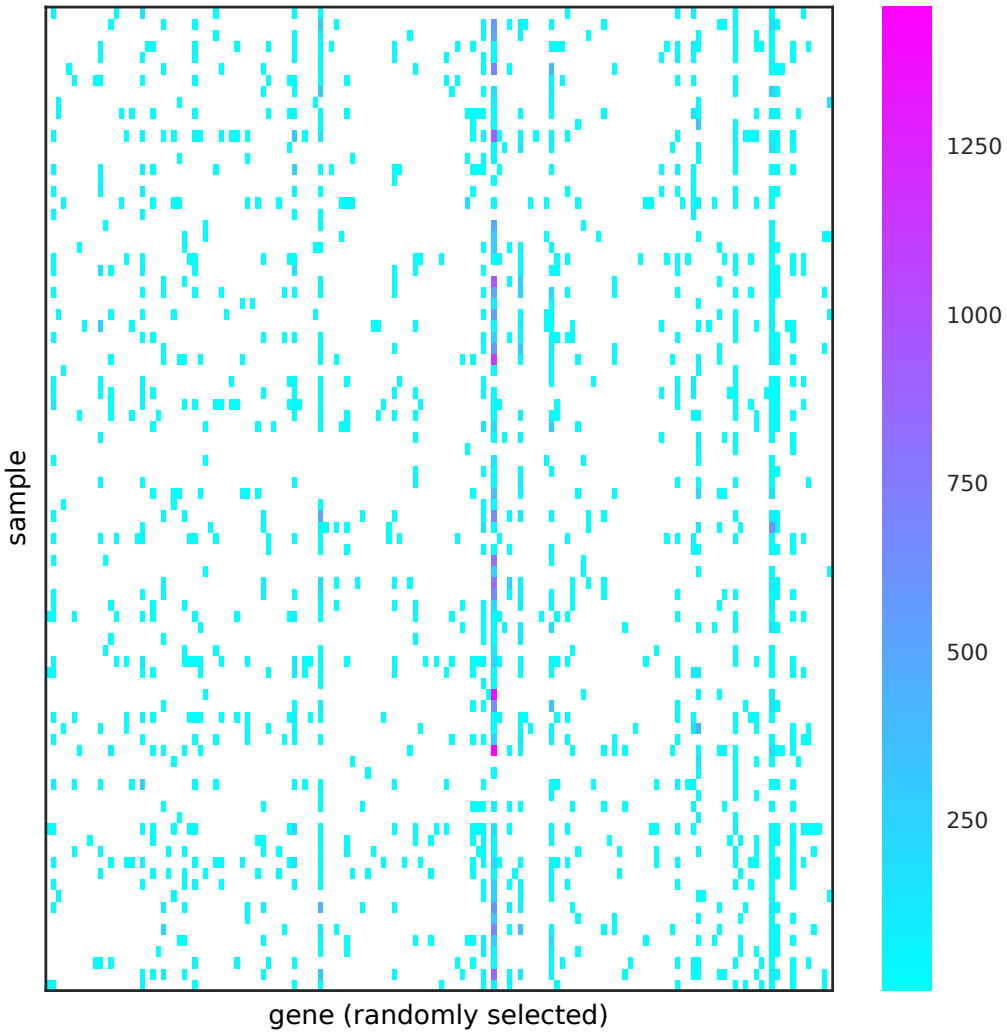


Figure 4.1: Illustration of sparsity in the RNA-seq count data. Nearly 90% of measurements of reads per gene per sample are zero, even after removing genes having zero reads mapped across samples.

4.2.2 General Steps

There are many types of machine learning approaches in the literature. There are, however, three nearly universal steps for common to all: data normalization/standardization, choice of strategy to limit model complexity, and cross validation. Each is described

below.

Standardization

First, standardization of the data is almost always done. Poor fits are often obtained when features/measurements are on different scales (e.g. inches and miles). Even if measurements are in comparable units, many algorithms perform better when the features are scaled to similar magnitudes. This chapter always scaled the features so the features for each gene had variance = 1, but found centering of features (shift so the mean = 0) to be harmful (see Section 4.5.2 and Figure 4.2).

Model Class Selection & Limiting Model Complexity

Next, an appropriate model class needs to be selected. As discussed above, consideration of N and d have large influence on model selection. With N = 88 and d = 0.7 million, extra care to limit model complexity is essential. Fancy algorithms with many parameters (e.g. neural networks) are inappropriate; models that can be limited to a small number of parameters were sought instead. The two most common penalties for limiting model complexity are L1 (a.k.a. "lasso") [151] (absolute values of coefficients are penalized) and L2 [152] (squared coefficients are penalized). Both limit the magnitude of model coefficients, but L1 leads to more truly zero coefficients [151] and is thus particularly appealing in the $d >> N$ regime, or when some features are expected to be uninformative. L1 penalties are used by algorithms in several of the methods explored herein.

Cross Validation

As emphasized previously, the greatest danger of machine learning is production of a model that will not generalize to future un-seen data. This can occur because of poorly selected model hyperparameters. The best practice approach for choosing hyperparameters is cross-validation, whereby the models are trained multiple times,

each time with a different portion of the data set aside as a validation set. The held-out validation set is used to assess the performance of the model trained on the the remaining fraction of the data. Such cross-validation provides the best framework for tuning parameters without touching the test data, thus reducing over-fitting.

4.3 Methods Explored

Machine Learning encompasses broad sets of tools that can be as simple as linear regression or as potentially complex as neural networks. This chapter uses canonical correlation analysis, a classical statistical tool, and various methods of obtaining and exploring graphs based on estimates of the gene-expression partial correlation matrix.

4.3.1 CCA

Canonical correlation analysis (CCA) is a classic method dating back to 1936 [153] that highlights relationships between two sets of variables [154]. In particular, CCA identifies linear combinations of variables that maximize the correlations [1] between measurements in two matrices. It can be used as a statistical learning algorithm when the vectors produced predict gene expression of new data that have not been used to fit or assess the model. In this chapter, if reasonable prediction at the cross-validation step is seen, the genes included in those vectors could highlight interesting biology to target with subsequent experiments. The utility of CCA was explored by testing for correlation between a linear combination of a set of methylotroph genes and a linear combination of a set of methanotroph genes. The data used were the tabulated RNA-seq reads after alignment to the 55 isolate genomes, allowing for labeling of species category.

The large number of features (d) relative to the number of samples (N) motivated use of a variation that produces sparse vectors. In other words, a method was sought that could use regularization to pick out a handful of genes (features) in each category, rather

[1] Pearson r, thus "Canonical Correlation" [154]

than all features. The R package PMA[2] [155] was selected. CCA variants that focused on both sparsity and compositional data were not found.

4.3.2 Gene Expression Partial Correlations

The observational studies of Chapter 3 were not able to answer the pressing question of what drives the changes in community composition and the patterns of organism distribution (see Figures 3.5, 3.6). What drives the higher abundance of methylotrophs in the high O_2 samples? What makes some samples more likely to have high abundances of Bacteroidetes or Burkholderiales? Do certain methylotrophs flourish when certain methanotrophs express particular sets of genes?

The trends of which genes tend to be co-expressed across samples can help answer these questions. Correlations are not the best measure to infer causality, as they detect independence rather than dependence [156]. A better measure is partial correlation, which measures the strength and directionality of a linear relationship between two variables while controlling for the effects of other variables. Partial correlations are more likely to be linked to causality than correlations, but are also more difficult to estimate when $d > N$ (see below).

The set of partial correlations can be represented as a graph describing how species interact. Genes can be represented by nodes, and edges can be added for genes with significant partial correlations [157]. Once a graph is in hand, community structure can be detected by searching for groups of nodes that are highly connected to each other but have fewer connections to nodes outside the group [158]. Edges that are labeled by organism type and/or taxonomy can be used to find sub-structures including multiple species.

A reduced version of the partial correlation matrix is desired. For a network of 0.7M genes (7.0E5), there are approximately $0.7M^2/2$ ($0.7E11$) potential partial correlation

[2] https://cran.r-project.org/web/packages/PMA/PMA.pdf

values. These partial correlations can be represented in a matrix, or equivalently as edges in a graph. Clearly the majority of edges are not expected to have meaningful partial correlations, so the goal is to identify the handful of edges that that highlight meaningful biology. Furthermore, in this $d >> N$ regime, the traditional partial correlation matrix is ill-formed [159] such that shrinkage [3] is required.

There are many potential algorithms that can estimate the reduced set of partial correlations and highlight interesting biology in $d >> N$ regimes. The Strimmer lab has made great advances in this area [160, 156, 161]. One Strimmer Lab method uses a shrinkage formula (see the R package `corpcor`[4]) and analytic approximations that leads to an estimation of the thinned partial correlation matrix from the gene counts matrix. They use a shrinkage parameter that shrinks the empirical correlations towards the identity matrix, while leaving empirical variances intact. While promising, their methods have not been applied to larger omics sets. The Strimmer group's 2005 paper uses 4,289 genes of *E. coli* and 8 RNA-seq samples [156]. The 2007 paper used 800 pre-selected *Aradopsis thaliana* genes [162]. This chapter explores the potential of using their methods on larger and higher-sparsity metatranscriptomics data.

Ledoit-Wolf is another route to that uses an empirical rule (no need for cross-validation) to estimate the inverse correlation matrix [163]. This inverse correlation matrix is proportional but not equal to partial correlations. Further thinning of the network with something such as false discovery rate corrections can be used to thin edges [162].

Graphical lasso is another promising method of producing the desired partial correlation network [164]. This technique uses L1 regularization to directly compute a sparse matrix of partial correlations. Cross-validation can be used to optimize the regularization strength.

---

[3] similar to regularization

[4] http://strimmerlab.org/software/corpcor/index.html

Mining the Partial Correlation Matrix

Even a sparse partial correlation matrix requires computational tools to investigate. If the top 5% of potential partial correlation values are kept, this is still $0.05 * (1.4E11) = 7.0E9$ edges. The graphs can be explored by a variety of methods, including NetworkX[5] (Python) [165], and the specialized graph database Neo4j[6].

4.3.3 Taxonomy of Genes and Contigs

Knowledge of the taxonomy of genes and contigs are necessary for finding interesting cross-species interactions. For CCA, taxonomy labels allow separation by microbe type (e.g. methanotrophs vs methylotrophs). Labels of species in gene-expression partial correlation graphs allow identification of cross-species edges and sub-graphs that include genes from multiple species. These taxonomic labels can be done at contig level, or the gene level. Ideally labels at both levels are consistent.

Several methods for obtaining taxonomic labels were explored. Most tools that assign taxonomic labels (e.g. MEGAN [166], Kaiju [167], Taxator-tk [168], MetaPhlAn [169]) are designed for individual raw sequencing reads rather than contigs. Caution must be used when extending these methods to contigs, due to the larger length of contigs and the variable sizes. For example, taxonomy calls may be made on short, say 100bp regions, of 100kb contigs. If that region is representative of the whole contig, than a short match works great. However, aggregating information about matches along the length of the contig should be more robust.

There are few tools designed for contig taxonomy assignments available. MyTaxa aims to classify contigs using an extension of the Average Amino Acid Identity (AAI) concept [170]. This tool was also tested, but appears to be broken and shows no signs of recent maintenance. Currently a tool for per-gene taxonomic classifications using BLAST best

[5] https://networkx.github.io/

[6] https://neo4j.com/

hits and large word sizes is being developed by David Beck. Additional tools will be required to infer taxonomy across whole contigs, as the labels for genes within a contig will not always agree.

4.4 Materials and Methods

The data used are RNA-seq counts described in Chapter 3.

4.4.1 Taxonomy Predictions

Kaiju [167] was tested for the ability to predict contig taxonomy. The settings used included `-a greedy -e 10000 -x`: greedy mode with up to 10,00 mismatches and filtering of query sequences containing low-complexity regions. As discussed below (Section 4.5.1), concern about predictions of the taxonomy of long contigs using small portions of their coding sequence led to exploration of BLAST-based methods. MEGAN was downloaded but was not pursued due to poor support for the command line interface.

4.4.2 Canonical Correlation Analysis (CCA)

CCA was applied to the mappings of the metatranscriptomes of Chapter 3 to the 55 isolate genomes (Table 3.1), in units of RPKM. Expression levels across genes with the same gene name were pooled. These data were then split into two matrices: one contained all genes from methanotrophic species and the other contained genes from the non-methanotrophic methylotrophic species. Gene products having zero variance were removed.

Each feature was normalized by scaling the variance to 1 with Python's StandardScaler[7], but was not centered to have zero mean due to sparsity concerns (see Figure 4.2). Cross-validation was done to assess predictive accuracy across regularization strengths.

[7] http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Penalties for the two matrices were kept at the same value, rather than use of a grid search. R code was called from Python through rpy2[8] [171]. All code is available on GitHub: CCA_Omics.

4.4.3 Estimating the Partial Correlation Matrix

Prepping Data: Trimming and Normalization

Data was trimmed and normalized before partial correlation calculations were performed. First, genes with zero reads assigned across all samples were removed, bringing down the number of features from ∼0.9 million to ∼0.75 million. Next, variants of these data with different total numbers of features were prepared by varying a cutoff that accounted for sequencing depth. The criteria for inclusion was contribution of a specified fraction of reads in at least one sample. This measure allowed features that were important in one sample, regardless of sequencing depth, to be included. Dividing by the fraction of reads mapped to coding sequences (e.g. RPKM) seemed reckless given low fractions of reads mapped to coding sequences (see Figure 3.10). These datasets of varying size were used to test how well various algorithms scale to this large dataset.

Three approaches were used to approximate the partial correlation matrix: (1) Ledoit-Wolf estimation [163], followed by removing small-magnitude edges, (2) Graphical Lasso [164] with varying regularization strengths, and (3) the GeneNet [9] package [160, 156]. For all three approaches, the variance of each feature was scaled to 1 prior to training the models. All three were tested on large-memory (244 GB) AWS instances.

---

[8] https://rpy2.readthedocs.io/en/version_2.8.x/index.html

[9] http://strimmerlab.org/software/genenet/

Ledoit-Wolf

Estimates of the pseudo-inverse matrix, a matrix proportional to the partial correlation matrix, was estimated via the Python implementation of Ledoit-Wolf[10]. Data sets of increasing size (increasing number of genes) were tried on an EC2 `r4.8xlarge` (244GB memory) instance. The fact that values produced are proportional but not equal to partial correlations prevented use of edge-thinning via false discovery rate tools such as fdrtool[11], which only accept test statistics. Instead, the partial correlations with the largest magnitude were extracted. Specifically, the 2.5% of edges with the most positive, and 2.5% of the edges with the most negative values were retained.

Graphical Lasso

Graphical Lasso was also tested, again with increasing file sizes. The Scikit-Learn implementation[12] [172] (Python) was tested, as well as the Huge[13] [173] package of R, which promises better scalability for large data.

GeneNet Package

GeneNet[14] was called from R. This package uses an analytic shrinkage estimator of the correlation matrix, so cross-validation was not required. Again, genes were trimmed out of the input data if they contributed less than a specified percent in all 88 samples. The top million edges, having the largest partial correlation magnitudes, were explored.

---

[10] http://scikit-learn.org/stable/modules/generated/sklearn.covariance.LedoitWolf.html

[11] https://cran.r-project.org/web/packages/fdrtool/fdrtool.pdf

[12] http://scikit-learn.org/stable/modules/generated/sklearn.covariance.GraphLasso.html

[13] https://cran.r-project.org/web/packages/huge/huge.pdf

[14] https://cran.r-project.org/web/packages/GeneNet/GeneNet.pdf

4.4.4 *Exploring Networks*

Different tools were used to explore the resulting matrix structures, in graph form. Regardless of the software used, the general graph structure included nodes as genes, and edges as partial correlations.

NetworkX

Tabular data with one row per edge were loaded into Pandas. Unique loci were identified, then loaded into NetworkX. Edges were added by iterating over edge dataframes. NetworkX haas built-in methods for computing basic network properties such as average connectedness. See Python code https://github.com/BeckResearchLab/meta4/blob/master/rnaseq/pcor_new/networkx/networkx_helpers.py.

Neo4j

Use of Neo4j was also explored (see https://github.com/JanetMatsen/Neo4j_meta4) through a mix of Python and Java code.

4.5 Results and Discussion

4.5.1 *Taxonomy Predictions*

Taxonomy calls at the gene and contig level were sought to enable applications of statistical learning that probe interactions between microbes. While Kaiju produced usable taxonomy calls for the majority of contigs, exploration of these data raised concerns about extrapolation of taxonomic assignments from small match regions to long contigs. Currently David Beck is developing a tool for BLAST-based taxonomic assignments of individual genes. These assignments will be rolled up into contig-level taxonomic predictions. The statistical learning technique results described below could not use contig taxonomy predictions, but future directions that use this rich layer of

information are proposed.

4.5.2 Avoid Centering Sparse Data Prior to Learning

Early work for this chapter highlighted the importance of not centering scaled features during preprocessing (Figure 4.2). The sparsity of metagenomic data leads to peaks at zero being shifted away from zero during centering. The heavier the tail, the greater the shift. These shifted peaks led to strange artifacts in modeling, so centering during preprocessing was always avoided.

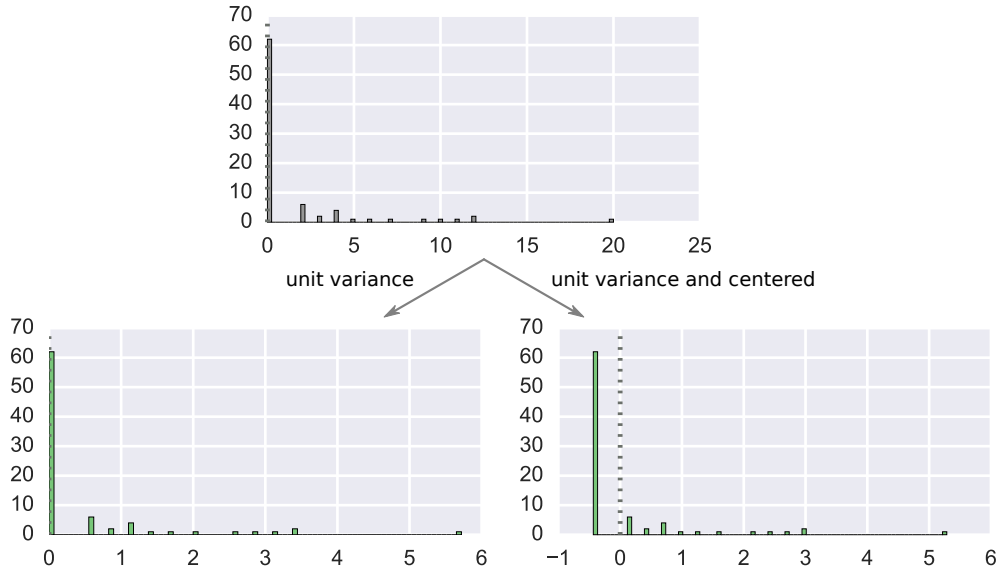


Figure 4.2: When preparing features for machine learning, centering sparse data is not advised. All plots have read counts (RPKM, with or without scaling) on the x-axis and the number of samples on the y-axis. The top plot shows the untransformed distribution of RPKM for a particular copy of a gene encoding "methane monooxygenase component A alpha chain". The gene is not expressed in many of the samples, giving the large peak at zero. Below, two different normalization schemes are shown: standardization so the variance is 1 (left), and standardization so the mean is zero and the variance is 1 (right). Most machine learning algorithms will produce models with problematic artifacts when fitting to such shifted-zero peaks.

The CCA results were particularly problematic when features were centered.

Specifically, very high correlations were predicted, but only using features where most measurements had been zero prior to centering. Variants of CCA that are designed to produce sparse results from sparse data were searched for, but not found.

4.5.3 Partial Correlation Matrix Estimation

Graphical Lasso

None of the methods for estimating the partial correlation matrix scaled well to the entire 0.75 million gene dataset. The general strategy for exploring algorithm scalability was to run each pipeline on input datasets of increasing size. The Scikit-learn implementation of Graphical lasso scaled particularly poorly, struggling even with the top 815 features and producing floating point errors: `FloatingPointError: Non SPD result: the system is too ill-conditioned for this solver`. In addition the necessity to tune the regularization strength via cross-validation added substantial compute requirements. Further exploration of this tool is not suggested, unless a few dozen genes have been pre-selected for exploration, as was done in the Strimmer Lab papers.

The variant of graphical lasso implemented in the Huge package scaled better to larger data than the Scikit-learn implementation, but still could not solve large networks. Instead, other techniques were pursued.

Ledoit-Wolf

The Scikit-learn Ledoit-Wolf algorithm scaled reasonably: it was able to approximate the partial correlation matrix for the top 26,968 genes. The values it reports, however are inverse covariance matrix values, not the partial correlation matrix. The inverse covariance matrix values are proportional to the partial correlations, but they are not true partial correlations.

Furthermore, genes that are expected to have positive partial correlations in fact had negative values. To verify this trend, a minimal example was prepared. The values of 9

genes corresponding to three *pmoCAB* clusters were used, and again the expected within-operon precision matrix values were negative (Figure 4.3). This may be the result of an unexpected sign convention that is not described in the Python documentation. Further reading of the Whittaker textbook [159] could elucidate the steps needed to get from inverse covariance estimates to partial correlations, and clarify this sign issue.

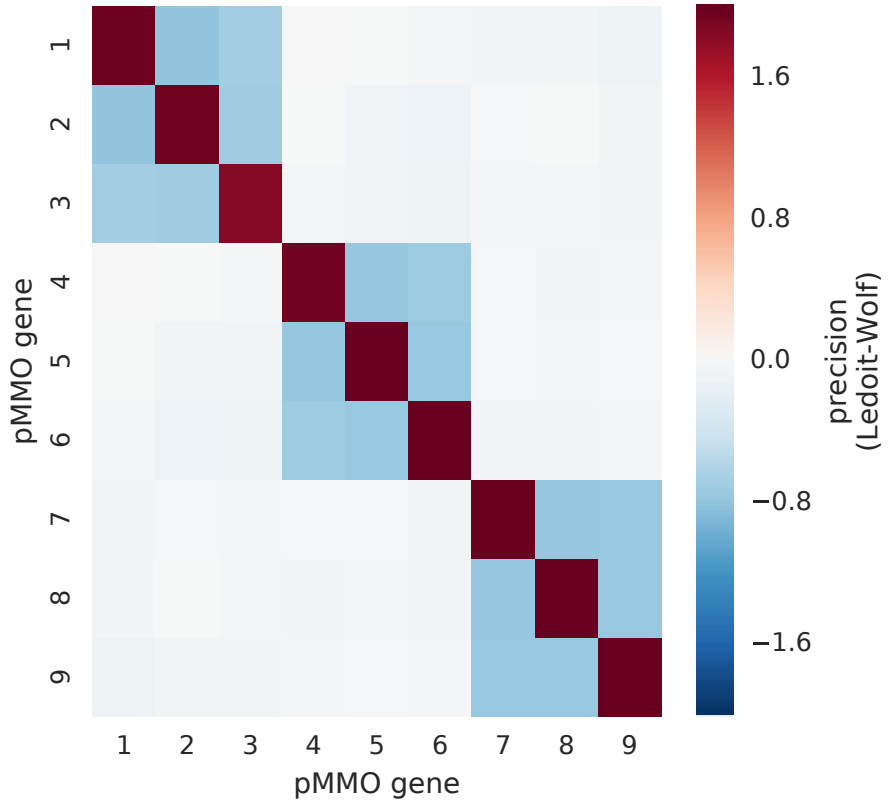


Figure 4.3: Ledoit-Wolf estimated precision matrix values (proportional to partial correlations) for a minimal example of 9 genes in three different three-gene *pmoCAB* clusters. The read counts were scaled to unit-variance prior to application of the Ledoit-Wolf function of Scikit-learn (Python).

GeneNet

GeneNet provided excellent estimates of the partial correlation matrix. The algorithms scaled well, despite previous publications having focused on tens or hundreds of genes.

Sequential genes in clusters proved to have the expected positive-valued partial correlations, and an input file of the top 26,968 genes produced partial correlation estimates in about 6 hours on an AWS `r4.8xlarge` instance (244GB memory). Larger networks produced memory errors.

Validation of the GeneNet Partial Correlation Graph

Partial correlation graphs were explored with NetworkX and Neo4j. NetworkX has the large advantage of being quick to program for Python developers, but Neo4j has a much more powerful query language. The Neo4j query language allows such queries as

```
MATCH (n)<- [e {cross_species:'True'}] -> (m)
    WHERE n.gene_product =~ '.*transport.*'
        AND e.pcor_abs > 0.02
    RETURN n, m
```

This query demonstrates restriction to cross-species edges where the first gene product of the first node contains the substring“transport” and the partial correlation magnitude is above 0.02. Neo4j may be used more in future analysis, but the short-term goals of probing the basic network structure were better served by NetworkX.

