EXPLORING ENVIRONMENTAL ADAPTATIONS AND HABITAT PREFERENCES IN THREE MICROBIAL LINEAGES USING COMPARATIVE (META)GENOMIC APPROACHES

By

ARCHANA YADAV

Bachelor of Technology in Biotechnology
Kathmandu University
Dhulikhel, Bagmati province, Nepal
2015

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
December, 2022

EXPLORING ENVIRONMENTAL ADAPTATIONS AND HABITAT PREFERENCES IN THREE MICROBIAL LINEAGES USING COMPARATIVE (META)GENOMIC APPROACHES

Dissertation Approved:

Dr. Mostafa Elshahed

Dissertation Adviser

Dr. Noha Youssef

Dr. Robert Burnap

Dr. Matthew Cabeen

Dr. Michael Anderson

ACKNOWLEDGEMENTS

I would like to start by thanking my advisor Dr. Mostafa Elshahed from the bottom of my heart for accepting me as a graduate student in his lab and providing an incredible mentorship and support during my Ph.D. journey. He made sure that I excelled in every aspect of my development as a scientist by encouraging me to read papers, write manuscripts, present and devise experiments. Equally commendable is the role of my "second" mentor Dr. Noha Youssef who shares a lab with us. I want to acknowledge her with the whole of my heart for guiding me with all the technical learning and overall support. One of the many things I learned from her is that there is always a solution to a bioinformatic program that is not working and that we should always keep trying. I am forever grateful to my mentors for the countless things they taught me over these years. I am also immensely thankful for the years I spent in their lab growing as a researcher. Joining their lab was one of the best decisions I made in my life.

To rest of my committee members, Dr. Robert Burnap, Dr. Matthew Cabeen, and Dr. Michael Anderson, thank you for your support, valuable feedback, and many fruitful discussions about my research projects we had during committee meetings. I would also like to acknowledge the past members of my lab Dr. Radwa Hanafy, Dr. Chelsea Murphy, and Ryan Hahn who were present for majority of my years at lab, for their precious friendship, and for making lab a great place where we could talk about everything and have fun discussions. Also, thanks to Tammy Austin and current lab members Carrie, Adrienne, and Casey for being fun companions and keeping our lab an interesting place to be.

Finally, I would like to thank my parents for always being there for me even being miles apart during my Ph.D. journey. Thank you for everything you taught me and raising me to work hard and live openly. Special thanks to my elder brother Rohan who has believed in me since the beginning of my Ph.D., even when everybody else had doubts. Also, thanks to a very special person, Raju for always listening to me and being there for me. There are many people at OSU who shared my journey and made this place feel like home and a place where I was loved and accepted. To my department friends Deepali, Ben, Addy, Ross, Samikshya as well my other friends Manika, Miruthula, Riva, Pratistha, Asma and countless others, thank you for sharing my journey here and giving me many fun memories to cherish for the rest of my life. My friends played a very important role in sharing good times at OSU and keeping me sane during difficult times. I also want to thank the OSU graduate housing for giving me the opportunity to connect to many amazing students from all over the world and providing a fun place to live all these years.

I am immensely blessed to have such loving people in my life who all made it possible for me to get through this journey.

Name: ARCHANA YADAV

Date of Degree: DECEMBER 2022

Title of Study: EXPLORING ENVIRONMENTAL ADAPTATIONS AND HABITAT PREFERENCES IN THREE MICROBIAL LINEAGES USING COMPARATIVE (META)GENOMIC APPROACHES

Major Field: MICROBIOLOGY, CELL, AND MOLECULAR BIOLOGY

Abstract:

The utilization of -omics based approaches (metagenomics, genomics, transcriptomics, proteomics, and metabolomics) in the field of microbiology has greatly advanced our understanding of the microbial world. The utilization of such approaches, either on pure cultures, or directly on environmental samples has provided novel insights into the role of microorganisms in earth biogeochemical cycles, microbial evolutionary dynamics, and their potential biotechnological applications. In the field of microbial pathogenesis, informatics-based methods have helped in uncovering several venues of pathogenesis including pathogens strain-specific characteristics, virulence genes, antimicrobial resistance, and understanding the landscape of various diseases.

Here, I present my 3 research projects based on exploiting various -omics based approaches to understand the ecology, evolution, and pathogenic determinants of various groups of cultured, and yet-uncultured microorganisms. In chapter I, I implemented genome-resolved metagenomics to elucidate the ecological roles, metabolic capabilities, and physiological preferences of a novel yet-uncultured microbial phylum recovered from enrichments of tertiary oil reservoir. I showed that this lineage is a slow-growing member of rare biosphere and an aminolytic halothermophilic organism. We proposed creating a new candidate phylum "Mcinerneybacteriota" to accommodate this organism. This work has been published in the journal “Systematic and Applied Microbiology”. In chapter II, I analyzed multiple genome-resolved metagenomes of uncultured Group 18 Acidobacteria to understand their biogeochemical roles and elucidate the key evolutionary innovations that enable Acidobacteria to thrive in soil ecosystems. I demonstrated that soil-dwelling genera were characterized by larger genomes, higher CRISPR loci, expanded CAZyme machinery, possession of a C_1 metabolism, and a sole dependence on aerobic respiration, whereas nonsoil genomes encoded a more versatile respiratory capacity and potential for utilizing the Wood-Ljungdahl (WL) pathway as an electron sink. This work is published in the journal “Applied and Environmental Microbiology”. Lastly, my third project (Chapter III) is about utilizing genomics and transcriptomics for an intracellular pathogen, *Coxiella burnetii,* to understand the changes in its genes crucial for intracellular success during long-term culturing in an axenic media. Here, I showed the expression changes and mutations in multiple genes that are known or most likely predicted to be crucial to their normal intracellular growth lifestyle or pathogenesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

CANDIDATUS MCINERNEYIBACTERIUM AMINIVORANS GEN. NOV., SP. NOV., THE FIRST REPRESENTATIVE OF THE CANDIDATE PHYLUM MCINERNEYIBACTERIOTA PHYL. NOV. RECOVERED FROM A HIGH TEMPERATURE, HIGH SALINITY TERTIARY OIL RESERVOIR IN NORTH CENTRAL OKLAHOMA, USA

1.1 Abstract

We report on the characterization of a novel genomic assembly (ARYD3) recovered from formation water (17.6% salinity) and crude oil enrichment amended by isolated soy proteins (0.2%), and incubated for 100 days under anaerobic conditions at 50°C. Phylogenetic and phylogenomic analysis demonstrated that the ARYD3 is unaffiliated with all currently described bacterial phyla and candidate phyla, as evident by the low AAI (34.7%), shared gene content (19.4%), and 78.9% 16S rRNA gene sequence similarity to *Halothiobacillus neapolitanus*, its closest cultured relative. Genomic characterization predicts a slow-growing, non-spore forming, and non-motile Gram-negative rod. Adaptation to high salinity is potentially mediated by the production of the compatible solutes cyclic 2,3-diphosphoglycerate (cDPG), α-glucosylglycerate, as well as the uptake of glycine betaine. Metabolically, the genome encodes primarily aminolytic capabilities for a wide range of amino acids and peptides. Interestingly, evidence of propionate degradation to succinate via methyl-malonyl CoA was identified, suggesting possible capability for syntrophic propionate degradation. Analysis of ARYD3 global distribution patterns identified its occurrence in a very small fraction of Earth Microbiome Project datasets examined (318/27,068), where it consistently represented an extremely rare fraction (maximum 0.28%, average 0.004%) of the overall community. We propose the *Candidatus* name Mcinerneyibacterium aminivorans gen. nov, sp. nov. for ARYD3, with the genome serving as the type material for the novel family Mcinerneyibacteriaceae fam. nov., order Mcinerneyibacteriales ord. nov., class Mcinerneyibacteria class nov., and phylum Mcinerneyibacteriota phyl. nov. The type material genome assembly is deposited in GenBank under accession number VSIX00000000.

1.2 Introduction

Our understanding of the scope of microbial diversity on Earth is rapidly expanding. New experimental and bioinformatic approaches allow for the direct, culture-independent recovery of genomes from environmental samples [1-3]. These advances, when coupled to recent breakthroughs in high throughput sequencing technologies and improved accessibility to supercomputing capacities, have been instrumental in expanding the tree of life and elucidating the metabolic characteristics, ecological roles, and physiological preferences of multiple yet-uncultured microbial taxa [4-6].

Amplicon-based 16S rRNA high throughput sequencing studies have often identified extremely diverse assemblies of microorganisms that are either temporarily or perpetually present in low abundance in a wide range of environments [7, 8]. Studies combining short high throughput and long-read (Sanger) sequencing have demonstrated that a fraction of the rare biosphere is phylogenetically novel [9, 10]. The rare biosphere could hence be regarded as a reservoir of hitherto uncharacterized novel microbial diversity. The recovery and characterization of genomes belonging to novel phylogenetic lineages from the rare biosphere could significantly expand the tree of life, improve our understanding of the global distribution patterns of various metabolic pathways and capacities, and allow for a more complete understanding of gene, pathways, and organismal evolution in the microbial world. Nevertheless, recovering complete or near complete genomes of rare members of the community, especially in highly diverse ecosystems, remains a daunting task; since it requires extremely deep sequencing effort and powerful computational capacities rarely available to the wider research community. Promotion of rare members into higher numbers through seasonal or anthropological environmental perturbations or through wide selective enrichment procedures occasionally provide serendipitous access to hitherto uncharacterized members of the rare biosphere.

Oil and natural gas (ONG) formations harbor and sustain diverse bacterial and archaeal communities [11, 12]. The microbial community structure and dynamics has been investigated in a wide range of ONG formations at various stages of production [13]. Microbial communities in belowground oil and natural gas formations are often constrained by the extreme environmental conditions prevalent *in-situ* (e.g., high temperature, high salinity, low concentrations of essential macronutrients such as N and P, extreme pressure, high concentration of heavy metals, and lack of light as an energy source). Taxa commonly encountered from such formations are often (poly)extremophiles e.g., members of the genera *Haloanaerobacter*, *Frackibacter*, *Aquifex*, and *Thermotoga* [14, 15]. Recent studies have provided excellent overviews on the impact of operational history and changes in the formation's geochemical characteristics on microbial community dynamics (e.g. Ref. [15]), as well as interspecies metabolic dependencies within such ecosystems (e.g. Ref. [16]). However, relatively less information is available on the presence, identity, proportion, and genomic/metabolic characteristics of the rare biosphere in ONG formations, as well as putative promotion/demotion dynamics occurring as a result of operational perturbation.

Efforts to improve energy recovery from tertiary oil reservoirs by microbial stimulation and/or inoculation have been an active area of research during the last two decades [12, 17]. One approach, microbially enhanced energy recovery (MEER), depends on the stimulation of anaerobic

hydrocarboniclastic microbial activity *in-situ* by the addition of nutrients [17] or microbial consortia [18, 19] to stimulate hydrocarbon breakdown to recoverable methane gas via methanogenesis. Within the context of these schemes, understanding the microbial community response to various types of nutrient stimulations is key for predicting MEER success. Nutrient stimulation during MEER schemes could theoretically lead to the enrichment of a rare member(s) of the *in-situ* community to higher levels of relative abundance, allowing for their identification and genomic recovery.

As part of a wider project to characterize the microbial community response to nutrient stimulation in ONG formations in north central Oklahoma, we encountered a novel, previously unidentified phylogenetic lineage in enrichments set up using production water (17.6% salinity), oil, and isolated soy proteins as a supplement, and incubated under conditions simulating the reservoir in-situ characteristics (50°C, anaerobic). Here, we present a detailed analysis of the lineage salient genomic features, physiological preferences, and metabolic capacities derived from its near complete genome recovered from such enrichments. We also report on its global distribution patterns via mining publicly available amplicon databases. We propose the name *Candidatus* Mcinerneyibacterium aminivorans gen. nov, sp. nov. to accommodate ARYD3, the type material is deposited under the BioProject accession number PRJNA558942, BioSample accession number SAMN12502798, and genome WGS accession number VSIX00000000. This also serves as the type material for the novel candidate family Mcinerneyibacteriaceae fam. nov., order Mcinerneyibacteriales ord. nov., class Mcinerneyibacteria class nov., and phylum Mcinerneyibacteriota phyl. nov.

1.3 Materials and methods

1.3.1 DNA samples

Samples were obtained as part of a wider effort to investigate the utility of nutrient stimulation for enhancing methanogenesis in tertiary oil reservoirs [17]. Enrichments analyzed in this study were set up using oil field and formation water samples from Cushing oil field (Well head of Yahola # 17 well: 35.77 N, 96.48 W) located in north-central Oklahoma [17]. Production water from this site is characterized by high salinity (17.6%) and moderately high temperature (50°C). The production waters had relatively low concentrations of sulfate (69.24 mg/L), nitrate (0.13mg/L), and iron (23.37mg/L). The enrichment was supplemented with isolated soy protein (0.2%, MP Biomedicals, LLC), an inexpensive protein-rich (90% w/w) source that could also provide the needed N, P, as well as essential trace metals lacking in oil field and potentially preventing growth and stimulation.

1.3.2 Sequencing, assembly, and binning

Samples from Yahola#17 were extracted using DNeasy PowerSoil kit (Qiagen, Valencia, CA, USA). Sequencing of the extracted DNA was conducted using the services of a commercial provider (Novogene, Beijing, China). A total of 24.2 Gbp of raw data were obtained. High quality reads were obtained using iu-merge-pairs (https://github.com/merenlab/illumina-utils/blob/master/scripts/iu-merge-pairs). Reads that passed quality control (91.94%) were assembled into contigs using default settings of MegaHit [20], and a minimum contig length of 1000 bp. The identified contigs were binned into metagenome-assembled genomes (MAGs) using MaxBin [3] with the default parameters. CheckM

[21] was used for estimation of genome completeness, strain heterogeneity, and contamination. Sequencing and assembly statistics are shown in Table 1.1.

Paired-end raw reads	36,781,215
Paired-end reads passing QC	33,818,355 (91.94%)
Median read length (bp)	125
Reads in assembled contigs	66,696,560
Number of contigs	15,891
Total assembly size (bp)	28,105,589
Longest contig (bp)	504,444
N50 (bp)	15,200

Table 1. 1 Sequencing and assembly statistics of Yaholla17 enrichment

1.3.3 Phylogenetic and phylogenomic analysis

16S rRNA genes encountered in the binned MAGs were utilized for phylogenetic analysis. Alignment with 16S rRNA gene sequences from pure cultures and culture-independent studies was conducted using SINA aligner v1.2.11 implemented in Silva webserver [22], and maximum likelihood trees were constructed using RaxML [23]. Phylogenomic analysis was conducted using a concatenated alignment of 120 single-copy markers implemented in GTDB-TK [5, 24]. Genomes were mined for the 120 single copy marker genes, that were then aligned to their respective HMM models in the 'identify' step of GTDB-TK. The resulting alignments were concatenated in the 'align' step of GTDB-TK and the concatenated alignment is then used by pplacer to find the maximum likelihood placement of each genome in the global GTDB-TK reference tree in the 'classify' step of GTDB-TK. Concatenated alignments of the predicted amino acids for the 120 marker genes from selected genomes were then extracted from the GTDB-TK concatenated alignment fasta files and used to construct maximum-likelihood trees in RaxML [23] using 100 bootstrap reiterations. Putative taxonomic ranks for obtained MAGs were also deduced using average Amino Acid Identity (AAI) and Shared Gene Content (SGC); calculated using AAI calculator [25].

1.3.4 Structural, physiological, and metabolic characterization

The general genomic features for the obtained MAGs were deduced using JGI IMG annotation pipeline. KEGG pathways and Metacyc database were used to predict the organisms' metabolic potential, putative growth preferences and other relevant physiological pathways. Annotations of key metabolic genes were confirmed by Blastp search against nr database [26] and phylogenetic assessment with reference taxa. Similarly, the absence of key relevant genes was further confirmed using local Blastp of a reference gene against the genomic assembly at hand. The subcellular protein localization pattern was predicted using PSORTb [27]. Other features, such as cell wall structure, extracellular structures (flagella, pili, fimbriae), cell shape, enzymes necessary for adaptation to harsh environment conditions were predicted by querying the relevant genes using local Blastp against the genomic assembly. MAGs were queried for putative transporters, and peptidases using Transporter Classification Database (TCBD), and MEROPS, respectively.

1.3.5 Ecological distribution

To identify the global ecological distribution pattern of candidate phylum ARYD3, we queried GenBank nucleotide (nt) database using ARYD3 near full-length 16S rRNA gene sequence (last accessed on July 30th, 2019). We further queried the occurrence of ARYD3 within the recently published Earth Microbiome Project [28] that contains 2.1 billion partial 16S rRNA gene sequences in 27,068 datasets (downloaded as individual studies from Qiitaa [https://qiita.ucsd.edu/emp/]) using local blastn, with the truncated V3 region of ARYD3 16S rRNA sequence used as query. A conservative cutoff of 85% sequence identity and 70% sequence coverage was utilized to identify sequences affiliated with the ARYD3 candidate phylum.

1.3.6 Enrichment and isolation attempts

Samples for metagenomic sequencing and recovery were unfortunately obtained as DNA samples to our research group. As such, efforts for visualization, enrichment, or isolation of this novel candidate phylum from the original sample were not feasible. After realizing that the original DNA samples provided harbored representatives of a novel lineage, we attempted to recreate these enrichments by going back to the environmental source and setting up the same enrichment samples once again. We managed to obtain a different water sample from the exact same oil production well from which the original enrichments were set up.

In an attempt to re-enrich novel lineage ARYD3, anaerobic enrichments were set up using production water, oil, and isolated soy protein as described earlier using modified Hungate technique at pH 7, 17.6% NaCl, and 50°C incubation temperatures [29]. As well, enrichments using substrates predicted from genomic analysis to support growth of ARYD3 were also set up. At regular intervals, 1 ml samples were obtained, and DNA was extracted using Qiagen PowerPlant DNA extraction kit. Small subunit rRNA gene amplification was conducted using general bacterial primers (338F-518R) as described previously [17]. PCR products obtained were sequenced using Illumina iSeq100 sequencing platform to investigate the community developing and determine whether the novel lineage was enriched. Sequence alignment, operational taxonomic unit (OTU) generation, and classification were conducted in Mothur [30] as described earlier [17]. Sequence classification against the Silva database was conducted after the addition of reference ARYD3 16S rRNA gene sequence to the database files.

1.3.7 Sequence deposition

MAGs from this effort were deposited at DDBJ/ENA/GenBank under the Whole Genome Shotgun BioProject accession number PRJNA558942, and BioSample accession numbers SAMN12502796-SAMN1202798 and WGS accession numbers VSIV00000000, VSIW00000000, and VSIX00000000.

1.4 Results

1.4.1 A defined microbial community enriched on isolated soy proteins

16S rRNA gene analysis of DNA samples from enrichments of Yaholla production water (17.6% salinity), oil, and 2% isolated soy protein incubated at 50°C for 100 days yielded a community dominated (98.7% of reads) by three OTUs, a drastic reduction in diversity from the community identified in production water prior to enrichments (Table S1). One OTU (ARYD1) exhibited high (99.8%) sequence similarity to a cultured representative, *Flexistipes sinusarabici* strain MAS 10 (DSM 4947) [31], the type species for the genus *Flexistipes* within the Phylum Deferribacteres (Figure 1.1A). The second enriched OTU (ARYD2) exhibited 93.5% sequence similarity to its closest cultured relative: *Kosmotoga pacifica* strain DSM 26,965 [32], a member of the phylum Thermotogae; and a higher similarity to 16S rRNA gene sequences recovered from culture-independent surveys of olive oil processing wastewater, and highly saline brine of a geothermal plant in North German Basin [33, 34] (Figure 1.1A). Surprisingly, the third member of the community (ARYD3) showed no clear affiliation with any specific bacterial lineage (78–79% sequence similarity to multiple phylogenetically disparate strains, e.g., the gamma Proteobacterium *Halothiobacillus neapolitanus*, as well as several Gram-positive species belonging to the genera *Bacillus*, *Lactobacillus*, and *Gracilibacillus*. Phylogenetic analysis (Figure 1.1A) placed this OTU as a novel, phylum-level monophyletic branch that is distinct from all known bacterial phyla, together with nine other 16S rRNA sequences from the sulfidogenic deep hypersaline Urania basin, salt ponds in wetland restoration areas in San Francisco, anaerobic digesters treating wastewater, hydrothermal vent microbial mats in Guaymas basin, and coastal marine sediments (Figure 1.1A).

1.4.2 Genomic recovery and phylogenomic analysis

Metagenomic sequencing, assembly, and binning recovered near complete genomes of the three dominant OTUs identified in the enrichment (Table 1.2). Phylogenomic analysis using a concatenated alignment of 120 single copy marker genes (Figure 1.1B), AAI values, and shared gene content supported the placement of ARYD1 genome as a member of *Flexistipes sinusarabici* species (AAI value 90.81% and shared gene content 37.26% with *Flexistipes sinusarabici* DSM 4947), and the placement of ARYD2 as a representative of a novel genus within the family Kosmotogaceae (average AAI of 54.71 ± 2.15% and average shared gene content of 28.36 ± 1.41% to other members of the family Kosmotogaceae) (Figure 1.1B). ARYD3, on the other hand, exhibited a unique placement in the GTDB-TK reference protein tree as a sister branch to the Spirochaetes and the uncultured phylum UBP7 (Figure 1.1B). GTDB-TK reports relative evolutionary divergence (RED) values for genomes, which is a measure of the relative placement between the root and the node's descending tips.

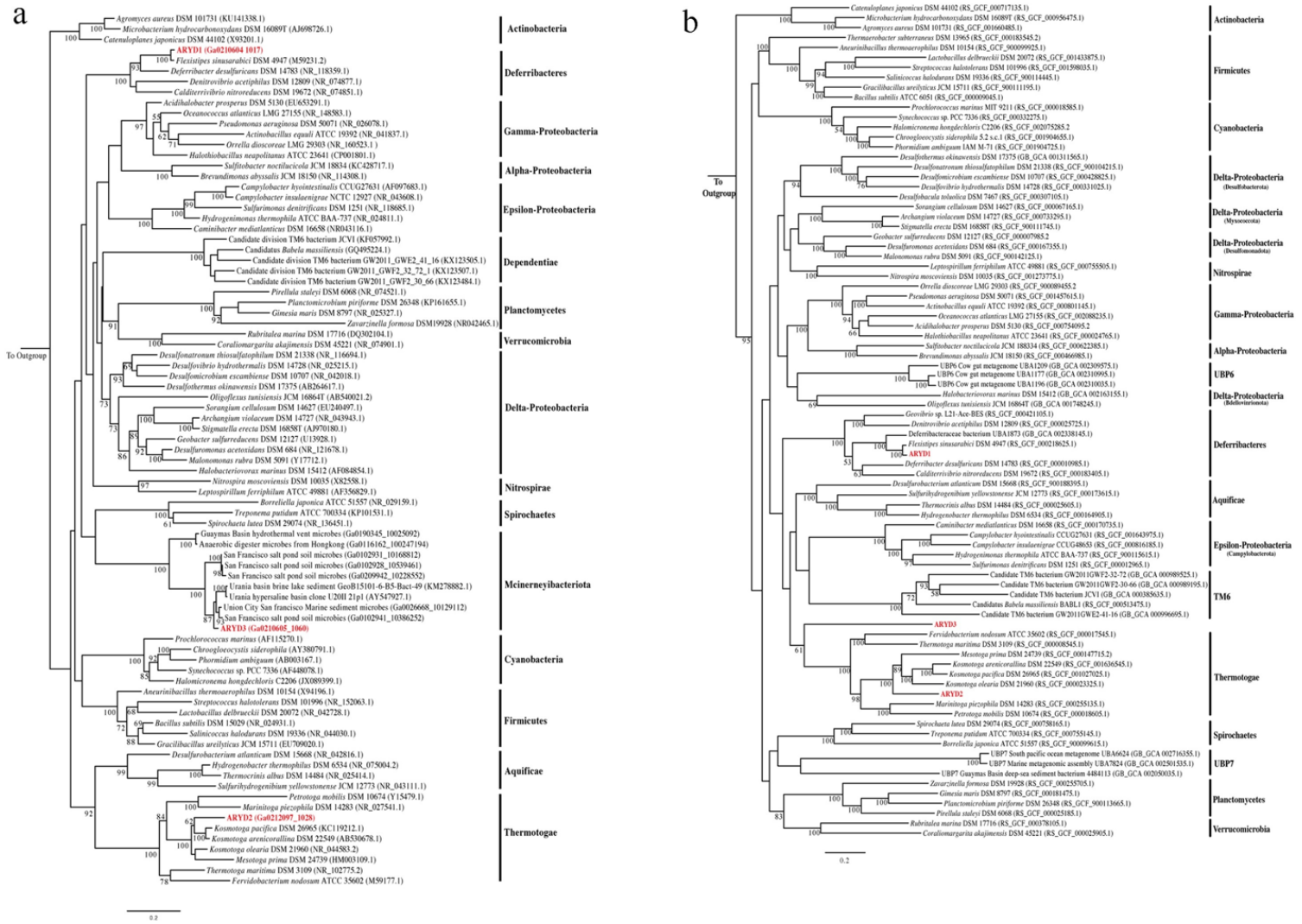


Figure 1. 1 Phylogenetic trees based on 16S rRNA genes (A) and concatenated alignments of 120 single-copy marker proteins (B) showing the phylogenetic position of ARYD1 (*Flexistipes sinusarabici*), ARYD2 (Kosmotogaceae), and ARYD3 (Mcinerneyibacteriota). Reference taxa are type strains of cultured isolates in both trees, uncultured 16S rRNA gene sequences related to lineages of interest, and genomes of yet-uncultured taxa. Some taxa (e.g., phyla UBP6 and UBP7 are not represented in 16S rRNA phylogeny due to the unavailability of 16S rRNA genes in the available genome assemblies. GenBank or IMG gene accession numbers are shown in parenthesis. Both trees were rooted with *Chloroflexus aggregans* DSM 9485 (NR_074226.1) as an outgroup. Bootstrap values (from 100 replicates) are shown for nodes with more than 50 bootstrap supports. Phyla are named according to NCBI taxonomy, except in cases where a phylum is solely recognized based on GTDB taxonomy (e.g., UBP7).

In general, median RED value thresholds of 0.32, 0.47, 0.63, 0.78, and 0.94 are indicative of phylum, class, order, family, and genus ranks, respectively. ARYD3 genome had a RED value of 0.304 to its closest relative in the tree (candidate phylum UBP7) suggesting its placement as a novel phylum. In addition, ARYD3 exhibited low AAI values and shared gene content to other phyla that clustered close to it in the protein tree (Table 1.3). Collectively, these 16S and genome-based similarity indices suggests that ARYD3 is a representative of a new phylum rather than a member of a previously characterized phylum.

Bin name	Number of contigs	N50	L50	Completeness (%)	Contamination (%)
ARYD1	269	12,411	39	76.94	0.86
ARYD2	159	70,472	11	99.84	5.17
ARYD3	167	49,318	11	95.51	1.12

Table 1. 2 Statistics of MAGs recovered from Yaholla17 enrichment

Phylum	ARYD3 genome		
	Number of genomes examined	AAI (average ± SD)	SGC (average ± SD)
Epsilonproteobacteria	10	36.22 ± 1.08	19.48 ± 1.25
Deferribacteres	3	37.28 ± 0.08	18.63 ± 2.77
Dependentiae	18	34.84 ± 0.64	16.61 ± 0.68
UBP6	4	37.00 ± 0.49	20.41 ± 0.54
Aquificae	3	37.55 ± 0.81	20.90 ± 1.15
UBP7	3	35.03 ± 0.31	19.50 ± 1.18
Spirochaetes	3	35.85 ± 0.40	18.20 ± 0.94
Deltaproteobacteria	5	37.43 ± 1.76	17.00 ± 2.19
Firmicutes	5	35.32 ± 0.88	17.96 ± 0.90
Actinobacteria	5	33.66 ± 0.89	13.73 ± 3.71
Tenericutes	3	35.44 ± 0.68	14.59 ± 2.19
Planctomycetes	3	33.81 ± 0.36	12.24 ± 2.09
Thermotogae	10	38.52 ± 0.36	22.07 ± 0.89

Table 1. 3 Average (±SD) amino acid identities and shared gene contents between Mcinerneyibacteriota ARYD3 genome and representatives of bacterial phyla exhibiting close relationship to ARYD3 in the protein tree.

The following genomes were used to calculate the AAI and SGC values in comparison to ARYD3: Campylobacterota (GCF_000620965.1, GCF_002119425.1, GCF_000744435.1, GCF_000523355.1, GCA_002328495.1, GCA_000276965.1), Deferribacteres (GCA_002348105.1, GCF_000487995.1, GCA_002452335.1), Dependentiae (GCA_003506475.1, GCA_003484685.1, GCA_000989955.1, GCA_000992305.1, GCA_000989185.1, GCF_000513475.1, GCA_003511635.1, GCA_003506475.1, GCA_000989185.1, GCA_000989195.1, GCA_000385635.1, GCA_001771315.1, GCA_001468855.1, GCA_001468865.1, GCA_002441565.1, GCA_001771315.1, GCA_000989185.1, GCA_000385635.1, GCA_000262655.1), UBP6 (GCA_002309575.1, GCA_002310035.1, GCA_002310995.1, GCA_002366955.1), Aquificae (GCF_900188395.1, GCF_000703085.1, GCF_000702425.1), UBP7(GCA_002501535.1, GCA_002456105.1, GCA_002050035.1), Spirochaetes (GCF_900099615.1, GCF_000758165.1, GCF_000755145.1), Deltaproteobacteria (GCA_002841865.1, GCA_001311565.1, GCA_000429325.1, GB_GCA_002771315.1, GCA_002771815.1), Firmicutes (GCF_000183545.2, GCA_002919105.1, GCA_003242675.1, GCA_000237975.1, GCF_001953275.1), Actinobacteria (GCA_001052995.1, GCA_001421565.1, GCF_000717135.1, GCA_000155145.1, GCA_000433855.1), Tenericutes (GCA_000433455.1, GCA_003298555.1, GCA_000965765.1), Planctomycetes (GCF_000255705.1, GCF_000181475.1, GCF_900113665.1), and Thermotogae (GCF_001636545.1, GCF_000023325.1, GCF_000008545.1, GCA_001509385.1, GCF_000147715.2, GCF_001719065.1, GCF_000017545.1, GCF_001027025.1, GCF_000255135.1, GCF_900129175.1). NCBI assembly accession numbers are in parenthesis. Phyla are named according to NCBI taxonomy, except in cases where a phylum is solely recognized based on GTDB taxonomy (e.g. UBB6 and UBP7).

1.4.3 Analysis of *Flexistipes sinusarabici* strain ARYD1 and Kosmotogaceae ARYD2 genomes

Analysis of *Flexistipes sinusarabici* strain ARYD1 genome revealed similar catabolic capacities to those reported for *Flexistipes sinusarabici* type strain MAS 10 (DSM 4947) [31, 35]. Complex proteins, rather than carbohydrates, appear to be the most plausible carbon and energy source. Anaerobic lifestyle was also predicted based on the absence of a complete aerobic respiratory chain and the identification of dissimilatory nitrate reduction to ammonium genes (*narGHI* and *nirBD*). This finding suggests that within the enrichment, strain ARYD1 initially couples the reduction of residual nitrate present in formation water to peptide degradation. Indeed, nitrate concentrations decreased from 3 mM in pre-enrichment production water to below detection limit after 100 days of enrichment. Subsequently, ARYD1 could grow fermentatively post nitrate depletion.

Analysis of the ARYD2 Kosmotogaceae genome suggested a predominantly saccharolytic role, with glucose, fructose, cellobiose, sucrose, trehalose, xylose, and xylitol utilization capabilities. The presence of genes encoding peptidases as well as the identification of complete pathways for some amino acid degradation suggest that a peptidolytic lifestyle is also possible. Absence of genes encoding respiratory chain proteins allude to a fermentative lifestyle. No genes for nitrate, sulfur, or cysteine (as previously shown for some *Kosmotoga* species [36, 37]), or sulfate reduction were identified in the genome indicating that anaerobic respiration is not feasible. This suggests that, within the enrichment, ARYD2 grows fermentatively on peptides and residual carbohydrates.

1.4.4 Analysis of ARYD3 genome

General genomic and structural features. ARYD3 possesses a small genome (2.19 Mbp assembly, estimate 2.31 Mbp), with an extremely low GC content (30.15%, Table 1.4). Structurally, ARYD3 is predicted to have an outer membrane of lipopolysaccharide based on the identification of genes encoding enzymes for lipid A and core oligosaccharide biosynthesis, as well as the outer membrane protein assembly complex Bam, and the outer membrane periplasmic chaperone Skp (Figure 1.2). An outer S-layer is also predicted based on the identification of an S-layer homology domain (pfam00395) (Figure 1.2). ARYD3 is predicted to possess non-flagellated non-chemotactic rod-shaped cells, based on the identification of the rod shape determining proteins MreBCD and RodA and the absence of flagellar assembly genes and chemotaxis-related genes. Genes encoding type IV pilus assembly were identified. The organism is predicted to have multiple defense mechanisms, including the CRISPR-Cas system, type I restriction endonucleases, and the oxidative stress enzymes superoxide reductase, alkyl hydroperoxide reductase, rubrerythrin, and rubredoxin.

Adaptive capabilities. ARYD3 was enriched at 50°C in the presence of 17.6% salt, necessitating the possession of adaptive mechanisms to cope with high salinity and elevated temperature. No evidence for the biosynthesis of unusual glycerophospholipids for thermal adaptation was detected in the genome; and we reason that this is a reflection of the relatively moderately thermal, rather than hyperthermophilic, growth temperature (50°C).

	ARYD1	ARYD2	ARYD3
Genome size (bp)	2,410,627	3,216,074	2,190,304
Genome completeness (%)	76.94	99.84	95.51
Genome contamination (%)	0.86	5.17	1.12
% Coding bases	90.07%	89.55%	90.35%
% GC content	38.54	35.93	30.15
Total number of genes	2594	3231	2169
Average gene length	854	915	927
Number of protein coding genes	2537	3147	2131
With function prediction	2051	2389	1635
With COG categories	1596	1885	1302
Subcellular localization			
Cytoplasmic	52.90%	53.6%	56.5%
Extracellular	0.39%	0.44%	0.61%
Periplasmic	1.38%	1.33%	0.37%
Membrane bound	2.28%	21.30%	20.70%
Unknown	23.29%	23.10%	21.60%
Number of non-protein coding genes	57	84	38
tRNA genes	41	60	31
rRNA genes	9	2	5
16S rRNA	3	0	2
23S rRNA	2	1	3
5S rRNA	4	1	0

Table 1. 4 General genomic features of ARYD1 (*Flexistipes sinusarabici*), Kosmotogaceae (ARYD2), and ARYD3 (McInerneybacteriota) MAGs.

Osmoadaptation to high salinity conditions usually involves a salting-in strategy (uptake and intracellular accumulation of molar concentrations of K^+), a compatible solute strategy (the uptake and/or synthesis of organic molecule(s) that do not interfere with intracellular enzymatic activities and cellular processes), or a combination of both strategies [38, 39]. Within ARYD3 genome, we identified the *trkAH* system for potassium uptake. While known to mediate K^+ import and facilitating K^+ intracellular accumulation, the *trkAH* system is widely spread in the majority of bacterial and archaeal phyla, and thus should not be taken as a definite proof for the use of a salting-in strategy. Indeed, the hypothetical average pI of the theoretical ARYD3 proteome is quite neutral (8.06), in contrast to the acidic proteome commonly associated with high intracellular K^+ level in microorganisms utilizing a salting-in strategy [38]. On the other hand, genes for the synthesis/uptake of multiple compatible solutes (osmoprotectants) were identified in ARYD3 genome, including 2-phosphoglycerate kinase and cyclic 2,3-diphosphoglycerate synthase for the synthesis of cyclic 2,3-diphosphoglycerate (cDPG), and glucosylglycerate synthase for synthesizing α-glucosylglycerate from UDP-glucose and 3 phospho-D-glycerate (Figure 1.2). In addition, the genome also suggests the capability of the biosynthesis of beta glutamine, commonly used as a compatible solute. Finally, glycine betaine transporters were identified suggesting its additional capacity for the uptake and utilization of glycine betaine as an osmoprotectant.

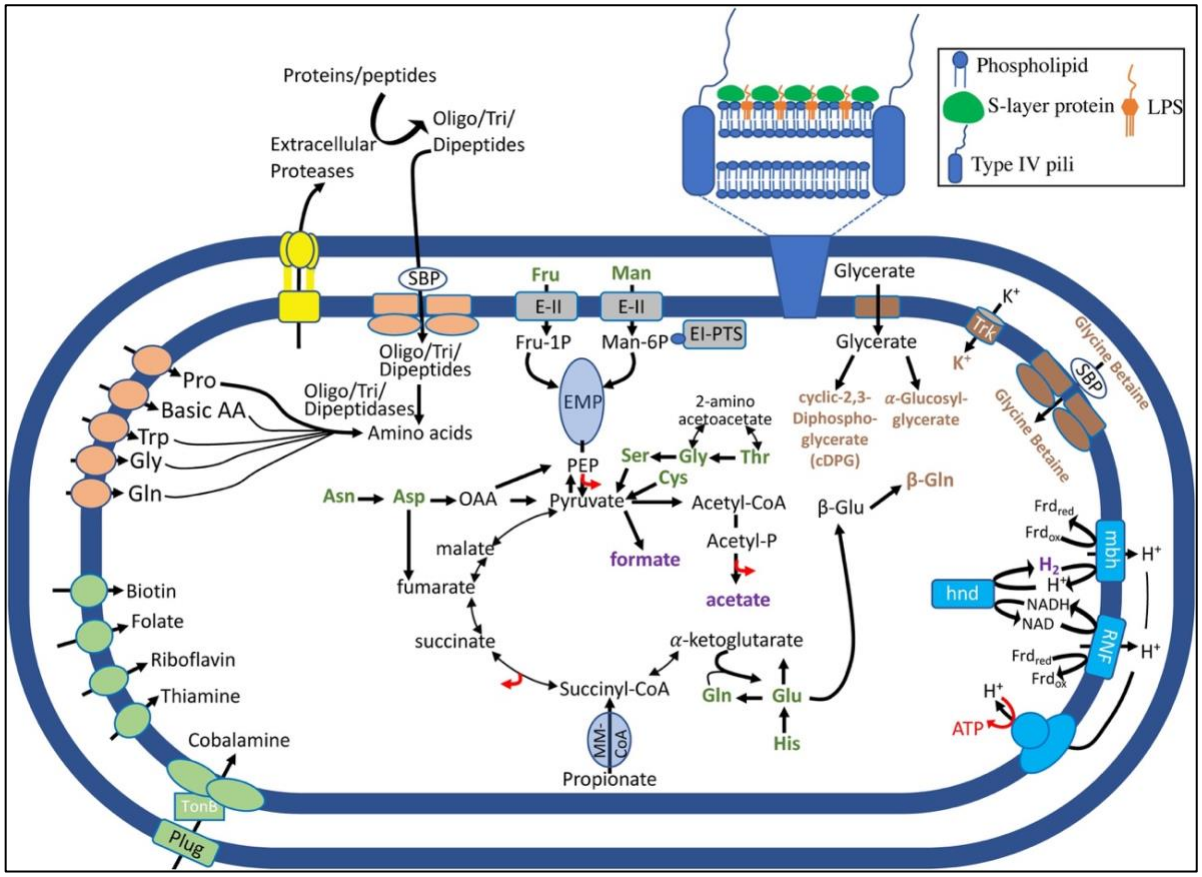


Figure 1. 2 Metabolic reconstruction of ARYD3 MAG. All possible substrates potentially supporting growth are shown in bold green, while predicted final products are shown in bold purple. Sites of ATP production at the substrate level are shown by a bent red arrow and include ATP production during glycolysis (EMP), the action of acetate kinase, and the action of succinyl-CoA synthase. Cell membrane (phospholipid bilayer), cell wall (Gram-negative cell wall with possible S-layer), and external structures (type IV pili) are shown in the inset. Membrane proteins are color coded as follows: amino acids and peptide transporters are shown in orange, cofactor transporters are shown in green, while sugar transporters are shown in grey. Compatible solute glycine betaine transporter as well as the potassium transporter (Trk) are shown in brown. Also, in brown, is the transporter involved in the uptake of glycerate, the precursor to the compatible solutes cyclic-2,3-diphospho-glycerate (cDPG) and α-glucosyl-glycerate. Membrane proteins involved in recycling reduced equivalents (e.g., reduced ferredoxin and NADH) are shown in blue and are labelled by the protein name as follows: Rnf, membrane-bound RNF complex coupling ferredoxin oxidation to NAD reduction; Mbh, membrane bound [Ni–Fe] hydrogenase involved in coupling ferredoxin oxidation to proton reduction; Hnd, cytoplasmic Fe-only hydrogenase involved in re-oxidizing NAD(P)H at the expense of H_2 production. Proton motive force created via the action of RNF and Mbh is shown as H^+ in the periplasmic space. Protons movement through the ATPase complex (also shown in blue) can then lead to ATP production (red arrow) via oxidative phosphorylation. Finally, type I secretion system, possibly used for the export of extracellular proteases, is shown in yellow. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Abbreviations: Asn, asparagine; Asp, aspartic acid; Cys, cysteine; EI-PTS, enzyme I of the phosphotransferase system shown with the Hpr enzyme as the blue circle; E-II, enzyme II of the PTS; EMP, Embden Myerhoff pathway; Frd (ox, and red), ferredoxin oxidized, and reduced, respectively; Fru, fructose; Fru-1P, fructose-1-phosphate; Gln, glutamine; Glu, glutamate; Gly, glycine; His, histidine; Man, mannose; Man-6P, mannose-6-phosphate; MM-CoA, methylmalonyl-CoA pathway for propionate conversion to succinyl-CoA; OAA, oxaloacetate; PEP, phosphoenol pyruvate; Pro, proline; SBP, substrate binding protein of the ABC transporter; Trp, tryptophan.