The top 1 million partial correlation values (edges) had the following distribution (Figure 4.4):

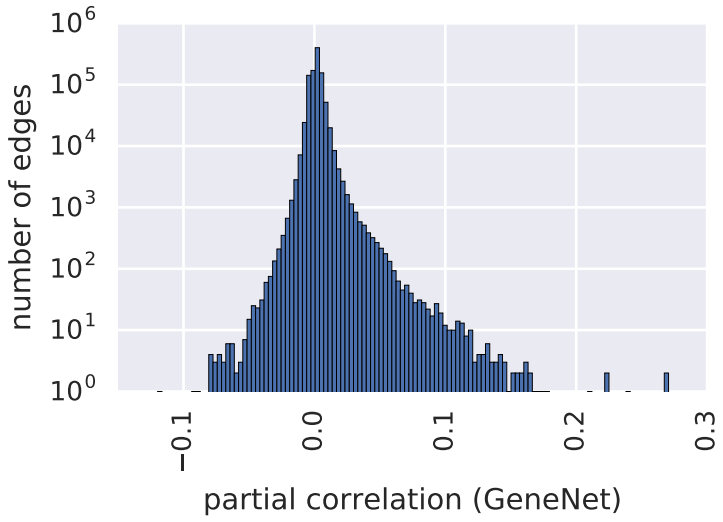


Figure 4.4: Distribution of the 1 million partial correlation values with the largest magnitudes, as estimated by GeneNet.

This set of one million edges corresponds to 19,850 nodes (genes). The average degree (edges per node) is 100.8, with the distribution shown in Figure 4.5.

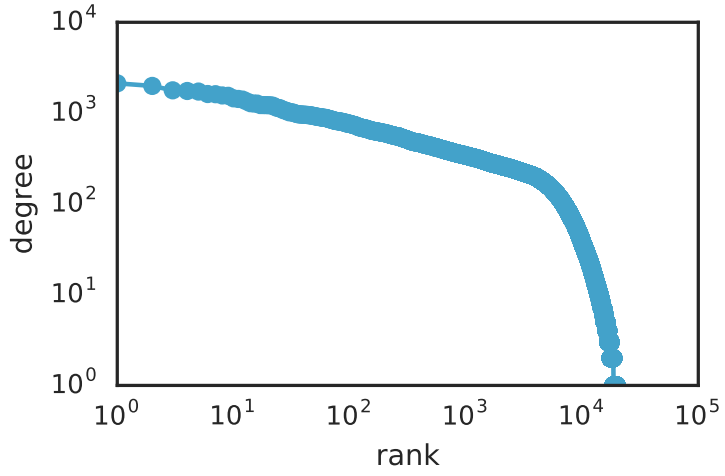


Figure 4.5: Degree rank of plot for nodes in GeneNet derived partial correlation network. The most connected nodes are shown on the left, and are each connected to more than 1,000 other genes. The least connected nodes have less than 10 connections.

Initial measures of the validity of the partial correlation matrix focused on the distribution of partial correlations for different sets of genes. For example, the partial correlations of genes on the same contig are expected to be higher than partial correlations of genes on different contigs. This turned out to be true (Figure 4.6).

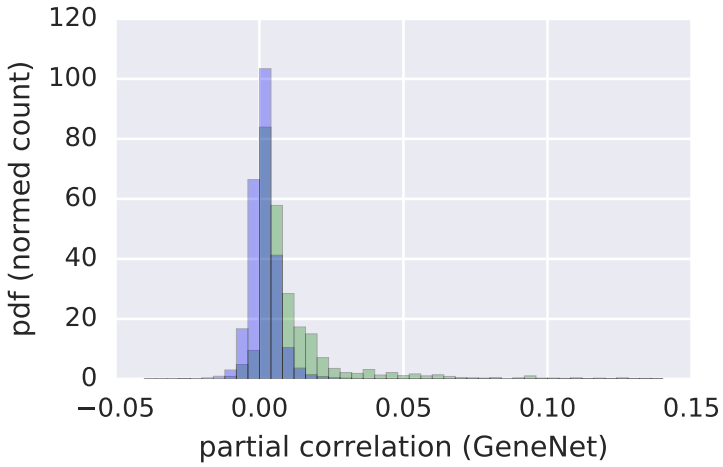


Figure 4.6: Distribution of partial correlations: same-contig versus across-contig gene pairs, with each distribution normalized to have unit area. Cross-contig gene pairs (blue bars, 997,276 edges) have lower partial correlations than pairs of genes found on the same contig (green bars, 2,724 edges).

Similarly, higher partial correlations are expected between subunits of the same protein, or enzymes that are under the same regulation. This hypothesis was tested by looking at the distribution of partial correlations between subunits of particulate methane monooxygenase *pmo*. All edges connecting pairs of *pmo* genes were gathered and split into to sets: edges corresponding to the same *pmo* cluster (sequential gene annotations on the same contig), and edges corresponding to two *pmo* genes that are on different contigs or are not sequential. The distributions (Figure 4.7) confirm that subunits of the same operon have higher partial correlation values, adding validity to the GeneNet model. The same is true for methanol dehydrogenase (*mdh*) subunit pairs (Figure 4.8) and hexulose-6-phosphate synthase:hexulose-6-phosphate isomerase gene pairs (Figure C.5).

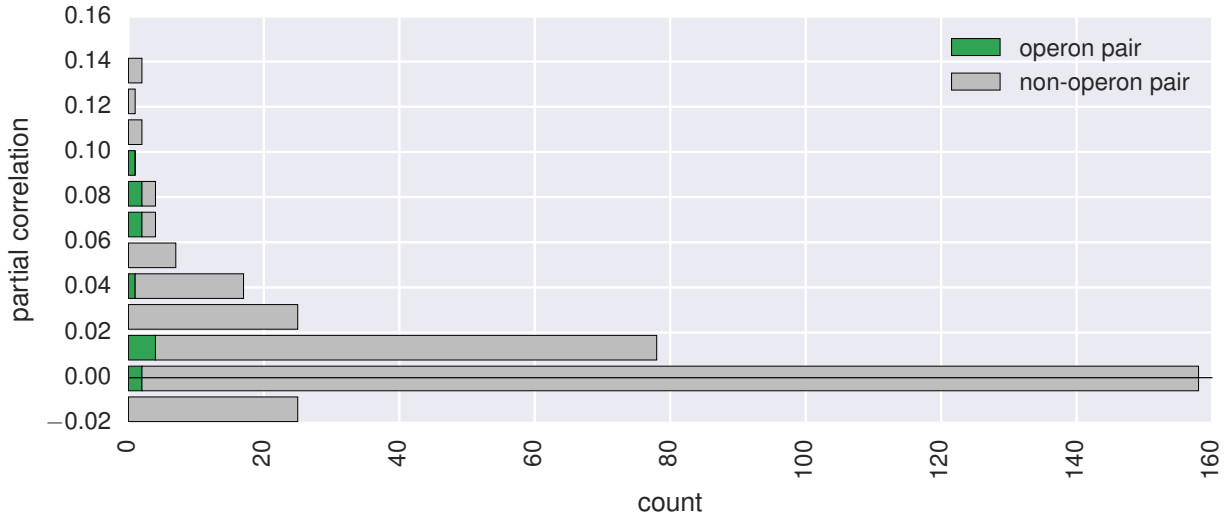


Figure 4.7: Partial correlation values for *pmo*:*pmo* subunit pairs in the GeneNet network. Green bars represent pairs that appear to be on the same operon, and gray bars represent all other pairs.

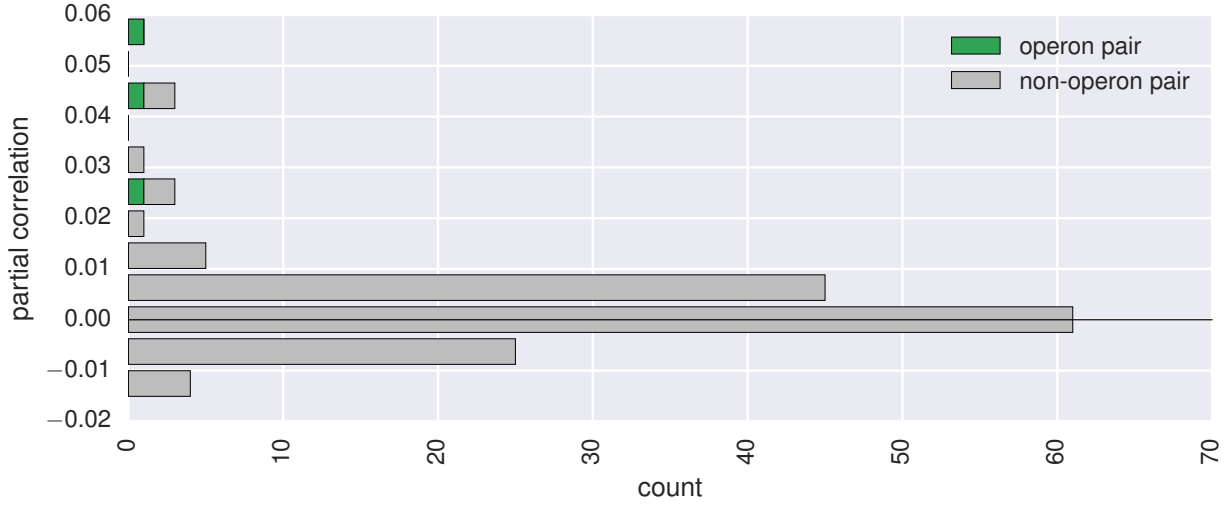


Figure 4.8: Partial correlation values for *mdh* subunit pairs in the GeneNet network. Green bars represent pairs that are sequential on a contig, and gray bars represent all other pairs.

Future Directions for GeneNet Analysis

This chapter identified great potential to use GeneNet for estimation of between-gene partial correlation values. The next step is to mine these graphs for interesting biology. Community structure can be detected by searching for groups of nodes that are highly connected to each other, but have fewer connections to nodes outside the group [174]. Layering in taxonomic labels for nodes will allow the ability to look at partial correlation values between species.

4.5.4 CCA

CCA was run to see if subsets of genes expressed by methanotrophs have expression that is correlated with subsets of non-methanotrophic methylotroph genes. The regularization path leading to the optimal objective function had penalties for both vectors equal to 0.026, indicating high regularization. The vectors it selected for the methanotrophy and methylotrophy genes each included 5 genes, but a correlation of only 0.656. The genes identified did not point toward meaningful biology, either (Tables 4.1, 4.2). This promising technique did not overcome the noise in these metatranscriptomes. Looking back to Figure 3.8, these null results are likely due to low input data quality. This method should be revisited when higher-confidence tabulated RNA-seq data becomes available, or used to target other questions.

Table 4.1: Canonical correlation analysis: top methanotroph features.

gene number	gene name	coefficient
1	PEP-CTERM protein-sorting domain-containing protein	0.977
2	heavy metal efflux pump, CzcA family	0.053
3	methionyl-tRNA synthetase	0.028
4	type I restriction enzyme M protein	0.203
5	zinc and cadmium transporter	0.017

Table 4.2: Canonical correlation analysis: top methylotroph features.

gene number	gene name	coefficient
1	3-oxoacyl-[acyl-carrier protein] reductase	0.214
2	GMP synthase (glutamine-hydrolysing)	0.735
3	chaperonin GroEL	0.544
4	citrate synthase	0.014
5	phosphoglucosamine mutase	0.341

4.6 Conclusions

This chapter described guidelines for applying statistical learning to metatranscriptomics datasets. It warns against the common practice of centering features in data preprocessing. Opportunities to use CCA and partial correlation graphs were highlighted. Given the small N and large d characteristic of metatranscriptomics data, these models are recommended for hypothesis generation rather than predictive modeling.

Two promising software tools were demonstrated for explorations of partial correlation graphs: NetworkX and Neo4j. NetworkX is powerful and easy to use for

Python programmers. It also has many algorithms built in. Neo4j has a more user friendly query language, but external packages are usually used when data mining algorithms are needed.

Applications of these techniques will be stronger after refining the tabulated RNA-seq data as is described in Chapter 3.

Chapter 5
CLOSING REMARKS

This thesis highlights experimental and computational approaches to understanding cellular metabolism. Chapter 2 demonstrates application of transcriptomics to a pure Type II methanotroph culture under methane-grown conditions. Strong conclusions about which metabolic pathways are active were possible given the high-quality reference genome available for RNA-seq alignments.

Chapter 3 extends these techniques to species rich microbial communities, and uses time-series and experimental differences to probe the driving factors for community composition. In this case, the only reference DNA available for alignments was a handful of isolate genomes, and contigs assembled from the pooled 88 metagenomes. The ability to derive high-quality inferences about the genetic potential and gene expression levels of these samples is much lower than for the pure culture case of Chapter 2. A high-level understanding of which microbes dominate the sample is provided, and steps to refine understanding are outlined.

The potential to extract additional information from the gene-expression levels found in Chapter 3 is explored in Chapter 4. The ability to use canonical correlation analysis to identify correlations in the gene-expression data was tested on a preliminary data set. Methods of estimating the gene-expression partial correlation matrix were tested, and promising results were obtained from GeneNet. Aims to identify cross-species gene pairs that co-occur across samples were proposed.

In total, this work demonstrates use of wet-lab and computational approaches to improve understanding of microbial metabolism. The skills I gained are directly useful for guiding metabolic engineering of these strains, or other bacteria. Furthermore, the

data science skills and ability apply machine learning to the unique challenges of biology have broad applications beyond studies of metabolism. I am thrilled to continue pushing the boundaries of sustainable chemical production using my combination of expertise in microbiology, wet-lab experimentation, and data science.

BIBLIOGRAPHY

[1] Justin B Siegel, Amanda Lee Smith, Sean Poust, Adam J Wargacki, Arren Bar-Even, Catherine Louw, Betty W Shen, Christopher B Eiben, Huu M Tran, Elad Noor, et al. Computational protein design enables a novel one-carbon assimilation pathway. *Proceedings of the National Academy of Sciences*, 112(12):3704–3709, 2015.

[2] Christopher Anthony. *The biochemistry of methylotrophs*. Academic Press Inc, New York, 1982.

[3] Ludmila Chistoserdova, Marina G Kalyuzhnaya, and Mary E Lidstrom. The expanding world of methylotrophic metabolism. *Annual review of microbiology*, 63:477–499, 2009.

[4] Marina G Kalyuzhnaya, Aaron W Puri, and Mary E Lidstrom. Metabolic engineering in methanotrophic bacteria. *Metabolic engineering*, 29:142–152, 2015.

[5] Janet B Matsen, Song Yang, Lisa Y Stein, David AC Beck, and Marina G Kalyuzhanaya. Global molecular analyses of methane metabolism in methanotrophic Alphaproteobacterium, *Methylosinus trichosporium* OB3b. part I: transcriptomic study. *Frontiers in microbiology*, 4:40, 2013.

[6] Song Yang, Janet B Matsen, Michael Konopka, Abigail Green-Saxena, Justin Clubb, Martin Sadilek, Victoria J Orphan, David Beck, and Marina G Kalyuzhanaya. Global molecular analyses of methane metabolism in methanotrophic Alphaproteobacterium, *Methylosinus trichosporium* OB3b. part II. metabolomics and 13C-labeling study. *Frontiers in microbiology*, 4:70, 2013.

[7] Richard S Hanson and Thomas E Hanson. Methanotrophic bacteria. *Microbiological reviews*, 60(2):439–471, 1996.

[8] J Colin Murrell and Mike SM Jetten. The microbial methane cycle. *Environmental microbiology reports*, 1(5):279–284, 2009.

[9] Yuri A Trotsenko and John Colin Murrell. Metabolic aspects of aerobic obligate methanotrophy. *Advances in applied microbiology*, 63:183–229, 2008.

[10] Naomi Ward, Øivind Larsen, James Sakwa, Live Bruseth, Hoda Khouri, A Scott Durkin, George Dimitrov, Lingxia Jiang, David Scanlan, Katherine H Kang, et al. Genomic insights into methanotrophy: the complete genome sequence of *Methylococcus capsulatus* (Bath). *PLoS Biol*, 2(10):e303, 2004.

[11] Peter F Dunfield, Anton Yuryev, Pavel Senin, Angela V Smirnova, Matthew B Stott, Shaobin Hou, Binh Ly, Jimmy H Saw, Zhemin Zhou, Yan Ren, et al. Methane oxidation by an extremely acidophilic bacterium of the phylum Verrucomicrobia. *Nature*, 450(7171):879–882, 2007.

[12] Shaobin Hou, Kira S Makarova, Jimmy HW Saw, Pavel Senin, Benjamin V Ly, Zhemin Zhou, Yan Ren, Jianmei Wang, Michael Y Galperin, Marina V Omelchenko, et al. Complete genome sequence of the extremely acidophilic methanotroph isolate V4, *Methylacidiphilum infernorum*, a representative of the bacterial phylum Verrucomicrobia. *Biology Direct*, 3(1):26, 2008.

[13] Yin Chen, Andrew Crombie, M Tanvir Rahman, Svetlana N Dedysh, Werner Liesack, Matthew B Stott, Maqsudul Alam, Andreas R Theisen, J Colin Murrell, and Peter F Dunfield. Complete genome sequence of the aerobic facultative methanotroph *Methylocella silvestris* BL2. *Journal of bacteriology*, 192(14):3840–3841, 2010.

[14] Lisa Y Stein, Sukhwan Yoon, Jeremy D Semrau, Alan A DiSpirito, Andrew Crombie, J Colin Murrell, Stéphane Vuilleumier, Marina G Kalyuzhnaya, Huub JM Op den Camp, Françoise Bringel, et al. Genome sequence of the obligate methanotroph *Methylosinus trichosporium* strain OB3b. *Journal of bacteriology*, 192(24):6497–6498, 2010.

[15] Lisa Y Stein, Françoise Bringel, Alan A DiSpirito, Sukkyun Han, Mike SM Jetten, Marina G Kalyuzhnaya, K Dimitri Kits, Martin G Klotz, Huub JM Op den Camp, Jeremy D Semrau, et al. Genome sequence of the methanotrophic Alphaproteobacterium, *Methylocystis* sp. Rockwell (ATCC 49242). *Journal of bacteriology*, 2011.

[16] Bomba Dam, Somasri Dam, Michael Kube, Richard Reinhardt, and Werner Liesack. Complete genome sequence of *Methylocystis* sp. strain SC2, an aerobic methanotroph with high-affinity methane oxidation potential. *Journal of bacteriology*, 194(21):6008–6009, 2012.

[17] Donovan P Kelly, Christopher Anthony, and J Colin Murrell. Insights into the obligate methanotroph *Methylococcus capsulatus*. *Trends in microbiology*, 13(5):195–198, 2005.

[18] Ahmad F Khadem, Arjan Pol, Adam Wieczorek, Seyed S Mohammadi, Kees-Jan Francoijs, Henk G Stunnenberg, Mike SM Jetten, and Huub JM Op den Camp. Autotrophic methanotrophy in Verrucomicrobia: *Methylacidiphilum fumariolicum* SolV uses the Calvin-Benson-Bassham cycle for carbon dioxide fixation. *Journal of bacteriology*, 193(17):4438–4446, 2011.

[19] R Whittenbury, KC Phillips, and JF Wilkinson. Enrichment, isolation and some properties of methane-utilizing bacteria. *Microbiology*, 61(2):205–218, 1970.

[20] Roelof Oldenhuis, Johannes Y Oedzes, JJ Van der Waarde, and Dick B Janssen. Kinetics of chlorinated hydrocarbon degradation by *Methylosinus trichosporium* OB3b and toxicity of trichloroethylene. *Applied and environmental microbiology*, 57(1):7–14, 1991.

[21] Emerging technology summary pilot-scale demonstration of a two-stage methanotrophic bioreactor for biodegradation of trichloroethylene in groundwater. http://nepis.epa.gov, 1993.

[22] Mark W Fitch, Gerald E Speitel, and George Georgiou. Degradation of trichloroethylene by methanol-grown cultures of *Methylosinus trichosporium* OB3b PP358. *Applied and environmental microbiology*, 62(3):1124–1128, 1996.

[23] Ching T Hou, Ramesh Patel, Allen I Laskin, and N Barnabe. Microbial oxidation of gaseous hydrocarbons: epoxidation of C2 to C4 n-alkenes by methylotrophic bacteria. *Applied and environmental microbiology*, 38(1):127–134, 1979.

[24] AM Williams. *The biochemistry and physiology of poly-beta-hydroxybutyrate metabolism in Methylosinus trichosporium OB3b*. PhD thesis, Cranefield University, 9 1998.

[25] NV Doronina, VA Ezhov, and Yu A Trotsenko. Growth of *Methylosinus trichosporium* OB3b on methane and poly-β-hydroxybutyrate biosynthesis. *Applied biochemistry and microbiology*, 44(2):182–185, 2008.

[26] Amanda S Hakemian and Amy C Rosenzweig. The biochemistry of methane oxidation. *Annu. Rev. Biochem.*, 76:223–241, 2007.

[27] Jeremy D Semrau, Alan A DiSpirito, and Sukhwan Yoon. Methanotrophs and copper. *FEMS microbiology reviews*, 34(4):496–531, 2010.

[28] Christopher J Oakley and J Colin Murrell. *nifH* genes in the obligate methane oxidizing bacteria. *FEMS microbiology letters*, 49(1):53–57, 1988.

[29] Ann J Auman, Catherine C Speake, and Mary E Lidstrom. *nifH* sequences and nitrogen fixation in type I and type II methanotrophs. *Applied and Environmental Microbiology*, 67(9):4009–4016, 2001.

[30] AJ Lawrence and JR Quayle. Alternative carbon assimilation pathways in methane-utilizing bacteria. *Microbiology*, 63(3):371–374, 1970.

[31] Terje Strøm, Thomas Ferenci, and J Rodney Quayle. The carbon assimilation pathways of *Methylococcus capsulatus*, *Pseudomonas methanica* and *Methylosinus trichosporium* (OB3B) during growth on methane. *Biochemical Journal*, 144(3):465–476, 1974.

[32] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[33] David Vallenet, Laurent Labarre, Zoe Rouy, Valerie Barbe, Stephanie Bocs, Stephane Cruveiller, Aurelie Lajus, Geraldine Pascal, Claude Scarpelli, and Claudine Medigue. MaGe: a microbial genome annotation system supported by synteny results. *Nucleic acids research*, 34(1):53–65, 2006.

[34] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–628, 2008.

[35] Nates An Elango, Ramaswamy Radhakrishnan, Wayne A Froland, Bradley J Wallar, Cathleen A Earhart, John D Lipscomb, and Douglas H Ohlendorf. Crystal structure of the hydroxylase component of methane monooxygenase from *Methylosinus trichosporium* OB3b. *Protein Science*, 6(3):556–568, 1997.

[36] Amanda S Hakemian, Kalyan C Kondapalli, Joshua Telser, Brian M Hoffman, Timothy L Stemmler, and Amy C Rosenzweig. The metal centers of particulate methane monooxygenase from *Methylosinus trichosporium* OB3b. *Biochemistry*, 47(26), 2008.

[37] Sunghoon Park, Leslie Hanna, Robert T Taylor, and Michael W Droege. Batch cultivation of *Methylosinus trichosporium* OB3b. I: Production of soluble methane monooxygenase. *Biotechnology and bioengineering*, 38(4):423–433, 1991.

[38] Patricia A Phelps, Sandeep K Agarwal, Gerald E Speitel, and George Georgiou. *Methylosinus trichosporium* OB3b mutants having constitutive expression of soluble methane monooxygenase in the presence of high levels of copper. *Applied and environmental microbiology*, 58(11):3701–3708, 1992.

[39] Allan K Nielsen, Kenn Gerdes, and J Colin Murrell. Copper-dependent reciprocal transcriptional regulation of methane monooxygenase genes in *Methylococcus capsulatus* and *Methylosinus trichosporium*. *Molecular microbiology*, 25(2):399–409, 1997.

[40] John S Lloyd, Ruth Finch, Howard Dalton, and J Colin Murrell. Homologous expression of soluble methane monooxygenase genes in *Methylosinus trichosporium* OB3b. *Microbiology*, 145(2):461–470, 1999.

[41] J Colin Murrell, Ian R McDonald, and Bettina Gilbert. Regulation of expression of methane monooxygenases by copper ions. *Trends in microbiology*, 8(5):221–225, 2000.

[42] Bettina Gilbert, Ian R McDonald, Ruth Finch, Graham P Stafford, Allan K Nielsen, and J Colin Murrell. Molecular analysis of the *pmo* (particulate methane monooxygenase) operons from two type II methanotrophs. *Applied and environmental microbiology*, 66(3):966–975, 2000.

[43] Andrew J Holmes, Andria Costello, Mary E Lidstrom, and J Colin Murrell. Evidence that participate methane monooxygenase and ammonia monooxygenase may be evolutionarily related. *FEMS microbiology letters*, 132(3):203–208, 1995.

[44] Sergei Stolyar, Andria M Costello, Tonya L Peeples, and Mary E Lidstrom. Role of multiple gene copies in particulate methane monooxygenase activity in the methane-oxidizing bacterium *Methylococcus capsulatus* Bath. *Microbiology*, 145(5):1235–1244, 1999.

[45] Bomba Dam, Michael Kube, Somasri Dam, Richard Reinhardt, and Werner Liesack. Complete sequence analysis of two methanotroph-specific *repABC*-containing plasmids from *Methylocystis* sp. strain SC2. *Applied and environmental microbiology*, 78(12):4373–4379, 2012.

[46] Paul M Berube and David A Stahl. The divergent AmoC3 subunit of ammonia monooxygenase functions as part of a stress response system in *Nitrosomonas europaea*. *Journal of bacteriology*, 194(13):3448–3456, 2012.

[47] Christopher Anthony. Methanol dehydrogenase, a PQQ-containing quinoprotein dehydrogenase. In *Enzyme-Catalyzed Electron and Radical Transfer*, pages 73–117. Springer, 2002.

[48] Kenji Yamada, Masahiro Shimoda, and Ichiro Okura. Purification and characterization of methanol dehydrogenase from *Methylosinus trichosporium* (OB3b). *Journal of molecular catalysis*, 73(3):381–386, 1992.

[49] C_ Anthony, M Ghosh, and CC Blake. The structure and function of methanol dehydrogenase and related quinoproteins containing pyrrolo-quinoline quinone. *Biochemical Journal*, 304:665, 1994.

[50] Sabrina Schmidt, Philipp Christen, Patrick Kiefer, and Julia A Vorholt. Functional investigation of methanol dehydrogenase-like protein XoxF in *Methylobacterium extorquens* AM1. *Microbiology*, 156(8):2575–2586, 2010.

[51] Elizabeth Skovran, Alexander D Palmer, Austin M Rountree, Nathan M Good, and Mary E Lidstrom. XoxF is required for expression of methanol dehydrogenase in *Methylobacterium extorquens* AM1. *Journal of bacteriology*, 193(21):6032–6038, 2011.

[52] Shondelle M Wilson, Marshall P Gleisten, and Timothy J Donohue. Identification of proteins involved in formaldehyde metabolism by *Rhodobacter sphaeroides*. *Microbiology*, 154(1):296–305, 2008.

[53] Hirohide Toyama, Ludmila Chistoserdova, and Mary E Lidstrom. Sequence analysis of *pqq* genes required for biosynthesis of pyrroloquinoline quinone in *Methylobacterium extorquens* AM1 and the purification of a biosynthetic intermediate. *Microbiology*, 143(2):595–602, 1997.

[54] Ramesh N Patel, Ching T Hou, P Derelanko, and A Felix. Purification and properties of a heme-containing aldehyde dehydrogenase from *Methylosinus trichosporium*. *Archives of biochemistry and biophysics*, 203(2):654–662, 1980.

[55] Julia A Vorholt, Ludmila Chistoserdova, Sergei M Stolyar, Rudolf K Thauer, and Mary E Lidstrom. Distribution of tetrahydromethanopterin-dependent enzymes in methylotrophic bacteria and phylogeny of methenyl tetrahydromethanopterin cyclohydrolases. *Journal of bacteriology*, 181(18):5750–5757, 1999.

[56] Arnold C Schwartz, Gaby Gockel, Julia Gross, Bernd Moritz, and Helmut E Meyer. Relations and functions of dye-linked formaldehyde dehydrogenase from *Hyphomicrobium zavarzinii* revealed by sequence determination and analysis. *Archives of microbiology*, 182(6):458–466, 2004.

[57] PJ Large and JR Quayle. Microbial growth on C1 compounds. 5. Enzyme activities in extracts of *Pseudomonas* AM1. *Biochemical Journal*, 87(2):386, 1963.

[58] Gregory J Crowther, George Kosály, and Mary E Lidstrom. Formate as the main branch point for methylotrophic metabolism in *Methylobacterium extorquens* AM1. *Journal of bacteriology*, 190(14):5057–5062, 2008.

[59] Ludmila Chistoserdova, Julia A Vorholt, Rudolf K Thauer, and Mary E Lidstrom. C1 transfer enzymes and coenzymes linking methylotrophic bacteria and methanogenic Archaea. *Science*, 281(5373):99–102, 1998.

[60] Marco A Caccamo, Courtney S Malone, and Madeline E Rasche. Biochemical characterization of a dihydromethanopterin reductase involved in tetrahydromethanopterin biosynthesis in *Methylobacterium extorquens* AM1. *Journal of bacteriology*, 186(7):2068–2073, 2004.

[61] Madeline E Rasche, Stephanie A Havemann, and Mariana Rosenzvaig. Characterization of two methanopterin biosynthesis mutants of *Methylobacterium extorquens* AM1 by use of a tetrahydromethanopterin bioassay. *Journal of bacteriology*, 186(5):1565–1570, 2004.

[62] Ludmila Chistoserdova, Madeline E Rasche, and Mary E Lidstrom. Novel dephosphotetrahydromethanopterin biosynthesis genes discovered via mutagenesis in *Methylobacterium extorquens* AM1. *Journal of bacteriology*, 187(7):2508–2512, 2005.

[63] Marina G Kalyuzhnaya, Natalia Korotkova, Gregory Crowther, Christopher J Marx, Mary E Lidstrom, and Ludmila Chistoserdova. Analysis of gene islands involved in methanopterin-linked C1 transfer reactions reveals new functions and provides evolutionary insights. *Journal of bacteriology*, 187(13):4607–4614, 2005.

[64] Julia A Vorholt, Christopher J Marx, Mary E Lidstrom, and Rudolf K Thauer. Novel formaldehyde-activating enzyme in *Methylobacterium extorquens* AM1 required for growth on methanol. *Journal of bacteriology*, 182(23):6645–6650, 2000.

[65] Ludmila Chistoserdova. Modularity of methylotrophy, revisited. *Environmental Microbiology*, 13(10):2603–2622, 2011.

[66] DR Jollie and John D Lipscomb. Formate dehydrogenase from *Methylosinus trichosporium* OB3b. Purification and spectroscopic characterization of the cofactors. *Journal of Biological Chemistry*, 266(32):21853–21863, 1991.

[67] SS Newaz and LOUIS B Hersh. Reduced nicotinamide adenine dinucleotide-activated phosphoenolpyruvate carboxylase in *Pseudomonas* MA: potential regulation between carbon assimilation and energy production. *Journal of bacteriology*, 124(2):825–833, 1975.

[68] Reiji Takahashi, Takashi Ohmori, Kenji Watanabe, and Tatsuaki Tokuyama. Phosphoenolpyruvate carboxylase of an ammonia-oxidizing chemoautotrophic bacterium, *Nitrosomonas europaea* ATCC 25978. *Journal of fermentation and bioengineering*, 76(3):232–234, 1993.

[69] Yasushi Kai, Hiroyoshi Matsumura, and Katsura Izui. Phosphoenolpyruvate carboxylase: three-dimensional structure and molecular mechanisms. *Archives of Biochemistry and Biophysics*, 414(2):170–179, 2003.

[70] Hiroyoshi Matsumura, Yong Xie, Shunsuke Shirakata, Tsuyoshi Inoue, Takeo Yoshinaga, Yoshihisa Ueno, Katsura Izui, and Yasushi Kai. Crystal structures of C4 form maize and quaternary complex of *E. coli* phosphoenolpyruvate carboxylases. *Structure*, 10(12):1721–1730, 2002.

[71] Robert C Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

[72] Mark A Larkin, Gordon Blackshields, NP Brown, R Chenna, Paul A McGettigan, Hamish McWilliam, Franck Valentin, Iain M Wallace, Andreas Wilm, Rodrigo Lopez, et al. Clustal W and Clustal X version 2.0. *bioinformatics*, 23(21):2947–2948, 2007.

[73] Ivica Letunic and Peer Bork. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1):127–128, 2007.

[74] Christopher Anthony. How half a century of research was required to understand bacterial growth on C1 and C2 compounds; the story of the serine cycle and the ethylmalonyl-CoA pathway. *Science progress*, 94(2):109–137, 2011.

[75] Rémi Peyraud, Patrick Kiefer, Philipp Christen, Stephane Massou, Jean-Charles Portais, and Julia A Vorholt. Demonstration of the ethylmalonyl-CoA pathway by using 13C metabolomics. *Proceedings of the National Academy of Sciences*, 106(12):4846–4851, 2009.

[76] Allison J Pieja, Eric R Sundstrom, and Craig S Criddle. Poly-3-hydroxybutyrate metabolism in the type II methanotroph *Methylocystis parvus* OBBP. *Applied and environmental microbiology*, 77(17):6012–6019, 2011.

[77] Yu A Trotsenko et al. Isolation and characterization of obligate methanotrophic bacteria. *Microbial Production and Utilization of Gases*, pages 329–336, 1976.

[78] Valentina N Shishkina and Yu A Trotsenko. Multiple enzymic lesions in obligate methanotrophic bacteria. *FEMS microbiology letters*, 13(3):237–242, 1982.

[79] AM Wadzinski and DW Ribbons. Utilization of acetate by *Methanomonas methanooxidans*. *Journal of bacteriology*, 123(1):380–381, 1975.

[80] IJ Higgins, DJ Best, RC Hammond, and De Scott. Methane-oxidizing microorganisms. *Microbiological reviews*, 45(4):380–381, 1981.

[81] Sunghoon Park, Nilesh N Shah, Robert T Taylor, and Michael W Droege. Batch cultivation of *Methylosinus trichosporium* OB3b: II. production of particulate methane monooxygenase. *Biotechnology and bioengineering*, 40(1):151–157, 1992.

[82] M Naguib. Alternative carboxylation reactions in type II Methylotrophs and the localization of carboxylase activities in the intra-cytoplasmic membranes. *Zeitschrift für allgemeine Mikrobiologie*, 19(5):333–342, 1979.

[83] RN Patel, S Louise Hoare, and DS Hoare. [14C] Acetate assimilation by obligate methylotrophs, *Pseudomonas methanica* and *Methylosinus trichosporium*. *Antonie Van Leeuwenhoek*, 45(3):499–511, 1979.

[84] Alexander S Reshetnikov, Olga N Rozova, Valentina N Khmelenina, Ildar I Mustakhimov, Alexander P Beschastny, J Colin Murrell, and Yuri A Trotsenko. Characterization of the pyrophosphate-dependent 6-phosphofructokinase from *Methylococcus capsulatus* Bath. *FEMS microbiology letters*, 288(2):202–210, 2008.

[85] ON Rozova, VN Khmelenina, and Yu A Trotsenko. Characterization of recombinant PPi-dependent 6-phosphofructokinases from *Methylosinus trichosporium* OB3b and *Methylobacterium nodulans* ORS 2060. *Biochemistry (Moscow)*, 77(3):288–295, 2012.

[86] TL Weaver, MA Patrick, and PR Dugan. Whole-cell and membrane lipids of the methylotrophic bacterium *Methylosinus trichosporium*. *Journal of bacteriology*, 124(2):602–605, 1975.

[87] James B Guckert, David B Ringelberg, David C White, Richard S Hanson, and Bonnie J Bratina. Membrane fatty acids as phenotypic markers in the polyphasic taxonomy of methylotrophs within the Proteobacteria. *Microbiology*, 137(11):2631–2641, 1991.

[88] Valentina N Shishkina and Yu A Trotsenko. Pathways of ammonia assimilation in obligate methane utilizers. *FEMS Microbiology Letters*, 5(3):187–191, 1979.

[89] J Colin Murrell and Howard Dalton. Ammonia assimilation in *Methylococcus capsulatus* (Bath) and other obligate methanotrophs. *Microbiology*, 129(4):1197–1206, 1983.

[90] J Colin Murrell and Howard Dalton. Purification and properties of glutamine synthetase from *Methylococcus capsulatus* (Bath). *Microbiology*, 129(4):1187–1196, 1983.

[91] Kung-Hui Chu and Lisa Alvarez-Cohen. Effect of nitrogen source on growth and trichloroethylene degradation by methane-oxidizing bacteria. *Applied and environmental microbiology*, 64(9):3451–3457, 1998.

[92] Hyung J Kim and David W Graham. Effect of oxygen level on simultaneous nitrogenase and sMMO expression and activity in *Methylosinus trichosporium* OB3b and its sMMO^{C} mutant, PP319: aerotolerant N_2 fixation in PP319. *FEMS microbiology letters*, 201(2):133–138, 2001.

[93] Aaron H Liepman and Laura J Olsen. Peroxisomal alanine: glyoxylate aminotransferase (AGT1) is a photorespiratory enzyme with multiple substrates in *Arabidopsis thaliana*. *The Plant Journal*, 25(5):487–498, 2001.

[94] William E Karsten, Takashi Ohshiro, Yoshikazu Izumi, and Paul F Cook. Initial velocity, spectral, and pH studies of the serine-glyoxylate aminotransferase from *Hyphomicrobiuim methylovorum*. *Archives of biochemistry and biophysics*, 388(2):267–275, 2001.

[95] Hyung J Kim, David W Graham, Alan A DiSpirito, Michail A Alterman, Nadezhda Galeva, Cynthia K Larive, Dan Asunskis, and Peter MA Sherwood. Methanobactin, a copper-acquisition compound from methane-oxidizing bacteria. *Science*, 305(5690):1612–1615, 2004.

[96] Ramakrishnan Balasubramanian and Amy C Rosenzweig. Copper methanobactin: a molecule whose time has come. *Current opinion in chemical biology*, 12(2):245–249, 2008.

[97] Ramakrishnan Balasubramanian, Stephen M Smith, Swati Rawat, Liliya A Yatsunyk, Timothy L Stemmler, and Amy C Rosenzweig. Oxidation of methane by a biological dicopper centre. *Nature*, 465(7294):115–119, 2010.

[98] Benjamin D Krentz, Heidi J Mulheron, Jeremy D Semrau, Alan A DiSpirito, Nathan L Bandow, Daniel H Haft, Stéphane Vuilleumier, J Colin Murrell, Marcus T McEllistrem, Scott C Hartsel, et al. A comparison of methanobactins from *Methylosinus trichosporium* OB3b and *Methylocystis* strain SB2 predicts methanobactins are synthesized from diverse peptide precursors modified to create a common core for binding and reducing copper ions. *Biochemistry*, 49(47):10117–10130, 2010.

[99] Sukhwan Yoon, Stephan M Kraemer, Alan A DiSpirito, and Jeremy D Semrau. An assay for screening microbial cultures for chalkophore production. *Environmental microbiology reports*, 2(2):295–303, 2010.

[100] Christopher Anthony. The prediction of growth yields in methylotrophs. *Microbiology*, 104:91–104, 1978.

[101] Daniel R Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18(5):821–829, 2008.

[102] Marcel H Schulz, Daniel R Zerbino, Martin Vingron, and Ewan Birney. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8):1086–1092, 2012.

[103] Ann J Auman and Mary E Lidstrom. Analysis of sMMO-containing type I methanotrophs in Lake Washington sediment. *Environmental microbiology*, 4(9):517–524, 2002.

[104] MG Kalyuzhnaya, ME Lidstrom, and L Chistoserdova. Utility of environmental primers targeting ancient enzymes: methylotroph detection in Lake Washington. *Microbial ecology*, 48(4):463–472, 2004.

[105] Andria M Costello, Ann J Auman, Jennifer L Macalady, Kate M Scow, and Mary E Lidstrom. Estimation of methanotroph abundance in a freshwater lake sediment. *Environmental Microbiology*, 4(8):443–450, 2002.

[106] Marina G Kalyuzhnaya, David AC Beck, Alexey Vorobev, Nicole Smalley, Denis D Kunkel, Mary E Lidstrom, and Ludmila Chistoserdova. Novel methylotrophic isolates from Lake Washington sediment and description of a new species in the genus *Methylotenera*, *Methylotenera versatilis* sp. nov. *International Journal of Systematic and Evolutionary Microbiology*, 2011.

[107] Tami L McTaggart, Gabrielle Benuska, Nicole Shapiro, Tanja Woyke, and Ludmila Chistoserdova. Draft genome sequences of five new strains of *Methylophilaceae* isolated from Lake Washington sediment. *Genome announcements*, 3(1), 2015.

[108] Marina G Kalyuzhnaya, Andrew E Lamb, Tami L McTaggart, Igor Y Oshkin, Nicole Shapiro, Tanja Woyke, and Ludmila Chistoserdova. Draft genome sequences of gammaproteobacterial methanotrophs isolated from Lake Washington sediment. *Genome announcements*, 3(2), 2015.

[109] Mary E Lidstrom and Leslie Somers. Seasonal study of methane oxidation in Lake Washington. *Applied and environmental microbiology*, 47(6):1255–1260, 1984.

[110] KM Kuivilal, JW Murray, AH Devol, ME Lidstrom, and Clare E Reimers. *Methane cycling in the sediments of Lake Washington*. American Society of Limnology and Oceanography, Inc., 1988.

[111] Ann J Auman, Sergei Stolyar, Andria M Costello, and Mary E Lidstrom. Molecular characterization of methanotrophic isolates from freshwater lake sediment. *Applied and Environmental Microbiology*, 66(12):5259–5266, 2000.

[112] Igor Y Oshkin, David AC Beck, Andrew E Lamb, Veronika Tchesnokova, Gabrielle Benuska, Tami L McTaggart, Marina G Kalyuzhnaya, Svetlana N Dedysh, Mary E Lidstrom, and Ludmila Chistoserdova. Methane-fed microbial microcosms show differential community dynamics and pinpoint taxa involved in communal response. *The ISME journal*, 9(5):1119–1129, 2015.

[113] Sascha MB Krause, Timothy Johnson, Yasodara Samadhi Karunaratne, Yanfen Fu, David AC Beck, Ludmila Chistoserdova, and Mary E Lidstrom. Lanthanide-dependent cross-feeding of methane-derived carbon is linked by microbial community interactions. *Proceedings of the National Academy of Sciences*, 114(2):358–363, 2017.

[114] Olivier Nercessian, Emma Noyes, Marina G Kalyuzhnaya, Mary E Lidstrom, and Ludmila Chistoserdova. Bacterial populations active in metabolism of C1 compounds in the sediment of Lake Washington, a freshwater lake. *Applied and environmental microbiology*, 71(11):6885–6899, 2005.

[115] Marina G Kalyuzhnaya, Sergey M Stolyar, Ann J Auman, Jimmie C Lara, Mary E Lidstrom, and Ludmila Chistoserdova. Methylosarcina lacus sp. nov., a methanotroph from Lake Washington, Seattle, USA, and emended description of the genus *Methylosarcina*. *International journal of systematic and evolutionary microbiology*, 55(6):2345–2350, 2005.

[116] Marina G Kalyuzhnaya, Sarah Bowerman, Jimmie C Lara, Mary E Lidstrom, and Ludmila Chistoserdova. *Methylotenera mobilis* gen. nov., sp. nov., an obligately methylamine-utilizing bacterium within the family *Methylophilaceae*. *International Journal of Systematic and Evolutionary Microbiology*, 56(12):2819–2823, 2006.

[117] David AC Beck, Tami L McTaggart, Usanisa Setboonsarng, Alexey Vorobev, Lynne Goodwin, Nicole Shapiro, Tanja Woyke, Marina G Kalyuzhnaya, Mary E Lidstrom, and Ludmila Chistoserdova. Multiphyletic origins of methylotrophy in Alphaproteobacteria, exemplified by comparative genomics of Lake Washington isolates. *Environmental microbiology*, 17(3):547–554, 2015.

[118] Tammi Kaeberlein, Kim Lewis, and Slava S Epstein. Isolating" uncultivable" microorganisms in pure culture in a simulated natural environment. *Science*, 296(5570):1127–1129, 2002.

[119] Eric J Stewart. Growing unculturable bacteria. *Journal of bacteriology*, 194(16):4151–4160, 2012.

[120] Zheng Yu, Sascha MB Krause, David AC Beck, and Ludmila Chistoserdova. A synthetic ecology perspective: How well does behavior of model organisms in the laboratory predict microbial activities in natural habitats? *Frontiers in Microbiology*, 7, 2016.

[121] David AC Beck, Marina G Kalyuzhnaya, Stephanie Malfatti, Susannah G Tringe, Tijana Glavina del Rio, Natalia Ivanova, Mary E Lidstrom, and Ludmila Chistoserdova. A metagenomic insight into freshwater methane-utilizing communities and evidence for cooperation between the methylococcaceae and the methylophilaceae. *PeerJ*, 1:e23, 2013.

[122] David AC Beck, Tami L McTaggart, Usanisa Setboonsarng, Alexey Vorobev, Marina G Kalyuzhnaya, Natalia Ivanova, Lynne Goodwin, Tanja Woyke, Mary E Lidstrom, and Ludmila Chistoserdova. The expanded diversity of methylophilaceae from lake washington through cultivation and genomic sequencing of novel ecotypes. *PLoS One*, 9(7):e102458, 2014.

[123] Maria E Hernandez, David AC Beck, Mary E Lidstrom, and Ludmila Chistoserdova. Oxygen availability is a major factor in determining the composition of microbial communities involved in methane oxidation. *PeerJ*, 3:e801, 2015.

[124] Marina G Kalyuzhnaya, Krassimira R Hristova, Mary E Lidstrom, and Ludmila Chistoserdova. Characterization of a novel methanol dehydrogenase in representatives of Burkholderiales: implications for environmental detection of methylotrophy and evidence for convergent evolution. *Journal of bacteriology*, 190(11):3817–3823, 2008.

[125] Victor Kunin, Alex Copeland, Alla Lapidus, Konstantinos Mavromatis, and Philip Hugenholtz. A bioinformatician's guide to metagenomics. *Microbiology and molecular biology reviews*, 72(4):557–578, 2008.

[126] Rose S Kantor, Robert J Huddy, Ramsunder M Iyer, Brian C Thomas, Christopher T Brown, Karthik Anantharaman, Susannah Green Tringe, Robert L Hettich, Susan TL Harrison, and Jillian F Banfield. Genome-resolved meta-omics ties microbial dynamics to process performance in biotechnology for thiocyanate degradation. *Environmental Science & Technology*, 2017.

[127] Naseer Sangwan, Fangfang Xia, and Jack A Gilbert. Recovering complete and draft population genomes from metagenome datasets. *Microbiome*, 4(1):8, 2016.

[128] Torsten Thomas, Jack Gilbert, and Folker Meyer. Metagenomics-a guide from sampling to data analysis. *Microbial informatics and experimentation*, 2(1):3, 2012.

[129] Michael Cantor, Henrik Nordberg, Tatyana Smirnova, Matthias Hess, Susannah Tringe, and Inna Dubchak. Elviz–exploration of metagenome assemblies with an interactive visualization tool. *BMC bioinformatics*, 16(1):130, 2015.

[130] Mads Albertsen, Philip Hugenholtz, Adam Skarshewski, Kåre L Nielsen, Gene W Tyson, and Per H Nielsen. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature biotechnology*, 31(6):533–538, 2013.

[131] Itai Sharon, Michael J Morowitz, Brian C Thomas, Elizabeth K Costello, David A Relman, and Jillian F Banfield. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome research*, 23(1):111–120, 2013.

[132] Donovan H Parks, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*, 25(7):1043–1055, 2015.

[133] Dongwan D Kang, Jeff Froula, Rob Egan, and Zhong Wang. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 3, 2015.

[134] Hsin-Hung Lin and Yu-Chieh Liao. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Scientific reports*, 6, 2016.

[135] Johannes Alneberg, Brynjar Smári Bjarnason, Ino De Bruijn, Melanie Schirmer, Joshua Quick, Umer Z Ijaz, Leo Lahti, Nicholas J Loman, Anders F Andersson, and Christopher Quince. Binning metagenomic contigs by coverage and composition. *Nature methods*, 11(11):1144–1146, 2014.

[136] Svetlana N Dedysh and Peter F Dunfield. Cultivation of methanotrophs. *Hydrocarbon and lipid microbiology protocols, Springer protocols handbooks. Berlin: Springer-Ferlag*, 2014. Sascha's paper used this.

[137] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. HTSeq - a python framework to work with high-throughput sequencing data. *Bioinformatics*, page btu638, 2014.

[138] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[139] Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 2015.

[140] Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, page 153, 2014.

[141] Neha J Varghese, Supratim Mukherjee, Natalia Ivanova, Konstantinos T Konstantinidis, Kostas Mavrommatis, Nikos C Kyrpides, and Amrita Pati. Microbial species delineation using whole genome sequences. *Nucleic acids research*, 2015.

[142] Nicola Segata, Daniela Börnigen, Xochitl C Morgan, and Curtis Huttenhower. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature communications*, 4, 2013.

[143] Günter P Wagner, Koryu Kin, and Vincent J Lynch. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences*, 131(4):281–285, 2012.

[144] Christian MK Sieber, Alexander J Probst, Allison Sharrar, Brian C Thomas, Matthias Hess, Susannah G Tringe, and Jillian F Banfield. Recovery of genomes from metagenomes via a dereplication, aggregation, and scoring strategy. *bioRxiv*, 2017.

[145] Yu-Wei Wu, Blake A Simmons, and Steven W Singer. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 2015.

[146] Gregory J Dick, Anders F Andersson, Brett J Baker, Sheri L Simmons, Brian C Thomas, A Pepper Yelton, and Jillian F Banfield. Community-wide analysis of microbial genome sequence signatures. *Genome biology*, 10(8), 2009.

[147] Frances Chu and Mary E Lidstrom. XoxF acts as the predominant methanol dehydrogenase in the type I methanotroph *Methylomicrobium buryatense*. *Journal of bacteriology*, 198(8):1317–1325, 2016.

[148] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[149] Matthew CB Tsilimigras and Anthony A Fodor. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Annals of epidemiology*, 26(5):330–335, 2016.

[150] John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 139–177, 1982.

[151] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[152] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[153] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[154] Alissa Sherry and Robin K Henson. Conducting and interpreting canonical correlation analysis in personality research: A user-friendly primer. *Journal of personality assessment*, 84(1):37–48, 2005.

[155] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, page kxp008, 2009.

[156] Juliane Schäfer, Korbinian Strimmer, et al. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1):32, 2005.