Metabolic capacities of ARYD3

Fermentative capacities. ARYD3 genome lacks a complete respiratory chain, as evident by the absence of succinate dehydrogenase, cytochrome reductase, or cytochrome oxidase complexes. This finding, coupled with the incomplete TCA cycle (no genes encoding conversion of oxaloacetate to 2-ketoglutarate) and the absence of the oxidative branch of the pentose phosphate pathway all argue for a predominantly anaerobic, heterotrophic, and fermentative mode of metabolism. A complete glycolysis/gluconeogenesis pathway was identified (Figure 1.2).

Analysis of the predicted metabolic pathways and transporter repertoire suggest a primarily aminolytic and a relatively narrow saccharolytic capacity. ARYD3 genome encodes extracellular proteases, a wide range of oligo/dipeptide, and single amino acids (proline, lysine, glycine, glutamine, asparagine, glutamate, arginine, histidine, cysteine) transporters, as well as cytoplasmic oligopeptidases/dipeptidases to potentially break down the imported oligo/dipeptides into amino acids, all of which suggest the organism's aminolytic capacity (Figure 1.2). Indeed, peripheral enzymes enabling the utilization of a wide range of at least 11 amino acids (alanine, asparagine, aspartate, glutamate, glutamine, glycine, serine, cysteine, methionine, histidine, and threonine) by channeling them to central metabolic pathways were identified. On the other hand, few sugars e.g., glucose, fructose, and mannose appear to support AYRD3 growth (Figure 1.2).

Reducing equivalents accumulated during operation of the EMP and the partial TCA cycle are most probably disposed of by reducing pyruvate to formate (via pyruvate formate lyase), and acetate (via pyruvate:ferredoxin oxidoreductase, phosphate acetyltransferase, and acetate kinase) as the main fermentation end products. Succinate is also thought to be produced as an end product of histidine, glutamine, and glutamate metabolism via the partial TCA cycle. ATP production could occur via substrate level phosphorylation during glycolysis, as well as the action of acetate kinase, and succinyl-CoA synthase (EC 6.2.1.5).

Syntrophic propionate degradation: While the genome does not suggest the capacity to degrade long chain fatty acids, it does encode genes required for the degradation of propionate via the methylmalonyl-CoA pathway to acetate. Hydrogen production is speculated to occur during the operation of the membrane-bound Ni–Fe hydrogenase (*mbh*), as well as the cytoplasmic NAD(P)-dependent Fe-only hydrogenase (*hnd*) identified in the genome. Propionate degradation coupled to hydrogen production in fermentative organisms is thermodynamically unfeasible ($\Delta G^{\circ\prime} = +72 - 76$ KJ per reaction under standard conditions [40, 41]). The process is only alleviated via syntrophic interaction of propionate degraders with H_2-users (e.g., hydrogenotrophic methanogens). Efficient hydrogen transfer between the two partners is known to occur through biotic conduits, e.g., flagella, pili, or nanowires [42, 43]. ARYD3 genome lacked evidence for homologs of the outer membrane multiheme cytochromes MtrC or OmcA of *Shewanella* previously shown to localize to extensions of the outer membrane and the periplasm and to act as nanowires for electron transfer [43]. Similarly, no flagellar assembly-encoding genes were identified in the genome. However, type IV pili seem to be encoded by ARYD3 and could potentially serve as the biotic conduit for inter-species hydrogen transfer. Another hallmark of syntrophic organisms is their energy conservation through the use of electron-bifurcating mechanisms that are employed for recycling of reducing equivalents. ARYD3 genome encodes an RNF complex known to oxidize reduced ferredoxin (produced via the action of

pyruvate:ferredoxin oxidoreductase) at the expense of NAD with the concomitant proton extrusion to the periplasmic space. Reduced ferredoxin can also be recycled via the action of the membrane-bound Ni–Fe hydrogenase (Mbh) here at the expense of proton reduction to molecular hydrogen and again with the concomitant proton extrusion to the periplasm. The action of these two membrane-bound complexes (Rnf and Mbh) lead to the creation of a proton motive force that could subsequently be used for ATP generation via oxidative phosphorylation using the F-type ATPase identified in the genome. Reduced NAD that is produced during RNF complex operation can then be re-oxidized with the concomitant H_2 production via the cytoplasmic NAD(P)-dependent Fe-only hydrogenase Hnd (Figure 1.2).

Anabolic capacities. Genomic analysis suggests an auxotrophy for a few amino acids (His, Arg, Val, Leu, Ile, Met, Trp, and Lys). These auxotrophies could theoretically be satisfied by the presence of specific transporters or via the action of cytoplasmic oligo/tri/dipeptidases on imported oligo/tri/dipeptides. Similarly, the organism seems to be auxotrophic for some cofactors (biotin, heme, and vitamin B_{12}) for which specific transporters were identified in the genome.

1.4.5 Global ecological distribution

As described above, few near full-length 16S rRNA gene sequences from GenBank NR and IMG-MG database were affiliated with ARYD3 (Figure 1.1A) and were mostly identified in saline habitats (Figure 1.1A). Such rarity suggests that members of ARYD3 lineage are seldom, if ever, present in relative abundance thresholds that readily enable their detection using 16S Sanger or shotgun metagenomics approaches.

Nevertheless, we reasoned that members of the ARYD3 might exhibit a wider distribution pattern as a component of the (extremely) rare biosphere in various habitats. To this end, we attempted to identify the occurrence of ARYD3 in the extensive Earth Microbiome Project. A total of 1088 ARYD3 sequences were identified (0.000053% of all EMP sequences), with at least one ARYD3 sequence encountered in 318 out of 27,068 datasets (Table S2) examined. In all datasets where ARYD3 sequences were identified, the lineage was invariably present in extremely low numbers. The highest relative abundance of ARYD3 (>0.01%) was mostly encountered in metagenomes obtained from saline habitats e.g., salt ponds in California and Puerto Rico (Table 1.5, Table S2). In addition, ARYD3 was also identified in extremely low abundance (<0.01%) in a wide range of non-saline and non-thermal environments, e.g., freshwater lakes, forest soil, grassland soil, bat feces, coral tissue, mammalian gut and skin, groundwater, hydrocarbon impacted environments, and high temperature microbial mats (Table S2).

Study name	GenBank Accession	Dataset Description	Habitat Classification	Total sequences
Halophilic communities from Puerto Rico and San Francisco as a source for novel lignocellulolytic enzymes	ERR1590017	Saline lake sediment from a salt pond (6.5% salinity) in California.	Aquatic/Marine	46089
Halophilic communities from Puerto Rico and San Francisco as a source for novel lignocellulolytic enzymes	ERR1590041	Saline lake sediment from a salt pond (9.5% salinity) in Cabo Rojo, PR.	Aquatic/Marine	105546
Halophilic communities from Puerto Rico and San Francisco as a source for novel lignocellulolytic enzymes	ERR1590018	Saline lake sediment from a salt pond (6.6% salinity) in California.	Aquatic/Marine	75329
A decade of seasonal dynamics and co-occurrences within freshwater bacterioplankton communities from eutrophic Lake Mendota, WI, USA	ERR1547119	Epilimnetic water sample from Mendota Lake, WI.	Temperate freshwater lakes	289500
Co-diversification of marine mammals and their skin microbiomes	ERR1593108	"Sterile" water from research facility in Boulder, CO, USA.	Urban biome	2730
Halophilic communities from Puerto Rico and San Francisco as a source for novel lignocellulolytic enzymes	ERR1590023	Saline lake sediment from a salt pond (33.2% salinity) in California.	Aquatic/Marine	92020
Gut microbiota and health in wild and captive colobine primates	ERR1511966	Fecal sample of spider monkey a West Lafayette, IN, USA Zoo.	Host associated/mammals	4783
Diversity of carbonate deposits and basement rocks in continental and marine serpentinite seeps in alkaline springs of Italy and California	ERR1524003	Sediment sample from alkaline habitat in Canada from an olivine colonization experiment. Salinity 2.1%, pH 12.	Non-marine Springs	5589
Halophilic communities from Puerto rico and San Francisco as a source for novel lignocellulolytic enzymes	ERR1590022	Saline lake sediment from a salt pond (33.2% salinity) in California.	Aquatic/Marine	93183
Halophilic communities from Puerto rico and San Francisco as a source for novel lignocellulolytic enzymes	ERR1590016	Saline lake sediment from a salt pond (7.0% salinity) in California.	Environmental/ Aquatic/Marine	99045
Gut microbiota and health in wild and captive colobine primates	ERR1511836	Spider monkey CPZX12C. Fecal sample of spider monkey from Columbian Park Zoo. Purdue-West Lafayette	host associated/mammals/ Feces	18610
Halophilic communities from Puerto Rico and San Francisco as a source for novel lignocellulolytic enzymes	ERR1590030	Filtered hypersaline water sample taken from salt pond in Cabo Rojo, Puerto Rico. Salinity 19%.	Environmental/ Aquatic/Marine	88728

Table 1. 5 Datasets with >0.01% Mcinerneybacteriota relative abundance in Earth Microbiome Project. A full list of datasets where ARYD3 sequences are identified is presented in Table S2.

1.4.6 Genome-guided attempts to enrich ARYD3

Samples for metagenomic sequencing and recovery were unfortunately delivered as extracted DNA to this research group. Multiple efforts to re-enrich ARYD3 were undertaken by going back to the environmental source (Yaholla oil field) and setting up enrichments using a new water sample from the same oil production well from which the original enrichments were derived. Further, additional enrichments were set up guided by physiological preferences and putative metabolic abilities predicted from its genomic content. Multiple enrichments were conducted at 50°C and 17.6% NaCl using isolated soy protein, yeast extract, casamino acids, serine, or mannose. As well, additional enrichments at lower salinity (10%) with the same substrates were undertaken, given the predicted halotolerant rather than halophilic nature of ARYD3. The enrichment was set up using frozen water samples as well as samples from adjacent wells as described previously [17], with isolated soy protein, yeast extract, or mannose as a carbon source. Progress was monitored using 16S rRNA gene sequencing.

Unfortunately, all efforts did not lead to any increase in ARYD3 relative abundance to levels achieved in previous enrichments, and the defined microbial community was not replicated, but rather communities dominated by Firmicutes with enrichment up to 95% was consistently obtained (Table S3). The relative proportion of ARYD3 in multiple enrichments never exceeded 0.09% of the total community.

1.5 Discussion

In this study, we identified a novel candidate phylum from enrichments derived from a tertiary oil reservoir, and characterized its structural features, physiological preferences, and metabolic capabilities using genome-resolved metagenomics. The genome was co-enriched with a strain of *Flexistipes sinusarabici*, and a representative of a novel genus within the family Kosmotogaceae under anaerobic, elevated temperature (50°C) and high salinity (17.6% NaCl) conditions, using formation water and crude oil amended with (0.2% w/v) isolated soy protein. Genomic analysis suggests a primarily fermentative and aminolytic organism with limited saccharolytic capabilities. The putative capacity for syntrophic propionate degradation, hitherto unreported in halothermophilic organisms, is also noted. Analysis of global occurrence identified ARYD3 lineage in a very limited number of datasets, where it constituted an extremely small fraction of the overall community.

Anaerobic fermentative growth capacity at elevated temperatures is ubiquitous across the tree of life. Organisms displaying such capability have been recovered from a wide range of ecosystems, including ONG formations, e.g., *Thermoanaerobacter* (order Clostridiales [44]), *Frackibacter* ([14] order Haloanaerobiales), *Thermotoga* (order Thermotogales, [45]). Many of these taxa are capable of growth under slightly elevated salinities (e.g., 5–10% NaCl). However, a relatively limited number of known isolates are capable of anaerobic fermentative halothermophilic growth at higher (e.g., >15% NaCl) salinities. Examples include *Flexistipes sinusarabici*, originally isolated from Atlantis II Deep brines in the Red Sea, a strain of which was co-enriched in this study (Figure 1.1A). Additional examples include *Halothermothrix orenii* and *Halanaerobaculum tunisiense* (order Haloanaerobiales, both isolated from sediments of a Tunisian hypersaline lake [46]), and *Thermohalobacter berrensis* (order Clostridiales) isolated from sediments of a solar saltern in southern France [47]. The paucity of microbial isolates possessing such capacity is a reflection

of the rarity of environments that provide the requisite polyextremophilic conditions for their growth e.g., some oil reservoirs, brine pools on the sea floor (e.g., Atlantis deep in the Red Sea), deeper sediments in hypersaline ponds, and soda lakes (for alkalihalothermoanaerobic polyextremophiles).

Analysis of the global amplicon and metagenomic datasets demonstrates an extremely limited occurrence of ARYD3 in diverse habitats with higher abundance in saline environments. The ecological distribution analysis also demonstrated that ARYD3 is member of the rare biosphere whenever encountered (Table S2). Indeed, ARYD3 was below the detection limit in original samples from oil field in Cushing (Table S1) and was only observed after its serendipitous enrichment on isolated soy proteins. This could be attributed to 1. The dearth of environments supporting its physiological growth preferences as described above, 2. It's out competition by anaerobic respiratory halothermoanaerobes. It is important to note that high energetic costs are associated with compatible solute production in high salinity ecosystems, and the relatively limited yield (ATP/mol substrate) of fermentative organisms could severely limit their competitive ability in highly salinity ecosystems, 3. It's out competition by faster-growing fermentative halothermoanaerobes. ARYD3 genome suggests a requirement for a specific set of substrates (mostly amino acids) for growth, and a limited capacity for sugar/polysaccharides uptake and metabolism. It is notable that many MEER or MEOR projects that involve substrate addition to oil reservoirs often involve addition of sugars/carbohydrates e.g., molasses, rather than amino acids.

Ever since observing that a large fraction of microbial cells exist in extremely low abundance in highly diverse habitats [48]; the identity, phylogenetic novelty, uniqueness, putative role in ecosystems, dynamics, rationale for occurrence, mechanisms of maintenance, and evolutionary history of rare members of the community have been extensively investigated [7, 8]. It has been suggested that a fraction of the rare biosphere exists in a state of dormancy or extremely low metabolic activity since the prevalent environmental conditions in the ecosystem is unconducive to its propagation [8]. These dormant cells could be resuscitated when conditions become favorable in the ecosystem [49, 50]. As such, they represent a seed bank of functional redundancy that aids in ecosystem response to changing environmental conditions. Lynch et al. [8] differentiates between periodic recruitment, i.e. instances where rare members are periodically promoted to high abundance due to recurring environmental changes such as spring thaw and fall leaf litter deposition; and occasional recruitment, i.e. instances where rare members of the community propagate to high numbers due to unexpected change in prevalent conditions within a specific environment, e.g. thermal treatment of arctic soils, and anthropogenic carbon input to subsurface environments. We postulate that the probability of enrichment, and hence detection, of phylogenetically novel members of the rare biosphere is higher in instances of occasional, rather than seasonal, recruitment, and that the more unexpected/ uncommon/unnatural/drastic the disturbance is, the higher the chance of recruitment of phylogenetically novel taxa. Under this scenario, and based on the metabolic capabilities, physiological preferences, and ecological preferences of ARYD3, we argue that ARYD3 is a member of the rare biosphere that could only propagate to higher abundances in response to a complex, multifaceted, and extremely uncommon type of disturbances, i.e. introduction of high levels of amino acids and peptides to an ecosystem with moderately elevated temperature, high salinity, low oxygen, and dearth of anaerobic electron acceptors. The rarity of environments satisfying these conditions, and the possibility that putatively faster growing lineages (e.g., *Halothermothrix, Thermohalobacter*) would outcompete it could explain its extreme rarity in prior amplicon-based studies, and its absence in environmental genomic surveys. The serendipitous enrichment of ARYD3 lineage in this study argues for the use of an array of complex environmental disturbances [17, 29] in efforts

to enrich for phylogenetically-novel lineages that have hitherto escaped detection, an approach that could significantly expand our understanding of the scope of high rank phylogenetic diversity on earth.

An intriguing trait in ARYD3 genome is its apparent ability to degrade propionate. Propionate degradation to acetate is thermodynamically unfavorable unless the generated reducing equivalents (hydrogen or electrons) are disposed, usually via hydrogen utilization by another organism. Given the paucity of high salinity, high temperature, anoxic environments, it is not surprising that very little, if any, is known regarding halophilic syntrophic fatty acid degradation. To date, the majority of studies on syntrophic propionate degradation have understandably been done in highly eutrophic environments where significant amounts of volatile fatty acids are generated during fermentative metabolism. The role of syntrophic metabolism in high salinity habitats is largely unclear, and progression of syntrophic metabolism is often hampered by salinity constrains limiting methanogenesis in subsurface habitats.

We propose the creation of a new bacterial phylum to accommodate the assembly ARYD3 obtained in this study. The designation is justified by its unique phylogenetic position in the bacterial tree of life in 16S rRNA gene and concatenated proteins phylogenetic trees. We propose the name Mcinerneyibacterium aminivorans to honor the contributions of Professor Michael J. McInerney in the fields of petroleum microbiology and syntrophic metabolism. The genome assembly ARYD3 serves as the type material and is deposited in GenBank under Bioproject accession number PRJNA558942, Biosample accession number SAMN12502798, and genome accession number VSIX00000000). Description of *Candidatus* Mcinerneyibacterium gen. nov.

(Mc.i.ner.ney.i.bac.te'ri.um. N.L. neut. n. bacterium a rod; N.L. neut. n. Mcinerneyibacterium a rod named after Professor Michael McInerney to recognize his decades long contribution to the fields of syntrophy and petroleum microbiology).

Genomic analysis predicts an anaerobic, Gram-negative, chemoorganoheterotrophic rod with an outer S layer. Substrates supporting growth based on genomic analysis include alanine, asparagine, aspartate, glutamate, glutamine, glycine, serine, cysteine, methionine, histidine, threonine, glucose, fructose, and mannose, and propionate. Fermentation end products include succinate, formate, acetate, and hydrogen. Genomic analysis identified auxotrophies for histidine, arginine, valine, leucine, isoleucine, methionine, tryptophane, and lysine. The type species is *Candidatus* Mcinerneyibacterium aminivorans.

Description of *Candidatus* Mcinerneyibacterium aminivorans sp. nov.

a.mi.ni.vo'rans. N.L. neut. n. aminum amine; L. v. vorare to devour, to eat; N.L. part. adj. aminivorans amine (amino acid)-consuming) pertaining to the predicted ability of the organism to utilize amino acids as a carbon and energy source based on genomic analysis.

Exhibits the following properties in addition to those given in the genus description. Possesses a small genome with a low G+C content, synthesizes the compatible solutes cyclic 2,3-diphosphoglycerate (cDPG), α-glucosylglycerate, and the uptake of glycine betaine for osmoadaptation. Predicted to have multiple defense mechanisms, including the CRISPR-Cas system, type I restriction endonucleases, and the oxidative stress enzymes superoxide reductase, alkyl hydroperoxide reductase, rubrerythrin, and rubredoxin. Appears to be a member of the rare biosphere, especially in high salinity habitats. The genome assembly ARYD3 serves as

the type material and is deposited in GenBank under Biosample accession number SAMN12502798, and genome accession number VSIX00000000.

Description of *Candidatus* Mcinerneyibacteriaceae fam. nov.
Mcinerneyibacteriaceae (Mc.i.ner.ney.i.bac.te.ri.a.ce'ae. N.L. neut. n. Mcinerneyibacterium a Candidatus generic name; N.L. suff. -aceae ending to denote a family; N.L. fem. pl. n. Mcinerneyibacteriaceae the family of the genus Mcinerneyibacterium). The description is the same as for the genus Mcinerneyibacterium. The type genus is Mcinerneyibacterium.

Description of *Candidatus* Mcinerneyibacteriales ord. nov.
Mcinerneyibacteriales ((Mc.i.ner.ney.i.bac.te.ri.a'les. N.L. neut. n. Mcinerneyibacterium a Candidatus generic name; N.L. suff. -ales ending to denote an order; N.L. fem. pl. n. Mcinerneyibacteriales the order of the genus McInerneyibacterium).). The description is the same as for the genus Mcinerneyibacterium. The type genus is Mcinerneyibacterium.

Description of *Candidatus* McInerneyibacteria classis nov.
Mcinerneyibacteria (Mc.i.ner.ney.i.bac.te'ri.a. N.L. fem. pl. n. Mcinerneyibacteriales type order of the class; N.L. suff. -ia ending to denote a class; N.L. neut. pl. n. Mcinerneyibacteria the class of the order Mcinerneyibacteriales). The description is the same as for the genus Mcinerneyibacterium. The type order is Mcinerneyibacteriales.

Description of *Candidatus* Mcinerneyibacteriota phyl. nov.
Mcinerneyibacteriota (Mc.i.ner.ney.i.bac.te.ri.o′ ta. N.L. neut. pl. n. Mcinerneyibacteria type class of the phylum; N.L. suff. -ota ending to denote a phylum; N.L. neut. pl. n. Mcinerneyibacteriota the phylum of the class Mcinerneyibacteria). The *Candidatus* phylum Mcinerneyibacteriota is defined by the genome assembly ARYD3 recovered using genome resolved metagenomics from formation water (17.6% salinity) and crude oil enrichment amended by isolated soy proteins (0.2%), and incubated for 100 days under anaerobic conditions at 50°C, as well as 16S rRNA gene sequences from uncultured representatives from several habitats, including sulfidogenic deep hypersaline Urania basin, salt ponds in wetland restoration areas in San Francisco, anaerobic digesters treating wastewater, hydrothermal vent microbial mats in Guaymas basin, and coastal marine sediments. The type material is deposited in GenBank under Biosample accession number SAMN12502798, and genome accession number VSIX00000000.

1.6 Acknowledgements

We thank Joshua York for technical assistance, and the Oklahoma State University High Performance Computing Center at Oklahoma State University for providing computational resources.

1.7 Figures and Tables

All Supplementary data for this work can be viewed online at:
https://www.sciencedirect.com/science/article/pii/S0723202020300059#bib0215

CHAPTER II

GENOMIC ANALYSIS OF FAMILY UBA6911 (GROUP 18 ACIDOBACTERIA) EXPANDS THE METABOLIC CAPACITIES OF THE PHYLUM AND HIGHLIGHTS ADAPTATIONS TO TERRESTRIAL HABITATS

2.1 Abstract

Approaches for recovering and analyzing genomes belonging to novel, hitherto-unexplored bacterial lineages have provided invaluable insights into the metabolic capabilities and ecological roles of yet-uncultured taxa. The phylum Acidobacteria is one of the most prevalent and ecologically successful lineages on Earth, yet currently, multiple lineages within this phylum remain unexplored. Here, we utilize genomes recovered from Zodletone Spring, an anaerobic sulfide and sulfur-rich spring in southwestern Oklahoma, as well as from multiple disparate soil and nonsoil habitats, to examine the metabolic capabilities and ecological role of members of family UBA6911 (group 18) Acidobacteria. The analyzed genomes clustered into five distinct genera, with genera Gp18_AA60 and QHZH01 recovered from soils, genus Ga0209509 from anaerobic digestors, and genera Ga0212092 and UBA6911 from freshwater habitats. All genomes analyzed suggested that members of Acidobacteria group 18 are metabolically versatile heterotrophs capable of utilizing a wide range of proteins, amino acids, and sugars as carbon sources, possess respiratory and fermentative capacities, and display few auxotrophies. Soil-dwelling genera were characterized by larger genome sizes, higher numbers of CRISPR loci, an expanded carbohydrate active enzyme (CAZyme) machinery enabling debranching of specific sugars from polymers, possession of a C_1 (methanol and methylamine) degradation machinery, and a sole dependence on aerobic respiration. In contrast, nonsoil genomes encoded a more versatile respiratory capacity for oxygen, nitrite, sulfate, and trimethylamine N-oxide (TMAO) respiration, as well as the potential for utilizing the Wood-Ljungdahl (WL) pathway as an electron sink during heterotrophic growth. Our results not only expand our knowledge of the metabolism of a yet-uncultured bacterial lineage but also provide interesting clues on how terrestrialization and niche adaptation drive metabolic specialization within the Acidobacteria.

2.2 Importance

Members of the Acidobacteria are important players in global biogeochemical cycles, especially in soils. A wide range of acidobacterial lineages remain currently unexplored. We present a detailed genomic characterization of genomes belonging to family UBA6911 (also known as group 18) within the phylum Acidobacteria. The genomes belong to different genera and were obtained from soil (genera Gp18_AA60 and QHZH01), freshwater habitats (genera Ga0212092 and UBA6911), and an anaerobic digestor (genus Ga0209509). While all members of the family shared common metabolic features, e.g., heterotrophic respiratory abilities, broad substrate utilization capacities, and few auxotrophies, distinct differences between soil and nonsoil genera were observed. Soil genera were characterized by expanded genomes, higher numbers of CRISPR loci, a larger carbohydrate active enzyme (CAZyme) repertoire enabling monomer extractions from polymer side chains, and methylotrophic (methanol and methylamine) degradation capacities. In contrast, nonsoil genera encoded more versatile respiratory capacities for utilizing nitrite, sulfate, TMAO, and the WL pathway, in addition to oxygen as electron acceptors. Our results not only broaden our understanding of the metabolic capacities within the Acidobacteria but also provide interesting clues on how terrestrialization shaped Acidobacteria evolution and niche adaptation.

2.3 Introduction

Our appreciation of the scope of phylogenetic and metabolic diversities within the microbial world is rapidly expanding. Approaches enabling direct recovery of genomes from environmental samples without the need for cultivation allow for deciphering the metabolic capacities and putative physiological preferences of yet-uncultured taxa [51-59]. Furthermore, the development of a genome-based taxonomic framework that incorporates environmentally sourced genomes [60] has opened the door for phylocentric (lineage-specific) studies. In such investigations, comparative analysis of genomes belonging to a target lineage is conducted to determine its common defining metabolic traits, the adaptive strategies of its members to various environments, and evolutionary trajectories and patterns of gene gain/loss across this lineage.

Members of the phylum Acidobacteria are one of the most dominant, diverse, and ecologically successful lineages within the bacterial domain [61-66]. Originally proposed to accommodate an eclectic group of acidophiles [67], aromatic compound degraders and homoacetogens [68], and iron reducers [69], it was subsequently identified as a soil-dwelling bacterial lineage in early 16S rRNA gene-based diversity surveys [70-72]. Subsequent 16S rRNA studies have clearly shown its near-universal prevalence in a wide range of soils, where it represents 5 to 50% of the overall community [73, 74].

Various taxonomic outlines have been proposed for the Acidobacteria. Genome-based classification by the Genome Taxonomy Database (GTDB [r95, October 2020]) [24] splits the phylum into 14 classes, 34 orders, 58 families, and 175 genera. This classification broadly, but not always, corresponds to 16S rRNA gene-based taxonomic schemes in SILVA [75], the 26-group (subdivision) classification scheme [76], and the most recently proposed refined class/order classification scheme [77] (see Table S1 in the supplemental material). Regardless, a strong concordance between habitat and phylogeny was observed within most lineages in the Acidobacteria. Some lineages, e.g., groups 1, 3 (both in class Acidobacteriae in the GTDB), and 6 (class Vicinamibacteria in the GTDB), have been predominantly encountered in soils, while others, e.g., groups 4

(class Blastocatellia in the GTDB), 8 (class Holophagae in the GTDB), and 23 (class Thermoanaerobaculia in the GTDB), are more prevalent in nonsoil habitats [62].

Genomic analysis of representatives of cultured [62] and uncultured metagenome-assembled genomes (MAGs) and single-cell genomes (SAGs) [78-80] of the phylum Acidobacteria has provided valuable insights into their metabolic capacities and lifestyle. However, the majority of genomic (and other -omics) approaches have focused mostly on cultured and yet-uncultured genomes of soil Acidobacteria [81-83]. Genomic-based investigations of nonsoil Acidobacteria pure cultures [84-87] or MAGs [88, 89] are more limited, and consequently, multiple lineages within the Acidobacteria remain unexplored.

We posit that genomic analysis of hitherto-unexplored lineages of Acidobacteria would not only expand our knowledge of their metabolic capacities but also enable comparative genomic investigation on how terrestrialization and niche adaptation shaped the evolutionary trajectory and metabolic specialization within the phylum. To this end, we focus on a yet-uncultured lineage in the Acidobacteria: family UBA6911 (subdivision 18 in reference [71], class 1-2 in reference [77], and group 18 in SILVA database release 138.1 [75]). We combine the analysis of genomes recovered from Zodletone Spring, an anaerobic sulfide- and sulfur-rich spring in southwestern Oklahoma, with available genomes from multiple disparate soil and nonsoil habitats. Our goal was to understand the metabolic capacities, physiological preferences, and ecological role of this yet-uncharacterized group and to utilize the observed genus-level niche diversification to identify genomic changes associated with the terrestrialization process.

2.4 Materials and Methods

2.4.1 Sample collection, DNA extraction, and metagenomic sequencing

Samples were collected from Zodletone Spring, a sulfide- and sulfur-rich spring in western Oklahoma's Anadarko Basin (N34.99562° W98.68895°). The ecology, geochemistry, and phylogenetic diversity of the various locations within the spring have been previously described [90-93]. Samples were collected 5 cm deep into the black sediments by completely filling sterile 50-ml polypropylene plastic tubes. Ten different samples were collected. The tubes were capped, sealed, and kept on ice until brought back to the lab (~2h drive), where they were immediately processed. DNA extraction was conducted on 0.5 g sediment from each of the 10 replicate samples using the DNeasy PowerSoil kit (Qiagen, Valencia, CA, USA) according to the manufacturer's protocols. On average, 0.15 to 0.3 μg DNA was obtained per extraction. All extractions were pooled and used for the preparation of sequencing libraries by use of the Nextera XT DNA library prep kit (Illumina, San Diego, CA, USA) per the manufacturer's instructions. Sequencing was conducted on the Illumina HiSeq 2500 platform and 150-bp pair-end technology using the services of a commercial provider (Novogene, Beijing, China). Sequencing produced 281.0 Gbp of raw data. Metagenomic reads were assessed for quality using FastQC, followed by quality filtering and trimming using Trimmomatic v0.38 [94]. High-quality reads were assembled into contigs using MegaHit (v.1.1.3) [20] with a minimum kmer of 27, maximum kmer of 127, kmer step of 10, and minimum contig length of 1,000 bp. Bowtie2 was used to calculate sequencing coverage of each contig by mapping the raw reads back to the contigs. Contigs were binned into draft genomes using both MetaBAT [95] and MaxBin2 [3], followed by selection of the highest-quality bins using DasTool [96]. GTDB-Tk [97] (v1.3.0) was used for the taxonomic classification of the bins using the classification workflow option -classify_wf, and 4 bins belonging to Acidobacteria family UBA6911 were selected for further analysis. In addition, 18 genomes belonging to family UBA6911 were

selected from the recently released 52,515 genomes in the Earth microbiome catalogue collection [98] available through the IMG/M database. These genomes were binned from peatland soils in Minnesota, USA (4 genomes) [99, 100], an anaerobic biogas reactor in Washington, USA (11 genomes) [101], a Cone Pool hot spring microbial mat in California, USA (2 genomes), and White Oak River estuary sediment in North Carolina, USA (1 genome) [102]. Five other genomes belonging to family UBA6911 were available from GenBank and were obtained in previous studies. These included 4 genomes binned from the Angelo Coastal Range Reserve in Northern California [80] and 1 genome binned from Noosa River sediments (Queensland, Australia) [103].

2.4.2 Genome quality assessment and general genomic features

Genome completeness and contamination were assessed using CheckM (v1.0.13) [21] by employing the lineage-specific workflow (lineage_wf flag). All genomes included in this study were of medium or high quality with >70% completion and <10% contamination (Table S2). Designation as medium- or high-quality drafts was based on the criteria set forth by MIMAGs [104]. The 5S, 16S, and 23S rRNA sequences were identified using Barrnap 0.9 (https:// github.com/tseemann/barrnap). tRNA sequences were identified and enumerated with tRNAscan-SE (v2.0.6, May 2020) [105]. Genomes were mined for CRISPR and Cas proteins using the CRISPR/CasFinder [106].

2.4.3 Phylogenomic analysis

Taxonomic classification using the GTDB taxonomic framework was conducted using the classification workflow option -classify_wf within the GTDB-Tk [97]. Phylogenomic analysis was conducted using the concatenated alignment of 120 single-copy marker genes [24] that is generated by GTDB-Tk [97]. Maximum likelihood phylogenetic trees were constructed in RAxML (v8.2.8) [23] with the PROTGAMMABLOSUM62 model and default settings. Representatives of all other Acidobacteria classes were included in the analysis (Figure 2.1), and *Chloroflexus aggregans* (GCF_000021945.1) was used as the outgroup. As expected, all 28 genomes were classified to the family UBA6911 within the UBA6911 class of the Acidobacteria, but only 14 genomes were classified by the GTDB to the genus level into two genera, gp18_AA60 and UBA6911, while the remaining genomes were unclassified at the genus level. To further assign these genomes to putative genera, average amino acid identity (AAI) and shared gene content (SGC) were calculated using the AAI calculator (http://enve-omics.ce.gatech.edu/). Based on these values, we propose assigning these 14 genomes to three putative novel genera. We propose the names QHZH01 (1 genome from grassland soil), Ga0212092 (2 genomes from a Cone Pool microbial mat), and Ga0209509 (10 genomes from an anaerobic gas digestor [Washington, USA] and 1 Zodletone sediment genome) based on the assembly accession number of the most complete genome within each genus (Table 2.1; Table S2).

2.4.4 Functional annotation

Protein coding genes were annotated using Prodigal (v2.50) [107]. BlastKOALA [108] was used to assign KEGG orthologies (KO) to protein coding genes, followed by metabolic pathway visualization in KEGG mapper [109]. In addition, all genomes were queried with custom-built HMM profiles for sulfur metabolism, electron transport chain components (for alternate complex III), C1 metabolism, and hydrogenases. To construct hmm profiles, UniProt reference sequences for genes with an assigned KO number were

downloaded and aligned with Mafft [110], and the alignment was used to construct hmm profiles using the hmmbuild function of HMMer (v3.1b2) [111]. For genes without a designated KO number, a representative protein was queried against the KEGG gene database using BLASTP, and hits with E values of $<1e^{-80}$ were downloaded, aligned, and used to construct an hmm profile as described above. Hydrogenase hmm profiles were built using alignments downloaded from the Hydrogenase Database (HydDB) [112]. The hmmscan function of HMMer [111] was used with the constructed profiles and a thresholding option of -T 100 to scan the protein coding genes for possible hits. Further confirmation was achieved through phylogenetic assessment and tree building procedures. For that, putatively identified Acidobacteria sequences were aligned with Mafft [110] against the reference sequences used to build the HMM database, and the alignment was then used to construct a maximum likelihood phylogenetic tree using FastTree (v2.1.10) [113]. Sequences that clustered with reference sequences were deemed to be true hits and were assigned a corresponding KO number or function. Carbohydrate active enzymes (CAZymes) (including glycoside hydrolases [GHs], polysaccharide lyases [PLs], and carbohydrate esterases [CEs]) were identified by searching all open reading frames (ORFs) from all genomes against the dbCAN hidden Markov models V9 [114, 115] (downloaded from the dbCAN web server in September 2020) using hmmscan. AntiSMASH 3.0 [116] was used with default parameters to predict biosynthetic gene clusters in the genomes. Canonical correspondence analysis (CCA) was used to identify the correlation between the genus, environmental source, and types of BGCs predicted in the genomes, using the function cca in the R package Vegan (https://cran.r-project.org/web/packages/vegan/index.html). To evaluate the novelty of the NRPS and PKS clusters identified in family UBA6911 genomes, we queried the synthetic genes from these clusters against the NCBI nucleotide database (downloaded in April 2020). A threshold of 75% identity over 50% of the query length was used to determine the novelty of these genes.

2.4.5 Phylogenetic analysis of XoxF methanol dehydrogenase and dissimilatory sulfite reductase DsrAB

Family UBA6911 predicted XoxF methanol dehydrogenase and dissimilatory sulfite reductase subunits A and B were compared to reference sequences for phylogenetic placement. Family UBA6911 predicted XoxF protein sequences were aligned to corresponding reference sequences from other methylotrophic taxa, while the dissimilatory sulfite reductase subunits A and B were aligned to corresponding subunits from sulfate-reducing taxa using Mafft [110]. DsrA and DsrB alignments were concatenated in Mega X [117]. The XoxF alignment and the DsrAB concatenated alignment were used to construct maximum-likelihood phylogenetic trees using FastTree (v2.1.10) [113].

2.4.6 Ecological distribution

To further examine the ecological distribution of family UBA6911 genera, we analyzed 177 near-full-length 16S rRNA gene sequences (>1,200bp) associated with this lineage in the SILVA database (r138.1) [75] (SILVA classification, Bacteria;Acidobacteriota;Subgroup 18). A nearly complete 16S rRNA gene from each genus (with the exception of genus QHZH01 represented by a single genomic assembly that unfortunately lacked a 16S rRNA gene) was selected as a representative and was included in the analysis. Sequences were aligned using the SINA aligner [22], and the alignment was used to construct maximum-likelihood phylogenetic trees with FastTree [113]. The environmental source of hits clustering with the appropriate

reference sequences was then classified with a scheme based on the GOLD ecosystem classification scheme [118]. All phylogenetic trees were visualized and annotated in iTol [119].

2.4.7 Data availability

Metagenomic raw reads for Zodletone Spring sediment are available under SRA accession no. SRX9813571. The Zodletone Spring whole-genome shotgun project was submitted to GenBank under Bioproject ID PRJNA690107 and Biosample ID SAMN17269717. The individual assembled Acidobacteria MAGs analyzed in this study have been deposited in DDBJ/ENA/GenBank under accession no. JAFGAO000000000, JAFGSS000000000, JAFGIY000000000, and JAFGTD000000000.

2.5 Results

2.5.1 Ecological distribution patterns of family UBA6911

Family UBA6911 genomes clustered into five genera based on amino acid identity (AAI) and shared gene content values (Figure 2.1; Table 2.1). Genomes of the genera Gp18_AA60 (n = 3) and QHZH01 (n = 1) were exclusively binned from a grassland meadow within the Angelo Coastal Range Reserve in Northern California [80]. Genus Ga0209509 genomes were mostly (10/11 genomes) binned from an anaerobic gas digestor (Washington, USA) [101]. Genus Ga0212092 (n = 2) genomes were binned from a Cone Pool hot spring microbial mat in California, USA. Finally, genus UBA6911 (10 genomes) displayed a broader distribution pattern, as its genomes were recovered from multiple, mostly freshwater habitats, e.g., river and estuary sediments and Zodletone Spring sediment (Table 2.1; Figure 2.1).