[157] Ana I Borthagaray, Matías Arim, and Pablo A Marquet. Inferring species roles in metacommunity structure from species co-occurrence networks. *Proceedings of the Royal Society of London B: Biological Sciences*, 281(1792):20141425, 2014.

[158] Alfred Hero and Bala Rajaratnam. Hub discovery in partial correlation graphs. *IEEE Transactions on Information Theory*, 58(9):6064–6078, 2012.

[159] Joe Whittaker. *Graphical models in applied multivariate statistics*. Wiley Publishing, 2009.

[160] Juliane Schäfer, Rainer Opgen-Rhein, and Korbinian Strimmer. Reverse engineering genetic networks using the genenet package. *J Am Stat Assoc*, 96:1151–1160, 2001.

[161] Rainer Opgen-Rhein and Korbinian Strimmer. Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

[162] Rainer Opgen-Rhein and Korbinian Strimmer. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC systems biology*, 1(1):37, 2007.

[163] Olivier Ledoit and Michael Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10(5):603–621, 2003.

[164] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[165] Daniel A Schult and P Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conferences (SciPy 2008)*, volume 2008, pages 11–16, 2008.

[166] DH Huson, Sina Beier, Benjamin Buchfink, Isabell Flade, Anna Górska, Mohamed El-Hadidi, Suparna Mitra, Hans-Joachim Ruscheweyh, and Rewati Tappu. Megan6-microbiome analysis involving hundreds of samples and billions of reads, 2015.

[167] Peter Menzel, Kim Lee Ng, and Anders Krogh. Fast and sensitive taxonomic classification for metagenomics with kaiju. *Nature communications*, 7, 2016.

[168] Johannes Dröge, Ivan Gregor, and Alice Carolyn McHardy. Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. *Bioinformatics*, page btu745, 2014.

[169] Nicola Segata, Levi Waldron, Annalisa Ballarini, Vagheesh Narasimhan, Olivier Jousson, and Curtis Huttenhower. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, 9(8):811–814, 2012.

[170] Konstantinos T Konstantinidis and James M Tiedje. Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(7):2567–2572, 2005.

[171] L Gautier. rpy2: A simple and efficient access to R from Python, 2008.

[172] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[173] Tuo Zhao, Han Liu, Kathryn Roeder, John Lafferty, and Larry Wasserman. The huge package for high-dimensional undirected graph estimation in R. *Journal of Machine Learning Research*, 13(Apr):1059–1062, 2012.

[174] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

Appendix A

SUPPLEMENTAL FOR CHAPTER 2

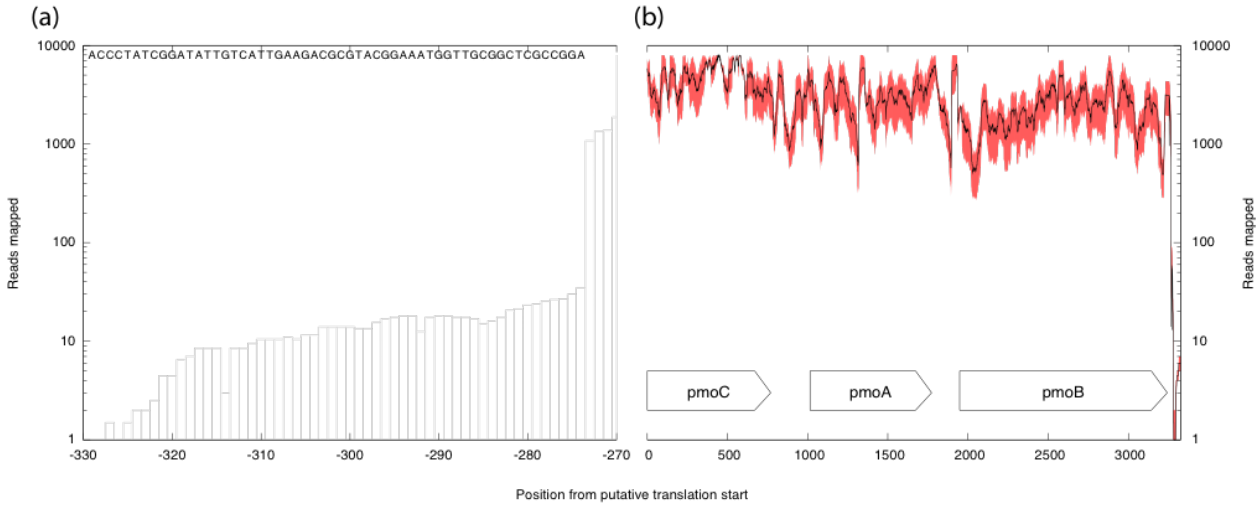


Figure A.1: RNA-Seq reads mapped per base relative to start of *pmo*-operon. Left, the average number of reads mapped to the region -330 to -270 upstream from the start of *pmoC* for two biological replicates. The sequence at each base is shown above the bars. Right, the range (shown in red) and mean of two biological replicates (shown as black line) for the number of reads mapped per base for the coding and downstream region of the *pmo*-operon.

Figure S2

Figure A.2: Genetic organization and relative expression (RPKM) of the *mxa* gene cluster (A), the *pqq* gene cluster (B) and cluster of genes encoding reactions of the H_4MPT-linked C1 transfer pathway (C).

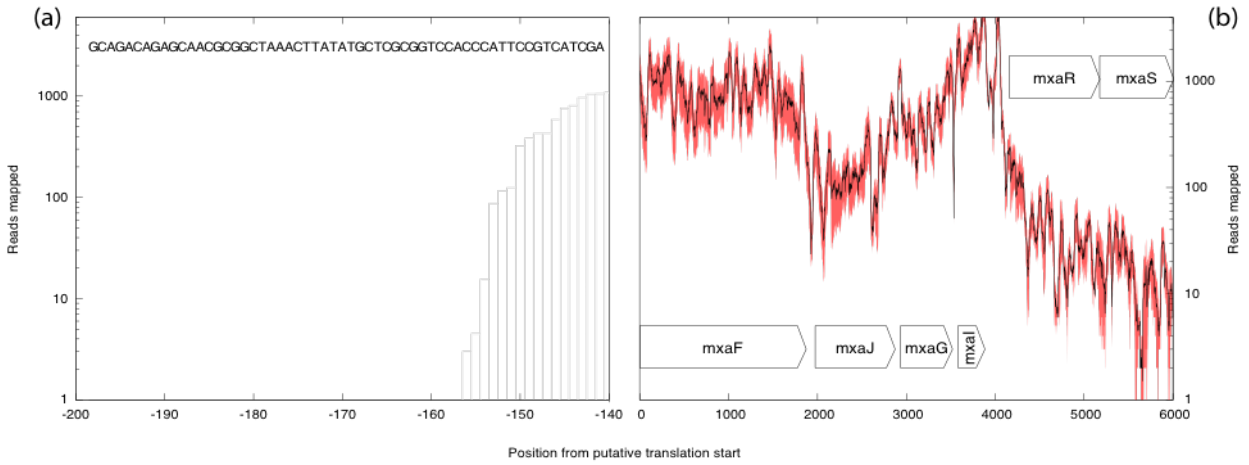


Figure A.3: RNA-Seq reads mapped per base relative to start of *mxaF* ORF. Left, the average number of reads mapped to the region -200 to -140 upstream from the start of *mxaF* for two biological replicates. The sequence at each base is shown above the bars. Right, the range (shown in red) and mean of two biological replicates (shown as black line) for the number of reads mapped per base for the coding and downstream region of the *mxaF* cluster.

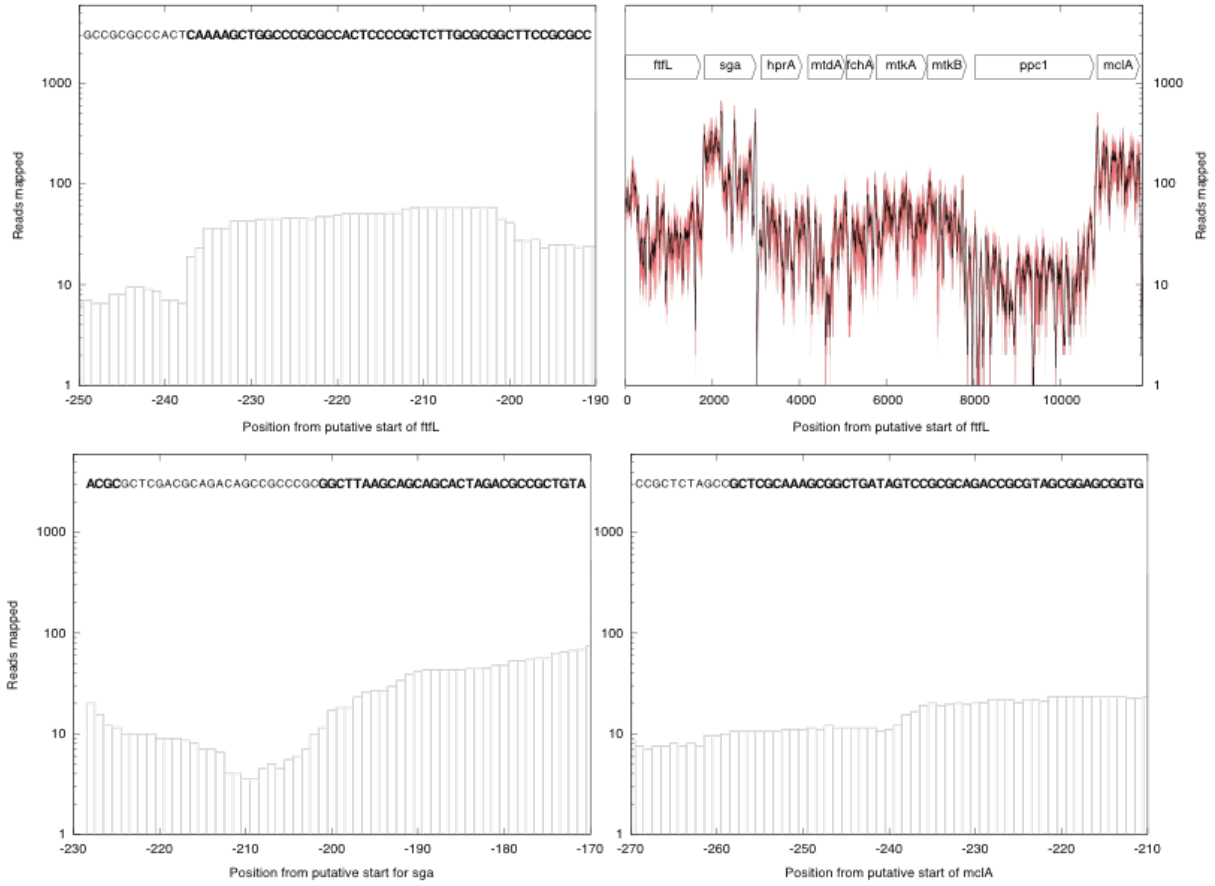


Figure A.4: RNA-Seq reads mapped per base relative to start of serine cycle gene operon. The log10 average number of reads mapped at each base from biological replicates one and two is shown. In (A), the upstream location spanning -250 to -190 from putative start of *ftfL*. (B) The expression over the entire operon. Note the several drop to near zero upstream indicating the operon is not co-transcribed. (C) The -230 to -170 region upstream from *sga*. Bases from the +strand are shown across the top of the figure. (D) The -270 to -210 region upstream of *mclA*.

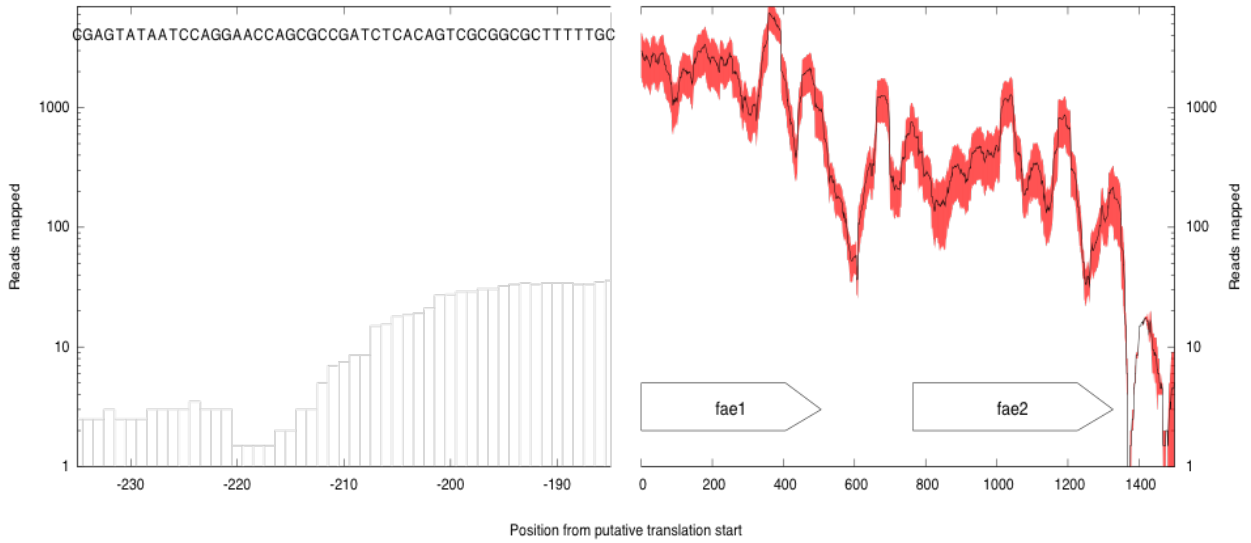


Figure A.5: RNA-Seq reads mapped per base relative to start of *fae1-1*. Left, the average number of reads mapped to the region -235 to -185 upstream from the start of *fae1-1* gene for two biological replicates. The sequence at each base is shown above the bars. Right, the range (shown in red) and mean of two biological replicates (shown as black line) for the number of reads mapped per base for the coding and downstream region of the *fae1-1* and *fae1-2* genes.

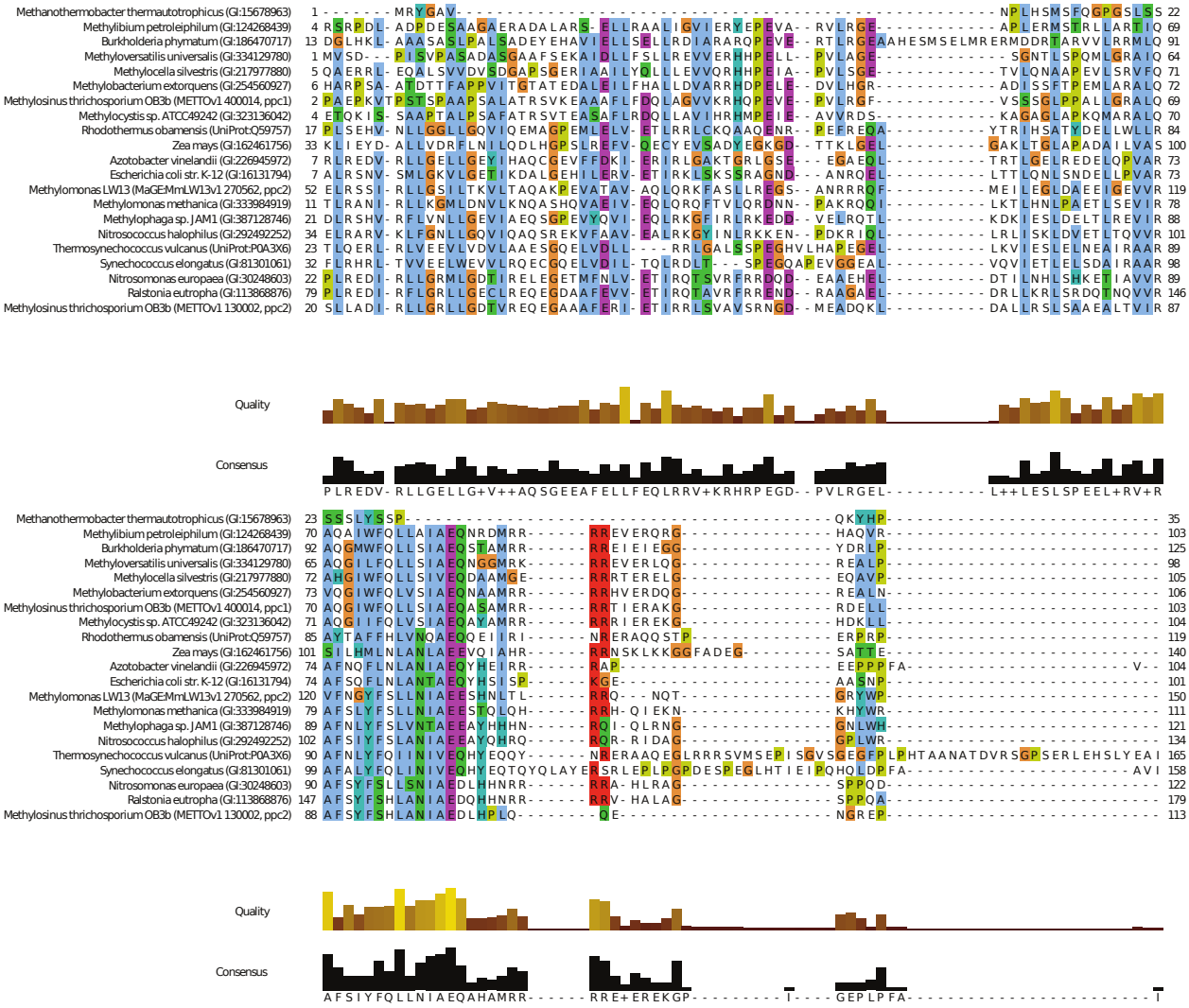
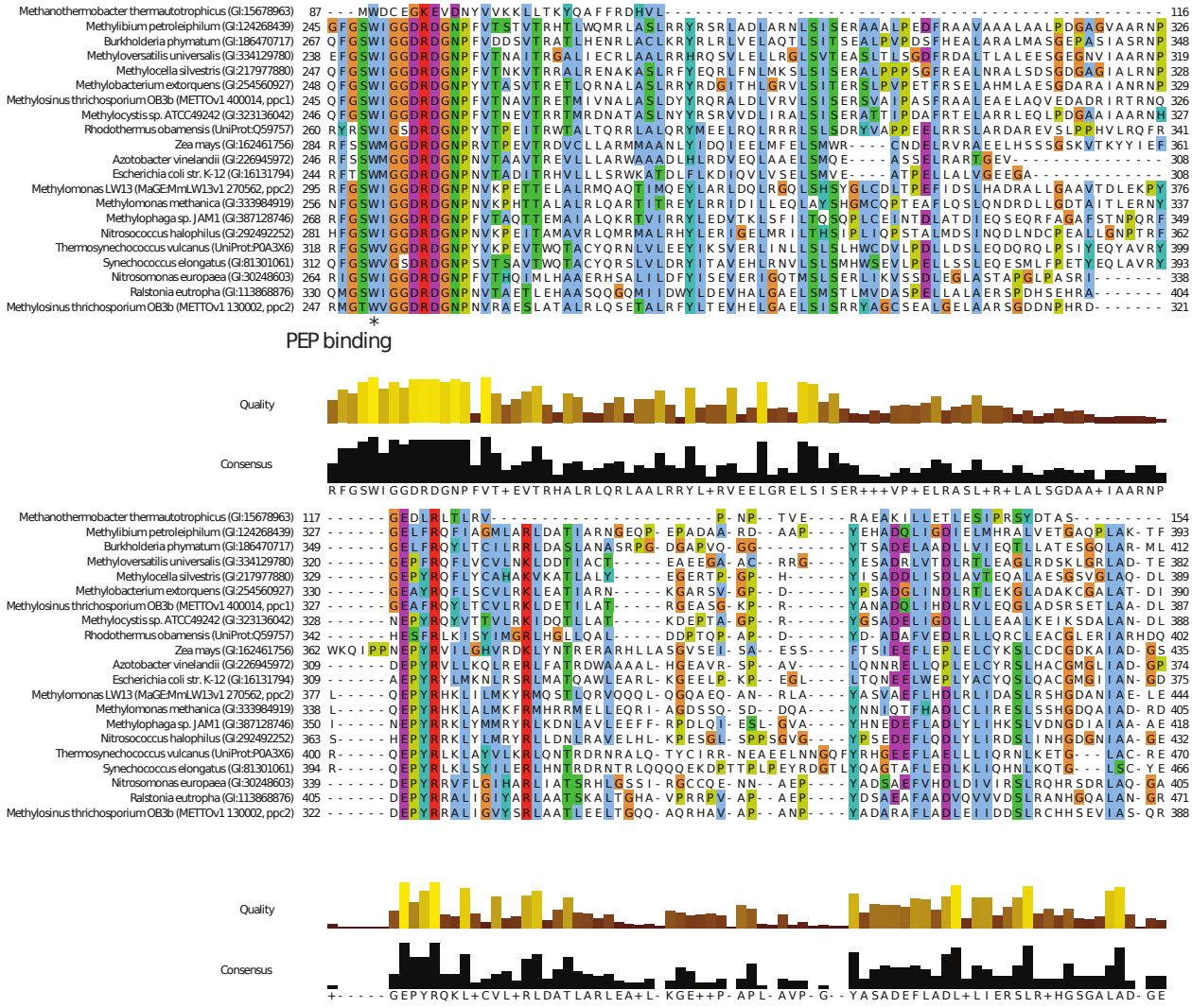
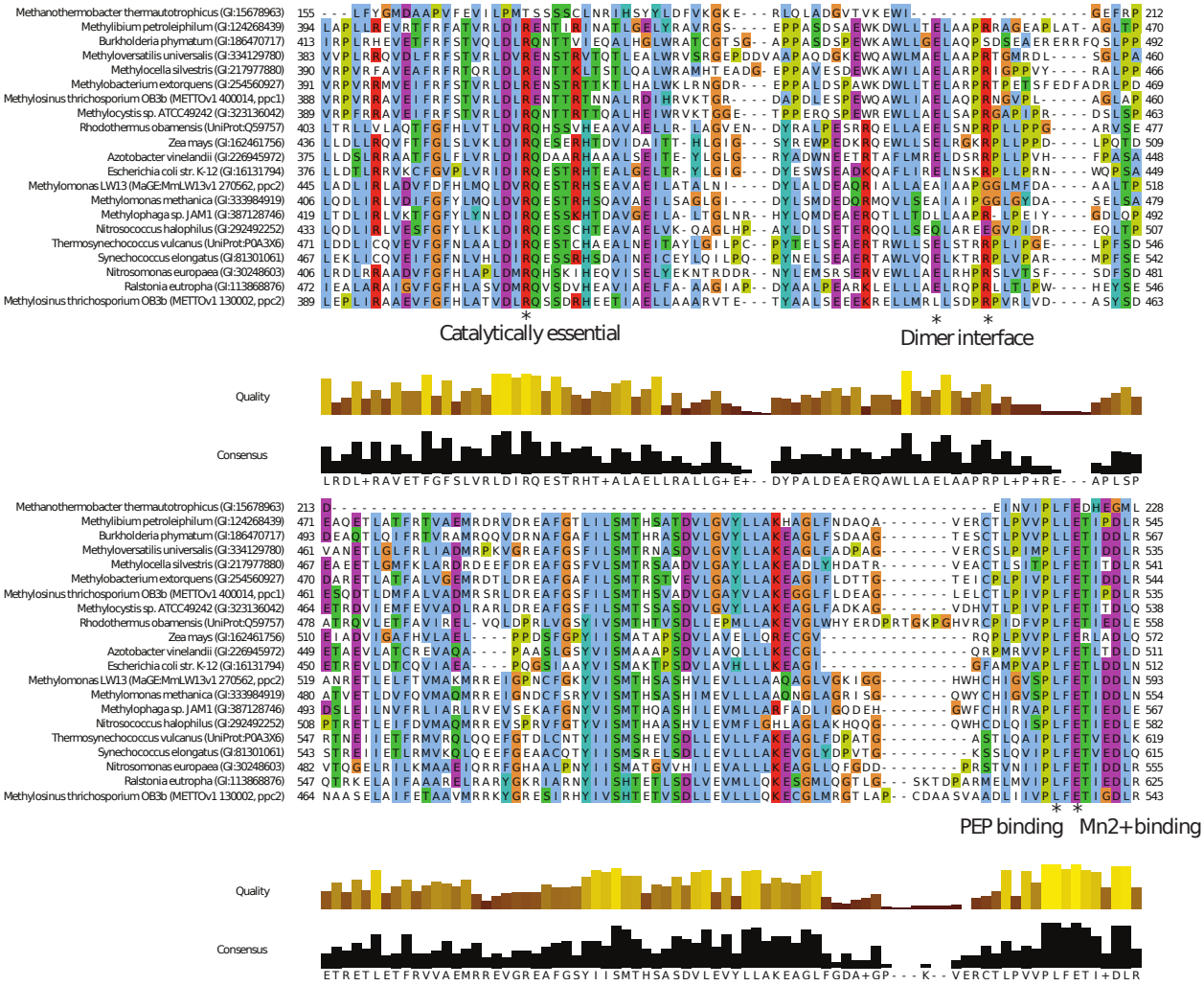
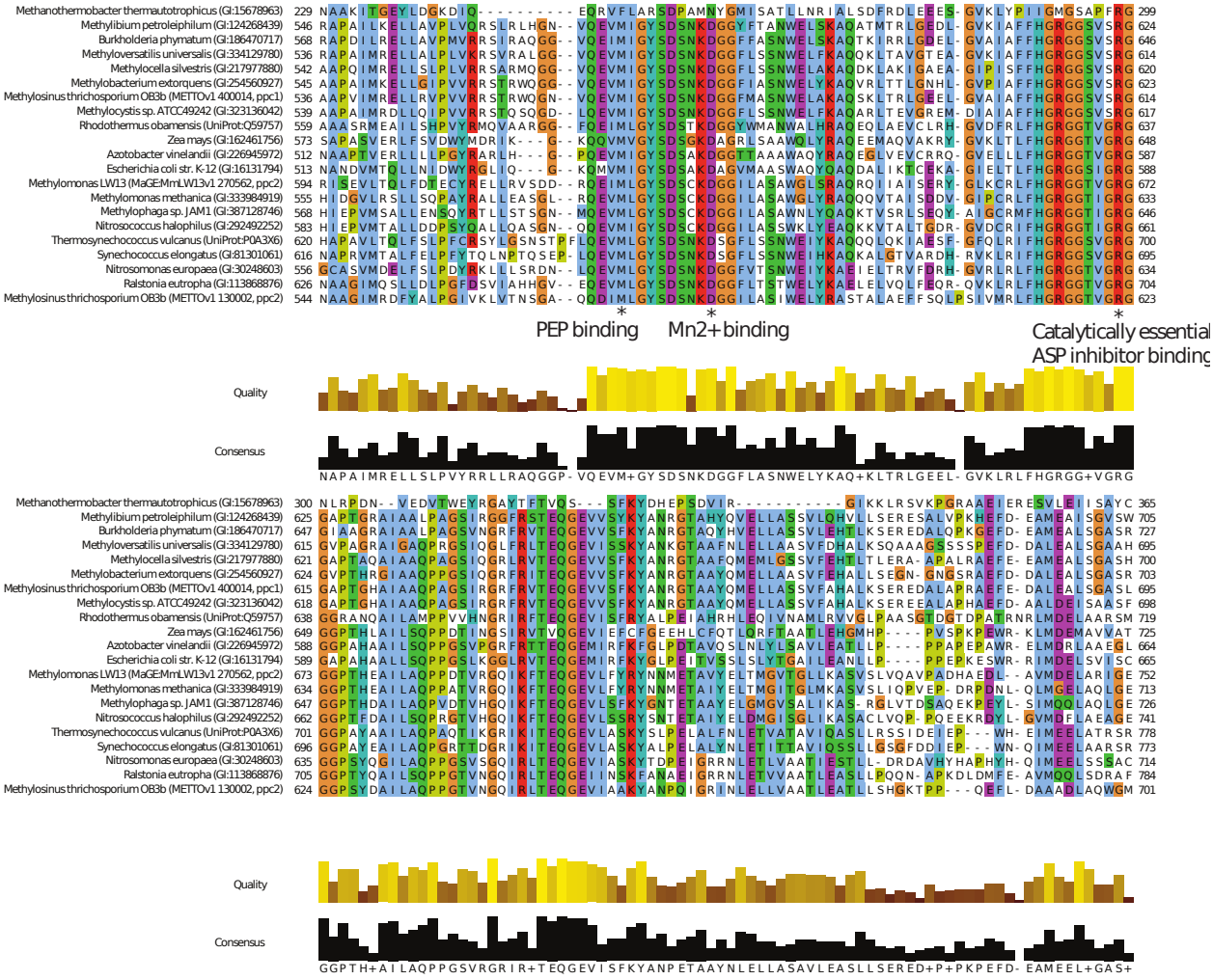
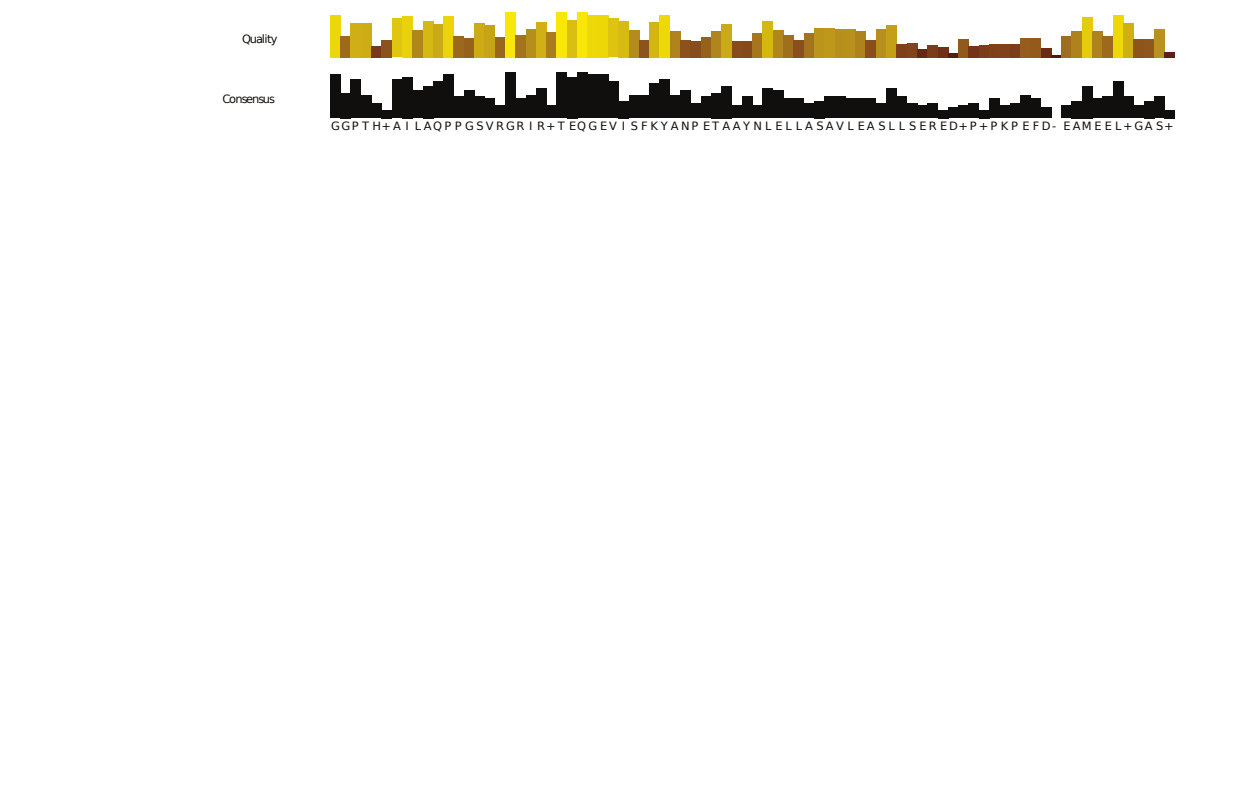