To further examine the ecological distribution patterns of family UBA6911, we analyzed 177 near-full-length 16S rRNA gene amplicons generated in multiple culture-independent amplicon-based diversity surveys and available through the SILVA database (r138.1) (Figure 2.2). Genus distribution patterns gleaned from the origin of MAGs were generally confirmed: 16S rRNA sequences affiliated with genus Gp18_AA60 were mostly recovered from soil (14/15, 93.3%), those affiliated with genus Ga0209509 were mostly encountered in anaerobic digestors (11/14, 85.7%), and the majority of 16S rRNA sequences affiliated with the genera Ga0212092 (90.9%) and UBA6911 (58.8%) were encountered in a wide range of freshwater environments (e.g., freshwater lake sediments, estuary sediments, and thermal springs). Collectively, the combined MAGs and 16S rRNA amplicon data suggest a preference for soil for genera Gp18_AA60 and QHZH01, a preference for anaerobic digestors for Ga0209509, and a wide occurrence of genera Ga0212092 and UBA6911 in freshwater habitats.

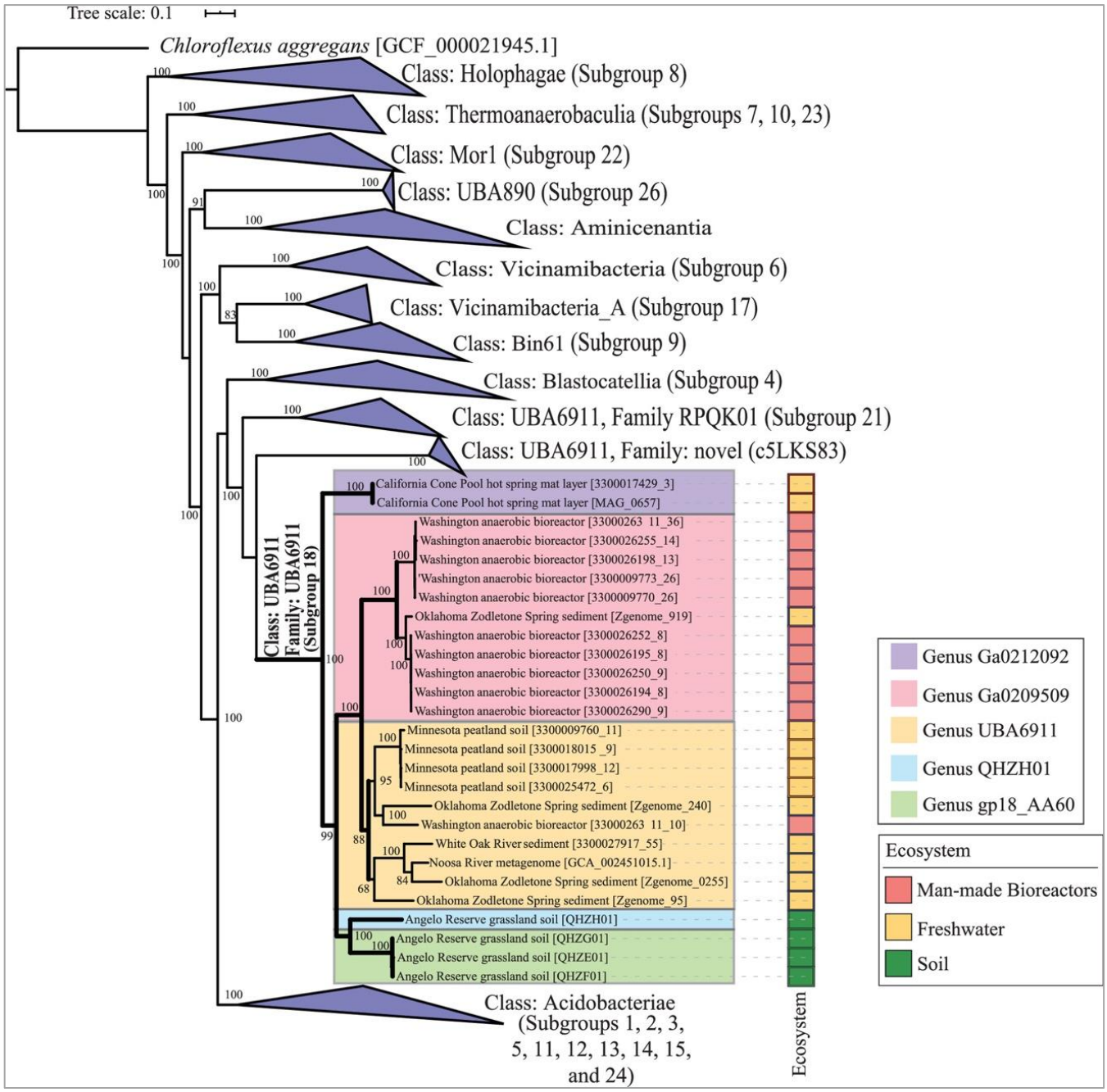


Figure 2. 1 Maximum likelihood phylogenetic tree based on concatenated alignment of 120 single-copy genes from all Acidobacteria classes in GTDB r95. Family UBA6911 (Acidobacteria subgroup 18), is shown unwedged with thick branches. Branches are named by the bin name in brackets (as shown in Table 2.1) and with the ecosystem from which they were binned. The five genera described here are color coded as shown in the legend. Bootstrap values (from 100 bootstraps) are displayed for branches with ≥70% support. The tree was rooted using *Chloroflexus aggregans* (GCF_000021945.1) as the outgroup. The tracks to the right of the tree represent the ecosystem from which the genomes were binned (color coded as shown in the legend). All other Acidobacteria classes are shown as wedges with the corresponding subgroup number(s) in parentheses.

Genus and bin names	Binning source	Similarity statistics		Binned genome size (Mbp)	Estimated genome size (Mbp)	% Protein coding bases	GC content (%)	No. of CRISPRs	Avg gene length (bp)	Total no. of genes	Total no. of protein coding genes	Accession no.	References
		AAI (avg ± SD)	SGC (avg ± SD)										
Genus Ga0212092													
3300022548-MAG_0657	Cone Pool hot spring mat layer	54.09 ± 0.98	40.34 ± 5.3	3.74	4.33	90.08	65.68	5	1,021	3,349	3,303	Ga0212092[a]	—[c]
3300017429_3				3.82	4.56	90.05	65.60	4	1,018	3,424	3,376	Ga0185343[a]	—
Genus QHZH01													
QHZH01	Angelo Reserve soil	54.63 ± 5.62	35.97 ± 5.12	6.00	8.01	90.41	66.73	2	895	6,088	6,058	QHZH01[b]	[80]
Genus Gp18_AA60													
QHZE01	Angelo Reserve soil	55.32 ± 5.08	37.1 ± 4.46	7.54	8.12	84.85	58.70	18	985	6,542	6,497	QHZE01[b]	[80]
QHZF01				7.06	8.42	84.49	58.85	25	916	6,546	6,508	QHZF01[b]	[80]
QHZG01				7.33	7.65	84.45	58.87	31	980	6,360	6,314	QHZG01[b]	[80]
Genus UBA6911													
3300026311_10	Anaerobic biogas reactor	53.59 ± 4.66	38.83 ± 7.76	4.38	4.71	92.16	56.50	3	1,002	4,076	4,029	Ga0209723[a]	[100]
3300009760_11	Peatland			3.10	5.17	84.98	52.10	0	827	3,218	3,184	Ga0116131[a]	[98, 99]
3300017998_12				4.18	5.15	87.18	52.70	4	866	4,262	4,211	Ga0187870[a]	[98, 99]
3300025472_6				4.15	5.61	86.20	52.60	4	838	4,309	4,267	Ga0208692[a]	[98, 99]
3300018015_9				4.69	5.91	87.26	52.80	2	831	4,956	4,923	Ga0187866[a]	[98, 99]
3300027917_55	White Oak River sediment			3.87	4.43	84.76	48.90	9	898	3,695	3,656	Ga0209536[a]	[101]
GCA_002451015	Noosa River			4.94	5.72	85.35	51.60	1	911	4,657	4,632	DKBB01[b]	[102]
Zgenome_0255	Zodletone Spring sediment			3.75	5.12	87.68	57.60	19	987	3,252	3,220	JAFGAO0[b]	This study
Zgenome_240				6.51	6.81	87.10	52.30	5	1,011	5,666	5,615	JAFGIY01[b]	This study
Zgenome_95				5.45	5.85	86.79	51.80	7	987	4,835	4,791	JAFGTD01[b]	This study
Genus Ga0209509													
3300009770_26	Anaerobic biogas reactor	54.58 ± 3.99	40.93 ± 8.14	2.50	2.70	90.94	65.20	1	1,008	2,291	2,253	Ga0123332[a]	[100]
3300009773_26				2.56	2.75	90.73	65.20	2	1,006	2,346	2,307	Ga0123333[a]	[100]
3300026194_8				3.76	3.88	91.48	66.60	2	1,035	3,370	3,326	Ga0209509[a]	[100]
3300026195_8				3.83	3.95	91.56	66.50	5	1,041	3,418	3,371	Ga0209312[a]	[100]
3300026198_13				2.52	2.82	91.52	65.70	1	997	2,344	2,310	Ga0209313[a]	[100]

3300026250_9				3.75	3.87	91.48	66.60	3	1,040	3,348	3,300	Ga0209612[a]	[100]
3300026252_8				3.75	3.86	91.47	66.60	2	1,041	3,337	3,291	Ga0209722[a]	[100]
3300026255_14				2.70	2.91	91.17	65.60	2	999	2,502	2,465	Ga0209613[a]	[100]
3300026290_9				3.86	3.98	91.35	66.20	3	1,045	3,425	3,376	Ga0209510[a]	[100]
3300026311_36				2.69	2.78	90.30	64.90	3	1,004	2,466	2,423	Ga0209723[a]	[100]
Zgenome_919	Zodletone Spring sediment			4.75	4.98	91.97	66.50	0	996	3,038	3,001	JAFGSS01[b]	This study

a Accession number corresponds to Gold analysis project number (for MAGs from the recently released 52,515 genomes in the Earth microbiome catalogue collection [98] that were deposited in the IMG/M database).

b NCBI (WGS Project) assembly accession number.

c —, unpublished, used with the principal investigator's permission.

Table 2. 1 Binning sources, similarity statistics, and general genomics features for the genomes analyzed in this study.

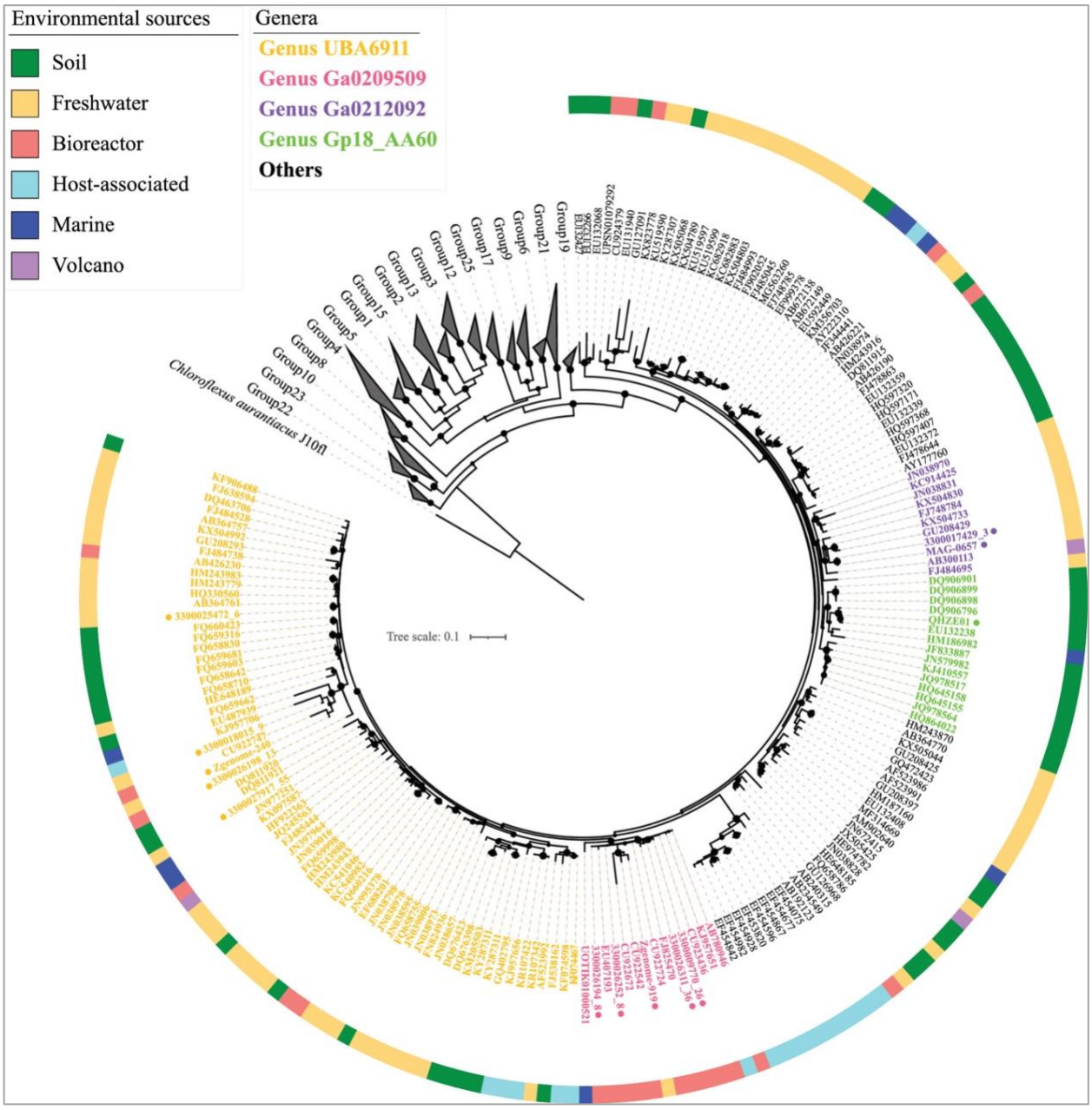


Figure 2. 2 Phylogenetic tree based on 16S rRNA gene for Acidobacteria classes. Groups other than family UBA6911 (subgroup 18) are shown as gray wedges. A total of 177 near-full-length (>1,200-bp) 16S rRNA sequences belonging to Acidobacteria group 18 in the SILVA database are shown with their GenBank accession numbers. Representative 16S rRNA sequences from the analyzed genomes are shown with their corresponding bin name (as in Table 2.1) and with a colored dot next to the name for ease of recognition. Branch labels are color coded by genus, as shown in the legend (using the same color scheme as that in Figure 2.1). The track around the tree shows the environmental classification of the ecosystem from which the sequence was obtained, and the corresponding color codes are shown in the legend. Sequences were aligned using the SINA aligner [22], and the alignment was used to construct a maximum likelihood phylogenetic tree with FastTree [113]. Bootstrap values (from 100 bootstraps) are displayed as bubbles for branches with ≥70% support. The tree was rooted using *Chloroflexus aggregans* (GenBank accession number M34116.1) as the outgroup.

2.5.2 General genomic and structural features of family UBA6911 genomes

Soil-affiliated genera (Gp18_AA60 and QHZH01) displayed larger estimated genome sizes (average ± SD of 8.05 ± 0.32 Mbp). In comparison, in the freshwater genus UBA6911, these values were 5.45 ± 0.68 Mbp, with only a single genome (Zgenome_240) exceeding 6 Mbp. Even smaller genomes were observed for genus Ga0212092 (4.45 ± 0.17 Mbp) and genus Ga0209509 (3.5 ± 0.75 Mbp) genomes (Table 2.1). GC content was generally high in all genomes (Table 2.1), with three genera (QHZH01, Ga0209509, and Ga0212092) exhibiting >65% GC content. Structurally, all genera in family UBA6911 are predicted to be Gram-negative rods with similar predicted membrane phospholipid compositions (Table 2.2; Table S3). Genes mediating type IV pilus formation, generally involved in a wide range of functions, e.g., adhesion and aggregation, twitching motility, DNA uptake, and protein secretion [120], were observed in most genera. Genes encoding secretion systems I and II were identified in all genomes. Interestingly, genes encoding the type VI secretion system, typically associated with pathogenic Gram-negative bacteria (mostly Proteobacteria) and known to mediate protein transport to adjacent cells as a mean of bacterial antagonism [121], was identified in all but a single (anaerobic digestor Ga0209509) genus. A search of the AnnoTree database [122] identified type VI secretion system genes in only 14 Acidobacteria genomes, all belonging to class Holophagae, suggesting the rare distribution of such trait in the phylum. Finally, genomes of the soil genus Gp18_AA60 possess an exceptionally large number of CRISPR genes (24.67 ± 6.5) (Table 2.1). In comparison, genomes from all other genera possessed 0 to 9 CRISPR genes per genome, with the notable exception of the Zgenome_0255 genome (genus UBA6911), which possessed 19 CRISPR genes.

Feature	**Results for indicated genus[a]**				
	Gp18_AA60	**QHZH01**	**UBA6911**	**Ga0209509**	**Ga0212092**
Predicted structural features					
Flagellar motility	n	n	Y	n	Y
Type IV pilus assembly	Y	n	Y	Y	Y
Chemotaxis	n	n	Y	n	n
Type VI secretion system	Y	Y	Y	n	Y
CRISPR count	24.67 ± 6.5	2	5.4 ± 5.4	2.18 ± 1.33	4.5 ± 0.71
Biosynthesis: biosynthetic gene clusters					
Terpenes	Y	Y	Y	Y	Y
Phenazine	Y	Y	n	n	n
NRPS/PKS	Y	Y	Y	Y	Y
Homoserine lactones	n	n	n	Y	n
Bacteriocin	n	Y	Y	n	n
Heterotrophic substrates predicted to support growth					
Sugars[b]					
CAZymes (GHs + PLs + CEs)	98.3 ± 11.7	166	48.2 ± 22.3	45.6 ± 13.4	42.5 ± 0.71
Hexoses	Glu, Man, Gal, Fru	Glu, Man	Glu, Man, Gal, Fru	Glu, Man, Gal, Fru	Glu, Man, Gal
Hexosamines	Y	Y	Y	Y	Y
Hexuronic acids	n	n	Y	Y	n
Sialic acid	n	Y	n	n	n

Sugar alcohols	Ino, Sorb, Xylitol	Ino, Sorb, Xylitol	Sorb, Xylitol	Sorb, Xylitol	Xylitol
Pentoses	Lyx, Xyl	Lyx, Xyl	n	Xyl	n
Amino acids					
Acidic and amides	D, N, E, Q	D, N, E, Q	D, N, E, Q	D, E, Q	D, N, E, Q
Aliphatic	A, G	A, G	A, G, V, L, I	A, G	A, G
Aromatic	n	n	Y	n	n
Basic	H	n	R, K	R	R, H
Hydroxy and S containing	S, T	S, T	M	M	S, T, M
Cyclic	P	P	P	P	P
C_1 compounds					
Methanol	n	Y	n	n	n
Methylamines	n	Y	n	n	n
Formaldehyde	n	Y	Y	n	n
Predicted respiratory capacities					
Aerobic (low affinity)	Y	Y	Y	Y	Y
Aerobic (high affinity)	Y	n	Y	Y	Y
Dissimilatory nitrate reduction to ammonium	n	n	Y		Y
Dissimilatory nitrite reduction to ammonium	n	n		Y	n
Dissimilatory sulfate reduction	n	n	Y	n	Y
TMAO respiration	n	n	n	n	Y
Predicted fermentation products					
Acetate	Y	Y	Y	Y	n
Ethanol	n	Y	Y	Y	Y
Formate	n	n	Y	Y	n
Acetone	n	n	Y	n	n
Acetoin and butanediol	Y	n	Y	Y	Y

a Y, full pathway identified; n, full pathway missing.

b Sugars are abbreviated as follows; Glu, glucose; Man, mannose; Gal, galactose; Fru, fructose; Ino, myo-inositol; Sorb, sorbitol; Lyx, lyxose; Xyl, xylose.

Table 2. 2 Salient defining features of family UBA6911 genera

2.5.3 Anabolic capabilities in family UBA6911 genomes

All UBA6911 genomes encoded a fairly extensive anabolic repertoire, with capacity for biosynthesis of the majority of amino acids from precursors (from 15 in genus Ga0212092 to 20 in genus UBA6911) (Table S3). Gluconeogenic capacity was encoded in all genera (Table S3). Cofactor biosynthesis capability was also widespread in all but the freshwater Cone Pool Ga0212092 genomes (Table S3). A complete assimilatory sulfate reduction pathway for sulfate uptake and assimilation was observed in all genomes. In addition, the presence of genes encoding taurine dioxygenase (in soil genus QHZH01 genome) and alkanesulfonate monooxygenase (in soil genus gp18_AA60 genomes) argue for the capacity for organosulfur compound assimilation. For nitrogen assimilation, in addition to NH_4 and amino acids, multiple pathways for N assimilation from organic substrates were identified. These include the presence of urease genes (*ureABC*) for urea assimilation in soil genus gp18_AA60 genomes, along with genes encoding urease Ni-delivering chaperones (*ureDEFG*). As well, genes for N extraction from arylformamides (arylformamidase, EC 3.5.1.9), N,N-dimethylformamide (N,N-dimethylformamidase, EC 3.5.1.56), and N-formylglutamate (formylglutamate deformylase, EC 3.5.1.68) were identified to various degrees (Table S3) in all genera, with the exception of freshwater Cone Pool genus Ga0212092 genomes. No evidence for assimilatory nitrate reduction or nitrogen fixation was identified in any of the genomes.

In addition to the above-described biosynthetic capacities, all examined genomes, regardless of environmental source or genus-level affiliation, included biosynthetic gene clusters (BGCs) ranging in number from 3 to 12. Such clusters encode polyketide synthases (PKS) and non-ribosomal peptide synthases (NPRS), homoserine lactones, terpenes, phenazines, and bacteriocin. While the numbers of BGCs per genome did not differ much between genera (Table S4), differences in the gene clusters, and hence putative products, were identified (Table 2.2; Figure S1). For example, phenazine biosynthetic clusters were identified only in the soil genera genomes, while BCGs for homoserine lactone biosynthesis were exclusive to the anaerobic digestor Ga0209509 genomes. While not exclusive, terpene biosynthetic clusters appear to be more enriched in anaerobic digestor Ga0209509 genomes, while NRPS and PKS clusters were enriched in the soil genus genomes and UBA6911 genomes. Finally, a bacteriocin biosynthesis cluster was identified only in genus QHZH01 genome and a single genome affiliated with genus UBA6911 (Figure S1). Finally, we queried NRPS and PKS clusters identified in all 28 genomes against the NCBI nucleotide database. Surprisingly, only 14 genes (3.8%) had significant hits to previously deposited sequences (Table S4).

2.5.4 Substrate utilization patterns in family UBA6911 genomes

Genomic analysis of all family UBA6911 genomes suggests a heterotrophic lifestyle with robust aminolytic and saccharolytic machineries. Genomes from all genera encoded a broad capacity to metabolize proteins and amino acids. An arsenal of endopeptidase, oligopeptidase, and dipeptidase genes were identified in all genomes (Table S5). Oligopeptide transporters (in UBA6911 genomes), peptide/Ni transporters (all genomes), and dedicated amino acid transporters (e.g., branched-chain amino acid transporters in QHZH01, UBA6911, and Ga0212092 genomes, and *trp/tyr* transporters in gp18_AA60 genomes) suggest the capacity for amino acid and oligo/dipeptide uptake. Furthermore, all genomes to various degrees showed degradation pathways for a wide range of amino acids, ranging from 9 (in genus Ga0212092) to 15 (in genus UBA6911) (Table 2.2; Table S3).

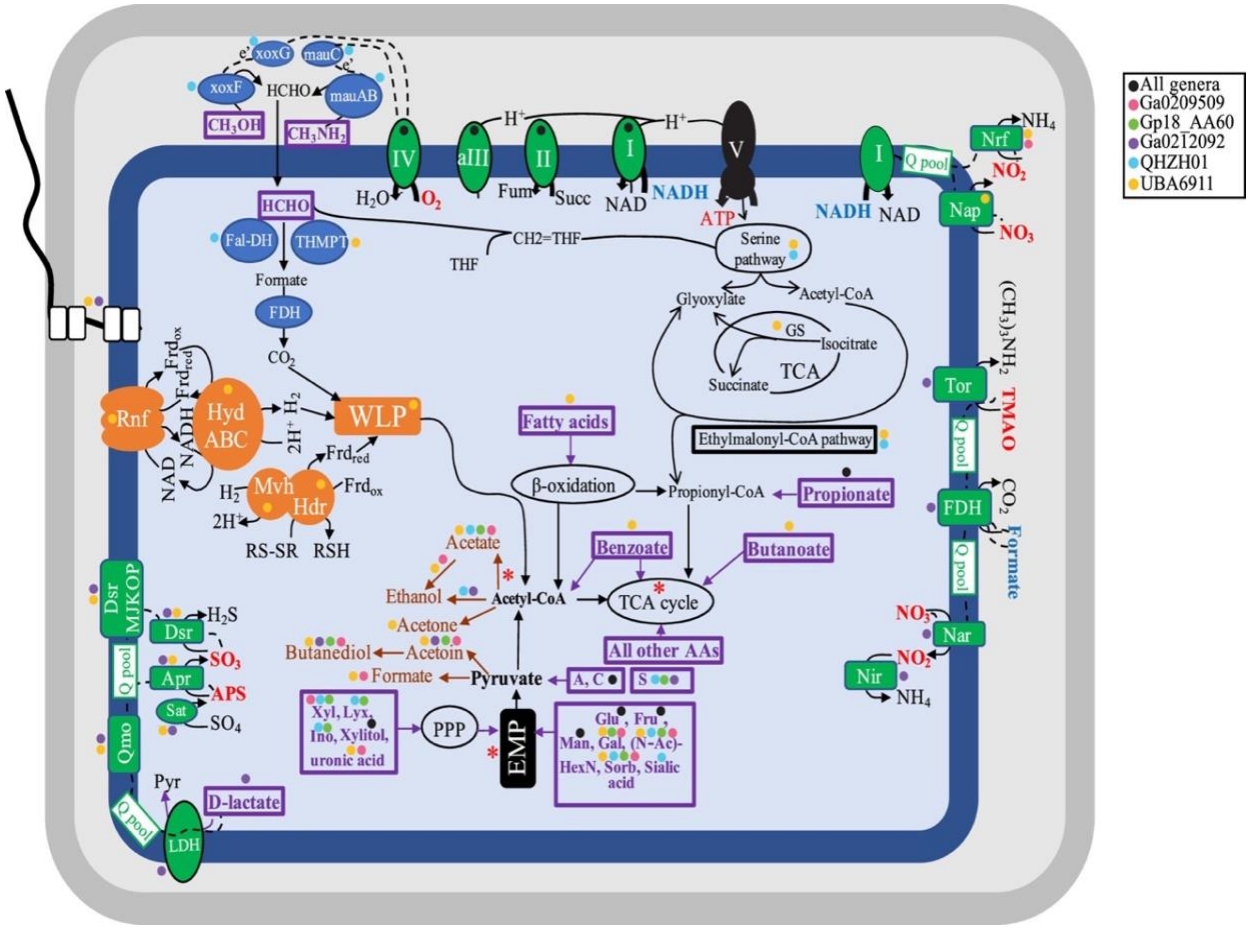


Figure 2. 3 Cartoon depicting different metabolic capabilities encoded in family UBA6911 genomes with capabilities predicted for different genera shown as colored circles (all orders, black; genus Ga0209509, pink; genus Gp18_AA60, green; genus Ga0212092, purple; genus QHZH01, cyan; genus UBA6911, yellow). Enzymes for C1 metabolism are shown in blue. Electron transport components are shown in green, and electron transfer is shown as dotted black lines from electron donors (shown in boldface blue text) to terminal electron acceptors (shown in boldface red text). The sites of proton extrusion to the periplasm are shown as black arrows, as is the F-type ATP synthase (V). Proton motive force generation, as well as electron carrier recycling pathways associated with the operation of the WLP, is shown in orange. All substrates predicted to support growth are shown in boldface purple text within thick purple boxes. Fermentation end products are shown in brown. Sites of substrate-level phosphorylation are shown as red asterisks. A flagellum is depicted, the biosynthetic genes of which were identified in genomes belonging to the genera UBA6911 and Ga0212092. The cell is depicted as rod shaped based on the identification of the rod shape-determining proteins RodA, MreB, and MreC in all genomes.

Abbreviations: Apr, the enzyme complex adenylylsulfate reductase (EC 1.8.99.2); APS, adenylyl sulfate; Dsr, dissimilatory sulfite reductase (EC 1.8.99.5); EMP, Embden-Meyerhoff-Parnas pathway; Fal-DH, formaldehyde dehydrogenase; FDH, formate dehydrogenase; Frdox/red, ferredoxin (oxidized/reduced); Fru, fructose; fum, fumarate; Gal, galactose; Glu, glucose; GS, glyoxylate shunt; Hdr, heterodisulfide reductase complex; HydABC, cytoplasmic [Fe Fe] hydrogenase; I, II, aIII, and IV, aerobic respiratory chain comprising complex I, complex II, alternate complex III, and complex IV; Ino, myo-inositol; LDH, l-lactate dehydrogenase; Lyx, lyxose; Man, mannose; mauABC; methylamine dehydrogenase; Mvh, Cytoplasmic [Ni Fe] hydrogenase; Nap, nitrate reductase (cytochrome); Nar, nitrate reductase; Nir, nitrite reductase (NADH); Nrf, nitrite reductase (cytochrome c-552); N-AcHexN, N-acetylhexosamines; PPP, pentose phosphate pathway; Pyr, pyruvate; Q pool, quinone pool; Qmo, quinone-interacting membrane-bound oxidoreductase complex; RNF, membrane-bound RNF complex; RSH/RS-SR, reduced/oxidized disulfide; Sorb, sorbitol; succ, succinate; TCA, tricarboxylic acid cycle; THMPT, tetrahydromethanopterin-linked formaldehyde dehydrogenase; TMAO, trimethylamine N-oxide; Tor, trimethylamine-N-oxide reductase (cytochrome c); V, ATP synthase complex; WLP, Wood-Ljungdahl pathway; *xoxFG*, methanol dehydrogenase; Xyl, xylose.

Similarly, with the notable exception of members of freshwater Cone Pool genus Ga0212092, a robust machinery for sugar metabolism was identified (Table 2.2; Table S3). Glucose, mannose, galactose, fructose, hexosamines e.g., N-acetyl hexosamines, uronic acids, sorbitol, and xylitol degradation capacities were identified in all other four genera (Table 2.2; Table S3). Further, soil genus genomes encoded the full degradation machinery for myo-inositol (IolBCDEG), an abundant soil component [123]. Interestingly, the QHZH01 genome also encoded the full machinery for sialic acid (an integral component of soil fungal cell walls [124]) degradation to fructose-6-P.

Analysis of the carbohydrate active enzyme (CAZyme) repertoire was conducted to assess family UBA6911 polysaccharide degradation capacities. A notable expansion of the CAZyme (carbohydrate esterase [CE] plus polysaccharide lyase [PL] plus glycoside hydrolase [GH]) repertoire in soil genomes was observed (166 in QHZH01 genome, 98.3 ± 11.7 in gp18_AA60 genomes), in comparison to 46.48 ± 17.01 in all other genera (Table S6). Higher numbers in soil genus genomes were brought about by the expansion of families GH33 (sialidase), GH165 (b-galactosidase), and GH56 (hyaluronidase) in genus QHZH01, of family GH13 (amylase) in gp18_AA, and of family GH109 (N-acetylhexosaminidase) in both genera (Table S6). As described above, genes encoding degradation machineries of the released monomer products for these enzymes (sialic acid, galactose, uronic acid, glucose, and N-acetylhexosamine) are encoded in the soil genomes (Table 2.2; Table S3).

Overall, all family UBA6911 genomes analyzed in this study displayed a notable absence or extreme paucity of CAZy families encoding/initiating breakdown of large polymers, e.g., endo- and exocellulases (only 6 copies of GH5 genes and 5 copies of GH9 genes in all 28 genomes examined), hemicellulases (only 19 copies of GH10 genes in all 28 genomes examined), and pectin degradation (only 22 copies of GH28 genes in all genomes examined), suggesting the limited ability of all members of the family for degrading these high-molecular-weight polysaccharides (Table S6). Collectively, the CAZyome and sugar degradation patterns of family UBA6911 argue for a lineage specializing in sugar, but not polymer, degradation, with added specialized capacities for debranching specific sugars from polymers in the soil-dwelling genera.

Further, methylotrophic C1 degradation capacity for methanol and methylamine utilization was encoded in the soil QHZH01 genome. The genome encoded a methanol dehydrogenase (EC 1.1.2.10) for methanol oxidation to formaldehyde (Figure 2.3 and Table 2.2; Table S3). QHZH01 methanol dehydrogenase belonged to the lanthanide-dependent pyrroloquinoline quinone (PQQ) methanol dehydrogenase XoxF type. More specifically, QHZH01 XoxF methanol dehydrogenase belonged to clade xoxF-2, previously described for "*Candidatus* Methylomirabilis" (NC10) (Figure 2.4). The accessory xoxG (c-type cytochrome) and xoxJ (periplasmic binding) genes were fused and encoded in QHZH01 genome downstream of the xoxF gene. A gene encoding quinohemoprotein amine dehydrogenase (EC 1.4.9.1) for methylamine oxidation to formaldehyde (Figure 2.3 and 2.4 and Table 2.2; Table S3) was also identified in the genome. Furthermore, the genome also encoded subsequent formaldehyde oxidation to formate via the glutathione-independent formaldehyde dehydrogenase (EC 1.2.1.46), as well as formate oxidation to CO2 via formate dehydrogenase (EC 1.17.1.9). Finally, for assimilating formaldehyde into biomass, genes encoding the majority of enzymes of the serine cycle were identified in the genome, as well as the majority of genes encoding the ethyl malonyl coenzyme A (ethylmalonyl-CoA) pathway for glyoxylate regeneration (Figure 2.3 and Table 2.2; Table S3). In addition, several genomes in genus UBA6911 encoded formaldehyde oxidation (via the tetrahydromethopterin-linked pathway, glutathione-independent formaldehyde dehydrogenase [EC 1.2.1.46], and glutathione-dependent formaldehyde dehydrogenase [EC 1.1.1.284, EC 3.1.2.12]) to formate, formate oxidation to CO2 via formate dehydrogenase (EC:1.17.1.9), formaldehyde assimilation into biomass (genes

encoding the majority of enzymes of the serine cycle were identified in some genomes), and glyoxylate regeneration (the majority of genes encoding the ethylmalonyl-CoA pathway were identified in some genomes, with only one genome containing glyoxylate shunt genes) (Figure 2.3 and Table 2.2; Table S3). Surprisingly, upstream genes for formaldehyde production from C1 compounds (e.g., methane, methanol, methylamine, C1 sulfur compounds) were missing from all genomes. Phylogenetic analysis of key C1 metabolism genes identified their close affiliation with group 1, 5, and 6 Acidobacteria genomes, previously demonstrated to mediate C1 metabolism and formaldehyde assimilation [80]. The closest relatives outside the phylum varied but were members of Chloroflexi, Desulfobacterota, and Actinobacteriota (Table S7).
Finally, in addition to proteins, carbohydrates, and C1 compounds, other potential carbon sources for family UBA6911 were identified. These include long-chain fatty acids (a complete beta-oxidation pathway) in genus UBA6911 genomes, short-chain aliphatic fatty acids (propionate and butyrate degradation to acetyl-CoA) in both UBA6911 and Ga0212092 genomes, and benzoate in genus UBA6911 genomes (Table S3).

2.5.5 Respiratory capacities

All family UBA6911 genomes encoded respiratory capacities. Genomic analysis suggested that the soil genera Gp18_AA60 and QHZH01 could potentially utilize O_2 as their sole electron acceptor, based on their possession of a respiratory chain comprising complex I, complex II, alternate complex III, and complex IV (cytochrome oxidase aa3) (Table 2.2; Table S3). On the other hand, all anaerobic digestor genus Ga0209509 genomes could mediate dissimilatory nitrite reduction to ammonium, based on the presence of nitrite reductase (cytochrome c-552) (EC 1.7.2.2), plus respiratory complexes I and II. Only 6 out of the 11 genomes in this genus possessed a complete aerobic respiratory chain. Freshwater genus Ga0212092 genomes from Cone Pool microbial sediments were notable in encoding a complete aerobic respiratory chain as well as the complete machinery for dissimilatory sulfate reduction. An additional plausible anaerobic electron acceptor for genus Ga0212092 is trimethylamine N-oxide (TMAO). Here, electron transfer is thought to occur from formate via the membrane-bound formate dehydrogenase (EC 1.17.1.9) through the quinone pool onto TMAO reductase (EC 1.7.2.3), eventually reducing trimethylamine N-oxide to trimethylamine (Figure 2.3 and Table 2.2; Table S3), as previously shown in *Escherichia coli* [125] and *Rhodopseudomonas capsulate* [126]. Finally, members of the freshwater genus UBA6911 demonstrated the most versatile respiratory capacities, with evidence for aerobic respiration (in all genomes except a single Oak River estuary sediment MAG), dissimilatory nitrite reduction to ammonium in genomes from Washington anaerobic gas digestor and Zodletone Spring sediment, and dissimilatory sulfate reduction to sulfide in Noosa River sediment and Zodletone Spring sediment genomes (Figure 2.3 and Table 2.2; Table S3).

Sulfate reduction machinery identified in the genomes of freshwater genera UBA6911 and Ga0212092 included sulfate adenylyltransferase (Sat; EC 2.7.7.4) for sulfate activation to adenylyl sulfate (APS), the enzyme complex adenylylsulfate reductase (AprAB; EC 1.8.99.2) for APS reduction to sulfite, the quinone-interacting membrane-bound oxidoreductase complex (QmoABC) for electron transfer, the enzyme dissimilatory sulfite reductase (DsrAB; EC 1.8.99.5) and its cosubstrate DsrC for dissimilatory sulfite reduction to sulfide, and the sulfite reduction-associated membrane complex DsrMKJOP for linking cytoplasmic sulfite reduction to energy conservation (Figure 2.3).