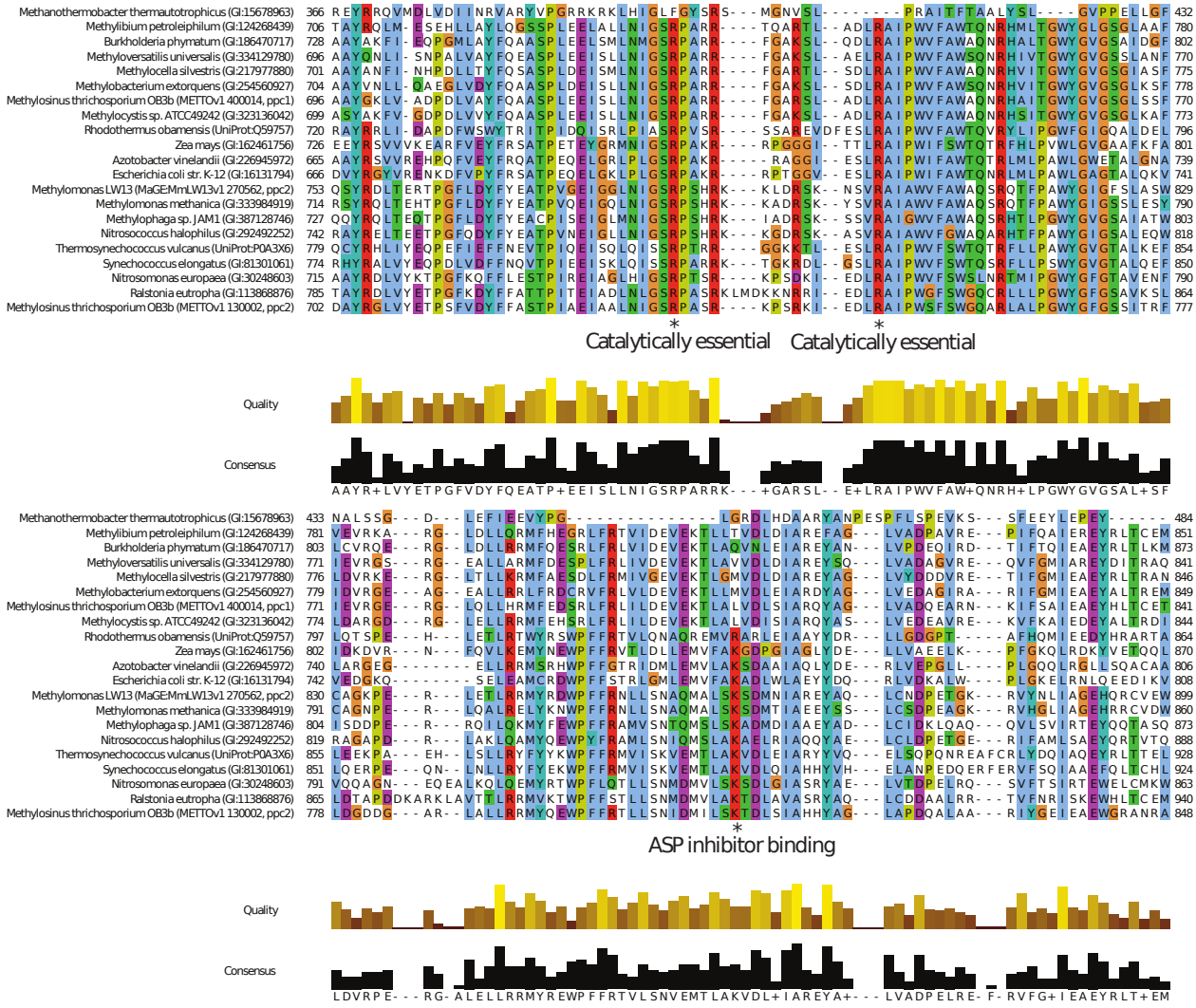
Figure A.6: Primary structure alignment for phosphoenolpyruvate carboxylases (Ppc1 and Ppc2) from *M. trichosporium OB3b* and Ppc-homologs (continued on subsequent pages).


PEP binding
Quality
Consensus
RFGSWIGGDRDGNPFVT+EVTRHALRLQRLAALRRYL+RVEELGRELSISER+++VP+ELRASL+R+LALSGDAA+IAARNP


Catalytically essential
Dimer interface
Quality
Consensus
PEP binding
Mn2+binding


PEP binding
Mn2+ binding
Catalytically essential
ASP inhibitor binding
Quality
Consensus
NAPAIMRELLSLPVYRRLLRAQGGP-VQEVM+GYSDSNKDGGFLASNWELYKAQ+KLTRLGEEL-GVKLRLFHGRGG+VGRG


Quality
Consensus
GGPTH+AILAQPPGSVRGRIR+TEQGEVISFKYANPETAAYNLELLASAVLEASLLSERED+P+PKPEFD-EAMEEL+GAS+


Catalytically essential
Catalytically essential
Quality
Consensus
ASP inhibitor binding
Quality
Consensus

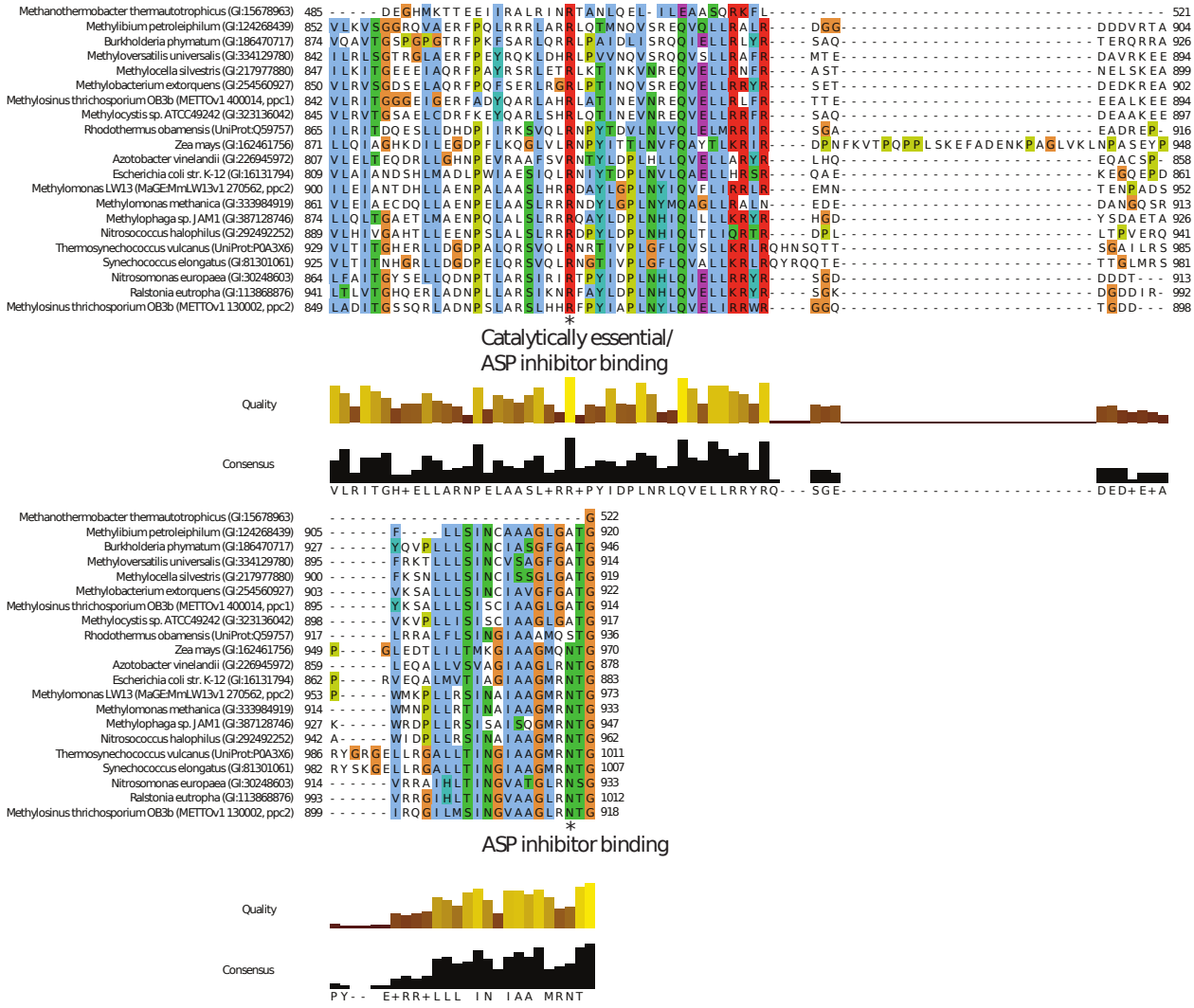

Catalytically essential/
ASP inhibitor binding
Quality
Consensus
ASP inhibitor binding
Quality
Consensus

Table A.1: Transcripts detected by *de novo* assembly RNA-seq data.

ID	Non-redundant	Description	Organism	Non-redundant length	Fraction of nr covered by alignment	Alignment % identity	Alignment e-value
8	319794846	hypothetical protein Varpa_4205	Variovorax paradoxus EPS	90	0.89	63	1.00E-23
48	323139488	SEC-C motif domain protein	Methylocystis sp. ATCC 49242	169	0.20	55	0.0004
48	323139488	SEC-C motif domain protein	Methylocystis sp. ATCC 49242	169	0.41	63	4.00E-18
90	307110660	hypothetical protein CHLNCDRAFT_50435	Chlorella variabilis	1535	0.05	24	4.5
90	307110660	hypothetical protein CHLNCDRAFT_50435	Chlorella variabilis	1535	0.05	24	4.6
120	309365749	hypothetical protein CBG_01089	Caenorhabditis briggsae AF16	812	0.12	32	0.7
121	310821854	hypothetical protein STAUR_4605	Stigmatella aurantiaca DW4/3-1	728	0.08	38	3.5
125	197103309	hypothetical protein PHZ_p0169	Phenylobacterium zucineum HLK1	223	0.55	60	8.00E-34
142	78213547	glycosyltransferase	Synechococcus sp. CC9605	1003	0.31	31	5.00E-14
142	325116388	tRNA uridine 5-carboxymethylaminomethyl modification enzyme GidA, related	Neospora caninum Liverpool	819	0.21	31	0.088
171	323136696	hypothetical protein Met49242DRAFT_1164	Methylocystis sp. ATCC 49242	92	0.64	53	6.00E-06
266	338532669	chloride channel	Myxococcus fulvus HW-1	628	0.72	63	3.00E-138
300	323136327	Patatin	Methylocystis sp. ATCC 49242	284	0.44	60	3.00E-38
303	227819692	transposase Y4ZB	Sinorhizobium fredii NGR234	493	0.35	42	8.00E-24
308	323449050	hypothetical protein AURANDRAFT_66784	Aureococcus anophagefferens	1445	0.48	25	3.00E-33
314	341615616	two-component response regulator	Citromicrobium sp. JLT1363	257	0.15	59	0.32
339	149918868	site-specific recombinase, phage integrase family protein	Plesiocystis pacifica SIR-1	120	0.53	36	7.9
360	46153	endo-glucanase	Ruminococcus flavefaciens	680	0.09	38	4.6
370	167901659	hypothetical protein BpseN_05228	Burkholderia pseudomallei NCTC 13177	474	0.32	29	4.9
381	203288753	BdrQ-like protein	Borrelia duttonii Ly	239	0.67	30	4.00E-05
401	323457221	hypothetical protein AURANDRAFT_70536	Aureococcus anophagefferens	1999	0.06	33	2.1
431	221486147	conserved hypothetical protein	Toxoplasma gondii GT1	702	0.16	30	5.9
469	16519794	transposase number 3 of uncharacterized insertion sequence	Sinorhizobium fredii NGR234	511	0.31	62	7.00E-51
518	323135994	2-oxoglutarate dehydrogenase, E2 subunit, dihydrolipoamide succinyltransferase	Methylocystis sp. ATCC 49242	410	0.10	68	0.0002

523	332821691	PREDICTED: hypothetical protein LOC100609280	Pan troglodytes	329	0.43	28	0.92
542	299133059	Pseudomurein-binding repeat protein	Afipia sp. 1NLS2	686	0.28	96	2.00E-102
542	299133059	Pseudomurein-binding repeat protein	Afipia sp. 1NLS2	686	0.17	94	3.00E-59
569	118340377	putative granule bound starch synthase	Sorghum bicolor	608	0.08	45	5.9
570	316932709	RNA-binding S4 domain-containing protein	Rhodopseudomonas palustris DX-1	717	0.06	41	4.7
570	325116236	putative retinitis pigmentosa GTPase regulator	Neospora caninum Liverpool	1413	0.04	43	0.41
582	255714473	KLTH0E00704p	Lachancea thermotolerans	1389	0.03	42	2.6
712	25029285	hypothetical protein CE2729	Corynebacterium efficiens YS-314	251	0.26	43	0.0003
721	88812339	succinate dehydrogenase catalytic subunit	Nitrococcus mobilis Nb-231	259	0.75	69	4.00E-72
737	222630362	hypothetical protein OsJ_17292	Oryza sativa Japonica Group	377	0.13	40	2
737	115462351	Os05g0171300	Oryza sativa Japonica Group	415	0.12	40	2
783	119174408	predicted protein	Coccidioides immitis RS	363	0.13	41	7.8
787	121710406	Ras GTPase activating protein, putative	Aspergillus clavatus NRRL 1	1671	0.04	35	5.9
788	323135945	molybdenum cofactor biosynthesis protein A	Methylocystis sp. ATCC 49242	346	0.09	87	1.00E-06
809	186510389	octicosapeptide/Phox/Bem1p domain-containing protein kinase	Arabidopsis thaliana	1117	0.09	30	2.7
810	83312534	hypothetical protein amb3435	Magnetospirillum magneticum AMB-1	86	0.38	61	0.029
810	83312534	hypothetical protein amb3435	Magnetospirillum magneticum AMB-1	86	0.42	58	0.006
823	154312862	hypothetical protein BC1G_05132	Botryotinia fuckeliana B05.10	315	0.13	44	7.9
829	304394601	conserved hypothetical protein	Ahrensia sp. R2A130	55	0.73	53	0.049
833	9631033	conotoxin-like protein	Lymantria dispar MNPV	92	0.68	38	7.9
844	298241184	conserved hypothetical protein	Ktedonobacter racemifer DSM 44963	204	0.16	55	2.1
895	198283398	XRE family transcriptional regulator	Acidithiobacillus ferrooxidans ATCC 53993	152	0.93	32	0.0004
909	315500368	transcriptional regulator, mucr family	Asticcacaulis excentricus CB 48	141	1.03	57	5.00E-36
929	23500301	IS66 family orf3	Brucella suis 1330	523	0.45	60	1.00E-70
929	288962634	transposase	Azospirillum sp. B510	420	0.29	63	6.00E-35
932	73980988	PREDICTED: similar to ALMS1	Canis familiaris	4146	0.01	42	7.9
933	307108552	expressed protein	Chlorella variabilis	838	0.18	27	0.41
954	299133061	DNA modification methyltransferase-related protein	Afipia sp. 1NLS2	917	0.16	94	8.00E-62
958	271965707	hypothetical protein Sros_4262	Streptosporangium roseum	238	0.20	40	6

			DSM 43021				
1074	323139155	replication protein C	Methylocystis sp. ATCC 49242	444	0.60	68	1.00E-94
1083	254502301	hypothetical protein SADFL11_2339	Labrenzia alexandrii DFL-11	993	0.12	83	3.00E-43
1097	85708226	putative DNA methylase	Erythrobacter sp. NAP1	909	0.66	76	0
1128	323139416	protein of unknown function DUF188	Methylocystis sp. ATCC 49242	152	0.34	85	9.00E-17
1128	323139416	protein of unknown function DUF188	Methylocystis sp. ATCC 49242	152	0.50	87	8.00E-18
1171	323139522	hypothetical protein Met49242DRAFT_3956	Methylocystis sp. ATCC 49242	444	0.60	91	2.00E-138
1200	299133058	Putative helicase A859L	Afipia sp. 1NLS2	408	0.48	81	1.00E-89
1250	224107413	PREDICTED: hypothetical protein, partial	Taeniopygia guttata	369	0.22	33	4.6
1259	225156592	conserved hypothetical protein	Opitutaceae bacterium TAV2	275	0.17	35	3.5
1290	298293273	succinate dehydrogenase, hydrophobic membrane anchor protein	Starkeya novella DSM 506	136	0.84	54	1.00E-24
1292	194761572	GF15723	Drosophila ananassae	514	0.14	37	4.5
1302	316934586	hypothetical protein Rpdx1_3258	Rhodopseudomonas palustris DX-1	253	0.27	72	1.00E-18
1310	154495963	hypothetical protein BACCAP_00246	Bacteroides capillosus ATCC 29799	166	0.23	49	4.5
1324	323137236	DNA gyrase, B subunit	Methylocystis sp. ATCC 49242	810	0.07	81	1.00E-15
1352	195442410	GK17748	Drosophila willistoni	421	0.10	45	5.9
1357	323137126	glycosyl transferase group 1	Methylocystis sp. ATCC 49242	408	0.14	70	2.00E-14
1369	323136028	major facilitator superfamily MFS_1	Methylocystis sp. ATCC 49242	528	0.15	65	1.00E-20
1374	90420316	phosphomethylpyrimidine kinase	Aurantimonas manganoxydans SI85-9A1	317	0.06	75	5.9
1436	238023318	hypothetical protein bglu_2p0290	Burkholderia glumae BGR1	476	0.29	36	5.00E-13
1498	330966470	hypothetical protein PSYAC_17830	Pseudomonas syringae pv. actinidiae str. M302091	437	0.10	47	0.0009
1615	338820076	hypothetical protein Agau_P200245	Agrobacterium tumefaciens F2	783	0.11	47	6.00E-14
1691	332560244	Mg chelatase-related protein	Rhodobacter sphaeroides WS8N	512	0.11	38	4.5
1734	323137126	glycosyl transferase group 1	Methylocystis sp. ATCC 49242	408	0.24	77	9.00E-39
1787	299133059	Pseudomurein-binding repeat protein	Afipia sp. 1NLS2	686	0.13	99	9.00E-39
1787	299133059	Pseudomurein-binding repeat protein	Afipia sp. 1NLS2	686	0.09	98	9.00E-25
1793	323137126	glycosyl transferase group 1	Methylocystis sp. ATCC 49242	408	0.08	88	6.00E-08
1793	323137126	glycosyl transferase group 1	Methylocystis sp. ATCC 49242	408	0.08	88	6.00E-08
1836	150378510	tyrosine-protein phosphatase non-receptor type 22	Danio rerio	887	0.05	34	7.6
1856	17232905	hypothetical protein alr8532	Nostoc sp. PCC 7120	304	0.25	48	4.00E-11
			Roseomonas cervicalis				

1922	296534676	conserved hypothetical protein	ATCC 49957	206	0.30	46	0.41
1947	163857359	LysR family transcriptional regulator	Bordetella petrii DSM 12804	295	0.15	68	7.00E-09
1953	332708113	penicillin-binding protein, family 1A	Lyngbya majuscula 3L	819	0.08	34	0.91
1961	323139158	integrase domain protein SAM domain protein	Methylocystis sp. ATCC 49242	347	0.10	74	2.00E-08
1970	154245910	succinate dehydrogenase cytochrome b556 subunit	Xanthobacter autotrophicus Py2	133	0.95	57	3.00E-32
1970	110635572	succinate dehydrogenase iron-sulfur subunit	Mesorhizobium sp. BNC1	259	0.25	88	2.00E-26
1976	197103308	hypothetical protein PHZ_p0168	Phenylobacterium zucineum HLK1	323	0.14	45	0.017
2036	291449786	LOW QUALITY PROTEIN: FscRII	Streptomyces albus J1074	880	0.11	36	3.5
2044	194733871	antirestriction protein	Salmonella enterica subsp. enterica serovar Schwarzengrund str. CVM19633	321	0.17	49	6.00E-05
2048	256377452	AraC family transcriptional regulator	Actinosynnema mirum DSM 43827	309	0.12	61	0.029
2048	256377452	AraC family transcriptional regulator	Actinosynnema mirum DSM 43827	309	0.12	61	0.029
2082	302539530	polyketide synthase type I	Streptomyces sp. C	1616	0.03	43	0.71
2087	299133058	Putative helicase A859L	Afipia sp. 1NLS2	408	0.11	82	5.00E-13
2108	159489659	predicted protein	Chlamydomonas reinhardtii	439	0.08	51	7.8
2149	296534676	conserved hypothetical protein	Roseomonas cervicalis ATCC 49957	206	0.32	38	4.6
2151	300865706	conserved hypothetical protein	Oscillatoria sp. PCC 6506	138	0.26	42	7.8
2167	209544899	hypothetical protein Gdia_2780	Gluconacetobacter diazotrophicus PAl 5	397	0.09	97	5.00E-05
2176	323138303	peptidase U62 modulator of DNA gyrase	Methylocystis sp. ATCC 49242	514	0.28	33	0.0002
2236	220922655	hypothetical protein Mnod_2696	Methylobacterium nodulans ORS 2060	173	0.23	75	6.00E-08
2392	316934586	hypothetical protein Rpdx1_3258	Rhodopseudomonas palustris DX-1	253	0.08	84	2.1
2503	323139732	putative transposase	Methylocystis sp. ATCC 49242	150	0.11	88	1.2
2606	159044846	hypothetical protein Dshi_2303	Dinoroseobacter shibae DFL 12	432	0.08	57	0.71
2614	258515430	hypothetical protein Dtox_2211	Desulfotomaculum acetoxidans DSM 771	377	0.10	74	3.00E-08
2641	125524722	hypothetical protein OsI_00707	Oryza sativa Indica Group	532	0.08	44	2.7
2690	222871727	predicted protein	Populus trichocarpa	398	0.10	46	6
2750	323136026	secretion protein HlyD family protein	Methylocystis sp. ATCC 49242	397	0.15	81	7.00E-20

2832	329889268	phosphotransferase domain-containing protein	Brevundimonas diminuta ATCC 11568	1279	0.03	54	0.01
2854	338778752	glutathione S-transferase	Achromobacter xylosoxidans AXX-A	201	0.22	73	5.00E-10
2901	323139705	hypothetical protein Met49242DRAFT_4133	Methylocystis sp. ATCC 49242	253	0.31	48	1.00E-11
2903	28897417	hypothetical protein VP0643	Vibrio parahaemolyticus RIMD 2210633	599	0.05	45	10
2975	323137126	glycosyl transferase group 1	Methylocystis sp. ATCC 49242	408	0.09	89	3.00E-11
2983	297620720	nitric oxide reductase, subunit B	Waddlia chondrophila WSU 86-1044	763	0.04	85	6.00E-10
3047	340030375	helix-turn-helix domain-containing protein	Paracoccus sp. TRP	81	0.64	60	6.00E-08
3093	302134922	hypothetical protein PsyrptN_26257	Pseudomonas syringae pv. tomato NCPPB 1108	988	0.04	53	0.92
3111	329889268	phosphotransferase domain-containing protein	Brevundimonas diminuta ATCC 11568	1279	0.03	51	0.022
3112	323139523	hypothetical protein Met49242DRAFT_3957	Methylocystis sp. ATCC 49242	482	0.09	88	7.00E-06
3123	297292023	PREDICTED: solute carrier family 22 member 3-like	Macaca mulatta	715	0.06	44	7.8
3128	334195582	putative osmotically inducible lipoprotein b1 transmembrane (osmB)	Ralstonia solanacearum Po82	296	0.17	45	3.5
3130	254500001	hypothetical protein SADFL11_39	Labrenzia alexandrii DFL-11	356	0.20	54	5.00E-12
3130	254500001	hypothetical protein SADFL11_39	Labrenzia alexandrii DFL-11	356	0.19	56	1.00E-11
3162	301780690	PREDICTED: lysine-specific demethylase 4D-like	Ailuropoda melanoleuca	487	0.08	49	7.8
3169	296123721	hypothetical protein Plim_3487	Planctomyces limnophilus DSM 3776	438	0.07	47	7.8
3190	149918280	RNA methyltransferase	Plesiocystis pacifica SIR-1	485	0.12	40	4.6
3190	149918280	RNA methyltransferase	Plesiocystis pacifica SIR-1	485	0.12	40	4.6
3191	91200065	conserved hypothetical protein	Candidatus Kuenenia stuttgartiensis	221	0.24	46	0.11
3192	302813543	hypothetical protein SELMODRAFT_427133	Selaginella moellendorffii	270	0.15	48	0.7
3236	333027859	putative ADP-ribosylation/Crystallin J1	Streptomyces sp. Tu6071	893	0.06	44	7.6
3258	110347006	hypothetical protein Meso_4194	Mesorhizobium sp. BNC1	366	0.16	62	4.00E-12
3315	288963127	transposase	Azospirillum sp. B510	267	0.25	67	2.00E-16
3340	3059133	transposase IS1355	Methylobacterium extorquens DM4	179	0.22	87	4.00E-12
3398	83592164	CRISPR-associated helicase Cas3 family protein	Rhodospirillum rubrum ATCC 11170	752	0.05	69	5.00E-05
3602	323137054	hypothetical protein Met49242DRAFT_1521	Methylocystis sp. ATCC 49242	73	0.59	77	4.00E-09
3602	197103644	RNA polymerase sigma 54 subunit, RpoN	Phenylobacterium zucineum HLK1	499	0.11	44	2.7

3605	218667529	hypothetical protein AFE_2089	Acidithiobacillus ferrooxidans ATCC 23270	253	0.42	29	4.5
3656	121596846	RNA polymerase sigma factor	Burkholderia mallei SAVP1	231	0.26	51	3.00E-08
3677	299133058	Putative helicase A859L	Afipia sp. 1NLS2	408	0.27	85	2.00E-28
3742	90424677	plasmid maintenance system killer protein	Rhodopseudomonas palustris BisB18	93	0.31	41	1.4
3786	229828371	hypothetical protein GCWU000342_00430	Shuttleworthia satelles DSM 14600	298	0.15	40	7.9
3786	229828371	hypothetical protein GCWU000342_00430	Shuttleworthia satelles DSM 14600	298	0.15	40	7.7
3891	220922092	NnrS family protein	Methylobacterium nodulans ORS 2060	392	0.09	68	2.00E-05
3901	221487851	conserved hypothetical protein	Toxoplasma gondii GT1	3229	0.02	39	4.5
3981	85711049	Predicted metal-dependent amidohydrolase with the TIM-barrel fold	Idiomarina baltica OS145	558	0.23	26	3.5
3981	89067955	type I secretion target repeat protein	Oceanicola granulosus HTCC2516	818	0.17	28	4.2
4000	323136327	Patatin	Methylocystis sp. ATCC 49242	284	0.44	61	3.00E-37
4020	328541621	hypothetical protein SL003B_p0053	Polymorphum gilvum SL003B-26A1	435	0.43	58	2.00E-56
4035	108763646	putative lipoprotein	Myxococcus xanthus DK 1622	447	0.13	35	7.8
4187	288957870	transposase	Azospirillum sp. B510	460	0.09	43	5.9
4195	335035794	hypothetical protein AGRO_3125	Agrobacterium sp. ATCC 31749	323	0.09	62	0.55
4205	90424421	transposase IS3/IS911	Rhodopseudomonas palustris BisB18	66	0.38	68	1.6
4305	329928833	hypothetical protein HMPREF9412_3689	Paenibacillus sp. HGF5	694	0.05	94	1.00E-10
4311	167621675	IS66 Orf2 family protein	Caulobacter sp. K31	117	0.35	59	3.00E-05
4338	209886756	hypothetical protein OCAR_7650	Oligotropha carboxidovorans OM5	362	0.10	64	2.00E-05
4381	284989640	hypothetical protein Gobs_1064	Geodermatophilus obscurus DSM 43160	285	0.11	59	2.7
4492	13475179	hypothetical protein mlr6199	Mesorhizobium loti MAFF303099	240	0.18	72	1.00E-09
4538	238061000	5'-nucleotidase domain-containing protein	Micromonospora sp. ATCC 39149	616	0.13	35	7.4
4646	302522431	LOW QUALITY PROTEIN: ABC transporter ATP-binding protein	Streptomyces sp. SPB78	564	0.15	30	4.5
4866	299133061	DNA modification methyltransferase-related protein	Afipia sp. 1NLS2	917	0.14	90	3.00E-58
4903	299741957	endoprotease	Coprinopsis cinerea okayama7#130	613	0.07	44	4.5
4908	323139523	hypothetical protein Met49242DRAFT_3957	Methylocystis sp. ATCC 49242	482	0.08	97	2.00E-14
4992	323135620	hypothetical protein Met49242DRAFT_0090	Methylocystis sp. ATCC 49242	153	0.23	91	2.00E-11

5003	299133061	DNA modification methyltransferase-related protein	Afipia sp. 1NLS2	917	0.04	95	6.00E-14
5069	316934586	hypothetical protein Rpdx1_3258	Rhodopseudomonas palustris DX-1	253	0.39	88	1.00E-44
5103	312883142	hypothetical protein VIBC2010_07634	Vibrio caribbenthicus ATCC BAA-2122	138	0.85	38	1.00E-14
5192	16125845	sulfate ABC transporter ATP-binding protein	Caulobacter crescentus CB15	359	0.12	55	0.007
5192	86739180	ABC transporter-like protein	Frankia sp. CcI3	385	0.13	51	0.017
5210	323139955	transposase IS204/IS1001/IS1096/IS1165 family protein	Methylocystis sp. ATCC 49242	549	0.05	54	4.7
5248	170743671	putative transposase	Methylobacterium sp. 4-46	373	0.08	71	0.001
5420	323139641	protein of unknown function DUF1403	Methylocystis sp. ATCC 49242	345	0.10	61	0.9

Table A.2: Gene expression profile in methane-grown cells of *M. trichosporium OB3b*. Values represent reads per kilobase of coding sequence per million (reads) mapped (RPKM). For the full 4,812-row dataset, see attached file.

GENE ID/LOCUS TAG	PUTATIVE FUNCTION	BIOLOGICAL REPLICATE 1 (RPKM)	BIOLOGICAL REPLICATE 2 (RPKM)
METTOv1_1180001	Protein of unknown function	444193	502003
METTOv1_1180006	Conserved protein of unknown function	114962	124378
METTOv1_1270004	Methane monooxygenase subunit PmoC	66689	67703
METTOv1_1270003	Protein of unknown function	56337	59538
METTOv1_1270002	Methane monooxygenase subunit PmoA	37102	31813
METTOv1_1270001	Methane monooxygenase subunit PmoB	27371	22917
METTOv1_720011	Exported protein of unknown function	25925	34114
METTOv1_240011	Methanol dehydrogenase beta subunit, MxaI	24552	28474
METTOv1_40013	Formaldehyde-activating enzyme, Fae	24353	24787
METTOv1_80041	Protein of unknown function	15629	17876
METTOv1_10085	Putative Flp/Fap pilin component (modular protein)	15210	17950
METTOv1_220005	Conserved protein of unknown function	13342	12873
METTOv1_100043	Flavodoxin FldA	12939	14717
METTOv1_10086	Putative Flp/Fap pilin component (modular protein)	12500	14508
METTOv1_pqqA	PqqA	11857	13927
METTOv1_350034	Exported protein of unknown function	10937	11059
METTOv1_10178	Conserved protein of unknown function	10911	11543
METTOv1_240014	PQQ-dependent methanol dehydrogenase, MxaF	9742	8367
METTOv1_360031	10 kDa chaperonin	8505	8095
METTOv1_50029	Conserved exported protein of unknown function	7799	8459
METTOv1_210021	Putative 31 kDa outer-membrane immunogenic protein precursor	7755	7718
METTOv1_50046	Antibiotic biosynthesis monooxygenase	7561	8400
METTOv1_220036	Protein of unknown function	6126	7735
METTOv1_380025	cold-shock DNA-binding domain protein	6056	7763
METTOv1_360030	60 kDa chaperonin	6040	5426
METTOv1_240012	Cytochrome c class I	5712	6117
METTOv1_240015	Protein of unknown function	5571	5393
METTOv1_250008	Protein of unknown function DUF465	5160	7729
METTOv1_370032	Protein of unknown function	5110	7733
METTOv1_310036	H+transporting two-sector ATPase C subunit	4898	4882
METTOv1_10101	50S ribosomal protein L7/L12	4728	4670
METTOv1_670020	Conserved protein of unknown function	4636	5951
METTOv1_90056	Putative outer-membrane immunogenic protein precursor	4497	4098
METTOv1_870006	Conserved protein of unknown function, putative MbtH-like protein	4176	5481
METTOv1_40014	Formaldehyde-activating enzyme, Fae	4024	3676
METTOv1_20142	50S ribosomal protein L34	3876	3678
METTOv1_1180005	Protein of unknown function	3865	4053
METTOv1_10102	50S ribosomal protein L10	3846	3875
METTOv1_50044	TonB-dependent heme/hemoglobin receptor family protein	3483	3631
METTOv1_60112	Protein of unknown function	3418	4716
METTOv1_680015	Cytochrome c class I	3331	3695
METTOv1_140005	Cold-shock DNA-binding domain protein	3308	3283
METTOv1_1590001	Elongation factor Tu (fragment)	3168	2796

METTOv1_150018	Conserved membrane protein of unknown function	3133	3356
METTOv1_140012	Protein of unknown function	3097	4163
METTOv1_270053	Putative outer membrane TonB-dependent receptor; putative receptor for iron transport	3058	2845
METTOv1_390020	Protein of unknown function	3058	3677
METTOv1_10003	Conserved exported protein of unknown function	3045	3088
METTOv1_60110	Flagellar hook protein FlgE	2996	3024
METTOv1_50060	Conserved exported protein of unknown function	2980	3047
METTOv1_50043	Protein of unknown function	2975	3223
METTOv1_340043	TonB-dependent siderophore receptor	2962	2659
METTOv1_10054	protein of unknown function	2885	3524
METTOv1_40026	protein of unknown function DUF88	2850	3035
METTOv1_80110	30S ribosomal protein S12	2744	2578
METTOv1_440010	ribosomal protein L29	2742	2676
METTOv1_130050	nitrogen regulatory protein P-II	2650	2735
METTOv1_710010	conserved protein of unknown function	2639	2582
METTOv1_740011	30S ribosomal protein S15	2582	2802
METTOv1_10009	Acyl carrier protein	2534	3554
METTOv1_840014	Histone family protein DNA-binding protein (fragment)	2484	2644
METTOv1_200048	Glutamine synthetase, type I	2452	2321
METTOv1_200047	Nitrogen regulatory protein P-II	2400	2277
METTOv1_440018	50S ribosomal protein L18	2297	2157
METTOv1_760004	conserved protein of unknown function	2255	2676
METTOv1_210044	exported protein of unknown function	2247	2538
METTOv1_440001	30S ribosomal protein S10	2232	2089
METTOv1_440004	Ribosomal protein L25/L23	2232	2042
METTOv1_760010	TonB-dependent receptor	2197	2174
METTOv1_250051	50S ribosomal protein L35	2195	2579
METTOv1_50068	30S ribosomal protein S13	2173	2167
METTOv1_440020	50S ribosomal protein L30	2157	1946
METTOv1_350038	50S ribosomal protein L33	2134	3246
METTOv1_60031	protein of unknown function	2120	3016
METTOv1_510015	30S ribosomal protein S1	2113	1995
METTOv1_440011	30S ribosomal protein S17	2053	2020
METTOv1_10124	30S ribosomal protein S6	2031	2505
METTOv1_130049	Ammonium transporter	2029	1766
METTOv1_430003	50S ribosomal protein L28	2025	1976
METTOv1_60030	Exported protein of unknown function	2023	2047
METTOv1_440006	30S ribosomal protein S19	2020	1922
METTOv1_20092	protein of unknown function	2013	2471
METTOv1_30061	protein of unknown function	1998	2438
METTOv1_680005	30S ribosomal protein S20	1966	1972
METTOv1_600007	Flagellin domain protein	1948	1994
METTOv1_240013	Extracellular solute-binding protein family 3	1942	1838
METTOv1_440009	50S ribosomal protein L16	1941	1692
METTOv1_50071	50S ribosomal protein L17	1893	1987
METTOv1_280037	glutaredoxin-like protein	1889	2282
METTOv1_440008	30S ribosomal protein S3	1880	1812
METTOv1_440017	50S ribosomal protein L6	1865	1673
METTOv1_180065	ATP synthase subunit alpha	1847	1622
METTOv1_400020	Serine-glyoxylate transaminase, Sga	1840	1969
METTOv1_360023	Exported protein of unknown function	1800	1818
METTOv1_440014	ribosomal protein L5	1790	1620

METTOv1_10173	30S ribosomal protein S4	1789	1786
METTOv1_420032	Glutathione peroxidase	1789	2445
METTOv1_220052	Bacterioferritin	1787	1970
METTOv1_240028	OmpW family protein	1767	1765
METTOv1_310020	Nitrite reductase (NAD(P)H), large subunit	1762	1562
METTOv1_440013	50S ribosomal protein L24	1750	1504
METTOv1_370004	Ferric uptake regulator, Fur family	1745	1828
METTOv1_310023	Major facilitator superfamily MFS_1	1743	1712
METTOv1_60068	Protein of unknown function UPF0005	1722	1670
METTOv1_400013	Malyl-CoA lyase/beta-methylmalyl-CoA lyase	1713	1615
METTOv1_50038	50S ribosomal protein L21	1691	1716
METTOv1_100044	protein of unknown function	1682	1492
METTOv1_140057	Nucleoside diphosphate kinase	1682	1500
METTOv1_440007	50S ribosomal protein L22	1674	1777
METTOv1_50037	50S ribosomal protein L27	1664	1794
METTOv1_710011	protein of unknown function	1655	1686
METTOv1_80112	Elongation factor G	1651	1691
METTOv1_10125	30S ribosomal protein S18	1647	1498
METTOv1_240035	50S ribosomal protein L25	1645	1617
METTOv1_340026	50S ribosomal protein L9	1638	1655
METTOv1_250050	50S ribosomal protein L20	1621	1595
METTOv1_80111	30S ribosomal protein S7	1618	1626
METTOv1_590017	50S ribosomal protein L11	1611	1524
METTOv1_440016	30S ribosomal protein S8	1602	1412
METTOv1_440005	50S ribosomal protein L2	1599	1392
METTOv1_10051	protein of unknown function	1598	2118
METTOv1_440012	50S ribosomal protein L14	1597	1473
METTOv1_410012	conserved exported protein of unknown function	1591	1734
METTOv1_80113	protein of unknown function	1582	1483
METTOv1_20091	globin	1580	1678
METTOv1_1710001	protein of unknown function	1577	1592
METTOv1_150057	50S ribosomal protein L31	1566	1767
METTOv1_140043	Outer membrane protein	1563	1450
METTOv1_870005	Thioesterase	1542	1473
METTOv1_300010	protein of unknown function	1538	1348
METTOv1_740013	conserved protein of unknown function	1528	1631
METTOv1_780012	nucleotide sugar dehydrogenase	1517	1399
METTOv1_440021	50S ribosomal protein L15	1512	1252
METTOv1_180063	F0F1 ATP synthase subunit beta	1478	1280
METTOv1_440015	30S ribosomal protein S14	1470	1254
METTOv1_60038	ribosomal protein L13	1461	1314
METTOv1_380017	Inorganic diphosphatase	1459	1471
METTOv1_CDS422275	methanobactin precursor	1439	2177
METTOv1_340044	conserved membrane protein of unknown function	1424	1254
METTOv1_200046	Glutamate--ammonia ligase	1413	1375
METTOv1_60075	Alkyl hydroperoxide reductase subunit C (Peroxiredoxin) (Thioredoxin peroxidase) (Alkyl hydroperoxide reductase protein C22) (SC	1411	1460
METTOv1_440019	30S ribosomal protein S5	1411	1175
METTOv1_660011	L-ornithine 5-monooxygenase (L-ornithine N5-oxygenase)	1410	1450
METTOv1_310035	H+transporting two-sector ATPase B/B' subunit	1405	1281
METTOv1_50069	30S ribosomal protein S11	1391	1375
METTOv1_40095	protein of unknown function	1387	1583
METTOv1_180064	F0F1 ATP synthase subunit gamma	1380	1236

METTOv1_200029	exported protein of unknown function	1375	1398
METTOv1_670002	phasin	1371	1276
METTOv1_310031	translation initiation factor IF-3	1368	1564
METTOv1_440003	50S ribosomal protein L4	1362	1174
METTOv1_310034	H+transporting two-sector ATPase B/B' subunit	1358	1362
METTOv1_180066	ATP synthase subunit delta	1355	1304
METTOv1_670019	serine hydroxymethyltransferase	1355	1209
METTOv1_50045	protein of unknown function	1352	1360
METTOv1_400031	transcriptional regulator, MucR family	1347	1461
METTOv1_60105	flagellar basal body rod modification protein	1346	1844
METTOv1_1440001	ATP-dependent metalloprotease FtsH (fragment)	1342	1308
METTOv1_60091	conserved exported protein of unknown function	1316	1398
METTOv1_220011	hemimethylated DNA binding protein	1301	1489
METTOv1_60092	flagellin domain protein	1299	1246
METTOv1_310045	phosphoribosylaminoimidazole-succinocarboxamide synthase	1289	1379
METTOv1_150031	exported protein of unknown function	1272	1162
METTOv1_760008	Sigma-24 (FecI-like) (modular protein)	1260	1522
METTOv1_100063	protein of unknown function	1259	1239
METTOv1_590016	50S ribosomal protein L1	1257	1187
METTOv1_280036	protein of unknown function	1252	1359
METTOv1_160008	putative outer-membrane immunogenic protein precursor	1248	1096
METTOv1_270004	heat shock protein Hsp20	1243	1377
METTOv1_310017	putative Alpha amylase, catalytic subdomain	1242	1235
METTOv1_550007	cytochrome c oxidase, subunit II	1241	1186
METTOv1_970002	30S ribosomal protein S21	1236	1515
METTOv1_10053	protein of unknown function	1226	1255
METTOv1_510003	succinyl-CoA synthetase, alpha subunit	1198	1135
METTOv1_310037	F0F1 ATP synthase subunit A	1197	1135
METTOv1_1110003	Electron transfer flavoprotein alpha/beta-subunit	1192	1075
METTOv1_60037	30S ribosomal protein S9	1189	1026
METTOv1_50070	DNA-directed RNA polymerase subunit alpha	1180	1066
METTOv1_550012	cytochrome c oxidase subunit III	1179	1219
METTOv1_80070	50S ribosomal protein L32	1170	1279
METTOv1_180031	outer membrane protein assembly complex, YaeT protein	1164	1083
METTOv1_550008	cytochrome c oxidase, subunit I	1164	1012
METTOv1_100080	acetoacetyl-CoA reductase	1160	1060
METTOv1_590019	preprotein translocase, SecE subunit	1159	1071
METTOv1_220033	DNA-binding protein HU-beta (NS1) (HU-1)	1155	1046
METTOv1_1340001	ATP-dependent metalloprotease FtsH (fragment)	1139	1260
METTOv1_240004	MxaD protein, putative	1137	1077
METTOv1_80082	50S ribosomal protein L19	1133	1068
METTOv1_1510001	Amino acid adenylation domain protein (fragment)	1121	1172
METTOv1_90054	30S ribosomal protein S2	1105	994
METTOv1_70096	protein of unknown function	1104	891
METTOv1_310038	conserved protein of unknown function	1100	1071
METTOv1_200019	Endopeptidase Clp (modular protein)	1098	1231
METTOv1_50033	PpiC-type peptidyl-prolyl cis-trans isomerase	1090	1137
METTOv1_380039	Integration host factor subunit beta	1080	1053
METTOv1_510021	exported protein of unknown function	1071	1054
METTOv1_10061	protein of unknown function	1067	1274
METTOv1_130056	methionine adenosyltransferase 1 (AdoMet synthetase)	1059	1050
METTOv1_220024	conserved exported protein of unknown function	1052	1076
METTOv1_440002	50S ribosomal protein L3	1045	948

Table A.3: Summary of putative transcription site mapping

Strain	Gene	Putative TS*	-35	-10
M. trichosporium OB3b	*pmoC*	-273	**TTG**T**CA**	A**A**TGGT**T**G
	pmoC	-324	**TTG**TGC	TCCTACCC TA
	mxaF	-156	CA**GACA**	**TATA**TG
	fae1	-214	**TT**CGAT	**TATAA**T
	pqqA	-90	**TTG**CAC	GC**TAAT**
Methylocystis sp. M[1]	*pmoC*		**TTG**T**CA**	GA**AT**GG**T**T G
M.extorquens AM1[2]	*mxaF*		AA**GACA**	**TA**G**AA**A
	mxaW		**TTG**GCA	ACCCAT
E. coli [3]	s70		**TTGACA**	**TATAAT**
	s32		TCTCNCCCTTGAA	
	s54		CTGGNA	TTGCA
	s28		CTAAA	CCGATAT

1 Holmes *et al.*, 1995 (see Ref)
2. Zhang, M., and Lidstrom, M.E. (2003) Microbiol 149:1033-40.
3. Harley, C.B., and Reynolds, R.P. 1987, Nucleic Acids Res. 11:2343-61

Table A.4: Summary of RNA-seq (Illumina) reads.