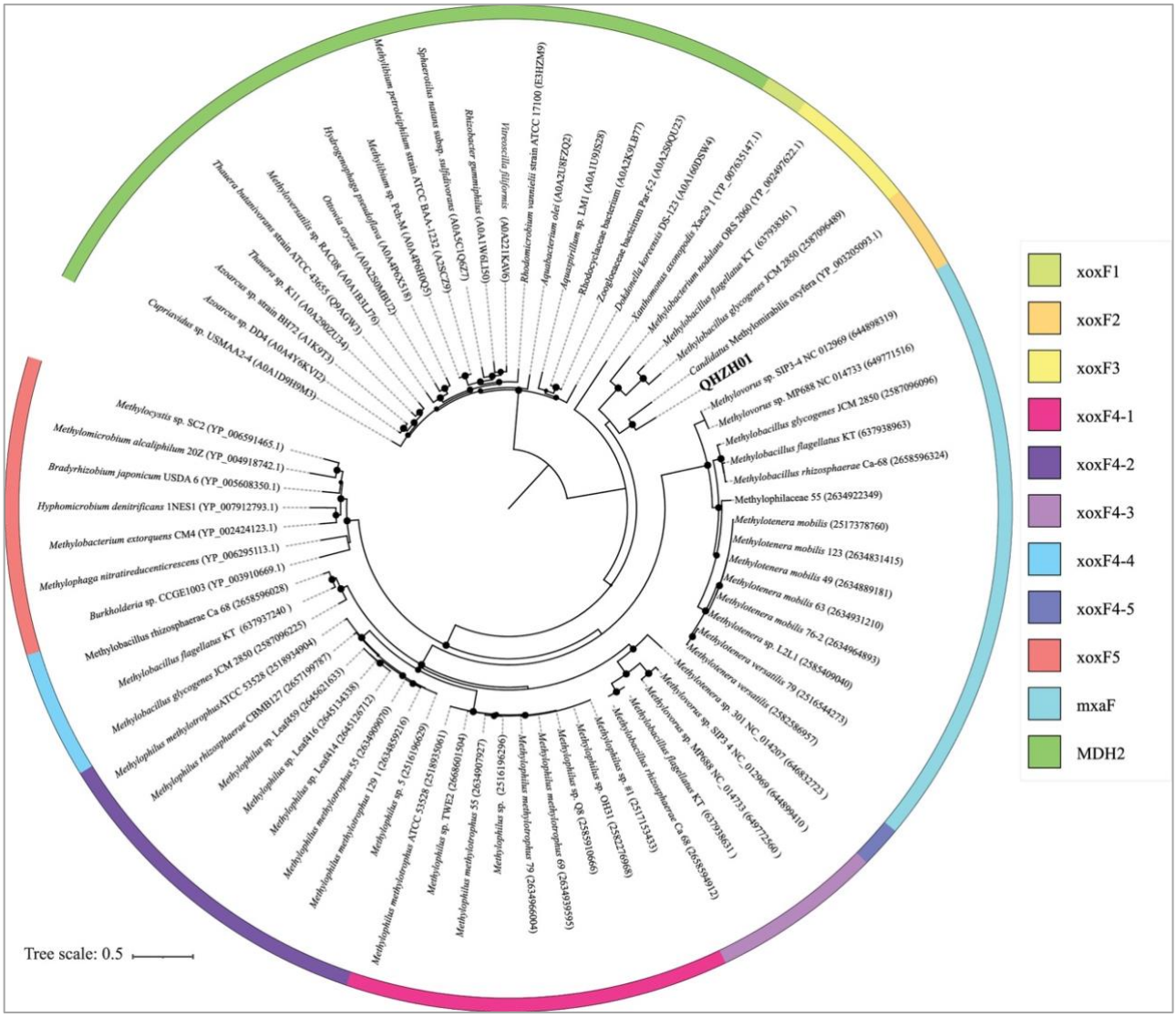


Figure 2. 4 Phylogenetic affiliation for family UBA6911 methanol dehydrogenase (XoxF) in relation to reference sequences. The genus QHZH01 sequence is shown in boldface. The track around the tree depicts the XoxF clade as shown in the key. The tree was rooted using methanol dehydrogenase sequences belonging to the MDH2 group previously described in members of Burkholderiales [127]. UniProt, IMG, and/or GenBank accession numbers are shown for reference sequences. Bootstrap support values are based on 100 replicates and are shown for nodes with >70% support.

Phylogenetic affiliation using a concatenated alignment of DsrA and DsrB proteins placed genus UBA6911 sequences close to Thermoanaerobaculia acidobacterial sequences from hydrothermal vents (Figure 2.5), while genus Ga0212092 sequences were close to group 3 Acidobacteria sequences from hydrothermal vents and peatland soil (Figure 2.5).

Finally, in addition to inorganic electron acceptors, the majority of genus UBA6911 genomes encoded a complete Wood-Ljungdahl pathway (WLP) (Figure 2.3). WLP is a versatile widespread pathway that is incorporated in the metabolic schemes of a wide range of phylogenetically disparate anaerobic prokaryotes, e.g., homoacetogenic bacteria, hydrogenotrophic methanogens, autotrophic sulfate-reducing prokaryotes, heterotrophic sulfate-reducing bacteria, and syntrophic acetate-oxidizing (SAO) bacteria, as well as acetoclastic methanogens. When operating in the reductive direction, the pathway can be used for carbon dioxide fixation and energy conservation during autotrophic growth or as an electron sink during heterotrophic fermentative metabolism. When operating in the oxidative direction, the pathway is used for acetate catabolism [128-130]. Syntrophic acetate oxidizers employing the oxidative WLP usually possess high-affinity acetate transporters to allow for the uptake of small concentrations of acetate, a competitive advantage in the presence of acetoclastic methanogens [131]. The absence of genes encoding high-affinity acetate transporters in any of the genomes argues against the involvement of WLP in syntrophic acetate catabolism. The possibility of its operation for autotrophic CO2 fixation is also unlikely, due to the absence of evidence for utilization of an inorganic electron donor (e.g., molecular H_2). Its most plausible function, therefore, is acting as an electron sink to reoxidize reduced ferredoxin, as previously noted in "*Candidatus* Bipolaricaulota" genomes [131]. The Rhodobacter nitrogen fixation (RNF) complex encoded in the majority of genus UBA6911 genomes would allow the reoxidation of reduced ferredoxin at the expense of NAD, with the concomitant export of protons to the periplasm, thus achieving redox balance between heterotrophic substrate oxidation and the WLP function as the electron sink. Additional ATP production via oxidative phosphorylation following the generation of the proton motive force is expected to occur via the F-type ATP synthase encoded in all genomes. Recycling of electron carriers would further be achieved by the cytoplasmic electron-bifurcating mechanisms HydABC and MvhAGD-HdrABC, both of which are encoded in the genomes (Figure 2.3).

2.5.6 Fermentative capacities

In addition to respiration, elements of fermentation were also encoded in all genomes. All MAGs encoded pyruvate dehydrogenase, as well as 2-oxoacid ferredoxin oxidoreductase for pyruvate oxidation to acetyl-CoA. Genes encoding fermentative enzymes included the acetate production genes (acetyl-CoA synthase [EC 6.2.1.1] in the soil genus and UBA6911 genomes, acetate:CoA ligase [ADP forming] [EC 6.2.1.13] in UBA6911 and Ga0209509 genus genomes, and phosphate acetyltransferase [EC 2.3.1.8] and acetate kinase [EC 2.7.2.1] [the latter two-gene pathway is associated with substrate-level phosphorylation and was encoded only in UBA6911 genomes]), genes for ethanol production from acetate (aldehyde dehydrogenase [EC 1.2.1.3] and alcohol dehydrogenases) in UBA6911 and Ga0209509 genus genomes, genes for ethanol production from acetyl-CoA (aldehyde dehydrogenase [EC 1.2.1.10] and alcohol dehydrogenases) in QHZH01 and Ga0212092 genomes, genes for formate production from pyruvate (formate C-acetyltransferase [EC 2.3.1.54] and its activating enzyme) in UBA6911 and Ga0209509 genus genomes, genes for acetoin and butanediol production from pyruvate in all but QHZH01 genomes, and genes for acetone production from acetyl-CoA (acetyl-CoA C-acetyltransferase [EC 2.3.1.9], acetate CoA/acetoacetate CoA-transferase alpha

subunit [EC 2.8.3.8], and acetoacetate decarboxylase [EC 4.1.1.4]) in UBA6911 genomes (Figure 2.3 and Table 2.2; Table S3).

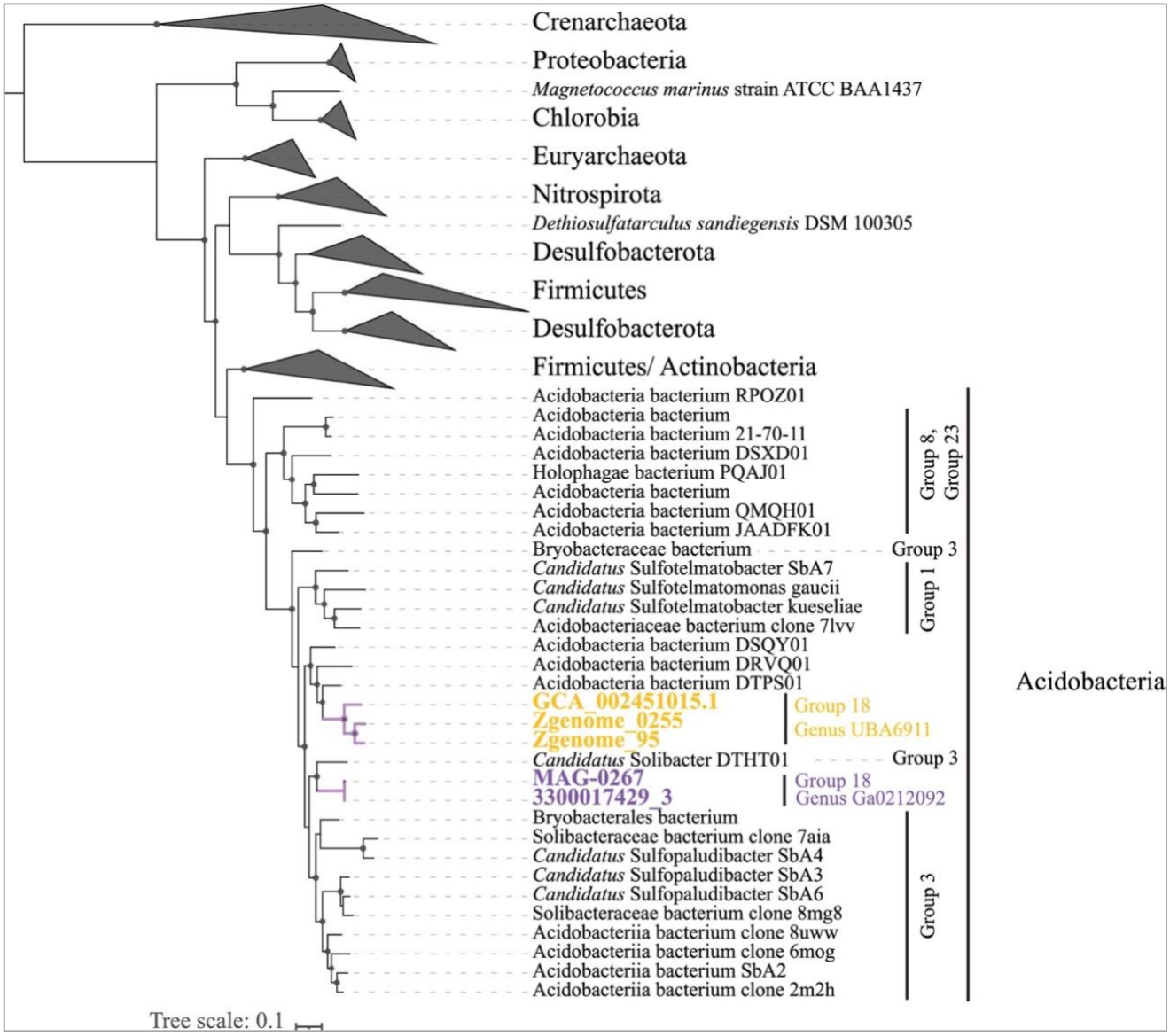


Figure 2. 5 Maximum likelihood phylogenetic tree based on the concatenated alignment of the alpha and beta subunits of dissimilatory sulfite reductase (DsrAB) from family UBA6911 in relation to reference sequences. Genus UBA6911 sequences are shown in yellow, while genus Ga0212092 sequences are shown in purple. Reference sequences from phyla other than Acidobacteria are shown as gray wedges. Acidobacteria references are labeled by the subgroup number. Bootstrap support values are based on 100 replicates and are shown for nodes with >70% support.

2.6 Discussion

Our analysis reveals multiple defining features in all examined family UBA9611 genomes regardless of their origin or genus-level affiliation. All members of the family are predicted to be heterotrophs, with an expected robust anabolic capacity as well as a high level of catabolic versatility. Catabolism of a wide range of sugars, proteins, and amino acids by members of this family is predicted. In addition, all members of this family appear to possess respiratory capacities. Such traits are similar to known metabolic capacities of the majority of examined genomes within the Acidobacteria [61, 83, 132]. However, our analysis uncovered multiple interesting differences between the examined genomes, all of which appear to be genus specific and given the strong preference of various genera for specific habitats. Such differences appear to differentiate the two soil genera from the three nonsoil (freshwater and anaerobic digestor) genera (Table 2.2 and Figure 2.3; Table S3) and hence provide clues to the genus-level adaptation to the terrestrialization process within family UBA6911. Three main differences are notable: genomic architecture, substrate-level preferences, and respiratory/electron-accepting processes.

Soil genera gp18_AA60 and QHZH01 possess significantly larger genome sizes (8.05 ± 0.32 Mbp), than freshwater genomes (5.28 ± 0.73 Mbp) and anaerobic digestor genomes (3.5 ± 0.75 Mbp) (Table 1). Soil Acidobacteria usually [83, 132, 133], but not always [83, 134, 135], exhibit large genome sizes, compared to nonsoil Acidobacteria (for examples, see references [85] and [86]; for a review, see reference [136]), strongly arguing that genome expansion is associated with terrestrialization in the Acidobacteria. The association between genome size expansion and adaptation to soils has been identified as a general trait across the microbial world [137], including Acidobacteria [132, 136]. Such expansion has been attributed to a larger number of paralogous genes as a mechanism for generation of new functions or optimal resource utilization under the different environmental conditions occurring in the highly complex and spatiotemporally dynamic soil ecosystem [138]. Our analysis suggests that such expansion in gp18_AA60 and QHZH01 genomes is due to a broad increase in the number of genes in soil genera across a wide range of cellular/metabolic functions, rather than gene duplication or differences in coding density. In addition to genome size expansion, soil genus gp18_AA60 exhibited a high number of CRISPR loci (24.67 ± 6.5) (Table 1). In bacteria, CRISPR repeats arise from invading genetic elements that are incorporated into the host's CRISPR locus. These short sequence tags (spacer sequences) are subsequently transcribed into small RNAs to guide the destruction of foreign genetic material [139]. Within the Acidobacteria, an examination of 177 genomes (including MAGs and SAGs) affiliated with the phylum Acidobacteria in the IMG database revealed that 67 possessed at least 1 CRISPR count, and only 4 of these possessed >5 CRISPRs. The expansion in the number of CRISPR loci in gp18_AA60 genomes is possibly a protective mechanism against higher potential for viral infection by Acidobacteria-specific viruses in soil, which is expected, given the higher relative abundance of members of this phylum in soil [73] in comparison to their low abundance in other habitats.

Multiple differences in expected catabolic capacities between soil and nonsoil family UBA6911 genera were identified in this study. First, a larger CAZyome was identified in the soil genera gp18_AA60 and QHZH01. This was mainly driven by the expansion of specific GH families, e.g., GH33 (sialidase), GH165 (b-galactosidase), and GH56 (hyaluronidase) in genus QHZH01, GH13 (amylase) in gp18_AA60, and GH109 (N-acetylhexosaminidase) in both genera. It is notable that the capacity for degradation of all released monomers (sialic acid, galactose, uronic acids, glucose, and N-acetylhexosamines) is indeed encoded in these

soil genomes. As such, acquisition of these specific CAZymes appears to be important for niche adaptation in family UBA6911 soil genomes. Second, the capacity for methanol and methylamine metabolism was predicted in members of the soil genus QHZH01. As previously noted [140], methylotrophy requires the possession of three metabolic modules for C_1 oxidation to formaldehyde, formaldehyde oxidation to CO_2, and formaldehyde assimilation. QHZH01 contains genes for all three modules. Formaldehyde assimilation via the serine cycle requires regeneration of glyoxylate from acetyl-CoA to restore glycine and close the cycle. In addition to all three modules described above, the QHZH01 genome also encodes the ethylmalonyl-CoA pathway for glyoxylate regeneration. Soil represents a major source of global methanol emissions [141], where demethylation reactions associated with pectin and other plant polysaccharide degradation contribute to the soil methanol pool. QHZH01 methanol dehydrogenase belongs to the XoxF family (Figure 2.4), previously detected in 187 genomes, including Acidobacteria SD1, SD5, and SD6 genomes [80], binned from Angelo Reserve soil, and identified as one of the most abundant proteins in a proteomics study from the same site [142]. In contrast, some genus UBA6911 genomes possessed capacities for formaldehyde degradation and assimilation but lacked any genes or modules for conversion of other C_1 substrates to formaldehyde. The functionality and value of such truncated C_1 machinery remain unclear. Formaldehyde in freshwater environments could be present as a contaminant in aquaculture facility effluent [143] or as a product of C_1 metabolism by other members of the community.

Finally, while all family UBA6911 genomes encoded respiratory chains components, distinct differences in predicted respiratory capacities were observed. Soil genera gp18_AA60 and QHZH01 encoded only an aerobic respiratory electron transport chain. In contrast, all anaerobic digestor genus Ga0209509 genomes contained genes for dissimilatory nitrite reduction to ammonium, suggesting a capacity and/or preference to grow in strict anaerobic settings. Freshwater genera Ga0212092 and UBA6911 were the most diverse, encoding aerobic capacity (in all but one genome), dissimilatory nitrite reduction to ammonium (in four UBA6911 genomes), trimethylamine N-oxide respiration (in both Cone Pool Ga0212092 genomes), potential use of WLP for electron acceptance (in genus UBA6911 genomes), and dissimilarity sulfate reduction capacities (in both Cone Pool Ga0212092 and some UBA6911 genomes). This versatility might be a reflection of the relatively broader habitats where these genomes were binned, many of which exhibit seasonal and diel fluctuation in O_2 and other electron acceptors levels. Of note is the presence of the complete machinery for dissimilatory sulfate reduction, a process long thought of as a specialty for Deltaproteobacteria (now Desulfobacterota [144]), some Firmicutes, and a few thermophilic bacteria and archaea [145]. However, with the recent accumulation of metagenomics data, genes for dissimilatory sulfate reduction were identified in a wide range of phyla [52, 144]. In the Acidobacteria, the presence and activity of genes for sulfate reduction have been reported from peatland samples (subdivisions 1 and 3) [78]. Subsequently, dissimilarity sulfate reduction capacities were identified in more Acidobacteria MAGs (subdivision 23) from mine drainage [52] and, recently, from marine fjord sediments of Svalbard [88]. Our study adds Acidobacteria family UBA6911 (subgroup 18) to the list of the dissimilatory sulfite reductase (DSR)-harboring Acidobacteria and suggests a role in freshwater habitats. The concurrent occurrence of dissimilatory sulfate reduction and aerobic respiration in the same MAG is notable but has been previously reported in the subdivision 23 Acidobacteria MAGs from marine fjord sediments of Svalbard [88]. All Acidobacteria DsrAB sequences (subdivisions 1, 3, 18, and 23) cluster within the reductive bacterial type branches in DsrAB concatenated phylogenetic trees, away from the oxidative Chlorobia and Proteobacteria (Figure 2.5). Also, the absence of DsrEFH homologs, known to be involved in the reverse DSR pathway, argue for the involvement of Acidobacteria DsrAB in the reductive dissimilatory sulfate reduction.

2.7 Acknowledgement

This work was supported by National Science Foundation grant no. 2016423 to N.H.Y. and M.S.E.

2.8 Supplemental Material

All Supplementary data for this work can be viewed online at:
https://journals.asm.org/doi/full/10.1128/AEM.00947-21

CHAPTER III

A "REVERSE EVOLUTION" APPROACH TO IDENTIFY STRATEGIES IN *COXIELLA BURNETII* INTRACELLULAR SURVIVAL

3.1 Abstract

Coxiella burnetii (Cb) is an obligate intracellular pathogen in nature and the causative agent of acute Q fever as well as chronic diseases. The bacterium infects macrophages and successfully propagates in a parasitophorous vacuole (PV) that has the properties of a phagolysosome. In an effort to identify genes and proteins crucial to their normal intracellular growth lifestyle, we applied a "reverse evolution" approach where the avirulent Nine Mile Phase II strain of Cb was grown for 67 passages in chemically defined ACCM-D media and gene expression patterns and genome integrity from various passages was compared to passage number 1 following intracellular growth. Transition and passaging of Cb into axenic media caused a decrease in Cb infectivity to HeLa cells. Transcriptomic analysis identified a marked downregulation of the structural components of the type 4B secretion system (T4BSS), as well as the general secretory (sec) pathway. Out of 118 previously identified genes encoding effector proteins, 14 were significantly downregulated, and these were involved in signal transductions, carbohydrate metabolisms, posttranslational modification, and lipid metabolisms. Additional downregulated pathogenicity determinants genes included several chaperones, LPS, and peptidoglycan biosynthesis. A general marked downregulation of central metabolic pathways such as Glycolysis, TCA cycle, Electron transport chain and FA biosynthesis is observed, which was balanced by a marked upregulation of genes encoding transporters. Such a pattern is a reflection of the richness of the media and diminishing anabolic and ATP-generation needs. Further, we identified 30 hypothetical proteins (13 cytoplasmic, 8 inner membrane, 1 extracellular, 1 periplasmic and 7 unknown localization) that were significantly downregulated. These may represent determinants with potential roles in intracellular survival. Finally, genomic analysis showed an extremely low mutation rate, with only 12 consensus mutations identified, only 1 of which could theoretically be implicated in the downregulation of its gene. These findings suggest that Cb gene expression changes significantly following acclimation to axenic media, although extensive genomic rearrangement does not occur.

3.2 Introduction

Coxiella burnetii (Cb), the causative agent of acute and chronic Q fever [146-151], is an obligate intracellular pathogen that infects macrophages, and successfully propagates in the parasitophorous vacuole (PV) [152-154]. Cb has evolved multiple strategies to tolerate and thrive in the PV, in spite of the prevailing low pH (≈ 4.5), low O_2 content, oxygen radicals, and high level of degradative host factors such as acid hydrolases and defensins [153, 155-157]. Such remarkable ability has been the subject of a wide range of studies that employed a plethora of biochemical, genetic, imaging, and –omics-based approaches. Further, Cb employs a type IVb secretion system (T4BSS) to deliver effector proteins into the host throughout infection [151-154, 158]. Cb effector proteins identified so far mediate a variety of biochemical activities and are known to target and modulate a broad array of host functions [151, 159-165]. Prior studies have employed bioinformatic tools [166], transposon mutagenesis [163, 164, 167-169], microscopic localization studies [166, 170-172] and cloning and infectivity testing to identify and characterize effector proteins. In addition, *Legionella pneumophila*, a close genetic neighbor of Cb with a very similar T4BSS [173-175], is known to use T4BSS-effector protein duality to infect its natural host cell, the amoeba. *L. pneumophila* has been extensively used as a proxy to identify putative effector proteins and propose molecular pathogenesis mechanisms in Cb [176-180]. Indeed, research on *L. pneumophila* has identified the structural features of the T4BSS, the nature of effector proteins secreted through the system, and possible function of some of these effectors.

Growth of Cb in an axenic media was first reported in 2009 using the undefined Acidified Citrate Cysteine Media (ACCM) media [157]. Increased replication rates in the somewhat more defined ACCM-2 medium soon followed in 2011 [181]. Subsequently, a nutritionally fully defined media (ACCM-D) with an even greater replication rate and physiologic parallels to intracellular bacteria was developed [182]. Growing Cb in axenic media is opening new venues for investigating mechanisms of Cb molecular pathogenesis [154, 164, 183-187]. Theoretically, when grown in axenic media, the expression of genes required for intracellular survival and host cell manipulation is no longer required for Cb viability. As such, continuous maintenance and passaging the bacterium for extended periods of times under axenic conditions could potentially remove the powerful selective pressure exerted by the host cell, thus potentially minimize/silence expression in such genes. As such, we posit that transcriptomic analysis of gene expression patterns as well as genomic identification of mutation and gene loss patterns in axenic grown versus Cb cultures derived from intracellular growth could be employed for identifying putative involvement of specific genes, as well as identification of novel genes necessary for Cb pathogenesis and survival in an intracellular environment. Similarly, continuous passaging could also lead to the propagation of mutations, DNA fragment losses, and rearrangements in genes/loci associated with intracellular survival, pathogenesis, and host cell manipulation. Such patterns could be regarded as "reverse evolution" i.e., the opposite of the natural evolution trajectory of Cb from a free-living ancestor to an obligate intracellular pathogen. Specifically, we hypothesized that: 1) changes in gene expression within the first few passages upon transition from intracellular to axenic media growth would be observed, and such differences would be more pronounced in genes involved in subverting and coopting host metabolism, as well as genes enabling general adaptation to physiological conditions prevalent in its intracellular vacuolar environment, and 2) Cb could acquire and accumulate DNA mutations upon transition from intracellular to axenic media growth after repetitive passages since certain bacterial genes/proteins are no longer required for successful growth.

In this study, we transitioned Cb Nine Mile phase II from cell cultures into axenic defined media ACCM-D and subcultured it into a long-term successive passage. We conducted transcriptomic and genomic sequencing on replicate samples at different time points (passages) to document temporal changes in gene expression patterns, and DNA mutations associated with adaptation to an axenic extracellular lifestyle.

3.3 Materials and Methods

3.3.1 Microorganism and growth conditions

Coxiella burnetii avirulent strain Nine Mile phase II (NMII), clone 4 (RSA439) was cultivated in rabbit epithelial RK13 cells (CCL-37; American Type Culture Collection) grown in Dulbecco's modified Eagle medium DMEM (ThermoFisher Scientific) supplemented with 5% fetal bovine serum in T75 culture flasks. This method of collecting cells was adapted from [188]. Briefly, the infected cell line was split into multiple non-vented and capped T150 culture flasks that were incubated at 37°C in 5% CO_2 for a week until confluent growth was observed. These flasks were then screwed tightly and left at room temperature for 2 weeks to induce cells to switch to the small cell variant (SCV) form. The cells were pelleted by ultracentrifugation (12,000 x g, 15 minutes) in 250 ml Nalgene round bottom tubes, scrapped off the round bottom tubes by using sterile 1X phosphate buffered saline (PBS) and then lysed by using Dounce homogenize. The lysed cells in PBS were then spun via centrifugation using Oakridge tubes in an ultracentrifuge at 12,000 x g for 15 minutes. The SCV pellets obtained were stored in SPG freezer media (0.7 M sucrose, 3.7 mM KH_2PO_4, 6.0mM K_2HPO_4, 0.15 M KCl, 5.0 mM glutamic acid, pH 7.4) at -80°C.

3.3.2 Axenic growth in defined ACCM-D media

Cb cultures propagated intracellularly in rabbit epithelial RK13 cells were used to inoculate ACCM-D media (Sunrise Science Products, San Diego, CA). Approximately 10^6 genome equivalents per mL was used as an inoculum (determined using the RT-PCR procedure in reference [189]. Cultures were grown in a T25 cell culture flasks at 5% O_2, 5% CO_2 and 37°C in a trigas incubator (Panasonic, MCO-170ML) for 7 days. Subsequent passages were achieved via a 1:1000 (6 µl into 6 ml) inoculum into freshly prepared ACCM-D media and incubation for 7 days. Axenically-grown Cb cultures were routinely (every 5 passages) subjected to contamination check by; inoculation into LB broth medium incubated under microaerophilic (5% O_2, 5% CO_2) conditions LB broth medium incubated aerobically at 37°C, as well as ACCM-D medium incubated aerobically at 37°C.

3.3.3 Measuring Growth and Host Cell Infectivity

To determine the infectivity of axenic- or intracellularly-grown Cb; HeLa cells (CCL-2; American Type Culture Collection) were seeded onto 96 well culture plates at a density of 10^4 in Roswell Park Memorial Institute (RPMI) medium containing 2% fetal bovine serum (FBS) for 16 hours. Cb cultures grown in ACCM-D were pelleted at 12000 x g at 4°C for 15 minutes. Serially passaged Cb were diluted in RPMI to normalize the number of genomes per volume, and 50 µl from various dilutions were inoculated onto the HeLa cell containing wells and centrifuged at 600 x g for 15 minutes at room temperature [190]. Immediately following centrifugation, the inoculating media was replaced with 200 µl of fresh RPMI containing 2% FBS. The plates

were incubated at 37°C and 5% CO_2 for 72 hours, fixed with ice-cold methanol for 10 minutes, then examined using indirect fluorescent antibody microscopy analysis as described previously [190]. Briefly, *C. burnetii* was stained using rabbit whole anti-*C. burnetii* NMII antibody diluted 1:1000 in PBS containing 3% bovine serum albumin (BSA) as a blocking agent. Primary antibodies were detected using Alexa Fluor 488 labeled goat anti-rabbit IgG antibodies diluted 1:1000 in PBS containing 3% BSA (Invitrogen). Total DNA was stained using 4',6-diamidino-2- phenylindole (DAPI) diluted 1:10000 in PBS containing 3% BSA (Molecular Probes) to illuminate host cell nuclei. The methanol fixed and stained cultures were visualized on a Nikon Eclipse TE2000-S and the number of maturing PVs were counted and calculations performed to ascertain the number of fluorescence forming units, which indicates the infectivity of the *C. burnetii* NMII in the ACCM-D samples.

3.3.4 Transcriptomics

3.3.4.1 RNA extraction

Cells from axenic media growth passages 1, 3, 5, 10, 12, 16, 21, 31, 42, 51, 61 and 67 were harvested for transcriptomic analysis. RNA was extracted using a combination of hot Trizol treatment [191] and the RNeasy Mini kit (Qiagen, Germany). Briefly, bacteria in 12 ml of ACCM-D culture (OD_{600} ~ 0.3-0.4) were pelleted, resuspended in 700 µl of Trizol (ThermoFisher Scientific), boiled at 90°C for 10 min, and vortexed vigorously. 200 µl Chloroform was then added, followed by centrifugation at 12,000 x g at 4°C for 10 min. After separation, 300 µl of 100% ethanol was added to the aqueous phase, which was then quickly transferred to the spin column provided in the RNeasy Mini kit. On-column DNA digestion was conducted by adding 80 µl (10 µl 1 Unit/µl RNase free DNase I, (ThermoFisher Scientific) in 70 µl reaction buffer from the Master Pure Yeast RNA Purification kit, Epicenter) of DNase preparation. The RNeasy Mini kit's protocol was followed for washing and eluting RNA. RNA quality was assessed visually on a gel as well as using RNA screen tape (Agilent) and RNA integrity number (RIN) value measurements using Tapestation and Bioanalyzer systems (Agilent).

3.3.4.2 Transcriptome sequencing and assembly

RNA sequencing (RNA-Seq) was conducted on the Illumina platform, using Nextseq 500 sequencer at Oklahoma State University Genomics and Proteomics core facility. Trimmomatic v0.38 [94] was used to process raw reads and remove Illumina adapter sequences. HISAT2 v2.1.0 [192] was used to map the trimmed reads to the Chromosome (GenBank accession number: CP020616.1) and Plasmid (GenBank accession number: CP020617.1) of Cb NMII RSA 439. StringTie v2.1.4 [193] was used to assemble reads alignments into potential transcript and to generate a non-redundant set of transcripts. The Python script prepDE.py supplemented with StringTie tool was used to convert transcripts per kilobase million (TPM) and fragments per kilobase million (FKPM) to gene level raw count matrix. The raw count table was imported to DESeq2 package [194] from Bioconductor in R programming language for further analysis.

3.3.4.3 Identification and analysis of Differentially Expressed Genes (DEGs)

The overall strategy for comparative transcriptomics analysis is outlined in Figure 3.1. DESeq2 was used to compute the fold change expression levels (reflected by logarithmic two-fold expression change i.e., L2fc) and its statistical significance (adjusted p-value, padj henceforth referred to as p-value) for every gene between passages when compared to passage one. DEseq2 tests the differential expression using negative binomial distribution and internally normalizes the counts by library size [195]. Genes with a p-value < 0.05 were labeled as significantly expressed. Only genes with TPM values > 10 in at least one passage were considered

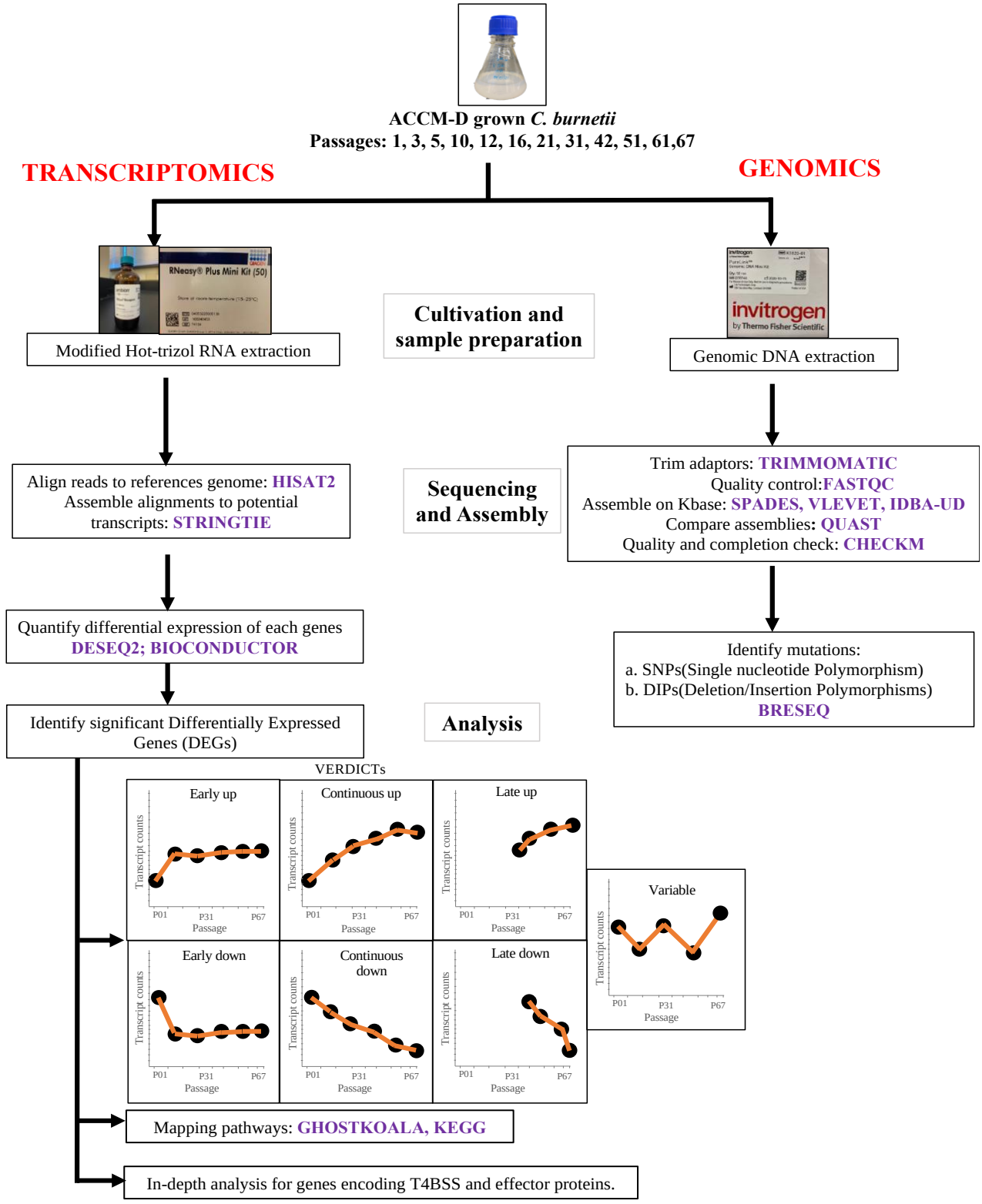


Figure 3. 1 Flowchart representing the overall comparative transcriptomics and genomics strategy employed in this study.

to minimize noise from minimally expressed genes. In most cases, differentially expressed genes in our temporal analysis were significantly expressed in more than one sampling point. In the few cases where differential expression was observed as a single spike in only one time point, a threshold of L2fc > 2 was considered as differentially expressed. Patterns of differential expression in DEGs was analyzed and visualized by constructing plotcount graphs using the function "plotCounts" in DESeq2 package. The package "EnhancedVolcano" was used to visualize expression patterns as volcano plots. Differentially expressed patterns are classified into 1- Early up/downregulated, i.e., differential expression occurred in early (before passage 31) and the levels were sustained in subsequent late passages (see Figure 3.1). 2- Continuously up/downregulated, i.e., a constant/gradual increase in the magnitude of L2fc was observed throughout the sampling process (see Figure 3.1). 3- Late up/downregulated, i.e., differential expression was observed at or after passage 31. 4- Variable, i.e., expression levels were significantly higher than passage 1 in some timepoints and significantly lower than passage 1 in other time points (see Figure 3.1).

3.3.4.4 Metabolic analysis and pathway mapping of DEGs

The subcellular protein localizations of proteins encoded by DEGs were predicted by using PSORTb [196]. Transporter Classification Database (TCBD) was queried to find the putative transporter proteins. Pfam database [197] was used to identify putative protein families for hypothetical proteins. BlastKOALA [108] was used for functional annotation and assign KEGG orthology (KO) numbers for the selected differentially expressed genes; and KEGG mapper [109] was then used to reconstruct metabolic pathway to visualize the differentially expressed genes in each pathway. The gene involvement in specific metabolic pathways were inferred from KEGG brite hierarchy file. Cluster of Orthologous genes (COGs) database (updated 2020) [198] downloaded from NCBI, was used to classify the effector proteins into functional categories.

3.3.5 Genomics

3.3.5.1 DNA extraction and sequencing

8 ml of Cb cultures grown in ACCM-D for 7 days were pelleted, by centrifugation at 12,000 x g and 4°C for 15 minutes. DNA extraction was conducted using Pure Link® Genomic DNA Kits (ThermoFisher Scientific) following the manufacturer's instructions. Sequencing was conducted at Oklahoma State University Genomics and Proteomics core facility using Illumina's NextSeq® 500 System. DNA quality was assessed visually on a gel as well as using DNA Screentape and Bioanalyzer systems (Agilent).

3.3.5.2 Genome assembly and quality control

The KBase platform [199], which implements and integrates multiple bioinformatic tools, was used for DNA sequence data handling. Trimmomatic v 0.36 [94] was used to trim the Illumina adapter sequences. Quality check was done using FastQC v0.11.5 [200]. Assembly of Illumina reads to contigs was attempted using four different assemblers (Spades v3.13.0, Velvet v1.2.10 and IDBA-UD v1.1.3 and Unicycler [201]). The quality of genome assemblies from these four assemblers were assessed using QUAST v1.4 [202] and the best assemblies were selected using metrices such as total length, largest N50, lesser number of contigs and less Ns. CheckM [21] was used to assess quality and completion of genomes (Figure 3.1).

3.3.5.3 Analysis of mutation frequencies

Breseq [203] was used to identify mutations/changes in the genome assemblies obtained, with Passage 01 used as a reference. The occurrence and frequency of both single nucleotide polymorphisms (SNPs) and deletion-insertion polymorphisms (DIPs) were examined (as outlined in Figure 3.1). Breseq was run in

polymorphism mode, which identifies the mutations occurring in a fraction of a population in addition to consensus mutations in the entire population in a sample. This allows for the visualization of the propagation of a particular mutation as a frequency of evolved alleles and genetic diversity in the population.

3.3.6 Nucleotide sequences accession number

The whole-transcriptome and genome shotgun sequences were deposited in GenBank under the BioProject PRJNA796300 and BioSample accession numbers SAMN24840407-SAMN24840437 and SAMN24847762-SAMN24847773. The 31 transcriptomic assemblies were deposited in the SRA under project accession number SRX13723330-SRX13723360. Reads for 12 genomic assemblies can be found under SRA with accession SRX13726189-SRX13726200.

3.4 Results

3.4.1 *Coxiella burnetii* infectivity but not viability decreases with continuous passaging in axenic media

Following anecdotal observations, we sought to quantitatively assess whether serially passaged *C. burnetii* infect cultured cells less readily than cell derived bacterial stocks. Using *C. burnetii* NMII serially passaged 1, 3, 5, and 10 times in ACCM-D, we initiated infections of Hela cells with bacterial dilutions normalized by the number of genomes in each sample. When the number of fluorescence forming units (FFU) per sample were calculated, they revealed a decrease in the number of *C. burnetii* filled vacuoles in tissue culture cells as the bacteria from subsequent passages were analyzed, respectively, resulting in a nearly two-log decrease between Passages 1 and 10 (Figure 3.2A). This indicated that there were fewer bacteria per genome that were capable of initiating a typical infection following multiple passages. Next, we sought to determine if the decrease in infectivity of tissue culture cells was associated with a decrease in in-vitro viability of the *C. burnetii* as measured by colony forming units on ACCM-D agar. To address this question, we plated dilutions of passages 1, 3, 5, and 10 on ACCM-D agar plates and performed colony counts. Contrary to the decrease in infectious units (Figure 3.2A), the colony counts indicated that there was no significant change in viable bacteria relative to genomes as the organism was serially passaged (Figure 3.2B). This indicated that the number of live and replicative bacteria did not change during axenic growth, and therefore the continuous in vitro propagation was not responsible for the decrease in Cb infectivity of the cultured eukaryotic cells observed.

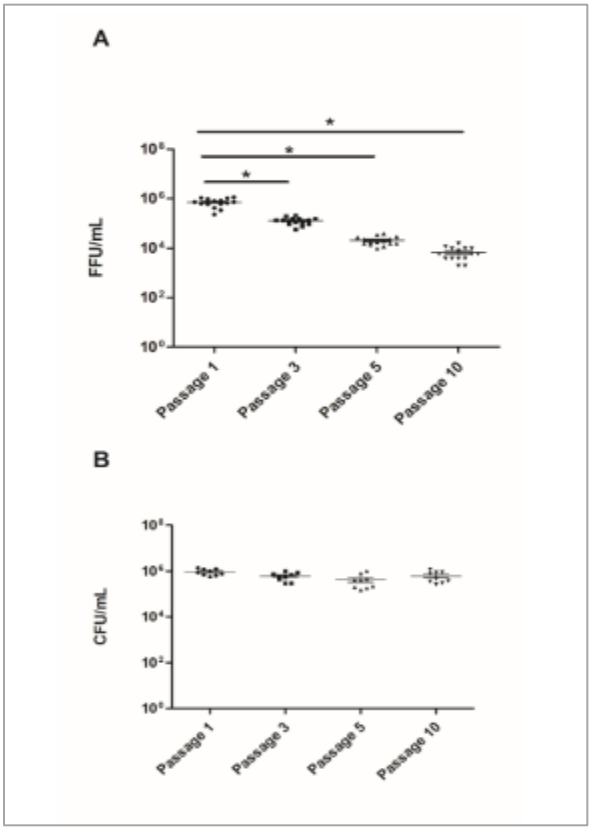


Figure 3. 2 Intracellular vs Axenic Growth Following Serial Passage of *C. burnetii* NMII. Intracellular and axenic growth from 3 biological replicates of passaged ACCM growth A) fluorescence forming units (FFU) counts of infections of HeLa cells normalized by Cb genomes from passages 1, 3, 5, and 10. Significance between different passages are indicated by lines and *p<0.001. B) CFU enumeration of passages 1, 3, 5, and 10 Cb spread on ACCM-D plates normalized to genomes. No statistically significant difference was observed between groups.

3.4.2 Transcriptional activity

RNAseq was conducted on 12 different passages (1, 3, 5, 10, 12, 16, 21, 31, 42, 51, 61, and 67). A total of 162.2 Gb data were obtained, with 6.05 – 22.57 million reads per sample (Average10.14 million reads). Transcripts representing each of the 2,217 genes in *C. burnetii* NMII strain (genome and plasmid) were identified in all samples, attesting to the depth of the sequencing effort conducted.

Expression level and overall pattern (Early up, Continuous up, Late up, Early down, Continuous down, Late down, Variable) of every gene in the Cb genome were observed. A total of 845 genes were differentially expressed in at least one passage, with 464 upregulated and 371 downregulated (Figure 3.3A). The number of differentially expressed genes (DEGs) per passage ranged between 25 and 807 (Figure 3.3B). A general pattern of an increasing number of differentially expressed genes per passage was observed through passage 51, after which the number of DEGs dropped in passage 61 and 67 (Figure 3.3B). The ratio of upregulated to downregulated genes in each passage ranged between 0.14 (in passage 5) and 1.26 (in passage 3). Of the 371 downregulated genes, 249 expressed early down pattern, 48 were continuous down, 43 were late down, and 31 were down in only one passage. Of the 464 upregulated genes, 288 were early up, 38 were continuous up, 85 were late up and 53 were up in only one passage (Figure 3.3C). Of the 845 DEGs, 81 were differentially regulated in 8-11 of the passages, 144 in 5-7 of the passages, 526 in 2-4 of the passages and 84 in only one passage (Figure 3.3D).

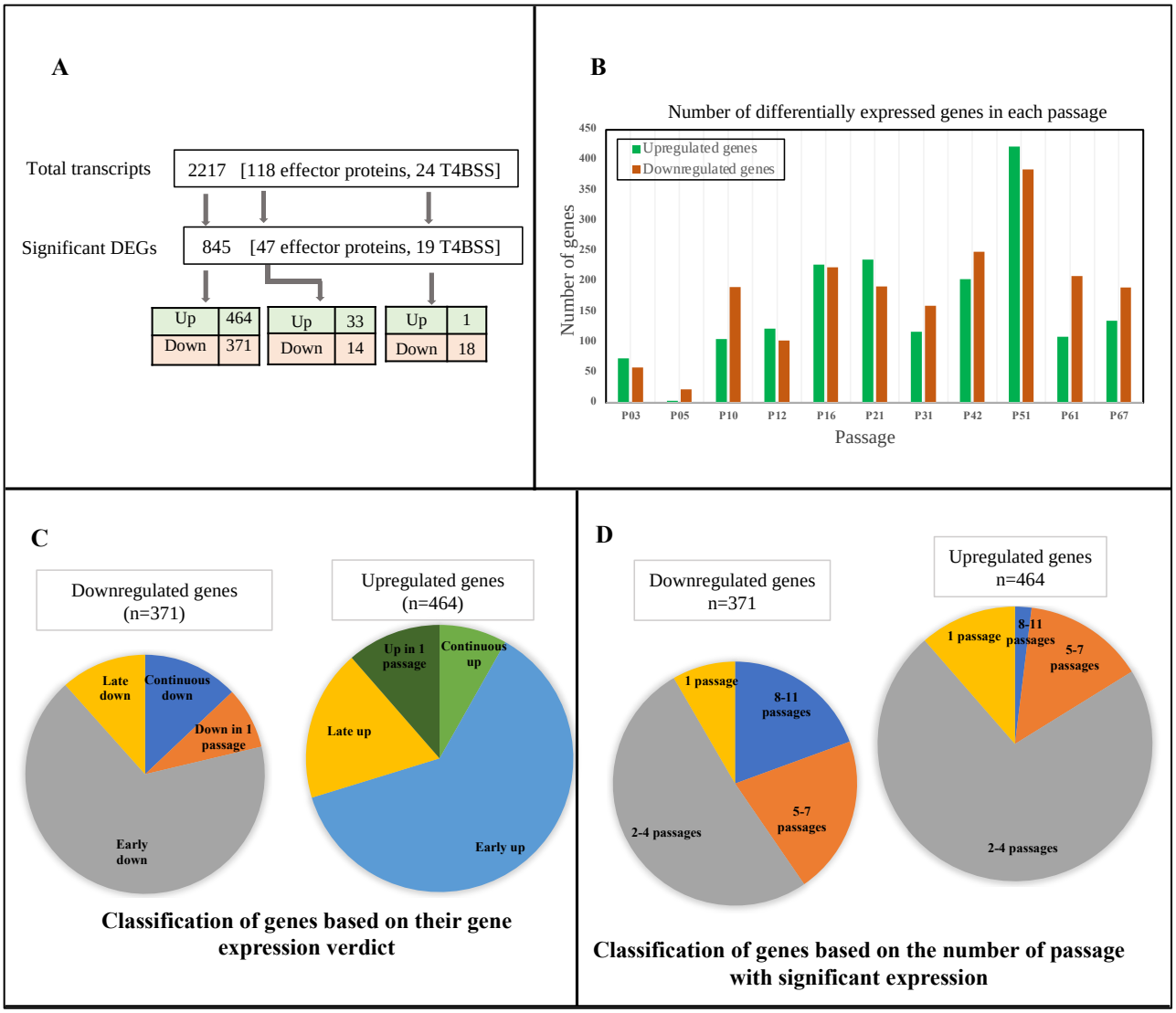


Figure 3. 3 Overview for Differential Expression patterns in axenically-grown Cb. A) Summary for the number of total and differentially expressed genes, T4BSS and effector proteins in this experiment. B) Bar graph showing the number of upregulated (p-value < 0.05 and $L_2fc > 0$) and downregulated (p-value < 0.05 and $L_2fc < 0$) genes in each passage. C) Classification of genes based on their gene expression verdicts. D) Classification of genes based on the number of passages showing significant expression.

Visualization of DEGs patterns using volcano plots was used to provide an overview of the overall level of expression changes (see Figure 3.4). Transcript expression levels from each passage were compared to passage 1. The labeled boxes within each plot analysis represents the 10 highest differentially expressed transcripts (i.e., smallest p-value). Visual inspections demonstrate that chaperons and T4BSS machinery proteins consistently represent an important component of highly downregulated genes in all passages. Below, we provide a more detailed assessment on differential expression patterns for various genes and pathways.

3.4.3 Secretory pathways are significantly downregulated in axenic growth media

The defective in organelle trafficking/intracellular multiplication (Dot/Icm) Type IVB secretion system (T4BSS) in Cb has been shown to secrete the effectors and other pathogenic determinants into the host cell, a process required for Cb intracellular growth and pathogenesis [151, 154, 158, 163, 167]. Interestingly, 19 out of the 24 components of the Cb T4BSS demonstrated significant differential expression (Figure 3.5A, Figure S1a). Out of these, 18 genes were downregulated and only 1 gene was upregulated (Figure 3.5A, Figure S1a). Indeed, T4BSS encoded gene transcripts were some of the most significantly downregulated across the passages (see Figure 3.4).

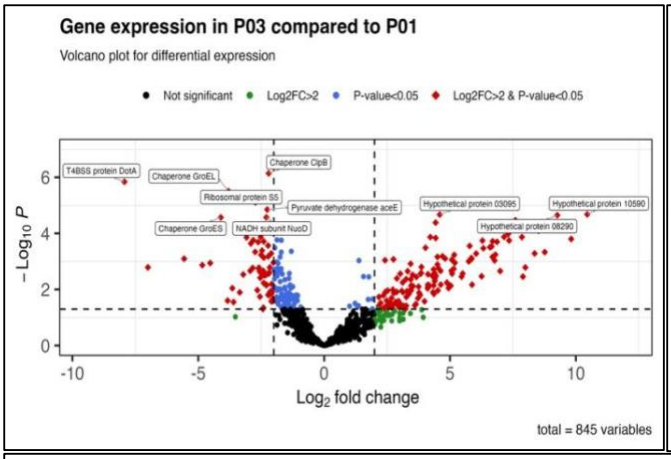
Within the T4BSS core transport complex, transcripts for genes *dotC, dotD, dotF*, and *dotG* were early downregulated, while only *dotH* indicated no significant change in gene expression in all passages.

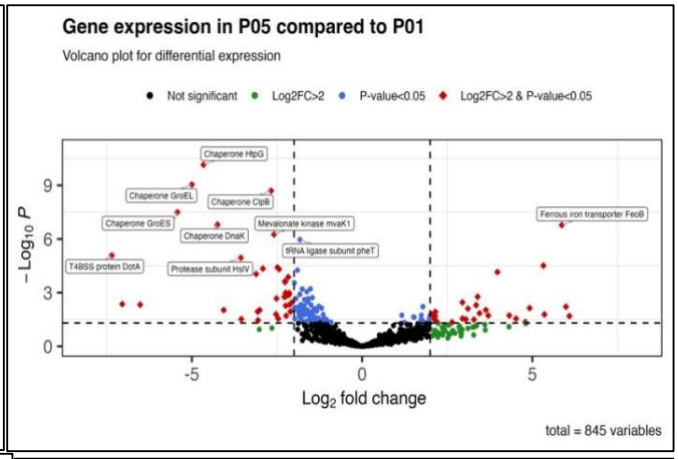
In addition, expression changes in genes encoding components of the T4BSS coupling protein complex (dotL, dotM and icmW) demonstrated early down or continuous down (dotN), whereas icmS was the only component with no significant gene expression changes. Transcripts of the gene *dotB* was continuous downregulated whereas *dotA* and *icmX* were found to be early downregulated (Figure S1a). Besides the two main complexes, other components of the Cb T4BSS that were transcriptionally downregulated during continuous axenic media passaging includes genes *dotE, dotP, dotK, dotI, dotJ, icmT* and *icmQ* (Figure 3.5a). *icmF*, located in a separate locus than the majority of the T4BSS genes (Figure 3.5b), was the only component that was transcriptionally upregulated in the system (Figure S1a).

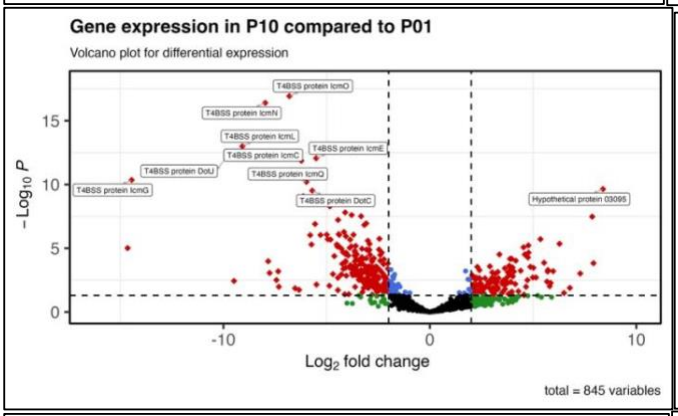
Transcript expression of genes within additional secretory pathways in Cb were also analyzed. Genes of the general secretary (Sec) pathway revealed a general trend of downregulation (Figure S1b). The Sec pathway provides a channel for polypeptide movement across the bacterial inner membrane [204]. It is comprised of the proteins SecY, SecE and SecG and an ATPase (SecA) that drives protein movement [204, 205]. This pathway is known to secrete proteins from the cytosol through the cytoplasmic membrane [206]. We identified all of the Cb Sec pathway components, as shown in Figure 3.5a. Transcripts for expression of the inner membrane proteins SecA, SecF and SecE were early down and SecY and YajC were continuously down, whereas the targeting proteins SecB and SecG were down at later passages.

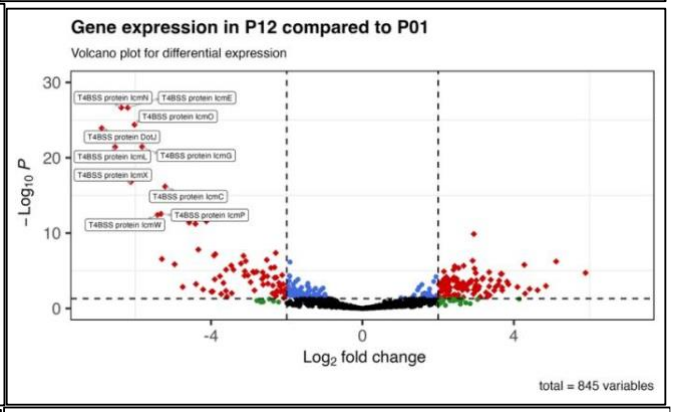
3.4.4 Expression patterns of T4BSS effector proteins previously implicated in Cb pathogenesis and intracellular survival

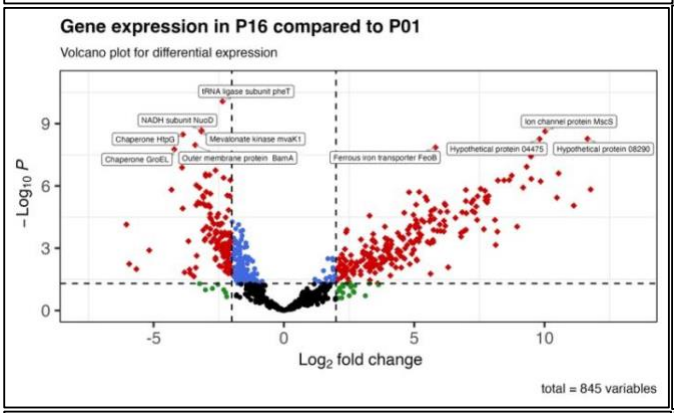
The differential expression in all 118 genes encoding T4BSS effector proteins previously identified in Cb through a variety of effector screens [166, 167, 169, 172, 207] was examined. Forty-seven effector proteins were differentially expressed (Figure 3.3a). Interestingly, more genes were upregulated (n=33) than downregulated (n= 14) (Figure 3.3a).

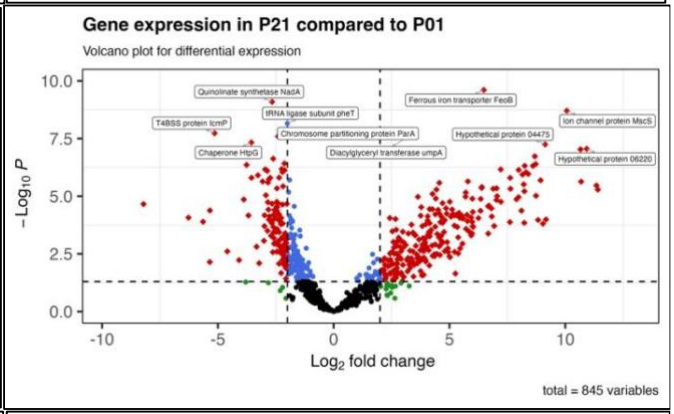

Gene expression in P03 compared to P01
Volcano plot for differential expression
Not significant
Log2FC>2
P-value<0.05
Log2FC>2 & P-value<0.05
T4BSS protein DotA
Chaperone GroEL
Chaperone ClpB
Ribosomal protein S5
Pyruvate dehydrogenase aceE
Chaperone GroES
NADH subunit NuoD
Hypothetical protein 03095
Hypothetical protein 10590
Hypothetical protein 08290
−Log10 P
Log2 fold change
total = 845 variables

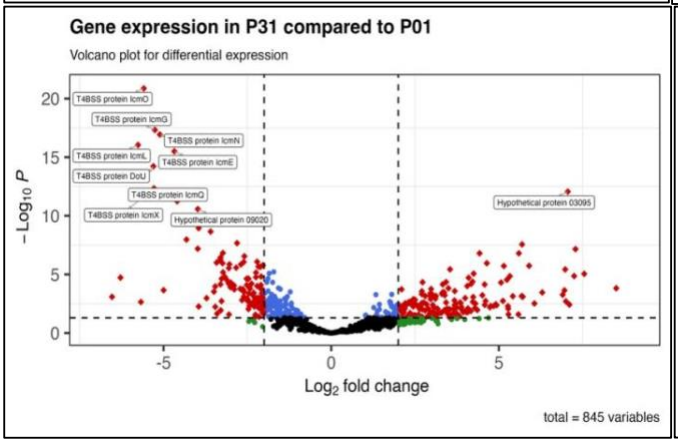

Gene expression in P05 compared to P01
Volcano plot for differential expression
Not significant
Log2FC>2
P-value<0.05
Log2FC>2 & P-value<0.05
Chaperone HtpG
Chaperone GroEL
Chaperone ClpB
Chaperone GroES
Chaperone DnaK
Mevalonate kinase mvaK1
tRNA ligase subunit pheT
T4BSS protein DotA
Protease subunit HslV
Ferrous iron transporter FeoB
−Log10 P
Log2 fold change
total = 845 variables

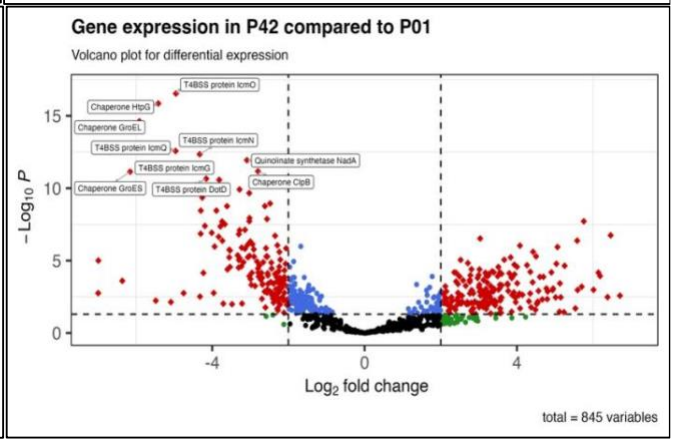

Gene expression in P10 compared to P01
Volcano plot for differential expression
T4BSS protein IcmO
T4BSS protein IcmN
T4BSS protein IcmL
T4BSS protein IcmC
T4BSS protein IcmE
T4BSS protein DotU
T4BSS protein IcmQ
T4BSS protein IcmG
T4BSS protein DotC
Hypothetical protein 03095
−Log10 P
Log2 fold change
total = 845 variables


Gene expression in P12 compared to P01
Volcano plot for differential expression
T4BSS protein IcmN
T4BSS protein IcmE
T4BSS protein IcmO
T4BSS protein DotU
T4BSS protein IcmL
T4BSS protein IcmG
T4BSS protein IcmX
T4BSS protein IcmC
T4BSS protein IcmP
T4BSS protein IcmW
−Log10 P
Log2 fold change
total = 845 variables


Gene expression in P16 compared to P01
Volcano plot for differential expression
tRNA ligase subunit pheT
NADH subunit NuoD
Ion channel protein MscS
Chaperone HtpG
Mevalonate kinase mvaK1
Chaperone GroEL
Outer membrane protein BamA
Ferrous iron transporter FeoB
Hypothetical protein 04475
Hypothetical protein 08290
−Log10 P
Log2 fold change
total = 845 variables


Gene expression in P21 compared to P01
Volcano plot for differential expression
Quinolinate synthetase NadA
Ferrous iron transporter FeoB
Ion channel protein MscS
tRNA ligase subunit pheT
T4BSS protein IcmP
Chromosome partitioning protein ParA
Hypothetical protein 04475
Chaperone HtpG
Diacylglyceryl transferase umpA
Hypothetical protein 06220
−Log10 P
Log2 fold change
total = 845 variables


Gene expression in P31 compared to P01
Volcano plot for differential expression
T4BSS protein IcmO
T4BSS protein IcmG
T4BSS protein IcmN
T4BSS protein IcmL
T4BSS protein IcmE
T4BSS protein DotU
T4BSS protein IcmQ
T4BSS protein IcmX
Hypothetical protein 09020
Hypothetical protein 03095
−Log10 P
Log2 fold change
total = 845 variables


Gene expression in P42 compared to P01
Volcano plot for differential expression
T4BSS protein IcmO
Chaperone HtpG
Chaperone GroEL
T4BSS protein IcmQ
T4BSS protein IcmN
T4BSS protein IcmG
Quinolinate synthetase NadA
Chaperone GroES
T4BSS protein DotD
Chaperone ClpB
−Log10 P
Log2 fold change
total = 845 variables

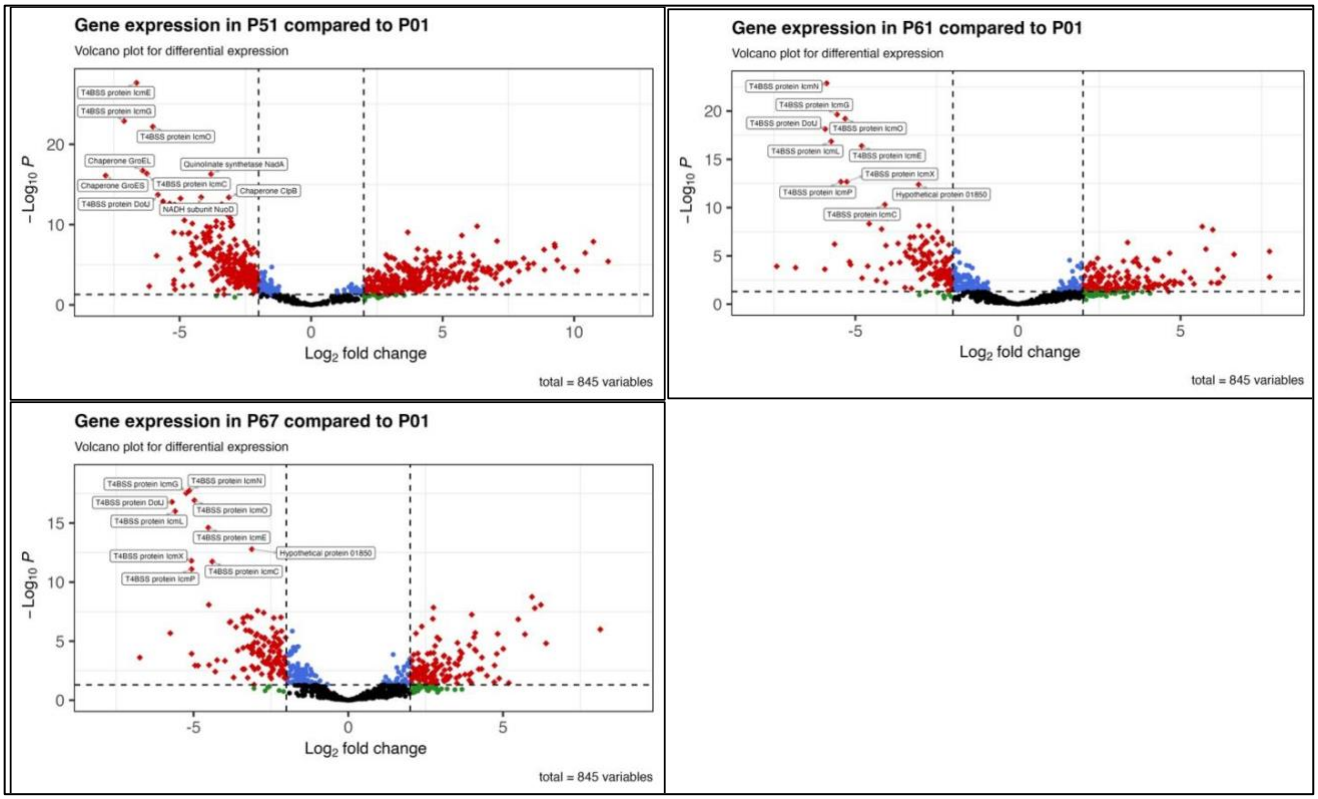


Figure 3. 4 Volcano plots for 845 significant DEGs in different passages when compared to passage 01. Black dots represent non-significant DEGs, green dots represent non-DEGs with L2fc > 2, blue dots represent significant DEGs with L2fc < 2 whereas the red diamonds represent significant DEGs with L2fc > 2. The 10 genes with the lowest p-value in every passage comparison are labeled in boxes.

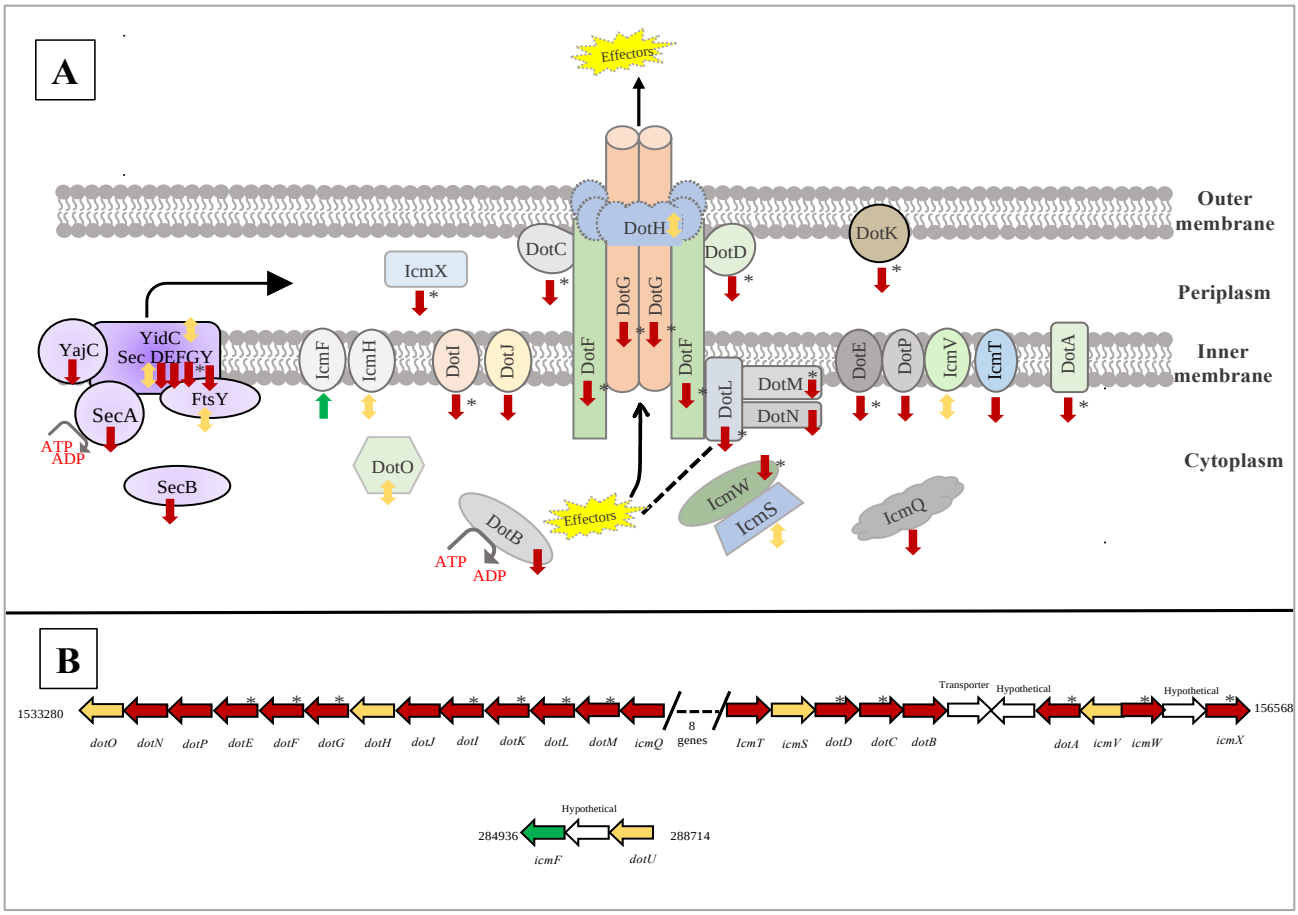


Figure 3. 5 Cb T4BSS machinery and Sec expression changes during axenic passaging. A. Membrane complex model with gene transcript expression profiles for components of Cb T4BSS (right side) and Sec protein export pathway (left side). DEG patterns are denoted by arrows (downregulated in maroon, no significant expression changes in yellow, and upregulated in green). * Represents genes that were highly downregulated (L2fc ≤ -3). B. Gene locus map for all the components in Cb NMII strain [CP020616.1]. The arrows are colored filled according to DEGs patterns as (downregulated in maroon, no significant expression change in yellow, and upregulated in green). The arrows not filled are genes not related to T4BSS pathway.

The 14 genes encoding effector proteins that were transcriptionally downregulated indicated an up to four-fold expression change, with the expression changes primarily beginning from passage 10 (Figure S2). These genes fell into COG functional groups of signal transduction mechanisms (*ankG*, *ankK* and *ankD*), carbohydrate transport and metabolism (B7L74_09020), posttranslational modification, protein turnover and chaperones (*cpeH*), replication, recombination, and repair (*cig57*), lipid transport and metabolism (B7L74_03275) and mobilome: prophages and transposons (B7L74_08400) (Table 3.1, Figure 3.6a, Figure 3.6b). Genes B7L74_08200 and *cpeF* were predicted as general function categories whereas there were no functional homologies for B7L74_07850, *cig2*, *cirC* and B7L74_03065 in the COG database (Table 3.1).

Genes	Corresppo nding CBU(Cb NM I RSA493) annotation	Gene annotation	Verdict from transcriptomics (Number of passages with significant expression)	COG categories	First verified of its potential effector functions based on	More information
B7L74_08200	Cbu1594	Hypothetical protein	Early down (3)	General function prediction only	T4BSS dependent translocation [169]	Using proteomics it is confirmed t localized in mitochondria [208]
B7L74_03065	Cbu0590	Hypothetical protein	Early down (2)	No hits on COG	Bioinformatic screening and experimental translocation assay [169]	
B7L74_11055	Cbua0034	Plasmid effector protein CpeH	Early down (2)	Posttranslational modification, protein turnover, chaperones	CyaA translocation assays of the plasmid [209]	
B7L74_03970	Cbu0781	AnkG	Early down (3)	Signal transduction mechanisms	T4SS translocation experiment [210]	[210] also shows how it localizes at host microtubule and thus might be very important for infection. [211] showed AnkG interfers with mammalian apoptosis pathway.
B7L74_10985	Cbua0023	CpeF	Early down (3)	General function prediction only	β-lactamase and adenylate cyclase translocation assays [172]	It is a member of plasmid effector family and was first detected in [172] where it was identified as effecotr based on β-lactamase and adenylate cyclase translocation assays, and identification of a C-terminal secretion signal Translocated to host cytosol. Mutation= Growth defect
B7L74_06660	Cbu1292	AnkK	Early down (3)	Signal transduction mechanisms	Presence of Eukaryotic like ORF or domain i.e., ankyrin repeats [210]	Predicted to be a effector protein because of the eukaryotic like ORF or domain i.e. ankyrin repeats. AnkK proteins are shown to be not secreted by T4SS inside the host cell [210] but it can work without being delivered to host cell. [212] showed its importance for growth inside macrophages
B7L74_03275	Cbu0635	Hypothetical membrane spanning protein	Early down (4)	Lipid transport and metabolism	BlaM translocation assay [167]	First reported in [167] where it was experimentally shown to be translocated using dot system using BlaM translocation assay. Expressed in mammalian Hela229 cells shows its distribution around golgi vesicles and potentially disturb mammalian secretory trafficking. It can interfere with host cell secretion [167].
B7L74_04780	Cbu0937	CirC	Early down (4)	No hits on COG	Localization and transposon insertion mutation studies [169]	[169] identified it as an effector by localization and transposon insertion mutation studies for intracellular replication and CCV formation. They showed that the mutant of this protein

						produced small vacuoles, which validated screens conducted by other groups that identified Cbu0937 as being an effector important for CCV biogenesis.
B7L74_07850	Cbu1525	Hypothetical protein	Early down (5)	No hits on COG	Cya transloaction studies and T4 translocation signals [167]	[167] identified it in cya transloaction studies and also by looking for T4 translocation signals. This is a frameshifted ORF.
B7L74_00115	Cbu0021	CvpB/Cig2	Early down (7)	No hits on COG	Translocation via L. pneumophila Dot/Icm system [213]	First reported in [213] where it is postulated to be an effector based on its translocation via L. pneumophila Dot/Icm system. [165] reports its mutation(by transposon insertion) displayed multi-vacuolar phenotype without an overall replication defect.. Mutation thus causes growth defect and CCV fusion defect.
B7L74_08400	Cbu1636	Hypothetical protein	Early down (8)	Mobilome: prophages, transposons	PmrA like domain containing protein and Dot/Icm dependent translocation [166].	Reported in article [166] (Table S03). Identified by Dot/Icm substrate homologus/putative PmrA regulated eukaryotic-like domain containing protein. It also has dot/Icm dependent translocation. 45 KDa protein with coiled coil structure.
B7L74_01850	Cbu0355	AnkD	Early down (8)	Signal transduction mechanisms	Presence of ankyrin repreats as well as F-box domains [210]	This protein has both ankyrin repeats as well as F-box domains [210]
B7L74_09015	Cbu1751	Cig57	Early down (9)	Replication, recombination, and repair	PmrA like domain containing protein and Dot/Icm dependent translocation [166].	From table S03 of [166], it is 49 KdA coiled coil protein. Identified by DotF binding protein/putative PmrA regulated eukaryotic-like domain containing protein. It also has dot/Icm dependent translocation. Can cause intracellular replication defects.
B7L74_09020	Cbu1752	Hypothetical protein	Early down (9)	Carbohydrate transport and metabolism	Loss-of-function mutations [164]	Was found to be important for vacuole biogenesis and its mutation resulted in a small-vacuole (CCV) phenotype. [164]
B7L74_08200	Cbu1594	Hypothetical protein	Early down (3)	General function prediction only	T4BSS dependent translocation [169]	Using proteomics, it is confirmed to be localized in mitochondria [208].

Table 3. 1 Detailed information on 14 downregulated effector proteins.

3.4.5 Expression patterns of additional pathogenic determinants in Cb

Expression patterns of additional pathogenic determinants unrelated to effector proteins were examined. Of these, we noted three interesting patterns. First, the general downregulation of a wide range of chaperone proteins. Chaperone proteins primarily function as protein folding catalyst, but many are considered virulence factors for many intracellular pathogens given that they encounter stress related to phagosome acidification and phagosome fusion with lysosomes [214]. Amongst the 16 genes annotated as chaperons in the Cb genome, 10 were transcriptionally downregulated during continuous in vitro passaging (Figure 3.7a). Notable downregulated chaperones include glutaredoxins (*grxC* and *grxD*) that have been shown to be involved in PV detoxification [215] and genes encoding heat shock proteins classes such as *dnaK, hptG, groEL* and *dnaJ* (Figure S3). These proteins are known to help bacteria adapt to stressful conditions [216, 217]. DnaK has been shown to be critical for survival of pathogenic bacteria inside the macrophage [218] and is induced in Cb in high acid condition, the condition similar to the phagolysosome [219].

The second observation is the downregulation of several genes involved in lipopolysaccharide biosynthesis. LPS layer has long been known as a pathogenic determinant and important for the host interaction in *C. burnetii* [220-222]. Out of 35 genes related to Lipopolysaccharide layer (LPS) synthesis and O-antigen nucleotide sugar biosynthesis, 11 were downregulated (Figure 3.7a, Figure S3). Out of 11 downregulated genes, 4 genes are involved in KDO2-lipid IVA Wbp pathway for LPS biosynthesis whereas 8 genes are involved in O-antigen nucleotide sugar biosynthesis. Genes involved in the first 3 steps i.e UDUDP-N-acetylglucosamine acyltransferase (*lpxA*), UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase (*lpxC*) and P-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (*lpxD*) and the gene D-glycero-D-manno-heptose 1,7-bisphosphate phosphatase (*gmhB*) in LPS biosynthesis pathway are early down or continuously down (Figure S3) (Figure 3.7C). In addition, 3 transporters related to LPS synthesis i.e., a lipoprotein releasing system ATP-binding protein (*lolD*), a lipid flippase important in cell membrane formation (*pglK*) and a probable O-antigen/lipopolysaccharide transport ATP-binding protein (*rfbE*) were also early down (Figure 3.7C). 8 downregulated genes including *wbpW, gmhB, galE, wbpD, galE, wbpI, cap1J* and *glmU* are involved in O-antigen nucleotide sugar biosynthesis pathway, a 14 genes pathway which is the first step in O-antigen biosynthesis where nucleotide sugars are assembled and activated by adding NTP [223].