	Reads Mapped		% of Reads Mapped	
	Replicate 1	Replicate 2	Replicate 1	Replicate 2
Coding sequences	$3.652*10^6$	$1.546*10^6$	13.3	5.44
Noncoding sequences	$2.935*10^5$	$1.396.*10^5$	1.07	0.49
rRNA	$2.359*10^7$	$2.671*10^7$	85.6	94.0
tRNA	$7.937*10^3$	$4.095*10^3$	0.029	0.0014

Table A.5: Genes removed from reference scaffold before alignment.

Locus tag	Comment
METTOv1_1600001	chimeric sequence/ METTOv1_870001 and METTOv1_870003
METTOv1_820016	protein of unknown function
METTOv1_150030	protein of unknown function
METTOv1_300034	protein of unknown function/gltD operon
METTOv1_1650001	chimeric sequence/part of METATv1_30343 and METATv1_30340
METTOv1_140042	chimeric protein/part of METTOv1_140043
METTOv1_1700001	chimeric sequence/METTOv1_870001 and METTOv1_870003
METTOv1_10052	chimeric protein/part of METTOv1_10051
METTOv1_200054	chimeric protein/part of METTOv1_200053
METTOv1_600024	exported protein of unknown function
METTOv1_370034	protein of unknown function
METTOv1_200053	protein of unknown function
METTOv1_720016	protein of unknown function
METTOv1_450003	protein of unknown function
METTOv1_590015	protein of unknown function
METTOv1_360021	protein of unknown function
METTOv1_800023	protein of unknown function//part of transposase IS3/IS911
METTOv1_760015	protein of unknown function
METTOv1_270007	protein of unknown function
METTOv1_210063	protein of unknown function
METTOv1_10194	protein of unknown function
METTOv1_200055	protein of unknown function
METTOv1_450005	protein of unknown function
METTOv1_1280001	protein of unknown function
METTOv1_1050002	protein of unknown function
METTOv1_160016	protein of unknown function
METTOv1_200057	protein of unknown function
METTOv1_260041	protein of unknown function
METTOv1_600022	exported protein of unknown function
METTOv1_870007	protein of unknown function
METTOv1_260040	protein of unknown function
METTOv1_90064	protein of unknown function
METTOv1_30137	exported protein of unknown function
METTOv1_590023	protein of unknown function
METTOv1_250030	protein of unknown function
METTOv1_360046	protein of unknown function

Appendix B

SUPPLEMENTAL FOR CHAPTER 3

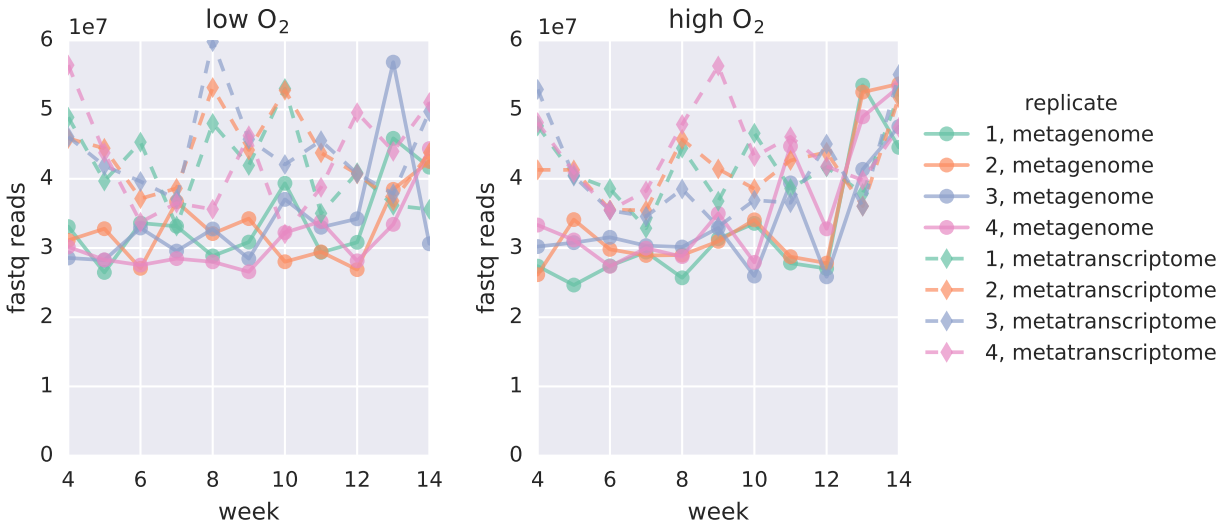


Figure B.1: Number of reads in metagenomes and metatranscriptomes, by sample.

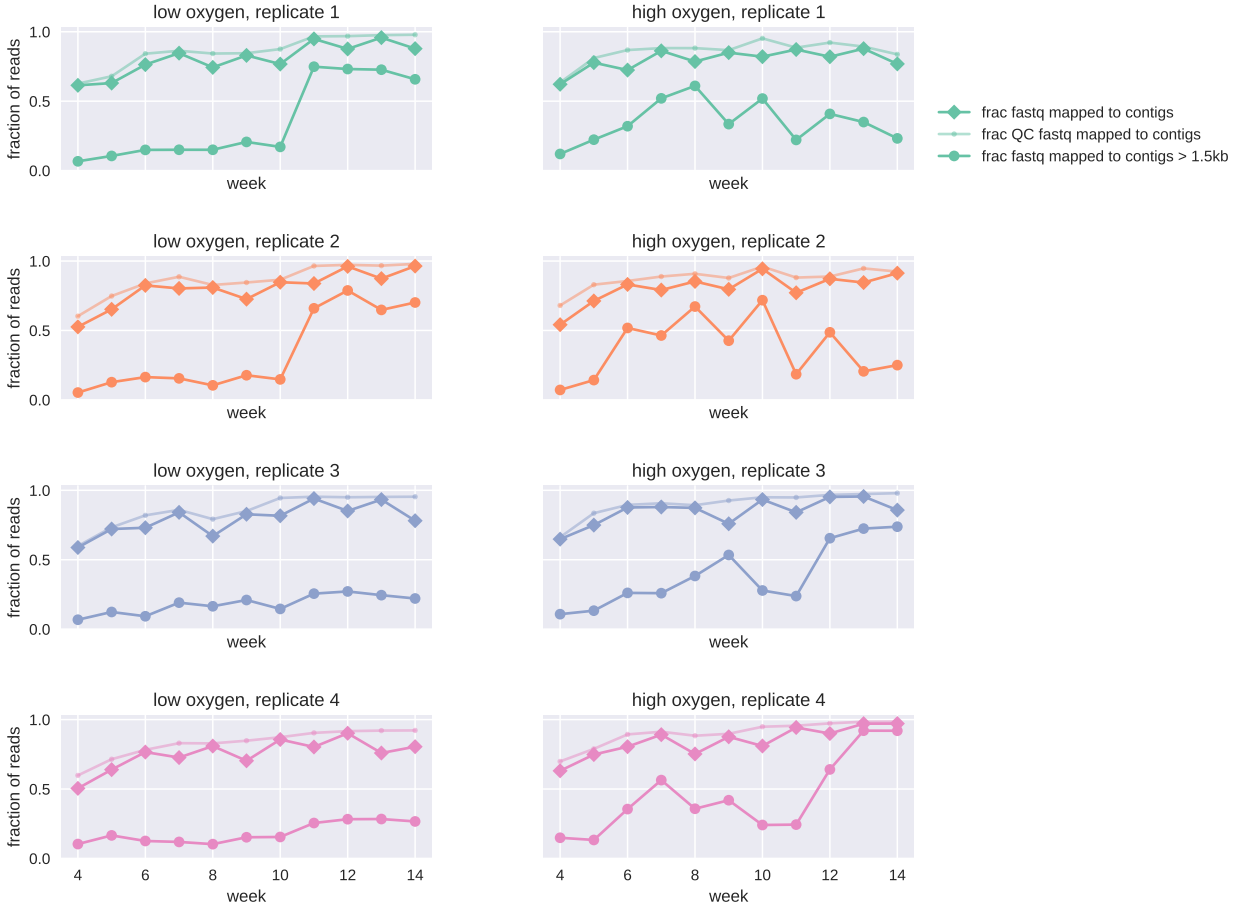


Figure B.2: Fraction of reads mapped to Elviz contigs. The top lines in each sub-plot show the fraction of reads that passed Elviz's QC filter. The diamond marks below indicate the fraction of total reads that mapped to contigs. The circle marks indicate the fraction that map to contigs with length $>$ 1.5kb, which are most promising for binning.

Table B.1: The top 300 proteins when summing reads across the 88 samples. Sample size variation and gene length are not controlled for. See Table B.2 for sample size variation. For gene sequences, locus tags, and counts in each sample, see supplementary files.

product	sum(RNA reads)	# gene copies
hypothetical protein	358,022,758	413,685
Particulate methane monooxygenase alpha subunit precursor	153,701,739	34
Ammonia monooxygenase/methane monooxygenase%2C subunit C	87,810,398	38
Sensor protein ZraS	81,403,544	613
Particulate methane monooxygenase beta subunit	75,048,024	32
Flagellin	12,453,840	155
Hydroxylamine reductase	11,249,260	13
S-layer protein	6,307,663	8
Cell wall-associated hydrolase	5,121,957	8
Capsid protein (F protein)	3,841,696	1
Methanol dehydrogenase [cytochrome c] subunit 2 precursor	3,405,479	27
Methanol dehydrogenase [cytochrome c] subunit 1 precursor	3,220,612	48
ATP-dependent zinc metalloprotease FtsH	3,200,325	363
Fimbrial protein precursor	3,181,026	638
Bacteriophage replication gene A protein (GPA)	2,954,718	49
Microvirus H protein (pilot protein)	2,810,060	1
Methane monooxygenase component C	2,769,830	30

Continued on next page

product	sum(RNA reads)	# gene copies
Methanol dehydrogenase [cytochrome c] subunit 1	2,751,386	69
Group II intron-encoded protein LtrA	2,444,067	133
3-hexulose-6-phosphate synthase	2,329,436	46
Transketolase 1	2,312,198	123
PEP-CTERM motif protein	2,148,893	652
Group 1 truncated hemoglobin GlbN	2,140,399	140
Formaldehyde-activating enzyme	2,072,071	147
Metal-binding protein SmbP precursor	2,064,649	77
Flp/Fap pilin component	2,014,075	100
ATP-dependent Clp protease ATP-binding subunit ClpA	1,907,981	121
Flagellar hook protein FlgE	1,853,924	130
DNA-binding protein HU-beta	1,689,620	176
Type II secretion system protein G precursor	1,659,638	384
Flagellar hook-associated protein 2	1,657,492	131
60 kDa chaperonin	1,621,356	167
Avirulence protein AvrBs3	1,601,853	4
DNA-directed RNA polymerase subunit beta′	1,586,236	152
Chaperone protein DnaK	1,544,707	277
Methyl-accepting chemotaxis protein II	1,527,723	525
Cyclic di-GMP phosphodiesterase Gmr	1,464,452	1,793
Elongation factor G 1	1,453,588	38
Phage Tail Collar Domain protein	1,434,560	159
Major spike protein (G protein)	1,368,598	1
Outer membrane porin F precursor	1,304,249	445

Continued on next page

product	sum(RNA reads)	# gene copies
putative TonB-dependent receptor BfrD precursor	1,289,784	190
Beta-lactamase TEM precursor	1,271,919	1
3-hexulose-6-phosphate isomerase	1,232,120	48
Phenol hydroxylase P5 protein	1,221,244	50
Lon protease	1,207,748	326
Aerobactin synthase	1,139,410	21
Colicin I receptor precursor	1,124,515	439
Cold shock protein ScoF	1,123,655	52
Cold shock-like protein CspA	1,117,436	40
Phytochrome-like protein cph2	1,105,919	1,558
putative peroxiredoxin	1,102,062	90
DNA-directed RNA polymerase subunit beta	1,084,375	110
Flagellar basal-body rod protein FlgG	1,076,811	128
Bacterial extracellular solute-binding proteins%2C family 3	1,073,251	470
Phenolphthiocerol synthesis polyketide synthase type I Pks15/1	1,069,727	23
Flagellar hook-associated protein 1	1,060,718	116
30S ribosomal protein S1	1,052,576	177
Chemotaxis protein CheA	1,049,177	280
Flagellar basal-body rod protein FlgF	1,044,793	93
RNA polymerase sigma factor RpoH	1,024,563	134
Cyclic di-GMP phosphodiesterase response regulator RpfG	1,009,170	931
Cytochrome c-L precursor	1,005,675	76
Poly(beta-D-mannuronate) C5 epimerase 1	984,141	17

Continued on next page

product	sum(RNA reads)	# gene copies
L%2CD-transpeptidase catalytic domain	979,975	212
Methyl-accepting chemotaxis protein I	976,610	309
ATP synthase subunit alpha	960,915	147
Elongation factor Tu	958,727	140
Chaperone protein ClpB	942,491	211
Right origin-binding protein	928,055	23
Outer membrane protein A precursor	927,434	148
ATP-dependent Clp protease ATP-binding subunit ClpX	901,929	192
Chemotaxis protein CheY	897,095	430
Modulator of FtsH protease YccA	892,776	137
Response regulator PleD	869,403	836
putative CtpA-like serine protease	863,297	162
RNA polymerase sigma factor RpoD	845,175	128
Biopolymer transport protein ExbB	838,111	570
Spore protein SP21	822,607	128
Polyketide cyclase / dehydrase and lipid transport	810,573	458
Multidrug resistance outer membrane protein MdtP precursor	806,559	101
Protease HtpX	804,805	204
Bacteriophage scaffolding protein D	794,345	1
Signal transduction histidine-protein kinase BarA	785,795	519
preprotein translocase subunit SecY	779,178	134
50S ribosomal protein L2	771,359	103
Superoxide dismutase [Fe]	764,292	138
Transcriptional regulatory protein ZraR	759,073	845

Continued on next page

product	sum(RNA reads)	# gene copies
Putative deoxyribonuclease RhsC	754,966	161
Chaperone protein HtpG	748,660	188
Minor curlin subunit precursor	748,058	12
Glutamine synthetase	746,921	91
Cold shock protein CspC	740,332	18
Cytochrome c oxidase subunit 1	740,292	171
Outer membrane protein W precursor	727,458	117
putative phospholipid-binding lipoprotein MlaA precursor	726,741	195
Transaldolase	724,887	145
30S ribosomal protein S4	716,789	141
Cytochrome c-551 precursor	714,542	132
ATP synthase subunit beta	706,925	154
Cell division protein FtsZ	682,952	157
Ribosome hibernation promoting factor	672,173	122
Modulator of FtsH protease HflK	670,892	239
30S ribosomal protein S2	667,779	135
Quercetin 2%2C3-dioxygenase	662,723	414
Peptidoglycan synthase FtsI precursor	655,407	39
Transposase DDE domain protein	644,635	575
Multidrug export protein EmrB	641,069	207
Ammonia channel precursor	638,801	258
Flagellar M-ring protein	637,658	112
ComE operon protein 1	625,964	88
putative periplasmic serine endoprotease DegP-like precursor	619,861	318
Cysteine desulfurase	617,548	464

Continued on next page

product	sum(RNA reads)	# gene copies
Cytochrome c oxidase subunit 2 precursor	617,471	138
Cytochrome c oxidase subunit 3	612,205	236
Sporulation related domain protein	606,722	182
50S ribosomal protein L3	598,078	111
Nitrate/nitrite transporter NarK	598,069	62
Flagellar P-ring protein precursor	596,197	113
DNA-directed RNA polymerase subunit alpha	587,351	135
Type II secretion system protein E	584,970	606
Cytochrome c551 peroxidase precursor	581,414	306
Copper resistance protein A precursor	579,604	174
50S ribosomal protein L25	574,166	130
Outer membrane porin protein 32 precursor	573,542	192
Flagellar protein FliT	549,417	30
50S ribosomal protein L1	545,307	124
Respiratory nitrate reductase 1 alpha chain	543,710	45
Na(+)-translocating NADH-quinone reductase subunit F	540,161	44
Linear gramicidin synthase subunit D	538,654	72
N-acetylmuramoyl-L-alanine amidase AmiC precursor	531,901	137
FecR protein	528,495	208
Cation efflux system protein CusA	525,055	267
Glucose-1-phosphate adenylyltransferase	519,029	150
Hypoxic response protein 1	514,246	168
Putative formate dehydrogenase	514,121	67
N%2CN'-diacetylchitobiose phosphorylase	503,811	68

Continued on next page

product	sum(RNA reads)	# gene copies
Polyribonucleotide nucleotidyltransferase	501,398	142
Ferrous iron transport protein B	497,854	158
Ribosomal RNA small subunit methyltransferase H	496,386	178
3-oxoacyl-[acyl-carrier-protein] synthase 1	495,702	136
Trigger factor	495,625	132
TonB dependent receptor	489,761	770
Copper-exporting P-type ATPase A	486,554	222
Membrane-bound lytic murein transglycosylase D precursor	485,735	279
Acyl carrier protein	475,902	248
Flagellum site-determining protein YlxH	475,541	86
Bifunctional hemolysin/adenylate cyclase precursor	474,908	224
50S ribosomal protein L20	472,974	117
Type II secretion system protein D precursor	466,040	360
Pyruvate kinase II	461,644	103
DNA primase	461,608	221
Transposase IS66 family protein	458,000	220
Ribose-phosphate pyrophosphokinase	454,049	201
Flagellar biosynthesis protein FlhF	451,863	83
30S ribosomal protein S15	450,754	121
3-phenylpropionate/cinnamic acid dioxygenase subunit alpha	450,573	16
RNA-binding protein Hfq	444,771	107
Maltodextrin phosphorylase	443,796	146
N(2)-citryl-N(6)-acetyl-N(6)-hydroxylysine synthase	440,152	15

Continued on next page

product	sum(RNA reads)	# gene copies
NAD(P)-dependent methylenetetrahydromethanopterin dehydrogenase	439,742	104
30S ribosomal protein S9	438,650	142
FeS cluster assembly protein SufB	436,043	110
Type IV pilus biogenesis and competence protein PilQ precursor	427,396	224
Ferredoxin-dependent glutamate synthase 1	426,544	85
30S ribosomal protein S12	423,584	101
Osmotically-inducible protein Y precursor	422,220	304
ATP-dependent Clp protease proteolytic subunit	421,818	178
Erythronolide synthase%2C modules 1 and 2	420,082	6
Pyruvate-flavodoxin oxidoreductase	419,139	48
Fe(3+) dicitrate transport protein FecA precursor	418,819	132
Na(+)-translocating NADH-quinone reductase subunit B	417,928	28
Glutamate synthase [NADPH] small chain	417,113	192
cell division protein MraZ	415,859	129
Flagellar hook-associated protein 3	413,545	95
Nitric oxide reductase subunit B	412,346	148
2-isopropylmalate synthase	411,299	246
Aconitate hydratase 2	410,938	109
50S ribosomal protein L27	410,073	127
30S ribosomal protein S7	409,297	104
Cyclic pyranopterin monophosphate synthase 1	407,726	56

Continued on next page

product	sum(RNA reads)	# gene copies
Bacterioferritin	406,109	166
Sensor protein FixL	405,443	475
Dihydrolipoyl dehydrogenase	404,743	334
putative multidrug resistance protein EmrK	404,355	65
Basal-body rod modification protein FlgD	404,309	107
3-oxoacyl-[acyl-carrier-protein] synthase 2	403,698	435
UDP-N-acetylmuramoyl-L-alanyl-D-glutamate–2%2C6-diaminopimelate ligase	403,230	148
preprotein translocase subunit SecA	400,638	232
Ribonuclease R	390,928	210
Fructose-bisphosphate aldolase class 2	389,633	63
50S ribosomal protein L10	385,597	110
DNA translocase FtsK	384,995	152
RNA polymerase sigma-54 factor	381,702	168
UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase	379,042	171
Cobalt-zinc-cadmium resistance protein CzcA	376,307	608
Modulator of FtsH protease HflC	373,571	135
putative lipoprotein YiaD precursor	373,448	332
Na(+)/H(+) antiporter NhaD	371,348	162
Squalene–hopene cyclase	366,723	41
Phosphoglucomutase	362,797	139
putative ABC transporter ATP-binding protein	362,766	553
Acetolactate synthase isozyme 3 large subunit	361,703	87
PilZ domain protein	361,543	389

Continued on next page

product	sum(RNA reads)	# gene copies
Phosphogluconate dehydratase	360,428	100
FMN-dependent NADH-azoreductase	358,941	47
Outer membrane efflux protein	355,020	711
ECF RNA polymerase sigma-E factor	354,402	327
Alpha/beta hydrolase family protein	354,213	1,330
3-oxoacyl-[acyl-carrier-protein] reductase FabG	353,830	1,263
Pyruvate dehydrogenase E1 component	353,583	129
NAD(P)H-quinone oxidoreductase chain 4 1	353,001	102
Bacteriohemerythrin	352,386	136
2%2C3-bisphosphoglycerate-independent phosphoglycerate mutase	351,523	142
putative phospholipid-binding protein MlaC precursor	351,178	146
ATP:dephospho-CoA triphosphoribosyl transferase	350,953	54
Na(+)-translocating NADH-quinone reductase subunit C	350,710	24
DNA-invertase hin	350,432	272
tol-pal system protein YbgF	349,970	181
D-erythrose-4-phosphate dehydrogenase	347,666	22
NADP-reducing hydrogenase subunit HndC	347,183	84
10 kDa chaperonin	347,019	113
6-phosphogluconolactonase	344,336	284
Peptidase propeptide and YPEB domain protein	344,261	141
Chemotaxis protein CheW	344,064	326

Continued on next page

product	sum(RNA reads)	# gene copies
Succinyl-CoA ligase [ADP-forming] subunit beta	343,575	175
Carbamoyl-phosphate synthase large chain	341,882	188
Polyketide synthase PksN	341,165	7
putative parvulin-type peptidyl-prolyl cis-trans isomerase precursor	340,103	132
Flagellar hook-length control protein	339,156	56
Methyltransferase domain protein	338,398	367
NAD-reducing hydrogenase HoxS subunit alpha	338,348	38
Inosine-5′-monophosphate dehydrogenase	337,284	207
flagellar protein FlaG	335,036	64
Glycosyl hydrolase family 57	334,643	60
50S ribosomal protein L4	333,085	111
ATPase family associated with various cellular activities (AAA)	332,936	296
Phosphoribosylformylglycinamidine synthase	331,827	142
Rod shape-determining protein MreB	331,122	173
Response regulator UvrY	329,883	400
Signal recognition particle protein	329,618	149
Tetratricopeptide repeat protein	329,174	1,427
Thermostable monoacylglycerol lipase	327,977	55
30S ribosomal protein S3	326,369	93
Tim44-like domain protein	322,952	105
Oxygen-independent coproporphyrinogen-III oxidase	320,953	201
ATP synthase subunit c	320,002	136

Continued on next page

product	sum(RNA reads)	# gene copies
50S ribosomal protein L15	319,012	116
RNA polymerase sigma factor SigA	317,688	83
FlgN protein	316,776	63
Ferrichrome receptor FcuA precursor	315,542	88
HDOD domain protein	313,422	460
dTDP-glucose 4%2C6-dehydratase	312,803	341
Flagellar motor switch protein FliM	312,289	104
ATP synthase subunit delta	312,084	107
tetratricopeptide repeat protein	311,426	444
N%2CN′-diacetyllegionaminic acid synthase	310,503	95
Formate dehydrogenase H	310,054	94
Dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex	309,903	254
MOSC domain protein	308,478	104
Na(+)-translocating NADH-quinone reductase subunit D	308,303	20
50S ribosomal protein L14	307,247	110
Transposase	307,033	581
Chaperone protein Skp precursor	305,658	115
UDP-N-acetylmuramoyl-tripeptide–D-alanyl-D-alanine ligase	304,306	183
Septum site-determining protein MinD	302,251	202
Tyrosine recombinase XerC	302,003	607
Malate dehydrogenase	301,547	151
50S ribosomal protein L6	300,590	120
ATP synthase epsilon chain	300,575	150

Continued on next page

product	sum(RNA reads)	# gene copies
50S ribosomal protein L5	299,638	112
transport protein TonB	299,458	222
Adenosylhomocysteinase	299,382	173
Chaperone protein DnaJ	298,386	374
50S ribosomal protein L17	296,464	143
50S ribosomal protein L7/L12	295,549	122
Dihydroxy-acid dehydratase	295,169	189
ATP-dependent Clp protease adapter protein ClpS	294,553	112
Multicopper oxidase	294,469	44
UDP-3-O-acylglucosamine N-acyltransferase	294,322	188
Type II secretion system protein F	293,302	371
Phospho-N-acetylmuramoyl-pentapeptide-transferase	292,626	171
30S ribosomal protein S5	291,627	114
L-2%2C4-diaminobutyrate decarboxylase	291,090	57
anti-sigma28 factor FlgM	290,901	71
Phytanoyl-CoA dioxygenase (PhyH)	289,636	142
30S ribosomal protein S13	289,348	109
Threonine–tRNA ligase	288,502	120
Iron-sulfur cluster insertion protein ErpA	287,287	204
NAD(P)H azoreductase	286,417	65
Flagellar biosynthesis protein FlhA	285,954	101
Nicotinate-nucleotide pyrophosphorylase [carboxylating]	284,572	144
Poly(beta-D-mannuronate) C5 epimerase 5	283,725	25
ATP synthase gamma chain	283,576	137

Continued on next page

product	sum(RNA reads)	# gene copies
Outer membrane efflux protein BepC precursor	282,205	131
Molybdenum-pterin-binding protein MopA	281,225	168
Cupin domain protein	277,550	631
Chemotaxis response regulator protein-glutamate methylesterase	276,474	192
UTP–glucose-1-phosphate uridylyltransferase	275,652	143
putative FAD-linked oxidoreductase	275,649	382

Figure B.3: Expression of a highly expressed phage capsid protein across samples.