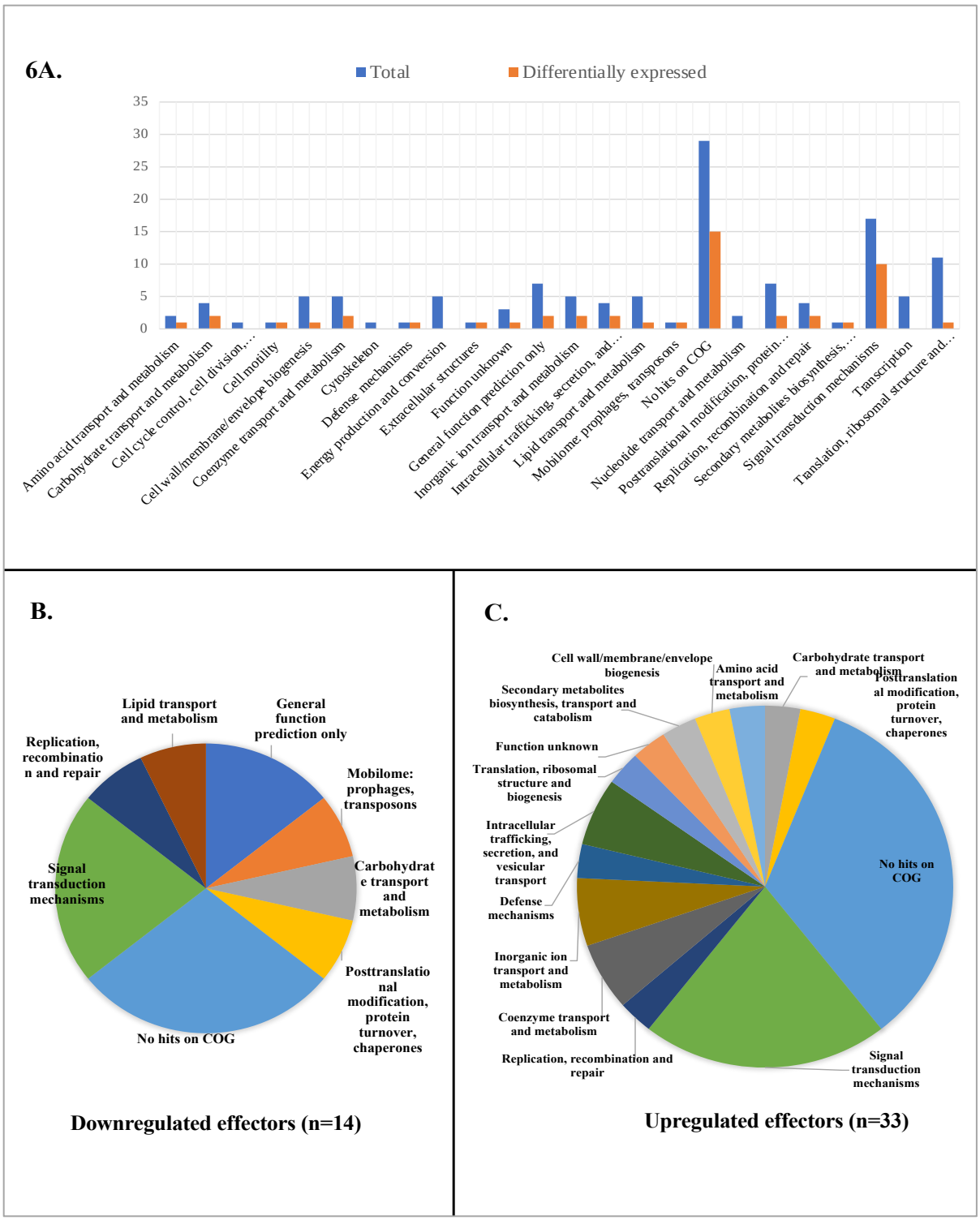


Figure 3. 6 Transcript expression changes for T4BSS effector proteins during continuous axenic passaging. A) COG classification graph for 118 identified effector proteins. The blue bars represent total effector proteins in each COG functional group (x-axis) and orange bars represents the effectors that showed differential gene expression. B) Pie chart showing COG classification for 14 downregulated effector proteins. C) Pie chart showing COG classification for 33 upregulated effector proteins.

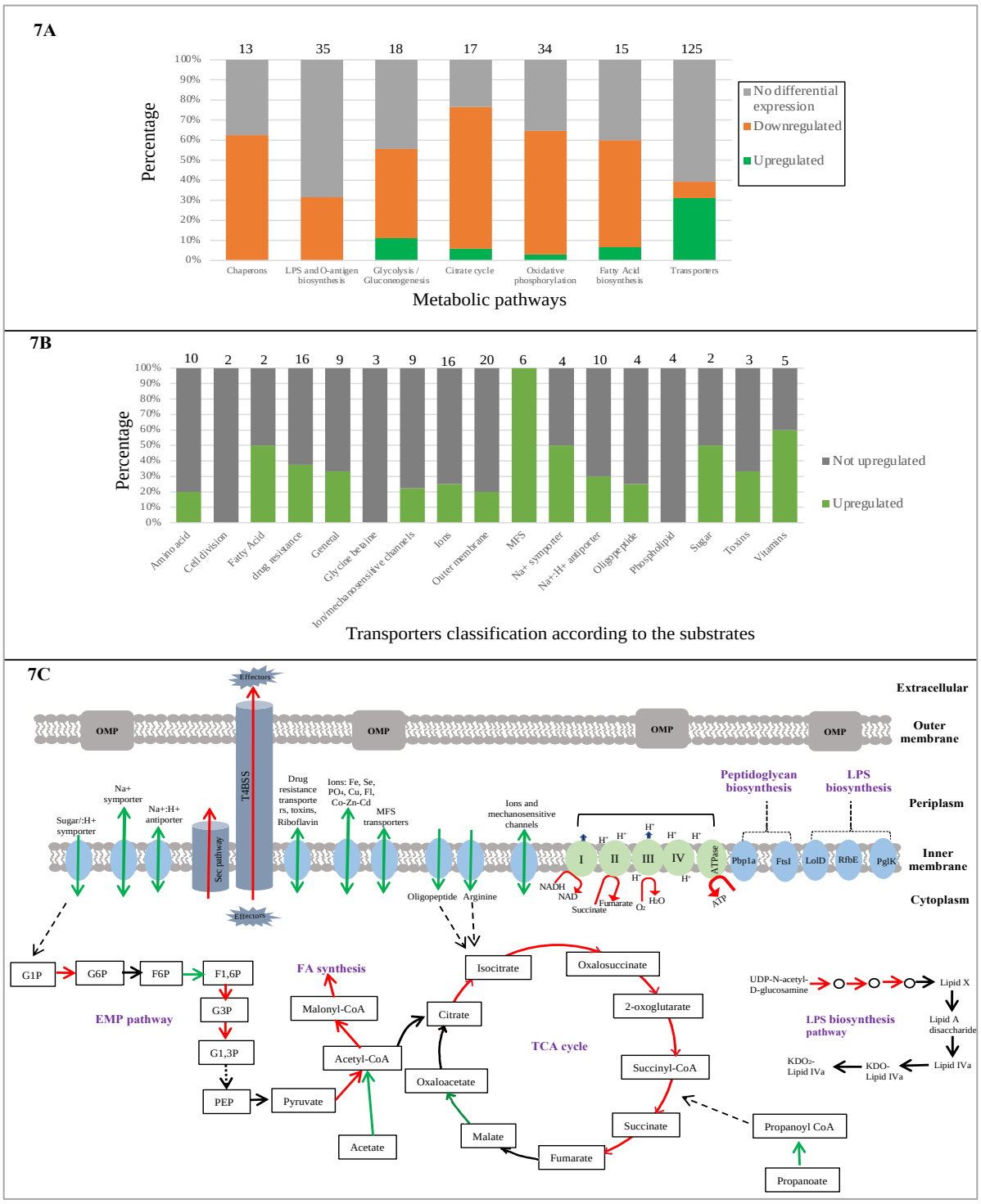


Figure 3. 7 Gene expression changes in central metabolic pathways. A) Bar graph showing numbers of DEGs in each central metabolic pathway. The numbers on top of the bars represent the total number of genes classified into that metabolic category by KEGG. B) Broad classification of differentially expressed transporters according to their substrates. The numbers on the top of the bars represents the total numbers classified as these specific transporters in Cb genome. C) Graphical representation of all metabolic pathways discussed in the manuscript text.

Finally, 3 of the 15 genes involved in peptidoglycan layer biosynthesis were downregulated. These genes are penicillin-binding protein *PBP3/ftsI*, penicillin binding protein *PBP1A/mrcA* and undecaprenyl diphosphate synthase *uppS*. The peptidoglycan layer in Cb is an immunogenicity determinant and thickens substantially during LCV to SCV transition to help in environment resistance [224, 225]. *ftsI* and *mrcA* are the only two genes encoding penicillin binding proteins in this genome that are involved in peptide cross linking. This suggests that the peptidoglycan layer may be even thinner than is in common intracellularly and could indicate a reduced requirement during axenic growth since the thick SCV peptidoglycan layer in Cb has been correlated to bacteria being more infectious [225].

3.4.6 Downregulation of multiple hypothetical proteins could suggest novel pathogenicity determinants

We posited that downregulated hypothetical proteins could represent previously unrecognized pathogenicity determinants, and that such identification could be a useful starting point for subsequent experimental validation. We identified 30 Cb genes encoding hypothetical proteins that were downregulated in more than 4 separate passages. Subcellular localization analysis software predicts that 13 are cytoplasmic, 8 inner membrane, 1 extracellular, 1 periplasmic and 7 with unknown localization (Table 3.2). Of the cytoplasmic proteins, Cig28 and an AMP binding protein (CBU_0787) possess a regulatory element recognized by PmrA, a sequence related to T4BSS expression and translocation and thus a potential predictor of effector proteins [160], although they were subsequently shown not to be translocated by the Cb T4BSS [160, 226]. Similarly, an uncharacterized protein, CBU_1234, has been shown to have a glutamate-rich C-terminal secretion signal (E-block), which is also a predictor of effector proteins [227]. Two Glycosyltransferase family 1 proteins (i.e. CBU_0839 and CBU_0841) have previously been linked to LPS mutations that lead to phase transitions [184]. The eight proteins localized in the inner membrane included Cig3, an immunoreactive peptidase CBU_0215 [227] previously shown to contain a regulatory element recognized by PmrA but not translocated by Dot/Icm system [160, 226], an immunoreactive protein CBU_1865 [228] and a DUF3971 domain-containing protein CBU_1468 that has been shown to be important for intracellular replication [165] (Table 2).

The one differentially expressed hypothetical protein identified by the localization software as an extracellular protein was CBU_0962; a predicted short chain dehydrogenase with a yet unknown specific function [229]. Lastly, proteins with unknown localization included an exported protein Cig40 [227] with a regulatory element recognized by PmrA, a hypothetical surface antigen Com1 [166] and a hypothetical protein CBUA0012 located in an ORF containing other plasmid effectors but has shown to be not secreted by Dot system [172] (Table 3.2).

Genes	Corresponding CBU (*Cb* NM I RSA493) annotation	Gene annotation	Verdict from transcriptomics	Number of passages with significant L2FC values	Subcellular localization	Protein family	More information
B7L74_04285	CBU_0841	Glycosyltransferase family 1 protein	Early down	5	Cytoplasmic	Glycos_transf_1, Glycosyl transferases group 1	This is one of the gene that accumulate mutation over a long period of culture and thus probably related to change in LPS [184].
B7L74_01595	CBU_0304	Reactive intermediate/ imine deaminase	Early down	9	Cytoplasmic	Ribonucleic-PSP, Endoribonuclease L-PSP	[230] mentions that this protein decreases after 14 days of growth in ACCM-D media.
B7L74_06350	CBU_1234	Uncharacterized protein	Early down	6	Cytoplasmic	AAA_14, AAA domain	It has been found as one of the *C. burnetii* T4BSS candidate effector proteins that possesses an E block motif (see result table [227]).
B7L74_05675	CBU_1098	Hypothetical cytosolic protein (Cig28)	Late down	4	Cytoplasmic	DUF762, *Coxiella burnetii* protein of unknown function (DUF762)	It was found to be one of the immunoreactive proteins [228]. It seems to have Pmr motif but shown to be not translocated by Dot system [160]. Has been named as Cig28.
B7L74_04275	CBU_0839	Glycosyl transferase	Early down	5	Cytoplasmic	Glycos_transf_1, Glycosyl transferases group 1	According to [184], the mutation in this protein changed the LPS from phase 1 to Phase 2.
B7L74_04000	CBU_0787	AMP-binding protein	Early down	6	Cytoplasmic	AMP-binding, AMP-binding enzyme	This protein had shown to be downregulated in the mutated PmrA regulatory sequence experiment [160]. It was suspected to be regulated by PmrA regulatory sequence but could not be confirmed if it is an effector protein. This is interesting and it might actually be an effector protein.
B7L74_03070	CBU_0591	Hypothetical cytosolic protein	Early down	10	Cytoplasmic	DUF2797, Protein of unknown function (DUF2797)	No information.
B7L74_03060	CBU_0589	Ferredoxin	Early down	7	Cytoplasmic	No hits found.	No information.
B7L74_03480	CBU_0672	Hypothetical protein	Early down	5	Cytoplasmic	No hits found.	No information.
B7L74_09270	CBU_1802	Hypothetical protein	Early down	4	Cytoplasmic	No hits found.	No information.
B7L74_06740	CBU_1308	Phosphohydrolase	Early down	4	Cytoplasmic	No hits found.	No information.
B7L74_01575	CBU_0300	YicC family protein	Early down	4	Cytoplasmic	YicC_N, YicC-like family, N-terminal region	No information on this protein in Cb. But in general, YicC protein seems to be involved in stress induced mutation in *E. coli*.
B7L74_04315	CBU_0847	3-hydroxyacyl-CoA	Early down	4	Cytoplasmic	adh_short, short chain dehydrogenase	No information.

		dehydrogenase					
B7L74_01105	CBU_0215	NlpC-P60 family protein. Peptidase	Early down	7	Cytoplasmic Membrane	SH3_6, SH3 domain (SH3b1 type)	It is a peptidase and one of the immunoreactive protein found in [229]. It is predicted to be regulated by PmrA sequences [227]. PmrA has been associated with the effector proteins.
B7L74_07575	CBU_1468	DUF3971 domain-containing protein	Early down	4	Cytoplasmic Membrane	DUF3971, Protein of unknown function	The transposon mutation in this gene caused moderate intracellular replication [165]. So, it might be one of the important gene.
B7L74_09610	CBU_1865	Hypothetical membrane associated protein	Early down	7	Cytoplasmic Membrane	bPH_2, Bacterial PH domain	In [228] it was found to be one of the immunoreactive proteins. This protein is applicable as candidate vaccines.
B7L74_00420	CBU_0084	K07014; Uncharacterized protein (Cig3)	Early down	5	Cytoplasmic Membrane	DUF3413, Domain of unknown function (DUF3413)	It is also one of the proteins that are affected by Pmr mutation but is not translocated by T4SS [160]. Might be an effector too. It also has a name Phosphoglycerol transferase MdoB, Cig3.
B7L74_03075	CBU_0592	Hypothetical membrane spanning protein	Continuous down	7	Cytoplasmic Membrane	No hits found.	No information.
B7L74_10740	CBU_2073	Acyltransferase	Early down	4	Cytoplasmic Membrane	Acyltransferase, Acyltransferase	No information.
B7L74_10350	CBU_2001	Hypothetical protein	Continuous down	4	Cytoplasmic Membrane	No hits found.	No information.
B7L74_01205	CBU_0230	Hypothetical protein	Early down	4	Cytoplasmic Membrane	No hits found.	No information.
B7L74_04930	CBU_0962	Short-chain dehydrogenase	Continuous down	4	Extracellular	adh_short, short chain dehydrogenase	This protein have been found in immunoprecipitation methods in [231].
B7L74_07865	CBU_1529	Peptidase	Early down	4	Periplasmic	Peptidase_S9, Prolyl oligopeptidase family	No information.
B7L74_10930	CBUA0012	Hypothetical protein	Early down	5	Unknown	DUF807, *Coxiella burnetii* protein of unknown function (DUF807)	It is found in same ORFs (with other 5 hypothetical proteins) as other plasmid effectors (which are translocated by Dot system). But this protein is not translocated Dot/Icm [172]
B7L74_07040	CBU_1366	Hypothetical exported protein (Cig40)	Early down	6	Unknown	No hits found.	This coiled-coil domain containing protein is regulated by PmrA in *L. pneumophila* [160, 226] but is not secreted outside *C.burnetii [232]*. It might be somehow working as an effector protein.
B7L74_09850	CBU_1910	Hypothetical outer membrane protein surface	Early down	7	Unknown	DSBA, DSBA-like thioredoxin domain	This protein is mentioned in these two papers [166, 169] but there is no mention of function. It is used as loading control western blot. Surface antigens in general helps bacteria to evade

		antigen Com1					into the epithelial cells (google). So, it might be important for pathogenesis.
B7L74_01355		Hypothetical protein	Continuous down	8	Unknown	No hits found.	No information.
B7L74_09265	CBU_1801	Hypothetical protein	Early down	5	Unknown	No hits found.	Doesn't have much information on except that they saw a lot of mutati operon [14].
B7L74_09795		DUF1658 domain-containing protein	Early down	4	Unknown	No hits found.	No information.
B7L74_09830	CBU_1906	Hypothetical protein	Early down	4	Unknown	No hits found.	No information.

Table 3. 2 Detailed information on 30 hypothetical proteins that are downregulated and could be potential effector proteins. The subcellular localizations of these proteins were identified using Psortb. The protein family of the genes were identified using pfam classification and only top hit is listed here. More information on these genes were mined from published and unpublished research articles with references mentioned in last column.

3.4.7 Transcriptional patterns of central metabolic pathways

Analysis of gene expression patterns of central metabolic pathways demonstrated a general trend of down-regulation in genes encoding enzymes in central catabolic, amphibolic, and anabolic pathways, coupled with a broad upregulation in genes encoding transporters. An overall pattern of downregulation of glycolysis genes (8/18 genes) was observed (Figure 3.7A), with several enzymes such as pyruvate dehydrogenases (*pdhC, pdhD*), fructose-bisphosphate aldolase (*fbaA*), glyceraldehyde 3-phosphate dehydrogenase (*gapA*), and phosphoenolpyruvate carboxykinase (*pckA*) downregulated early in the passaging (Figure S3). Gene *pmm-pgm*, which encodes the enzyme phosphomannomutase/phosphoglucomutase and is involved in the first step of glycolysis, was down early as well (Figure S3). Similarly, analysis of genes encoding enzymes of the tricarboxylic acid (TCA) cycle, demonstrated an overall downregulation (12/17 genes), with a downregulation of ~2 fold in isocitrate dehydrogenase (IDH2), the rate-limiting enzyme in the TCA cycle (Figure S3). As well, the downregulations in genes encoding multiple carbohydrate dehydrogenases (e.g., pyruvate dehydrogenases *pdhC* and *pdhD*), succinate dehydrogenases (*sdhA*, *sdhB*, *sdhD*), as well as genes that are necessary for oxidation of glycolytic and TCA cycle sugar intermediates was observed. Finally, genes encoding components of the electron transport chain (ETC) were also downregulated. These include Complex I: NADH-quinone oxidoreductase (*nuoB, nuoD, nuoE, nuoF, nuoH, nuoI, nuoK, nuoL, nuoM, nuoN*), Complex II:Succinate dehydrogenase (*sdhA, sdhB, sdhD*), Complex III:Cytochrome oxidoreductase (*cyoC, cyoD, cydA, cydB, cydX*), and Complex V:F-type ATPase (*atpA, atpB, atpE, atpD, atpF, atpG*) (Figure 3.7A, Figure S3). The downregulations in two complexes (*cyoC and cyoD*) of cytochrome o oxidase, which is known to be induced in oxygen rich growth conditions in bacteria, [233] suggests a decreased affinity and/or competition for oxygen in the cell-free growth environment, as previously suggested [234]. Cytochrome d oxidase, which is shown to be expressed more in oxidative and nitrosative stress conditions [233], also has expression of two of its components (i.e., *cydA* and *cydX*) early down and late down, respectively. Finally, an overall downregulation of fatty acids biosynthesis genes transcription (8/15) was also observed (Figure 3.7A, Figure S3).

In contrast to the general trend of downregulation of the central metabolic machinery of Cb, a marked upregulation of genes encoding transporters was observed. Out of 125 general transporters, the transcription of 39 were upregulated and 10 were downregulated (Figure 3.7a). Transporters that were upregulated have double fold expression change (L_2fc) ranging from 3-10 (Figure S3). Upregulated primary transporters included transporters for amino acid arginine, oligopeptides, fatty acids, and vitamins such as riboflavin and thiamin (Figure 3.7b) as well as a small number of transporters (4 out of 20 present) related to synthesis and maintenance of outer membrane. On the other hand, upregulated secondary transporters included MFS transporters, symporters, antiporters, and mechanosensitive ion channels. Of these, a notable observation was made where all six MFS transporters and two out of four Na^+ symporters found in Cb genome were found to be early upregulated. These MFS transporters transports various compounds such as monosaccharides, oligosaccharides, amino acids, peptides, vitamins, cofactors, drugs, nucleobases, nucleosides, and organic and inorganic anions and cations. In addition, a large proportion of transporters related to drug resistance (6 /16 present in the genome) and ion transporters mediating the uptake of ions such as copper, iron, fluoride, selenite, cobalt-cadmium-zinc, and phosphate were also upregulated (Figure 3.7b). In addition to transporters mediating substrate transport, transporters involved in pH homeostasis such as ions/mechanosensitive channels and Na^+:H^+ antiporter were also upregulated (Figure 3.7b). Na^+:H^+ antiporter functions to utilizes the proton motive force to efflux intracellular sodium ions for intracellular pH homeostasis [235] and these antiporters along with ion/mechanosensitive channels have been proposed to play an important role in pH homeostasis and survival within the acidic PL [236]. Lastly, 3 out of 9 transporters classified under general or unknown functions were upregulated as well (Figure 3.7b).

3.4.8 Genomics reveals a stable Cb genome

For all 12 passages, genomes with 100% completeness (assessed by identifying all 265 housekeeping marker genes specific for the Proteobacteria [21]) were obtained. N50 of genomic assemblies ranged between 49,903 and 75,629, N90 ranged between of 15,966 and 20,406, and the number of contigs per genome ranged between 56 and 64 (Table 3.3). Using Passage 01 as a reference, we identified 842 unique single nucleotide polymorphisms (SNPs) and 118 unique deletions/insertion polymorphisms (DIPs) (Figure 3.8A).

Of 842 unique SNPs, only 9 were identified in consensus mode i.e., present in 100% of population in one or more passages while the remaining 833 SNPs were identified in population mode, i.e., occurring in a fraction of the community when sequenced (Figure 3.8). Further only 69 unique SNPs were identified to occur in all i.e., passage 3-61, and only 43 SNPs were maintained in later passages (Figure 3.8B). More importantly, only one consensus mutation occurred in a gene that was downregulated in transcriptomic analysis. This gene GTP pyrophosphokinase *spoT* (B7L74_01590) had a one amino acid (aa) substitution (T to A) at position 262 which propagates to 100% population in last 9 passages and is also noticeably early down in gene expression. SpoT is a signal transduction component and transcriptional regulator with a role in helping *Coxiella* cope with the low-nutrient and high stress condition [237].

Passage	Completeness	Contamination	Number of contigs	N50	L50	N90	Total length
Passage 10	100	0	64	49903	11	15966	1975973
Passage 21	100	0	64	75629	9	16820	1980957
Passage 61	100	0	63	63550	11	16844	1975487
Passage 51	100	0	63	51753	10	18488	1981816
Passage 3	100	0	58	66036	9	19293	1975591
Passage 42	100	0	60	66036	10	19293	1975994
Passage 67	100	0	59	66036	10	19293	1976118
Passage 16	100	0	56	71520	9	19389	1976105
Passage 1	100	0	59	69758	9	20406	1975592
Passage 5	100	0	57	69778	9	20406	1976018
Passage 13	100	0	59	71520	9	20406	1976303
Passage 31	100	0	57	71520	9	20406	1976112

Table 3. 3 Sequencing and assembly statistics for genomes of all passages of *Coxiella burnetii* in this study.

For the 118 unique DIPs, only 3 DIPs were identified in consensus mode and 115 in population mode (Figure 3.8). Lengths of insertions and deletions were always very minor with 93% of DIPs representing an insertions or deletions of a single bp. The multi base pair deletions included deletions of 2, 3, 7 and 12 bp that occurred in coding region and the longest 32 base pair deletion occurring in intergenic region. However, none of these genes were affected by deletion since they had no significant transcriptomic changes. Of the genes that were downregulated, 7 had DIPs mutations but all of them being in only a fraction of populations (mostly 5-10% populations). Collectively, the low level and lack of possible effect, suggest a very stable and minor level in genomic mutation in modulating transcriptional levels.

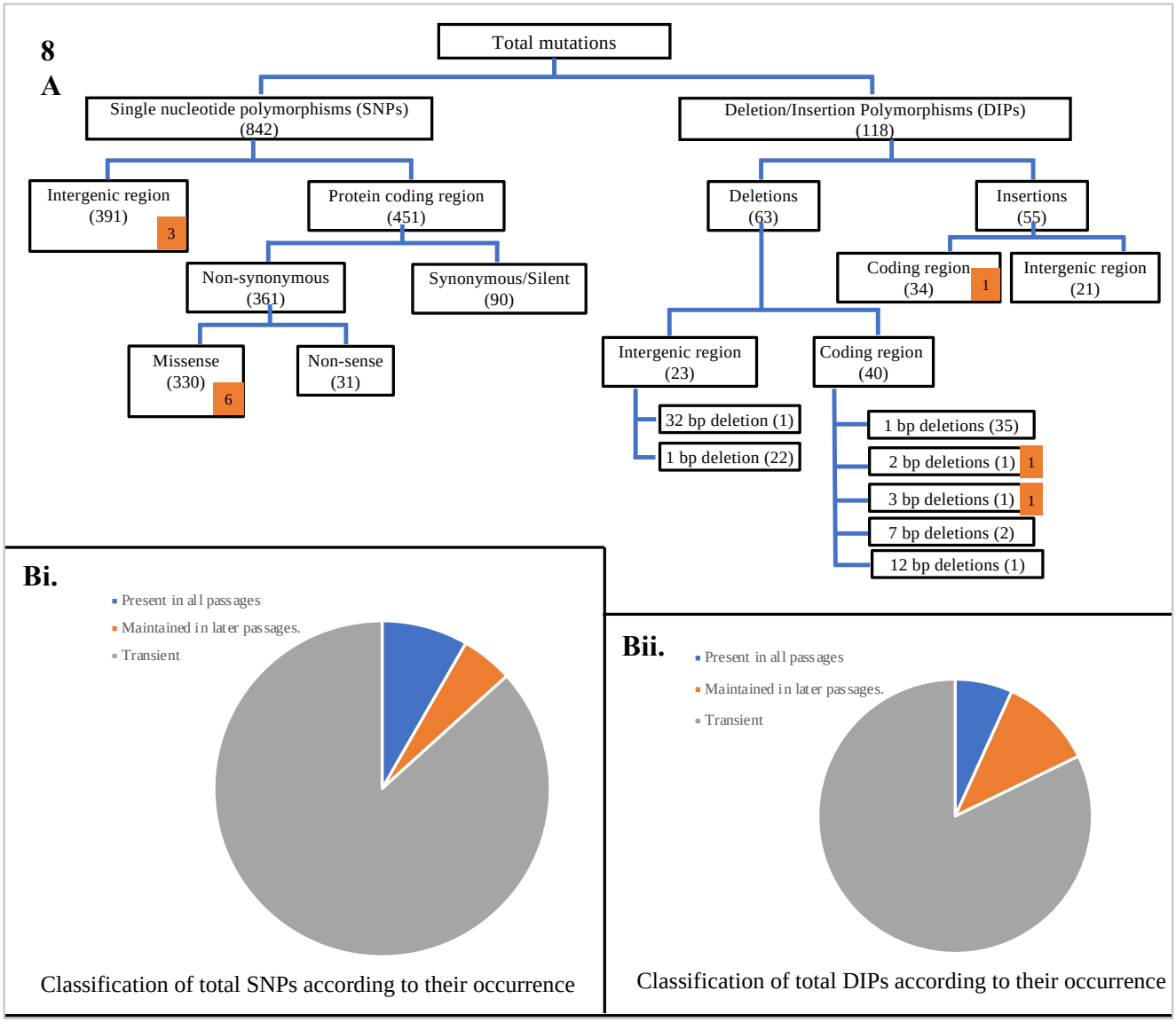


Figure 3. 8 A) Flowchart for classification of different types of Single nucleotide polymorphisms (SNPs) and Deletion/Insertion polymorphisms (DIPs) found in this experiment. The numbers at the bottom right corners that are highlighted in orange represents the number of genes that had mutation in all 100% of the populations in a passage. Bi) Classification of SNPs according to its occurrence in number of passages. Bii) Classification of DIPs according to the occurrence in number of passages.

3.5 Discussion

Here, we attempted to identify genes and proteins crucial to *Coxiella burnetii* intracellular growth lifestyle using a "reverse evolution" approach. We transitioned Cb Nine Mile phase II from cell cultures into axenic defined media ACCM-D and subcultured it into a long-term successive passage. Temporal changes in gene expression patterns, and DNA mutations associated with adaptation to an axenic extracellular lifestyle were identified. In general, we observe a significant number of differential expression (464 up, 371 down, 38% of overall Cb genes) through 67 passages. It is interesting to note that the majority (288 upregulated and 249 downregulated) of differentially expressed genes expressed an "early up" or "early down" expression pattern (Figure 3.3), suggestive of a relatively rapid adaptation (within 31 passages out of 61 total passages) into this new axenic environment.

Differentially expressed genes identified in this study could be grouped into multiple structural and functional categories (secretory apparatus, effector proteins, other pathogenicity determinants, hypothetical proteins, and central metabolic pathways). In general, a broad (19/24 genes encoding secretory T4BSS components showed significant expression change with 18 genes showing decrease in the expression whereas only 1 gene that was upregulated. T4BSS is the most crucial conduit for pathogenicity and effector proteins in Cb [151, 158, 167]. Components of T4BSS span both membranes and periplasm and are bridged by the core transport complex comprising proteins DotC, DotD, DotF, DotG and DotH, which provides a channel for export of substrates (Figure 3.5a) [238]. The coupling protein complex provides a link between substrates and transport complex and includes DotL, DotM, DotN, IcmS and IcmW [239]. DotB is an essential cytoplasmic protein with an ATPase activity and unknown function, but its mutation has been linked to failure in secreting effector proteins to infect host cells [187]. DotA and IcmX has been shown to be secreted out from cell [190]. Besides these, other components of T4BSS includes DotO localized in cytoplasm, IcmX in periplasmic space, DotK in outer membrane whereas IcmF, IcmH, DotI, DotJ, DotA, DotE, DotP, IcmV and IcmT in inner membrane. (Figure 3.5a). The genes involved in T4BSS in Cb are clustered in a single locus except the genes *icmF* and *dotU* which were a part of a separate operon (Figure 3.5b). This is similar to the gene rearrangement shown in [236]. Gene *icmF*, which has been shown to be involved in intra-macrophage replication and inhibition of phagosome-lysosome fusion in *L. pneumophila* [240, 241] and stabilization of the secretion complex [242] was the only T4BSS component that showed upregulation.

The observed downregulation of this experimentally verified central pathogenic determinant is hence well justified and provide a general overall credence and justification that gene downregulation under the experimental setting employed could be regarded as a reasonable proxy for requirement for intracellular survival in cell-cultures. In addition to T4BSS, other secretory pathway such as general secretory (sec) pathway and a component of type I secretary pathway i.e., TolC also exhibited a general trend of overall downregulation (Figure 3.5A, Figure S1a, Figure S1b). In general, we interpret such overall lower expression of structural secretory apparatuses as a reflection of the lower need for interaction between Cb and the outside environment in an axenic setting when compared to an intracellular setting.

Interestingly, while genes encoding the production of secretary pathways were downregulated, expression patterns of effector proteins were mixed, with 33 upregulated and 14 downregulated. Of the 14 downregulated genes (all of which were early downregulated), nine have been experimentally verified based on experimental evidence of their translocation by the Dot/Icm system [166, 167, 169, 172, 209, 210, 213], three genes

containing ankyrin repeat domains (*ankG, ankD* and *ankK*) were considered effectors based on the presence of eukaryotic like domains and subsequently shown to be translocated by Dot/Icm system[210], gene *cirC* (*Coxiella* effector for intracellular replication) was verified as effector by transposon insertion mutation studies where its mutation was associated with defect in *Coxiella* containing vacuole (CCV) biogenesis [169] and lastly a hypothetical protein B7L74_09020 was verified to be an effector based on loss-of-function mutation where its mutation was related to smaller CCV phenotype [164](Table 1).

AnkG, AnkD and AnkK are ankyrin repeat-containing effector proteins in Cb [236]. The eukaryotic type Ank domain in this protein family might have a role in host-cell attachment and allows the interaction of bacteria with a spectrum of host cell proteins and thus are particularly important in the pathogenic process [210, 243-245] AnkD has both eukaryotic like domain and F-box domain, but the function is not yet clear [210]. AnkG has been shown to localize at the host microtubules and interferes with host apoptosis pathway by interacting with the host protein gC1qR (p32) [210, 211]. AnkK has been shown to have important role for the bacterial growth inside macrophages [212] although it is not delivered to the host cell via T4BSS [210](Table 1). *Coxiella* plasmid effector proteins (CpeF and CpeH) are the two proteins in plasmid effector family proteins, the family important for disrupting host cell mechanisms. CpeF doesn't have any specialized localization in host cell but has shown to cause growth defect when mutated [168, 172] whereas CpeH localizes in host cell's cytoplasm [209]. Cig57 mutation has been linked to intracellular replication defect whereas Cig2 mutation causes growth defect and CCV fusion defect [165]. These two proteins are early downregulated in 9 and 7 passages respectively. CirC has been shown to be important for CCV biogenesis [169] and it is early downregulated in 4 passages. The remaining 6 downregulated effectors are hypothetical proteins with unknown functions, except Cbu1752 which has been shown to be important for vacuole biogenesis and Cbu0635 important for host cell secretion (Table 1).

Of the 33 upregulated effector proteins, the majority fall into COG categories of unclassified (n=11), signal transduction mechanisms (n=7), transportation and metabolism of coenzyme and inorganic ions (n=6) (Figure 3.6). These upregulated, and no expression change, effector proteins (n=71) could also be involved in mediating general survival functions or cellular metabolisms besides their involvement in maneuvering pathogenesis process. This could be one of the explanations behind their upregulation, or no expression change in this particular setting.

Such pattern where genes encoding the formation of the structural conduits (i.e., secretary pathways) are downregulated, but numerous genes encoding proteins secreted through these conduits (i.e., effector proteins) are upregulated is puzzling. We put forth the possibility that the expression of these effector proteins is controlled by Cb intracellular conditions, where high concentrations of intracellular metabolites (amino acids, inorganic salts, ATP/ADP ratio) regulate their expression. Under this scenario, high level of intracellular precursors in Cb is associated with growth inside the cell, as opposed to the more stressed, more starved condition within the extracellular small cell variant (SCV) conditions. It remains to be seen whether translation of these effector transcripts to protein products and subsequent secretion occurs in Cb grown in axenic media. To the best of our knowledge, no prior study have demonstrated the secretion of effector proteins in Cb axenic media.

Multiple additional pathogenic determinants were also downregulated in axenic media. Specifically, chaperons, LPS, and peptidoglycan synthesis. Amongst the 16 genes annotated as chaperons in the Cb genome, 10 were transcriptionally downregulated starting at early passages (Figure 3.7a). The

downregulation could be explained by the fact that chaperons play important roles for withstanding stress associated with intracellular survival e.g., PV detoxification [215], survival inside the macrophage [218] and high acidity in the phagolysosome [219].

Analysis of central metabolic pathways showed a clear trend of downregulation of multiple catabolic (e.g., glycolysis and electron transport chain), amphibolic (e.g., citric acid cycle) and anabolic (e.g., FA synthesis) pathways, with a parallel upregulation of genes encoding transporters (Figure 3.7C). Such pattern could readily be explained by the nutrient rich growth environment (ACCM-D media) where Cb is grown. This media is a defined axenic medium for CB growth and contains all 20 amino acids, salts (sodium phosphate and sodium bicarbonate), vitamins, minerals, and trace elements [185]. As such, the need for expression of genes encoding enzymes that are components of these biosynthetic pathways decreases and subsequently, the overall need for ATP generation for biosynthetic purposes (hence decrease in respiratory activity). Finally, 30 hypothetical proteins were downregulated, and their predicted localization, predicted role in pathogenesis and their general analysis provide some possible explanations for such pattern (see results section). Regardless, we suggest that these could be important, hitherto untested possible pathogenicity determinants. Future biochemical and genetic efforts to test such assumptions is certainly needed.

In conclusion, we present a detailed temporal analysis on how Cb transition from intracellular growth, (where a wide range of cellular processes is required to maintain survival and growth), to a defined, rich axenic media (where many of such processes are theoretically, no longer needed). As any genome-wide transcriptomics survey, the approach is useful for uncovering patterns, confirming prior observations, and generating new insights and hypothesis. We stress that while downregulation in axenic media compared to cell culture could broadly be associated with importance for survival in cell cultures, the precise nature of such correlation is yet unclear, and that differential expression patterns could further be modified on the translational and post translational levels. Experimental assessments and validation of many of the observed patterns is certainly needed. Nevertheless, our analysis could be extremely beneficial to provide information on how specific genes and pathways in Cb could be important for its intracellular survival, as well as to identify putatively novel pathogenicity determinants in Cb.

3.7 Supplemental materials

The supplement figures Figure S1, Figure S2 and Figure S3 are provided in the appendices.

REFERENCES

1. Imelfort, M., et al., *GroopM: an automated tool for the recovery of population genomes from related metagenomes.* PeerJ., 2014. **2**: p. e603.
2. Rinke, C., et al., *Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics.* Nat Protoc., 2014. **9**(5): p. 1038-48.
3. Wu, Y.W., B.A. Simmons, and S.W. Singer, *MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets,.* Bioinformatics, 2016. **32**: p. 605-607.
4. Anantharaman, K., et al., *Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system.* Nature Comm, 2016. **7**: p. 13219.
5. Parks, D.H., et al., *Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life.* Nature microbiology, 2017. **2**(11): p. 1533-1542.
6. Sharon, I. and J.F. Banfield, *Genomes from metagenomics.* Science, 2013. **342**: p. 1057-1058.
7. Jousset, A., et al., *Where less may be more: how the rare biosphere pulls ecosystems strings.* ISME J., 2017. **11**: p. 853-862.
8. Lynch, M.D., A.K. Bartram, and J.D. Neufeld, *Targeted recovery of novel phylogenetic diversity from nextgeneration sequence data.* ISME J., 2012. **6**: p. 2067–2077.
9. Lynch, M.D. and J.D. Neufeld, *Ecology and exploration of the rare biosphere.* Nature Rev. Microbiol., 2015. **13**: p. 217-229.
10. Youssef, N., B.L. Steidley, and M.S. Elshahed, *Novel High-Rank Phylogenetic Lineages within a Sulfur Spring (Zodletone Spring, Oklahoma), Revealed Using a Combined Pyrosequencing-Sanger Approach.* Applied and Environmental Microbiology, 2012. **78**(8): p. 2677-2688.
11. Pannekens, M., et al., *Oil reservoirs, an exceptional habitat for microorganisms.* New Biotechnology, 2019. **49**: p. 1-9.
12. Youssef, N., M.S. Elshahed, and M.J. McInerney, *Chapter 6 Microbial Processes in Oil Fields: Culprits, Problems, and Opportunities*, in *Advances in Applied Microbiology*. 2009, Academic Press. p. 141-251.
13. Struchtemeyer, C.G., *Microbiology of Oil- and Natural Gas-Producing Shale Formations: An Overview*, in *Consequences of Microbial Interactions with Hydrocarbons, Oils, and Lipids: Biodegradation and Bioremediation*, R. Steffan, Editor. 2018, Springer International Publishing: Cham. p. 1-18.
14. Booker, A.E., et al., *Draft genome sequences of multiple Frackibacter strains isolated from hydraulically fractured shale environments.* Genome Announc., 2017. **5**: p. e00608.
15. Vigneron, A., et al., *Succession in the petroleum reservoir microbiome through an oil field production lifecycle.* ISME J, 2017. **11**: p. 2141-2154.
16. Borton, M.A., et al., *Coupled laboratory and field investigations resolve microbial interactions that underpin persistence in hydraulically fractured shales.* Proceedings of the National Academy of Sciences., 2018. **115**(28): p. E6585-E6594.
17. Vilcaez, J., et al., *Stimulation of methanogenic crude oil biodegradation in depleted oil T reservoirs.* Fuel, 2018. **232**: p. 581-590.
18. Berdugo-Clavijo, C. and L.M. Gieg, *Conversion of crude oil to methane by a microbial consortium enriched from oil reservoir production waters.* Front Microbiol., 2014. **5**: p. 197.
19. Gieg, L.M., K.E. Duncan, and J.M. Suflita, *Bioenergy production via microbial conversion of residual oil to natural gas.* Appl. Environ. Microbiol., 2008. **74**: p. 3022-3029.
20. D. Li, C.M.L., R. Luo, K. Sadakane, T.W. Lam,, *MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph.* Bioinformatics, 2015. **31**(10): p. 1674-1676.