Figure B.4: Expression of a highly expressed phage pilot protein across samples.

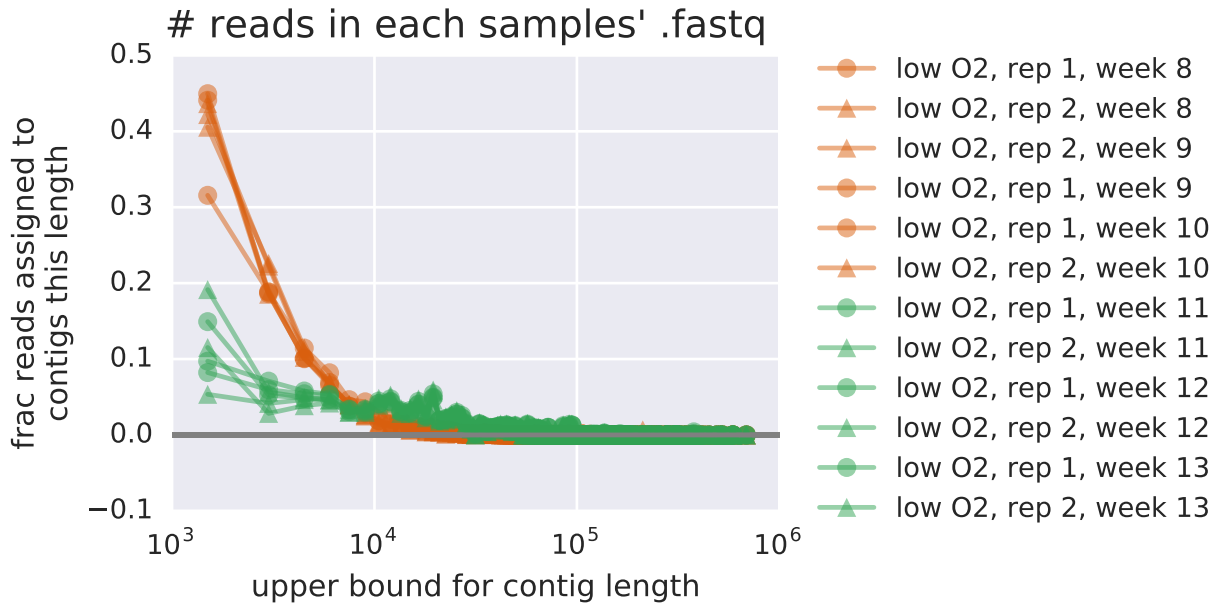


Figure B.5: Samples with metagenomes that are well represented by MetaBAT bins have more reads drawn to longer contigs.

Table B.2: CheckM results for isolate genomes, a positive control of CheckM efficacy for methylotrophs.

Bin Id	Marker lineage	Completeness	Contamination	Strain heterogeneity
Ancylobacter_sp_FA202	o__Rhizobiales (UID3642)	100.00	0.00	0.00
Arthrobacter_sp_31Y	f__Micrococcaceae (UID1631)	99.71	0.00	0.00
Arthrobacter_sp_35W	f__Micrococcaceae (UID1623)	99.77	0.46	0.00
Arthrobacter_sp_MA-N2	f__Micrococcaceae (UID1631)	99.62	0.00	0.00
Bacillus_sp_37MA	c__Bacilli (UID259)	98.68	1.55	0.00
Bacillus_sp_72	c__Bacilli (UID259)	98.68	2.43	33.33
Bosea_sp_117	o__Rhizobiales (UID3642)	99.68	0.00	0.00
Flavobacterium_sp_83	f__Flavobacteriaceae (UID2817)	99.65	0.71	0.00
Flavobacterium_sp_Fl	f__Flavobacteriaceae (UID2817)	99.65	0.42	0.00
Hoeflea_sp_108	o__Rhizobiales (UID3450)	99.43	0.96	0.00
Hyphomicrobium_sp_802	o__Rhizobiales (UID3447)	100.00	0.20	0.00
Hyphomicrobium_sp_99	o__Rhizobiales (UID3447)	99.39	0.61	50.00
Janthinobacterium_sp_RA13	o__Burkholderiales (UID4002)	99.25	1.94	0.00
Methylobacter_tundripaludum_21_22	c__Gammaproteobacteria (UID4274)	99.37	2.82	9.09
Methylobacter_tundripaludum_31_32	c__Gammaproteobacteria (UID4274)	99.37	3.39	7.14
Methylobacterium_sp_10	o__Rhizobiales (UID3654)	99.84	0.00	0.00
Methylobacterium_sp_77	o__Rhizobiales (UID3654)	100.00	0.31	0.00
Methylobacterium_sp_88A	o__Rhizobiales (UID3654)	99.84	0.00	0.00
Methylocystis_sp_LW5	o__Rhizobiales (UID3654)	100.00	0.31	0.00
Methylomonas_sp_11b	c__Gammaproteobacteria (UID4274)	99.93	0.42	0.00
Methylomonas_sp_LW13	c__Gammaproteobacteria (UID4274)	99.91	0.41	0.00
Methylomonas_sp_MK1	c__Gammaproteobacteria (UID4274)	99.92	1.10	0.00
Methylophilaceae_bacterium_11	c__Betaproteobacteria (UID3888)	100.00	0.00	0.00
Methylophilus_sp_1	c__Betaproteobacteria (UID3888)	100.00	0.00	0.00
Methylophilus_sp_42	c__Betaproteobacteria (UID3888)	100.00	0.00	0.00

Continued on next page

Bin Id	Marker lineage	Completeness	Contamination	Strain heterogeneity
Methylophilus_sp_5	c__Betaproteobacteria (UID3888)	100.00	0.11	0.00
Methylophilus_sp_Q8	c__Betaproteobacteria (UID3888)	100.00	0.03	0.00
Methylopila_sp_73B	o__Rhizobiales (UID3642)	100.00	0.16	0.00
Methylopila_sp_M107	o__Rhizobiales (UID3642)	100.00	0.32	0.00
Methylosarcina_lacus_LW14	c__Gammaproteobacteria (UID4274)	99.59	0.95	0.00
Methylosinus_sp_LW3	o__Rhizobiales (UID3654)	100.00	0.31	0.00
Methylosinus_sp_LW4	o__Rhizobiales (UID3654)	99.69	0.00	0.00
Methylosinus_sp_PW1	o__Rhizobiales (UID3654)	100.00	0.30	0.00
Methylotenera_mobilis_13	c__Betaproteobacteria (UID3888)	99.57	0.00	0.00
Methylotenera_mobilis_JLW8	c__Betaproteobacteria (UID3888)	99.57	0.00	0.00
Methylotenera_sp_1P_1	c__Betaproteobacteria (UID3888)	99.57	0.00	0.00
Methylotenera_sp_73s	c__Betaproteobacteria (UID3888)	99.57	0.00	0.00
Methylotenera_sp_G11	c__Betaproteobacteria (UID3888)	100.00	0.03	0.00
Methylotenera_sp_L2L1	c__Betaproteobacteria (UID3888)	99.57	0.00	0.00
Methylotenera_sp_N17	c__Betaproteobacteria (UID3888)	100.00	0.00	0.00
Methylotenera_versatilis_301	c__Betaproteobacteria (UID3888)	99.57	1.28	0.00
Methylotenera_versatilis_7	c__Betaproteobacteria (UID3888)	99.57	0.43	100.00
Methylotenera_versatilis_79	c__Betaproteobacteria (UID3888)	99.57	0.00	0.00
Methyloversatilis_sp_FAM1	f__Rhodocyclaceae (UID3972)	99.17	0.00	0.00
Methyloversatilis_sp_RZ18-153	f__Rhodocyclaceae (UID3972)	99.59	0.00	0.00
Methyloversatilis_universalis_FAM5	f__Rhodocyclaceae (UID3972)	99.59	0.00	0.00
Methyloversatilis_universalis_Fam500	f__Rhodocyclaceae (UID3972)	99.59	0.00	0.00
Methylovorus_glucosetrophus_SIP3-4	c__Betaproteobacteria (UID3888)	99.57	0.00	0.00
Mycobacterium_sp_141	o__Actinomycetales (UID1815)	99.95	0.09	0.00
Mycobacterium_sp_155	o__Actinomycetales (UID1815)	100.00	0.00	0.00
Paracoccus_sp_N5	f__Rhodobacteraceae (UID3340)	99.39	0.15	0.00
Pseudomonas_sp_11_12A	g__Pseudomonas (UID4576)	99.66	0.44	0.00
Xanthobacter_sp_126	o__Rhizobiales (UID3642)	99.96	0.32	0.00
Xanthobacter_sp_91	o__Rhizobiales (UID3642)	99.81	0.32	100.00
Xanthobacteraceae_bacterium_501b	o__Rhizobiales (UID3642)	100.00	0.00	0.00

Appendix C

SUPPLEMENTAL FOR CHAPTER 4

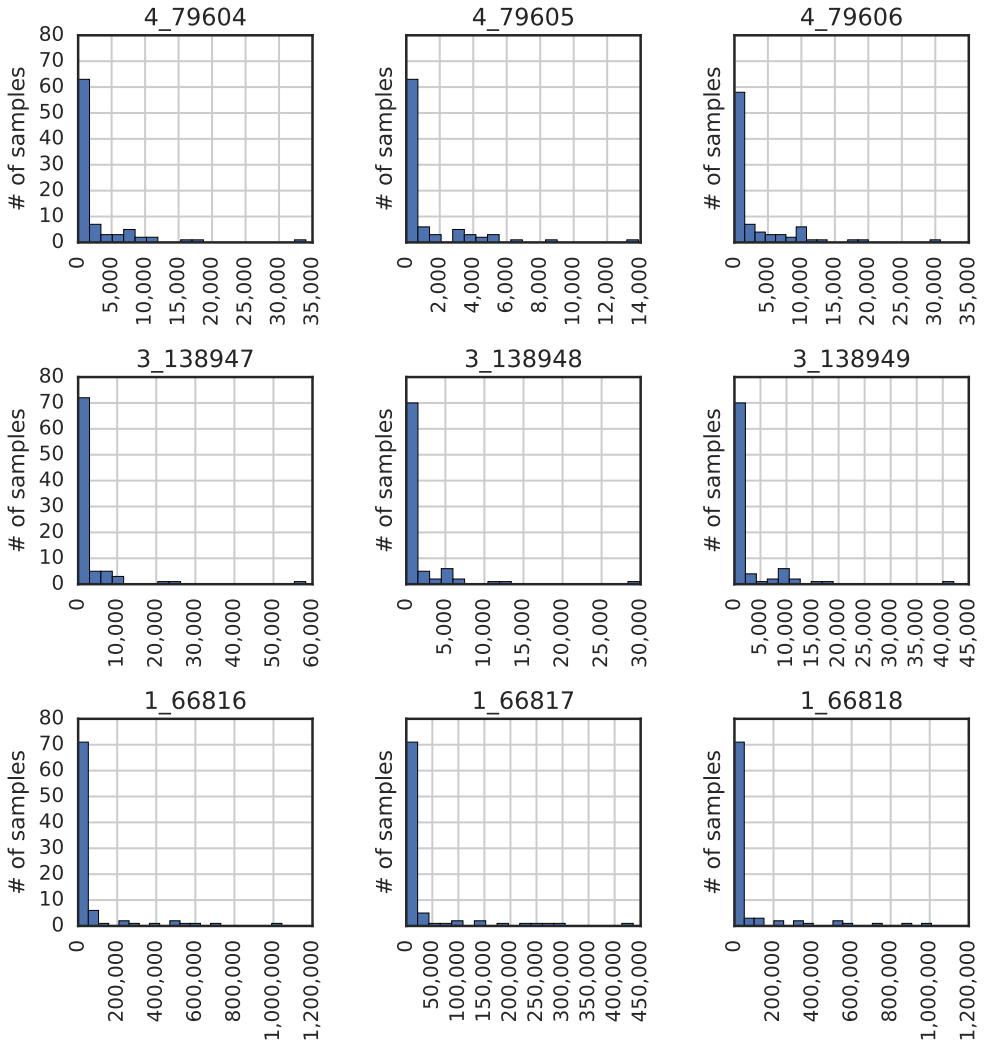


Figure C.1: Histograms of read counts for three sets of *pmoCAB* gene clusters. These data do not represent normal distributions.

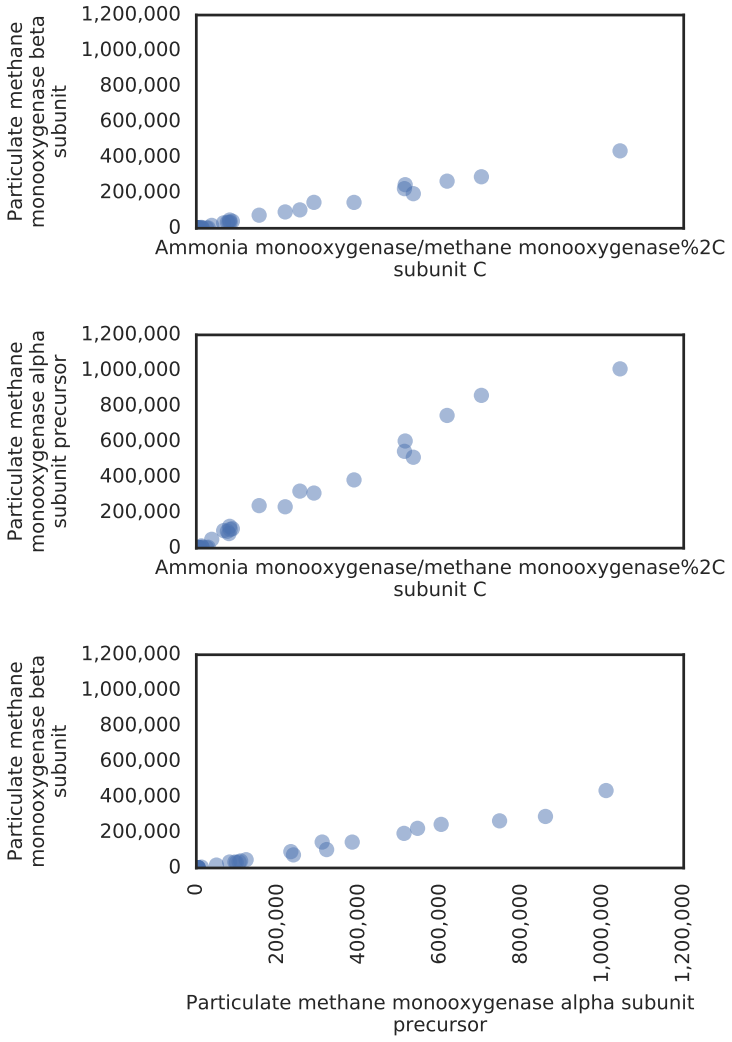


Figure C.2: Demonstration of correlation between *pmo* subunits in cluster 1.

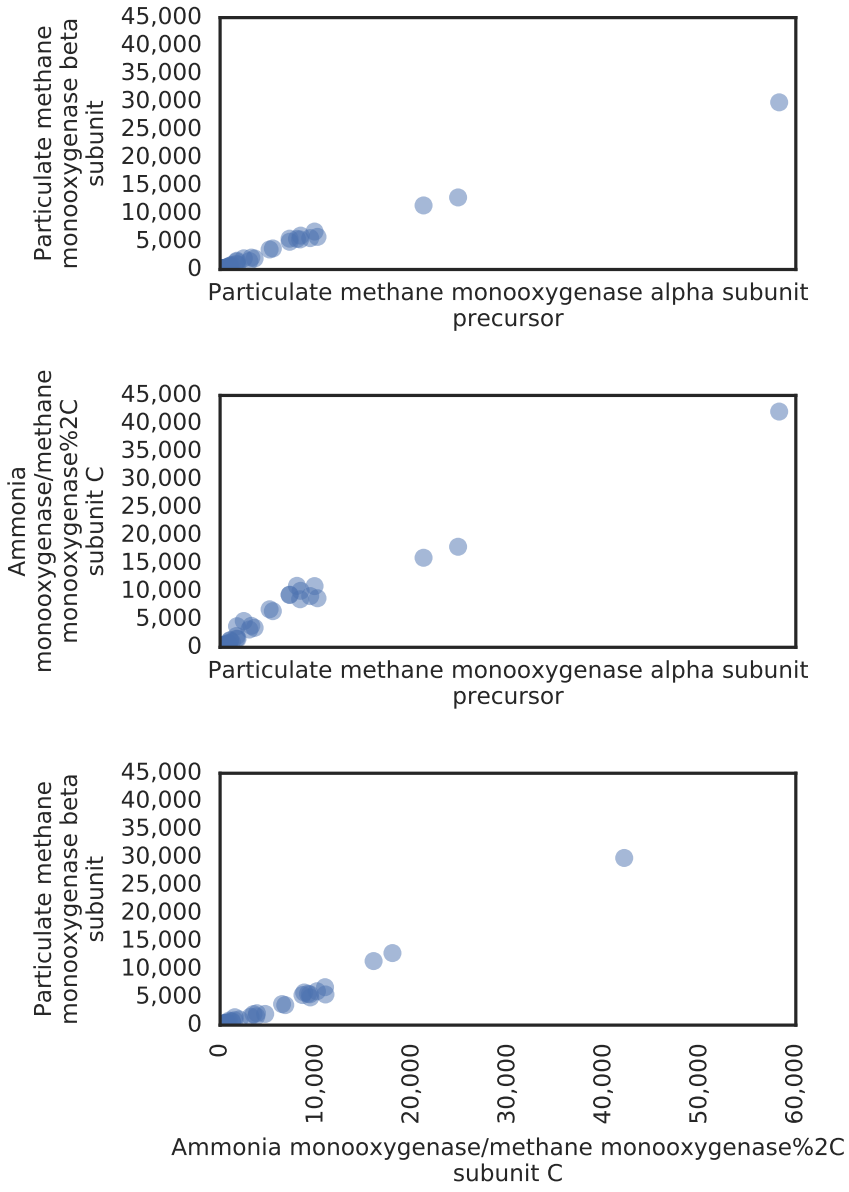


Figure C.3: Demonstration of correlation between *pmo* subunits in cluster 2.

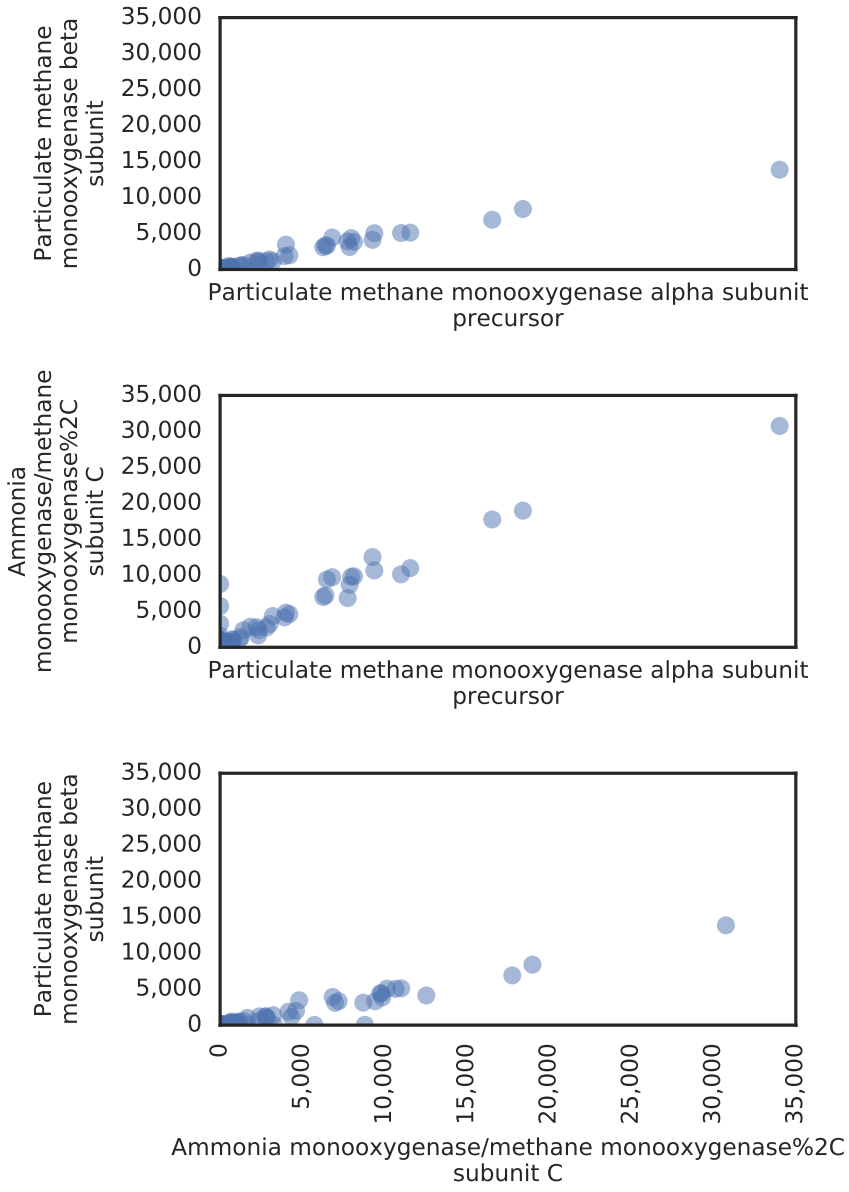


Figure C.4: Demonstration of correlation between *pmo* subunits in cluster 3.

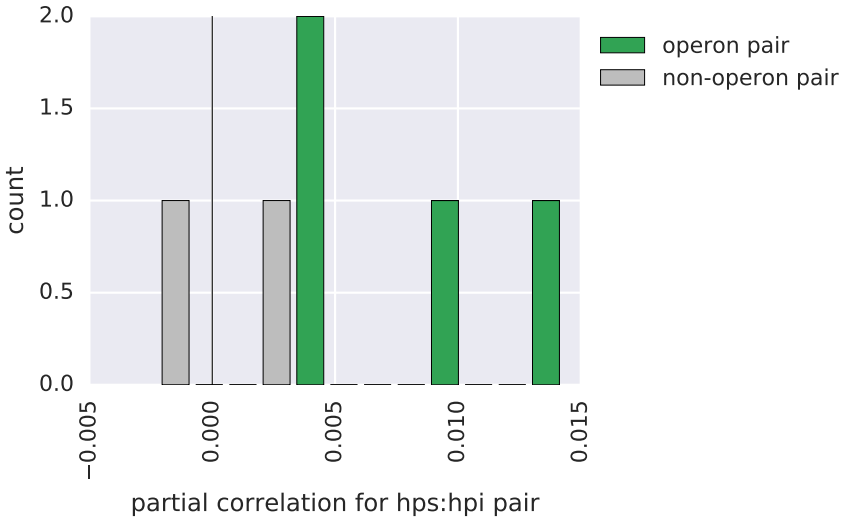


Figure C.5: Partial correlation values for *hps*:*hpi* (hexulose-6-phosphate synthase:hexulose-6-phosphate isomerase) gene pairs in the GeneNet network. Green bars represent pairs that are sequential on a contig, and gray bars represent all other pairs.