21. D.H. Parks, M.I., C.T. Skennerton, P. Hugenholtz, G.W. Tyson, *CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes.* Genome Res., 2015. **25**: p. 1043-1055.
22. Pruesse, E., J. Peplies, and F.O. Glöckner, *SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes.* Bioinformatics, 2012. **28**(14): p. 1823-1829.
23. Stamatakis, A., *RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.* Bioinformatics, 2014. **30**(9): p. 1312-1313.
24. Parks, D.H., et al., *A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life.* Nat Biotechnol., 2018. **36**: p. 996-1004.
25. Rodriguez-R, L.M. and K.T. Konstantinidis, *The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes*. 2016, PeerJ Preprints.
26. Camacho, C., et al., *BLAST+: architecture and applications.* BMC Bioinformatics, 2009. **10**(1): p. 421.
27. Yu, N.Y., et al., *PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes.* Bioinformatics, 2010. **26**(13): p. 1608-1615.
28. Thompson, L.R. and e. al., *A communal catalogue reveals Earth's multiscale microbial diversity.* Nature, 2017. **551**: p. 457-463.
29. Coveley, S., M.S. Elshahed, and N.H. Youssef, *Response of the rare biosphere to environmental stressors in a highly diverse ecosystem (Zodletone spring, OK, USA).* PeerJ., 2015. **3**: p. e1182.
30. Schloss, P.D., et al., *Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities.* Appl Environ Microbiol., 2009. **75**(23): p. 7537-41.
31. Fiala, G., et al., *Flexistipes sinusarabici, a novel genus and species of eubacteria occurring in the Atlantis II Deep brines of the Red Sea.* Arch. Microbiol., 1990. **154**: p. 120-126.
32. L'Haridon, S., et al., *Kosmotoga pacifica sp. nov., a thermophilic chemoorganoheterotrophic bacterium isolated from an East Pacific hydrothermal sediment.* Extremophiles, 2014. **18**(1): p. 81-88.
33. Marone, A., et al., *Bioelectrochemical treatment of table olive brine processing wastewater for biogas production and phenolic compounds removal.* Water Res., 2016. **100**: p. 316-325.
34. Westphal, A., et al., *Effects of plant downtime on the microbial community composition in the highly saline brine of a geothermal plant in the North German Basin.* Appl Microbiol Biotechnol, 2016. **100**(7): p. 3277-90.
35. Lapidus, A., et al., *Genome sequence of the moderately thermophilic halophile Flexistipes sinusarabici strain (MAS10).* Standards in genomic sciences, 2011. **5**(1): p. 86-96.
36. DiPippo, J.L., et al., *Kosmotoga olearia gen. nov., sp. nov., a thermophilic, anaerobic heterotroph isolated from an oil production fluid.* Int. J. Syst. Evol. Microbiol., 2009. **59**(12): p. 2991-3000.
37. Nunoura, T., et al., *Kosmotoga arenicorallina sp. nov. a thermophilic and obligately anaerobic heterotroph isolated from a shallow hydrothermal system occurring within a coral reef, southern part of the Yaeyama Archipelago, Japan, reclassification of Thermococcoides shengliensis as Kosmotoga shengliensis comb. nov., and emended description of the genus Kosmotoga.* Archives of Microbiology, 2010. **192**(10): p. 811-819.
38. Grant, W.D., *Life at low water activity.* Philos Trans Roy Soc B 2004. **359**: p. 1249–1267.
39. Youssef, N.H., et al., *Trehalose/2-sulfotrehalose biosynthesis and glycine-betaine uptake are widely spread mechanisms for osmoadaptation in the Halobacteriales.* The ISME J., 2014. **8**: p. 636-649.
40. Thauer, R.K., K. Jungermann, and K. Decker, *Energy conservation in chemotrophic anaerobic bacteria.* Bacteriol Rev, 1977. **41**(1): p. 100-80.
41. Worm, P., et al., *A genomic view on syntrophic versus non-syntrophic lifestyle in anaerobic fatty acid degrading communities.* Biochimica et Biophysica Acta (BBA) - Bioenergetics, 2014. **1837**(12): p. 2004-2016.
42. Kouzuma, A., S. Kato, and K. Watanabe, *Microbial interspecies interactions: recent findings in syntrophic consortia.* Front Microbiol., 2015. **6**: p. 477.

43. Pirbadian, S., et al., *Shewanella oneidensis MR-1 nanowires are outer membrane and periplasmic extensions of the extracellular electron transport components.* Proceedings of the National Academy of Sciences, 2014. **111**(35): p. 12883-12888.
44. Fardeau, M.L., et al., *Thermoanaerobacter subterraneus sp. nov., a novel thermophile isolated from oilfield water.* Int. J. Syst. Evol. Microbiol., 2000. **50**: p. 2141-2149.
45. Takahata, Y., et al., *Thermotoga petrophila sp. nov. and Thermotoga naphthophila sp. nov., two hyperthermophilic bacteria from the Kubiki oil reservoir in Niigata, Japan.* Int J Syst Evol Microbiol, 2001. **51**: p. 1901-1909.
46. Cayol, J.L., et al., *Thermohalobacter berrensis gen. nov., sp. nov., a thermophilic, strictly halophilic bacterium from a solar saltern.* Int J Syst Evol Microbiol., 2000. **50**: p. 2559-2564.
47. Cayol, J.L., et al., *Isolation and characterization of Halothermothrix orenii gen. nov., sp. nov., a halophilic, thermophilic, fermentative, strictly anaerobic bacterium.* Int J Syst Evol Microbiol., 1994. **44**: p. 534-540.
48. Sogin, M.L., et al., *Microbial diversity in the deep sea and the underexplored "rare biosphere"*. Proc. Nat. Acad. Sci. USA, 2006. **103**: p. 12115-12120.
49. Jørgensen, B.B., *Deep subseafloor microbial cells on physiological standby.* Proc Natl Acad Sci., 2011. **108**: p. 18193-18194.
50. Morono, Y., et al., *Carbon and nitrogen assimilation in deep subseafloor microbial cells.* Proc Natl Acad Sci., 2011. **108**: p. 18295–18300.
51. Doud, D.F., et al., *Function-driven single-cell genomics uncovers cellulose-degrading bacteria from the rare biosphere.* The ISME journal, 2020. **14**(3): p. 659-675.
52. Anantharaman, K., et al., *Expanded diversity of microbial groups that shape the dissimilatory sulfur cycle.* The ISME Journal, 2018. **12**(7): p. 1715-1728.
53. Becraft, E.D., et al., *Rokubacteria: genomic giants among the uncultured bacterial phyla.* Frontiers in microbiology, 2017. **8**: p. 2264.
54. Rinke, C., et al., *A phylogenomic and ecological analysis of the globally abundant Marine Group II archaea (Ca. Poseidoniales ord. nov.).* The ISME journal, 2019. **13**(3): p. 663-675.
55. Farag, I.F., et al., *Global patterns of abundance, diversity and community structure of the Aminicenantes (candidate phylum OP8).* PloS one, 2014. **9**(3): p. e92139.
56. Hu, P., et al., *Simulation of Deepwater Horizon oil plume reveals substrate specialization within a complex community of hydrocarbon degraders.* Proceedings of the National Academy of Sciences, 2017. **114**(28): p. 7432-7437.
57. Zhou, Z., et al., *Genome diversification in globally distributed novel marine Proteobacteria is linked to environmental adaptation.* The ISME journal, 2020. **14**(8): p. 2060-2077.
58. Rinke, C., et al., *Insights into the phylogeny and coding potential of microbial dark matter.* Nature, 2013. **499**(7459): p. 431-437.
59. Beam, J.P., et al., *Ancestral absence of electron transport chains in Patescibacteria and DPANN.* Frontiers in microbiology, 2020: p. 1848.
60. Parks, D.H., et al., *A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life.* Nature Biotechnology, 2018. **36**(10): p. 996-1004.
61. Kielak, A., et al., *Phylogenetic diversity of Acidobacteria in a former agricultural soil.* Isme j, 2009. **3**(3): p. 378-82.
62. Kielak, A.M., et al., *The Ecology of Acidobacteria: Moving beyond Genes and Genomes.* Front Microbiol, 2016. **7**: p. 744.
63. Quaiser, A., et al., *Acidobacteria form a coherent but highly diverse group within the bacterial domain: evidence from environmental genomics.* Mol Microbiol, 2003. **50**(2): p. 563-75.
64. Fierer, N., et al., *Assessment of soil microbial community structure by use of taxon-specific quantitative PCR assays.* Appl Environ Microbiol, 2005. **71**(7): p. 4117-20.
65. Kim, J.-S., et al., *Bacterial diversity of terra preta and pristine forest soil from the Western Amazon.* Soil Biology and Biochemistry, 2007. **39**(2): p. 684-690.

66. Zhang, Y., et al., *Community structure and elevational diversity patterns of soil Acidobacteria.* J Environ Sci (China), 2014. **26**(8): p. 1717-24.
67. Kishimoto, N., Y. Kosako, and T. Tano, *Acidobacterium capsulatum gen. nov., sp. nov.: An acidophilic chemoorganotrophic bacterium containing menaquinone from acidic mineral environment.* Current Microbiology, 1991. **22**(1): p. 1-7.
68. Liesack, W., et al., *Holophaga foetida gen. nov., sp. nov., a new, homoacetogenic bacterium degrading methoxylated aromatic compounds.* Arch Microbiol, 1994. **162**(1-2): p. 85-90.
69. Coates, J.D., et al., *Geothrix fermentans gen. nov., sp. nov., a novel Fe(III)-reducing bacterium from a hydrocarbon-contaminated aquifer.* Int J Syst Bacteriol, 1999. **49 Pt 4**: p. 1615-22.
70. Borneman, J. and E.W. Triplett, *Molecular microbial diversity in soils from eastern Amazonia: evidence for unusual microorganisms and microbial population shifts associated with deforestation.* Applied and environmental microbiology, 1997. **63**(7): p. 2647-2653.
71. Barns, S.M., S.L. Takala, and C.R. Kuske, *Wide distribution and diversity of members of the bacterial kingdom Acidobacterium in the environment.* Appl Environ Microbiol, 1999. **65**(4): p. 1731-7.
72. Dunbar, J., et al., *Levels of bacterial community diversity in four arid soils compared by cultivation and 16S rRNA gene cloning.* Appl Environ Microbiol, 1999. **65**(4): p. 1662-9.
73. Janssen, P.H., *Identifying the Dominant Soil Bacterial Taxa in Libraries of 16S rRNA and 16S rRNA Genes.* Applied and Environmental Microbiology, 2006. **72**(3): p. 1719.
74. Liles, M.R., et al., *A census of rRNA genes and linked genomic sequences within a soil metagenomic library.* Appl Environ Microbiol, 2003. **69**(5): p. 2684-91.
75. Quast, C., et al., *The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.* Nucleic Acids Res, 2013. **41**(Database issue): p. D590-6.
76. Barns, S.M., et al., *Acidobacteria phylum sequences in uranium-contaminated subsurface sediments greatly expand the known diversity within the phylum.* Appl Environ Microbiol, 2007. **73**(9): p. 3113-6.
77. Dedysh, S.N. and P. Yilmaz, *Refining the taxonomic structure of the phylum Acidobacteria.* Int J Syst Evol Microbiol, 2018. **68**(12): p. 3796-3806.
78. Hausmann, B., et al., *Peatland Acidobacteria with a dissimilatory sulfur metabolism.* The ISME Journal, 2018. **12**(7): p. 1729-1742.
79. Crits-Christoph, A., et al., *Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis.* Nature, 2018. **558**(7710): p. 440-444.
80. Diamond, S., et al., *Mediterranean grassland soil C–N compound turnover is dependent on rainfall and depth, and is mediated by genomically divergent microorganisms.* Nature Microbiology, 2019. **4**(8): p. 1356-1367.
81. Podar, M., et al., *Complete genome sequence of Terriglobus albidus strain ORNL, an acidobacterium isolated from the Populus deltoides rhizosphere.* Microbiology Resource Announcements, 2019. **8**(46): p. e01065-19.
82. Rawat, S.R., et al., *Comparative genomic and physiological analysis provides insights into the role of Acidobacteria in organic carbon utilization in Arctic tundra soils.* FEMS microbiology ecology, 2012. **82**(2): p. 341-355.
83. Ward, N.L., et al., *Three genomes from the phylum Acidobacteria provide insight into the lifestyles of these microorganisms in soils.* Appl Environ Microbiol, 2009. **75**(7): p. 2046-56.
84. Anderson, I., et al., *Genome sequence of the homoacetogenic bacterium Holophaga foetida type strain (TMBS4T).* Standards in genomic sciences, 2012. **6**(2): p. 174-184.
85. Garcia Costas, A.M., et al., *Complete genome of Candidatus Chloracidobacterium thermophilum, a chlorophyll-based photoheterotroph belonging to the phylum Acidobacteria.* Environmental Microbiology, 2012. **14**(1): p. 177-190.
86. Stamps, B.W., et al., *Genome sequence of Thermoanaerobaculum aquaticum MP-01T, the first cultivated member of Acidobacteria subdivision 23, isolated from a hot spring.* Genome announcements, 2014. **2**(3): p. e00570-14.

87. Ward, L.M., S.E. McGlynn, and W.W. Fischer, *Draft genome sequence of Chloracidobacterium sp. CP2_5A, a phototrophic member of the phylum Acidobacteria recovered from a Japanese hot spring.* Genome Announcements, 2017. **5**(40): p. e00821-17.
88. Flieder, M., et al., *Novel taxa of Acidobacteriota implicated in seafloor sulfur cycling.* The ISME journal, 2021. **15**(11): p. 3159-3180.
89. Wegner, C.-E. and W. Liesack, *Unexpected dominance of elusive Acidobacteria in early industrial soft coal slags.* Frontiers in microbiology, 2017. **8**: p. 1023.
90. Elshahed, M.S., et al., *Bacterial diversity and sulfur cycling in a mesophilic sulfide-rich spring.* Applied and Environmental Microbiology, 2003. **69**(9): p. 5609-5621.
91. Senko, J.M., et al., *Barite deposition resulting from phototrophic sulfide-oxidizing bacterial activity*. Geochimica et Cosmochimica Acta, 2004. **68**(4): p. 773-780.
92. Buhring, S., et al., *Hinrichs 717 KU. 2011. Insights into chemotaxonomic composition and carbon cycling of 718 phototrophic communities in an artesian sulfur-rich spring (Zodletone, Oklahoma, USA), 719 a possible analog for ancient microbial mat systems.* Geobiology. **9**: p. 166-179.
93. Spain, A.M., et al., *Metatranscriptomic analysis of a high-sulfide aquatic spring reveals insights into sulfur cycling and unexpected aerobic metabolism.* PeerJ, 2015. **3**: p. e1259.
94. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-2120.
95. Kang, D.D., et al., *MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies.* PeerJ, 2019. **7**: p. e7359.
96. Sieber, C.M., et al., *Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy*. Nature microbiology, 2018. **3**(7): p. 836-843.
97. Chaumeil, P.-A., et al., *GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database.* Bioinformatics, 2019. **36**(6): p. 1925-1927.
98. Nayfach, S., *A Genomic Catalogue of Earth's Microbiomes.* Nature Biotechnology, 2020.
99. Kluber, L.A., et al., *SPRUCE deep peat heating (DPH) to whole ecosystem warming (WEW) metagenomes for peat samples collected June 2016*. 2018, Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States).
100. Klumber, L.A., Z.K. Yang, and C.W. Schadt, *SPRUCE deep peat heat (DPH) metagenomes for peat samples collected June 2015*. 2016, ORNLTESSFA (Oak Ridge National Lab's Terrestrial Ecosystem Science ….
101. Ziels, R.M., et al., *DNA-SIP based genome-centric metagenomics identifies key long-chain fatty acid-degrading populations in anaerobic digesters with different feeding frequencies.* The ISME journal, 2018. **12**(1): p. 112-123.
102. Baker, B.J., et al., *Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria.* Microbiome, 2015. **3**(1): p. 1-12.
103. D.H. Parks, C.R., M. Chuvochina, P.A. Chaumeil, B.J. Woodcroft, P.N. Evans, P. Hugenholtz, G.W. Tyson, , *Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life*, . Nat. Microbiol.,, 2017 **2**: p. 1533-1542.
104. Bowers, R.M., et al., *Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea.* Nature Biotechnology, 2017. **35**(8): p. 725-731.
105. Chan, P.P. and T.M. Lowe, *tRNAscan-SE: searching for tRNA genes in genomic sequences*, in *Gene prediction*. 2019, Springer. p. 1-14.
106. Couvin, D., et al., *CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins.* Nucleic Acids Research, 2018. **46**(W1): p. W246-W251.
107. Hyatt, D., et al., *Prodigal: prokaryotic gene recognition and translation initiation site identification.* BMC bioinformatics, 2010. **11**(1): p. 1-11.

108. Kanehisa, M., Y. Sato, and K. Morishima, *BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences.* J Mol Biol, 2016. **428**(4): p. 726-731.
109. Kanehisa, M. and Y. Sato, *KEGG Mapper for inferring cellular functions from protein sequences.* Protein Science, 2020. **29**(1): p. 28-35.
110. Nakamura, T., et al., *Parallelization of MAFFT for large-scale multiple sequence alignments.* Bioinformatics, 2018. **34**(14): p. 2490-2492.
111. Mistry, J., et al., *Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions.* Nucleic acids research, 2013. **41**(12): p. e121-e121.
112. Søndergaard, D., C.N. Pedersen, and C. Greening, *HydDB: a web tool for hydrogenase classification and analysis.* Scientific reports, 2016. **6**(1): p. 1-8.
113. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments.* PLOS ONE, 2010. **5**(3): p. e9490.
114. Zhang, H., et al., *dbCAN2: a meta server for automated carbohydrate-active enzyme annotation.* Nucleic acids research, 2018. **46**(W1): p. W95-W101.
115. Huang, L., et al., *dbCAN-seq: a database of carbohydrate-active enzyme (CAZyme) sequence and annotation.* Nucleic Acids Research, 2018. **46**(D1): p. D516-D521.
116. Medema, M.H., et al., *antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences.* Nucleic acids research, 2011. **39**(suppl_2): p. W339-W346.
117. Kumar, S., et al., *MEGA X: molecular evolutionary genetics analysis across computing platforms.* Molecular biology and evolution, 2018. **35**(6): p. 1547.
118. Mukherjee, S., et al., *Genomes OnLine database (GOLD) v. 7: updates and new features.* Nucleic acids research, 2019. **47**(D1): p. D649-D659.
119. Letunic, I. and P. Bork, *Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees.* Nucleic acids research, 2016. **44**(W1): p. W242-W245.
120. Ligthart, K., et al., *Bridging bacteria and the gut: functional aspects of type IV pili.* Trends in microbiology, 2020. **28**(5): p. 340-348.
121. Gallique, M., M. Bouteiller, and A. Merieau, *The Type VI Secretion System: A Dynamic System for Bacterial Communication?* Frontiers in Microbiology, 2017. **8**(1454).
122. Mendler, K., et al., *AnnoTree: visualization and exploration of a functionally annotated microbial tree of life.* Nucleic acids research, 2019. **47**(9): p. 4442-4448.
123. Giles, C., B. Cade-Menun, and J. Hill, *The inositol phosphates in soils and manures: Abundance, cycling, and measurement.* Canadian Journal of Soil Science, 2011. **91**(3): p. 397-416.
124. Gruteser, N., et al., *Sialic acid utilization by the soil bacterium Corynebacterium glutamicum.* FEMS microbiology letters, 2012. **336**(2): p. 131-138.
125. Abaibou, H., G. Giordano, and M.-A. Mandrand-Berthelot, *Suppression of Escherichia coli formate hydrogenlyase activity by trimethylamine N-oxide is due to drainage of the inducer formate.* Microbiology, 1997. **143**(8): p. 2657-2664.
126. Cox, J.C., et al., *Redox mechanisms in "oxidant-dependent" hexose fermentation by Rhodopseudomonas capsulata.* Archives of Biochemistry and Biophysics, 1980. **204**(1): p. 10-17.
127. Kalyuzhnaya, M.G., et al., *Characterization of a Novel Methanol Dehydrogenase in Representatives of Burkholderiales: Implications for Environmental Detection of Methylotrophy and Evidence for Convergent Evolution.* Journal of Bacteriology, 2008. **190**(11): p. 3817-3823.
128. Müller, B., L. Sun, and A. Schnürer, *First insights into the syntrophic acetate-oxidizing bacteria–a genetic study.* MicrobiologyOpen, 2013. **2**(1): p. 35-53.
129. Ragsdale, S.W. and E. Pierce, *Acetogenesis and the Wood–Ljungdahl pathway of CO2 fixation.* Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 2008. **1784**(12): p. 1873-1898.
130. Schuchmann, K. and V. Müller, *Energetics and application of heterotrophy in acetogenic bacteria.* Applied and environmental microbiology, 2016. **82**(14): p. 4056-4069.

131. Youssef, N.H., et al., *The Wood–Ljungdahl pathway as a key component of metabolic versatility in candidate phylum Bipolaricaulota (Acetothermia, OP1).* Environmental microbiology reports, 2019. **11**(4): p. 538-547.
132. Eichorst, S.A., et al., *Genomic insights into the Acidobacteria reveal strategies for their success in terrestrial environments.* Environmental Microbiology, 2018. **20**(3): p. 1041-1063.
133. Huang, S., et al., *First complete genome sequence of a subdivision 6 Acidobacterium strain.* Genome Announcements, 2016. **4**(3): p. e00469-16.
134. Domeignoz-Horta, L.A., K.M. DeAngelis, and G. Pold, *Draft genome sequence of Acidobacteria group 1 Acidipila sp. strain EB88, isolated from forest soil.* Microbiology Resource Announcements, 2019. **8**(1): p. e01464-18.
135. Eichorst, S.A., et al., *One complete and seven draft genome sequences of subdivision 1 and 3 Acidobacteria isolated from soil.* Microbiology resource announcements, 2020. **9**(5): p. e01087-19.
136. Kalam, S., et al., *Recent understanding of soil acidobacteria and their ecological significance: a critical review.* Frontiers in Microbiology, 2020. **11**: p. 580024.
137. Maistrenko, O.M., et al., *Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity.* The ISME journal, 2020. **14**(5): p. 1247-1259.
138. Challacombe, J.F., et al., *Biological consequences of ancient gene acquisition and duplication in the large genome of Candidatus Solibacter usitatus Ellin6076.* PloS one, 2011. **6**(9): p. e24882.
139. Karginov, F.V. and G.J. Hannon, *The CRISPR system: small RNA-guided defense in bacteria and archaea.* Molecular cell, 2010. **37**(1): p. 7-19.
140. Chistoserdova, L., *Modularity of methylotrophy, revisited.* Environmental microbiology, 2011. **13**(10): p. 2603-2622.
141. Kolb, S., *Aerobic methanol-oxidizing bacteria in soil.* FEMS Microbiology Letters, 2009. **300**(1): p. 1-10.
142. Butterfield, C.N., et al., *Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone.* PeerJ, 2016. **4**: p. e2687.
143. Lalonde, B.A., W. Ernst, and C. Garron, *Formaldehyde concentration in discharge from land based aquaculture facilities in Atlantic Canada.* Bulletin of Environmental Contamination and Toxicology, 2015. **94**(4): p. 444-447.
144. Waite, D.W., et al., *Proposal to reclassify the proteobacterial classes Deltaproteobacteria and Oligoflexia, and the phylum Thermodesulfobacteria into four phyla reflecting major functional capabilities.* International Journal of Systematic and Evolutionary Microbiology, 2020. **70**(11): p. 5972-6016.
145. Müller, A.L., et al., *Phylogenetic and environmental diversity of DsrAB-type dissimilatory (bi) sulfite reductases.* The ISME journal, 2015. **9**(5): p. 1152-1165.
146. Raoult, D., T. Marrie, and J. Mege, *Natural history and pathophysiology of Q fever.* Lancet Infect Dis, 2005. **5**(4): p. 219-26.
147. Miller, J.D., E.I. Shaw, and H.A. Thompson, *Coxiella burnetii, Q Fever, and Bioterrorism, in Microorganisms and Bioterrorism*, in *Microorganisms and Bioterrorism*, B. Anderson, H. Friedman, and M. Bendinelli, Editors. 2006, Springer US: Boston, MA. p. 181-208.
148. Arricau-Bouvery, N. and A. Rodolakis, *Is Q fever an emerging or re-emerging zoonosis?* Vet Res, 2005. **36**(3): p. 327-49.
149. Maurin, M. and D. Raoult, *Q fever.* Clin Microbiol Rev, 1999. **12**(4): p. 518-53.
150. McQuiston, J.H., J.E. Childs, and H.A. Thompson, *Q fever.* Journal of the American Veterinary Medical Association, 2002. **221**(6): p. 796-799.
151. van Schaik, E.J., et al., *Molecular pathogenesis of the obligate intracellular bacterium Coxiella burnetii.* Nat Rev Microbiol, 2013. **11**(8): p. 561-73.
152. Voth, D.E. and R.A. Heinzen, *Lounging in a lysosome: the intracellular lifestyle of Coxiella burnetii.* Cell Microbiol, 2007. **9**(4): p. 829-40.

153. Heinzen, R.A., T. Hackstadt, and J.E. Samuel, *Developmental biology of Coxiella burnettii.* Trends Microbiol, 1999. **7**(4): p. 149-54.
154. Rudolf Toman, R.A.H., James E. Samuel, Jean-Louis Mege, *Coxiella burnetii: Recent Advances and New Perspectives in Research of the Q Fever Bacterium*. Advances in Experimental Medicine and Biology. 2012: Springer Netherlands. XIV, 406.
155. Brennan, R.E., et al., *Both inducible nitric oxide synthase and NADPH oxidase contribute to the control of virulent phase I Coxiella burnetii infections.* Infection and immunity, 2004. **72**(11): p. 6666-6675.
156. Hackstadt, T. and J.C. Williams, *Biochemical stratagem for obligate parasitism of eukaryotic cells by Coxiella burnetii.* Proc Natl Acad Sci U S A, 1981. **78**(5): p. 3240-4.
157. Omsland, A., et al., *Host cell-free growth of the Q fever bacterium Coxiella burnetii.* Proc Natl Acad Sci U S A, 2009. **106**(11): p. 4430-4.
158. Voth, D.E. and R.A. Heinzen, *Coxiella type IV secretion and cellular microbiology*. Current opinion in microbiology, 2009. **12**(1): p. 74-80.
159. Newton, H.J., J.A. McDonough, and C.R. Roy, *Effector Protein Translocation by the Coxiella burnetii Dot/Icm Type IV Secretion System Requires Endocytic Maturation of the Pathogen-Occupied Vacuole*. PLOS ONE, 2013. **8**(1): p. e54566.
160. Beare, P.A., et al., *Essential role for the response regulator PmrA in Coxiella burnetii type 4B secretion and colonization of mammalian host cells.* Journal of bacteriology, 2014. **196**(11): p. 1925-1940.
161. Larson, C.L., et al., *Coxiella burnetii effector proteins that localize to the parasitophorous vacuole membrane promote intracellular replication.* Infect Immun, 2015. **83**(2): p. 661-70.
162. Larson, C.L. and R.A. Heinzen, *High-Content Imaging Reveals Expansion of the Endosomal Compartment during Coxiella burnetii Parasitophorous Vacuole Maturation.* Front Cell Infect Microbiol, 2017. **7**: p. 48.
163. Beare, P.A., et al., *Dot/Icm Type IVB Secretion System Requirements for Coxiella burnetii Growth in Human Macrophages.* mBio, 2011. **2**(4): p. e00175-11.
164. Crabill, E., et al., *Dot/Icm-Translocated Proteins Important for Biogenesis of the Coxiella burnetii-Containing Vacuole Identified by Screening of an Effector Mutant Sublibrary*. Infect Immun, 2018. **86**(4).
165. Newton, H.J., et al., *A screen of Coxiella burnetii mutants reveals important roles for Dot/Icm effectors and host autophagy in vacuole biogenesis.* PLoS Pathog, 2014. **10**(7): p. e1004286.
166. Chen, C., et al., *Large-scale identification and translocation of type IV secretion substrates by Coxiella burnetii.* Proc Natl Acad Sci U S A, 2010. **107**(50): p. 21755-60.
167. Carey, K.L., et al., *The Coxiella burnetii Dot/Icm system delivers a unique repertoire of type IV effectors into host cells and is required for intracellular replication.* PLoS Pathog, 2011. **7**(5): p. e1002056.
168. Martinez, E., et al., *Identification of OmpA, a Coxiella burnetii protein involved in host cell invasion, by multi-phenotypic high-content screening.* PLoS Pathog, 2014. **10**(3): p. e1004013.
169. Weber, M.M., et al., *Identification of Coxiella burnetii type IV secretion substrates required for intracellular replication and Coxiella-containing vacuole formation.* J Bacteriol, 2013. **195**(17): p. 3914-24.
170. HOWE, D., et al., *Fusogenicity of the Coxiella burnetii Parasitophorous Vacuole.* Annals of the New York Academy of Sciences, 2003. **990**(1): p. 556-562.
171. Morgan, J.K., B.E. Luedtke, and E.I. Shaw, *Polar localization of the Coxiella burnetii type IVB secretion system.* FEMS Microbiology Letters, 2010. **305**(2): p. 177-183.
172. Voth, D.E., et al., *The Coxiella burnetii cryptic plasmid is enriched in genes encoding type IV secretion system substrates.* J Bacteriol, 2011. **193**(7): p. 1493-503.
173. Nagai, H. and T. Kubori, *Type IVB Secretion Systems of Legionella and Other Gram-Negative Bacteria.* Frontiers in Microbiology, 2011. **2**.
174. Sexton, J.A. and J.P. Vogel, *Type IVB Secretion by Intracellular Pathogens.* Traffic, 2002. **3**(3): p. 178-185.

175. Segal, G. and H.A. Shuman, *Possible origin of the Legionella pneumophila virulence genes and their relation to Coxiella burnetii.* Molecular Microbiology, 1999. **33**(3): p. 669-670.
176. Zamboni, D.S., et al., *Coxiella burnetii express type IV secretion system proteins that function similarly to components of the Legionella pneumophila Dot/Icm system.* Molecular Microbiology, 2003. **49**(4): p. 965-976.
177. Segal, G., M. Feldman, and T. Zusman, *The Icm/Dot type-IV secretion systems of Legionella pneumophila and Coxiella burnetii.* FEMS Microbiology Reviews, 2005. **29**(1): p. 65-81.
178. Vogel, J.P., *Turning a tiger into a house cat: using Legionella pneumophila to study Coxiella burnetii.* Trends in Microbiology, 2004. **12**(3): p. 103-105.
179. Pan, X., et al., *Ankyrin repeat proteins comprise a diverse family of bacterial type IV effectors.* Science, 2008. **320**(5883): p. 1651-4.
180. Zusman, T., G. Yerushalmi, and G. Segal, *Functional similarities between the icm/dot pathogenesis systems of Coxiella burnetii and Legionella pneumophila.* Infect Immun, 2003. **71**(7): p. 3714-23.
181. Omsland, A., et al., *Isolation from animal tissue and genetic transformation of Coxiella burnetii are facilitated by an improved axenic growth medium.* Appl Environ Microbiol, 2011. **77**(11): p. 3720-5.
182. Sanchez, S.E., E. Vallejo-Esquerra, and A. Omsland, *Use of Axenic Culture Tools to Study Coxiella burnetii.* Curr Protoc Microbiol, 2018. **50**(1): p. e52.
183. Martinez, E., F. Cantet, and M. Bonazzi, *Generation and multi-phenotypic high-content screening of Coxiella burnetii transposon mutants.* J Vis Exp, 2015(99): p. e52851.
184. Beare, P.A., et al., *Genetic mechanisms of Coxiella burnetii lipopolysaccharide phase variation.* PLOS Pathogens, 2018. **14**(3): p. e1006922.
185. Sandoz, K.M., et al., *Complementation of Arginine Auxotrophy for Genetic Transformation of Coxiella burnetii by Use of a Defined Axenic Medium.* Appl Environ Microbiol, 2016. **82**(10): p. 3042-51.
186. Beare, P.A. and R.A. Heinzen, *Gene inactivation in Coxiella burnetii*, in *Host-Bacteria Interactions*. 2014, Springer. p. 329-345.
187. Beare, P.A., et al., *Two systems for targeted gene deletion in Coxiella burnetii.* Applied and environmental microbiology, 2012. **78**(13): p. 4580-4589.
188. Coleman, S.A., et al., *Temporal analysis of Coxiella burnetii morphological differentiation.* J Bacteriol, 2004. **186**(21): p. 7344-52.
189. Brennan, R.E. and J.E. Samuel, *Evaluation of Coxiella burnetii Antibiotic Susceptibilities by Real-Time PCR Assay.* Journal of Clinical Microbiology, 2003. **41**(5): p. 1869-1874.
190. Luedtke, B.E., et al., *The Coxiella Burnetii type IVB secretion system (T4BSS) component DotA is released/secreted during infection of host cells and during in vitro growth in a T4BSS-dependent manner.* Pathogens and Disease, 2017. **75**(4).
191. Moormeier, D.E., et al., *Coxiella burnetii RpoS Regulates Genes Involved in Morphological Differentiation and Intracellular Growth.* J Bacteriol, 2019. **201**(8).
192. Kim, D., et al., *Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype.* Nature Biotechnology, 2019. **37**(8): p. 907-915.
193. Kovaka, S., et al., *Transcriptome assembly from long-read RNA-seq alignments with StringTie2.* Genome Biology, 2019. **20**(1): p. 278.
194. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biology, 2014. **15**(12): p. 550.
195. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biol, 2010. **11**(10): p. R106.
196. Yu, N.Y., et al., *PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes.* Bioinformatics, 2010. **26**(13): p. 1608-15.
197. Mistry, J., et al., *Pfam: The protein families database in 2021.* Nucleic Acids Research, 2020. **49**(D1): p. D412-D419.

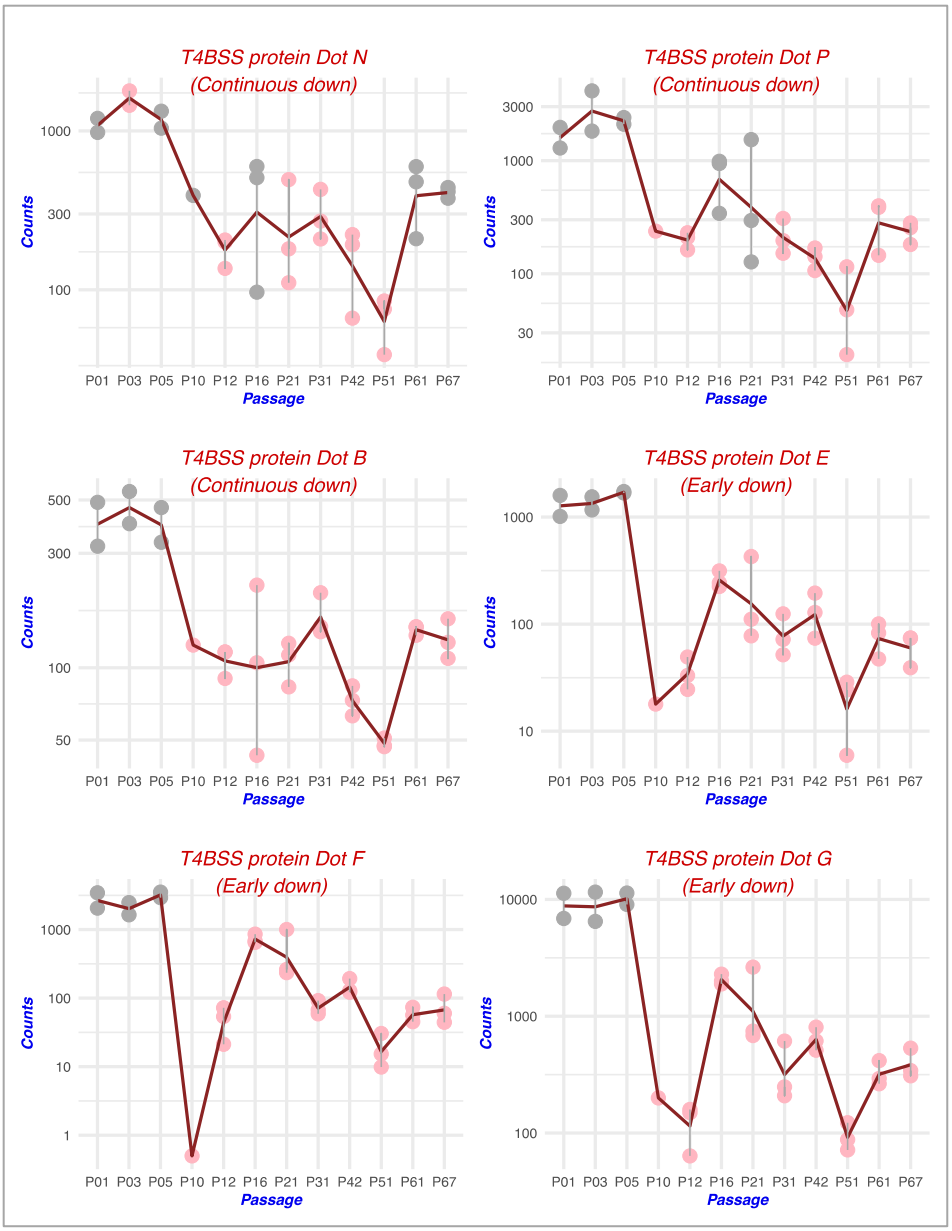
198. Galperin, M.Y., et al., *COG database update: focus on microbial diversity, model organisms, and widespread pathogens.* Nucleic Acids Res, 2021. **49**(D1): p. D274-d281.
199. Allen, B., et al., *Using KBase to Assemble and Annotate Prokaryotic Genomes.* Curr Protoc Microbiol, 2017. **46**: p. 1E 13 1-1E 13 18.
200. Andrews, S., *FastQC: a quality control tool for high throughput sequence data.* . 2010.
201. Davis, J.J., et al., *The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities.* Nucleic Acids Res, 2020. **48**(D1): p. D606-D612.
202. Gurevich, A., et al., *QUAST: quality assessment tool for genome assemblies.* Bioinformatics, 2013. **29**(8): p. 1072-1075.
203. Deatherage, D.E. and J.E. Barrick, *Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq.* Methods Mol Biol, 2014. **1151**: p. 165-88.
204. Green, E.R. and J. Mecsas, *Bacterial Secretion Systems: An Overview.* Microbiol Spectr, 2016. **4**(1).
205. Tsirigotaki, A., et al., *Protein export through the bacterial Sec pathway.* Nature Reviews Microbiology, 2017. **15**(1): p. 21-36.
206. Mori, H. and K. Ito, *The Sec protein-translocation pathway.* Trends in Microbiology, 2001. **9**(10): p. 494-500.
207. Larson, C.L., et al., *Right on Q: genetics begin to unravel Coxiella burnetii host cell interactions.* Future Microbiol, 2016. **11**: p. 919-39.
208. Fielden, L.F., et al., *Proteomic Identification of Coxiella burnetii Effector Proteins Targeted to the Host Cell Mitochondria During Infection.* Mol Cell Proteomics, 2021. **20**: p. 100005.
209. Maturana, P., et al., *Refining the plasmid-encoded type IV secretion system substrate repertoire of Coxiella burnetii.* J Bacteriol, 2013. **195**(14): p. 3269-76.
210. Voth, D.E., et al., *The Coxiella burnetii ankyrin repeat domain-containing protein family is heterogeneous, with C-terminal truncations that influence Dot/Icm-mediated secretion.* J Bacteriol, 2009. **191**(13): p. 4232-42.
211. Luhrmann, A., et al., *Inhibition of pathogen-induced apoptosis by a Coxiella burnetii type IV effector protein.* Proc Natl Acad Sci U S A, 2010. **107**(44): p. 18997-9001.
212. Habyarimana, F., et al., *Role for the Ankyrin eukaryotic-like genes of Legionella pneumophila in parasitism of protozoan hosts and human macrophages.* Environ Microbiol, 2008. **10**(6): p. 1460-74.
213. Lifshitz, Z., et al., *Computational modeling and experimental validation of the Legionella and Coxiella virulence-related type-IVB secretion signal.* Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(8): p. E707-E715.
214. Neckers, L. and U. Tatu, *Molecular chaperones in pathogen virulence: emerging new targets for therapy.* Cell host & microbe, 2008. **4**(6): p. 519-527.
215. Beare, P.A., et al., *Comparative genomics reveal extensive transposon-mediated genomic plasticity and diversity among potential effector proteins within the genus Coxiella.* Infect Immun, 2009. **77**(2): p. 642-56.
216. Genevaux, P., C. Georgopoulos, and W.L. Kelley, *The Hsp70 chaperone machines of Escherichia coli: a paradigm for the repartition of chaperone functions.* Mol Microbiol, 2007. **66**(4): p. 840-57.
217. Arnold, D.L., et al., *Evolution of microbial virulence: the benefits of stress.* Trends Genet, 2007. **23**(6): p. 293-300.
218. Takaya, A., et al., *The DnaK/DnaJ chaperone machinery of Salmonella enterica serovar Typhimurium is essential for invasion of epithelial cells and survival within macrophages, leading to systemic infection.* Infect Immun, 2004. **72**(3): p. 1364-73.
219. Macellaro, A., et al., *Identification of a 71-kilodalton surface-associated Hsp70 homologue in Coxiella burnetii.* Infection and immunity, 1998. **66**(12): p. 5882-5888.
220. Williams, J.C. and D.M. Waag, *Antigens, virulence factors, and biological response modifiers of Coxiella burnetii: strategies for vaccine development.* Q fever: the biology of *Coxiella burnetii.*, 1991: p. 175-222.

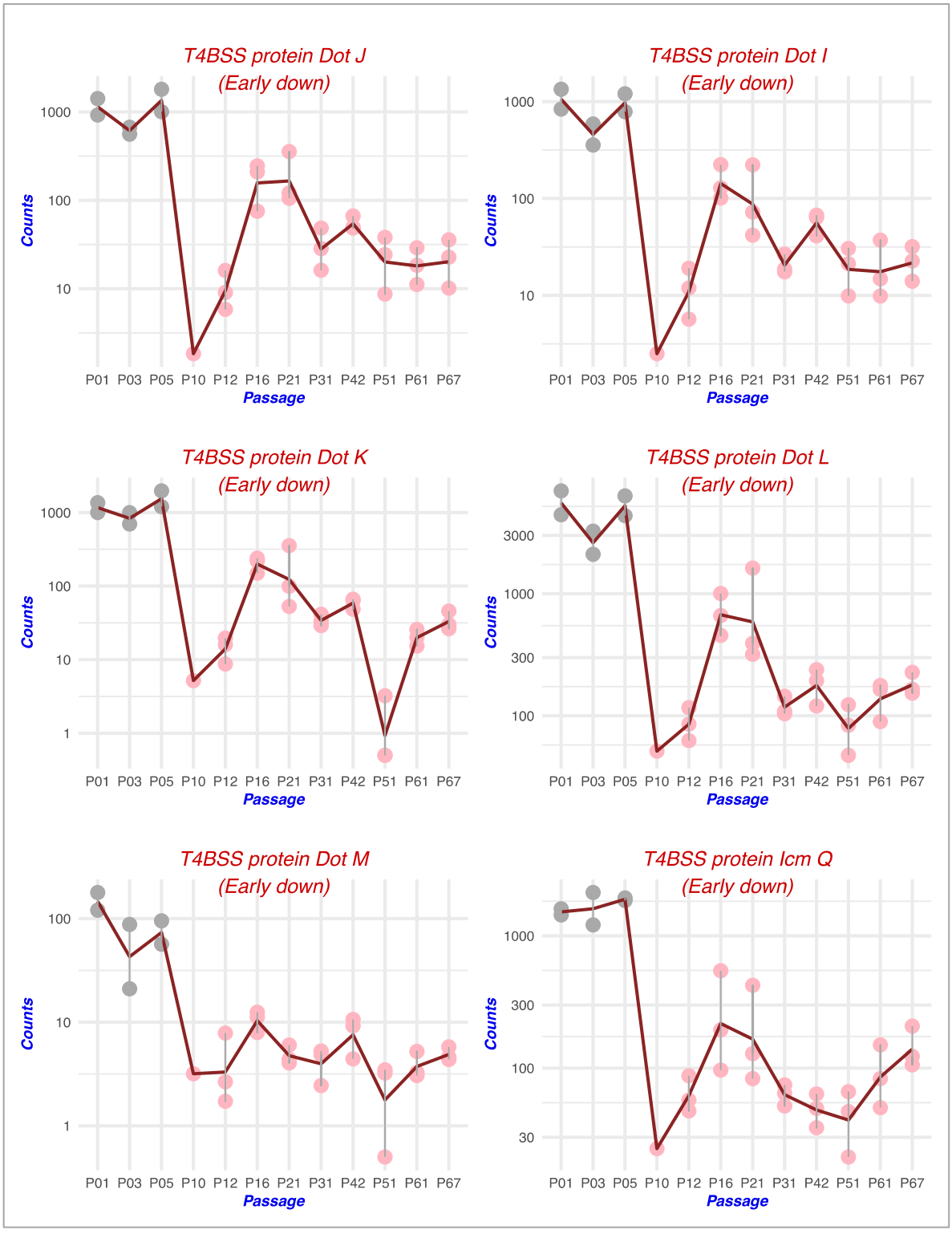
221. Gajdosova, E., et al., *Immunogenicity of Coxiella burnetii whole cells and their outer membrane components.* Acta virologica, 1994. **38**(6): p. 339-344.
222. Hussein, A., E. Kovacova, and R. Toman, *Isolation and evaluation of Coxiella burnetii O-polysaccharide antigen as an immunodiagnostic reagent.* Acta virologica, 2001. **45**(3): p. 173-180.
223. Samuel, G. and P. Reeves, *Biosynthesis of O-antigens: genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly.* Carbohydrate Research, 2003. **338**(23): p. 2503-2519.
224. Amano, K., et al., *Biochemical and immunological properties of Coxiella burnetii cell wall and peptidoglycan-protein complex fractions.* Journal of bacteriology, 1984. **160**(3): p. 982-988.
225. Sandoz, K.M., et al., *Transcriptional Profiling of Coxiella burnetii Reveals Extensive Cell Wall Remodeling in the Small Cell Variant Developmental Form.* PloS one, 2016. **11**(2): p. e0149957-e0149957.
226. Zusman, T., et al., *The response regulator PmrA is a major regulator of the icm/dot type IV secretion system in Legionella pneumophila and Coxiella burnetii.* Mol Microbiol, 2007. **63**(5): p. 1508-23.
227. Weber, M.M., *Identification Of C. burnetii Type IV Secretion Substrates required for Intracallular replication and Coxiella-containing vacoule formation* 2014, Texas A&M University p. 126.
228. Beare, P.A., et al., *Candidate antigens for Q fever serodiagnosis revealed by immunoscreening of a Coxiella burnetii protein microarray.* Clinical and vaccine immunology : CVI, 2008. **15**(12): p. 1771-1779.
229. Bewley, K.R., *The identification of immune-reactive proteins recognised in response to Coxiella burnetii infection*, in *School of Pharmacy and Biomedical Sciences*. 2015, University of Portsmouth. p. 230.
230. Sandoz, K.M., et al., *Developmental transitions of Coxiella burnetii grown in axenic media.* J Microbiol Methods, 2014. **96**: p. 104-10.
231. Bewley, K.R., *The identification of immune-reactive proteins recognised in response to Coxiella burnetii infection*, in *School of Pharmacy and Biomedical Sciences* 2015, University of Portsmouth
232. Burette, M., *Intracellular replication and persistence strategies of the Q Fever pathogen Coxiella burnetii*, in *Agricultural sciences*. 2020, Université Montpellier.
233. Cotter, P.A., et al., *Cytochrome o (cyoABCDE) and d (cydAB) oxidase gene expression in Escherichia coli is regulated by oxygen, pH, and the fnr gene product.* J Bacteriol, 1990. **172**(11): p. 6333-8.
234. Kuley, R., et al., *Major differential gene regulation in Coxiella burnetii between in vivo and in vitro cultivation models.* BMC Genomics, 2015. **16**: p. 953.
235. Ito, M., M. Morino, and T.A. Krulwich, *Mrp Antiporters Have Important Roles in Diverse Bacteria and Archaea.* Front Microbiol, 2017. **8**: p. 2325.
236. Seshadri, R., et al., *Complete genome sequence of the Q-fever pathogen Coxiella burnetii.* Proc Natl Acad Sci U S A, 2003. **100**(9): p. 5455-60.
237. Minnick, M.F. and R. Raghavan, *Developmental biology of Coxiella burnetii.* Adv Exp Med Biol, 2012. **984**: p. 231-48.
238. Vincent, C.D., et al., *Identification of the core transmembrane complex of the Legionella Dot/Icm type IV secretion system.* Mol Microbiol, 2006. **62**(5): p. 1278-91.
239. Vincent, C.D., et al., *Identification of the DotL coupling protein subcomplex of the Legionella Dot/Icm type IV secretion system.* Mol Microbiol, 2012. **85**(2): p. 378-91.
240. Zusman, T., et al., *Characterization of the icmH and icmF genes required for Legionella pneumophila intracellular growth, genes that are present in many bacteria associated with eukaryotic cells.* Infect Immun, 2004. **72**(6): p. 3398-409.
241. VanRheenen, S.M., G. Duménil, and R.R. Isberg, *IcmF and DotU are required for optimal effector translocation and trafficking of the Legionella pneumophila vacuole.* Infection and immunity, 2004. **72**(10): p. 5972-5982.
242. Sexton, J.A., et al., *Legionella pneumophila DotU and IcmF are required for stability of the Dot/Icm complex.* Infect Immun, 2004. **72**(10): p. 5983-92.

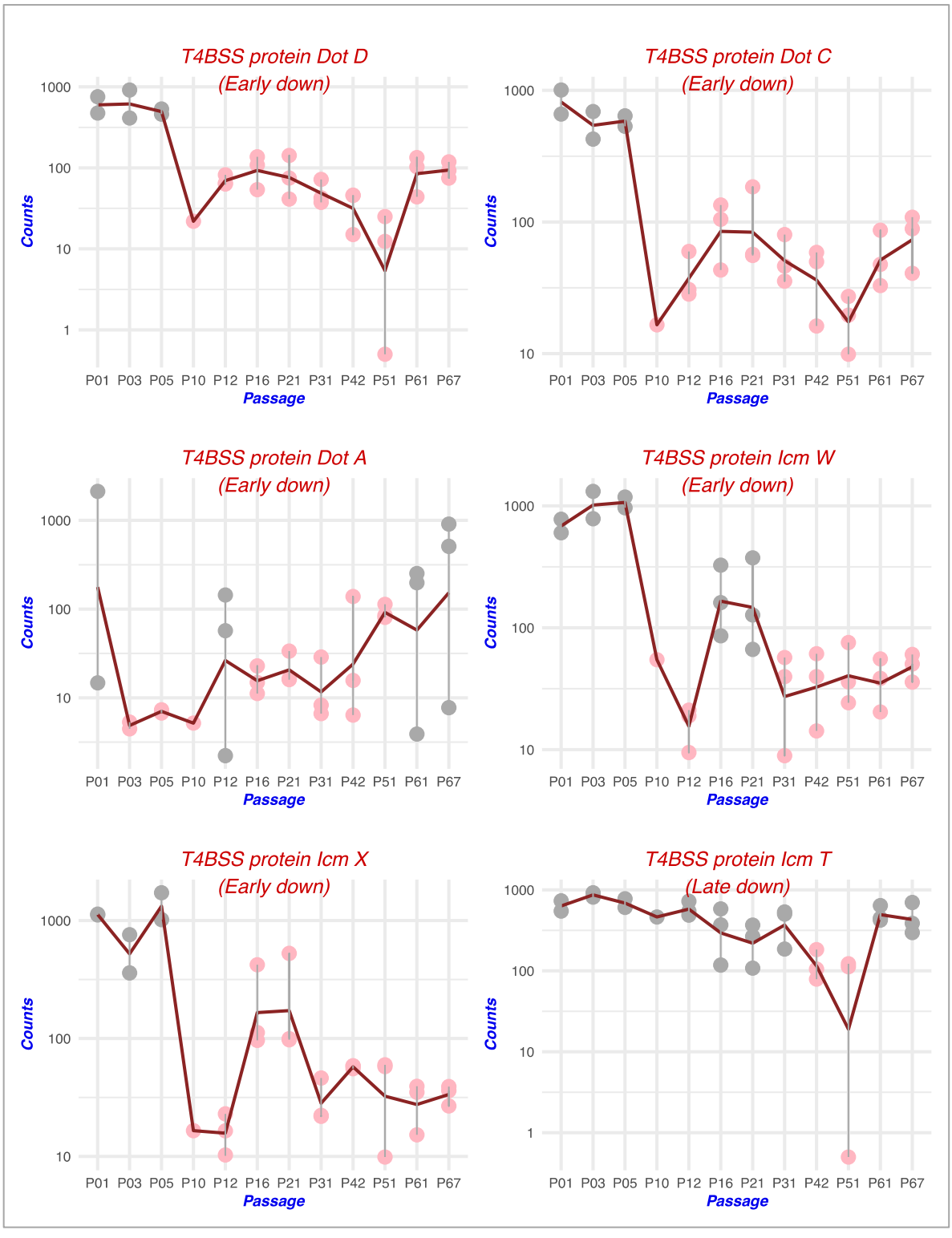
243. Batrukova, M.A., et al., *Ankyrin: structure, properties, and functions.* Biochemistry (Mosc), 2000. **65**(4): p. 395-408.
244. Pechstein, J., et al., *The Coxiella burnetii T4SS Effector AnkF Is Important for Intracellular Replication.* Frontiers in Cellular and Infection Microbiology, 2020. **10**.
245. Cordsmeier, A., et al., *The Coxiella burnetii T4SS effector protein AnkG hijacks the 7SK small nuclear ribonucleoprotein complex for reprogramming host cell transcription.* PLOS Pathogens, 2022. **18**(2): p. e1010266.

APPENDICES

Figure S1. A. Plot count graphs for all 19 DEG T4BSS genes. S1.B. Plot count graphs for all 7 DEGs in the sec. pathway. The y-axis represents log of counts normalized by estimated size factors and x-axis represents the passage. Dots represent replicates in a passage, pink dots represent passages in which L2Fc was found to be significant, and the line(red) represents the mean of the counts in each passage.




T4BSS protein Dot J
(Early down)
T4BSS protein Dot I
(Early down)
T4BSS protein Dot K
(Early down)
T4BSS protein Dot L
(Early down)
T4BSS protein Dot M
(Early down)
T4BSS protein Icm Q
(Early down)
Counts
Passage
P01 P03 P05 P10 P12 P16 P21 P31 P42 P51 P61 P67


T4BSS protein Dot D
(Early down)
T4BSS protein Dot C
(Early down)
T4BSS protein Dot A
(Early down)
T4BSS protein Icm W
(Early down)
T4BSS protein Icm X
(Early down)
T4BSS protein Icm T
(Late down)
Counts
Passage
1000
100
10
1
P01 P03 P05 P10 P12 P16 P21 P31 P42 P51 P61 P67

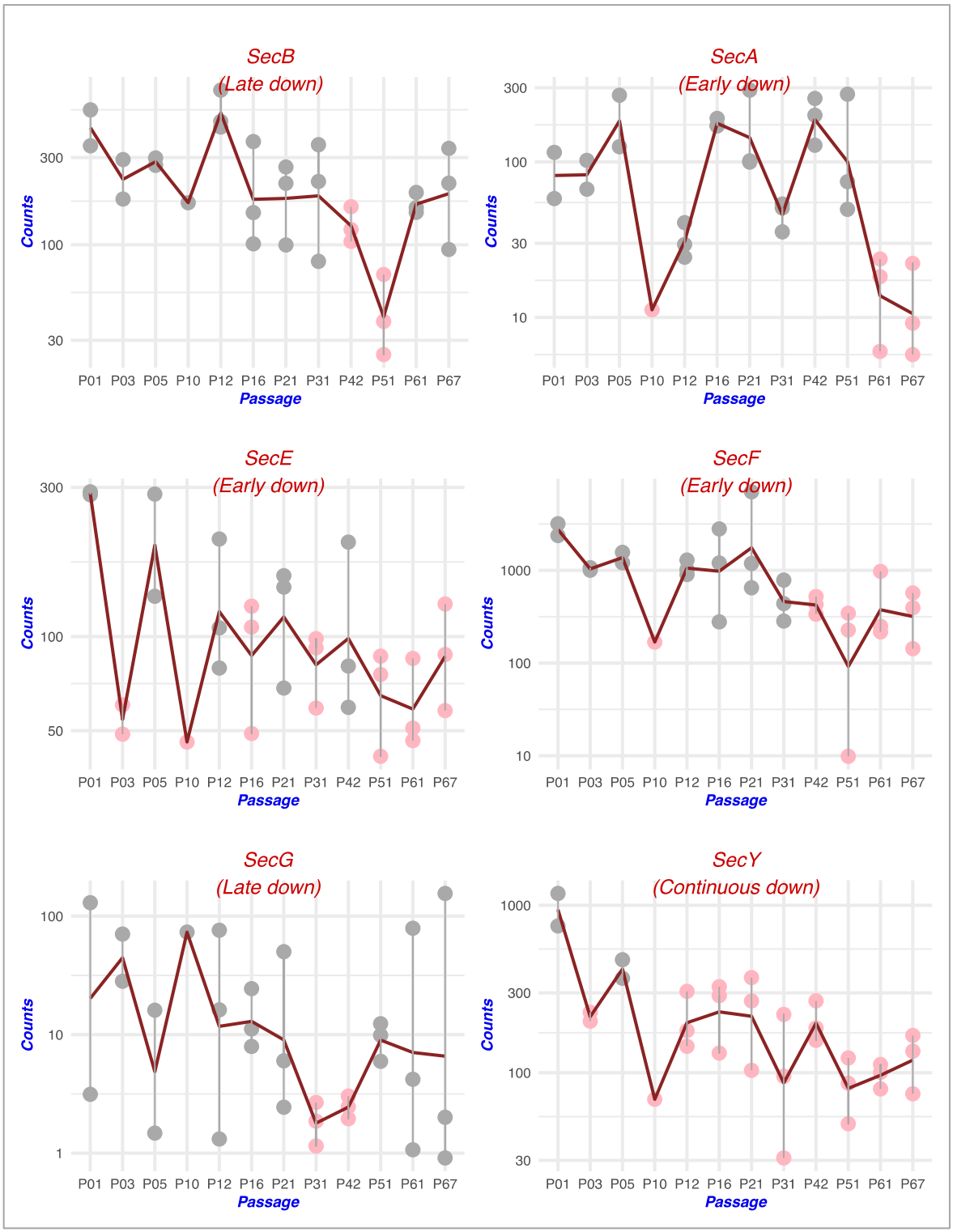
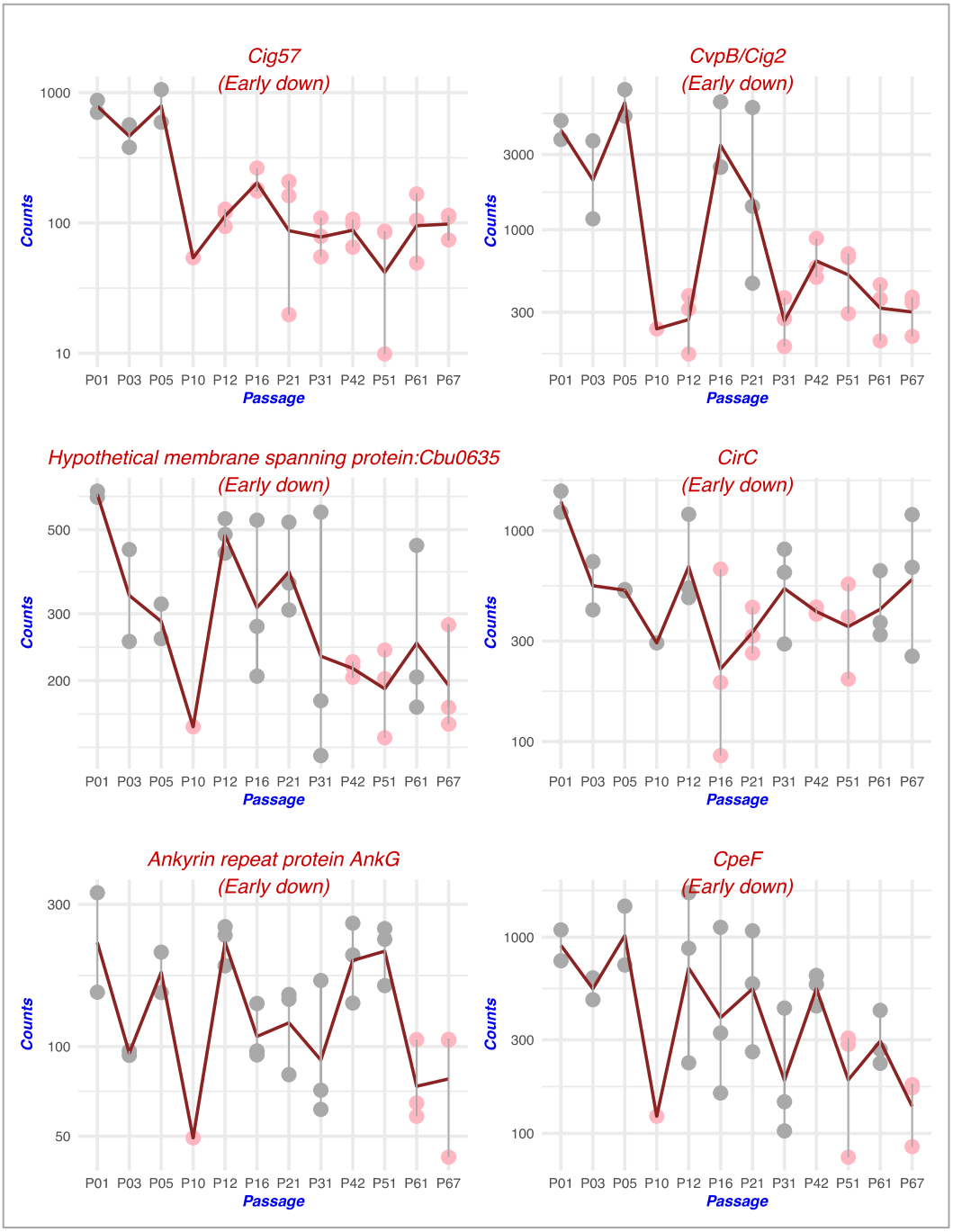

SecB
(Late down)
SecA
(Early down)
SecE
(Early down)
SecF
(Early down)
SecG
(Late down)
SecY
(Continuous down)
Counts
Passage
P01 P03 P05 P10 P12 P16 P21 P31 P42 P51 P61 P67

Figure S2: Plot count graphs for 14 downregulated effector proteins. The y-axis represents log of counts normalized by estimated size factors and x-axis represents the passage. Dots represent replicates in a passage, pink dots represent passages in which L2Fc was found to be significant, and the red line represents the mean of the counts in each passage.



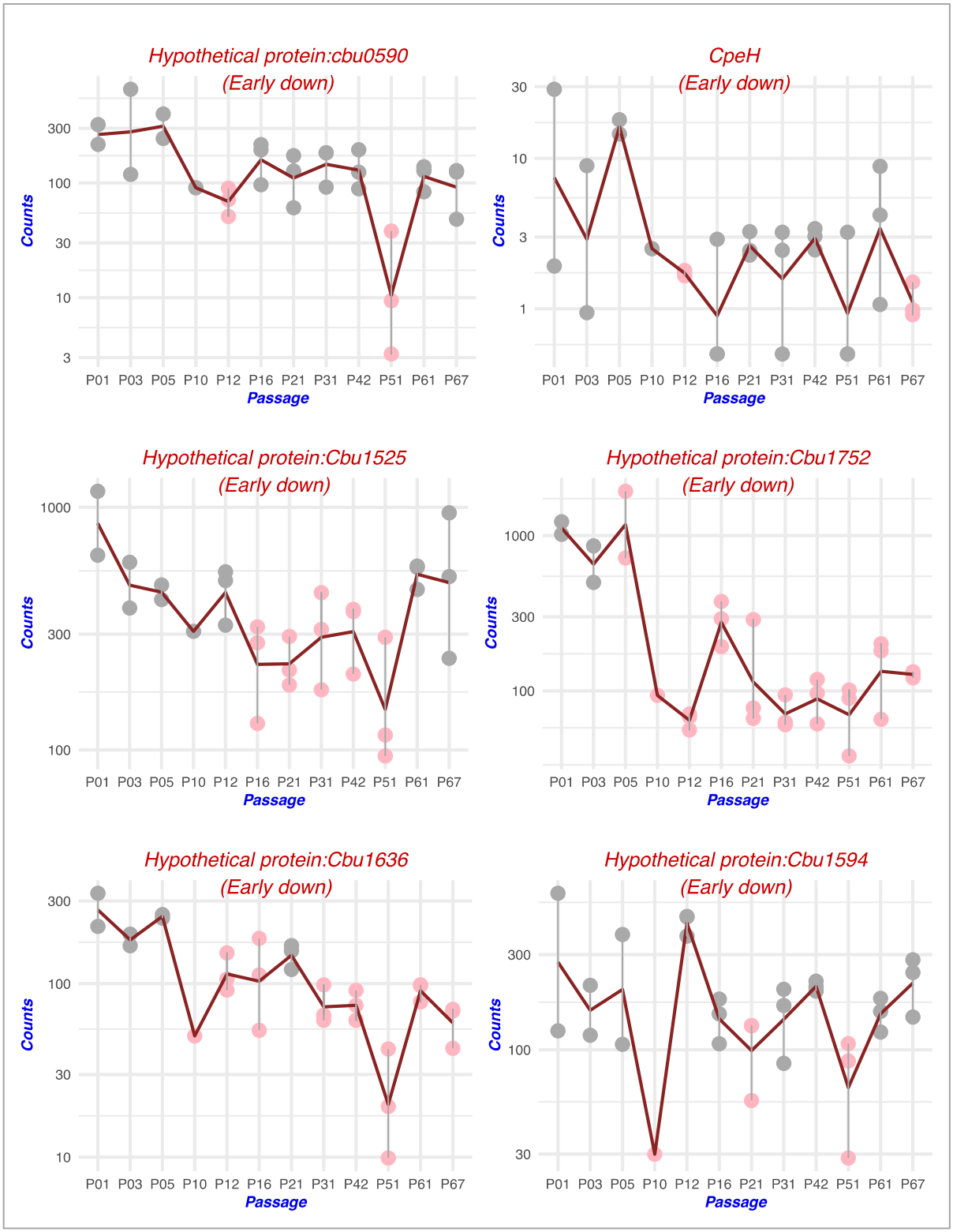
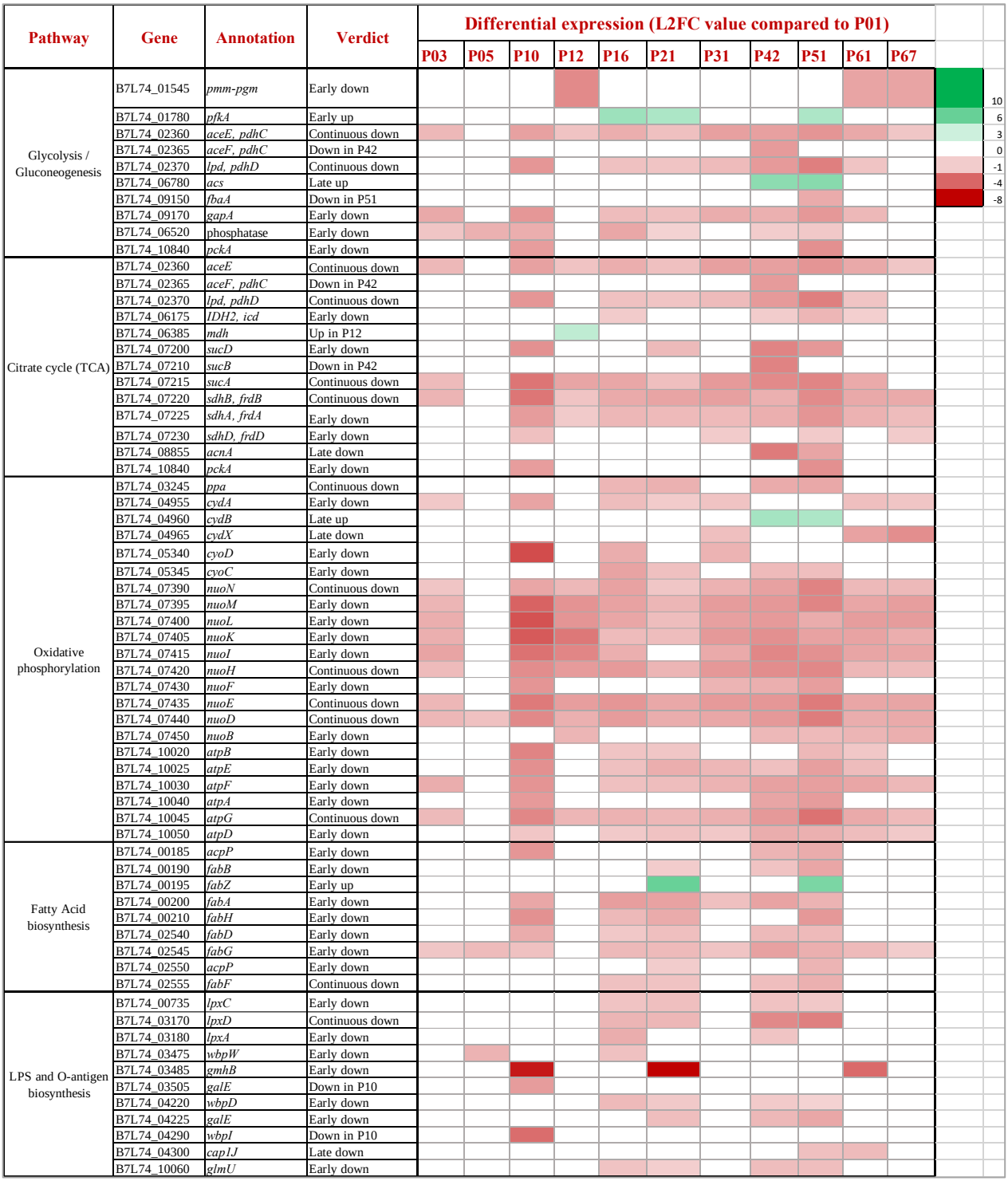
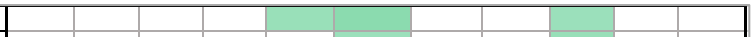
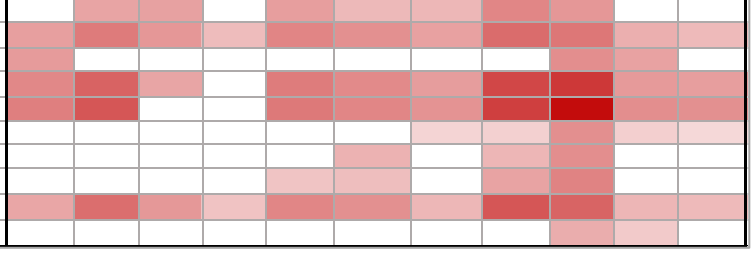

Hypothetical protein:cbu0590
(Early down)
Counts
300
100
30
10
3
P01 P03 P05 P10 P12 P16 P21 P31 P42 P51 P61 P67
Passage
CpeH
(Early down)
Counts
30
10
3
1
P01 P03 P05 P10 P12 P16 P21 P31 P42 P51 P61 P67
Passage
Hypothetical protein:Cbu1525
(Early down)
Counts
1000
300
100
P01 P03 P05 P10 P12 P16 P21 P31 P42 P51 P61 P67
Passage
Hypothetical protein:Cbu1752
(Early down)
Counts
1000
300
100
P01 P03 P05 P10 P12 P16 P21 P31 P42 P51 P61 P67
Passage
Hypothetical protein:Cbu1636
(Early down)
Counts
300
100
30
10
P01 P03 P05 P10 P12 P16 P21 P31 P42 P51 P61 P67
Passage
Hypothetical protein:Cbu1594
(Early down)
Counts
300
100
30
P01 P03 P05 P10 P12 P16 P21 P31 P42 P51 P61 P67
Passage

Figure S3: Heatmap for DEGs involved in central metabolic pathways and transporters. The L2fc values used to plot differential expression of the genes ranges from -8 to 10 over different passage.



Pathway	Gene	Annotation	Verdict	Differential expression (L2FC value compared to P01)										
				P03	P05	P10	P12	P16	P21	P31	P42	P51	P61	P67
Glycolysis / Gluconeogenesis	B7L74_01545	*pmm-pgm*	Early down											
	B7L74_01780	*pfkA*	Early up											
	B7L74_02360	*aceE, pdhC*	Continuous down											
	B7L74_02365	*aceF, pdhC*	Down in P42											
	B7L74_02370	*lpd, pdhD*	Continuous down											
	B7L74_06780	*acs*	Late up											
	B7L74_09150	*fbaA*	Down in P51											
	B7L74_09170	*gapA*	Early down											
	B7L74_06520	phosphatase	Early down											
	B7L74_10840	*pckA*	Early down											
Citrate cycle (TCA)	B7L74_02360	*aceE*	Continuous down											
	B7L74_02365	*aceF, pdhC*	Down in P42											
	B7L74_02370	*lpd, pdhD*	Continuous down											
	B7L74_06175	*IDH2, icd*	Early down											
	B7L74_06385	*mdh*	Up in P12											
	B7L74_07200	*sucD*	Early down											
	B7L74_07210	*sucB*	Down in P42											
	B7L74_07215	*sucA*	Continuous down											
	B7L74_07220	*sdhB, frdB*	Continuous down											
	B7L74_07225	*sdhA, frdA*	Early down											
	B7L74_07230	*sdhD, frdD*	Early down											
	B7L74_08855	*acnA*	Late down											
	B7L74_10840	*pckA*	Early down											
Oxidative phosphorylation	B7L74_03245	*ppa*	Continuous down											
	B7L74_04955	*cydA*	Early down											
	B7L74_04960	*cydB*	Late up											
	B7L74_04965	*cydX*	Late down											
	B7L74_05340	*cyoD*	Early down											
	B7L74_05345	*cyoC*	Early down											
	B7L74_07390	*nuoN*	Continuous down											
	B7L74_07395	*nuoM*	Early down											
	B7L74_07400	*nuoL*	Early down											
	B7L74_07405	*nuoK*	Early down											
	B7L74_07415	*nuoI*	Early down											
	B7L74_07420	*nuoH*	Continuous down											
	B7L74_07430	*nuoF*	Early down											
	B7L74_07435	*nuoE*	Continuous down											
	B7L74_07440	*nuoD*	Continuous down											
	B7L74_07450	*nuoB*	Early down											
	B7L74_10020	*atpB*	Early down											
	B7L74_10025	*atpE*	Early down											
	B7L74_10030	*atpF*	Early down											
	B7L74_10040	*atpA*	Early down											
	B7L74_10045	*atpG*	Continuous down											
	B7L74_10050	*atpD*	Early down											
Fatty Acid biosynthesis	B7L74_00185	*acpP*	Early down											
	B7L74_00190	*fabB*	Early down											
	B7L74_00195	*fabZ*	Early up											
	B7L74_00200	*fabA*	Early down											
	B7L74_00210	*fabH*	Early down											
	B7L74_02540	*fabD*	Early down											
	B7L74_02545	*fabG*	Early down											
	B7L74_02550	*acpP*	Early down											
	B7L74_02555	*fabF*	Continuous down											
LPS and O-antigen biosynthesis	B7L74_00735	*lpxC*	Early down											
	B7L74_03170	*lpxD*	Continuous down											
	B7L74_03180	*lpxA*	Early down											
	B7L74_03475	*wbpW*	Early down											
	B7L74_03485	*gmhB*	Early down											
	B7L74_03505	*galE*	Down in P10											
	B7L74_04220	*wbpD*	Early down											
	B7L74_04225	*galE*	Early down											
	B7L74_04290	*wbpI*	Down in P10											
	B7L74_04300	*cap1J*	Late down											
	B7L74_10060	*glmU*	Early down											

Transporters	B7L74_02485	*artQ*	Early up	
	B7L74_03540	*wzm, rfbA*	Early up	
	B7L74_03545	*wzt, rbfB*	Early down	
	B7L74_05125	*lolD*	Early down	
	B7L74_05130	*lolC_E*	Early up	
	B7L74_09820	*ftsE*	Early down	
	B7L74_10570	*lapB*	Early up	
	B7L74_00460	*pal*	Down in P51	
	B7L74_00480	*yrbG*	Early up	
	B7L74_01455	MarC family	Early up	
	B7L74_01835	*czcD, zitB*	Late down	
	B7L74_01880	*PHO84*	Continuous up	
	B7L74_01910	*bcr, tcaB*	Early up	
	B7L74_02215	MFS transporter	Early up	
	B7L74_02490	*artM*	Early up	
	B7L74_02665	MFS transporter	Early up	
	B7L74_02695	*dedA*	Early up	
	B7L74_03280	EamA family	Down in P10	
	B7L74_03345	*perM*	Up in P16	
	B7L74_03565	*perM*	Late up	
	B7L74_03840	*bamD*	Up in P12	
	B7L74_03905	Multidrug transpor	Early up	
	B7L74_04030	MFS transporter	Late up	
	B7L74_04070	*emrA*	Early down	
	B7L74_04245	*pglK*	Early down	
	B7L74_04625	MFS transporter	Early up	
	B7L74_04765	*ABC-2.A*	Late up	
	B7L74_05540	*ynaI, mscMJ*	Early up	
	B7L74_05815	Oligopeptide transp	Early up	
	B7L74_06065	*cmlA, cmlB, floR*	Late up	
	B7L74_06390	*fadL*	Late up	
	B7L74_06400	MFS transporter	Early up	
	B7L74_06620	MFS transporter	Late up	
	B7L74_06940	*glcP*	Late up	
	B7L74_07280	*omp31*	Early up	
	B7L74_07620	*ahpC*	Early up	
	B7L74_07760	*copA, ctpA*	Early down	
	B7L74_08085	MFS transporter	Early up	
	B7L74_08145	*mnhF, mrpF*	Down in P10	
	B7L74_08165	*mnhB, mrpB*	Early up	
	B7L74_08170	*mnhA, mrpA*	Continuous up	
	B7L74_08255	*mscK, kefA*	Early up	
	B7L74_08895	*pnuC*	Early up	
	B7L74_08915	Na+ symporter;NS	Up in P21	
	B7L74_08920	Na+ symporter;NS	Continuous up	
	B7L74_09100	*feoB*	Early up	
	B7L74_09260	*yhhQ*	Early up	
	B7L74_09330	*TC.OOP;OmpA-O*	Early up	
	B7L74_10175	Bcr/CflA family ef	Continuous up	
Chaperons	B7L74_06645	*dnaJ*	Early down	
	B7L74_06650	*dnaK, HSPA9*	Early down	
	B7L74_03255	*fklB*	Early down	
	B7L74_08845	*groEL, HSPD1*	Continuous down	
	B7L74_08850	*groES*	Continuous down	
	B7L74_07830	*grxC*	Late down	
	B7L74_03030	*grxD*	Early down	
	B7L74_10420	*hslU*	Continuous down	
	B7L74_01615	*htpG, HSP90A*	Continuous down	
	B7L74_10235	*surA*	Late down	

VITA

Archana Yadav

Candidate for the Degree of

Doctor of Philosophy

Thesis: EXPLORING ENVIRONMENTAL ADAPTATIONS AND HABITAT PREFERENCES IN THREE MICROBIAL LINEAGES USING COMPARATIVE (META)GENOMIC APPROACHES

Major Field: Microbiology, Cell, and Molecular Biology

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Microbiology at Oklahoma State University, Stillwater, Oklahoma in December 2022.

Completed the requirements for the Bachelor of Technology in Biotechnology at Kathmandu University, City, Dhulikhel, Nepal in 2015.

Experience:

Graduate Research/Teaching Assistant at Oklahoma State University. 2017-2020.
Research Assistant at Research Institute of Bioscience and Biotechnology (RIBB). Nepal. 2016-2017
Certified Instructor at Software Carpentries and OSU carpentries. 2021-Present.

Professional Memberships:

Research Institute of bioscience and Biotechnology alumni network. 2022-Present
American Society for Microbiology (ASM). 2017-Present
ASM Missouri Valley Branch. 2017-2022
Microbiology and Molecular Genetics Graduate Student Association (MMGGSA), OSU. Secretary. 2018-2019
Nepalese Student Association. OSU. Cultural coordinator and Vice President. 2018-2020.